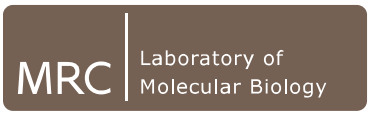
Exploiting network-based approaches for understanding gene regulation and function

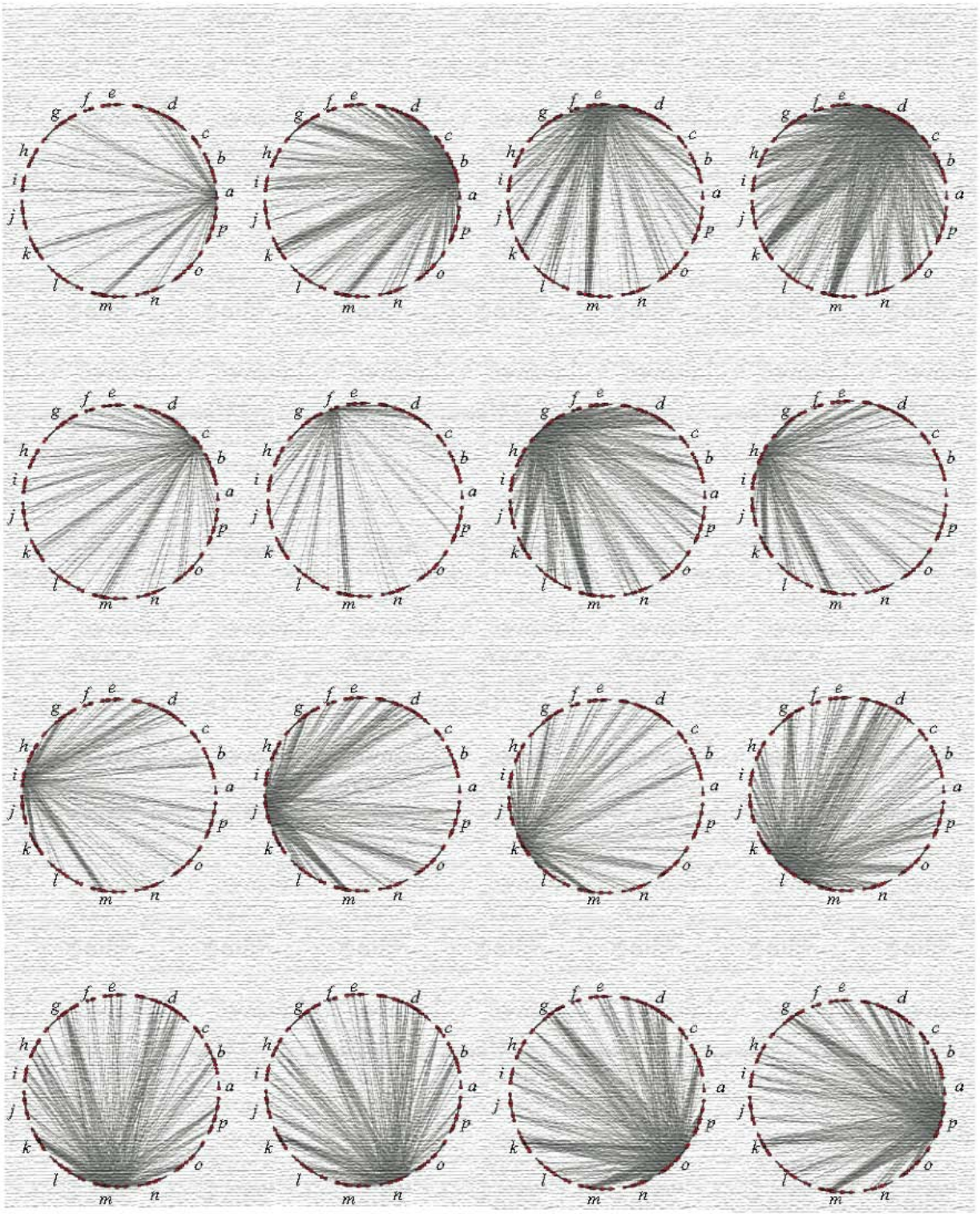
Sarath Chandra Janga

A dissertation submitted to the University of Cambridge in candidature for the degree of Doctorate of Philosophy



April 2010

Darwin College, University of Cambridge
MRC Laboratory of Molecular Biology
Cambridge, United Kingdom


a b c d e f g h i j k l m n o p

Previous page: A portrait of the transcriptional regulatory network of the budding yeast, *Saccharomyces Cerevisiae.* Each circle represents the network of transcriptional interconnections between all other chromosomes to one of the chromosomes. Evidently all chromosomes are transcriptionally controlled by factors encoded on many of the 16 chromosomes in this organism marked by the letters ‘a’ through ‘p’.

Declaration of originality

This dissertation describes work I carried out at the Medical Research Council Laboratory of Molecular Biology in Cambridge between January 2008 and April 2010. The contents are my original work, although much has been influenced by the collaborations in which I took part. I have not submitted the work in this dissertation for any other degree or qualification at any other university.

Sarath Chandra Janga
April, 2010
Cambridge, United Kingdom

Acknowledgements

First of all I would like to express my gratitude to Dr. Madan Babu with out whose continuous support all along my doctoral work, it would have just remained a dream for me to carry out my thesis work at MRC Laboratory of Molecular Biology. Madan has not only been an excellent supervisor but a good friend who was always supportive of my research interests, by allowing me to work independently on a wide range of problems during my stay here. He has been a source of great inspiration on various occasions and a great scientific colleague to work with. In short, I probably could not have had a more understanding and motivating supervisor.

I am also very grateful to Dr. Sarah Teichmann whose equivalently supporting words from time to time have been a motivation to finish my doctoral work in a short time. I have learnt from her the art of adventuring into unchartered territories of molecular biology with out fear.

I am also thankful for the kind support and warm welcome that I received from Dr. Cyrus Chothia from the first day that I came to LMB.

I consider myself very fortunate to be in a wonderful lab with a lot of energetic and highly motivating people working on fundamental problems of molecular biology. Indeed, I must admit that I have learnt at least as much from my colleagues and seminars at LMB, as I have learnt from reading books and papers, not to mention the fun that I had during numerous lunch and dinner breaks with various members of the lab and TCB group in particular. I especially would like to thank A Wuster, B Lang, AJ Venkatakrishnan, D Hebenstreit, D Wilson, E Levy, G Chalancon, J Su, N Mittal, P Kota, R Janky, S De, T Perica, V Charoensawan and J Gsponer for making my stay at LMB a memorable experience.

I am also greatly indebted to all my scientific friends, collaborators and mentors, both in the past and during my PhD, for having helped me learn and adventure diverse areas of molecular biology. In no defined order, I would like to sincerely thank Agustino Martinez-Antonio (Irapuato, Mexico) for his confidence in my abilities, Ernesto Perez-Rueda (Cuernavaca, Mexico) for his kind hospitality during my visits to mexico, Gabriel Moreno-Hagelsieb (Waterloo, Canada) for being a great mentor and an excellent scientific friend, Heladia Salgado (Cuernavaca, Mexico) for her energy and patience to my requests to data, Andrew Emili (Toronto, Canada) for giving me the opportunity to work on an unsolved mystery, Denis Thieffry (Marseille, France) for making me learn to focus on important ideas and many other colleagues for scientific discussions over the years which made me a mature and independent scientist. I would also like to take this opportunity to offer my gratitude to all colleagues, administrative staff and heads of division, Venki Ramakrishnan and Kiyoshi Nagai at LMB whose continuous support have made it possible for me to develop a career in science.

I am also grateful to the financial support that I received from Cambridge Commonwealth Trust (CCT) and the Medical Research Council during my PhD.

Last, but not the least, I am most indebted to my family (my parents and sister) as well as near and dear who have been continuously supportive of my adventures in science and for understanding my reasons to be in silence for months. My very presence on this planet would not have been possible if not for my mother who expired long before I knew what maths and science is all about. I dedicate this thesis on her name.

Abbreviations

3C	Chromosome Confirmation Capture
ArcA	Aerobic respiration control protein A
BDBH	Bi-Directional Best Hits
BLAST	Basic Local Alignment Search Tool
cAMP	cyclic Adenosine MonoPhosphate
ChIP	Chromatin immunoprecipitation
CLIP	Cross Linking and Immuno-Precipitation
COGs	Clusters of Orthologous Groups
CRP	cAMP Receptor Protein
CT	Chromosomal Territory
DBTBS	DataBase of Transcriptional regulation in Bacillus Subtilis
DNA	DeoxyriboNucleic Acid
EC	Enzyme Commission
FDR	False Discovery Rate
FIS	Factor for Inversion Stimulation
FISH	Fluorescent In Situ Hybridization
FFL	Feed Forward Loop
FNR	regulator of Fumarate and Nitrate Reduction
GBA	Guilt By Association
GC	Genomic Context
GO	Gene Ontology
GR	Global Regulator
GRN	Gene Regulatory Network
HMM	Hidden Markov model
hnRNP	heterogeneous nuclear RiboNucleoProtein
HNS	Histone-like Nucleoid Structuring protein
HU	Heat Unstable protein
IHF	Integration Host Factor
LAD	Lamina Associated Domain
LCMS	Liquid Chromatography-Mass Spectrometry
LCR	Locus Control Region
MALDI	Matrix-Assisted Laser Desorption/Ionization
MCL	Markov CLuster algorithm
mRNA	Messenger RNA
NAP	Nucleoid Associated Protein
PAB	PolyAdenylate-Binding protein
PI/PPI	Protein Interactions
PTM	Post-Translational Modification
PTN	Post-Transcriptional Network
PTS	PhosphoTransferase System
RBD	RNA Binding Domain
RBP	RNA Binding Protein
RIP	RNP ImmunoPrecipitation
RNA	RiboNucleic Acid
RNP	RiboNucleo Protein complex
RRM	RNA Recognition Motif
TAP	Tandem Affinity Purification
TF	Transcription Factor
TG	Target Gene
TPI	Target Proximity Index
TRN	Transcriptional Regulatory Network

Summary

It is increasingly becoming clear in the post-genomic era that proteins in a cell do not work in isolation but rather work in the context of other proteins and cellular entities during their life time. This has lead to the notion that cellular components can be visualized as wiring diagrams composed of different molecules like proteins, DNA, RNA and metabolites. These systems-approaches for quantitatively and qualitatively studying the dynamic biological systems have provided us unprecedented insights at varying levels of detail into the cellular organization and the interplay between different processes. The work in this thesis attempts to use these systems or network-based approaches to understand the design principles governing different cellular processes and to elucidate the functional and evolutionary consequences of the observed principles.

Chapter 1 is an introduction to the concepts of networks and graph theory summarizing the various properties which are frequently studied in biological networks along with an overview of different kinds of cellular networks that are amenable for graph-theoretical analysis, emphasizing in particular on transcriptional, post-transcriptional and functional networks.

In Chapter 2, I address the questions, how and why are genes organized on a particular fashion on bacterial genomes and what are the constraints bacterial transcriptional regulatory networks impose on their genomic organization. I then extend this one step further to unravel the constraints imposed on the network of TF-TF interactions and relate it to the numerous phenotypes they can impart to growing bacterial populations.

Chapter 3 presents an overview of our current understanding of eukaryotic gene regulation at different levels and then shows evidence for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation. The results emphasize that specific organization of genes across and within chromosomes that allowed for efficient control of transcription within the nuclear space has been selected during evolution.

Chapter 4 first summarizes different computational approaches for inferring the function of uncharacterized genes and then discusses network-based approaches currently employed for predicting function. I then present an overview of a recent high-throughput study performed to provide a ‘systems-wide’ functional blueprint of the bacterial model, *Escherichia coli* K-12, with insights into the biological and evolutionary significance of previously uncharacterized proteins.

In Chapter 5, I focus on post-transcriptional regulatory networks formed by RBPs. I discuss the sequence attributes and functional processes associated with RBPs, methods used for the construction of the networks formed by them and finally examine the structure and dynamics of these networks based on recent publicly available data. The results obtained here show that RBPs exhibit distinct gene expression dynamics compared to other class of proteins in a eukaryotic cell.

Chapter 6 provides a summary of the important aspects of the findings presented in this thesis and their practical implications.

Overall, this dissertation presents a framework which can be exploited for the investigation of interactions between different cellular entities to understand biological processes at different levels of resolution.

TABLE OF CONTENTS

Chapter 1: Introduction

Chapter 2: Functional, structural and dynamic constraints on bacterial regulatory networks

Chapter 3: Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes

Chapter 4: Uncovering the functional architecture of uncharacterized proteins in *Escherichia coli*

Chapter 5: Structure and dynamics of post-transcriptional regulatory networks directed by RNA-binding proteins

Chapter 6: Conclusions and Perspectives

Appendix

1 Introduction

CONTENTS OF CHAPTER 1

PREAMBLE

Reductionism, which has been the paradigm in biological research for more than a century, has provided us with a wealth of knowledge about the individual cellular components, their functions and mechanisms. Despite its huge success in the last century, post-genomic biology has increasingly made it clear that discrete biological function can only rarely be attributed to an individual molecule. Instead, most biological outcomes in a cell arise from a complex interplay between different cellular entities such as proteins, DNA, RNA and metabolites. Therefore, a key challenge for biology in the twenty-first century is to understand the structure and dynamics of the complex web of interactions in a cell that contribute to its proper functioning. Although, we can not answer this question in full, the analyses, concepts and frameworks outlined in this thesis, will help the scientific community to interpret and better understanding the logic behind the several layers of complex web of interactions happening in the cell.

In the last few years there has been a rapid development in various high-throughput technologies which has lead to the accumulation of a large amount of data from different areas of molecular and cellular biology. These developments together with increasing interest in the community for gaining a systems-wide understanding of the cellular machinery have provided us unprecedented insights into the structure, organization and dynamics of various major cellular processes such as transcription, translation, degradation etc. Likewise, efforts to understand the interaction of the cell with external environment have generated global phenotypic maps such as those due to small-molecule perturbations. Despite the growing amount of data representing each of these processes it should be admitted that none of these cellular processes work in isolation but rather form an integrated network of different wiring diagrams which is responsible for the observed behavior of the cell. In this thesis, I provide evidence that each of these networks of associations associated with a particular cellular process can be studied in detail to provide meaningful insights into how they contribute to the functioning of the cell, factors that constrain their structure and how they influence the genomes on which they are encoded. Nevertheless, an open challenge of the contemporary biology is to integrate these diverse cellular programs to first understand and model in quantitative terms the topological and dynamic properties of such a unified cellular network and then to exploit it for the therapeutic benefit of mankind.

OUTLINE OF THE INTRODUCTION

An emerging notion in post-genomic biology is that cellular components can be visualized as a network of associations between different molecules like proteins, DNA, RNA and metabolites. This has led to the application of network theory and network-based approaches to a wide range of biological problems from understanding regulation of gene expression to prediction of gene's function and phenotype to drug discovery settings. In this chapter, I first introduce the notion of networks and the basic principles of network biology together with an overview of different kinds of networks that are being widely studied in biological sciences at the systems level. In particular, I introduce the transcriptional and post-transcriptional networks in which trans-acting elements like TFs, RBPs and sigma factors form one set of nodes and their target genes or RNAs, of which they control the activity, form the other set of nodes. The links between them which have directionality from the trans-acting elements to their target genes, controlled by their cis-regulatory elements, form a complex and directional network of interactions. In contrast, functional linkage networks constructed in function prediction pipelines typically comprise of undirected networks where all the nodes are treated essentially the same and there is no directionality between nodes. These networks aim to uncover the broad functional role of the uncharacterized genes using the annotations of already characterized members to which they are connected to. I then give a brief overview of other classes of networks such as small-molecule protein interaction networks which are also referred to as the drug-target networks, to extend the generality and applicability of the network-guided approaches in understanding biological systems.

1.1 BASICS OF GRAPH THEORY AND NETWORKS

Complex networks describe a wide range of dynamical systems in nature and society. In simplistic terms, a network comprises of a set of nodes with connections between them called edges. Most real world systems can be visualized in the form of networks also called graphs in mathematical literature. Examples include that of internet, World Wide Web (WWW), social networks of acquaintances between individuals, food webs, metabolic networks, transcriptional networks, signaling networks, neural networks and many others. Although the study of networks, in the form of mathematical graph theory, is one of the fundamental areas of discrete mathematics, much of our understanding about their underlying organizational principles has come to light only recently. While traditionally most complex networks have been modeled as random graphs, it is increasingly recognized that the topology and evolution of real networks are governed by robust design principles.

A number of biological systems ranging from metabolic to neuronal and food webs to ecosystems can be usefully represented as networks. More generally, the behavior of most complex systems emerges from the orchestrated activity of a many components that interact with each other through pairwise interactions. As such at a highly abstract level, the components can be reduced to a series of nodes that are connected to each other by edges, with each edge representing the interactions between two components. The nodes and links together form a network, or in more formal mathematical language, a graph and these definitions can be extended to any sub-system of a complex system under study. Since understanding the network of cellular interactions as a whole is impractical at the moment for at least two major reasons, namely incompleteness of the data representing the wide variety of interactions that are possible in a cell and variations in the mode as well as type of interactions. Theoreticians have been studying networks by dissecting the biological processes into different levels with the most commonly studied being the physical interactions between molecules, such as protein-protein, protein-nucleic acids and protein-metabolite, all of which can be conceptualized using the node-link nomenclature. Nevertheless, more complex functional interactions can also be considered within this representation. A classic example of such a representation is the network of metabolic pathways, where in metabolic substrates and products are connected with directed edges joining them if a known metabolic reaction exists that acts on a given substrate and produces a given product.

Depending on the nature of the interactions, networks can be directed or undirected. In directed networks, the interaction between any two nodes has a well-defined direction, which

represents, for example, the direction of material flow from a substrate to a product in a metabolic reaction or the direction of information flow from a transcription factor to the gene that it regulates. In undirected networks, the links do not have an assigned direction. For example, in protein interaction networks a link represents a mutual binding relationship and hence do not have a directionality in their association.

Another important class of biological networks is the genetic regulatory network. The expression of a gene, i.e., the production by transcription and translation of the protein for which the gene encodes for, can be controlled by the presence of other proteins called transcription factors (TFs) which can control the expression of the gene both positively or negatively. In the former case, TFs are considered to act as activators and in the later as repressors. It is due to the regulatory network the genome can co-ordinate its response to both external and internal stimuli by controlling the expression of thousands of genes in appropriate amounts under appropriate conditions and time. Genetic regulatory networks were in fact one of the first networked dynamical systems for which large-scale modeling attempts were made. The early work on random Boolean nets by Kauffman (Kauffman, 1969; Kauffman, 1971; Kauffman, 1993) is a classic in this field before substantial advance has come more recently. The structure of transcriptional regulatory networks has been the focus of several recent studies (Babu et al., 2004; Farkas, 2003; Guelzim et al., 2002; Janga and Collado-Vides, 2007; Thieffry et al., 1998).

1.1.1 Local level

A number of properties can be defined for a network representation and these properties can be grouped into three major classes namely local, module and global levels. In the following sections, I will summarize the major quantitative properties which can be used to define the structure of complex networks at each of these three levels. The first of them is at the local level and as the name suggests refers to the local properties of a node. For instance, as discussed above, networks can be directed or undirected depending on the nature of the interactions and as such directed networks comprise of both an out-going degree as well as in-coming degree while undirected networks only comprise of one degree associated with their nodes (see Table 1-1 for a list of local properties of networks). Degree or connectivity of a node in a network corresponds to the total number of connections it has with other nodes in the network. As is evident, in directed networks degree or connectivity of a node is the sum of in-coming and out-going degrees. Highly connected nodes i.e, nodes with high degree in biological networks are often referred to as hubs in the network. Degree distribution, P(k), is another property derived from degree of nodes in a network, which gives the probability that a selected node has exactly

Table 1-1. Different local properties which can be defined for a node in complex networks.

Property	Definition
Indegree or incoming degree	In directed networks where directionality of an interaction is taken into account, indegree refers to the number of incoming connections to a node of interest. In other words, indegree is the number of arrows that flow into the node under investigation.
Outdegree or outgoing degree	Out degree refers to the number of edges which start from a node of interest and point to other nodes in the network and is valid for directed networks where there is direction associated with each edge represented.
Degree or Connectivity	Degree or connectivity of a node refers to the total number of interactions it has in a network – the higher the connectivity (i.e., hub nodes) the more the number of targets it interacts with. In directed networks degree simply corresponds to the sum of in and out degrees of a node.
Clustering coefficient	Clustering coefficient of a node reflects the extent to which the neighbors of a given node are interconnected among themselves to what is expected theoretically and indicates the cohesiveness or local modularity of the network. An extension of this metric to the complete network defined as the average clustering coefficient of all nodes, tells whether the network is modular or is sparsely connected.
Betweenness	Betweenness centrality of a node measures the number of shortest paths between all pairs of nodes in the network that pass through a node of interest – the higher the number of paths that pass through a node, the more important it is.
Average path length	Average length of the shortest paths between all pairs of nodes in the network.
Closeness	Closeness centrality is defined as the inverse of the average length of all the shortest paths from a node of interest to all other nodes in the network - note that closeness centrality defined this way implies that higher the closeness value, the higher the importance (centrality) of a node.
Diameter	The diameter of a network is the length of the longest path among all the shortest paths defined between two nodes. It gives an estimation of the distance between the farthest nodes in the network.
Graph density	The density of a network is the ratio of the number of edges to the number of total possible edges.
Power law fit (exponent-alpha)	Fitting a power-law distribution function to the degree distribution of the network to study whether the network is likely to exhibit a scale-free network structure.

k links. P(*k*) is obtained by counting the number of nodes N(*k*) with k=1,2.. links and dividing by the total number of nodes N. The degree distribution allows us to distinguish between different classes of networks. For example, a poissonian degree distribution is seen when P(*k*) is plotted against *k* for random networks indicating that most nodes have roughly equal number of links with little deviation from the average degree of a node in the network. By contrast, a power-law degree distribution indicates that a few nodes interact with numerous other nodes while most interact with rather few nodes (see Global Level).

Another important property at the local level is the clustering coefficient of a node which tells how interconnected are the neighbors of a given node to what is expected if all the neighbors are full connected. Mathematically, it is defined as the ratio of the number of observed links between the neighbors of a node of interest to the total number of feasible links between all the immediate neighbors. Average clustering coefficient of a network calculated as the mean of the clustering coefficients of all the nodes in the network gives a measure of cohesiveness in the network which is also commonly referred to as the extent of modularity. The higher the clustering coefficient greater is the modular nature of the network. To compare the extent of cohesiveness in a network often clustering coefficients of the real networks are compared with random networks with similar size and degree distribution.

So far all the properties which are discussed concern the nodes in the network, however a number of properties have also been defined for edges in a network. Most important of these which needs mention is the path length between two nodes, which refers to the number of edges that one needs to traverse between two nodes of interest. Since there can be many alternative paths between two nodes, the shortest path i.e, the path with the smallest number of links between the selected nodes is often referred to as the path length. In directed networks, the path length between two nodes A and B may not be the same as that between nodes B and A reflecting the directionality in the network. Another important global property which stems from path length is the average or mean path length of a network and refers to the average of all the shortest paths between all pairs of nodes and offers a measure of a network's overall reach.

In addition to the degree of a node which tells how central or important a node in a network is, a number of other centrality measures have also been defined in the literature. These include betweenness and closeness centrality among other less popular definitions (Junker et al., 2006). Betweenness centrality, which is the number of shortest paths going through a node is typically calculated using the brandes algorithm (Brandes, 2001). Closeness, is measured as the inverse of the average length of the shortest paths from a node of interest to all other nodes in the network. Since the centrality measures, betweenness and closeness use the shortest path lengths between all pairs of nodes in a graph, for cases where no path exists between a particular pair of nodes, shortest path length is usually taken as one less than the maximum number of nodes in the graph.

While a number of these properties have been studied in diverse kinds of cellular networks and these will be discussed in the respective chapters or as appropriate, I summarize below some of the observations to give a flavor of their importance in understanding complex networks. Studies on the statistical properties of metabolic networks revealed that the

distributions of the outgoing and incoming degrees have been found to follow power law (Jeong et al., 2000). It was also shown using undirected versions of these metabolic graphs that they have short average path length and a large clustering coefficient (Fell and Wagner, 2000). In protein-protein interaction networks it was shown that the degree distribution follows a power law and that highly connected proteins are more likely to be lethal than lowly connected ones (Jeong et al., 2001) and that links between highly connected proteins tend to be suppressed while those between highly connected and low-connected proteins are abundant, which was proposed as an attribute of cellular networks to attain robustness and decrease cross talk between different functional modules (Maslov and Sneppen, 2002). This property of highly connected proteins avoiding interactions with other highly connected proteins in a network has been referred to as dissociative property. On the other hand, the observation that most real world networks have extremely small average path lengths is referred to as the small world effect (Watts and Strogatz, 1998).

1.1.2 Modular level

Another important level at which network organization is often studied is that of modules. Modules are seen in all kinds of complex systems from groups of friends in social networks, websites that are dedicated to similar topics in the internet, to groups of organisms which survive in a similar niche in an ecological food web. Modules are also evident in several engineered systems, from a simple computer chip to a more sophisticated super computer, where in they are employed to create an order and to organize the tasks dedicated to each of these fundamental units. Likewise, cellular processes have been proposed to be carried out in a highly modular manner (Hartwell et al., 1999). More generally, modules in biological networks refer to a group of genes/proteins or other cellular entities that work together to achieve a common task for the proper functioning of the cell (Alon, 2003; Hartwell et al., 1999; Ravasz and Barabasi, 2003; Ravasz et al., 2002). In fact, there are numerous examples of modules in a cellular context such as protein-protein and protein-RNA complexes which form physical modules or co-expressed gene clusters which work together in a given biochemical process or signaling modules which gather extracellular cues to prepare an organism for variations in the environment. Evidence for the existence of modularity in cellular networks has mostly come from the calculation of average clustering coefficient (see Table 1-1) of a wide variety of networks, which indicates the occurrence of a high number of interconnections between the neighbors of a node of interest. Average clustering coefficient which is the mean of the clustering coefficients of all the nodes is considered a proxy for modularity in networks. In the

absence of modularity, the clustering coefficient of the real and the randomized network are comparable. The average clustering coefficient of most real networks is significantly larger than that of a random network of equivalent size and degree distribution. For instance, existence of modularity defined in this fashion has been convincingly shown for a number of biological networks including metabolic, protein-protein and transcriptional (Guelzim et al., 2002; Ravasz et al., 2002; Wagner, 2001; Wuchty, 2001). Although there is no definitive agreement on how modules in cellular networks can be best identified and what set of genes would constitute a module (Wolf and Arkin, 2003), it is now a common knowledge that most biological systems can be divided into groups of genes which form discrete biological functions. Part of the problem in our ability to precisely determine the components of a module in cellular networks is that biological networks are hierarchical and scale-free structures (Ravasz and Barabasi, 2003; Ravasz et al., 2002) (see below) and therefore modularity in these settings indicates that the network can be split into either many modules each of which containing only few genes or a set of few modules where in each module can harbor many genes. It is therefore intuitive that the hierarchical modular nature of cellular systems naturally permits the definition of a module to be plastic depending the choice of the granularity one wishes to dissect a system into.

The high clustering in the cellular networks indicates that they are generally locally grouped with various subgraphs of highly interconnected groups of nodes forming the core – evidence supporting the occurrence of isolated functional modules. Subgraphs capture specific patterns of interconnections that characterize a given network at the modular level. However, not all subgraphs are equally significant in real networks, as indicated by a series of recent observations (Milo et al., 2002; Shen-Orr et al., 2002). Some subgraphs or patterns of interconnections between nodes in a network appear more often than expected by chance in random networks with the same topology and these are often referred to as network motifs. Motifs in networks are analogous to sequence motifs in a set of homologous sequences which are defined as the patterns of amino acids or DNA stretches which occur more conserved than expected by chance. Different networks have been shown to be abundant for various motifs (Milo et al., 2002). For instance, transcriptional networks have been shown to harbor the Feed-Forward Loops (FFLs) as the most abundant motif while protein interaction networks have been shown to comprise of fully connected cliques i.e, subgraphs in which all the nodes are connected to each other (Shen-Orr et al., 2002; Wuchty et al., 2003). The identification of motifs not only provides information about the type of local interconnections in the network but also allows one to understand their interplay with the rest of the network. Several evidences support the biological relevance for the occurrence of motifs in networks. For example, the high degree

of evolutionary conservation of motif constituents within the yeast protein interaction network and the convergent evolution of motifs observed in the transcription regulatory network of diverse species all support their biological relevance (Conant and Wagner, 2003; Madan Babu et al., 2006; Wuchty et al., 2003).

In case of a transcriptional regulatory network, a module is typically defined as a set of genes that are regulated by a common set of Transcription Factors (TFs). Under this definition, it is intuitive to expect that various cellular processes can be conveniently regulated by discrete and separable modules which can coordinate the activities of many genes and carry out complex functions. Therefore, identifying transcriptional modules is useful for understanding cellular responses to internal and external signals under different cellular conditions. Datasets of genome-wide gene expression and location analysis (ChIP-chip) are frequently used to identify transcriptional modules controlling a variety of cellular processes (Bar-Joseph et al., 2003; Ihmels et al., 2002; Segal et al., 2003; Stuart et al., 2003; Wu et al., 2006). Several of these studies have focused on yeast and other model organisms due to the availability of extensive datasets on gene expression and transcriptional regulatory interactions together with their binding site information. From a computational perspective, typical approaches for module discovery involved the use of clustering and motif-discovery algorithms to gene expression data to find sets of co-regulated genes with variations in methods to include previously known information of cellular functions or promoter sequences. Some studies also used model based approaches such as Bayesian networks to infer modules and understand regulatory network architectures (Segal et al., 2003). Despite several methods which have been developed to identify regulatory modules from expression data, most frequently used implementations take into account that genes co-expressed in similar conditions are likely to belong to the same set of regulatory modules (Ihmels et al., 2004; Ihmels et al., 2002; Segal et al., 2003) while more sophisticated approaches integrate additional data sources like TF binding data, motif information or functional annotation (Bar-Joseph et al., 2003; Ihmels et al., 2002; Pilpel et al., 2001).

Although there have been several different approaches to identifying modules and have provided distinct outcomes in terms of the number and size of the resulting modules, the general consensus has been that regulatory networks are highly interconnected and very few modules are entirely separable from the rest of the network. Therefore, the major conclusion has been that modules are frequently nested within each other in a hierarchical fashion at different levels. In fact, an analysis of the distribution of the commonly seen motifs across the identified modules in transcriptional networks, suggests that network motifs themselves do not

exist in isolation but rather integrate to form part of the modules by sharing some of their edges (Dobrin et al., 2004; Resendis-Antonio et al., 2005). Thus, many small, highly connected motifs group into a few larger modules, which in turn integrate into even larger ones. These nested modules are interconnected through local regulatory hubs. Such an organization not only explains the hierarchical organization, which is seen in other cellular networks (Ravasz and Barabasi, 2003) but also intuitively suggests the capacity for rapid regulatory changes through regulatory hubs, with integration and fine tuning of the regulatory processes by downstream TFs, thereby linking several modules in a hierarchical manner.

As the components of a specific motif often interact with nodes that are outside the motif, it is important to understand how different motifs interact with each other and with the rest of the network for different kinds of networks. While recent work shows that different motifs aggregate to form large motif clusters in transcriptional networks, the generality of these findings is still under debate. However, since motifs are present in all kinds of biological networks that have been examined till date (Milo et al., 2002), it is likely that the aggregation of motifs into motif clusters and modules is a generic property of most biological and real world networks.

1.1.3 Global level

One of the most important developments in our understanding of complex systems is the observation that despite the remarkable diversity in the variety of complex networks in nature, their architecture was found to be governed by a few simple principles. For example, most complex networks have been long believed to follow the degree distributions like that proposed by the Erdos-Renyi model, according to which a plot of the degree distribution, P(*k*), against the degree *k* of a complex network should follow a poisson distribution. However, it is now clear that most real world complex systems including biological networks follow a scale-free topology with a power-law degree distribution where in degree distribution, P(*k*), against the degree *k* on a log-log plot shows a straight line with a negative slope γ which varies between 2 and 3. It has also been shown that in both Erdos-Renyi model as well as scale-free model proposed by Barabasi and Albert (Barabasi and Albert, 1999), distribution of clustering coefficient was found to be independent of the degree (Barabasi and Oltvai, 2004). Nevertheless, a major difference between the two network models is that in the former most nodes have approximately equal number of links with all of them being close to the average degree in the network - indicative of a gaussian/poissonian degree distribution while the later is determined by the presence of a large number of nodes which are poorly connected and a relatively small number of nodes which are highly connected (also referred to as hubs). Due the scaling nature in the degree

distribution Barabasi-Albert or scale-free model exhibits a straight line on a log-log plot between the degree distribution and the degree of a node.

Yet another class of networks which have been proposed in the literature are the hierarchical scale-free networks which comprise of all the properties of scale-free networks and in addition also exhibit a slope of -1 when the distribution of clustering coefficient is plotted against the degree of a node on a log-log scale, indicating an organization where in sparsely connected nodes are part of highly clustered areas, with communication between the different highly clustered neighborhoods being maintained by a few hubs. It is increasingly believed that most real world complex networks obey this hierarchical scale-free modular structure (Ravasz and Barabasi, 2003; Ravasz et al., 2002; Yu and Gerstein, 2006).

Although the hierarchical nature of networks has not been extensively explored for all the cellular networks, there is extensive evidence that most of them including protein-protein, transcriptional regulatory to metabolic linkages at least exhibit a scale-free topology (Giot et al., 2003; Guelzim et al., 2002; Jeong et al., 2001; Wagner, 2001). In such networks, most proteins or cellular entities participate in only a few interactions while a few participate in disproportionately large number of interactions – a signature of scale-free networks with inherent power-law degree distribution. Although a large number of cellular networks have been shown to observe the scale-free topology in the recent years, not all of them are scale-free graphs. For instance, in the case of transcriptional regulatory networks the incoming connectivity which is defined as the number of transcription factors regulating a target gene, which quantifies the combinatorial effect of gene regulation, was observed to follow an exponential distribution in both *Escherichia coli* and *Saccharomyces cerevisiae* (Guelzim et al., 2002; Thieffry et al., 1998). The exponential behaviour indicates that most target genes are regulated by similar number of factors and could reflect the limits on the number of transcription factors that can affect a target gene due to the constraints on the intergenic spacing available and the number of proteins that can simultaneously effect a promoter region. On the other hand, the outgoing connectivity, which is the number of target genes regulated by each transcription factor, was found to be distributed according to a power law, contrary to the incoming connectivity parameter. This is indicative of a hub-containing network structure, in which a select set of transcription factors participate in the regulation of a disproportionately large number of target genes. These hubs can be viewed as ‘global regulators’, as opposed to the remaining transcription factors that can be considered as ‘fine tuners’.

In case of transcriptional regulatory networks it has been shown, by both a top-down and bottom-up approaches for determining hierarchy, that they possess a multi-layer hierarchical

modular structure (Ma et al., 2004; Yu and Gerstein, 2006). Interestingly, transcription networks do not seem to possess feedback regulation at the level of transcription meaning transcriptional regulation of TFs at the top by TFs at the bottom of this hierarchial structure is not frequent, indicating the prevalence for alternative forms of feedback control of transcription. Typically such a feedback occurs through the usage of protein-protein interactions at post-translational level or due to a complex interplay of cellular entities which control the activity of TFs by changing their conformation depending on the continuously varying intra- and extra-cellular conditions (Martinez-Antonio et al., 2006; Yu and Gerstein, 2006). It has also been observed that the TFs in the middle of this hierarchy (often from the levels 2 and 3 measured from the bottom) regulate more direct targets than those at the top suggesting that these middle level TFs act as managers and are indeed control-bottlenecks for cellular transcriptional response (Yu and Gerstein, 2006).

While a number of other properties such as diameter, graph density etc of a network have also been defined in network biology (see Table 1-1) they would not be of immediate relevance to the work discussed in this thesis and hence have not been discussed in detail.

1.2 NETWORKS IN MOLECULAR BIOLOGY

1.2.1 Methods to construct transcriptional regulatory networks

At an abstract level regulatory interactions linking TFs to their transcriptionally controlled target genes (TGs) in an organism can be viewed as a directed graph, in which the TFs and TGs represent the nodes while the regulatory interactions that connect them as the edges. Typically the resulting network is a complex, hierarchical, multilayered graph that can be studied at several levels of detail. However at a more fundamental level the organization of transcriptional regulatory machinery and the principles involved are considerably different in the two major kingdoms of life, bacteria and eukarya. In bacteria, transcription and translation happen in the same compartment i.e cytoplasm and transcriptional control can be considered to be mostly at the DNA sequence level through the use of cis-regulatory elements and organization of contiguous genes on the same strand of DNA into operons. However in eukaryotic genomes, the process of transcriptional regulation is highly complex and is co-ordinated at three major hierarchical levels. The first is at the DNA sequence level, i.e. the linear organization of transcription units and regulatory sequences. Co-regulated genes organized into clusters in the genome constitute part of these individual functional units. The second is at the chromatin level, which allows switching between different functional states, i.e between a state that suppresses

transcription and one that is permissive for gene activity. This level involves the changes in the chromatin structure that are controlled by the interplay between histone modification, DNA methylation, and a variety of repressive and activating mechanisms. This regulatory level is linked with the control mechanisms from level one that switch individual genes in the cluster to on and off, depending on the properties of the promoter. The third level is the nuclear level, which includes the dynamic 3D spatial organization of the genome inside the cell nucleus. The nucleus is structurally and functionally compartmentalized and epigenetic regulation of gene expression may involve repositioning of loci in the nucleus through changes in large-scale chromatin structure. All these differences add a layer of complexity and sophisticated control to the inherent structure, functionality and dynamics of transcriptional networks in eukarya in comparison to their bacterial counterparts. Despite these fundamental differences several basic principles in their organization and structure from a network perspective have been shown to be similar in both the kingdoms (Guelzim et al., 2002; Lee et al., 2002; Milo et al., 2002; Shen-Orr et al., 2002; Thieffry et al., 1998; Yu and Gerstein, 2006).

Despite enormous interest in understanding transcriptional networks across organisms our knowledge on transcriptional interaction graphs for a genome has been very limited and is mostly restricted to model organisms like *Escherichia coli* and *Saccharomyces cerevisiae* for which extensive information is available (Gama-Castro et al., 2008; Lee et al., 2002). Transcriptional interactions in an organism have been traditionally identified from small scale assays which are documented in regulatory network databases through extensive manual curation efforts (Baumbach et al., 2007; Gama-Castro et al., 2008; Makita et al., 2004; Matys et al., 2006) or are obtained from high-throughput screens like ChIP-chip or ChIP-seq which allow the identification of regulatory interactions for a vast set of TFs in an organism (Grainger et al., 2005; Lee et al., 2002). Yet another lower resolution high-throughput approach to screen in the whole genome, targets for a TF, is through the knock-out of TF genes and performing a whole genome microarray expression analysis (Devaux et al., 2001). Table 1-2 summarizes a list of these frequently employed low and high-throughput experimental techniques for the identification of regulatory interactions in an organism in an unambiguous manner.

(Space left for an enhanced layout of the table)

Table 1-2. Different low and high-throughput strategies for studying and probing protein-DNA interactions. High-throughput technologies such as ChIP-chip, ChIP-seq and PBMs are frequently employed for the elucidating of regulatory networks on a genome-wide scale.

Method	Description
Band shift	Since DNA molecules are more flexible than proteins, they tend to exhibit much higher mobility in a polyacrylamide gel. Thus, under favourable conditions, free DNA can be distinguished from DNA bound to proteins due to the difference in molecular weight (Garner and Revzin, 1981).
DNA footprinting	In DNA footprinting, a 5' end labeled double stranded DNA is partially degraded by DNAase both in the presence and absence of the TF. Degraded fragments are then loaded on to a gel to visualize by autoradiography. Since the region where the protein has bound the DNA will be protected from DNAase, no fragments are seen in those regions. Therefore, by comparing lanes, one can identify the binding site (Galas and Schmitz, 1978).
FRET based binding site identification	In this method a library of double stranded DNA with one of the two fluorophores attached to its end is used. Protein binding to two pieces of DNA , one from each library where each comprises half of the binding site's sequence, induces FRET signal which can then be used to find protein bound to DNA (Heyduk and Heyduk, 2002).
Binding site detection using unnatural base analog	In this approach a library of DNA sequences with an unnatural base analog (one for each base) is used. Following selection for protein-bound DNA molecules, the DNA is cleaved specifically at the modified base. The site of incorporation can be identified by gel electrophoresis by running fragments generated from unbound sample next to the fragments generated from the bound sample. Since the presence of an analog in the binding site impedes protein binding, this results in a depletion of the protein-bound pool (Storek et al., 2002).
(ChIP-chip) and (ChIp-seq) techniques	The DNA binding protein is tagged with an epitope and is expressed in a cell. The bound protein is covalently linked to DNA by using an in vivo cross-linking agent such as formaldehyde. After cross-linking, DNA is sheared and the protein–DNA complex is pulled down using an antibody for the tag. Reversal of the cross-link releases the bound DNA, allowing the sequence of the fragments to be determined by hybridization to a microarray (ChIP-chip) or by sequencing (ChIP-seq). In ChIP-chip experiments, intergenic regions are spotted on to a microarray chip. Following a chromatin immunoprecipitation step, the bound fragments are reverse cross-linked and hybridized onto the microarray chip (Lee et al., 2006). In ChIP-seq experiments, the bound fragments are directly sequenced using 454/Solexa/Illumina sequencing technology. The sequences are then computationally mapped back to the genome sequence (Johnson et al., 2007).
DNA adenine methyl transferase Identification (DamID)	In DamID technique, protein of interest is fused to an *E. coli* protein, DNA adenine methyl transferase (Dam). Dam methylates the N6 position of the adenine in the sequence GATC, which occurs at reasonably high frequency in any genome (1 site in 256 bases). Upon binding DNA, the Dam protein preferentially methylates adenine in the vicinity of binding. Subsequently, the genomic DNA is digested by the DpnI and DpnII restriction enzymes that cleave within the non-methylated GATC sequence, and remove fragments that are not methylated. The remaining methylated fragments are amplified by selective PCR and quantified using a microarray (Greil et al., 2006).
Protein binding universal DNA microarrays (PBMs)	This is an invitro method to probe protein–DNA interactions. A DNA binding protein of interest is epitope tagged, purified and bound directly to a double-stranded DNA microarray spotted with a large number of potential binding sites. Labeling with fluorophore conjugated antibody for the tag allows detection of binding sites from the significantly bound spots (Bulyk et al., 2004).

1.2.2 Methods to construct functional linkage networks

Traditionally function of a protein was defined using a number of low-throughput approaches like mutagenesis of residues or whole proteins which allowed the identification of the phenotypes for follow up analysis. However, it is increasingly becoming clear that this rational is limited in its ability to infer the function of proteins; failing for those which exhibit mild phenotype or those which are not expressed under standard experimental conditions. In addition, since most proteins associate dynamically with a number of other cellular entities during their life time, the traditional notion of identifying function of a protein by isolating it from the rest of the cellular machinery can be misleading for a majority. This notion followed by the availability of experimentally determined protein-protein interaction maps for diverse model organisms have given rise to the use of these datasets for delineating the biological processes, pathways and complexes that proteins take part in (Aranda et al., ; Bader et al., 2003; Breitkreutz et al., 2008). Indeed, there is now observable overlap and informative variation between different types of low- and high-throughput experiments (Shoemaker and Panchenko, 2007) which provides a convincing reason for exploiting them as complementary approaches in unraveling the functions of proteins. Indeed, recent years have seen an explosion in the number of methods and databases which provide functional associations (both direct physical and indirect contextual interactions) between proteins using both experimental and computational means. I present an extensive list of these resources in Table 4-2 of Chapter 4, where in I also provide a more in depth discussion of network-based approaches for function prediction.

Briefly, experimental approaches employed for constructing functional association networks mostly comprise of data from protein-protein interaction screens followed by co-expression networks comprising of gene pairs showing significant correlation in their expression profiles across conditions, derived from microarray datasets (Luo et al., 2007; Ruan et al., ; Wang et al., 2009). More recently, genetic interactions- measuring the fitness defects of the double mutants compared to that of the individual mutants, are also being employed for constructing these functional linkage networks (Butland et al., 2008; Costanzo et al.). These high-throughput experimental approaches not only increase the confidence of an association but also give cellular context of the protein providing complementary view to the traditional functional prediction paradigm.

In addition to the experimental methods, several computational methods have been proposed for constructing protein-protein associations from sequence data alone. These include the genome context methods namely gene fusion, gene cluster or gene order conservation,

operon arrangements and protein phylogenetic profiles. The gene fusion approach tries to detect the fusion of two genes into a single protein coding gene in one of the sequenced genomes and thereby links them as a strong functional association (Enright et al., 1999; Marcotte et al., 1999a). The method of gene order conservation aims to identify pairs of genes which consistently show a tendency to cluster in immediate vicinity in a number of genomes-suggesting a strong functional link in prokaryotic genomes which are abundant in operons (Dandekar et al., 1998; Overbeek et al., 1999). The method of operon rearrangement tries to identify a link between any pair of genes on a genome as long as their orthologs are predicted to be organized in an operon with a high confidence in at least one sequenced genome (Janga et al., 2005; Rogozin et al., 2002; Snel et al., 2002). The power of this approach depends on the predictive quality of operon prediction methods which have been shown to reach ~90% accuracy in most sequenced genomes (Brouwer et al., 2008; Moreno-Hagelsieb and Collado-Vides, 2002). Yet another approach not based on genomic proximity is phylogenetic profiles. In this method a vector of presence/absence profile of a gene across all the analyzed genomes is constructed and compared to identify genes which show the most correlated profiles, as a measure of functional link. The rational here is that two proteins showing similar profiles i.e, coordinated in their evolutionary gain and loss, are expected to be functionally related (Gaasterland and Ragan, 1998; Pellegrini et al., 1999). Modified versions of this approach take into account the phyogenetic signal of the genomes employed and/or the redundancy in the genome sequence information (Barker and Pagel, 2005; Date and Marcotte, 2003; Moreno-Hagelsieb and Janga, 2008).

Recently, the integration of different types of interaction data into genome-wide functional linkage maps has gained much popularity for functional inference as these integrated maps not only boost coverage but also confidence of an association when assessing protein function. One of the first studies which demonstrated the power of integrating different types of interaction data was by Marcotte and colleagues where they have put together diverse kinds of computational genome context inferences (Marcotte et al., 1999b). This was followed by a number of other methods such as those implemented in the STRING and PROLINKS databases, among other focused studies (Bowers et al., 2004; Hu et al., 2009; Jensen et al., 2009; Massjouni et al., 2006). Typically, in these networks edge weights correspond to the integrated interaction probability values obtained by first scoring each of the methods independently against a set of gold standard interactions, which are then used in a bayesian fashion assuming the scores obtained in each method are independent of each other. More complex methods take into account the dependence and correlation between methods to

develop a regression model for scoring the integrated interactome (Linghu et al., 2008; Zhao et al., 2008). Nevertheless, all of them boil down to constructing a network with either weighted or unweighted edges which are then used for propagating annotations to uncharacterized members using network-based approaches discussed in Chapter 4.

1.2.3 Methods to construct post-transcriptional regulatory networks

Gene expression is a highly controlled process which is known to occur at several levels in eukaryotic organisms. Although traditionally messenger RNAs have been viewed as passive molecules in the pathway from transcription to translation there is increasing evidence that their metabolism is controlled by a class of proteins called RNA-binding proteins (RBPs) (Glisovic et al., 2008; Keene, 2007; Mata et al., 2005). In eukaryotes, since transcription and translation occur in different compartments, it allows for a plethora of options to control RNA at the post-transcriptional level, including their splicing, polyadenylation, transport, mRNA stability, localization and translational control (Glisovic et al., 2008; Keene, 2007). Although some early studies revealed the involvement RBPs in the transport of mRNA from nucleus to the site of their translation, increasing evidence now suggests that RBPs regulate almost all of the post-transcriptional steps.

Development of several high throughput approaches has increased the amount of data for targets of RBPs in diverse organisms (See Table 5-3 in Chapter 5 for a detailed overview of these methods and techniques). These techniques have not been discussed here to avoid redundancy. This data of RBPs and their targets could be utilized to construct RBP-RNA interaction network which is also typically referred to as post-transcriptional regulatory network. This post-transcriptional network is represented in the form of a directional network with each edge corresponding to a regulatory link between the nodes (RBP and the target RNA) similar to directed networks discussed above for transcriptional regulatory networks. In this directed network, one set of nodes are RBPs forming the regulatory proteins while the other set of nodes are RNAs encoded by either protein-coding or non-protein coding genes referred to as the target nodes. These two nodes (regulator node and target node) are joined by an arrow starting from regulator node and directing towards target node. The target RNA may belong to diverse functional proteins including other RBPs. This network can also contain loops as a link starting from RBP and targeting itself, typically referred to as autoregulation of an RBP. This loop structure suggests that RBP can bind to its own RNA and control its metabolism at transcript level. There are several examples suggesting the auto-regulation of RBPs at post-transcriptional level. For instance, in humans, RBPs such as AUF1, HuR, KSRP, NF90, TIA-1

and TIAR were reported to associate with their own mRNA and other RBPs (Pullmann et al., 2007).

Due to the availability of the network of post-transcriptional interactions for a considerable fraction of RBPs in model systems such as *S. cerevisiae* (Hogan et al., 2008), it has become possible to address several questions concerning the structure and organization of post-transcriptional networks directed by RBPs. Chapter 5 focuses on studying these properties by directly analyzing the currently available post-transcriptional regulatory network in the budding yeast.

1.2.4 Methods to construct other classes of cellular and biological networks

Development of several high throughput approaches in the last decade have not only increased the amount of information that we could gather to reveal important insights on the transcriptional, post-transcriptional or functional organization of an organism but they have also enabled us to start our journey to uncover the principles which hold them together. This is mainly because of the extent of information that has been possible to be collected by interrogating the cell's environment at different levels of detail. For instance, availability of modern techniques now enable us to identify the set of protein-protein interactions, genetic interactions, metabolic maps and small molecule interactions at a whole-organism level. While a complete discussion of all the methods and techniques used to identify their respective interactomes is beyond the scope of this thesis. I outline below some of the commonly employed approaches for identifying the interaction graphs for each of these types of interactions occurring in the cell.

Perhaps the most common form of interaction graphs which have been studied since the early days of genome sequencing are protein interactions. A number of approaches for studying them have been reported in the literature and these include the yeast two hybrid (Y2H) (Fields and Song, 1989), protein fragment complementation assay (PCA) (Pelletier et al., 1998), affinity purification coupled with mass spectrometry (AP-MS) (Babu et al., 2009a; Babu et al., 2009b; Gavin et al., 2002), protein chips (Fasolo and Snyder, 2009; Kung and Snyder, 2006), phage display (McCafferty et al., 1990), fluorescence energy transfer (FRET) (Jares-Erijman and Jovin, 2003) and surface plasmon resonance (SPR) (Slavik and Homola, 2006). For a more extensive discussion on the protocols and methods for identifying protein interactions as well as for new developments in this area the reader is referred to recent reviews (Levy and Pereira-Leal, 2008; Shoemaker and Panchenko, 2007).

Another class of networks which are commonly studied is that of metabolic networks. They comprise of representing the metabolites and enzymes involved in catalyzing metabolic reactions as the nodes and edges in a directed network. Most of the work on understanding metabolic networks relies on either manually curated or semi-automated metabolic databases such as the kyoto encyclopedia of genes and genomes (KEGG) and Metacyc which are available for a wide range of model organisms (Caspi et al., 2008; Grossetete et al., ; Kanehisa et al., 2008). In addition to the metabolic maps available for diverse organisms, several groups also study and compile the metabolic reactions for a model organism of interest which are then used for follow up analysis of the metabolic circuitry (Duarte et al., 2007; Durot et al., 2009; Ma et al., 2007).

Organisms respond to continuous variations in internal and external cellular conditions by orchestrating their responses depending on the environmental challenges they are faced with. This involves the usage of a complex network of interactions among different proteins, RNA, metabolites and several other cellular entities, which undergo rewiring when perturbed by small molecules such as chemicals or drugs. The interaction between different chemicals and cellular entities can be represented in the form of a network- so called Drug-Target network. Recent years have seen the development of a number of approaches both computational and experimental for the identification and elucidation of the molecular targets of a drug on a genomic scale (Apsel et al., 2008; Brewerton, 2008; Fabian et al., 2005; Hillenmeyer et al., 2008; Ho et al., 2009; Jacob and Vert, 2008; Kuhn et al., 2008; Paolini et al., 2006; Whitehurst et al., 2007; Yamanishi et al., 2008). This cellular target space which contains the targets of drugs, can be considered to predominantly comprise of three components namely protein-protein, metabolic and transcriptional interaction networks. While the vast majority of the drugs target the protein-protein and metabolic components, limited number of targets have been identified till date for the transcriptional pool (Brennan et al., 2008; Goh et al., 2007; Lage et al., 2007; Lee et al., 2008; Yildirim et al., 2007). Indeed, most common therapeutic targets for established drugs belong to either protein kinase or receptor families with enzymes and ion channels forming the second most predominant class of targets (Wishart et al., 2008). This explains the reasons for the increased attention towards understanding the biophysics of protein-protein contacts in the context of drug targets as these protein classes form major players in protein-protein interactions (Archakov et al., 2003). Table 1-3 shows different methods which are used for the construction of Drug-Target networks and can be broadly classified into genetics-based, proteomics-based and knowledge-driven approaches.

Table 1-3. Different methods available for identifying drug-targets on a genomic scale. Methods can be broadly classified into Proteomics-based, Genetics-based and Knowledge-driven. Although most methods traditionally are based on experimental screening, there is an increase in the number of computational techniques available for small molecule target discovery (grouped as knowledge-driven approaches).

Proteomics-based Methods	Description
Activity based protein profiling (ABPP) (Speers and Cravatt, 2004)	This is a functional proteomic technology that uses chemical probes that react with mechanistically related classes of enzymes. The basic unit of ABPP is a probe that typically consists of a reactive group (electrophile or a photoreactive group) that covalently binds to the active site of an enzyme (nucleophilic residue) and a tag. The tag can either be a reporter (i.e. fluorophore, radioactive group) or a handle (i.e. affinity tags such as biotin). A tag-free strategy for activity-based protein profiling has also been introduced that utilizes the copper(I)-catalyzed azide-alkyne cycloaddition reaction (click chemistry) and gives the advantage of not interfering with biological activity or binding affinities of the probes. The activity-based protein profiling and multidimensional protein identification technologies (ABPP-MudPIT) can provide profiling of inhibitor selectivity, as the potency of an inhibitor can be tested against hundreds of targets simultaneously. (Jessani et al., 2005)
Affinity chromatography (Katayama and Oda, 2007)	This is a protein separation method based on the interaction between target proteins and specific immobilized ligands. Traditionally, the ligand is tethered on a solid support via a spacer arm followed by the addition of a cellular lysate or tissue extract. Only target proteins binding tightly to the ligand are selectively purified, eluted off (denaturation or competition with free ligand) and subsequently identified by mass spectroscopy. To minimize the identification of nonspecifically bound proteins, the protein profile that is obtained with an inactive ligand analogue is also determined and compared with the relevant profile, determined with the desired analogue. More recently, an improved method for the identification of proteins that can bind to small-molecules and drugs has been established which uses quantitative mass spectrometry (MS)-based proteomics (utilizing stable isotope labeling with amino acids in cell culture (SILAC)) and affinity chromatography. (Ong et al., 2009)
Microarrays (Kingsmore, 2006; Ma and Horiuchi, 2006; Salcius et al., 2007; Wingren and Borrebaeck, 2006)	Microarrays in drug target discovery provide miniaturized high-throughput tools to study binding of specific molecules to immobilized proteins or small molecules. In protein microarrays, different recombinant proteins or antibodies that are immobilized on a solid substrate are exposed to a drug solution to identify the target protein(s) which can bind to the small molecule. In chemical microarrays, immobilized drug compounds can be screened for candidate drug-target interactions with purified proteins (Ma and Horiuchi, 2006). When the target protein is known, small molecule arrays can be also used to identify off-target interactions that could have implications for side-effects.
Genetics-based Methods	**Description**
Synthetic lethality/ Gene knock-out (Hillenmeyer et al., 2008; Ho et al., 2009)	Single gene knock-out strains on a genomic scale or for a selected set are exposed to small molecules at different concentrations to evaluate the fitness defects and fitness levels are compared to wild-type populations exposed to the same conditions. This provides an easy means to identify targets on a large scale. (Hillenmeyer et al., 2008; Ho et al., 2009)
RNAi	RNA interference pathways in mammalian systems are used for silencing genes and similar approaches as above are employed to study the fitness defects of cell lines to identify potential drug targets in higher eukaryotes (Turner et al., 2008; Whitehurst et al., 2007)

Knowledge-driven approaches	Description
Literature derived interactions. (Chen et al., 2008; Frijters et al., 2007; Tsui et al., 2007; Yildirim et al., 2007)	In these approaches, manually curated set of interactions are obtained from the literature to generate high confidence set of drug-target relationships to either study their overall structure (Yildirim et al., 2007) or focus on specific disease of interest. (Chen et al., 2008; Frijters et al., 2007; Tsui et al., 2007; Yildirim et al., 2007)
Network-based approaches. (Apsel et al., 2008; Hopkins, 2008)	In these approaches, literature derived interactions are exploited to predict new interactions based on the principles governing the structure of the networks so that new disease targets are identified using comparative genomics or other informatics-based methods followed possibly by experiments to improve the chemicals. (Apsel et al., 2008; Hopkins, 2008)
in silico chemogenomics (Rognan, 2007)	In predictive chemogenomics one predicts relationships between genes/proteins and compounds. In silico approaches that are used can be classified into ligand-based approaches (ligand comparison for target prediction), target-based approaches (target comparison for ligand prediction) or ligand-target based approaches (Rognan, 2007)

1.3 OUTLINE OF THE THESIS

Now that I have summarized the different tools which can quantify the networks at varying levels and the numerous kinds of interaction graphs operating in the cell there are enormous possibilities to understand a cell's internal organization and dynamics. In the following chapters, I will attempt to address some of these open questions.

In particular in Chapter 2, I address the questions, how and why are genes organized on a particular fashion on bacterial genomes and what are the constraints bacterial transcriptional regulatory networks impose on their genomic organization. I then extend this one step further to unravel the constraints imposed on the network of TF-TF interactions and relate it to the numerous phenotypes they can impart to growing bacterial populations.

In contrast to prokaryotes, regulation of gene expression in eukaryotes is much more complex and is known to occur at many different levels even at the stage of transcription. In Chapter 3, I first present an overview of our current understanding of eukaryotic gene regulation at different levels and then present evidence for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation. These results demonstrate that specific organization of genes across and within chromosomes that allowed for efficient control of transcription within the nuclear space has been selected during evolution.

Determining the functions of proteins encoded by genome sequences represents a major challenge in contemporary biology. With traditional methods for annotation of a genome reaching their saturation there is an increasing need to develop alternate and complementary

approaches for solving the genomic function prediction challenge. As a result, alternate computational methods for inferring the protein function such as those which exploit the context of a protein in protein association networks have come to be sought after. These network-based approaches aim to integrate diverse kinds of functional interactions as a means of boosting coverage as well as confidence level of an association. In Chapter 4, I first present an overview of different computational approaches for inferring the function of uncharacterized genes and discuss network-based approaches currently employed for predicting function. I then summarize a recent high-throughput study performed to provide a 'systems-wide' functional blueprint of the bacterial model, *Escherichia coli* K-12, with insights into the biological and evolutionary significance of previously uncharacterized proteins. Given the volume of high-throughput data that is being reported for understanding diverse model systems, the network-based approaches presented here would become be a useful addition to unravel the functions of an increasing number of uncharacterized proteins accumulating in the genomic databases.

While control of gene expression in eukaryotes first occurs at the level of transcription, there is accumulating evidence that RNA-binding proteins play major roles in controlling the expression of a protein by regulating expression at post-transcriptional level. In Chapter 5, I attempt to provide a comprehensive overview and preliminary insights on this rapidly developing area of post-transcriptional regulatory networks formed by RBPs. I discuss the sequence attributes and functional processes associated with RBPs, methods used for the construction of the networks formed by them and finally discuss the structure and dynamics of these post-transcriptional networks based on recent publicly available data. The results obtained from this study show that RBPs exhibit distinct gene expression dynamics compared to other class of proteins in a eukaryotic cell and that these properties are also reflected from an analysis of the post-transcriptional networks formed by them.

In Chapter 6, I first summarize the key findings of all the above chapters and then discuss their broader implications in light of recent findings.

REFERENCES

Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science* **301**, 1866-7.

Apsel, B., Blair, J. A., Gonzalez, B., Nazif, T. M., Feldman, M. E., Aizenstein, B., Hoffman, R., Williams, R. L., Shokat, K. M. and Knight, Z. A. (2008). Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nature Chemical Biology* **4**, 691-699.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J. et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-31.

Archakov, A. I., Govorun, V. M., Dubanov, A. V., Ivanov, Y. D., Veselovsky, A. V., Lewi, P. and Janssen, P. (2003). Protein-protein interactions as a target for drugs in proteomics. *Proteomics* **3**, 380-91.

Babu, M., Butland, G., Pogoutse, O., Li, J., Greenblatt, J. F. and Emili, A. (2009a). Sequential peptide affinity purification system for the systematic isolation and identification of protein complexes from Escherichia coli. *Methods Mol Biol* **564**, 373-400.

Babu, M., Krogan, N. J., Awrey, D. E., Emili, A. and Greenblatt, J. F. (2009b). Systematic characterization of the protein interaction network and protein complexes in Saccharomyces cerevisiae using tandem affinity purification and mass spectrometry. *Methods Mol Biol* **548**, 187-207.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14**, 283-91.

Bader, G. D., Betel, D. and Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. et al. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**, 1337-42.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-12.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-13.

Barker, D. and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* **1**, e3.

Baumbach, J., Wittkop, T., Rademacher, K., Rahmann, S., Brinkrolf, K. and Tauch, A. (2007). CoryneRegNet 3.0--an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and Escherichia coli. *J Biotechnol* **129**, 279-89.

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O. and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**, R35.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* **25**, 163-177.

Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V. et al. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40.

Brennan, P., Donev, R. and Hewamana, S. (2008). Targeting transcription factors for therapeutic benefit. *Mol Biosyst* **4**, 909-19.

Brewerton, S. C. (2008). The use of protein-ligand interaction fingerprints in docking. *Curr Opin Drug Discov Devel* **11**, 356-64.

Brouwer, R. W., Kuipers, O. P. and van Hijum, S. A. (2008). The relative value of operon predictions. *Brief Bioinform* **9**, 367-75.

Bulyk, M. L., McGuire, A. M., Masuda, N. and Church, G. M. (2004). A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. *Genome Res* **14**, 201-8.

Butland, G., Babu, M., Diaz-Mejia, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O. et al. (2008). eSGA: E. coli synthetic genetic array analysis. *Nat Methods.* **5**, 789-95.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C. et al. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623-31.

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M. and Friedman, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* **15**, 87-98.

Conant, G. C. and Wagner, A. (2003). Convergent evolution of gene circuits. *Nat Genet* **34**, 264-6.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S. et al. The genetic landscape of a cell. *Science* **327**, 425-31.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-8.

Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-62.

Devaux, F., Marc, P., Bouchoux, C., Delaveau, T., Hikkel, I., Potier, M. C. and Jacq, C. (2001). An artificial transcription activator mimics the genome-wide properties of the yeast Pdr1 transcription factor. *EMBO Rep* **2**, 493-8.

Dobrin, R., Beg, Q. K., Barabasi, A. L. and Oltvai, Z. N. (2004). Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics* **5**, 10.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R. and Palsson, B. O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* **104**, 1777-82.

Durot, M., Bourguignon, P. Y. and Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* **33**, 164-90.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

Fabian, M. A., Biggs, W. H., 3rd, Treiber, D. K., Atteridge, C. E., Azimioara, M. D., Benedetti, M. G., Carter, T. A., Ciceri, P., Edeen, P. T., Floyd, M. et al. (2005). A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol* **23**, 329-36.

Farkas, I. J., Jeong, H., Vicsek, T., Barabasi, A.-L., and Oltvai, Z.N.,. (2003). The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae. *Physica A* **381**, 601-612.

Fasolo, J. and Snyder, M. (2009). Protein microarrays. *Methods Mol Biol* **548**, 209-22.

Fell, D. A. and Wagner, A. (2000). The small world of metabolism. *Nat Biotechnol* **18**, 1121-2.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.

Frijters, R., Verhoeven, S., Alkema, W., van Schaik, R. and Polman, J. (2007). Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics* **8**, 1521-34.

Gaasterland, T. and Ragan, M. A. (1998). Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics* **3**, 199-217.

Galas, D. J. and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157-70.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H. et al. (2008). RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**, D120-4.

Garner, M. M. and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* **9**, 3047-60.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E. et al. (2003). A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-36.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-86.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabasi, A. L. (2007). The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-90.

Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. and Busby, S. J. (2005). Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. *Proc Natl Acad Sci U S A* **102**, 17693-8.

Greil, F., Moorman, C. and van Steensel, B. (2006). DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. *Methods Enzymol* **410**, 342-59.

Grossetete, S., Labedan, B. and Lespinet, O. FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics* **11**, 81.

Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31**, 60-3.

Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.

Heyduk, T. and Heyduk, E. (2002). Molecular beacons for detecting DNA binding proteins. *Nat Biotechnol* **20**, 171-6.

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., Proctor, M., St Onge, R. P., Tyers, M., Koller, D. et al. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362-5.

Ho, C. H., Magtanong, L., Barker, S. L., Gresham, D., Nishimura, S., Natarajan, P., Koh, J. L., Porter, J., Gray, C. A., Andersen, R. J. et al. (2009). A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol* **27**, 369-77.

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. and Brown, P. O. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6**, e255.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4**, 682-690.

Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P. et al. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol* **7**, e96.

Ihmels, J., Bergmann, S. and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993-2003.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-7.

Jacob, L. and Vert, J. P. (2008). Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **24**, 2149-56.

Janga, S. C. and Collado-Vides, J. (2007). Structure and evolution of gene regulatory networks in microbial genomes. *Res Microbiol* **158**, 787-94.

Janga, S. C., Collado-Vides, J. and Moreno-Hagelsieb, G. (2005). Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res* **33**, 2521-30.

Jares-Erijman, E. A. and Jovin, T. M. (2003). FRET imaging. *Nat Biotechnol* **21**, 1387-95.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412-6.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-2.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651-4.

Jessani, N., Niessen, S., Wei, B. Q., Nicolau, M., Humphrey, M., Ji, Y., Han, W., Noh, D. Y., Yates, J. R., 3rd, Jeffrey, S. S. et al. (2005). A streamlined platform for high-content functional proteomics of primary human specimens. *Nat Methods* **2**, 691-7.

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-502.

Junker, B. H., Koschutzki, D. and Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **7**, 219.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-4.

Katayama, H. and Oda, Y. (2007). Chemical proteomics for drug discovery based on compound-immobilized affinity chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* **855**, 21-7.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* **22**, 437-67.

Kauffman, S. A. (1971). Gene regulation networks: A theory for their structure and global behaviour, in A. Moscana and A. Monroy (eds.), Current Topics in Developmental Biology 6, pp.145-182. New York: Academic Press.

Kauffman, S. A. (1993). The Origins of Order. Oxford: Oxford University Press.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-43.

Kingsmore, S. F. (2006). Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nat Rev Drug Discov* **5**, 310-20.

Kuhn, M., Campillos, M., Gonzalez, P., Jensen, L. J. and Bork, P. (2008). Large-scale prediction of drug-target relationships. *FEBS Lett* **582**, 1283-90.

Kung, L. A. and Snyder, M. (2006). Proteome chips for whole-organism assays. *Nat Rev Mol Cell Biol* **7**, 617-22.

Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tumer, Z., Pociot, F., Tommerup, N. et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309-16.

Lee, D. S., Park, J., Kay, K. A., Christakis, N. A., Oltvai, Z. N. and Barabasi, A. L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* **105**, 9880-5.

Lee, T. I., Johnstone, S. E. and Young, R. A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* **1**, 729-48.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804.

Levy, E. D. and Pereira-Leal, J. B. (2008). Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol* **18**, 349-57.

Linghu, B., Snitkin, E. S., Holloway, D. T., Gustafson, A. M., Xia, Y. and DeLisi, C. (2008). High-precision high-coverage functional inference from integrated data sources. *BMC Bioinformatics* **9**, 119.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K. and Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* **8**, 299.

Ma, H. and Horiuchi, K. Y. (2006). Chemical microarray: a new tool for drug screening and discovery. *Drug Discov Today* **11**, 661-8.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O. and Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**, 135.

Ma, H. W., Buer, J. and Zeng, A. P. (2004). Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**, 199.

Madan Babu, M., Teichmann, S. A. and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* **358**, 614-33.

Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004). DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res* **32**, D75-7.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-3.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6.

Martinez-Antonio, A., Janga, S. C., Salgado, H. and Collado-Vides, J. (2006). Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. *Trends Microbiol* **14**, 22-7.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* **296**, 910-3.

Massjouni, N., Rivera, C. G. and Murali, T. M. (2006). VIRGO: computational prediction of gene functions. *Nucleic Acids Res* **34**, W340-4.

Mata, J., Marguerat, S. and Bahler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* **30**, 506-14.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-10.

McCafferty, J., Griffiths, A. D., Winter, G. and Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552-4.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* **298**, 824-7.

Moreno-Hagelsieb, G. and Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl 1**, S329-36.

Moreno-Hagelsieb, G. and Janga, S. C. (2008). Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins* **70**, 344-52.

Ong, S. E., Schenone, M., Margolin, A. A., Li, X., Do, K., Doud, M. K., Mani, D. R., Kuai, L., Wang, X., Wood, J. L. et al. (2009). Identifying the proteins to which small-molecule probes and drugs bind in cells. *Proc Natl Acad Sci U S A* **106**, 4617-22.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-901.

Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S. and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nat Biotechnol* **24**, 805-15.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-8.

Pelletier, J. N., Campbell-Valois, F. X. and Michnick, S. W. (1998). Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc Natl Acad Sci U S A* **95**, 12141-6.

Pilpel, Y., Sudarsanam, P. and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **29**, 153-9.

Pullmann, R., Jr., Kim, H. H., Abdelmohsen, K., Lal, A., Martindale, J. L., Yang, X. and Gorospe, M. (2007). Analysis of turnover and translation regulatory RNA-binding protein expression through binding to cognate mRNAs. *Mol Cell Biol* **27**, 6265-78.

Ravasz, E. and Barabasi, A. L. (2003). Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**, 026112.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-5.

Resendis-Antonio, O., Freyre-Gonzalez, J. A., Menchaca-Mendez, R., Gutierrez-Rios, R. M., Martinez-Antonio, A., Avila-Sanchez, C. and Collado-Vides, J. (2005). Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet* **21**, 16-20.

Rognan, D. (2007). Chemogenomic approaches to rational drug design. *Br J Pharmacol* **152**, 38-52.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A. and Koonin, E. V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30**, 2212-23.

Ruan, J., Dean, A. K. and Zhang, W. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Syst Biol* **4**, 8.

Salcius, M., Michaud, G. A., Schweitzer, B. and Predki, P. F. (2007). Identification of small molecule targets on functional protein microarrays. *Methods Mol Biol* **382**, 239-48.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-76.

Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-8.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* **3**, e42.

Slavik, R. and Homola, J. (2006). Optical multilayers for LED-based surface plasmon resonance sensors. *Appl Opt* **45**, 3752-9.

Snel, B., Bork, P. and Huynen, M. A. (2002). The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A* **99**, 5890-5.

Speers, A. E. and Cravatt, B. F. (2004). Chemical strategies for activity-based proteomics. *Chembiochem* **5**, 41-7.

Storek, M. J., Ernst, A. and Verdine, G. L. (2002). High-resolution footprinting of sequence-specific protein-DNA contacts. *Nat Biotechnol* **20**, 183-6.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-55.

Thieffry, D., Huerta, A. M., Perez-Rueda, E. and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays* **20**, 433-40.

Tsui, I. F., Chari, R., Buys, T. P. and Lam, W. L. (2007). Public Databases and Software for the Pathway Analysis of Cancer Genomes. *Cancer Inform* **3**, 389-407.

Turner, N. C., Lord, C. J., Iorns, E., Brough, R., Swift, S., Elliott, R., Rayter, S., Tutt, A. N. and Ashworth, A. (2008). A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor. *EMBO J* **27**, 1368-77.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**, 1283-92.

Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E. E. and Zhu, J. (2009). Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput Biol* **5**, e1000616.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440-2.

Whitehurst, A. W., Bodemann, B. O., Cardenas, J., Ferguson, D., Girard, L., Peyton, M., Minna, J. D., Michnoff, C., Hao, W., Roth, M. G. et al. (2007). Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature* **446**, 815-9.

Wingren, C. and Borrebaeck, C. A. (2006). Antibody microarrays: current status and key technological advances. *OMICS* **10**, 411-27.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B. and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901-6.

Wolf, D. M. and Arkin, A. P. (2003). Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6**, 125-34.

Wu, W. S., Li, W. H. and Chen, B. S. (2006). Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics* **7**, 421.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Mol Biol Evol* **18**, 1694-702.

Wuchty, S., Oltvai, Z. N. and Barabasi, A. L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* **35**, 176-9.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232-40.

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. and Vidal, M. (2007). Drug-target network. *Nat Biotechnol* **25**, 1119-26.

Yu, H. and Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A* **103**, 14724-31.

Zhao, X. M., Chen, L. and Aihara, K. (2008). Protein function prediction with the shortest path in functional linkage graph and boosting. *Int J Bioinform Res Appl* **4**, 375-84.

2 Functional, structural and dynamic constraints on bacterial regulatory networks

CONTENTS OF CHAPTER 2

OUTLINE

One of the most important developments in our understanding of biological systems in the past decade is the application of network theory to biological problems. This is particularly true for the case of regulation of gene expression. Taking advantage of the currently available transcriptional regulatory networks of the model bacteria, *Escherichia coli* and *Bacillus subtilis,* a comprehensive genomic and structural analysis was performed. It was found that while the mode of regulatory interaction between transcription factors (TFs) is predominantly positive, TFs are frequently negatively auto-regulated. Furthermore, feedback loops, regulatory motifs and regulatory pathways are unevenly distributed in this network. Short pathways, multiple feed-forward loops and negative auto-regulatory interactions are particularly predominant in the sub-network controlling metabolic functions such as the use of alternative carbon sources. In contrast, long hierarchical cascades and positive auto-regulatory loops are over-represented in the sub-networks controlling developmental processes for biofilm and chemotaxis. We propose that these long transcriptional cascades coupled with regulatory switches (positive loops) for external sensing enable the coexistence of multiple bacterial phenotypes. We also provide a link between the transcriptional hierarchy of regulons (TFs) and their genome organization. We show that, to drive the kinetics and concentration gradients, TFs belonging to big and small regulons, depending on the number of genes they regulate, organize themselves differently on the genome with respect to their targets. We then propose a conceptual model that can explain how the hierarchical structure of TRNs might be ultimately governed by the dynamic biophysical requirements for targeting DNA-binding sites by transcription factors. Our results suggest that the main parameters defining the position of a TF in the network hierarchy are the number and chromosomal distances of the genes they regulate and their protein concentration gradients. These observations give insights into how the hierarchical structure of transcriptional networks can be encoded on the chromosome to drive the kinetics and concentration gradients of TFs depending on the number of genes they regulate and could be a common theme valid for other prokaryotes, proposing the role of transcriptional regulation in shaping the organization of genes on a chromosome.

CONTRIBUTION TO THE WORK IN THIS CHAPTER

Please note that the work presented in this chapter is the result of the following three publications during my doctoral period and my contribution to the work excludes the construction and layout of the cross-regulatory network of *E. coli* discussed in section 2.2.1 and defining the conceptual model discussed in section 2.2.2. I am grateful for the input and collaboration of Dr. Agustino Martinez-Antonio at CINVESTAV, IPN, Mexico, Dr. Dennis Thieffry at University of Marseille, France and Heladia Salgado at CCG, UNAM, Mexico for this analysis.

1) Transcriptional regulatory networks
Sarath Chandra Janga and M. Madan Babu
Book chapter for *Cambridge University Press* for an edited book on "*Networks in Cell Biology*"
2) Functional organization of *Escherichia coli* transcriptional regulatory network
Agustino Martinez-Antonio, Sarath Chandra Janga and Dennis Thieffry
Journal of Molecular Biology, 2008, Vol. 381(1):238-247
3) Transcriptional regulation shapes the organization of genes on bacterial chromosomes
Sarath Chandra Janga, Heladia Salgado and Agustino Martinez-Antonio
Nucleic Acids Research, 2009, Vol.37, No. 11, 3680-3688

2.1 INTRODUCTION

One of the most important developments in our understanding of biological systems in the past decade is the application of network theory to biological problems. This is particularly true for the case of regulation of gene expression. The accumulation of data on many factors that control the expression of genes or groups of genes, together with the increased use of high-throughput techniques, such as DNA arrays and proteomics, has generated an overwhelming amount of data that has to be understood to infer relationships between genes, and between genes and signals. The reductionist approaches of molecular biology have made it impractical to deal with large amounts of information giving rise to the increasing use of the notion of networks in biology. Typically in network approaches to understand a biological system, elements are represented as nodes in the graph, which are connected by edges that represent biological interactions. This approach allows ill-defined descriptions of complexity to be replaced by objectively quantifiable, numerical parameters, such as connectivity or strengths of interactions (Jeong et al., 2000; Ronen et al., 2002).

Most network analysis of transcriptional regulatory events in an organism involves representing genes and the proteins they encode as nodes. However it should be noted that in contrast to protein-protein networks, the links in transcription networks have directionality, meaning that connections have a starting node and a target node. Normally, an edge in such a network goes from a transcription factor to the genes it regulates. In most cases such regulation occurs through a direct effect i.e, the regulator binds the promoter regions upstream of the protein coding genes they control the activity off. More complex representations include the incorporation of other entities like small molecules, RNA encoding genes, signal-transduction pathways or interacting proteins. However most of our understanding of transcriptional networks to date has been limited to the holistic view of Transcription Factors (TFs) and Target Genes (TGs) as nodes and the regulatory interactions between them as edges. It is this graph in an organism that is usually referred to as the Gene Regulatory Network (GRN). This network is also referred to as “Transcriptional Regulatory Network (TRN)” or simply “transcriptional network”.

One of the most important and obvious pieces of information that can be obtained is the distribution of connectivity, i.e how many connections a node has and how many nodes have a particular number of connections. In the case of transcriptional networks these parameters actually have two sides, as incoming and outgoing connections must be considered separately. The incoming connectivity is the number of transcription factors regulating a target gene, which

gives a sense of the combinatorial effect of gene regulation. The fraction of target genes with a given incoming connectivity was observed to follow an exponential distribution in both *Escherichia coli* and *Saccharomyces cerevisiae* (Guelzim et al., 2002; Thieffry et al., 1998). The exponential behaviour indicates that most target genes are regulated by a similar number of factors and apparently reflects the limits on the size of the multiprotein complexes that can be bound near the promoter as well as by the amount of DNA sequence in upstream regions of genes. On the other hand, the outgoing connectivity, which is the number of target genes regulated by each transcription factor, was found to be distributed according to a power law, contrary to the incoming connectivity distribution. This is indicative of a hub-containing network structure, in which a select set of transcription factors participate in the regulation of a disproportionately large number of target genes.

At a local level, in transcriptional networks certain sub-networks appear more often than expected by chance and have been referred to as motifs, analogous to sequence motifs which occur repeatedly in sequences. Motifs were originally described in *E. coli* transcriptional regulatory network but were subsequently found in yeast and other organisms (Alon, 2007b; Shen-Orr et al., 2002). Three network motifs were found to be predominantly occurring in most transcriptional networks: 1) Feed-Forward Loop (FFL), in which a transcription factor regulates the expression of another transcription factor which, in turn, regulates a gene that is also regulated by the first transcription factor; 2) Single-Input Module (SIM), in which a single transcription factor regulates several genes, which is usually also called a simple regulon (Gutierrez-Rios et al., 2003); 3) Dense Overlapping Regulons (DORs) in which several TFs regulate overlapping sets of genes and these groups are also called a complex regulon. FFL appears to be the most abundant motif among the best studied transcriptional networks. FFLs have been further classified into eight motif sub-types and two of them namely coherent type-1 and incoherent type-1 FFL appear to be much more predominant than others (Alon, 2007b; Mangan and Alon, 2003). The former was shown to act as a sign-sensitive delay element and a persistence detector while the later was demonstrated to function as a pulse generator and response accelerator (Mangan et al., 2006; Mangan et al., 2003). Although motifs form over-represented sub-graphs in the entire network of transcriptional regulation, they do not appear independently but rather integrate to form super-structures or modules that carry out a common biological function by sharing some of their edges (Dobrin et al., 2004a; Resendis-Antonio et al., 2005).

At a global level, transcriptional regulatory networks have been shown to possess a scalefree multi-layer hierarchical modular structure using both, a top-down and bottom-up,

approaches for determining hierarchy (Ma et al., 2004a; Yu and Gerstein, 2006). Interestingly, transcription networks do not seem to possess feedback regulation at the level of transcription meaning transcriptional regulation of TFs at the top by TFs at the bottom of this hierarchial structure is not frequent, indicating the prevalence for alternative forms of feedback control of transcription. Typically such a feedback occurs through the usage of protein-protein interactions at post-translational level or due to a complex interplay of cellular entities which control the activity of TFs by changing their conformation depending on the continuously varying intra- and extra-cellular conditions (Martinez-Antonio et al., 2006a; Yu and Gerstein, 2006). The pyramid shaped multi-layer hierarchical transcriptional networks builds the basis for modules and motifs which have been determined using a number of approaches for decomposing the regulatory networks (Bar-Joseph et al., 2003; Dobrin et al., 2004a; Ihmels et al., 2004; Ihmels et al., 2002; Ma et al., 2004a; Milo et al., 2002; Resendis-Antonio et al., 2005; Segal et al., 2003; Shen-Orr et al., 2002; Wu et al., 2006) (see Figure 2-1 below for a summary of these properties).

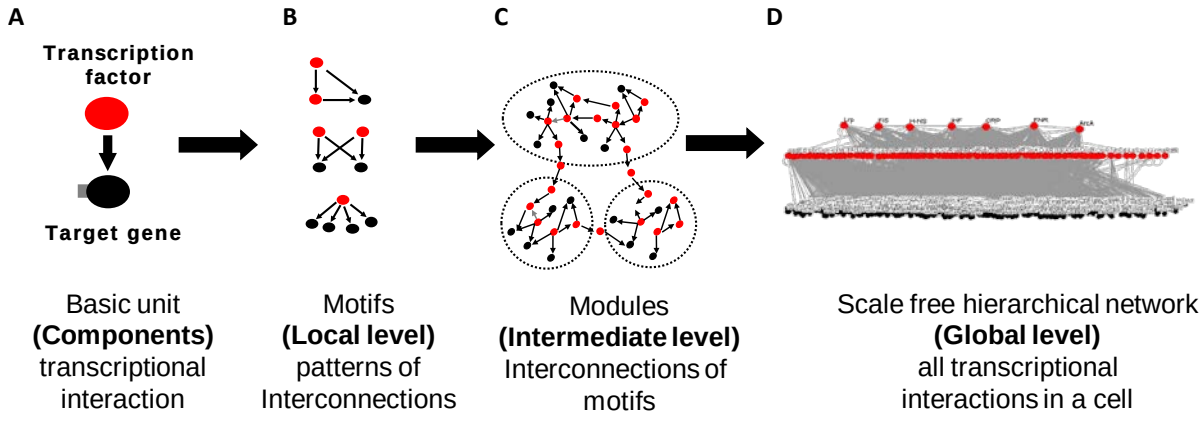


Figure 2-1: Structure of the transcriptional regulatory network. Nodes represent transcription factors (red nodes) or target genes (black nodes) and directed edge indicates a regulatory interaction between the two. (A) Components of a regulatory interaction (B) Local structure of the network consists of patterns of inter-connections called network motifs. The three frequently occurring motifs are the feed-forward motif (top), single input motif (middle) and multiple input motif (bottom) (C) Motifs are interconnected to form groups of highly connected genes, referred to as regulatory modules (dashed circles) (D) the set of all regulatory interactions in a cell is referred to as the transcriptional network.

Chromosomal proximity of functionally related genes has been observed as early as 1960 when Jacob and Monod first reported that genes in bacteria are organized into polycistronic messenger RNAs (Jacob et al., 1960) suggesting the importance of chromosomal organization. It is now known that genes in bacterial genomes are largely organized into operons in order to enable co-regulation (Price et al., 2005) and it is due to this reason that the conservation of gene order in prokaryotic genomes is highly non-random (Ermolaeva et al.,

2001; Korbel et al., 2004). However our progress in understanding, how and what governs the organization of these functional modules on bacterial chromosomes has been very limited, despite the enormous number of bacterial genomes that have been sequenced in recent years. In particular, these observations lead to several unanswered basic questions like, is transcriptional regulation constraining genome structure beyond local operon structure? Are their constraints on the organization and evolution of transcriptional networks?

Products of genes have different functional roles and hence not all genes are used at the same time and for the same purpose. This explains why groups of genes are differentially expressed. For instance, genes encoding for enzymes in krebs's cycle are constitutively expressed in response to most growing conditions while genes responsible for using alternative carbon sources are sporadically required. The decision about which genes should be *turned on* or *off* is executed by transcription factors (TFs) that use metabolites/signals as input information from the environmental state and give a transcriptional response as output (Jacob, 1970; Martinez-Antonio et al., 2006b; Ptashne and Gann, 1997). As a result, the notion that TFs are expressed in varying concentrations came into existence. For instance, LacI, a repressor of the operon for lactose consumption, is expressed in the order of tens' of molecules per cell, while global regulators such as CRP (cAMP receptor protein) or IHF (integration host factor) occur in the order of thousands of molecules in the course of the cell cycle (Elf et al., 2007; Luijsterburg et al., 2006). In bacterial cells, where transcription and translation are coupled to happen in the same compartment these considerations become especially important for regulating gene expression. During transcription, regulatory proteins (TFs) should find and bind to specific DNA sequences on the operator region of their target genes to repress or induce their transcription (Browning and Busby, 2004). The protein-DNA interaction is a critical step in gene regulation as TFs find their DNA-binding sites as result of a passive process. Furthermore, TFs do not use energy (e.g. ATP hydrolysis) to get DNA-sequence information (Hu et al., 2008), which forces these systems to use additional strategies for the optimal performance of different TFs. In the early era of molecular biology, brownian diffusion was thought to be the determining step in DNA-binding site recognition by TFs. However, this assumption was challenged when it was reported that the LacI repressor finds its DNA-targets 90-100 times faster than that predicted by a mere diffusive mechanism (Riggs et al., 1970; Wang et al., 2006). This observation led to the suggestion of 'facilitated diffusion' mechanism. In such a mechanism, TFs alternate between a three dimensional (3-D) diffusion in the cell jumping between DNA-strands and one-dimensional (1-D) sliding along the DNA to rapidly locate their binding sites (Berg et al., 1981; Richter and Eigen, 1974; Winter et al., 1981). This hypothesis was corroborated by several works mostly

with single molecule studies in which the authors visualized individual TFs interacting with the DNA (Elf et al., 2007; Shimamoto, 1999; Wang et al., 2006; Xie et al., 2008). Several groups have also mathematically modeled the sliding process along the DNA and shown it to be a plausible way of making the search significantly faster than 3-D diffusion alone, in particular for TFs in low cellular concentrations (Cherstvy et al., 2008; Gowers et al., 2005; Hu et al., 2008; Murugan, 2007). However, it is unclear what factors govern a TF to adopt one or the other strategy discussed above and if there is an interplay between nucleoid structure, genome organization and the biophysical aspects of transcriptional regulation in bacterial systems. In what follows, I will present a summary of the work that was performed to understand the functional and structural constraints that are imposed on bacterial transcriptional networks.

2.2 RESULTS

2.2.1 Constraints imposed on the network of transcription factors in bacteria

In bacteria, coupling of gene expression with external conditions is achieved through two molecular functions; (i) association of transcription factors (TFs) at specific sites in the genome and (ii) recognition of a relevant effector signal or metabolite (Jacob, 1970; Martinez-Antonio et al., 2006b). Typically these functions are performed by different domains of a single polypeptide, but there are also cases where two interacting proteins are responsible for these functions, as in two-component systems (Ulrich et al., 2005).

At the phenotypic level, there are evidences for the coexistence of multiple phenotypes in bacterial cultures, e.g., of cells with different morphological and physiological abilities like motility, biofilm formation, drug-resistance etc (Balaban et al., 2004; Ehrlich et al., 2005). In particular, biofilm formation and chemotaxis are considered as multi-stage developmental processes and, in mature biofilms, a mixture of bacterial population from different developmental-stages coexist (O'Toole et al., 2000; Stoodley et al., 2002). In an attempt to understand whether these distinct phenotypes in growing populations of bacteria can be linked and explained in the context of transcriptional regulation, wealth of experimental data on transcriptional regulation for the best-characterized bacterium, *Escherichia coli*, was analyzed (Gama-Castro et al., 2008; Martinez-Antonio et al., 2008). In particular, a detailed analysis of the structure of the transcriptional regulatory network of transcription factors allowed us to unravel several constraints imposed on this network. In the following sections, I summarize the results of this published study where we performed a comprehensive analysis of regulatory

interactions between all the experimentally characterized TFs (defined as cross-regulatory network) in *E. coli* (Martinez-Antonio et al., 2008). Our structural analysis of the transcriptional cross-regulatory network in *Escherichia coli* suggests that, regulatory interactions between TFs are predominantly positive while auto-regulatory interactions are mostly negative. However, this general trend seems to be reversed in the case of most downstream TFs involved in the regulation of biofilm/chemotaxis modules.

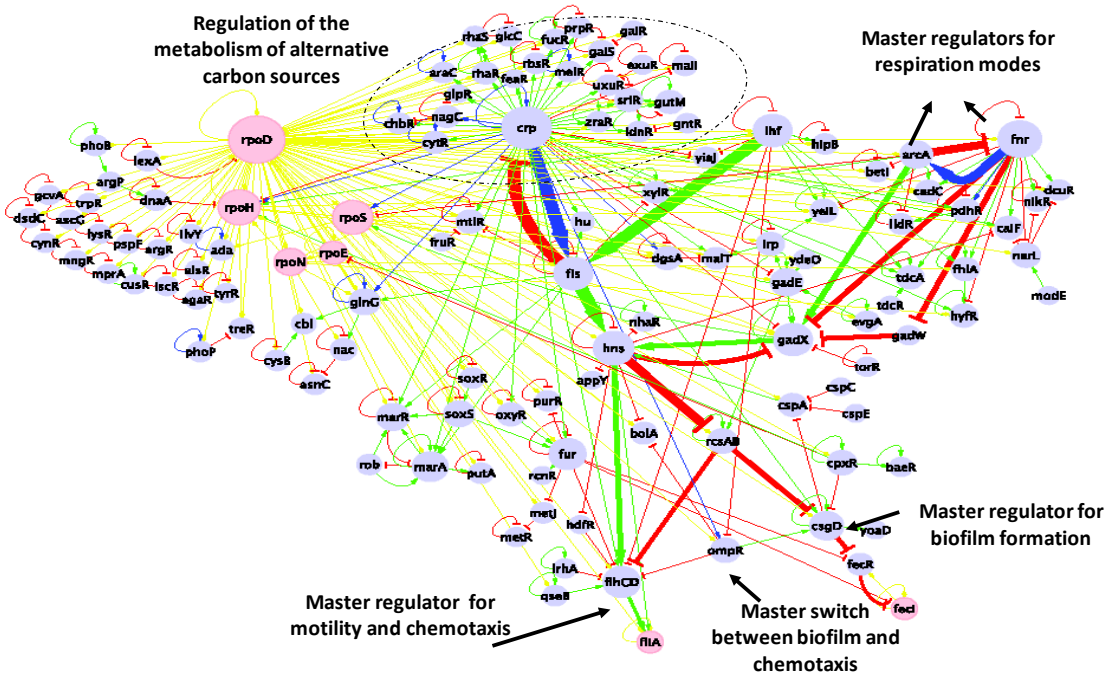


Figure 2-2: Core transcriptional regulatory network of *E. coli*. Light blue and pink nodes represent genes encoding for transcription factors and sigma factors, respectively; edges represent regulatory interactions among TFs and sigma factors (green for activation, red for repression, blue for dual interactions and yellow for sigma transcription), whereas loops represent transcriptional auto-regulation. Specific sub-networks responsible for the regulation different processes are delimitated with a dashed line including the section showing the regulation of carbon sources.

We also note that, there are striking topological differences between the sub-networks controlling carbon metabolism and developmental processes; the former compose of many parallel short transcriptional cascades, encompassing multiple feed-forward loops, each enabling the use of one alternative carbon source, while the later involve long and intertwined regulatory cascades (see Figure 2-2). These long transcriptional cascades typically include multiple auto-activated intermediate TFs, as well as regulatory circuits between TFs and sigma factors. We further observe that transcription factors acting at the end of these regulatory cascades often belong to two-component systems. This topology suggests that on one hand, cell homeostasy is maintained through multiple regulatory cascades with commonly auto-

repressed TFs, while the regulatory memory is preserved by the sequential activation of TFs at the core of the network. On the other hand, downstream of the hierarchical network, two-component systems can memorise transient external signals through auto-activation loops, thus acting as molecular switches enabling the coexistence of alternative phenotypes.

2.2.1.1 Topology of *Escherichia coli* cross-regulatory transcriptional network

Available experimental data point to more than 3000 regulatory interactions between TFs and their regulated genes in *E. coli*. This information is integrated and documented in a specialised database called RegulonDB (Salgado et al., 2006). Global analyses of this huge network have already been published, emphasizing a hierarchical organisation and statistically over-represented regulatory motifs (Dobrin et al., 2004b; Ma et al., 2004b; Shen-Orr et al., 2002). However, our aim here is to analyse the flow of regulatory information within the network of transcriptional interactions among TFs and sigmas (defined here as the *E. coli* transcriptional cross-regulatory network). This network encompasses 115 TFs and 7 sigma factors, i.e., around one third of the total predicted TF proteins in this bacterium (Figure 2-2) (Madan Babu and Teichmann, 2003; Perez-Rueda and Collado-Vides, 2000). On average, every TF is connected to two other TFs (i.e., more technically, the mean degree of the regulatory graph is 2.74). However, the connectivity distribution of TFs is not uniform, with a small fraction of global TFs with high out-degrees dominating the network (Martinez-Antonio and Collado-Vides, 2003).

In order to better visualise the informational flow through the network, the following graphical conventions have been used in this figure (see also legend): (i) the size of the nodes representing TFs is proportional to the number of genes they regulate (e.g., CRP regulates 413 genes and is represented by the second biggest node, after the housekeeping sigma factor rpoD); (ii) arrows and colours refer to the direction and sign of the regulatory interaction; (iii) arrow thickness is proportional to the impact of the interaction, computed as the number of genes thereby (in)directly regulated.

Majority of the TFs in this network are auto-regulated (~70%), of which about two-third account for negative loops (Table 2-1). This finding is consistent with the results of an analysis performed with a much smaller number of TFs more than ten years ago (Thieffry et al., 1998). This predominance of negative auto-regulatory loops contrasts with the predominance of positive arcs between different TFs (about 54%, see Table 2-1). The dominance of positive regulatory interactions in the regulatory network of *E. coli* is not limited to those among TFs, as a comparison of the regulation of all the target genes (3017 arcs/edges) shows that about 54%

(1630) are positively regulated, 40% (1206) of them are repressed while about 6% (171) are dual regulated. This is especially interesting because majority of the TFs in bacteria have been reported to act as repressors (Moreno-Campuzano et al., 2006; Perez-Rueda and Collado-Vides, 2000; Struhl, 1999). The conventions used in Figure 2-2 clearly display the hierarchical organisation of the network, with master regulators such as CRP, FNR or IHF each (in)directly regulating a large number of other transcription factors. Furthermore, the layout emphasises important variations regarding the length of the transcriptional cascades.

Table 2-1. Distributions of positive, negative and dual (auto-) regulatory interactions and mean path length computed for the *E. coli* transcriptional cross-regulatory network displayed in Figure 2-2 and for the sub-networks controlling the regulation of alternative carbon sources as well as biofilm and chemotaxis development processes. Note that arcs are synonymous to edges in the network.

Network section	No. of TFs	Auto-regulatory interactions	Positive auto-regulatory interactions	Negative auto-regulatory interactions	Dual auto-regulatory interactions	Regulatory arcs[1]	Positive arcs	Negative arcs	Dual arcs	Average path lenghth[2]
All	115	80 (70%)	24 (30%)	48 (60%)	8 (10%)	166	90 (54%)	67 (40%)	9 (6%)	2.74
Carbon sources	24	19 (79%)	5 (26%)	9 (48%)	5 (26%)	27	20 (74%)	6 (22%)	1 (4%)	1.53
Biofilm & motility[3]	32	18 (56%)	9 (50%)	9 (50%)	0	52	22 (42%)	27 (52%)	3 (6%)	3.12

[1]Regulatory interactions from TFs to others TFs or towards sigma factors
[2]Average path lengths in the (sub-)network(s) were calculated with the ViSANT program (Hu *et al.*, 2007)
[3]Only the TFs forming cascades ending on biofilm and chemotaxis modules were computed, the autoregulation of CRP was included in the carbon sources module.

Although functional annotations on transcription factors are still limited, it is possible to classify the cross-regulating TFs into broad categories according to the physiological functions of the target structural genes: carbohydrate initial catabolism, respiration, biofilm formation and chemotaxis, etc,. As shown in this figure, these broad classes correspond to different local network topologies. Due to their contrasting topologies, in what follows, we will focus our discussion on short regulatory cascades observed in the case of carbohydrate catabolism as opposed to long regulatory cascades seen in the case of biofilm and chemotaxis pathways (marked in the figure and shown in table). CRP resides at the top of both sub-networks. CRP is the only global TF acting hierarchically over local TFs for the usage of carbohydrates, whereas CRP's activity is comparable to the activity of other global regulators in the rest of the network.

Note that the concentration of its effector metabolite, cAMP (cyclic adenosine monophosphate), is at par with that of ATP, (adenosine triphosphate), which acts as the energetic currency of the cell (Bettenbrock et al., 2007). This suggests that CRP not only regulates the use of these substrates for producing ATP, but also senses the energetic status of the cell to decide the execution of other cellular programs.

While it is evident from this figure and the table that there are differences in the topologies of the subnetworks controlling metabolism versus motility and chemotaxis, which will be the focus of the following sections, there are other subnetworks which are also worth mentioning. In particular, all 9 TFs controlling the expression of genes for amino acid biosynthesis seem to be expressed constitutively by sigma 70. Each TF regulates the transcription of the required genes for producing different amino acids. At high concentrations of the amino acids, allosteric modifications of TFs follow binding to their respective amino acids, resulting in TF auto-repression, as well as to the repression of the corresponding biosynthetic genes. Interestingly, the logic behind negative autoregulation in this case is different to that of the catabolism of carbohydrates. While in the latter case TFs are autorepressed until the substrate is available, in the case of amino acids, TFs are autorepressed only in the presence of an excess of the synthesized final product. Another interesting subnetwork is that for alleviating the stresses by drugs, solvents and weak organic acids. The regulatory logic in this complex subnetwork is peculiar as their components form multi-element circuits and their inputs are directed by Rob and SoxR, two small proteins constitutively expressed but with very short half lives (1-2 min). Their stability/degradation depends on the presence/absence of their effector signals (Griffith and Wolf, 2004; Martin et al., 2000; Shah and Wolf, 2004).

2.2.1.2 Multiple parallel feed-forward loops regulate the use of different carbon sources

Cellular feeding, which includes the uptake of carbon and energy sources and their metabolism, can be considered as one of the main physiological processes in bacterial systems. The regulation of these processes directly affects cellular fitness. The selection of carbon sources is regulated by CRP and about 20 more specific TFs (Figure 2-2). The hierarchical organisation of the corresponding sub-network is characterized by a short average path length (Table 2-1). Regulatory interactions between CRP and the specific TFs result in the occurrence of multiple Feed-Forward Loops (FFL) for the use of alternative sugar sources. FFL is a network motif recurrently found in transcriptional networks and is defined as a three-gene pattern composed of two input transcription factors, one of which regulates the other, both jointly regulating a target

gene (Mangan and Alon, 2003; Shen-Orr et al., 2002). Based on the mode of regulation of each TF, this motif is sub-divided into 8 different sub-types (Mangan and Alon, 2003). Coherent FFL type 1 corresponds to all the regulatory interactions in the motif being positive while in incoherent type 1 FFL the first TF regulates positively both the targets although second TF represses the expression of the target gene, thereby reversing the final effect. Majority of the feed-forward loops present in the subnetwork for carbon catabolism belong to coherent and incoherent type 1 groups (Mangan and Alon, 2003), with both TFs working cooperatively, as a result of a persistent signal (in this case, cyclic adenosine monophosphate) affecting the global TF and the presence of a signal affecting a TF corresponding to a sugar alternative to glucose (Alon, 2007b; Janga et al., 2007b; Mangan and Alon, 2003). This motif structure enables the filtering of short pulses of the signal affecting the global TF (cAMP) in case of transient glucose deprivation. Consequently, the target structural genes are activated only in the persistent absence of glucose and in the presence of an alternative carbon source.

The phosphotransferase system (PTS) system typically transports and phosphorylates certain sugars, including glucose, a preferred carbon source for *E. coli*, and this condition ultimately results in low levels of cAMP. Consequently, CRP does not activate the transcription of the genes responsible for the degradation of alternative sugars. Note that most structural genes involved in the transport and initial catabolism of alternative carbon sources are encoded in operons, each specifically repressed in the absence of the inducing sugar. However, when glucose is lacking, cAMP level increases and CRP can activate the transcription of genes responsible for degrading alternative carbon sources (Deutscher et al., 2006). Simultaneously, sugars (or processed variant thereof) present in the cell bind their specific TF; allosteric interactions then result in TF unbinding from DNA, alleviating the repression and permitting the transcription of the corresponding target genes. This organisation involving multiple parallel feed-forward loops coupled to PTS activity appears optimal to enable rapid transcriptional responses to sudden lack of glucose in the presence of alternative carbon sources in the milieu (Dekel et al., 2005; Mangan et al., 2005).

2.2.1.3 Long hierarchical cascades regulate developmental processes

Biofilm formation and bacterial mobility can be seen as the outcome of specialised cell differentiation pathways. Biofilm formation involves subsequent cellular changes at the morphological and physiological levels resulting in bacterial populations with multiple phenotypes (Ehrlich et al., 2005; Hooshangi et al., 2005; Shapiro, 1998). Furthermore, bacteria living in biofilm communities are present in different developmental stages (at least four defined

stages) as has been observed in *Cryptoccocus*, *Pseudomonas*, *Staphylococcus*, *Xanthomonas*, etc. (Aldridge and Hughes, 2002; Chantratita et al., 2007; Guerrero et al., 2006; Handke et al., 2004; Kamoun and Kado, 1990; Massey et al., 2001). The part of *E. coli* transcriptional cross-regulatory network involved in the control of biofilm formation and motility exhibits a relatively complex topology with several long cascades from CRP, IHF and FNR to downstream specialised TFs (Table 2-1 and Figure 2-2). Several of these cascades converge on the master regulators for motility and biofilm formation in the downstream (FhlCD and CsgD, respectively). For instance, CsgD, the master regulator for biofilm formation, is directly involved in several long circuits, suggesting a particular tight coupling of CsgD activity with the intracellular status. In contrast, FlhCD, the master compound regulator for motility and chemotaxis, is known to be regulated by nine other TFs but has not yet been reported to regulate any other TF. In fact, a relatively high proportion of the downstream specialised transcription factors also auto-activate themselves, a feature which is rare at the level of the whole transcriptional network.

Note that the motility module has its own sigma factor, FliA, regulated by FlhCD. FliA is required for the transcription of the genes required for the last part of flagella development and for chemotaxis machinery (Kalir and Alon, 2004; Kalir et al., 2001). In contrast, the genes for biofilm development are transcribed by the housekeeping sigma 70 and RpoS, the sigma factor expressed in response to general stress (Hengge-Aronis, 2002).

The execution of such long regulatory cascades requires time. Indeed, complete flagella assembly may take a generation time or longer (Aizawa and Kubori, 1998; Macnab, 2003; Pruss and Matsumura, 1997). The occurrence of positively auto-regulated TFs at several intermediate steps enables informed decisions about the cellular/environmental condition. In some conditions, cellular duplication might be faster than the conclusion of a long regulatory cascade. This implies that bacterial populations likely consist of mixture of bacteria with transcriptional programs at different levels in long regulatory cascades.

2.2.2 Constraints imposed on bacterial genome organization by transcriptional network

Simple regulons comprise of transcription factors (TFs) and the set of genes they regulate and were defined as early as 1964 (Maas et al., 1964). The functional properties of these set of genes can be diverse, vary in number and be encoded dispersedly on the chromosome. However, it is unclear if there is any relationship between regulon size and the chromosomal positioning of their genetic components, nor is it known how TFs, constituting large and small regulons, coordinate their activity in the context of regulatory networks.

It is now well accepted that most biological networks are scale-free in their structure (Barabasi and Albert, 1999; Cases and de Lorenzo, 2005) and modular in their function (Hartwell et al., 1999; Kashtan and Alon, 2005; Resendis-Antonio et al., 2005; Slonim et al., 2006; Snel and Huynen, 2004) but our understanding on how this scale-free structure is reflected on the chromosomal organization is very limited. Thus, addressing the design behind these architectures in the context of genome organization can provide important insights to a better understanding of genome structure and function. In bacterial genomes where transcription and translation occur in the same compartment these questions become especially important ,as the positioning of TFs on the chromosome could depend on concentration (Cai et al., 2006; Golding et al., 2005; Yu et al., 2006). Recent works have suggested the importance of chromosomal distance in bacterial genomes from diverse perspectives; from transcription units and operon organization to divergent and convergent transcriptional control (Janga et al., 2007a; Korbel et al., 2004; Menchaca-Mendez et al., 2005; Warren and ten Wolde, 2004), however no analysis has focused on a link between regulon sizes, their genome organization and how this relates to the hierarchical transcriptional network structure (Lagomarsino et al., 2007; Ma et al., 2004a; Yu and Gerstein, 2006). In a attempt to address the question of how these factors interplay and relate in the larger context of transcriptional networks of bacteria, currently available TRNs of best characterised gram-negative bacterial model, *Escherichia coli* (Salgado et al., 2006) and the not as well-characterized gram-positive representative, *Bacillus subtilis* (Makita et al., 2004) were used.

In short, we studied the dependency between regulon sizes and their chromosomal positioning and show that regulons can be classified into 3 distinct groups based on average chromosomal distance between TFs and their respective target genes (Janga et al., 2007a) into big, intermediate and small regulons (Figure 2-3). We note that regulatory flux is generally driven from big to small regulons in both *E. coli* and *B. subtilis*. Finally, using data from two independently reported studies we show that the higher a TF is in the transcriptional hierarchy, more are its detected number of mRNA molecules per cell reflecting their need to be expressed in higher concentrations to regulate targets located distantly on the chromosome. In contrast to big regulons, local or dedicated TFs (lower in the network hierarchy) were found to be expressed in much lower concentrations explaining the reasons for their proximity on the chromosome to their target genes. These observations show how scale-free structure of transcriptional networks can be encoded on the chromosome to drive the kinetics and concentration gradients of TFs depending on the number of genes they regulate and could facilitate the horizontal transfer of local environment-specific transcriptional modules.

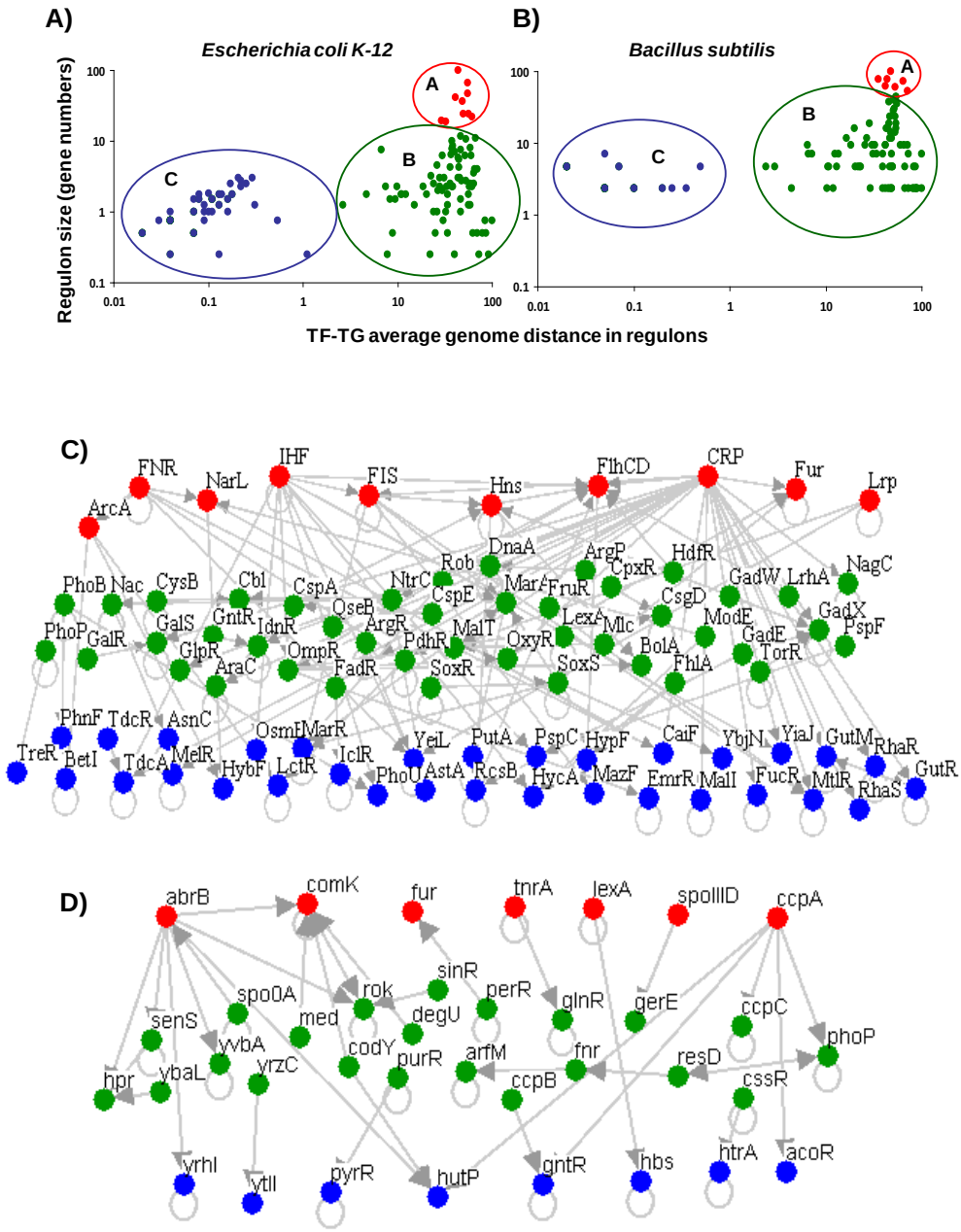


Figure 2-3: Relationship between size (defined as the number of target genes) and average chromosomal distance for all known regulons in (A) *E. coli* and (B) *B. subtilis*. Regulon size is plotted on Y-axis and is normalized with respect to size of the biggest regulon in each genome for the sake of comparison across genomes and the average chromosomal distance between the TF encoding gene and their respective target genes is shown on X-axis. Chromosomal distances were calculated as defined earlier (Janga et al., 2007a) with the maximum distance being half the number of protein coding genes on a circular chromosome. Note that both regulon sizes and average chromosomal distances are normalized with respect to the maximum and both the axes are shown on a logarithmic scale. Flow of regulatory interactions between the TFs heading the regulons, grouped according to their size and chromosomal distance in (C) *E. coli* and (D) *B. subtilis*; it can be noted that the regulatory flux among TFs typically follows the order, big to intermediate to small regulons, coloured respectively in red (big), green (intermediate) and blue (small).

2.2.2.1 Genomic co-localization of TFs and target genes is observed in small regulons

In a previous study we reported a distinct organization of genes coding for transcription factors (TFs) and their effector genes (whose products control TFs), depending on whether the effector proteins sense signals from endogenous or exogenous origin in *Escherichia coli* (Janga et al., 2007a). In this study, we analyze if this observed distance, when extended to all members of a regulon, shows any trends depending on the size of the regulon. It should be noted that there is a clear distinction between TF-effector gene pairs and TF-target gene pairs. While the product of the former controls the activity of the TFs the later correspond to the set of genes transcriptionally regulated by the TF (forming part of a regulon). In this work our interest is to understand how the chromosomal distances (measured as number of intervening protein coding genes on a circular genome) between TF and its target genes in different regulons can explain or reflect the network structure. To address these questions, we obtained all regulons wherein transcription factors regulate at least two genes (excluding auto-regulation) in *E. coli* and in *B. subtilis*, taken from regulonDB (Salgado et al., 2006) and DBTBS (Ishii et al., 2001), respectively. We included heterodimeric TFs and excluded auto regulatory interactions. In *E. coli K12*, our final dataset contained 141 regulons comprising of 1597 regulatory interactions between TFs and their regulated genes; in *B. subtilis* the dataset contained 54 regulons comprising of 499 genes. First we asked if there is any link between regulon size (number of regulated genes by each TF) and the average chromosomal distance (calculated as the number of intervening protein coding genes on the circular chromosome as described earlier (Janga et al., 2007a)) between the TF and its target genes in each case. As a result of clustering (see Methods), regulons in both organisms can be grouped into three main categories (see Figure 2-3 panel A, B and Table 2-2): (A) a few big regulons (10 in *E. coli* and 7 in *B. subtilis*) regulating more that 50% of the genes in their transcriptional networks (group A in the Figure). (B) An intermediate and heterogeneous group of regulons consisting of varying regulon sizes and chromosomal distances (group B); and (C) a group of small regulons having short chromosomal distances (group C). Notably, small regulons (group C) are smaller than the biggest operons of *E. coli* (15 genes) and *B. subtilus* (22 genes), possibly suggesting limitations on their sizes to act as functional modules either in the context of co-expression or for horizontal transfer (Korbel et al., 2004; Pal et al., 2005). The group of 10 TFs in *E. coli* having the most number of regulated genes, all are classified as global regulators according to one or more previous studies (Martinez-Antonio and Collado-Vides, 2003) while most of the TFs constituting small

regulons were found to sense external fluctuant signals resembling local genetic modules (Martinez-Antonio et al., 2006b). In particular, we found that highly connected TFs were either Nucleoid Associated Proteins (NAPs) like IHF, FIS, HNS or growth condition specific regulators like Crp (aerobic), Fnr and NarL (anaerobic), central intermediary regulators like Lrp, ferric uptake regulator (Fur) or developmental pathway associated factors like FlhDC responsible for biofilm formation, suggesting that these regulators indeed have key functional roles in controlling the transcriptional responses of the cell depending on the condition of growth. It is interesting to note that several NAPs which are known to act as bacterial analogs of chromatin remodeling factors are enriched in this class (see below). Similarly, a functional analysis of the TFs from group B suggested that several of them are involved in basic cellular activities like regulation of the biosynthesis of amino acids, regulation of cell division and repair, regulation of the uptake of elements, cellular stress and response to antibiotics, indicating a limited functional role of these TFs compared to those from group A. Finally, an analysis of TFs from group C suggested that they are involved in the uptake of carbon sources, degradation of small molecules and are abundant in two component response regulators. To estimate if the average chromosomal distance seen in each group is significant, we compared this distance against those seen in randomly generated sets as described in Methods. We found that the observed distances for each of the three groups are significantly smaller than expected by chance, with regulons from group C being the closest (Table 2-2).

Table 2-2. Properties of the main groups of regulons identified in the regulatory network of *E. coli,* based on average chromosomal distance between TF and its target genes.

Regulon group	Number of regulons (% of total regulons)	Regulon size (average no. of genes/regulon)	Total number of regulated genes (% of total regulated genes)	Average distance (in gene numbers) between the TF and the target genes	P-value significance (Z-score)
A	10 (7%)	76-399 (159)	1595 (99)	1059.45	P< 0.001 (-9.98)
B	73 (52%)	2-48 (13)	953 (59)	889.69	P< 0.001 (-10.74)
C	58 (41%)	<12 (4.8)	281 (17.5)	2.42	P< 0.001 (-20.44)

2.2.2.2 Transcriptional regulatory flow in the network of TFs

To find out if there is any coordination between the TFs heading the different groups of regulons identified above, we analyzed the regulatory flow among the TFs constituting the regulatory

network (Dobrin et al., 2004b; Ma et al., 2004a; Yu and Gerstein, 2006). Figures 2-3C and 2-3D show the regulatory interactions present between at least two TFs in *E. coli* and *B. subtilis*. Note that, all the TFs of group A are at the top of the network hierarchy initiating the regulatory interactions in the network of TFs. The regulatory flow follows an order, from TF members of group A to B to C, and there are no regulatory interactions from members of group C directed to B or A, indicating no feedback at the level of transcriptional regulation from the bottom to the top. However, there are some regulatory interactions between members of the same group and from members of group B towards members of group A. Other approaches for constructing hierarchical networks, such as the bottom-up strategy (Yu and Gerstein, 2006), using TF-TF network did not change our observations that group A shows a preference to occur at the top of the hierarchy while group C appears at the bottom of the hierarchical network. The partitioning of transcriptional network into big, intermediate and small regulons illustrates how the network components could be structured on a chromosome in a scale-free distribution, observed in various biological networks (Barabasi and Albert, 1999; Hartwell et al., 1999). It is possible to generalize from our observations, that the TFs at the bottom of this hierarchy often correspond to very specific functional roles like those sensing specific environmental conditions (Lagomarsino et al., 2007).

2.2.2.3 Absolute and average mRNA abundance of TFs suggests correlation with regulon size and network hierarchy in *E. coli*

It is believed that global regulators should be present in higher concentrations in the cell compared to local or dedicated TFs (Elf et al., 2007). In fact, it is known to be valid for nucleoid-associated proteins and other global regulators like CRP, Lrp and Fur in *E. coli,* whose protein concentrations reach more than 1000 units per cell (Chen et al., 2001; Luijsterburg et al., 2006). On the other hand, the number of TF proteins of LacI, a dedicated TF for lactose utilization, rises from around 5 to a maximum of 20 upon induction of lactose (Droge and Muller-Hill, 2001). Indeed, early genomic approaches to study gene expression patterns on a genomic scale which exploited the codon frequency bias of highly expressed cellular machinery like ribosomal, transcription and cheparone associated classes, have shown that sequence specific TFs are generally poorly expressed (Karlin and Mrazek, 2000). However, so far, no global analysis has been performed to compare TF protein concentration with their connectivity and network hierarchy. Therefore to address this, we used mRNA profile data from two experiments performed in the M9+glucose medium, in which the absolute number of mRNA molecules were quantified (Covert et al., 2004; Liu et al., 2005; Lu et al., 2007). We obtained the number of

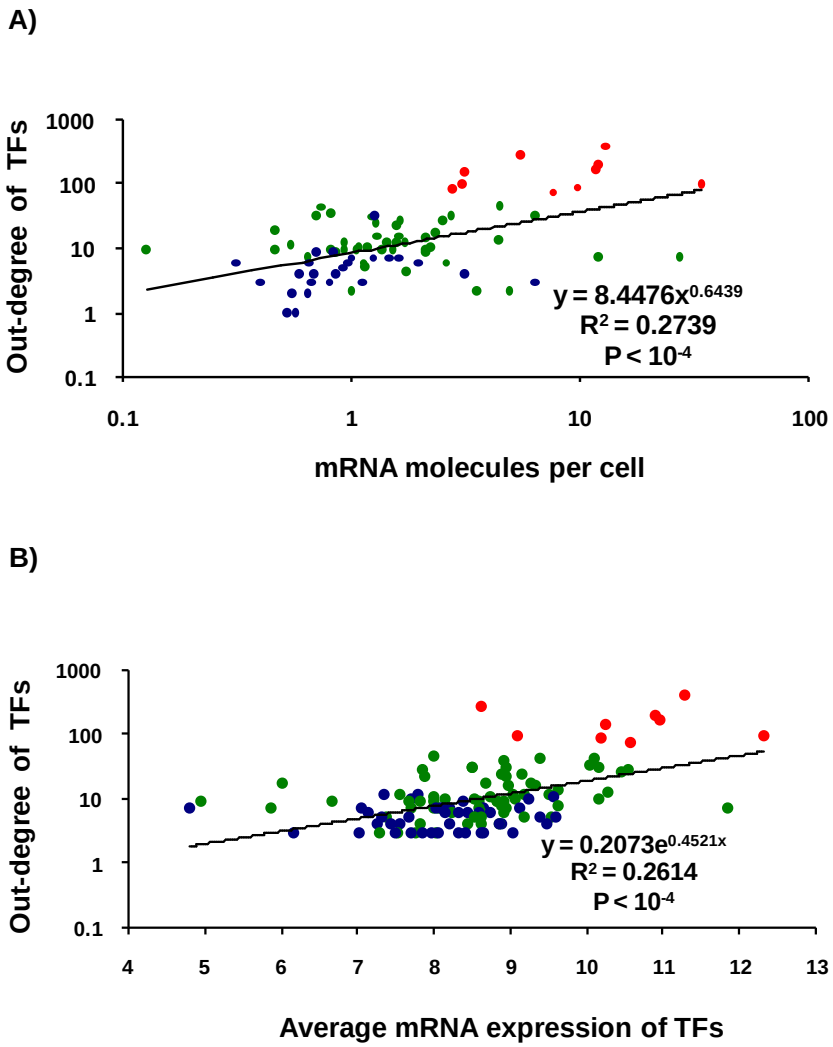


Figure 2-4: A) Relationship between mRNA abundance and out-degree of a TF in the regulatory network of *E. coli*. TFs are colored as per their grouping in Figure 2-3 with big regulons in red, intermediate ones in green and small regulons in blue. Bigger the regulon, stronger is its tendency to be expressed in higher concentrations. B) Relationship between out-degree of a TF and its average mRNA level, calculated after processing and normalizing the expression data according to RMA normalization, as reported by the authors (Faith et al., 2007).

mRNA molecules (per cell) of genes encoding for TFs from this dataset, to see if it correlates with their connectivity and grouping as identified in Figure 2-3 (see Figure 2-4). We found that TFs higher in the network hierarchy had greater number of mRNA molecules per cell associated with them, suggesting that more protein molecules are produced (Figure 2-4A). To investigate further, the relationship between concentration of a TF and its network hierarchy, we compared TF's outdegree against its average gene expression using a large compendium of *E. coli* microarrays reported recently (Faith et al., 2007). We found that TF's outdegree and its average mRNA level across experiments follows the hierarchy described above (Figure 2-4B). Our

results suggest that several regulators from group C identified in Figure 2-3A are poorly expressed, consistent with previous observations that two-component systems which are enriched in group C and are proximal on the chromosome show poor predicted expression values using codon usage measures (Janga et al., 2007a; Karlin and Mrazek, 2000). If we assume that mRNA formation is a determining step in protein synthesis, these data might correspond to the absolute protein concentrations of the respective TFs per bacterial cell implying a correlation between a TF's out-degree and its concentration, extending upon previous studies (Lozada-Chavez et al., 2008; Martinez-Antonio et al., 2008; Seshasayee et al., 2009).These observations clearly indicate that the concentration of a TF is related to the way it is encoded on the chromosome with respect to its target genes, with local TFs regulating few genes present in physical proximity to their target genes and global TFs facilitating the regulation of many genes by increasing their cellular concentration. Indeed it has been postulated using simulations that low copy number TFs need to colocalize with their targets to enable a rapid and reliable gene regulation, confirming the need to place low copy local TFs in physical proximity to their targets in the genome (Kolesov et al., 2007). Proteome profiles for TFs were limited to a countable number until recently when two massive proteomic experiments were reported for *E. coli* (Ishihama et al., 2008; Lu et al., 2007). Excluding the nucleoid-associated proteins which are discussed below, we could obtain protein concentrations for 25 TFs belonging to different levels of *E. coli* network from these experiments (Figure 2-5). Consistent with our observations at mRNA level, TFs with high intracellular levels corresponded with high out-degree when their protein concentration is plotted as a function of the number of target genes. With respect to NAPs, these high-throughput experiments confirm their high abundance reported almost ten years ago using quantitative western blot analyses (Ali Azam et al., 1999). Indeed, a closer look at the peak expression by the same authors suggested that the production of these NAPs is distributed along the bacterial growth-phases (see Figure 2-6). The high cellular levels of these proteins with concentrations varying from 20000 and 50000 units made it possible to estimate that on an average each monomer may bind every 500 bp along the genomic DNA (Ali Azam et al., 1999). In summary, in agreement with the data for mRNAs, we observe that protein abundance correlates with the out-degree of a TF in the network, with NAPs being particularly abundant and expressed in a growth-phase dependent manner, possibly to re-structure the nucleoid, facilitating the running of particular transcriptional programs depending on growth phase status (see below), (Ali Azam et al., 1999; Luijsterburg et al., 2006; Marr et al., 2008). Therefore, one can hypothesize from these results that the scalefree structure of bacterial transcriptional regulatory networks is encoded in the

chromosome itself and that genome organization of bacterial chromosomes might indeed be influenced by their TRNs.

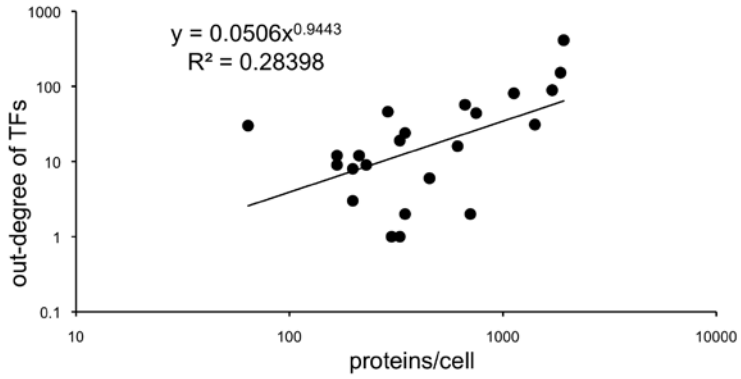


Figure 2-5: Number of proteins/cell for TFs, as a function of the number of genes transcriptionally regulated by it (excluding NAPs as their protein levels are shown in Figure 2-6D and ArcA and NarL which are known to be poorly expressed in aerobic condition where the experiment is performed). Note that the proteomic data is available for only 25 TFs.

2.2.2.4 A conceptual model for the structuring of regulatory networks in bacteria

In the integrated model we propose here (Figure 2-6), the biophysical aspects of TFs for reaching their DNA-binding sites might be the main driving force for structuring the regulatory networks in bacteria as we know presently. This conceptual model is supported by the following observations and evidences:

1) TFs governing small regulons are located close to their regulated genes on the chromosome and this spatial arrangement together with the fact that transcription and translational mechanisms occur simultaneously, should favor that the newly synthesized protein can contact quickly its target DNA through the sliding and hopping mechanism as was shown in the case of LacI (Elf et al., 2007; Wang et al., 2006) (Figure 2-6C). These local regulators are normally expressed in lower cellular concentrations as they would be required sporadically to regulate few operons whose products have dedicated functions. For instance, regulation of alternative carbon sources in *E. coli* is mainly governed by the global regulator CRP and a group of local TFs controlling small regulons which are located proximally on the chromosome with respect to their target genes (Figure 2-6B). The role of the products encoded in these small regulons is to transport and carry out the first catabolic steps of alternative sugars until their catabolism converges in the glycolysis pathway. Additionally, note that most of these TFs in bacteria are autoregulated (Martinez-Antonio et al., 2008). Thus, this sliding mechanism could be a generalized strategy for a quicker and tighter control of TFs over their own expression (Alon, 2007a).

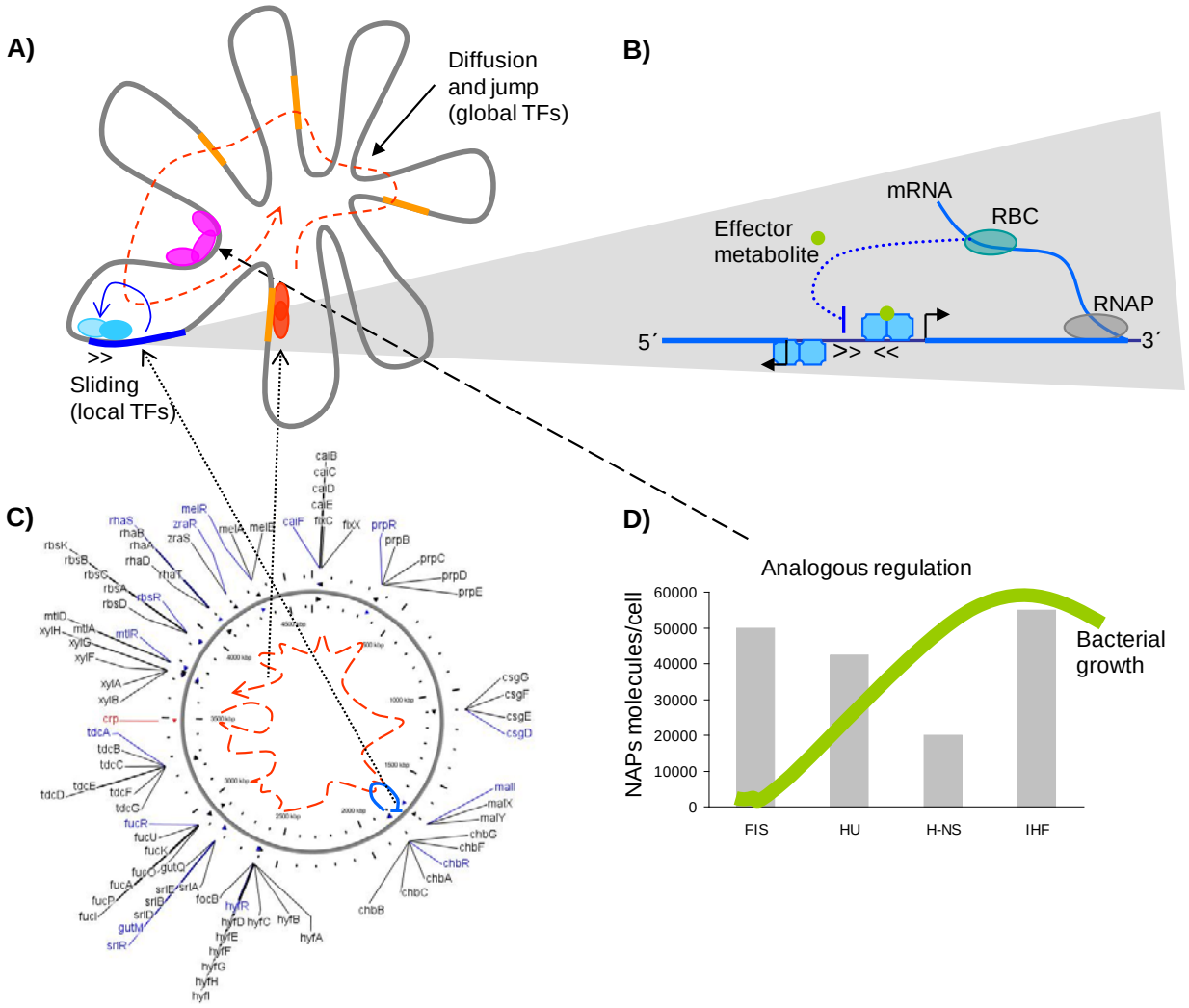


Figure 2-6: Integrated model of transcriptional regulatory network in bacteria (A) combined model representing various factors involved (B) activity and mechanistic basis for the functioning of local TFs (C) an example of global and local TFs co-regulating genes involved in the uptake of carbon sources in E. coli (D) protein abundance of different nucleoid-associated proteins along the growth-phases, acting as analog regulators.

2) In contrast, global regulators which are distantly located with respect to the large number of genes they regulate employ a different strategy. Targeting DNA seems to be accurately managed by raising the concentration of the respective TFs and the actual mechanism used for binding DNA would be 3-D diffusion and jumping between the DNA strands (Figure 2-6A and CRP path in Figure 2-6B). The large cellular concentrations of these proteins might be maintained, in part, given that most global regulators are autoregulated in both positive and negative manner (Martinez-Antonio et al., 2008). Such a mechanism would also make sure that the concentrations of these proteins are maintained at high intracellular levels.

3) A third major player for gene regulation in bacteria is the way the DNA molecule is packed into nucleoids (Ali Azam et al., 1999; Luijsterburg et al., 2006; Zimmerman, 2006). Recent studies provide evidence that the DNA molecule is organized into loops of different lengths (10-100 kbp) which make it possible for some DNA regions to be spatially proximal which would otherwise be distant on a linear molecule of DNA (Kepes, 2004; Marenduzzo et al., 2007; Postow et al., 2004; Riva et al., 2008). Although the exact co-ordinates of these DNA-loops is yet to be unveiled even in well-studied systems like *E. coli,* it is known that nucleoid associated proteins (NAPs) are specifically engaged in structuring DNA depending on the growth condition. These proteins bridge or bend the DNA molecule facilitating DNA loops and nucleoid´s structuring (Luijsterburg et al., 2006; Zimmerman, 2006). In particular, NAPs are shown to express in growth-phase dependent manner with FIS at the beginning of stationary phase, HNS in the mid-exponential and IHF in the arrested phase (see Figure 2-6D) (Ali Azam et al., 1999). These observations suggest that NAPs might structure the DNA molecule in a different way depending on the growth phase and this action should facilitate or predispose off only a section of the DNA-template for the activity of global and local regulators and the running of specific transcriptional programs. Accordingly, it has been suggested that NAPs act as analog regulators whereas the rest of the TFs responding to specific conditions (e.g. by binding signal effectors) act as digital regulators (Marr et al., 2008; Travers and Muskhelishvili, 2007) (Figure 2-6C).

2.3 DISCUSSION & CONCLUSION

Our structural analysis of the transcriptional cross-regulatory network in *E. coli* suggests that regulatory interactions between TFs are predominantly positive, while autoregulatory interactions are mostly negative. We also note that there are striking topological differences between the subnetworks controlling metabolic activities, such as carbon metabolism, and that controlling developmental processes; the former encompasses many parallel short transcriptional cascades and multiple FFLs, each enabling the use of one alternative carbon source, while the latter involves long and intertwined regulatory cascades. These long transcriptional cascades typically include multiple autoactivated intermediate TFs, as well as regulatory circuits between TFs and sigma factors in the case of biofilm formation.

We further observe that TFs acting at the end of these regulatory cascades often belong to two-component systems. This topology suggests that cell homeostasy is maintained through multiple regulatory cascades with commonly autorepressed TFs, while the regulatory memory within the network is preserved by the sequential activation of TFs and by multi-element circuits at the core of the network. Downstream of the hierarchical network, two-component systems

can memorise transient external signals through autoactivation loops, thus acting as molecular switches enabling the coexistence of alternative phenotypes.

As shown in a recent study, the *E. coli* cross-regulatory network appears to be robust to tolerate the rewiring between members high and low in the network hierarchy (Isalan et al., 2008). This study also indicated that the allosteric signals are the mandatory input elements for network function. Thus, TFs present in a condition different from the natural one(s) would have limited activity due to the absence of their effector signals. In this respect, a proper global understanding of the organisation of the *E. coli* transcriptional network (combining sigma and TFs) could contribute to the interpretation of network-rewiring experiments as well as foster more efficient design of synthetic regulatory circuits.

It is important to note that the generality of the observed organization of the *E. coli* transcriptional cross-regulatory network remains to be assessed. Nevertheless, a more comprehensive picture of the network organisation in bacteria will progressively be drawn as additional regulatory elements such as small RNAs, anti-sigma factors and riboswitches are integrated (Gama-Castro et al., 2008). In addition, the combination of transcriptional and metabolic networks should provide important insights by linking effector metabolites and regulatory elements. Clearly, variations in regulatory network topology might be expected in the case of bacteria with asymmetric cell division (mostly alpha-proteobacteria), where the offspring asymmetric cells cause a transient genetic asymmetry that triggers different developmental processes, such as the formation of stalked and swarmer cells in Caulabacter or vegetative and spore-forming cells in Bacillus (Ausmees and Jacobs-Wagner, 2003; Dworkin, 2003; Dworkin and Losick, 2001; Hilbert and Piggot, 2004; Yudkin and Clarkson, 2005). Future comparisons between network topologies for different model systems should further enhance our understanding of regulatory network organization and its conservation or variations among different bacterial phyla.

Our analysis linking the transcriptional hierarchy, genome organization and expression dynamics of TFs suggest that TFs high up in the hierarchy are detected in higher mRNA and protein molecules per cell, reflecting their need to be expressed in higher concentrations to regulate target genes located dispersedly on the chromosome. In contrast to big regulons, local or dedicated TFs (lower in the network hierarchy) were found to be expressed in much lower concentrations explaining the reasons for their proximity on the chromosome to their target genes. These observations give insights into how the scale-free structure of transcriptional networks can be encoded on the chromosome to drive the kinetics and concentration gradients of TFs, depending on the number of genes they regulate and could facilitate the horizontal

transfer of local environment-specific transcriptional modules. Although our distance calculations do not take into account the three dimensional topology of the chromosome under a given cellular condition, it is easy to note that the chromosomal proximity of TFs to their targets in the case of small regulons can not be explained due to chance alone. While in the case of global TFs one can argue that as they regulate several genes, their average linear chromosomal distance could be an over-estimation of intracellular proximity considering the dynamic nature of the nucleoid. However, global TFs with their fuzzy binding sites in contrast to local TFs could complement their affinity to targets by increasing their concentrations to a sufficient degree when needed (Kolesov et al., 2007; Lozada-Chavez et al., 2008). Thus, our results suggest that transcriptional regulatory networks play an important role in genome organization by shaping the organization of genes in genomes. These observations illustrate how bacteria as simple biological systems fit predicted theoretical principles in order to optimize their cellular performance in a compacted genome.

2.4 METHODS

2.4.1 Identification of regulon groups

To identify the different regulon groups based on normalized regulon size and normalized average chromosomal distance between TF and its TGs in a regulon, we used K-means clustering implemented in cluster (de Hoon et al., 2004). To find the number of distinct clusters present in the data we first varied the number of clusters (parameter - number of clusters in K-means clustering) to identify how many times the optimal solution has been found in 1000 runs using euclidean distance as the similarity metric. We found that when the number of clusters was set to 3 the optimal solution was found in 350 times out of 1000 runs while when the numbers of clusters was set to 2,4,5 the optimal solution was reached in 120, 167 and 84 times respectively, suggesting that the number of clusters in the set is indeed 3. Similar approaches have been used by others in calculating the significance of clusters with other clustering approaches, as principled clustering frequently results in suboptimal solutions in a single run (Slonim et al., 2006).

To determine the composition of the clusters, we ran the K-means clustering algorithm using 3 as the number of clusters and 1000 as the number of runs. However, since different runs of the k-means clustering algorithm may not give the same final clustering solution, we repeated this experiment 10 times and finally took a consensus of the groupings identified in these runs. We repeated the whole approach to identify the distinct clusters in *B. subtilis*.

2.4.2 Estimating the statistical significance of the regulon groups

To calculate the probability of expecting the chromosomal distances seen in each regulon group by chance, we compared the average chromosomal distance observed in each regulon group against the average chromosomal distances seen in 1000 randomly generated regulon groups obtained by preserving the number of regulatory interactions for each TF in a regulon group. Such a randomization preserves the number of TFs and the interactions in a regulon group but still associates to randomly selected genes in the complete genome thus preserving the topology of the regulon group while shuffling the genomic organization of the targets with respect to their regulating TF.

Statistical significance was assessed based on (i) Z-score, calculated as the number of standard deviations the observed value is away from the randomly expected mean. This is obtained as the ratio between the differences of the observed, x, and random expected, μ, values to the standard deviation, σ i.e., $Z = (x - \mu)/\sigma$) and (ii) p-values, defined as the fraction of the 1000 random trails which showed a value≥ what was observed in the real dataset.

REFERENCES

Aizawa, S. I. and Kubori, T. (1998). Bacterial flagellation and cell division. *Genes Cells* **3**, 625-34.

Aldridge, P. and Hughes, K. T. (2002). Regulation of flagellar assembly. *Curr Opin Microbiol* **5**, 160-5.

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. and Ishihama, A. (1999). Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J Bacteriol* **181**, 6361-70.

Alon, U. (2007a). An Introduction to Systems Biology: Design Principles of Biological Circuits. London. UK.: Chapman & Hall/CRC.

Alon, U. (2007b). Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**, 450-61.

Ausmees, N. and Jacobs-Wagner, C. (2003). Spatial and temporal control of differentiation and cell cycle progression in Caulobacter crescentus. *Annu Rev Microbiol* **57**, 225-47.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. and Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science* **305**, 1622-5.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. et al. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**, 1337-42.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-12.

Berg, O. G., Winter, R. B. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20**, 6929-48.

Bettenbrock, K., Sauter, T., Jahreis, K., Kremling, A., Lengeler, J. W. and Gilles, E. D. (2007). Correlation between growth rates, EIIACrr phosphorylation, and intracellular cyclic AMP levels in Escherichia coli K-12. *J Bacteriol* **189**, 6891-900.

Browning, D. F. and Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57-65.

Cai, L., Friedman, N. and Xie, X. S. (2006). Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358-62.

Cases, I. and de Lorenzo, V. (2005). Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol* **3**, 105-18.

Chantratita, N., Wuthiekanun, V., Boonbumrung, K., Tiyawisutsri, R., Vesaratchavest, M., Limmathurotsakul, D., Chierakul, W., Wongratanacheewin, S., Pukritiyakamee, S., White, N. J. et al. (2007). Biological relevance of colony morphology and phenotypic switching by Burkholderia pseudomallei. *J Bacteriol* **189**, 807-17.

Chen, S., Hao, Z., Bieniek, E. and Calvo, J. M. (2001). Modulation of Lrp action in Escherichia coli by leucine: effects on non-specific binding of Lrp to DNA. *J Mol Biol* **314**, 1067-75.

Cherstvy, A. G., Kolomeisky, A. B. and Kornyshev, A. A. (2008). Protein--DNA interactions: reaching and recognizing the targets. *J Phys Chem B* **112**, 4741-50.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92-6.

de Hoon, M. J., Imoto, S., Nolan, J. and Miyano, S. (2004). Open source clustering software. *Bioinformatics* **20**, 1453-4.

Dekel, E., Mangan, S. and Alon, U. (2005). Environmental selection of the feed-forward loop circuit in gene-regulation networks. *Phys Biol* **2**, 81-8.

Deutscher, J., Francke, C. and Postma, P. W. (2006). How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiol Mol Biol Rev* **70**, 939-1031.

Dobrin, R., Beg, Q. K., Barabasi, A. L. and Oltvai, Z. N. (2004a). Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics* **5**, 10.

Dobrin, R., Beg, Q. K., Barabasi, A. L. and Oltvai, Z. N. (2004b). Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics* **5**, 10.

Droge, P. and Muller-Hill, B. (2001). High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *Bioessays* **23**, 179-83.

Dworkin, J. (2003). Transient genetic asymmetry and cell fate in a bacterium. *Trends Genet* **19**, 107-12.

Dworkin, J. and Losick, R. (2001). Differential gene expression governed by chromosomal spatial asymmetry. *Cell* **107**, 339-46.

Ehrlich, G. D., Hu, F. Z., Shen, K., Stoodley, P. and Post, J. C. (2005). Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin Orthop Relat Res*, 20-4.

Elf, J., Li, G. W. and Xie, X. S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191-4.

Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**, 1216-21.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H. et al. (2008). RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**, D120-4.

Golding, I., Paulsson, J., Zawilski, S. M. and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025-36.

Gowers, D. M., Wilson, G. G. and Halford, S. E. (2005). Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proc Natl Acad Sci U S A* **102**, 15883-8.

Griffith, K. L. and Wolf, R. E., Jr. (2004). Genetic evidence for pre-recruitment as the mechanism of transcription activation by SoxS of Escherichia coli: the dominance of DNA binding mutations of SoxS. *J Mol Biol* **344**, 1-10.

Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31**, 60-3.

Guerrero, A., Jain, N., Goldman, D. L. and Fries, B. C. (2006). Phenotypic switching in Cryptococcus neoformans. *Microbiology* **152**, 3-9.

Gutierrez-Rios, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R. and Collado-Vides, J. (2003). Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res* **13**, 2435-43.

Handke, L. D., Conlon, K. M., Slater, S. R., Elbaruni, S., Fitzpatrick, F., Humphreys, H., Giles, W. P., Rupp, M. E., Fey, P. D. and O'Gara, J. P. (2004). Genetic and phenotypic

analysis of biofilm phenotypic variation in multiple Staphylococcus epidermidis isolates. *J Med Microbiol* **53**, 367-74.

Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.

Hengge-Aronis, R. (2002). Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. *Microbiol Mol Biol Rev* **66**, 373-95, table of contents.

Hilbert, D. W. and Piggot, P. J. (2004). Compartmentalization of gene expression during Bacillus subtilis spore formation. *Microbiol Mol Biol Rev* **68**, 234-62.

Hooshangi, S., Thiberge, S. and Weiss, R. (2005). Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc Natl Acad Sci U S A* **102**, 3581-6.

Hu, L., Grosberg, A. Y. and Bruinsma, R. (2008). Are DNA Transcription Factor Proteins Maxwellian Demons? *Biophys J*.

Hu, Z., Ng, D. M., Yamada, T., Chen, C., Kawashima, S., Mellor, J., Linghu, B., Kanehisa, M., Stuart, J. M. and DeLisi, C. (2007). VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res* **35**, W625-32.

Ihmels, J., Bergmann, S. and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993-2003.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-7.

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M. and Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840-5.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M. J. and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* **9**, 102.

Ishii, T., Yoshida, K., Terai, G., Fujita, Y. and Nakai, K. (2001). DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Res* **29**, 278-80.

Jacob, F. (1970). *La Logique du Vivant, Une Histoire de L'Hérédité*. Paris: Gallimard.

Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960). [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci* **250**, 1727-9.

Janga, S. C., Salgado, H., Collado-Vides, J. and Martinez-Antonio, A. (2007a). Internal versus external effector and transcription factor gene pairs differ in their relative chromosomal position in Escherichia coli. *J Mol Biol* **368**, 263-72.

Janga, S. C., Salgado, H., Martinez-Antonio, A. and Collado-Vides, J. (2007b). Coordination logic of the sensing machinery in the transcriptional regulatory network of Escherichia coli. *Nucleic Acids Res* **35**, 6963-6972.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651-4.

Kalir, S. and Alon, U. (2004). Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117**, 713-20.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M. G. and Alon, U. (2001). Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292**, 2080-3.

Kamoun, S. and Kado, C. I. (1990). Phenotypic Switching Affecting Chemotaxis, Xanthan Production, and Virulence in Xanthomonas campestris. *Appl Environ Microbiol* **56**, 3855-3860.

Karlin, S. and Mrazek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* **182**, 5238-50.

Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A* **102**, 13773-8.

Kepes, F. (2004). Periodic transcriptional organization of the E.coli genome. *J Mol Biol* **340**, 957-64.

Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S. and Mirny, L. A. (2007). How gene order is influenced by the biophysics of transcription regulation. *Proc Natl Acad Sci U S A* **104**, 13948-53.

Korbel, J. O., Jensen, L. J., von Mering, C. and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22**, 911-7.

Lagomarsino, M. C., Jona, P., Bassetti, B. and Isambert, H. (2007). Hierarchy and feedback in the evolution of the Escherichia coli transcription network. *Proc Natl Acad Sci U S A* **104**, 5516-20.

Liu, M., Durfee, T., Cabrera, J. E., Zhao, K., Jin, D. J. and Blattner, F. R. (2005). Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. *J Biol Chem* **280**, 15921-7.

Lozada-Chavez, I., Angarica, V. E., Collado-Vides, J. and Contreras-Moreira, B. (2008). The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *J Mol Biol* **379**, 627-43.

Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117-24.

Luijsterburg, M. S., Noom, M. C., Wuite, G. J. and Dame, R. T. (2006). The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *J Struct Biol* **156**, 262-72.

Ma, H. W., Buer, J. and Zeng, A. P. (2004a). Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**, 199.

Ma, H. W., Kumar, B., Ditges, U., Gunzer, F., Buer, J. and Zeng, A. P. (2004b). An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res* **32**, 6643-9.

Maas, W. K., Maas, R., Wiame, J. M. and Glansdorff, N. (1964). Studies On The Mechanism Of Repression Of Arginine Biosynthesis In Escherichia Coli. I. Dominance Of Repressibility In Zygotes. *J Mol Biol* **78**, 359-64.

Macnab, R. M. (2003). How bacteria assemble flagella. *Annu Rev Microbiol* **57**, 77-100.

Madan Babu, M. and Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **31**, 1234-44.

Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004). DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res* **32**, D75-7.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* **100**, 11980-5.

Mangan, S., Itzkovitz, S., Zaslaver, A. and Alon, U. (2005). The Incoherent Feed-forward Loop Accelerates the Response-time of the gal System of Escherichia coli. *J Mol Biol*.

Mangan, S., Itzkovitz, S., Zaslaver, A. and Alon, U. (2006). The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. *J Mol Biol* **356**, 1073-81.

Mangan, S., Zaslaver, A. and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* **334**, 197-204.

Marenduzzo, D., Faro-Trindade, I. and Cook, P. R. (2007). What are the molecular ties that maintain genomic loops? *Trends Genet* **23**, 126-33.

Marr, C., Geertz, M., Hutt, M. T. and Muskhelishvili, G. (2008). Dissecting the logical types of network control in gene expression profiles. *BMC Syst Biol* **2**, 18.

Martin, R. G., Gillette, W. K. and Rosner, J. L. (2000). Promoter discrimination by the related transcriptional activators MarA and SoxS: differential regulation by differential binding. *Mol Microbiol* **35**, 623-34.

Martinez-Antonio, A. and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**, 482-9.

Martinez-Antonio, A., Janga, S. C., Salgado, H. and Collado-Vides, J. (2006a). Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. *Trends Microbiol* **14**, 22-7.

Martinez-Antonio, A., Janga, S. C., Salgado, H. and Collado-Vides, J. (2006b). Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. *Trends Microbiol* **14**, 22-27.

Martinez-Antonio, A., Janga, S. C. and Thieffry, D. (2008). Functional organisation of Escherichia coli transcriptional regulatory network. *J Mol Biol* **381**, 238-47.

Massey, R. C., Buckling, A. and Peacock, S. J. (2001). Phenotypic switching of antibiotic resistance circumvents permanent costs in Staphylococcus aureus. *Curr Biol* **11**, 1810-4.

Menchaca-Mendez, R., Janga, S. C. and Collado-Vides, J. (2005). The network of transcriptional interactions imposes linear constrains in the genome. *Omics* **9**, 139-45.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* **298**, 824-7.

Moreno-Campuzano, S., Janga, S. C. and Perez-Rueda, E. (2006). Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes--a genomic approach. *BMC Genomics* **7**, 147.

Murugan, R. (2007). Generalized theory of site-specific DNA-protein interactions. *Phys Rev E Stat Nonlin Soft Matter Phys* **76**, 011901.

O'Toole, G., Kaplan, H. B. and Kolter, R. (2000). Biofilm formation as microbial development. *Annu Rev Microbiol* **54**, 49-79.

Pal, C., Papp, B. and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**, 1372-5.

Perez-Rueda, E. and Collado-Vides, J. (2000). The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res* **28**, 1838-47.

Postow, L., Hardy, C. D., Arsuaga, J. and Cozzarelli, N. R. (2004). Topological domain structure of the Escherichia coli chromosome. *Genes Dev* **18**, 1766-79.

Price, M. N., Huang, K. H., Arkin, A. P. and Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15**, 809-19.

Pruss, B. M. and Matsumura, P. (1997). Cell cycle regulation of flagellar genes. *J Bacteriol* **179**, 5602-4.

Ptashne, M. and Gann, A. (1997). Transcriptional activation by recruitment. *Nature* **386**, 569-77.

Resendis-Antonio, O., Freyre-Gonzalez, J. A., Menchaca-Mendez, R., Gutierrez-Rios, R. M., Martinez-Antonio, A., Avila-Sanchez, C. and Collado-Vides, J. (2005). Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet* **21**, 16-20.

Richter, P. H. and Eigen, M. (1974). Diffusion controlled reaction rates in spheroidal geometry. Application to repressor--operator association and membrane bound enzymes. *Biophys Chem* **2**, 255-63.

Riggs, A. D., Bourgeois, S. and Cohn, M. (1970). The lac repressor-operator interaction. 3. Kinetic studies. *J Mol Biol* **53**, 401-17.

Riva, A., Carpentier, A. S., Barloy-Hubler, F., Cheron, A. and Henaut, A. (2008). Analyzing stochastic transcription to elucidate the nucleoid's organization. *BMC Genomics* **9**, 125.

Ronen, M., Rosenberg, R., Shraiman, B. I. and Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A* **99**, 10555-60.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J. et al. (2006). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34**, D394-7.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-76.

Seshasayee, A. S., Fraser, G. M., Babu, M. M. and Luscombe, N. M. (2009). Principles of transcriptional regulation and evolution of the metabolic system in E. coli. *Genome Res* **19**, 79-91.

Shah, I. M. and Wolf, R. E., Jr. (2004). Novel protein--protein interaction between Escherichia coli SoxS and the DNA binding determinant of the RNA polymerase alpha subunit: SoxS functions as a co-sigma factor and redeploys RNA polymerase from UP-element-containing promoters to SoxS-dependent promoters during oxidative stress. *J Mol Biol* **343**, 513-32.

Shapiro, J. A. (1998). Thinking about bacterial populations as multicellular organisms. *Annu Rev Microbiol* **52**, 81-104.

Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-8.

Shimamoto, N. (1999). One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements. *J Biol Chem* **274**, 15293-6.

Slonim, N., Elemento, O. and Tavazoie, S. (2006). Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol* **2**, 2006 0005.

Snel, B. and Huynen, M. A. (2004). Quantifying modularity in the evolution of biomolecular systems. *Genome Res* **14**, 391-7.

Stoodley, P., Sauer, K., Davies, D. G. and Costerton, J. W. (2002). Biofilms as complex differentiated communities. *Annu Rev Microbiol* **56**, 187-209.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**, 1-4.

Thieffry, D., Huerta, A. M., Perez-Rueda, E. and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays* **20**, 433-40.

Travers, A. and Muskhelishvili, G. (2007). A common topology for bacterial and eukaryotic transcription initiation? *EMBO Rep* **8**, 147-51.

Ulrich, L. E., Koonin, E. V. and Zhulin, I. B. (2005). One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* **13**, 52-6.

Wang, Y. M., Austin, R. H. and Cox, E. C. (2006). Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys Rev Lett* **97**, 048302.

Warren, P. B. and ten Wolde, P. R. (2004). Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. *J Mol Biol* **342**, 1379-90.

Winter, R. B., Berg, O. G. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **20**, 6961-77.

Wu, W. S., Li, W. H. and Chen, B. S. (2006). Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics* **7**, 421.

Xie, X. S., Choi, P. J., Li, G. W., Lee, N. K. and Lia, G. (2008). Single-molecule approach to molecular biology in living bacterial cells. *Annu Rev Biophys* **37**, 417-44.

Yu, H. and Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A* **103**, 14724-31.

Yu, J., Xiao, J., Ren, X., Lao, K. and Xie, X. S. (2006). Probing gene expression in live cells, one protein molecule at a time. *Science* **311**, 1600-3.

Yudkin, M. D. and Clarkson, J. (2005). Differential gene expression in genetically identical sister cells: the initiation of sporulation in Bacillus subtilis. *Mol Microbiol* **56**, 578-89.

Zimmerman, S. B. (2006). Shape and compaction of Escherichia coli nucleoids. *J Struct Biol* **156**, 255-61.

3 Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes

CONTENTS OF CHAPTER 3

OUTLINE

Recent advances in molecular techniques and high-resolution imaging are beginning to provide exciting insights into the higher order chromatin organization within the cell nucleus and its influence on eukaryotic gene regulation. This improved understanding of gene regulation also raises fundamental questions about how spatial features might have constrained the organization of genes on eukaryotic chromosomes and how re-arrangements that affect these processes might contribute to disease conditions. In this chapter, I discuss recent studies that highlight the role of spatial components in gene regulation and their impact on genome evolution. I then present evidence for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation. In particular, I show that the target genes of transcription factors for the yeast, *Saccharomyces cerevisiae*, are encoded in a highly ordered manner both across and within the sixteen chromosomes by demonstrating that the target genes of a (i) majority of the TFs are not randomly distributed across chromosomes but show a strong preference to be encoded on specific chromosomes, (ii) significant fraction of the TFs are not randomly distributed within a chromosome, but display a strong preference (or avoidance) to be encoded in regions containing particular chromosomal landmarks such as telomeres and centromeres (iii) majority of the TFs are not randomly scattered but are positionally clustered within a chromosome. These results demonstrate that specific organization of genes that allowed for efficient control of transcription within the nuclear space has been selected during evolution. The framework developed here can be exploited to uncover such higher-order organizational principles in other eukaryotes to provide insights into chromosomal territories, their role in cellular differentiation and transformation, and will have implications for understanding disease conditions that involve chromosomal aberrations.

CONTRIBUTION TO THE WORK IN THIS CHAPTER

Please note that the work presented in this chapter is the result of the following two publications and my contribution to the work excludes the collaboration with Dr. Ana Pombo and Ines De Santiago at MRC Clinical Sciences Centre, London towards the review on eukaryotic genome organization and transcriptional regulation. I performed all other analyses. I would also like to thank Dr. Julio Collado-Vides at UNAM, Mexico for helpful discussions in developing this work.

1) Eukaryotic gene regulation in three dimensions and its impact on genome evolution
M. Madan Babu, Sarath Chandra Janga, Ines de Santiago and Ana Pombo
Curr. Opin. Genet. Dev., 2008, Vol. 18(6):571-582
2) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes
Sarath Chandra Janga, Julio Collado-Vides and M. Madan Babu
Proc. Natl. Acad. Sci. U S A. 105(41): 15761-6, 2008

3.1 INTRODUCTION

Since the discovery of chromatin in 1974 (Kornberg, 1974; Olins and Olins, 1974), it is now well known that eukaryotic genomes are compactly packed into chromatin, the fundamental unit of which is the nucleosome. Such an organization appears to serve two important purposes: (i) they allow for compaction to fit the DNA in the nucleus and (ii) they avoid unnecessary transcription of genes by preventing the RNA polymerase from accessing the promoter regions of genes. Apart from these general functions, chromatin structure is also known to play an important role in DNA replication and repair (Loizou et al., 2006). Nucleosomes consists of ~146 bp of DNA wrapped twice around the core histone octamer (Luger et al., 1997) whose components and additional chromatin proteins can interact to form higher order chromosomal structures. Apart from providing a structural basis, components of the histone octamer could themselves be post-translationally modified by several different proteins. For instance, a class of proteins called the nucleosome remodeling enzymes, either remove the histone octamer from the nucleosome by chemically modifying them or by physically changing the position of the nucleosome to provide access. Importantly, several studies (both recently and in the past) have shown that the individual subunits of the histone octamer in a nucleosome could be chemically modified by an acetyl group, methyl (mono, di, or tri) group, phosphorylation, ADP ribosylation, ubiquitinylation, and sumoylation (Allfrey et al., 1964; Millar and Grunstein, 2006; Nightingale et al., 2006). Hence, such an organization of DNA into nucleosomes and the plethora of combinatorial possibilities of the modified state of the nucleosome is believed to provide an opportunity to regulate expression of relevant genes in a more sophisticated way, resulting in discrete biological outcomes. This combination of modification states that results in distinct effects in a cell has been conventionally referred to as the histone code (Turner, 1993; Turner, 2007). Thus, nucleosomes are critical to the organization and maintenance of genetic material and their position and modification state can profoundly influence genetic activities such as regulation of gene expression (Kouzarides, 2002; Narlikar et al., 2002).

More generally, the eukaryotic genome compared to its bacterial counterpart is a highly complex system, which is regulated at three major hierarchical levels (Lee and Young, 2000; van Driel et al., 2003). The first is at the DNA sequence level, i.e. the linear organization of transcription units and regulatory sequences. Co-regulated genes organized into clusters in the genome constitute part of these individual functional units. The second is at the chromatin level, which allows switching between different functional states. This level involves the changes in the chromatin structure that are controlled by the interplay of histones and remodeling factors

along with a variety of repressive and activating mechanisms. This regulatory level is linked with the control mechanisms from level one that switch individual genes in the cluster to on and off, depending on the properties of the promoter. The third level is the nuclear level, which includes the dynamic 3D spatial organization of the genome inside the cell nucleus. The nucleus is structurally and functionally compartmentalized and epigenetic regulation of gene expression may involve repositioning of loci in the nucleus through changes in large-scale chromatin structure. There is increasing evidence that such a higher order organization of chromatin arrangement contributes essentially to the regulation of gene expression and other nuclear functions (see (Cremer and Cremer, 2001; Lanctot et al., 2007; van Driel et al., 2003; Zinner et al., 2006)). The territorial organization of chromosomes was known from very early experiments, in which damaged regions of micro-irradiated cell nuclei, visualized in the subsequently prepared metaphase chromosomes, were found to be locally clustered (Zorn et al., 1979). The chromosome territories were later visualized directly by means of in situ hybridization in interspecies somatic hybrid cells (Manuelidis, 1985). There is now convincing evidence that chromosomes in most eukaryotic nuclei occupy distinct volumes in the nuclear space called chromosomal territories separated by intra-chromosomal regions providing evidence for the dynamic nature of the positions occupied by the chromosomes (Cremer et al., 2000; Gasser, 2002; Heun et al., 2001; Kurz et al., 1996; Taddei et al., 2004). Hence, the cross-talk between different chromosomes and genes located within them in the context of metabolic, transcriptional and signaling mechanisms could provide additional layer of complexity to our understanding of the proper functioning of the cell.

Apart from the three dimensional architecture of cell nucleus discussed above (and shown in Figure 3-1) a number of regulatory mechanisms control the movement, organization and regulation of different loci with in the nucleus. Functions and our current understanding of some of these regulatory elements responsible for the regulation at different levels have been discussed in detail in the first half of this chapter. All these observations raise some fundamental questions: has the requirement for transcriptional regulation and their spatial considerations constrained the way in which genes are organized on chromosomes? If yes, in what ways does it affect genome evolution? In the second half of this chapter I discuss the investigations involving the understanding of constraints placed by transcriptional regulation on the organization of genes on the chromosomes in eukaryotic organisms.

(Space left for an enhanced layout of the figure)

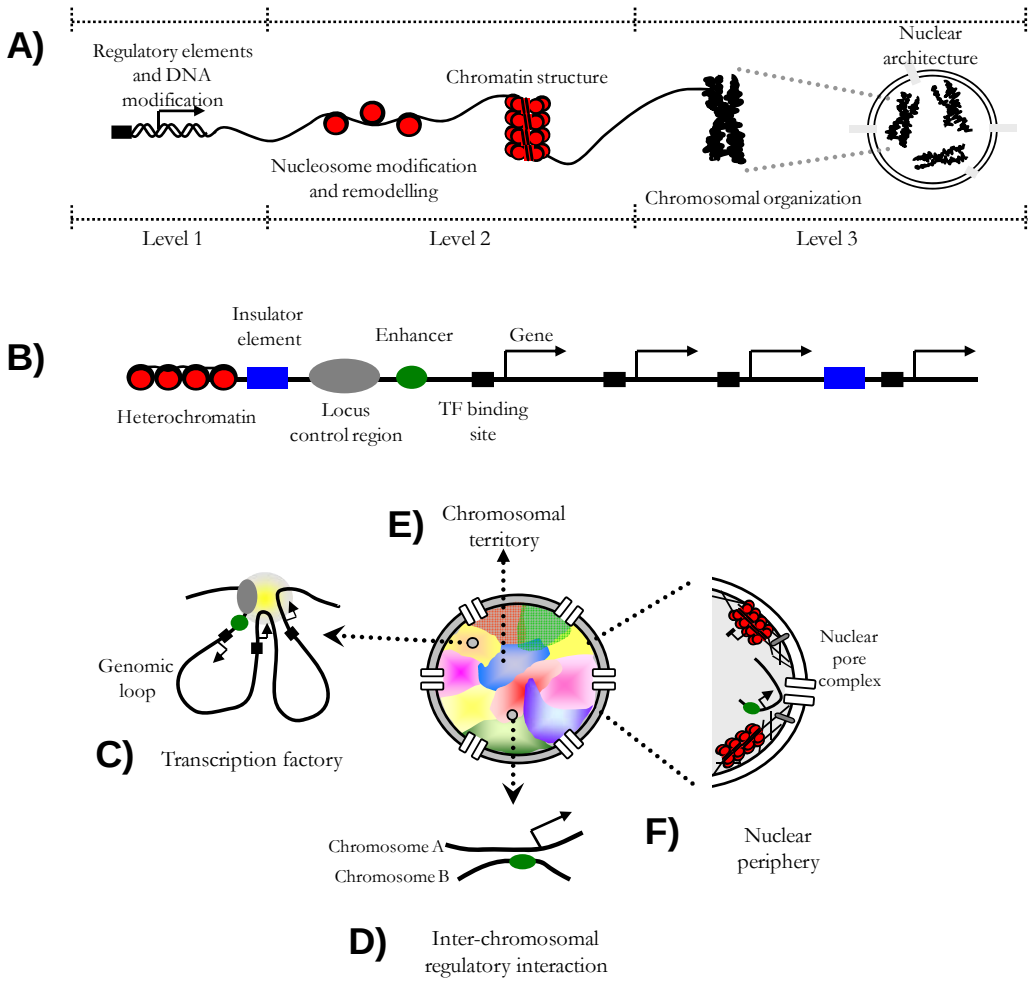


Figure 3-1: (A) Hierarchical organization of eukaryotic genetic material. DNA is wrapped into nucleosomes, which form the chromatin and is ultimately packaged into a chromosome that resides within the nucleus. The first level of regulation includes regulatory elements (e.g., enhancers and insulators), DNA methylation and DNA structure. The second level includes post-translational modification of nucleosomes and remodelling of nucleosomes. The third level of regulation includes chromosomal organization and the nuclear architecture. Features of genome architecture in 3D showing (B) DNA is shown as a black line, a gene is represented as an arrow and the different classes of regulatory elements are shown in various shapes and colors. Insulator elements (blue rectangles) block spread of heterochromatin (red circles) and prevent inappropriate interaction between enhancers (green oval) and unrelated genes. Enhancers can facilitate regulation of nearby genes that may still be a few kilobases away. Locus control region (gray oval) can bring genomic loci that are several kilobases away close to each other to co-ordinate gene expression. The bottommost panel shows various aspects of the spatial component in eukaryotic gene regulation. The nucleus is shown in the center. (C) Different active regions of the same or different chromosomes can associate with the same transcription factory. (D) Enhancers from one chromosome may regulate the expression of genes present on another chromosome via inter-chromosomal interactions. (E) Chromosomes occupy defined volume within the nucleus, called as chromosome territories which are depicted in different colours with significant intermingling mostly at the edges. (F) Genetic material residing near the nuclear periphery has been correlated with gene silencing. One theme which stands out is that regions of the chromosome that interact with lamin and the nuclear inner membrane are largely inactive, in both mammals and yeast, whereas loci that interact with the components of the nuclear pore appear to be transcriptionally active, mostly observed in Drosophila and yeast.

3.2 RESULTS

3.2.1 Eukaryotic genome organization and transcriptional regulation

Though we observe an amazing diversity in the number of chromosomes that eukaryotic organisms encode (*e.g.*, 32 chromosomes in yeast, over 200 in butterflies and 46 in humans), they are packaged in similar ways: each DNA molecule is wrapped around histone proteins to form nucleosomes, which are then condensed in a complex hierarchical manner to make up an entire chromosome (Figure 3-1A). Such an intricate organization of genetic material within the eukaryotic nucleus provides ample opportunities to regulate expression of the encoded genes at many different hierarchical levels. For instance, eukaryotic transcription is dynamically regulated at least at three major levels as shown in Figure 3-1A. The first is at the level of DNA sequence where DNA binding proteins (*e.g.*, transcription factors; TFs) associate with *cis*-regulatory elements (*e.g.*, TF binding sites) to regulate transcription. The second is at the level of chromatin, which allows segments within a chromosomal arm to switch between different transcriptional states, *i.e.*, those that suppress transcription (heterochromatin) and those that allow for gene activation (euchromatin). This involves changes in chromatin structure and nucleosome occupancy, both of which are controlled by the interplay between several factors such as nucleosome remodeling complexes, histone modifications, and a variety of repressive and activating mechanisms (Millar and Grunstein, 2006; Razin et al., 2007). The third is at the level of the entire chromosome (Figure 3-1A) and includes positioning of chromosomes within the nuclear space (*e.g.*, closer to the nuclear periphery or next to internal nuclear compartments) and spatial organization of specific chromosomal loci within the nucleus, both of which are known to influence gene expression (de Laat and Grosveld, 2007; Fraser and Bickmore, 2007; Misteli, 2007; Pombo and Branco, 2007; Schneider and Grosschedl, 2007).

Several studies have investigated these mechanisms in detail and have revealed that such processes involve extensive physical and spatial association between distantly located genomic elements and widespread crosstalk between the different levels. Advancements in molecular techniques and high-resolution imaging (see Table 3-1) have facilitated investigation of the role of spatial component in gene regulation and have provided valuable insights into its importance in gene regulation (de Laat and Grosveld, 2007; Fraser and Bickmore, 2007; Misteli, 2007; Pombo and Branco, 2007; Schneider and Grosschedl, 2007). In this chapter I first discuss recent studies that highlight the importance of spatial component in gene regulation and then present a detailed analysis that addresses how the requirements for gene regulation could have constrained genome organization. Finally, I discuss implications and outline open questions.

Table 3-1: Experimental and computational approaches to study eukaryotic transcriptional regulation.

Experimental approaches	Description
ChIP-chip	Chromatin-bound proteins are covalently linked to DNA by using an *in vivo* crosslinking agent such as formaldehyde (histones can be detected in unfixed chromatin preparations in native ChIP). Chromatin is then sheared and immunoprecipitated (ChIP) using an antibody for a native protein, a tagged version, or a specific post-translational modification. Reversal of the crosslink releases the bound DNA, allowing the enrichment of specific DNA fragments, whose identity is determined by hybridization to a microarray (chip).
ChIP-seq	In ChIP-seq experiments, the immunoprecipitated DNA is directly sequenced using high-throughput sequencing technologies (*e.g.*, Solexa or 454). The sequences are then computationally mapped back to the reference genome. Fragments that were bound by the protein will be more abundant and sequenced several times, providing a direct measure of enrichment.
DamID	The DNA binding protein of interest is fused to an *E. coli* protein, Dam. Dam methylates the N^6 position of the adenine in the sequence GATC, which is expected to occur once in every ~256 bases. Upon binding DNA, the Dam protein preferentially methylates adenine in the vicinity of binding. The DNA is digested by DpnI and DpnII restriction enzymes, which cleave within the non-methylated GATC sequence, and remove fragments that are not methylated. The remaining methylated fragments are amplified by selective PCR and quantified using a microarray.
RNA-TRAP	Newly-made transcripts are detected in crosslinked cells by RNA-FISH using biotinylated probes and probe-RNA-chromatin complexes are amplified with tyramide or directly immunoprecipitated, before PCR analyses.
Chromosome Confirmation Capture (3C)	3C is used to determine which DNA sequences lie close together in 3D space in fixed cells. This typically involves fixation to crosslink DNA sequences that lie next to each other (usually through DNA–protein–DNA links), before cutting with a restriction enzyme, dilution and ligation at low concentration. This favours the ligation of pairs of DNA sequences that are crosslinked after which the reversing of crosslinks allows the ligated DNA to be detected by PCR.
4C	4C technology [chromosome conformation capture on chip (3C-on-chip) or circular chromosome conformation capture (circular-3C)] allows for an unbiased genome-wide search for DNA loci that contact a given locus.
5C	Chromosome Conformation Capture Carbon Copy (5C) is a massively parallel technique, which involves mapping physical interactions between genomic elements and sequencing or microarray analysis of the ligated end products of the 3C technique. 3C typically converts physical chromatin interactions into specific ligation products, which are quantified using high-throughput microarrays or quantitative DNA sequencing using 454-technology as detection methods.
6C	Combines ChIP for a specific chromatin bound protein with 3C-based methods to correlate specific long-range chromatin interactions with the presence of a specific bound protein.
FISH	Fluorescent in situ hybridization (FISH) detects specific DNA sequences and localizes them on cytogenetic preparations of chromosomes or interphase cell nuclei. Cells are hypotonically swollen and dropped on glass slides before hybridization, such that fine structural details might be lost. It uses tagged probes amplified from specific DNA fragments up to single chromosomes, to detect the target sequences. The genomic regions bound by the probe are visualized by fluorescence microscopy.
3D-FISH	A modified FISH procedure that improves the preservation of 3D nuclear structure (3D-FISH), important for spatial mapping of the position of specific genomic sequences within the interphase nuclei. This technique can be slow as

	it requires imaging of multiple image stacks on a small number of nuclei and 3D reconstruction. It can also be combined with protein and RNA localization.
Cryo-FISH	A modified FISH procedure that uses ultra-thin cryosections from sucrose-embedded fixed samples. Sections are 100-200 nm thick. Preservation of ultrastructure is optimized, signal-to-noise ratios are improved and imaging artifacts are minimized. It is ideal for imaging short-range interactions between specific loci or their associations with specific landmarks with higher resolution and faster data collection. Specific cells in their tissue context can be easily investigated. It is not suitable to measure general 3D genomic positioning over large distances.
Single Molecule Imaging (Fluorescence Microscopy)	In Single Molecule Imaging (SMI) of live cells, the molecules of interest are conjugated with fluorophores and introduced into cells. The behavior of multiple fluorescent molecules in cells is then visualized using high-sensitivity video microscopy. The observables in SMI are the position or movement of the fluorescent spots, the fluorescence intensity of individual spots, the fluorescence spectrum or color of individual spots, and the number and distribution of the spots.
Lac-binding-site array	In this approach, *in vivo* visualization of chromatin dynamics is based on lac repressor recognition of direct repeats of the lac operator. The method allows tagging of specific chromosomal sites and thus *in situ* localization *in vivo*. Detection by light microscopy, using GFP-lac-repressor fusion proteins or immunofluorescence, can be complemented by higher-resolution electron microscopy using immunogold staining. This method facilitates the investigation of interphase chromosome dynamics, as well as chromosome segregation during cell division in organisms that lack cytologically condensed chromosomes.
Computational approaches	**Description**
Boolean modelling	In qualitative modelling, kinetic processes are simulated by tracking over discrete time, the state of the system, defined in terms of a coarse range for each variable. The weak specification of such models conserves computer resources needed to explore the space of possible behaviours. Moreover, it provides high-level predictions applying to a whole family of systems. Although simulation of qualitative models can be fast, even a rough exploration of parameter space can become intractable as the size of the system increases, highlighting the need for increasing computer resources and methods to accelerate the parameters' search space. For genes that are naturally found in only two states (*e.g.*, on or off), the trade-off in accuracy may not be high. On the other hand, simple models can, in some cases, predict behaviours that are far from reality.
Deterministic modelling	Deterministic modeling falls into the class of quantitative models. The most popular formalism is the deterministic ordinary differential equations (ODEs) which, when extended to model space, is referred to as partial differential equations (PDEs). Each equation in a set typically represents the rate of change of a species' continuous concentration as a sum or product of, more or less, empirical terms. This accounts for the effect of biological events on the concentration of the species. By definition, the initial state of the system in a deterministic model uniquely sets all future states. As analytical solutions seldom exist, numerical solutions need to be computed (once for each set of parameter values and initial conditions explored).
Stochastic modelling	Molecular interactions involving a small number of objects in a large volume are intrinsically random and cellular behaviour itself sometimes seems to reflect this randomness. Indeed, occurrences of "noise" have been found to be exploited by cells—for instance, to survive a variety of environmental changes or to increase sensitivity in signal transduction processes. To model such stochastic systems, two main methods are used. The first comprises using stochastic differential equations (SDEs; derived from ODEs by adding noise terms to the equations),

	the solutions for which can be numerically obtained either by computing many trajectories (Monte Carlo methods) or approximating their probability distribution and then calculating statistical measures (such as mean and variance). The second is an exact method which can cope with different reaction time-scales or spaces. Within this approach, molecules are modelled individually and reaction events are calculated by their probability.
Monte Carlo simulation	Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo methods are often used when simulating physical and mathematical systems. Because of their reliance on repeated computation and random or pseudo-random numbers, Monte Carlo methods are most suited for computer simulations and tend to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm. There is no single Monte Carlo method; instead, the term describes a large and widely-used class of approaches.
Multi-scale modelling	Multi-scale modeling refers to the modeling of a system at several levels of detail to increase the accuracy and representation of the system as close to reality. For instance, modeling of a chromatin unit, a nucleosome, using a simplified model for rapid discrete molecular dynamics simulations and an all-atom model for detailed structural investigation, would correspond to this class of modelling.
Statistical correlations	Statistical models search for patterns in experimental data. Correlation, regression and cluster analysis are all powerful statistical tools that can identify relationships among measured variables that probably are not attributable to chance. Statistics is also a powerful tool for uncovering the prevalence of a phenomena and evidence for potentially new mechanisms.
Spatial modelling	Spatial modelling takes into account that biological processes take place in heterogeneous and highly structured environments regulating cellular processes in both space and time. While recent technological advances are addressing the dearth of spatial data, theoretical advances are improving computational methods, making it now possible to simulate spatio-temporal models of biological processes in coarse-grained or realistic geometries.
Kinetic modelling	Kinetic modelling supports quantitative hypothesis testing by first translating a diagram into a mechanistic kinetic model. Diagrams typically consist of molecules, complexes, cellular locations and processes. As molecules and complexes can exist in several locations, it is often necessary to define several states for a single molecule — each state is a set of chemical species in a physical place.
Comparative genomics	Comparative genomics permits addressing questions at the sequence level both within and across organisms and their variations across diverse phylogenetic groups. Evolutionary aspects of several cellular elements from genes, regulatory elements, organellar macromolecular complexes to their chromosomal organization can be addressed using computational genome-scale approaches.
Network based approaches	In a network approach, objects are represented as nodes and interactions between objects are represented as links. This permits representation of genome-scale information in a convenient way to identify interesting topological features. One of the features is the presence of hub nodes which are objects connected to an extremely high number of other objects in a system. With the amount of data from several high-throughput technologies, representing interactions between biological molecules as networks has provided us with a general framework to address fundamental biological questions at a systems level. Examples of molecular interactions represented as networks include protein-DNA (transcription network), protein-RNA (post-transcriptional network), protein-protein (post-translational network, signaling and protein complexes) and protein-metabolite (metabolic network) interactions.

and discuss how computational approaches can be helpful in investigating the prevalence of spatial regulatory mechanisms and in understanding their impact on genome evolution. It is important to note that the studies discussed have been carried out in different model systems and that further research is necessary to assess whether particular spatial mechanisms are universal or specific to each system.

3.2.1.1 Long-range interactions involving distal regulatory elements

Regulatory elements in eukaryotes can be spread over several kilobases away from the associated gene. These include binding sites for specific TFs, enhancer elements, locus control regions (LCRs) and insulator elements (Figure 3-1B). TF binding sites are generally close to promoter regions, but enhancer elements, LCRs and insulator elements can be present far away on the chromosome and may influence the expression of more than one gene simultaneously. Enhancers affect expression of nearby genes, whereas LCRs can affect several genes that are distantly located within a genomic locus spanning several kilobases (Dean, 2006). Insulator elements can block promiscuous enhancer-promoter interaction or act as a barrier against the spreading of heterochromatin. The former class of insulators function by forming genomic loops via long-range interactions and the latter class prevents inappropriate gene expression by recruiting nucleosome modifying enzymes (Dorman et al., 2007).

The formation of loops mediated by proteins bound to specific elements along a chromosome appears to have a central role in several processes as it can affect the expression of several genes in a neighborhood (O'Sullivan et al., 2004). Although only a few loops have been analyzed in detail and the nature of the molecular forces that maintain them remain unclear, recent evidence suggests that they are found in several eukaryotes (Dean, 2006) and that the transcriptional machinery itself could be a molecular tie (Grimaud et al., 2006; Marenduzzo et al., 2007; Osborne et al., 2004) (Figure 3-1C). Several studies that have used 3D-FISH, chromosome conformation capture (3C) (Dekker et al., 2002) and its variants 4C, 5C and 6C (Simonis et al., 2007) and live-cell imaging (Muller et al., 2007) support the idea that active transcription units are in close contact within the nuclear space (Osborne et al., 2004; Pombo et al., 1999; Pombo et al., 2000) (see Table 3-1). The results are consistent with a model for genome organization in which active polymerases cluster into transcription 'factories' bringing together distal genes (Figure 3-1C) and where active genes are dynamically organized into shared nuclear subcompartments (Muller et al., 2007; Osborne et al., 2004; Pombo et al., 1999; Pombo et al., 2000). They are also consistent with these *cis*-regulatory elements functioning as insulators, enhancers or LCRs, depending on their positions relative to other

genes. Interestingly, in a recent study, it has been shown that specific 'factories' produce only a particular kind of transcript depending on the promoter type and whether or not the gene contains an intron, supporting the presence of 'specialized' transcription factories (Pombo et al., 1999; Pombo et al., 2000; Xu and Cook, 2008).

The genomic loops involving regulatory elements are dynamic, depend on the transcriptional status of a gene, vary between cell-types in the same organism and may involve several proteins. The beta-globin locus in mouse is the most studied and involves the Hbb-b1 gene (which encodes beta-globin), its LCR and the Eraf gene (encoding an alpha-globin-stabilizing protein) on the same chromosome. This LCR is thought to nucleate a chromatin hub which correlates with expression of globin-related genes. It has been confirmed, by 3C, 4C and RNA-TRAP that the contacts between Hbb-b1, the LCR and Eraf are seen only in erythroid nuclei (in which all three are transcribed) but not in brain cell nuclei (in which Hbb-b1 is inactive) (Carter et al., 2002; Osborne et al., 2004; Simonis et al., 2006). Moreover, the contacts that Hbb-b1 makes with other genomic regions depend on its transcriptional activity; in erythroid nuclei, 80% of contacts are with other active genes, but, in brain cells, this falls to only 13% (Simonis et al., 2006). Interestingly, the LCR region itself is also transcribed and this might even be required for its function (Ho et al., 2006). A range of TFs have been implicated in mediating genomic loops in the case of the Hbb-b1 locus and these include Eklf (erythroid Kruppel-like factor; aka Klf1), Gata1 (GATA-binding protein 1) and the zinc-finger protein Fog1 (aka Zfpm1) (Drissen et al., 2004; Vakoc et al., 2005).

3.2.1.2 Inter-chromosomal interactions

While long-range regulatory interactions involving loci from the same chromosome have been known for some time, it is only recently that inter-chromosomal regulatory interactions (trans-interactions) were discovered (Figure 3-1D). Inter-chromosomal interactions may involve enhancer elements and genes from different chromosomes and can be cell-type specific. The first example of a trans-interaction between chromosomes in mammals (identified by 3C and FISH) is the association between the T helper 2 cytokine locus on mouse chromosome 11 and the promoter of IFN-gamma gene on chromosome 10 in the nuclei of naïve CD4+ T-cells (Spilianakis et al., 2005). This interaction is thought to hold the two loci in a poised state and might facilitate a quicker response upon T-cell activation to differentiate into Th1 and Th2 lineages by expression of either gene-locus. Another interesting example is the regulation of olfactory receptor genes. Dual RNA and DNA FISH revealed that the expression of a specific olfactory gene is accompanied by inter- or intra-chromosomal interactions between the active

gene and a genomic region on chromosome 14 containing an enhancer sequence, referred to as the H element (Lomvardas et al., 2006).

Recently, two different studies showed that inter-chromosomal interactions may involve several factors and can be induced upon exposure to specific stimuli or upon viral infection. The dynamics of gene association with transcription factories was investigated during immediate early gene induction in mouse B lymphocytes and was shown to result in a rapid relocation of the Myc proto-oncogene on chromosome 15 to the same factory that transcribed the Igh gene located on chromosome 12 (Osborne et al., 2007). The study on the investigation of the Interferon (IFN-beta) gene locus upon viral infection reported that the stochastic and monoallelic expression of the IFN-beta gene depends on inter-chromosomal associations with distinct genetic loci that could mediate binding of the transcription factor NF-kappaB to the IFN-beta enhancer, thus triggering transcription from this allele (Apostolou and Thanos, 2008).

Another prominent example comes from the mammalian X-chromosome inactivation process, which involves a specific trans-association of the X-inactivation center (Xic) between the X chromosomes during early development. Female cells carry two X chromosomes, one of which is mostly silenced so that expression levels of X-linked genes are comparable to those in male cells. Recent studies detected a transient co-localization of the X inactivation centers of the homologous chromosomes that precedes the initiation of inactivation of one of the two chromosomes (Augui et al., 2007) and that the chromatin insulator protein (CTCF) is involved in mediating this interaction (Xu et al., 2007). CTCF also colocalizes with cohesin at specific sites in human and mouse chromosomes (Parelho et al., 2008) and raises the possibility that protein-chromatin interaction involving genomic loci from different chromosomes could possibly stabilize, at least transiently, a network of inter-chromosomal interactions within the cell nucleus.

3.2.1.3 Chromosomal territories, movement and nuclear organization

Despite the growing evidence on inter-chromosomal interactions, it is known that chromosomes occupy territories with preferred and non-random positions in the nucleus of mammalian cells, so-called chromosome territories (CTs) (Cremer et al., 2006; Meaburn and Misteli, 2007) (Figure 3-1E). FISH experiments have revealed the relocation of chromatin domains containing activated genes to substantial distances outside their chromosome territory, suggesting that positional organization of chromatin domains within the nucleus could impinge on the regulation of gene expression. This finding, together with the observation of extensive intermingling of DNA from different chromosomes (Branco and Pombo, 2006) raises the issues of how and why genes move relative to their chromosome territories and whether the looping out regulates, or is

regulated by transcriptional activity. Evidences in favour of a role for looping out in the regulation of gene expression has come from studies that show the colocalization of genes in the nucleus for co-expression or co-regulation (reviewed in (Fraser and Bickmore, 2007)). Active genes on decondensed chromatin loops extend outside chromosome territories and can colocalize both in *cis* and in *trans* at sites within the nucleus to share the same transcription factories (Osborne et al., 2004; Osborne et al., 2007) or to sites adjacent to splicing-factor enriched speckles (Chuang et al., 2006) or Cajal bodies (Dundr et al., 2007). Extensive relocalization of large genomic regions in response to gene activation can depend on actin (Chuang et al., 2006; Dundr et al., 2007) and myosin (Chuang et al., 2006), suggesting that intranuclear movements of genomic regions are, at least in some cases, more directed than previously thought (Kumaran et al., 2008).

The spatial organization of chromosome territories in mammalian cells can be described by their radial positioning relative to the center of the nucleus, as was recently done for the three-dimensional (3D) map of all chromosomes in human male fibroblast nuclei (Bolzer et al., 2005). The radial positioning of chromosomes correlates with their gene density in spherical nuclei such as lymphocytes (Kupper et al., 2007) and is evolutionarily conserved (Mora et al., 2006; Neusser et al., 2007) but nevertheless tissue-specific to a certain degree (Parada et al., 2004). Gene-rich regions tend to occupy more interior positions, while gene-poor and late-replication regions tend to be associated with the nuclear periphery (Kupper et al., 2007; Neusser et al., 2007). In addition, similar non-random chromatin arrangements with respect to the local gene density or GC content have been observed for different cell types (*e.g.*, fibroblasts, bone-marrow cells and cell lines) from several eukaryotic lineages such as amphibians, reptiles, birds and mammals (Federico et al., 2006; Neusser et al., 2007). Surprisingly, a detailed analysis of the position of chromosomes in mouse lymphocytes has shown that chromosomes are more likely to form ‘heterologous’ neighbourhoods, where homologous chromosomes are preferentially separated from each other, which might facilitate more extensive trans-interactions between heterologous chromosomes (Khalil et al., 2007).

Though the general patterns appear to be evolutionarily conserved, they are nevertheless dynamic and are altered during cellular differentiation. Changes in the transcriptional program of a cell correlate with specific changes in the organization of individual CTs, at the level of intermingling, CT volume and radial position during lymphocyte activation (Branco et al., 2008), possibly reflecting an adaptation to the new transcriptional program. Similarly, other recent studies have shown that the architecture of chromosome territories

changes during differentiation (*e.g.*, human adipocyte differentiation (Kuroda et al., 2004) and mouse T-cell differentiation (Kim et al., 2004)).

3.2.1.4 Association of the genomic loci with the nuclear periphery

In metazoan nuclei, the nuclear envelope is underlaid by a continuous meshwork of lamins and lamin-associated proteins, which preferentially associate with inactive chromatin regions and facilitate chromatin organization (Akhtar and Gasser, 2007). Pickersgill *et al* (Pickersgill et al., 2006) characterised the regions that interacted with the nuclear lamina in *Drosophila melanogaster* and showed an enrichment for gene-poor regions and repressed genes. More recently, Guelen *et al* (Guelen et al., 2008a) mapped the interaction sites of the entire genome with the nuclear lamina components in human fibroblasts and described over 1,300 lamina-associated-domains (LAD) which were again enriched for genes with low expression levels. Though an association of silenced genes with the nuclear periphery is demonstrated, what was unclear from these studies was whether the requirement for gene repression causes association or if repression is an effect of association with the nuclear lamina. Experimentally induced repositioning of human chromosomal regions to the nuclear periphery in Finlan *et al* (Finlan et al., 2008) suggests a causative role of the nuclear periphery in suppressing the expression of some (but not all) genes as repositioning to the periphery is still compatible with active transcription. Another study investigating the consequences of repositioning the immunoglobulin loci in mouse fibroblasts to the nuclear periphery supports the notion that such molecular interactions may be a mechanism to limit the accessibility to proteins that facilitate recombination or transcription (Reddy et al., 2008).

While the nuclear periphery has been generally associated with repressed genes, several studies have shown a correlation with active genes being associated with components of the nuclear pore complexes (NPCs), which serve as gates for the transport of molecules between the nucleus and cytoplasm. ChIP experiments in yeast for NPC components revealed an enrichment for active genes (Casolari et al., 2004). Several inducible genes such as *INO1*, *HXK1*, *GAL1*, *GAL2*, and *HSP104* become stably positioned at the nuclear periphery when activated and remain there after transcription is shut off (Brickner et al., 2007; Cabal et al., 2006; Casolari et al., 2004; Taddei et al., 2006). In the case of Gal1 and Ino1, the relocalization to pores was found to be dependent upon the SAGA acetyl-transferase complex (Brickner and Walter, 2004; Cabal et al., 2006). In humans and *Drosophila*, the MSL complex can recruit transcriptionally active loci to the nuclear pore (Mendjan et al., 2006), although another study revealed that the association of silent genes is just as likely as for active genes (Brown et al.,

2008). Most of these results have to be reconciled with the observation that many (if not most) transcribed genes, in both yeast (Gartenberg et al., 2004) and mammals (Janicki et al., 2004), do not associate stably with pores. Despite this, one common theme that is emerging is a tendency for the inner nuclear membrane to be associated with less active genes, whereas the NPCs tend to associate with transcriptionally active loci, at least in yeast, possibly in order to facilitate efficient transport of mRNAs (Figure 3-1F).

3.2.2 Transcriptional regulation constrains genome organization

Unlike in prokaryotes, where the genetic material is primarily packaged in a single circular chromosome, the genome of eukaryotes is contained in the nucleus, condensed in a complex, hierarchical manner and is encoded in several different linear chromosomes. These distinctions, together with the fact that the transcriptional apparatus are largely different, enforce very different ways by which genes are transcribed from the chromosomes. In the case of most prokaryotes, the absence of a nucleus and the organization of functionally related genes into operons facilitate coupled transcription and translation of polycistronic transcripts. In contrast, the presence of a nucleus in eukaryotes imposes the constraint that the transcribed monocistronic mRNA needs to be transported to the cytoplasm before translation can occur.

Although transcription in both prokaryotes and eukaryotes involves the evolutionarily conserved core RNA polymerase subunit, the whole process of transcriptional regulation is fundamentally different. In contrast to prokaryotes where transcription primarily relies on the *cis*-regulatory DNA sequences alone (Browning and Busby, 2004), eukaryotic transcription is regulated at many levels (Lee and Young, 2000; van Driel et al., 2003). Therefore unlike in prokaryotes, transcription in eukaryotes is an energy-intensive, multi-step process, involving a large number of molecular events to be coordinated both in space and time. Given the intricacy involved in a single transcriptional regulatory interaction, one can ask whether or not the complexity of the whole network of transcriptional interactions has imposed a significant constraint on the organization of genes across the different eukaryotic chromosomes. This becomes particularly interesting in the light of a recent work, which demonstrated that tuning the expression level of a single gene can provide an enormous fitness advantage to an individual in a population of cells (Dekel and Alon, 2005). Thus one could extrapolate that optimization of transcriptional regulation on a global scale, such as the efficient expression of relevant genes under specific conditions, would have significant advantage on the fitness of an individual in a genetically heterogeneous population.

Though several studies have described that genes with similar expression pattern cluster on the genome and that gene order is conserved, no study has investigated how genes are organized across and within the chromosomes: given that eukaryotes contain several chromosomes, are the set of genes regulated by a given TF (i) randomly distributed across different chromosomes or encoded on specific chromosomes? (ii) distributed in an unbiased manner within a chromosomal arm or display preference to be encoded in regions containing particular chromosomal landmarks? (iii) positionally clustered within a chromosome or not? Here, we investigate these questions by using the recently available genome-scale data on 13,853 high-confidence regulatory interactions (see Materials and Methods). This data covers 156 TFs and 4495 target genes for the model eukaryote *Saccharomyces cerevisiae*, whose genetic material is organized into 16 linear chromosomes.

3.2.2.1 The majority of TFs show a strong preference to regulate genes on specific chromosomes

Several elegant studies have elucidated that the organization of chromosomes within the eukaryotic cell nucleus is non-random and that they occupy distinct volumes called chromosomal territories (Cremer and Cremer, 2001; Gasser, 2002). In yeast, in addition to the ordered movements during cell division, it has been demonstrated that interphase chromosomes undergo large rapid movements (over 0.5 μm in a 10 seconds interval; nuclear diameter of ~2μm) and that such movements could reflect the metabolic state of the cell (refs (Akhtar and Gasser, 2007; Gasser, 2002) and references therein). These observations have suggested that the non-random organization of the chromosomes could (i) allow functional compartmentalization of the nuclear space, thus potentially enhancing or repressing expression of specific genes and (ii) bring co-regulated genes into physical proximity in order to co-ordinate gene expression. The above-mentioned observations (also see previous sections) on the non-random nuclear architecture and chromosomal dynamics together with the fact that transcriptional regulation in eukaryotes is an energy-intensive, highly coordinated and time-intensive process motivated us to ask if such considerations have constrained the positioning of genes in specific chromosomes during the course of evolution.

Given that eukaryotes encode several linear chromosomes, we first investigated if the targets of TFs tend to be preferentially encoded on specific chromosomes, or randomly distributed on different chromosomes. We therefore analyzed the chromosomal location of the targets for each TF in the currently available map of protein-DNA interactions for yeast (see Figure 3-2A and methods). We first created a ‘chromosome preference profile’ for every TF,

which is a vector that contains the number of target genes on each of the 16 chromosomes. By comparing this vector to what is expected by chance (see Methods), we identified the TFs which displayed a significant preference to have their targets on specific chromosomes more often than what is expected by chance.

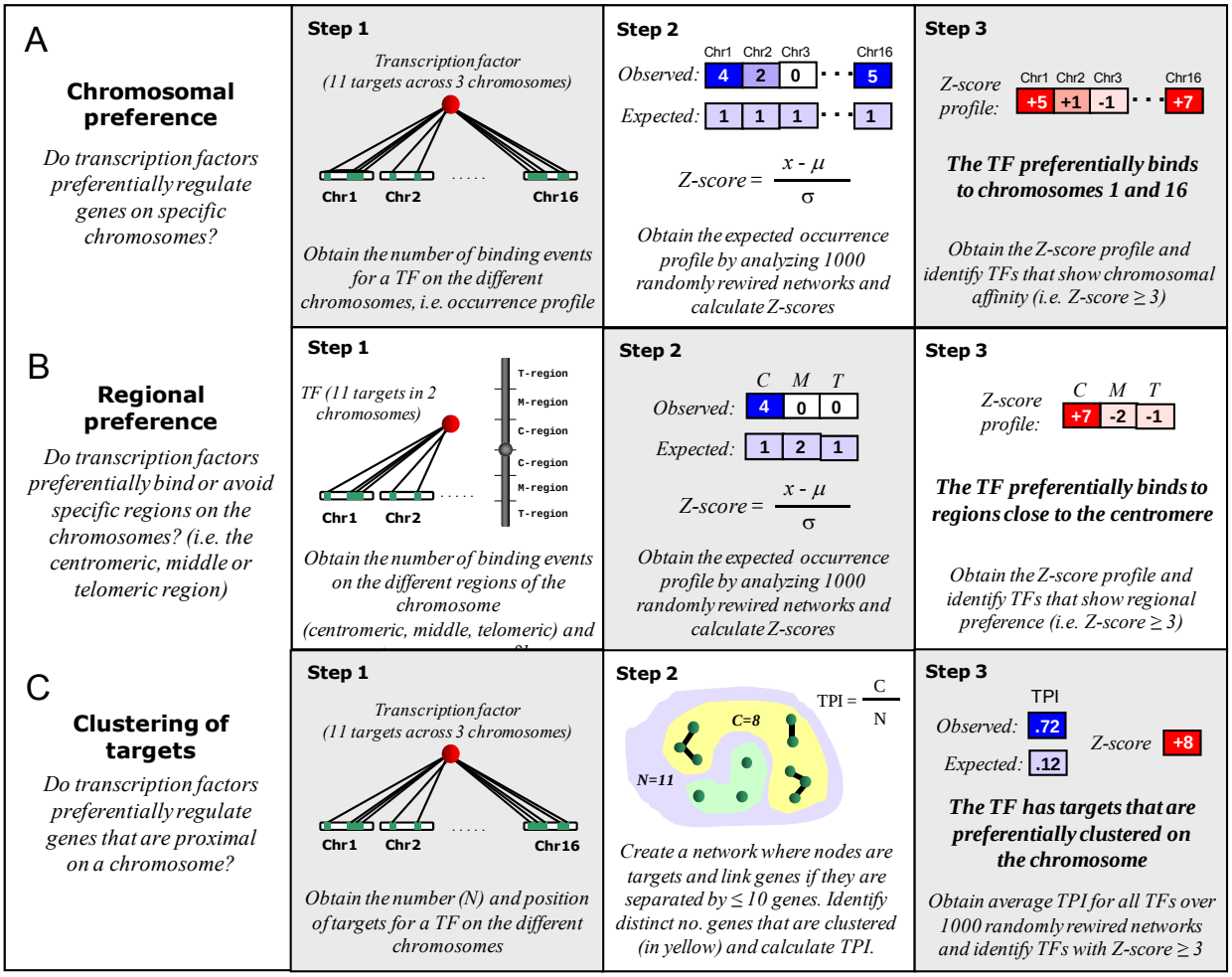


Figure 3-2: Schematic showing the methods employed to estimate the significance for (A) chromosomal preference (B) regional preference and (C) clustering of target genes. Please see methods for details. *x*: observed value; μ mean: σ standard deviation.

Since the null model is critical to obtain statistical significance, we ensured that the random networks are as close as possible to the real network in terms of the topology and the gene distribution on the chromosomes. The random networks were therefore obtained by employing a re-wiring procedure, preserving the connectivity distribution and the inherent chromosomal distribution of the genes. In other words, the number of targets for each TF and the number of TFs regulating a given target gene (TG) in the random networks will be the same

as what is seen in the real network but the interactions between them are randomly re-wired. As this procedure does not randomize the chromosomal position of a gene, any inherent, non-random clustering of genes on the genome is explicitly maintained. Furthermore, this procedure treats every chromosome independently by maintaining the same gene density and the same number of genes as seen in the real yeast chromosomes. This therefore allows us to assess any preference for binding by the TFs. For all observations reported here, statistical significance was assessed based on p-value and Z-score. Only TFs with $p \leq 10^{-3}$ and $|Z| \geq 3$ were considered to show a significant difference in comparison to the null model. To correct for multiple testing, we calculated q-values as a measure of significance using the q-value package in R. We estimate a false discovery rate (FDR) of 0.3% when calling all $p \leq 10^{-3}$ as significant.

Through this analysis, we found that a majority of the TFs (84 TFs, $p < 10^{-3}$ and $Z \geq 3$) showed a striking preference to encode a significant fraction of target genes on at least one particular chromosome. Of these, 78% (66 TFs) showed preference to only one chromosome, 18% (15 TFs) showed preference to two chromosomes and a smaller fraction (4%) of the TFs showed preference to three or more chromosomes. Figure 3-3A shows all the 16 chromosomes of *S. cerevesiae* along with the TFs which have been identified to preferentially bind to the target loci on them. Our investigation identified several TFs to have a strong preference to regulate genes on specific chromosomes. Some of these include (i) the global regulatory hub Sok2p, showing a significant preference for binding to chromosome 15 (observed, x: 67, expected, μ: 32, Z: 6.7, $p < 10^{-3}$), regulating genes important for pseudohyphal differentiation and vesicle trafficking, (ii) Phd1p, showing a preference for binding to chromosome 5 (x: 52, μ: 23, Z: 7.0, $p < 10^{-3}$) and chromosome 9 (x: 32, μ: 14, Z: 5.1, $p < 10^{-3}$), controlling expression of genes required for differentiation and (iii) Msn4p, showing preference for chromosome 13 (x: 32, μ: 13, Z: 5.4, $p < 10^{-3}$), regulating expression of genes involved in stress response. While it is interesting to note that all of the 16 chromosomes have a preferred set of TFs binding them (Figure 3-3B), the number of TFs showing preference to a particular chromosome does not correlate with the physical size of the chromosome (in bp), gene content or the gene density. Taken together, these observations indicate that the targets of most TFs are not randomly distributed across the different chromosomes. Instead, they are highly ordered and show a preference to be encoded on specific chromosomes, independent of the size and the gene density of the chromosome.

Our finding that such a pattern of organization exists for the distribution of targets of TFs motivated us to methodically analyze (i) if the TFs themselves show a preference to be encoded on specific chromosomes, and in particular, if global regulatory proteins show any such

preference and (ii) if there are any patterns of higher-order organization of regulatory interactions between chromosomes. Our investigation on the first question unambiguously revealed that TFs and particularly the global regulatory hubs do not show any preference to be encoded on specific chromosomes. Instead the distribution was similar to what is expected by chance. However, we identified the existence of a higher-order organization of regulatory interactions wherein several TFs which are encoded on specific chromosomes tend to preferentially regulate or avoid regulating genes on distinct chromosomes. Figure 3-3C shows the links between chromosomes which display statistically significant tendency to either interact (red line; $p < 10^{-3}$; $Z \geq 3$) or avoid interaction (blue line; $p < 10^{-3}$; $Z \leq -3$) in the context of transcriptional regulation. These observations suggest that TFs encoded in specific chromosomes can show distinct preferences to regulate targets encoded on particular chromosomes and might reflect a coordinated and possibly a combinatorial, effect between TFs that are encoded in the same chromosome.

(Space left for an enhanced layout of the figure)

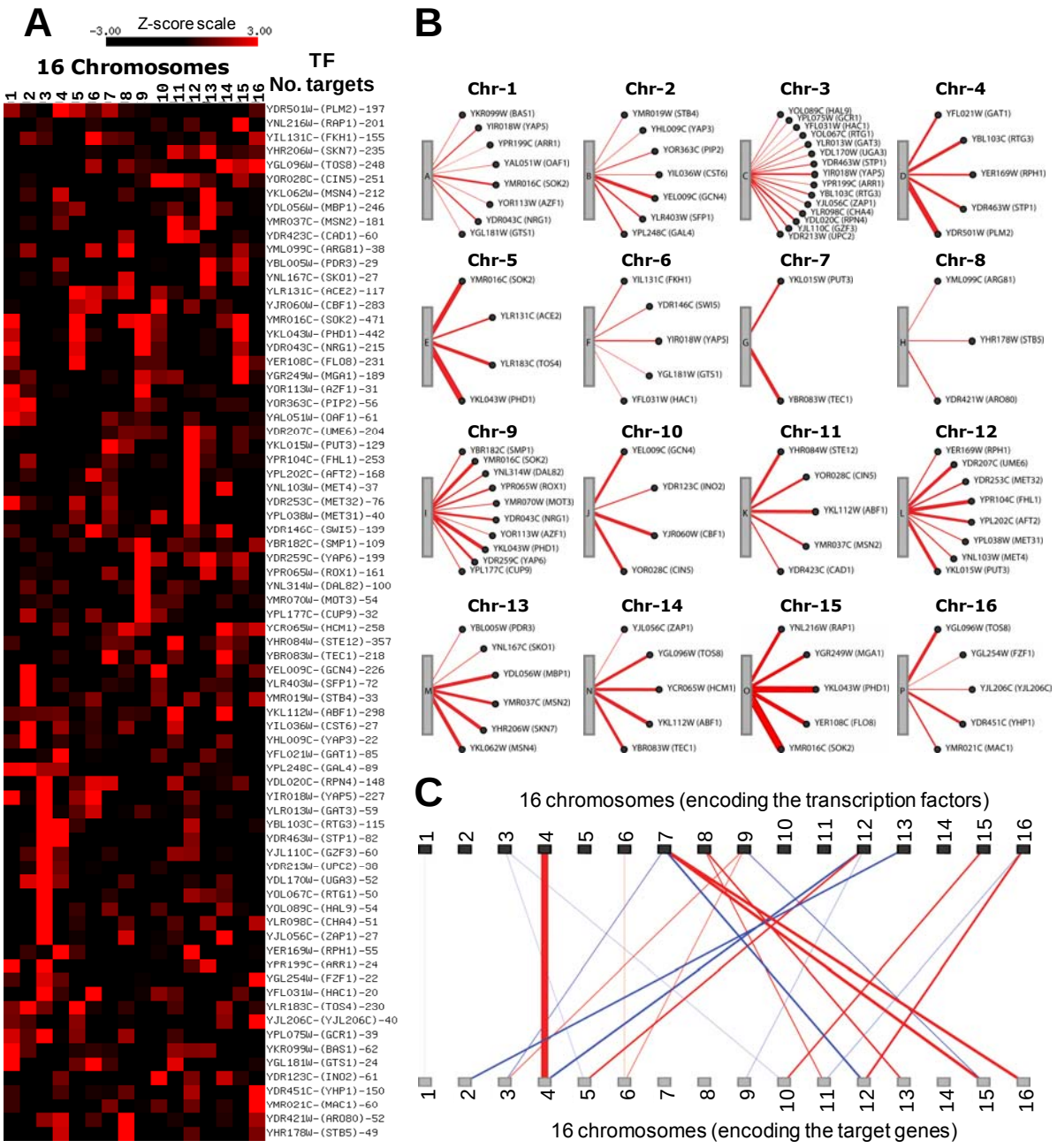


Figure 3-3: Chromosomal preference for binding by TFs. (A) Each column in the matrix represents one of the 16 chromosomes. Each row represents the Z-score significance profile of a particular TF to have its targets on the different chromosomes. The top 75 TFs (selected by p-value and higher Z-scores) are ordered after hierarchically clustering their Z-score profiles. The number of target genes is shown next to the gene name (B) TFs with target preference for each of the 16 chromosomes. Only those TFs which show significant preference and regulate more than 16 genes are shown. Each chromosome has a set of TFs that tend to preferentially bind them. The thickness of the red line is proportional to the absolute number of target genes for that TF on the chromosome. (C) Higher order organization of regulatory interactions. The top and bottom columns denote the chromosomes where the TFs and TGs. Red and blue lines connecting the two chromosomes mean that TFs originating from a specific chromosome tend to preferentially encode or avoid targets on a particular chromosome, respectively. The thickness is proportional to the Z-score.

3.2.2.2 A significant fraction of the TFs tend to have targets on specific regions of the chromosomal arm

Apart from the fact that the nucleus is organized into sub-compartments, creating microenvironments that facilitate distinct nuclear functions, several studies that visualized precise chromosomal loci have revealed that specific regions of the chromosomes display restricted displacement to varying degrees (Akhtar and Gasser, 2007; Gasser, 2002). For instance, in yeast, chromosomal 'landmarks' such as the telomeres and centromeres show marked constraints in their movements within the nuclear space when compared to other chromosomal loci. In addition, live microscopy studies have revealed that centromeres tend to cluster near the spindle pole body (SPB) whereas the telomeres tend to be tethered to the nuclear envelope (Akhtar and Gasser, 2007; Gasser, 2002). Moreover it has been shown that yeast chromosomes form chromosomal loops, where the telomeric ends come closer to each other than to the centromeres. Such anchoring of chromosomal regions is thought to be reversible and is known to involve microtubules that associate with the SPB (for centromeres) and the yKu heterodimeric protein, Esc1p and Sir4p (for telomeres)(Akhtar and Gasser, 2007; Gasser, 2002). This phenomenon of periodic attachment of distinct regions of the chromosomal arms to the nuclear periphery appears to be a conserved mechanism and is believed to regulate patterned gene expression, possibly by separating transcriptionally active and inactive chromosomal domains (Finlan et al., 2008; Guelen et al., 2008b). These observations motivated us to assess if such phenomena, during the course of evolution, could have constrained the target genes of TFs to be encoded within distinct regions of the chromosomal arm.

In particular, we asked if TFs tend to preferentially bind or avoid specific regions on the linear chromosomes, such as regions closer to the centromere, the telomere, or the regions in-between. To investigate this question, we first divided each chromosomal arm into three equal regions (in bp): C, containing the centromere, M, the middle region and T, containing the telomere. For each TF, we then created a 'regional preference profile', which contains the number of targets in each of the three regions. Comparing these results with random expectation by performing the same calculations on 1000 random networks allowed us to assess the statistical significance (see Figure 3-2B and Methods). This enabled the discovery of TFs which display a significant bias to bind to particular regions of the chromosomal arm independent of the specific chromosome. We found that 29 TFs (Figure 3-4A) showed a statistically significant preference ($p < 10^{-3}$; $Z \geq 3$ at a FDR of 0.5%) to bind to a particular region over others, thus providing the first evidence for the prevalence for such an effect. We show that

several TFs display a strong preference to bind specific regions on chromosomal arms. For instance, Hsf1p, the trimeric heat shock regulatory protein and Msn2p, the multicopy suppressor of SNF1 mutation protein tend to preferentially regulate genes that are encoded in regions closer to the centromere, whereas the bZIP domain containing TFs Yap5p and Yap6p which are required under stress conditions tend to bind to regions closer to the telomere. Additional evidence which reinforced our observations that certain TFs do show preference to bind to specific regions on the chromosome came from our inspection of the TFs which avoided binding to a particular region (Figure 3-4B). We found that certain TFs like the osmosis dependent regulator Skn7p and Msn2p clearly avoided binding to the T-region (containing the telomere) while the pleiotropic drug regulator Pdr1p and Smp1p avoided regulating genes in the C-region (containing the centromere). Interestingly, the suppressor of kinase Sok2p, which regulates genes involved in cellular differentiation, avoids binding to both the C and M regions of the chromosomes, displaying a clear preference to bind to the region containing the telomere. Taken together, these observations suggest that events which allowed clustering of certain functionally related genes, based on their usage, accessibility and transcriptional activity, have been selected during evolution. Consistent with this proposal, it is interesting to note that regions that cluster at the nuclear periphery such as the telomeres, as well as the mating-type loci are generally transcriptionally silent, whereas internally located regions encoding metabolic enzymes on the chromosomal arm get recruited to nuclear pores upon transcriptional activation (Cabal et al., 2006; Casolari et al., 2004; Ishii et al., 2002; Taddei et al., 2006).

We then investigated if (i) the loci encoding TFs, and in particular global regulatory proteins, show any regional preference and (ii) there are patterns of higher-order organization of regulatory interactions involving specific chromosomal regions, *i.e.*, if TFs encoded in specific regions tend to preferentially regulate genes on other chromosomal regions. Though our investigation along these lines revealed the absence of any such preferential organizational pattern for the loci encoding TFs, we discovered that genes encoding global regulatory hubs tend to strongly avoid being encoded in regions closer to the telomere ($p = 0.004$). Investigations to uncover the presence of higher-order interactions between specific chromosomal regions revealed that TFs encoded elsewhere in the genome regulate genes within the T-region whereas TFs within the T-region appear to preferentially avoid regulating genes in the same region ($p = 0.007$; Figure 3-4C). These observations are consistent with the fact that genes on telomeric and sub-telomeric regions are largely repressed. Given the dynamic nature of the different chromosomal regions and the differential transcriptional activity associated with specific regions, such organization of loci encoding TFs within specific regions

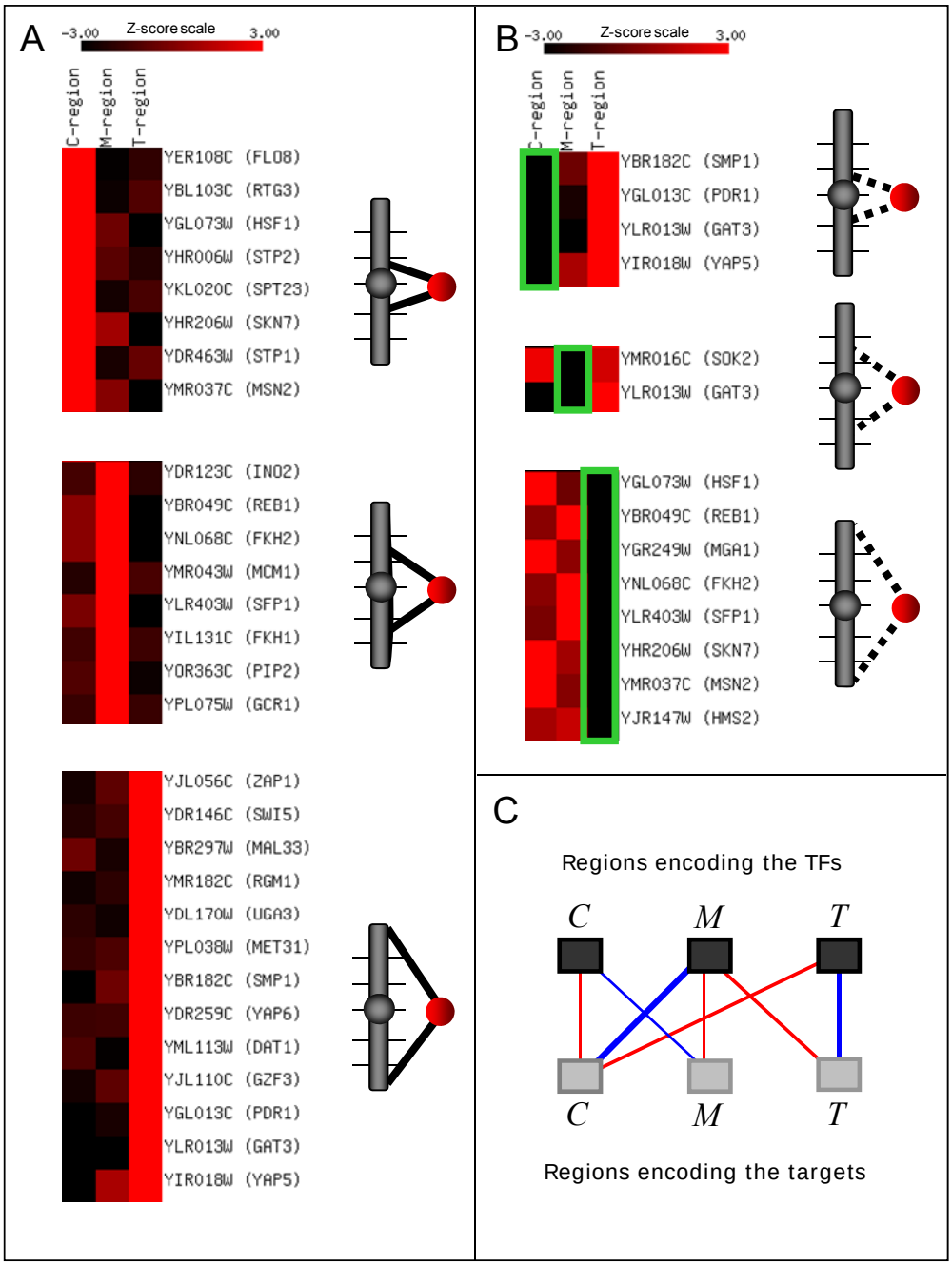


Figure 3-4: TFs showing significant regional preference or avoidance for binding on the chromosomes (see Figure 3-2B). (A) TFs which show a strong tendency to have their targets on the C-region (containing the centromere), M-region (containing the middle region) or T-region (containing the telomere) on the chromosome. (B) TFs which show a strong avoidance to have their targets on the three regions. Green boxes highlight the group of TFs which show significant regional avoidance for one of the three regions. In the cartoon next to the matrices, thick black lines indicate preference and broken black lines indicate avoidance. Only TFs with $p < 10^{-3}$ and $|Z| \geq 3$ are shown in both cases. (C) Higher order organization of regulatory interactions. The top column denotes regions on the chromosomal arm where the TFs are encoded and the bottom column denotes the regions where the targets are encoded. Lines connecting the two regions mean that TFs originating from a specific region tend to preferentially have (red lines) or avoid (blue lines) targets on a particular region of the chromosome. The thickness is proportional to the Z-score.

of the chromosomes, and patterns of higher order regulatory interactions may have been selected during evolution. Taken together, the findings reported here strongly suggest that such regional preferences are not only seen for the targets of specific TFs, but also for global regulatory hubs and the regulatory interactions affecting expression of genes in specific chromosomal regions.

3.2.2.3 Most TFs show a strong preference to positionally cluster their targets within a chromosome

Though we report the prevalence of chromosomal preference and regional bias in the distribution of the targets of a large fraction of the TFs, it does not answer if the regulated genes are proximal to each other on the chromosome or if they are relatively far apart within the same region. While several studies have revealed that genes with similar expression profiles (co-expressed genes) cluster on the chromosome (Cohen et al., 2000; Hurst et al., 2004; Spellman and Rubin, 2002), no study has addressed if the targets of the same TF, cluster on the chromosome on a genomic scale. Although previous studies have unambiguously revealed the existence of chromosomal domains that contain genes with similar expression pattern (co-expressed genes), it should be kept in mind that clustering of co-expressed genes need not always imply regulation by the same TF because co-expressed genes maybe clustered due to several reasons such as mechanisms involving chromatin remodeling, transcriptional read-through, regulation of genes by the same TF or regulation by different TFs in the same transcriptionally active euchromatinic domain (Batada et al., 2007). Therefore, we initiated a systematic investigation and analyzed if the targets of most TFs display positional clustering on a given chromosome or not.

We first defined and calculated the Target Proximity Index (TPI) for each TF (see methods and Figure 3-2C). In short, the TPI for a TF represents the fraction of all the regulated genes that show proximal clustering on the chromosome. In our study we defined proximity, D, as the number of genes that separate two targets of a TF. We then compared the TPI values for the observed and the random networks to obtain the statistical significance. From our analysis, we found that most TFs (>75%) showed high TPI values (TPI > 0.6, $p < 10^{-3}$; at a FDR of 0.1% for $D \leq 20$), suggesting a strong preference for target genes to be clustered within a distance range of ~20 genes. On the contrary, TPI values in random networks for the same distance threshold were found to be significantly lower than 0.2. To ensure that the observations are (i) not biased by tandem gene duplications controlled by the same TF or (ii) not biased by divergent, bi-directional genes which could artificially increase the TPI score, relevant control

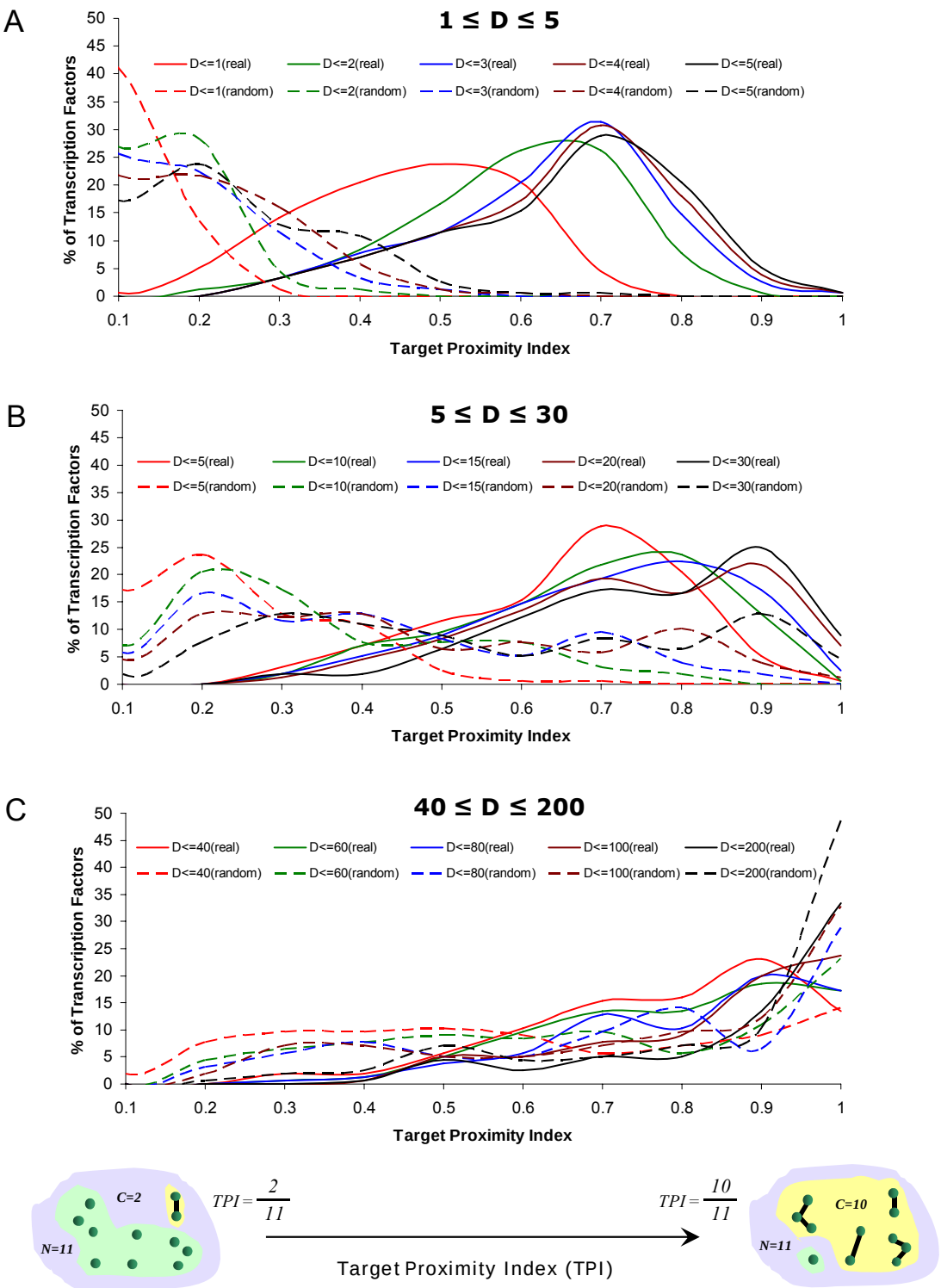


Figure 3-5: Frequency distribution of TPI values. Distribution of Target Proximity Index (TPI) for all TFs in the real and randomly constructed networks at different proximity values *i.e.*, D values (see Methods) are shown in (A) D ≤ 1 to D ≤ 5; (B) D ≤ 5 to D ≤ 30 and (C) D ≤ 40 to D ≤ 200. Note that in the real network, the maximum proportion of TFs have TPI values which are much higher than what is seen for the random networks (at around 0.8 for real network and 0.2 for random networks at D ≤ 20), demonstrating that most TFs show clustering of their targets in this distance range.

calculations were performed. In the filtered network, we removed (i) all tandem duplicates from our dataset and (ii) randomly chose a target gene from a divergent, bi-directional gene pair and calculated the TPI score. Our results did not change after controlling for tandem duplicates and bi-directionally transcribed genes, suggesting that what we observe are truly attributable to positional clustering of targets on a chromosome. An investigation of how many genes are positionally clustered within the window of 20 genes revealed that on an average, such a window only contains 2.6 genes regulated by the same TF. This is striking and suggests that all three mechanisms, *i.e.*, (a) chromatin remodeling, (b) regulation by different TFs in the same euchromatinic domain and (c) regulation by the same TF within a euchromatinic domain, may contribute to the previously observed domains of co-expressed genes.

In order to validate the robustness of our definition of proximity on the TPI values, we systematically varied this parameter (D) from 1 to 200 and compared them against what was obtained in random networks (Figure 3-5). We found that significant separation between real data and random networks occurred for the definition of proximity (D) as being less than 20 genes, suggesting that this could reflect the average size of a possible open euchromatinic domain that is available for transcription in yeast. Our results therefore suggest that evolution might have favored certain recombination events which allowed genes that need to be regulated by the same TF to be encoded close to each other. Another distinct possibility given that transcriptional regulatory networks are likely to be plastic (Borneman et al., 2007) would be that selection could have first driven clustering of genes that need to be co-regulated and then new transcriptional regulatory interactions could have evolved afterwards. Regardless of the driving force, the evolutionary advantages are clear: such a clustering of targets would not demand high concentrations of TFs in the nucleus which are generally expressed in low quantities and prevent inappropriate regulation of unrelated target genes. Such an organization has the added advantage of minimizing noise in expression levels, which has been recently proposed to be an additional driving force for gene order conservation (Batada and Hurst, 2007).

3.3 DISCUSSION & CONCLUSION

In conclusion, our study demonstrates that the complexity of transcriptional regulation constrains genome organization at several levels. Our findings beyond those discussed in detail here, such as TFs encoded in specific chromosomes and within distinct regions show a strong preference to regulate genes on distinct chromosomes and regions open up several questions and expand our need to understand eukaryotic gene regulation at a higher level. The findings reported here are consistent with several molecular mechanisms, such as the genome-wide

loop model of chromosomes (Francastel et al., 2000), the presence of expression hubs (Kosak and Groudine, 2004) and transcription factories (Cook, 1999; Osborne et al., 2004) and the nuclear gating hypothesis (Blobel, 1985).

With the development of experimental methods such as 3D chromosome capture, 4C and 5C and the availability of genome-scale data on protein-DNA interactions from high-throughput experiments in other eukaryotes (shown in Table 3-1), our work provides a fundamental framework by which such questions can be systematically studied for higher eukaryotes. In fact, a preliminary analysis in mammalian systems using stem cell differentiation factors Sox2, Oct4 and Nanog have indeed revealed a striking preference for these TFs to encode their targets on specific chromosomes (SCJ, MMB, Unpublished). We therefore believe that our work, which demonstrates that gene organization is constrained by the process of transcriptional regulation in yeast, is likely to be a paradigm that is also applicable to other eukaryotes.

The findings reported here has several direct applications. For instance, the map that we describe for yeast in this study can serve as a guide and be exploited in genetic engineering experiments for identifying the most appropriate region (on the 16 chromosomes) to incorporate a gene of interest – particularly if it has to be regulated under the control of a specific TF. Describing such maps for higher eukaryotes will have implications in gene therapy and in rationally identifying suitable sites to incorporate reporter genes while producing transgenic organisms. We anticipate that revealing the presence of such patterns of organization of genes within the linear chromosomes of eukaryotes, such as humans, would have significant implications in our understanding of transcriptional regulation, chromosomal territories, their role in cellular differentiation and of specific chromosomal disorders, such as recombination events and copy number variations that are prevalent in diverse diseases such as cancer.

3.4 MATERIALS AND METHODS

3.4.1 Dataset of Transcription factors in *S. cerevisiae* and their regulatory interactions

The transcriptional regulatory network for *S. cerevisiae* was assembled from the results of literature curation, and ChIP-chip experiments (Harbison et al., 2004; Horak et al., 2002; Lee et al., 2002; Svetlov and Cooper, 1995). This network consists of 4527 genes, which include 156 DNA-binding TFs, 4495 target genes and 13,853 regulatory interactions. 31 TFs qualified as hubs, which were defined as the top 20% of the TFs with high out-going connectivity.

Chromosomal positions of all the protein coding genes on the yeast genome were obtained from http://www.yeastgenome.org. Tandem duplicates and bi-directionally transcribed genes were identified by employing pair-wise blast using an e-value cut-off of 10^{-2} and using chromosomal position of the genes in the network.

3.4.2 Estimation of statistical significance

To estimate statistical significance of the properties described here, the reported values for the real network of protein-DNA interactions were compared against 1000 randomly generated networks obtained by employing the re-wiring procedure. The re-wiring procedure randomly reconnects TFs with target genes but ensures that any inherent gene distribution on the chromosome and the overall connectivity distribution of the network is maintained. As this procedure does not randomize the chromosomal position of a gene, any inherent, non-random clustering of genes on the genome is explicitly maintained. Furthermore, it is important to note that this procedure maintains the same gene density and the same number of genes on a chromosome as what is seen in the real yeast chromosomes. This therefore allows us to assess any preference for binding by the TFs reported in our study. To assess if TFs and hubs were preferentially encoded in different chromosomes, we carried out 1000 trials, where we randomly picked the same number of genes as the number of TFs and hubs seen in the real network and analyzed the chromosomal distribution of them. For all observations reported in our study, statistical significance was assessed based on (i) p-value, defined as the fraction of the 1000 random networks which showed a value ≥ what was observed in the real network and (ii) Z-score, calculated as the number of standard deviations the observed value is away from the mean of the 1000 random networks. This is obtained as the ratio of the difference between the observed, x, and random expected, σ, values to the standard deviation, σ *i.e.*, $Z = (x-\sigma)/\sigma$. TFs with $p \leq 10^{-3}$ and $|Z\text{-scores}| \geq 3$ (unless stated otherwise) were considered to show a significant difference in comparison to the null model described above. All significance values were corrected for multiple testing using the q-value package in R (Arava et al., 2003). In particular, the Benjamini & Hochberg step-wise p-value method implemented in the package was used. The same package was used to assess the False Discovery Rate (FDR) at a p-value threshold of 10^{-3} and to estimate the corresponding q-values.

3.4.3 Calculation of chromosomal preference

To test whether a TF has a preference to bind a specific chromosome more often than expected by chance, we first constructed a ‘chromosomal binding profile’. This is a 16 dimensional (one

for each chromosome) vector describing the number of binding events in each chromosome. We then obtained an expected 'chromosomal binding profile' by using 1000 randomly re-wired networks, and taking it through a similar procedure. The preference for a TF to bind a particular chromosome was measured using p-value and Z-score profiles. The p-value for each TF for each chromosome was estimated as the fraction of the 1000 random networks that showed an equal or higher number of binding events than in the real network. The p-value profile was obtained in a similar manner across the different chromosomes for all TFs. The Z-score profile was calculated based on average binding frequency and standard deviation from the 1000 random networks (see Figure 3-2A). Only those TFs which showed a preference to bind to at least one chromosome with $p \leq 10^{-3}$ and $Z \geq 3$ were considered significant. A p-value cut-off of 10^{-3} results in an estimated FDR of 0.3%.

3.4.4 Calculation of regional preference

To assess if TFs preferentially bind to specific regions of the chromosomes more often than expected by chance, we first obtained a 'regional binding profile'. Every chromosomal arm was divided into three regions of equal size (in bp, see Figure 3-2B) to obtain the C-region (containing the centromere), M-region (in the middle) and the T-region (containing the telomere). Thus the 'regional binding profile' is a 3 dimensional (one for each region) vector that captures the number of binding events of a TF on all the chromosomes. We then obtained the expected 'regional binding profile' by using the 1000 randomly re-wired networks and taking it through the same set of calculations. The p-value and Z-score profiles were obtained as described above. Only those TFs with $p \leq 10^{-3}$ were considered to show regional preference (or avoidance) for binding. We estimate a FDR of 0.5% at a p-value threshold of 10^{-3} for TFs showing regional preference.

3.4.5 Calculation of target proximity

To assess the positional clustering of targets of a given TF across chromosomes, we calculated the Target Proximity Index (TPI) for each TF. This is defined as the ratio of the number of the targets that are within a particular distance (proximity is measured as D, the number of genes that physically separate two genes regulated by the same TF on the chromosome) to the total number of targets regulate by that TF (see Figure 3-2C). The TPI values lie between 0 and 1 where TFs with high TPI values would indicate high clustering of their targets. In order to test the significance of clustering of targets for each TF, we obtained the expected TPI values by computing the same for 1000 randomly re-wired networks. P-values and Z-score were

computed as described above. TFs were considered to display a preference to cluster their binding sites if $p \leq 10^{-3}$ and $Z \geq 3$. At these thresholds we estimated a FDR of 0.1%. Since most TFs were found to show significant clustering of targets, the TPI score distribution of all the TFs was used to demonstrate the differences between the observed and expected behavior.

REFERENCES

Akhtar, A. and Gasser, S. M. (2007). The nuclear envelope and transcriptional control. *Nat Rev Genet* **8**, 507-17.

Allfrey, V. G., Faulkner, R. and Mirsky, A. E. (1964). Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci U S A* **51**, 786-94.

Apostolou, E. and Thanos, D. (2008). Virus Infection Induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. *Cell* **134**, 85-96.

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O. and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **100**, 3889-94.

Augui, S., Filion, G. J., Huart, S., Nora, E., Guggiari, M., Maresca, M., Stewart, A. F. and Heard, E. (2007). Sensing X chromosome pairs before X inactivation via a novel X-pairing region of the Xic. *Science* **318**, 1632-6.

Batada, N. N. and Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**, 945-9.

Batada, N. N., Urrutia, A. O. and Hurst, L. D. (2007). Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* **23**, 480-4.

Blobel, G. (1985). Gene gating: a hypothesis. *Proc Natl Acad Sci U S A* **82**, 8527-9.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M. R. et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol* **3**, e157.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M. and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815-9.

Branco, M. R., Branco, T., Ramirez, F. and Pombo, A. (2008). Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. *Chromosome Res* **16**, 413-26.

Branco, M. R. and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**, e138.

Brickner, D. G., Cajigas, I., Fondufe-Mittendorf, Y., Ahmed, S., Lee, P. C., Widom, J. and Brickner, J. H. (2007). H2A.Z-mediated localization of genes at the nuclear periphery confers epigenetic memory of previous transcriptional state. *PLoS Biol* **5**, e81.

Brickner, J. H. and Walter, P. (2004). Gene recruitment of the activated INO1 locus to the nuclear membrane. *PLoS Biol* **2**, e342.

Brown, C. R., Kennedy, C. J., Delmar, V. A., Forbes, D. J. and Silver, P. A. (2008). Global histone acetylation induces functional genomic reorganization at mammalian nuclear pore complexes. *Genes Dev* **22**, 627-39.

Browning, D. F. and Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57-65.

Cabal, G. G., Genovesio, A., Rodriguez-Navarro, S., Zimmer, C., Gadal, O., Lesne, A., Buc, H., Feuerbach-Fournier, F., Olivo-Marin, J. C., Hurt, E. C. et al. (2006). SAGA interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope. *Nature* **441**, 770-3.

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. F. and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nat Genet* **32**, 623-6.

Casolari, J. M., Brown, C. R., Komili, S., West, J., Hieronymus, H. and Silver, P. A. (2004). Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell* **117**, 427-39.

Chuang, C. H., Carpenter, A. E., Fuchsova, B., Johnson, T., de Lanerolle, P. and Belmont, A. S. (2006). Long-range directional movement of an interphase chromosome site. *Curr Biol* **16**, 825-31.

Cohen, B. A., Mitra, R. D., Hughes, J. D. and Church, G. M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**, 183-6.

Cook, P. R. (1999). The organization of replication and transcription. *Science* **284**, 1790-5.

Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301.

Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I. and Fakan, S. (2006). Chromosome territories--a functional nuclear landscape. *Curr Opin Cell Biol* **18**, 307-16.

Cremer, T., Kreth, G., Koester, H., Fink, R. H., Heintzmann, R., Cremer, M., Solovei, I., Zink, D. and Cremer, C. (2000). Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture. *Crit Rev Eukaryot Gene Expr* **10**, 179-212.

de Laat, W. and Grosveld, F. (2007). Inter-chromosomal gene regulation in the mammalian cell nucleus. *Curr Opin Genet Dev* **17**, 456-64.

Dean, A. (2006). On a chromosome far, far away: LCRs and gene expression. *Trends Genet* **22**, 38-45.

Dekel, E. and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588-92.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002). Capturing chromosome conformation. *Science* **295**, 1306-11.

Dorman, E. R., Bushey, A. M. and Corces, V. G. (2007). The role of insulator elements in large-scale chromatin structure in interphase. *Semin Cell Dev Biol* **18**, 682-90.

Drissen, R., Palstra, R. J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. and de Laat, W. (2004). The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* **18**, 2485-90.

Dundr, M., Ospina, J. K., Sung, M. H., John, S., Upender, M., Ried, T., Hager, G. L. and Matera, A. G. (2007). Actin-dependent intranuclear repositioning of an active gene locus in vivo. *J Cell Biol* **179**, 1095-103.

Federico, C., Scavo, C., Cantarella, C. D., Motta, S., Saccone, S. and Bernardi, G. (2006). Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. *Chromosoma* **115**, 123-8.

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J. R. and Bickmore, W. A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet* **4**, e1000039.

Francastel, C., Schubeler, D., Martin, D. I. and Groudine, M. (2000). Nuclear compartmentalization and gene activity. *Nat Rev Mol Cell Biol* **1**, 137-43.

Fraser, P. and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413-7.

Gartenberg, M. R., Neumann, F. R., Laroche, T., Blaszczyk, M. and Gasser, S. M. (2004). Sir-mediated repression can occur independently of chromosomal and subnuclear contexts. *Cell* **119**, 955-67.

Gasser, S. M. (2002). Visualizing chromatin dynamics in interphase nuclei. *Science* **296**, 1412-6.

Grimaud, C., Bantignies, F., Pal-Bhadra, M., Ghana, P., Bhadra, U. and Cavalli, G. (2006). RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124**, 957-71.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W. et al. (2008a). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W. et al. (2008b). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J. et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104.

Heun, P., Laroche, T., Shimada, K., Furrer, P. and Gasser, S. M. (2001). Chromosome dynamics in the yeast interphase nucleus. *Science* **294**, 2181-6.

Ho, Y., Elefant, F., Liebhaber, S. A. and Cooke, N. E. (2006). Locus control region transcription plays an active role in long-range gene activation. *Mol Cell* **23**, 365-75.

Horak, C. E., Luscombe, N. M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M. and Snyder, M. (2002). Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev* **16**, 3017-33.

Hurst, L. D., Pal, C. and Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**, 299-310.

Ishii, K., Arib, G., Lin, C., Van Houwe, G. and Laemmli, U. K. (2002). Chromatin boundaries in budding yeast: the nuclear pore connection. *Cell* **109**, 551-62.

Janicki, S. M., Tsukamoto, T., Salghetti, S. E., Tansey, W. P., Sachidanandam, R., Prasanth, K. V., Ried, T., Shav-Tal, Y., Bertrand, E., Singer, R. H. et al. (2004). From silencing to gene expression: real-time analysis in single cells. *Cell* **116**, 683-98.

Khalil, A., Grant, J. L., Caddle, L. B., Atzema, E., Mills, K. D. and Arneodo, A. (2007). Chromosome territories have a highly nonspherical morphology and nonrandom positioning. *Chromosome Res* **15**, 899-916.

Kim, S. H., McQueen, P. G., Lichtman, M. K., Shevach, E. M., Parada, L. A. and Misteli, T. (2004). Spatial genome organization during T-cell differentiation. *Cytogenet Genome Res* **105**, 292-301.

Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-71.

Kosak, S. T. and Groudine, M. (2004). Gene order and dynamic domains. *Science* **306**, 644-7.

Kouzarides, T. (2002). Histone methylation in transcriptional control. *Curr Opin Genet Dev* **12**, 198-209.

Kumaran, R. I., Thakar, R. and Spector, D. L. (2008). Chromatin dynamics and gene positioning. *Cell* **132**, 929-34.

Kupper, K., Kolbl, A., Biener, D., Dittrich, S., von Hase, J., Thormeyer, T., Fiegler, H., Carter, N. P., Speicher, M. R., Cremer, T. et al. (2007). Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma* **116**, 285-306.

Kuroda, M., Tanabe, H., Yoshida, K., Oikawa, K., Saito, A., Kiyuna, T., Mizusawa, H. and Mukai, K. (2004). Alteration of chromosome positioning during adipocyte differentiation. *J Cell Sci* **117**, 5897-903.

Kurz, A., Lampel, S., Nickolenko, J. E., Bradl, J., Benner, A., Zirbel, R. M., Cremer, T. and Lichter, P. (1996). Active and inactive genes localize preferentially in the periphery of chromosome territories. *J Cell Biol* **135**, 1195-205.

Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**, 104-15.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**, 77-137.

Loizou, J. I., Murr, R., Finkbeiner, M. G., Sawan, C., Wang, Z. Q. and Herceg, Z. (2006). Epigenetic information in chromatin: the code of entry for DNA repair. *Cell Cycle* **5**, 696-701.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J. and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403-13.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-60.

Manuelidis, L. (1985). Individual interphase chromosome domains revealed by in situ hybridization. *Hum Genet* **71**, 288-93.

Marenduzzo, D., Faro-Trindade, I. and Cook, P. R. (2007). What are the molecular ties that maintain genomic loops? *Trends Genet* **23**, 126-33.

Meaburn, K. J. and Misteli, T. (2007). Cell biology: chromosome territories. *Nature* **445**, 379-781.

Mendjan, S., Taipale, M., Kind, J., Holz, H., Gebhardt, P., Schelder, M., Vermeulen, M., Buscaino, A., Duncan, K., Mueller, J. et al. (2006). Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Mol Cell* **21**, 811-23.

Millar, C. B. and Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast. *Nat Rev Mol Cell Biol* **7**, 657-66.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800.

Mora, L., Sanchez, I., Garcia, M. and Ponsa, M. (2006). Chromosome territory positioning of conserved homologous chromosomes in different primate species. *Chromosoma* **115**, 367-75.

Muller, W. G., Rieder, D., Karpova, T. S., John, S., Trajanoski, Z. and McNally, J. G. (2007). Organization of chromatin and histone modifications at a transcription site. *J Cell Biol* **177**, 957-67.

Narlikar, G. J., Fan, H. Y. and Kingston, R. E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475-87.

Neusser, M., Schubel, V., Koch, A., Cremer, T. and Muller, S. (2007). Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates. *Chromosoma* **116**, 307-20.

Nightingale, K. P., O'Neill, L. P. and Turner, B. M. (2006). Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev* **16**, 125-36.

O'Sullivan, J. M., Tan-Wong, S. M., Morillon, A., Lee, B., Coles, J., Mellor, J. and Proudfoot, N. J. (2004). Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* **36**, 1014-8.

Olins, A. L. and Olins, D. E. (1974). Spheroid chromatin units (v bodies). *Science* **183**, 330-2.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W. et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-71.

Osborne, C. S., Chakalova, L., Mitchell, J. A., Horton, A., Wood, A. L., Bolland, D. J., Corcoran, A. E. and Fraser, P. (2007). Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol* **5**, e192.

Parada, L. A., McQueen, P. G. and Misteli, T. (2004). Tissue-specific spatial organization of genomes. *Genome Biol* **5**, R44.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T. et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-33.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M. and van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet* **38**, 1005-14.

Pombo, A. and Branco, M. R. (2007). Functional organisation of the genome during interphase. *Curr Opin Genet Dev* **17**, 451-5.

Pombo, A., Jackson, D. A., Hollinshead, M., Wang, Z., Roeder, R. G. and Cook, P. R. (1999). Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *Embo J* **18**, 2241-53.

Pombo, A., Jones, E., Iborra, F. J., Kimura, H., Sugaya, K., Cook, P. R. and Jackson, D. A. (2000). Specialized transcription factories within mammalian nuclei. *Crit Rev Eukaryot Gene Expr* **10**, 21-9.

Razin, S. V., Iarovaia, O. V., Sjakste, N., Sjakste, T., Bagdoniene, L., Rynditch, A. V., Eivazova, E. R., Lipinski, M. and Vassetzky, Y. S. (2007). Chromatin domains and regulation of transcription. *J Mol Biol* **369**, 597-607.

Reddy, K. L., Zullo, J. M., Bertolino, E. and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243-7.

Schneider, R. and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev* **21**, 3027-43.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-54.

Simonis, M., Kooren, J. and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* **4**, 895-901.

Spellman, P. T. and Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* **1**, 5.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637-45.

Svetlov, V. V. and Cooper, T. G. (1995). Review: compilation and characteristics of dedicated transcription factors in Saccharomyces cerevisiae. *Yeast* **11**, 1439-84.

Taddei, A., Hediger, F., Neumann, F. R. and Gasser, S. M. (2004). The function of nuclear architecture: a genetic approach. *Annu Rev Genet* **38**, 305-45.

Taddei, A., Van Houwe, G., Hediger, F., Kalck, V., Cubizolles, F., Schober, H. and Gasser, S. M. (2006). Nuclear pore association confers optimal expression levels for an inducible yeast gene. *Nature* **441**, 774-8.

Turner, B. M. (1993). Decoding the nucleosome. *Cell* **75**, 5-8.

Turner, B. M. (2007). Defining an epigenetic code. *Nat Cell Biol* **9**, 2-6.

Vakoc, C. R., Letting, D. L., Gheldof, N., Sawado, T., Bender, M. A., Groudine, M., Weiss, M. J., Dekker, J. and Blobel, G. A. (2005). Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17**, 453-62.

van Driel, R., Fransz, P. F. and Verschure, P. J. (2003). The eukaryotic genome: a system regulated at different hierarchical levels. *J Cell Sci* **116**, 4067-75.

Xu, M. and Cook, P. R. (2008). Similar active genes cluster in specialized transcription factories. *J Cell Biol* **181**, 615-23.

Xu, N., Donohoe, M. E., Silva, S. S. and Lee, J. T. (2007). Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. *Nat Genet* **39**, 1390-6.

Zinner, R., Albiez, H., Walter, J., Peters, A. H., Cremer, T. and Cremer, M. (2006). Histone lysine methylation patterns in human cell types are arranged in distinct three-dimensional nuclear zones. *Histochem Cell Biol* **125**, 3-19.

Zorn, C., Cremer, C., Cremer, T. and Zimmer, J. (1979). Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. *Exp Cell Res* **124**, 111-9.

4 Uncovering the functional architecture of uncharacterized proteins in *E. coli*

CONTENTS OF CHAPTER 4

OUTLINE

Determining the functions of proteins encoded by genome sequences represents a major challenge in modern biology. Whole-genome sequencing projects are a major source of proteins of unknown function. Annotation of a genome involves assignment of functions to gene products, in most cases on the basis of amino-acid sequence alone. Structure-based identification of homologues often succeed where sequence-alone-based methods fail, due to the conservation of folding patterns long after sequence similarity becomes undetectable. Nevertheless, prediction of protein function from sequence and structure is still a difficult problem, because homologous proteins often have different functions and these traditional approaches have already started to reach an optimum. As a result, alternative computational methods for inferring the protein function such as those which exploit the context of a protein in protein association networks have come to be sought after. These methods, often referred to as network-based functional inference techniques, provide a first hand guess of the functional role and provide complementary insights to traditional methods in understanding the function of uncharacterized proteins. Most recent network-based approaches aim to integrate diverse kinds of functional interactions as it not only boosts coverage but also confidence level of an association, thereby improving the assessment of protein function. In a recent study we attempted to characterize one-third of the 4,225 protein-coding genes of *Escherichia coli* K-12 which remain functionally unannotated (functional orphans). In particular, to elucidate their biological roles, we performed an extensive proteomic survey using affinity-tagged *E. coli* strains and generated comprehensive genomic context inferences to derive a high-confidence compendium for virtually the entire proteome consisting of 5,993 putative physical interactions and 74,776 putative functional associations, most of which are novel. Clustering of the respective probabilistic networks revealed putative orphan membership in discrete multiprotein complexes and functional modules, while a machine-learning strategy based on network integration implicated the orphans in specific biological processes. In the second half of this chapter, I highlight this resource which provides a ‘systems-wide’ functional blueprint of a model microbe, with insights into the biological and evolutionary significance of previously uncharacterized proteins. Given the volume of high-throughput data that is being reported for understanding diverse model systems the time is ripe to employ these network-based approaches which can be used on a whole-organism level to unravel the functions of an increasing number of proteins accumulating in the genomic databases.

CONTRIBUTION TO THE WORK IN THIS CHAPTER

Please note that the work presented in this chapter is the result of the following two publications. Uncovering the functional roles of previously uncharacterized *E. coli* proteins is an ongoing collaborative project with the groups of Dr. Andrew Emili at University of Toronto and Dr. Gabriel Moreno-Hagelsieb at Wilfred Laurier University, Canada. My contribution to this high-throughput study included, but was not limited to, developing novel computational frameworks for understanding genome-context functional associations, analyzing raw protein-protein interaction data generated by Dr. Emili's group, integrating data using computational approaches and inferring function from such networks. Please note that some of the work on functional associations have resulted in other publications in the past, these studies are cited in the appendix but not discussed here.

1) Network-based function prediction in post-genomic era : Metabolic enzymes as a case study
Sarath Chandra Janga and Gabriel Moreno-Hagelsieb
Metabolic Engineering (Submitted)

2) Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins
Pingzhao Hu[†], Sarath Chandra Janga[†], Mohan Babu[†], J. Javier Díaz-Mejía[†], Gareth Butland[†], Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G, Emili A
PLoS Biology 2009, 7(4): e96

4.1 INTRODUCTION

Determining the functions of proteins encoded by genome sequences represents a major challenge in modern biology. As of March 23, 2010, the TrEMBL database contained 10,618,387 sequences (http://www.ebi.ac.uk/uniprot/TrEMBLstats/). The GOLD database (http://www.genomesonline.org) reports more than 1000 published genomes with over 3700 genome projects underway; the database also reports more than 100 metagenome projects with the venter's marine microbial communities project alone contributing more than 6,000,000 proteins to the already accumulating list of protein repertoire. Although the pace at which sequencing technologies are able to generate the genome sequence data is increasing, our ability to unravel the functional roles of the encoded proteins in these genomes has been rather limited.

Historically proteins identified from genome sequencing projects were annotated mostly using the aid of BLAST (Altschul et al., 1997) or other sequence comparison tools followed by manual intervention (Gotoh, 1999; Pearson, 1995; Procter et al.). A principal reason behind researchers BLASTing protein sequences against databases is to learn about some aspect of their function. The researcher aims to answer this question by finding a significant sequence similarity to another protein that is already in the database and whose function was experimentally characterized. This is essentially the most widely used form of computational function prediction and is commonly referred to as annotation transfer by sequence similarity or simply homology-based transfer. The rationale behind homology-based annotation transfer is that, if two sequences have a high degree of similarity, then they have evolved from a common ancestor and they have similar, if not identical functions. This might appear an obvious statement however with increasing number of sequences as well as duplications observed in different lineages, the power of homology-based annotation transfer is being challenged. Adding to this is the problem of errors in annotation even in human curated databases, which spread mis-annotations when homology-based approaches are used. All these factors have made it evident that the traditional approaches for annotating genes with their functional descriptions is nearly impossible with the exponential increase in the number of proteins. In addition, most of the newly identified proteins do not show a high sequence similarity with an already characterized protein leading to the failure or rather saturation of the homology-based approaches and making it impossible to keep up with the influx of data for manually curated annotation. All of these factors have been responsible for an increase in a varied number of automated function inference approaches in the recent years (see Table 4-1) (Godzik et al.,

2007; Han et al., 2006; Rentzsch and Orengo, 2009; Zhao et al., 2008a). These automated function inference methods are based on a number of features, starting from nucleotide or amino acid sequence, sequence patterns/profiles and protein structure patterns to chromosomal location, phylogenetic information, expression profiles, molecular interaction data, functional associations and gene co-evolution and are summarized in Table 4-1.

4.2 RESULTS

4.2.1 Overview of network-based function prediction

The very definition of biological function is ambiguous with its exact meaning depending on the context in which it is used and the classification it is based on (Rison et al., 2000; Whisstock and Lesk, 2003). It is obvious in the post-genomic era that biological function has many aspects associated with it. For instance, a protein kinase; in the biochemical context can simply be defined as an enzyme or more precisely a kinase's function would be the phosphorylation of the hydroxyl group of a specific substrate. While the former gives a very coarse annotation of the protein under study the later gives finer details about its function. A totally different way to understand the role of a protein with in the cell is to ask where exactly it occurs in the cell. This aspect is equally important information especially for entities which occur with in a cell as they can potentially occur in a number of sub-cellular localizations. In this particular case, kinases can be identified either in the cytoplasm or nucleus and this information is crucial in gathering its role and interactions with other proteins with in the cellular environment. Likewise, a mutation in the kinase can be associated with a disease phenotype. Therefore, it is increasingly becoming clear that when speaking of a protein's function, we must always specify the aspect or aspects of the functional description. In particular, when setting out to develop a function prediction tool we must keep in mind which functional aspect or aspects we are trying to predict and use the appropriate vocabulary.

Once functional aspects of a protein are defined, the question is how function can be interpreted in computational terms. For instance, protein sequences for a long time have been represented as character strings that enable their use for many computational tasks including pairwise comparisons and multiple sequence alignments, motif searching, database searching and several other tasks aimed at extracting biological information from the sequence. In fact, our ability to express protein sequence information as a character string amenable for computational processing followed by the availability of algorithms which can exploit this information for meaningful interpretation has changed our view and understanding of cellular

Table 4-1. Resources currently available for protein function prediction grouped according to the predominant method or approach implemented in them. Note that the list may be incomplete as some resources which are not directly relevant to the methods discussed here might have escaped their mention in this table.

Approach	Resource	Webpage
Sequence similarity based	GOtcha	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php
	PFP	http://dragon.bio.purdue.edu/pfp/
	GOsling	https://www.sapac.edu.au/gosling/
	OntoBlast	http://functionalgenomics.de/ontogate/
	GOblet	http://goblet.molgen.mpg.de
	Blast2GO	http://www.blast2go.de
Phylogenomics based	SIFTER	http://sifter.berkeley.edu
	AFAWE	http://bioinfo.mpiz-koeln.mpg.de/afawe/
	RIO	http://www.rio.wustl.edu/
	OrthoStrapper	http://www.cgb.ki.se/OrthoGUI
Domain/pattern/profile based	InterProScan	http://www.ebi.ac.uk/tools/interproscan/
	Pfam	http://pfam.sanger.ac.uk
	SUPERFAMILY	http://supfam.cs.bris.ac.uk/superfamily/
	PROSITE	http://www.expasy.ch/prosite/
	PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
	SMART	http://smart.embl-heidelberg.de/
	Gene3D	http://gene3d.biochem.ucl.ac.uk/gene3d/
	PANTHER	http://www.pantherdb.org/
	TIGRFAMs	http://www.tigr.org/TIGRFAMs/
	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
	CATH	http://www.cathdb.info/
	CatFam	http://www.bhsai.org/downloads/catfam.tar.gz
Sequence clustering based	ProtoNet	http://www.protonet.cs.huji.ac.il/
	CluSTr	http://www.ebi.ac.uk/clustr/
	eggNOG	http://eggnog.embl.de
	COGs	http://www.ncbi.nlm.nih.gov/COG/
	InParanoid	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
	MultiParanoid	http://multiparanoid.sbe.su.se/index.html
	OrthoMCL	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi
Machine Learning based	ProtoFun	http://www.cbs.dtu.dk/services/ProtFun/
	GOPET	http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar
	SVM-Prot	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
	ffPred	http://bioinf.cs.ucl.ac.uk/ffpred/
	EzyPred	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/
Network based	MCODE	http://baderlab.org/Software/MCODE
	MCL	http://www.micans.org/mcl/
	SAMBA	http://acgt.cs.tau.ac.il/samba/
	PRODISTIN	http://crfb.univ-mrs.fr/webdistin/
	Cytoscape	http://www.cytoscape.org/
	STRING	http://string.embl.de/
	VisANT	http://visant.bu.edu/
	VIRGO	http://whipple.cs.vt.edu/virgo/welcome.cgi

entities. However, in contrast to sequence information, the annotation of a protein until recently has been written in human language, conveying the complex descriptions and intricacies of its function as well as experimental evidence in support of it, in terms of custom non-standard format varying between different groups. As a result, vocabulary went on to be invented and re-invented, with many terms being synonymous. This synonymy not only raises confusion among human curators (re)annotating the annotations but also increases the chances of additional errors due to a non-standard format for annotating the function. Therefore, over the years a need to convey this information in a more controlled and well-defined fashion has emerged especially due to the requirements to make the annotations processed automatically. One of the first group of people who appreciated this were the biochemists to come up with the Enzyme Commission (EC) classification (Tipton, 1994). EC classifies metabolic reactions in a four-level hierarchy which are noted by a four-position identifier, going from the most general in the first position to the most specific function of the enzyme in the last position. This classification not only addresses the need for a controlled vocabulary but also a well-defined relationship between terms thereby allowing the comparison between annotations. While enzymes form one of the most commonly occurring protein classes in the cell, they are definitely not the only kind, so these definitions are not sufficient for annotating functions of all the proteins in a cell. Therefore, following this classification, Monica riley and colleagues in 1993 came up with the Riley or Multifun classification system for *E. coli* (Riley, 1993; Serres and Riley, 2000). Other annotation systems came into existence following this which include Clusters of Orthologous Groups (COG) (Tatusov et al., 1997) – based on manual annotation of a group of orthologous proteins by hierarchically organizing the functional descriptions, swissprot annotations based on human curation efforts on well-annotated proteins (Apweiler, 2001; Kretschmann et al., 2001), and more recently Gene Ontology (GO) (Ashburner et al., 2000). The common theme among these schemes is the establishment of a controlled vocabulary and in many cases a categorization that proceeds from the general to the specific. The Gene Ontology (GO) currently serves as the dominant cross-specie approach for machine-legible functional annotation and covers three major aspects of gene products' function, namely molecular function, biological process and cellular component. Each ontology is implemented as a directed acyclic graph (DAG) where terms are represented as nodes in the graph and are arranged from the general to the specific. The DAG arrangement means that each node may have more than a single parent which enables the description of functions that are associated with more than one biological activity or process. By standardizing an annotation and defining the relationships between terms using a graph, annotations can be computationally processed. For instance, given a GO-

annotated genome a researcher can computationally identify the set of all genes with a given annotation and likewise one can predict functional labels of proteins using such a controlled vocabulary. Naturally, such standardized annotations also limit the flexibility in the amount of detail an annotation can be made.

Having defined function and the means of describing function, one can start discussing function prediction. In particular, function prediction using network-based approaches which is the topic of this chapter essentially requires two seed components: a) a network of functional associations which are amenable for graph theory analysis b) a network-based function prediction algorithm for predicting functional labels for uncharacterized genes in the graph. In what follows, I will first discuss different approaches for constructing and integrating functional association networks and then outline currently available computational methods for inferring function based on them.

4.2.1.1 Methods and databases for constructing functional association networks

Traditionally function of a protein was defined using a number of low-throughput approaches like mutagenesis of residues or whole proteins which allowed the identification of the phenotypes for follow up analysis. However, it is increasingly becoming clear that this rational is limited in its ability to infer the function of proteins; failing for those which exhibit mild phenotype or those which are not expressed under standard experimental conditions. In addition, since most proteins associate dynamically with a number of other cellular entities during their life time, the traditional notion of identifying function of a protein by isolating it from the rest of the cellular machinery can be misleading for a majority. This notion followed by the availability of experimentally determined protein-protein interaction maps for diverse model organisms have given rise to the use of these datasets for delineating the biological processes, pathways and complexes that proteins take part in (Aranda et al., ; Bader et al., 2003; Breitkreutz et al., 2008). Indeed, there is now observable overlap and informative variation between different types of low- and high-throughput experiments (Shoemaker and Panchenko, 2007a) which provides a convincing reason for exploiting them as complementary approaches in unraveling the functions of proteins. Indeed, recent years have seen an explosion in the number of methods and databases which provide functional associations (both direct physical and indirect contextual interactions) between proteins using both experimental and computational means (Table 4-2).

(Space left for an enhanced layout of the table)

Table 4-2. Different approaches for generating functional linkage maps or networks. Typically, these networks either independently or integrated versions of them form the input for network-based functional inference algorithms.

Approach	Description	Data sources
Protein-protein interactions	Physical interactions between proteins identified either by mass spectrometry or one of the hybrid approaches are used to generate protein interaction maps on a large-scale which are used as input for function prediction algorithms.(Shoemaker and Panchenko, 2007a)	**HPRD** (http://www.hprd.org) **IntAct** (http://www.ebi.ac.uk/intact/site/index.jsf) **MINT** (http://cbm.bio.uniroma2.it/mint/index.html) **BioGRID** (http://www.thebiogrid.org) **DIP** (http://dip.doe-mbi.ucla.edu/dip/Main.cgi) **MPPI** (http:// mips.gsf.de/proj/ppi)
Co-expression networks	In these approaches gene co-expression above a significant correlation threshold is considered as a presence of a functional linkage between genes. Genome-wide inspection of these gene co-expression networks provides an intuitive way to represent complex co-expression patterns between many genes providing functional insights into uncharacterized processes. (Aoki et al., 2007; Huber et al., 2007)	**GEO** (http://www.ncbi.nlm.nih.gov/geo) **SMD** (http://genome-www5.stanford.edu) **ArrayExpress** (http://www.ebi.ac.uk/arrayexpress) **caArray** (http://caarraydb.nci.nih.gov/caarray) **M3D** (http://m3d.bu.edu/)
Genetic interaction networks	(Lasko, 2000) In these approaches interactions between genes are constructed by linking gene pairs which show significantly reduced fitness when both the genes are knocked out compared to when each gene is knocked out independently. These lethality assays are carried out on a high-throughput scale to construct genome-scale interactions. (Butland et al., 2008; Costanzo et al.)	**BioGRID** (http://www.thebiogrid.org) **DRYGIN** (http://drygin.ccbr.utoronto.ca) **IM Browser** (http://proteome.wayne.edu/PIMdb.html)
Genome context networks	These approaches include the gene fusion, gene cluster or gene order conservation, phylogenetic profile and operon rearrangement methods (Dandekar et al., 1998; Enright et al., 1999; Janga et al., 2005; Pellegrini et al., 1999). See text for further discussion.	**STRING** (http://string.embl.de) **ProLinks** (http://prolinks.mbi.ucla.edu/) **VisANT** (http://visant.bu.edu)
Integration of data sources	These approaches integrate different kinds of functional association data using machine learning techniques and then construct high-confidence functional linkage networks which are then used for function prediction (Hu et al., 2009; Linghu et al., 2008; Marcotte et al., 1999b; Zhao et al., 2008b).	**STRING** (http://string.embl.de) **ProLinks** (http://prolinks.mbi.ucla.edu/) **VisANT** (http://visant.bu.edu) **Virgo** (http://whipple.cs.vt.edu:8080/virgo)

To summarize, experimental approaches employed for constructing functional association networks mostly comprise of data from protein-protein interaction screens followed by co-expression networks comprising of gene pairs showing significant correlation in their expression profiles across conditions, derived from microarray datasets (Luo et al., 2007; Ruan et al., ; Wang et al., 2009). More recently, genetic interactions- measuring the fitness defects of the double mutants compared to that of the individual mutants, are also being employed for constructing these functional linkage networks (Butland et al., 2008; Costanzo et al.). These high-throughput experimental approaches not only increase the confidence of an association but also give cellular context of the protein providing complementary view to the traditional functional prediction paradigm.

In addition to the experimental methods, several computational methods have been proposed for constructing protein-protein associations from sequence data alone. These include the so-called genome context methods namely gene fusion, gene cluster or gene order conservation, operon arrangements and protein phylogenetic profiles. The gene fusion approach tries to detect the fusion of two genes into a single protein coding gene in one of the sequenced genomes and thereby links them as a strong functional association (Enright et al., 1999; Marcotte et al., 1999a). The method of gene order conservation aims to identify pairs of genes which consistently show a tendency to cluster in immediate vicinity in a number of genomes- suggesting a strong functional link in prokaryotic genomes which are abundant in operons (Dandekar et al., 1998; Overbeek et al., 1999). The method of operon rearrangement tries to identify a link between any pair of genes on a genome as long as their orthologs are predicted to be organized in an operon with a high confidence in at least one sequenced genome (Janga et al., 2005; Rogozin et al., 2002; Snel et al., 2002). The power of this approach depends on the predictive quality of operon prediction methods which have been shown to reach ~90% accuracy in most sequenced genomes (Brouwer et al., 2008; Moreno-Hagelsieb and Collado-Vides, 2002). Yet another approach not based on genomic proximity is phylogenetic profiles. In this method a vector of presence/absence profile of a gene across all the analyzed genomes is constructed and compared to identify genes which show the most correlated profiles, as a measure of functional link. The rational here is that two proteins showing similar profiles i.e, coordinated in their evolutionary gain and loss, are expected to be functionally related (Gaasterland and Ragan, 1998; Pellegrini et al., 1999). Modified versions of this approach take into account the phyogenetic signal of the genomes employed and/or the redundancy in the genome sequence information (Barker and Pagel, 2005; Date and Marcotte, 2003; Moreno-Hagelsieb and Janga, 2008).

Recently, the integration of different types of interaction data into genome-wide functional linkage maps has gained much popularity for functional inference as these integrated maps not only boost coverage but also confidence of an association when assessing protein function. One of the first studies which demonstrated the power of integrating different types of interaction data was by Marcotte and colleagues where they have put together diverse kinds of computational genome context inferences (Marcotte et al., 1999b). This was followed by a number of other methods such as those implemented in the STRING and PROLINKS databases, among other focused studies (Bowers et al., 2004; Hu et al., 2009; Jensen et al., 2009; Massjouni et al., 2006). Typically, in these networks edge weights correspond to the integrated interaction probability values obtained by first scoring each of the methods independently against a set of gold standard interactions, which are then used in a bayesian fashion assuming the scores obtained in each method are independent of each other. More complex methods take into account the dependence and correlation between methods to develop a regression model for scoring the integrated interactome (Linghu et al., 2008; Zhao et al., 2008b). Nevertheless, all of them boil down to constructing a network with either weighted or unweighted edges which are then used for propagating annotations to uncharacterized members using approaches discussed in the section below.

4.2.1.2 Computational methods for predicting function from network context

Any set of functional associations, whether experimentally derived or predicted by the above methods can be depicted as a network of nodes connected by edges, with nodes representing proteins and edges denoting the interactions between these nodes. As such most network-based functional inference algorithms work under the premise that the closer the two nodes are in the network higher is the functional similarity between them (Sharan et al., 2007). Indeed, most computational approaches for predicting function from network simply exploit the context of a protein with in the local or global network-neighborhood analogous to traditional sequence or genomic context methods. These approaches also generally tend to infer the broader function such as biological process a protein is in involved in, as opposed to the molecular function which is typically inferred by homology-based approaches – making network-based approaches complementary methods for annotating genomes. These methods can be grouped into two major classes namely those which use direct network-context and those which are assisted by module prediction. The former infer the function of a protein based on its connections (direct or indirect) in the network while the later first identify the modules of related

proteins and then annotate each protein in the module based on the known functions of its members using one of the direct methods (see Table 4-3 for a summary of the methods belonging to either class).

Table 4-3. Different methods currently available for network-based function prediction.

Method	Description	References
Direct	In simpler versions of these methods function of a protein is assigned based on the number of annotated protein neighbors in the immediate network neighborhood which are associated with a particular function. Advanced approaches take into account overall network topology and are able to give confidence scores for predictions. Techniques such as flow simulation and graph theoretic based have shown to yield high accuracies on some model systems. Other methods in this category involve the use of probabilistic markov random models.	(Chua et al., 2006; Deng et al., 2003; Hishigaki et al., 2001; Karaoz et al., 2004; Letovsky and Kasif, 2003; Nabieva et al., 2005; Schwikowski et al., 2000; Vazquez et al., 2003)
Module based	In these approaches, two major steps are involved: 1) Identification of modules which are functionally coherent using any clustering technique 2) predicting function of uncharacterized members in a cluster using any of the direct methods or by computing enrichment for characterized functions in a given module and then transferring the annotations to other members. The first step follows the notion that genes which work in the same biological process should be homogenous in their functional roles and hence plays a crucial role in these methods. So majority of the methods in this category differ in the approach taken to identify modules.	(Altaf-Ul-Amin et al., 2006; Bader and Hogue, 2003; Brun et al., 2003; King et al., 2004; Pereira-Leal et al., 2004; Rives and Galitski, 2003; Samanta and Liang, 2003; Spirin and Mirny, 2003)

Among the direct methods, the simplest and perhaps the most intuitive method for function prediction determines the function of a protein based on the known function of proteins lying in the immediate neighborhood and is commonly referred to as the majority consensus or Guilt-By-Association (GBA) method (Schwikowski et al., 2000). Although simple and can be effective for dense networks, the method does not take into account the complete topology of the network and neither does provide a score for predicted functional label. Therefore, over the years more sophisticated methods like those developed by Hishigaki et. al, (Hishigaki et al., 2001) and Chua et. al, (Chua et al., 2006) tried to address these limitations. Other direct

methods involve the use of graph theoretical principles such as cuts and flow-simulation in the networks in order to take advantage of the global and/or local topology of the network under consideration (Karaoz et al., 2004; Nabieva et al., 2005; Vazquez et al., 2003). In doing so, these methods also aim at maximizing the number of edges (for a protein of interest) which connect to other proteins assigned with the same function. Some authors also employed probabilistic approaches to address the caveats of the original methods and follow the premise that the function of a protein is independent of all other proteins given the functions of its immediate neighbors- thereby leading to the use of markov random field models for solving the problem of function prediction (Deng et al., 2003; Letovsky and Kasif, 2003)(also see (Sharan et al., 2007)).

Biological systems are inherently modular in their functions with groups of genes being associated with a particular biological process/pathway (Hartwell et al., 1999). This has resulted in the development of module-based functional inference approaches. In these approaches, first coherent groups' of genes which are predicted to work together to achieve a common biological task are identified by clustering methods and then the functions of genes with in the group are assigned. Once modules are identified, simple methods like GBA or hypergeometric enrichment computed for every function associated with the module are used for transferring the annotations to the uncharacterized members. Therefore, in these approaches the initial clustering method employed is crucial in determining the quality of the functional predictions. As a result, different module-assisted techniques differ in the module detection technique employed. Module finding algorithms typically depend on the network topology information which is used as a distance metric, resulting in the use of clustering techniques for identifying either a defined number of clusters, as in *k-means* clustering or some times hierarchical clustering of the data. Some of the module detection techniques also have the ability to detect overlapping clusters as a means of revealing the inherent plasticity in biological systems. Table 4-3 summarizes some of the module-assisted techniques employed for functional inference.

4.2.2 Uncovering the cellular roles of functional orphans in *E. coli*

Because of its central position in the microbial research community, the Gram-negative bacterium *Escherichia coli* plays a leading role in investigations of the fundamental molecular biology of bacteria (Arifuzzaman et al., 2006; Baba et al., 2006; Barrett et al., 2005; Butland et al., 2005; Faith et al., 2007; Feist et al., 2007; Joyce et al., 2006; Riley et al., 2006). This experimentally-tractable microbe is a workhorse in basic and applied research aimed at elucidating the mechanistic basis of prokaryotic processes and traits, including those of

pathogens. The ever-expanding availability of genomic resources makes *E. coli* particularly well-suited to systematic investigations of microbial protein components and functional relationships on a global scale. These include a genome-wide collection of single gene deletion strains (Baba et al., 2006) along with extensive knowledge of regulatory circuits (Barrett et al., 2005; Faith et al., 2007; Gama-Castro et al., 2008; Joyce et al., 2006) and metabolic pathways (Feist et al., 2007; Kanehisa and Goto, 2000; Keseler et al., 2005).

Yet despite being the most highly studied model bacterium, a recent comprehensive community annotation effort for the fully sequenced reference K-12 laboratory strains (Riley et al., 2006) indicated that only half (~54%) of the protein-coding gene products of *E. coli* currently have experimental evidence indicative of a biological role. The remaining genes have either only generic, homology-derived functional attributes (e.g. 'predicted DNA-binding') or no discernable physiological significance. Some of these functional 'orphans' (not to be confused with 'ORFans', which are genes present within only single or closely-related species) may have eluded characterization in part because they exhibit mild mutant phenotypes, are expressed at low or undetectable levels, or have limited homology to annotated genes.

A key feature of the molecular organization of all organisms, including bacteria, is the tendency of gene products to associate into macromolecular complexes, biochemical pathways and functional modules that in turn mediate all the major cellular processes. Elaboration of these interaction networks via proteomic, genomic and bioinformatic approaches can reveal previously overlooked components and unanticipated functional associations (Hawkins and Kihara, 2007). For example, a recent integrative analysis of phenotypic, phylogenetic and physical interaction data led to the discovery of an evolutionarily conserved set of novel bacterial motility-related proteins (Rajagopala et al., 2007). However, while systematic integration of diverse high-throughput interaction datasets is routinely performed to reveal new functional relationships in model eukaryotes such as yeast, worm and fly (Bandyopadhyay et al., 2008; Gunsalus et al., 2005; Lee et al., 2008; Myers et al., 2005; Reguly et al., 2006; Sharan and Ideker, 2006), few analogous studies of the global functional architecture of *E. coli,* and any prokaryote for that matter, have been reported to date (Campillos et al., 2006; Slonim et al., 2006; Yellaboina et al., 2007).

To this end, we have combined complementary, highly-sensitive computational and experimental procedures to derive extensive high-quality maps of the functional interactions inferred by genomic context (GC) methods and physical interactions (PI) deduced by proteomics of *E. coli*. Our results indicate that many previously unannotated bacterial proteins are components of functionally cohesive modules and multiprotein complexes linked to well

known biological processes. A substantive fraction of these associations could be verified by independent experimentation and were found to be broadly conserved across prokaryotic phyla, indicating homologous systems in other microbes, while others are seemingly restricted to the *E. coli* lineage. However, in what follows I present a summary of this large-scale study where in we characterize the broad biological processes of these functional orphans using an integration of computational and experimental means. The entire data collection is publicly accessible via a searchable web-browser interface (http://ecoli.med.utoronto.ca/) to stimulate exploration of both conserved and specialized bacterial proteins within the context of biological processes of particular interest.

4.2.2.1 The extent of existing functional annotation for *E. coli* proteins

Since the functional characterization of *E. coli*, and bacteria in general, has largely been guided historically by scientific interests and technical considerations, some bias is expected in terms of the coverage and depth of existing biological knowledge as reflected in current gene annotations. This biased coverage is likely due to multiple reasons, ranging from the low expression of certain proteins to the lack of homologs in other organisms including humans. To evaluate the degree to which the physiological functions of the 4,225 putative protein-coding sequences of *E. coli* K-12 are characterized presently, we examined the scope of literature reference records curated in the UniProt annotation system (Apweiler et al., 2004). After excluding PubMed references corresponding to genomic mapping studies, the average total number of papers associated with each of the proteins of *E. coli* K-12 is surprisingly limited (Figure 4-1A), with many proteins apparently still uncited.

We next examined recent *E. coli* K-12 (sub-strains W3110 and MG1655) gene annotations in the public databases RefSeq (Pruitt et al., 2005), MultiFun (Serres et al., 2004), and EcoCyc (Keseler et al., 2005). Since W3110 is commonly used for high-throughput studies, we devoted the bulk of our subsequent analysis to this sub-strain. In total, we found that 2,794 (66%) of *E. coli*'s proteins had either proper mnemonic names (Rudd, 1998), experimentally-derived annotations in the MultiFun multifunction schema, or literature documentation to a well-defined pathway or multiprotein complex in EcoCyc (Figure 4-1B). This left 1,431 proteins (34%) as currently functionally uncharacterized (which constitute our 'orphans' set). Of these, 446 (31%) have at least one putative molecular function defined on the basis of sequence (such as the presence of a predicted DNA-binding domain or an enzymatic motif) in the Clusters of Orthologous Groups of proteins (COGs) catalog (Tatusov et al., 1997).

4.2.2.2 Properties of the functional orphans of *E. coli*

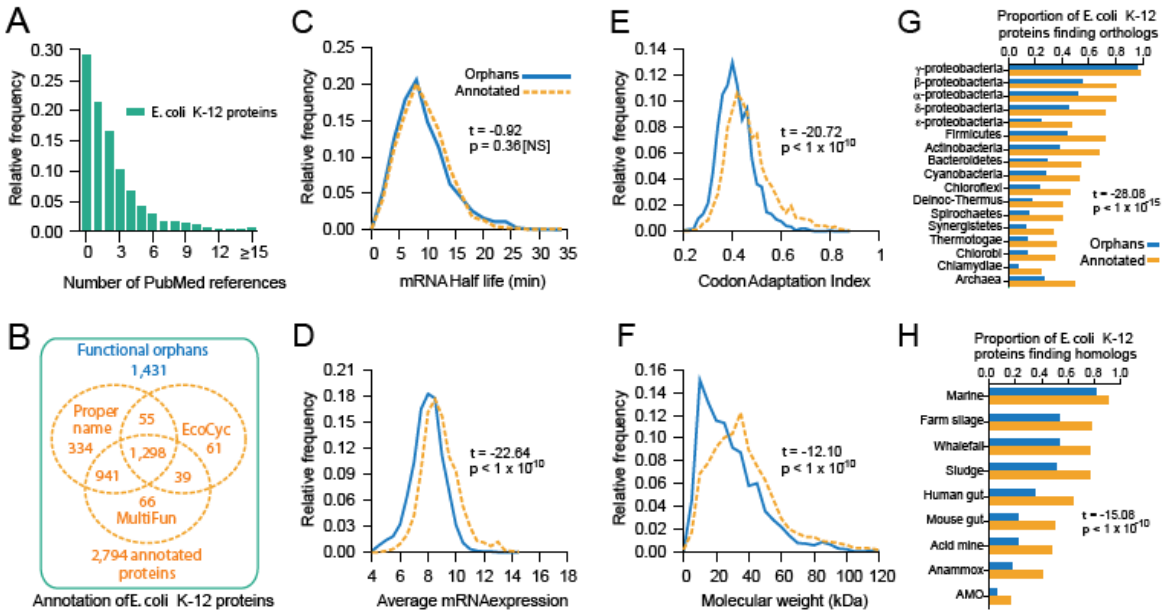


Figure 4-1. Annotated and functional orphan genes of the *E. coli* K-12 reference strain

(**A**) Frequency distribution of supporting publications per *E. coli* protein-coding gene. (**B**) Summary of existing annotations for *E. coli*, showing proteins of unknown function (*orphans*) lacking proper names or functional annotations in MultiFun or EcoCyc. (**C**) Although the functional orphans are encoded by transcripts with half-lives comparable to those of annotated genes, they tend to be expressed at lower levels based on (**D**) microarray analysis of mRNA and (**E**) Codon Adaptation Index scores, and (**F**) have lower molecular weights on average. Orthologs of orphans are also less prevalent in sequenced genomes than those of annotated genes (**G**). However, examination of environmental metagenomic libraries (**H**) indicates that the orphans are not necessarily exclusive to the *Escherichia* lineage. *AMO*: methane oxidizing Archaea; *Anammox*: anaerobic ammonium oxidation bacteria. *t*, T-test; *p*, P-value; NS, not statistically significant.

The genes lacking annotation appear to be translated into bona fide proteins as their corresponding transcripts (Selinger et al., 2003) were not significantly ($p = 0.36$) less stable than the products of annotated genes (Figure 4-1C). However, some differences were evident in terms of their biophysical attributes and evolutionary scope relative to annotated genes. Most notably, only 21 orphans (1.5%) are required for viability under standard laboratory conditions (Baba et al., 2006) in contrast with the 280 annotated genes (10%) previously deemed essential. The orphans were also significantly ($p < 1e\text{-}10$) less abundant at both the transcript [Figure 4-1D; avg. normalized mRNA expression over 400 microarray experiments (Faith et al., 2007): 8.0 (orphans) *vs*. 8.9 (annotated)] and protein levels (Figure 4-1E; avg. codon adaptation index: 0.41 *vs.* 0.47). Furthermore, they tend to encode somewhat smaller proteins (Figure 4-

1F; avg. M_W: 29.4 *vs.* 38.2 kDa; p < 1e-10) with fewer domain assignments (44%) than for annotated proteins (74%) according to the SUPERFAMILY database (Madera et al., 2004).

Orphans also generally find fewer orthologs in a non-redundant genome dataset, defined by filtering at 90% similarity based on the frequency of shared orthologs among genomes (Figure 4-1G), with an average of 0.22 as compared with 0.48 for annotated genes (p < 1e-10) using a maximum-score E-value cutoff of $1x10^{-6}$ for BLAST bi-directional best hits (BDBHs). Nevertheless, broader sequence comparisons against currently available metagenomes (Figure 4-1H) indicated that orphan homologs (one way BLAST hits) are often widely distributed in diverse environments (See online protocols accompanying this published study for more detailed description of the Materials and Methods); for example, a high proportion (0.80) of orphans have homologs present in marine metagenomes, anaerobic bacterial populations (farm silage, 0.51; whalefall, 0.50; sludges, 0.49), and even in the residents of the mammalian gut (union of human and mouse, 0.35), implying participation in core bacterial processes. Furthermore, the same high proportion (~99%) of orphan and annotated genes have orthologs in the other sequenced *E. coli* isolates, including pathogenic variants and closely-related *Shigella* strains. Taken together, this argues that the functional significance of the orphans is more pervasive than the current annotations suggest.

4.2.2.3 A systematic approach to elucidate biological function

The scarce existing knowledge regarding the biological roles of the orphans is likely due to multiple reasons, ranging from the lower expression, non-essentiality, or smaller sizes of certain orphan proteins to their lack of obvious homologs in other organisms including humans. Accordingly, integration of multiple data sources is warranted to decipher the specific biological roles of this uncharacterized repertory. Since the elucidation of physical and functional interaction networks can provide insights into bacterial protein function based on the concept of guilt-by-association (Yao and Ruzzo, 2006), we took a multi-pronged approach. We performed large-scale proteomic analysis to determine orphan participation as components of stable multimeric protein complexes, and inferred functional relationships based on genomic context inference, which exploits the patterns of gene conservation across bacterial genomes (Shoemaker and Panchenko, 2007b; von Mering et al., 2005). We then predicted the functions of the orphans using an integrative machine-learning procedure. Finally, independent low-throughput experiments were also performed to validate a subset of high confidence predictions related to core biological processes which will not be discussed in here. Key steps (mostly computational) in this pipeline are outlined schematically in Figure 4-2.

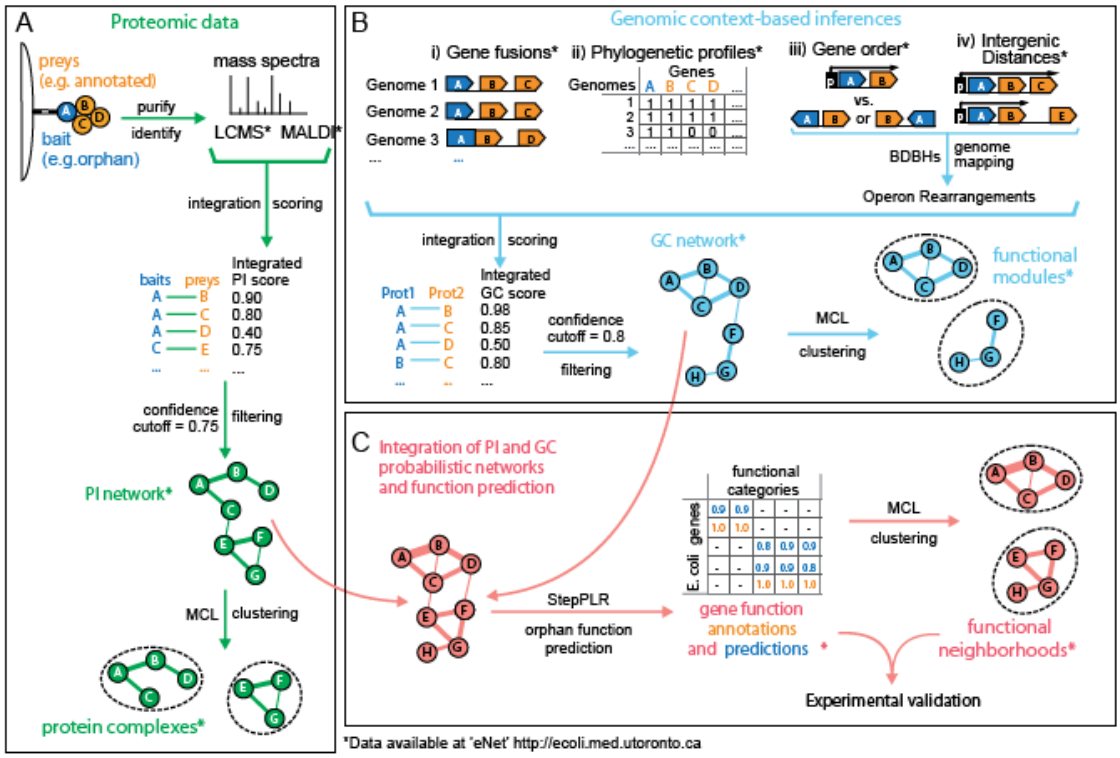


Figure 4-2. Generation, integration of different networks and orphan function prediction

(**A**) Construction of a PI network based on protein co-purification and detection by mass spectrometry. For the confidence scoring by logistic regression, datasets consisting of PI from low-throughput studies curated in DIP, BIND and IntAct (gold positives) and proteins in different subcellular localizations (gold negatives) were used for benchmarking. The resulting PI network, with edge weights corresponding to likelihood ratios, was clustered using MCL to delimit 'multiprotein complexes'. (**B**) Integration of four GC methods into a single functional interaction network using a probabilistic model (von Mering et al., 2005), whose resulting scores (edge weights) were inputted to MCL to delimit 'functional modules'. (**C**) Orphan function prediction was conducted using a 'guilt-by-association' procedure. After integration of PI and GC interactions into a single probabilistic network, a machine learning algorithm (StepPLR) newly developed for this study was used to assign functions based on the binary associations of orphans with annotated proteins, the respective interaction edge weights and the overall network topology. Correlations between vectors of these function predictions (orphans) and the annotations were then used as input to delimit 'functional neighborhoods' by clustering using MCL.

4.2.2.4 Experimental definition of the physical interaction network of the soluble proteome

We performed systematic large-scale tandem-affinity purifications of all endogenous soluble orphan and annotated proteins detectably expressed in *E. coli* W3110 under standard culture conditions [see Materials and Methods below and protocols linked with the manuscript for further details]. We used an optimized Sequential Peptide Affinity (SPA)-tagging system to

isolate multiprotein complexes (Zeghouf et al., 2004). This procedure is based on the integration of marker cassette bearing a dual-affinity tag, consisting of three FLAG sequences and a calmodulin binding peptide separated by a protease cleavage site, fused to the C-termini of targeted open reading frames in *E. coli* DY330 (W3110 background) via λ-phage "Red" mediated homologous recombination. This system enables recovery of native bacterial protein complexes at near-endogenous levels (Butland et al., 2005), minimizing spurious non-specific protein associations. Stably interacting polypeptides were subsequently detected using a highly-sensitive combination of tandem mass spectrometry (LCMS) and peptide mass fingerprinting procedures (MALDI) to increase detection coverage and accuracy, just as had been previously done in a focused investigation for highly-conserved essential *E. coli* proteins by my collaborators (Butland et al., 2005). We successfully chromosomally-tagged 1,241 new baits, aiming to verify putative interactions by reciprocal tagging where possible, for a total of 1,476 large-scale protein purifications (after including the 235 reported previously), of which 552 represented orphans.

Since proteomic datasets typically contain noise in the form of non-specific associations, we performed a careful statistical analysis and quality filtering to determine biologically meaningful physical interactions. We considered that the specificity and affinity between any two putatively interacting proteins should be correlated with the consistency of co-purification over all the experiments in which the proteins were identified (i.e. co-complexed). We therefore used an established co-purification metric (Zhang et al., 2008) to assess interaction specificity based on the similarity of the protein co-purification patterns. We then generated a single consolidated confidence score for each putative pair-wise physical interaction based on the co-purification metric together with the primary interaction evidence to penalize inconsistent or promiscuous binders (i.e. possible false-positives) using alternatively a logistic regression model and bayesian inference (Suthram et al., 2006).

The logistic regression model was trained using a reference set of curated gold-standard Protein Interactions (PIs), which represents the union of experimentally-verified physical interactions derived from low-throughput experiments extracted from the Database of Interacting Proteins (DIP) (Xenarios et al., 2000), the Biomolecular Interaction Network Database (BIND) (Bader et al., 2003) and the IntAct database (Kerrien et al., 2007). For the negative gold standards, we compiled pairs of proteins annotated with different subcellular localizations (i.e. one cytoplasmic, the other periplasmic or outer membrane-bound (Diaz-Mejia et al., 2009).

Despite its relative simplicity, the logistic regression model offered better performance than the Bayesian method (see Figure 4-3A). We therefore applied the former to our global PI

network, assigning a probabilistic confidence score for each pair of putatively interacting proteins. To minimize false positives without incurring excessive false negatives, we further filtered our network using a stringent minimum confidence cutoff of ≥ 0.75 as a high proportion (71%) of PI verified by reciprocal purification had likelihood scores at or above this threshold. Finally, we removed from consideration the ten most-highly connected 'hub' proteins which were deemed particularly abundant non-specific contaminants.

The resulting final network consisted of 5,993 high-confidence, non-redundant pair-wise interactions among 1,757 distinct *E. coli* proteins, including 451 orphans, or roughly two thirds of the predicted soluble cytoplasmic proteome. As summarized in Figure 4-3B, most (3,193, or 53%) of these physical interactions are novel, while only 47% were already reported in either the DIP, BIND or IntAct interaction databases, or previous large-scale proteomic studies (Arifuzzaman et al., 2006; Butland et al., 2005). Importantly, our filtered dataset had a comparable level of accuracy as for the much smaller set of 716 'validated' PI reported previously (Butland et al., 2005) and a genome-scale dataset of 7,123 PI (median confidence of 0.69) generated using an analogous affinity purification schema in yeast (Krogan et al., 2006).

The reliability of this dataset was also evident by two additional independent criteria. First, the mRNA expression patterns of the putatively interacting proteins were nearly as highly correlated as those of the presumably more abundant curated protein pairs determined by low-throughput experiments (Figure 4-3C). Second, despite the more limited evolutionary distribution of the orphans, the putatively interacting proteins exhibited an elevated degree of co-occurrence of the respective orthologs across other bacterial species, evident from the high mutual information of the corresponding phylogenetic profiles (see Methods), again comparable to that of interacting pairs derived from low-throughput experiments (Figure 4-3D). Collectively, these results indicate that our physical interaction network is very likely to be informative about orphan protein function.

4.2.2.5 Orphan membership within multiple protein complexes

Since macromolecular assemblies mediate biological function in cells, we partitioned our high confidence physical interaction network using the Markov clustering algorithm (MCL; see Materials and Methods) to define orphan membership as subunits of discrete multiprotein complexes. MCL simulates random walks (i.e. flux) to delimit highly connected sub-networks based on both the connectivity and the weight of the graph edges (Enright et al., 2002). In this case, the weights reflect the interaction likelihood ratios obtained by logistic regression (Figure 4-2A). The higher the flux within in a region, the more likely MCL will delimit the region as a

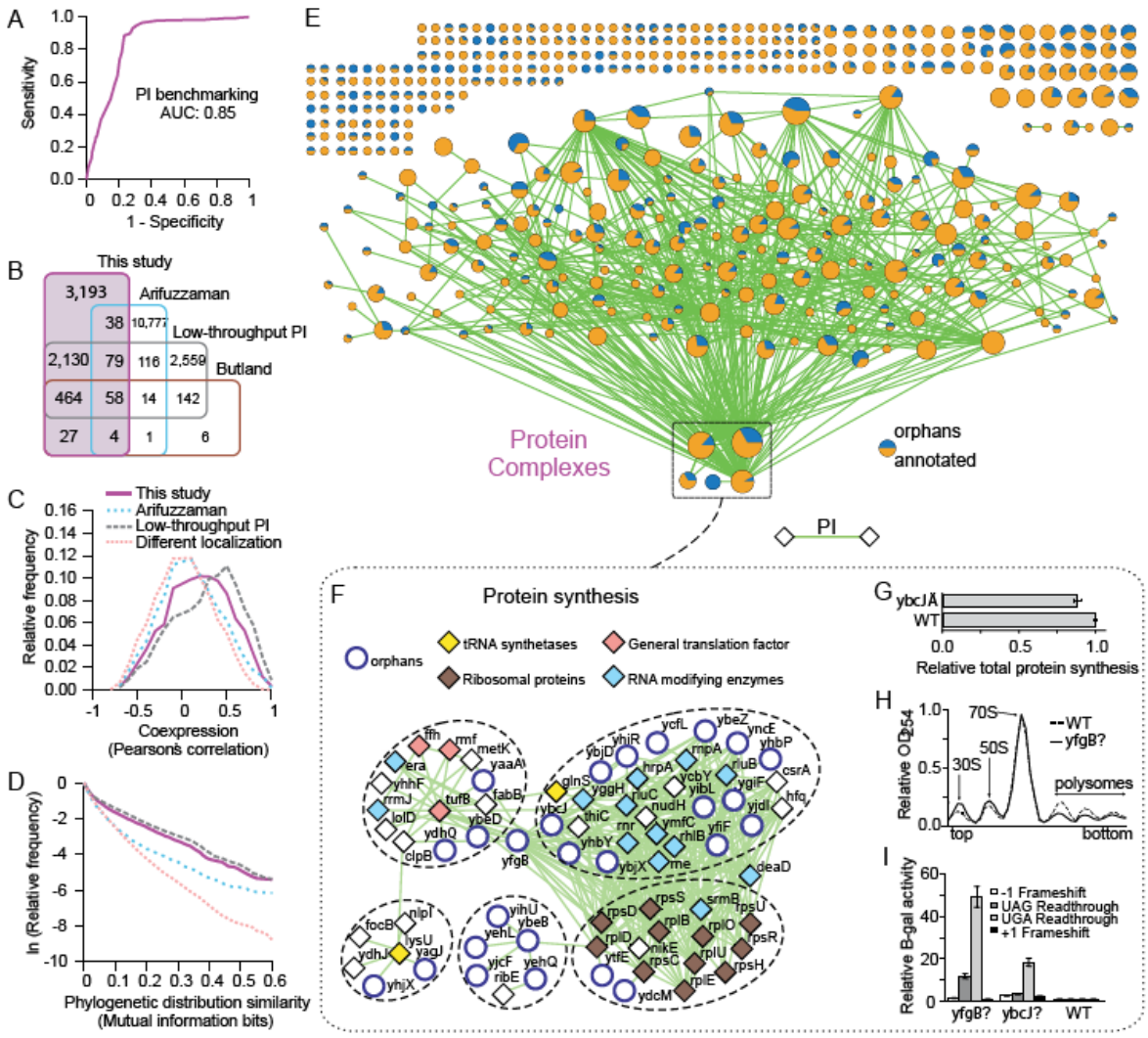


Figure 4-3. High-confidence physical interactions and putative multiprotein complexes

(**A**) Benchmarking of the experimentally-derived PI network in *E. coli* against positive and negative gold standards by ROC-curve analysis. (**B**) Overlap of PI identified in this study with previous proteomic reports (Arifuzzaman et al., 2006; Butland et al., 2005) and low-throughput PI obtained from DIP, BIND and IntAct. (**C**) Putatively interacting proteins have highly-correlated gene expression patterns and (**D**) similar phylogenetic profiles based on mutual information as for low-throughput curated PI and in contrast to control protein pairs derived from different sub-cellular compartments.. (**E**) Graphical schematic of putative stable, soluble multiprotein complexes using the GenePRO Cytoscape plugin (Vlasblom et al., 2006). Each node represents a complex, whose size reflects the number of contained proteins; edge widths reflect the number of interactions between subunits of different complexes. (**F**) Multiprotein complexes implicated in the bacterial translation apparatus; orphans mentioned in the main text are highlighted in bold. (**G**) Reduced rate of total protein synthesis in a strain lacking *ybcJ* relative to wild-type cells (WT). (**H**) Perturbed ribosome profiles in an *yfgB* deletion strain. (**I**) Elevated rates of frame-shifting and stop-codon readthrough in *yfgB* and *ybcJ* deletion strains relative to wild-type (WT). *β*-gal activity is only produced after the corresponding translational defect has occurred; error bars indicate standard deviation.

cluster (in this case, a putative multimeric protein complex). A recent comparative study (Brohee and van Helden, 2006) found that MCL is often superior to other clustering algorithms in identifying functionally-related groupings in probabilistic molecular interaction graphs and is remarkably resilient to spurious graph perturbations (e.g. missing edges).

We optimized the MCL parameters (see Materials and Methods) to partition the 5,993 PI network, generating a set of 443 putative multiprotein complexes (Figure 4-3E), most of which consist of 2-4 polypeptides. In agreement with previous reports (Brohee and van Helden, 2006), alternative clustering algorithms comparable to MCL in terms of accuracy, like Restricted Neighborhood Search Cluster algorithm (King et al., 2004), produced similar groupings (data not shown). Moreover, as was found in a proteomic survey of yeast multiprotein complexes (Krogan et al., 2006), both the subunit number and degree connectivity of the MCL clusters followed a power-law distribution. In particular, two hundred and forty four (55%) of these *E. coli* multiprotein complexes contained at least one orphan as a putative subunit, with mechanistically suggestive linkages suggestive of a concerted biological function (Figure 4-3E). The complexes also showed a significant ($p < 0.001$) enrichment in terms of functional homogeneity implying that both the annotated components and the associated orphans tend to participate in the same biological processes.

For example, 25 orphans were detected as part of a large sub-network of putative complexes involved in protein synthesis (Figure 4-3F). These include the orphans YbcJ and YncE, which physically interacted with the pseudouridylate synthase RluB, the RNA helicases SrmB and DeaD, the exoribonucleases E (Rne) and R (Rnr), and other components of the ribonucleolytic 'degradosome' responsible for mRNA degradation, suggesting a probable role in RNA processing and/or turnover. Likewise, YfgB co-purified with three translation-related complexes, including the ribosome. Consistent with these observations, the expression of YncE, which has similarity to the non-ribosomal peptide synthase AfuA of *Aspergillus fumigatus,* is reduced >9-fold upon exposure of *E. coli* to the translational inhibitor puromycin (Sabina et al., 2003). We also determined that deletion of *ybcJ* results in a significant reduction in the incorporation of ^{35}S- labeled methionine *in vivo* relative to wild-type (Figure 4-3G), indicating a decrease in the global rate of protein synthesis. Similarly, ribosome profile analysis (Figure 4-3H) showed that inactivation of *yfgB* decreased the level of mature polysomes actively engaged in mRNA translation and altered the cellular ratios of 30S and 50S ribosomal subunits relative to 70S monosomes. Moreover, both the *ybcJ* and *yfgB* mutants exhibited reduced translation fidelity (Figure 4-3I) as assayed by four reporter plasmids that measure the frequency of frameshifts and stop codon readthrough.

Other orphans is this translation sub-network include YibL, which co-purified both with YfgB and YbcJ, and with RNA processing factors involved in ribosome biogenesis, such as the RNA pseudouridine synthetases RluB/RluC and the RNA helicase DeaD, and with RppH (formerly NudH), which was recently identified as a regulator of 5'-end-dependent mRNA degradation (Barkan et al., 2007; Deana et al., 2008; Jiang et al., 2006). Similarly, the orphan YdhQ co-purified with translation elongation factor Tu, while YagJ interacted with lysine tRNA synthetase (LysU), and YjcF, which has similarity to phenylalanyl-tRNA synthetase PheT of *Bacteroides vulgatus,* bound ribosomal release factor 2 and another orphan, YbeB, which in turn was found to associate with the 50S ribosome subunit, as recently reported (Jiang et al., 2007). These results confirm that our high-confidence physical interaction network is informative about the function of at least certain orphans.

4.2.2.6 Functional interactions predicted by genomic-context methods

Although we attempted to tag and purify the entire soluble *E. coli* interactome, we failed to detect 469 orphan proteins by MALDI or LCMS, presumably because they are both membrane-associated (~35%**)** and hence not soluble**,** or are of particularly low in abundance (~40). To bypass this limitation, we applied computational methods to discern a network of high-confidence pair-wise functional interactions for all *E. coli* proteins, including those not detectable by proteomic methods, by examining the natural chromosomal clustering of bacterial genes. As illustrated in Figure 4-2B, we used four different genomic context (GC) methods, namely: (i) Gene Fusions (Enright et al., 1999; Marcotte et al., 1999a); (ii) similarity between Phylogenetic Profiles (Gaasterland and Ragan, 1998; Pellegrini et al., 1999; Tatusov et al., 1997); (iii) evolutionary conservation of Gene Order (Dandekar et al., 1998; Janga and Moreno-Hagelsieb, 2004; Overbeek et al., 1999); and (iv) Intergenic Distances (Janga et al., 2005; Rogozin et al., 2002; Snel et al., 2002) (see Materials and Methods for details). The latter two methods are independent approaches to detect operons and their subsequent rearrangements across prokaryotic genomes. In particular, the Intergenic Distances method, leads to considerably more high-quality predicted functional associations compared with the first three classic GC methods (Janga et al., 2005), and does not depend critically on the detection of orthologs in evolutionarily distant genomes, making it potentially better suited for detecting functional interactions involving orphans.

The pair-wise interactions generated by each of these prediction methods were independently evaluated by benchmarking using gold standards. Positive gold standards were defined as pairs of *E. coli* genes belonging to the same biological pathway as defined in

EcoCyc, while the negative gold standards represented pairs of annotated *E. coli* genes whose products participate in different pathways. The results of each GC method were subsequently combined to create a single unified functional association score (Figure 4-2B). Although different data integration algorithms have been developed (Chua et al., 2006; Lee et al., 2004; Nabieva et al., 2005; von Mering et al., 2005), most of these have a similar probabilistic basis and assumptions. For this study, we opted for the integration procedure used by Bork and colleagues (von Mering et al., 2005) to construct the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database. This approach treated the reliability of the associations generated by each GC method as independent probabilities, such that the likelihood of an interaction is proportional to the number of times it was observed and the degree to which each GC method contributed to the overall network reliability. Finally, we applied a stringent filter to the unified functional network to obtain a set of 74,776 high-confidence (probabilities≥ 0.80) non-redundant interactions (Figure 4-4A).

Despite the tendency of the orphans to exhibit more limited conservation notwithstanding the dependency of GC methods on homologs in multiple species (except for operon predictions based on intergenic distances (Janga et al., 2005)), our combined GC network implicated virtually all (1,367, or 96%) of the orphans in 23,365 pair-wise functional interactions. Moreover, relatively few (<18%) of our predicted interactions appear to have been reported previously (Figure 4-4B). While we could not meaningfully compare our results to an alternate set of putative functional links generated recently (Yellaboina et al., 2007) because of a lack of publicly accessible dataset scores, we found that less than 5% (3,368) of our predicted interactions are listed in the PROLINKS comparative genomics databank (Bowers et al., 2004) while only ~16% (11,842, of which only 2,613 involve an orphan) were present in STRING (v. 7.1) at a more liberal 0.7 confidence threshold. More critically, greater than 85% of our predicted orphan interactions involve a functionally-annotated *E. coli* protein, indicating a good potential to make functional inferences. The fact that PROLINKS has 1,657 predictions not attained by our integrative approach may reflect our use of a higher confidence threshold as well as differences in implementation of the GC measures and the identification of putative orthologs. For instance, whereas we used BLAST-BDBHs as criteria to detect orthologs between pairs of genomes, STRING uses COG-based definitions of orthology, while PROLINKS uses one-way BLAST hits (not necessarily orthologs). Conversely, most of the 16,585 predictions exclusive to the STRING database were compiled using text mining or alternate experimental criteria such as protein-protein interactions, whereas the highest numbers of predictions exclusive to our GC datasets come from operon rearrangements (Janga et al., 2005).

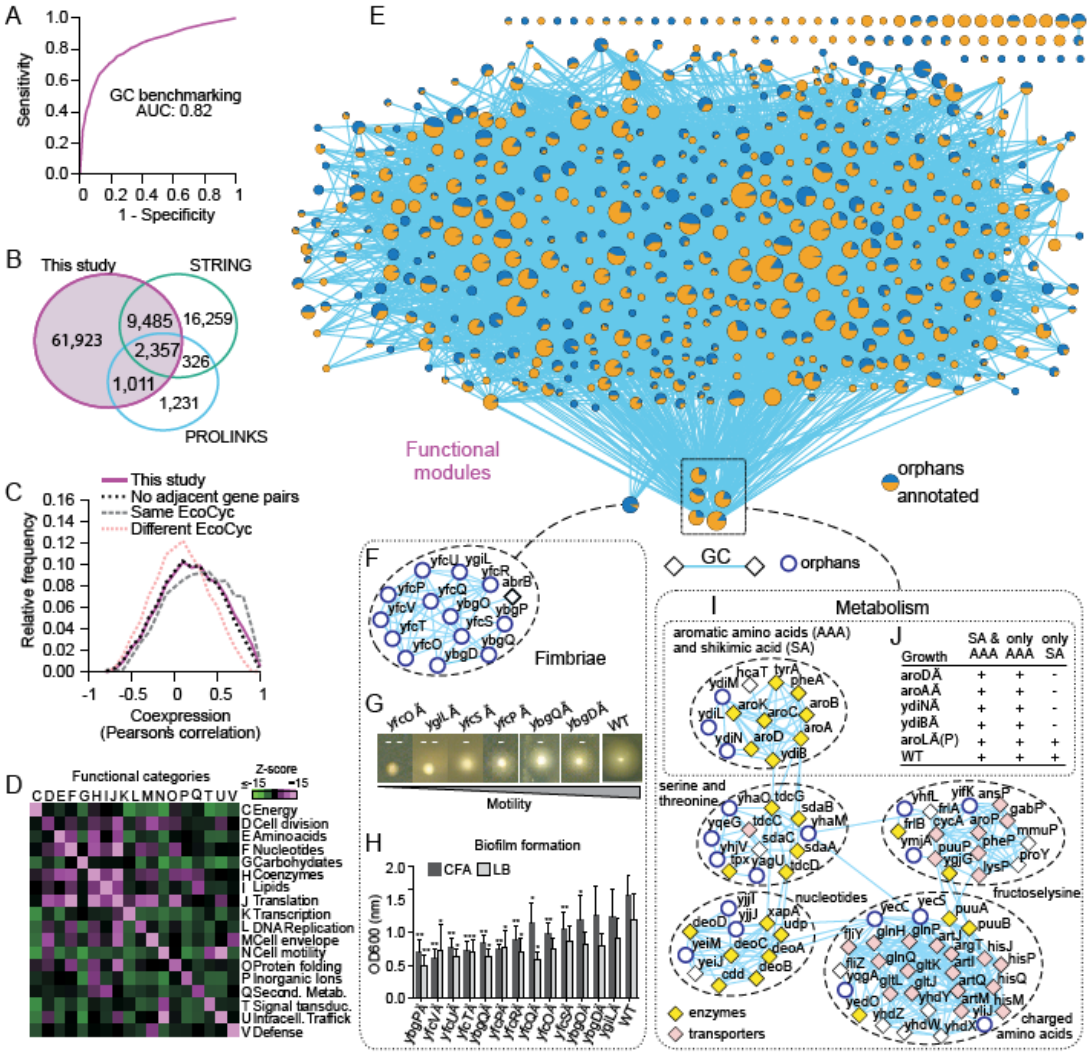


Figure 4-4. High-confidence genomic context associations and putative functional modules

(**A**) Benchmarking of unified GC interactions in *E. coli* against positive and negative gold standards by receiver operating characteristic (ROC)-curve analysis; cumulative area-under-the-curve (AUC) is shown as an overall performance measure. (**B**) Overlap of high-confidence functional interactions predicted in this study with two other public GC databases. (**C**) Even after eliminating adjacent gene pairs to control for known and predicted *E. coli* operons, functionally-linked genes have highly-correlated patterns of mRNA expression comparable to components of the same curated EcoCyc pathways rather than different pathways. (**D**) Functionally-linked genes are enriched for annotations to the same COG functional categories. (**E**) Graphical representation of putative *E. coli* functional modules; node size and colors are proportional to the number and fraction of orphan and annotated subunits, respectively, while lines represent interactions connecting modules. (**F**) Putative fimbriae-related module. (**G**) Defective motility of mutant strains deleted for orphans linked to fimbriae (as in panel F); single dashes indicate moderately impaired motility, while double dashes represent strong repression. Other mutants displaying a normal phenotype comparable to the wild-type strain BW25113 (WT) are not shown. (**H**) Defective biofilm formation by mutants deleted for fimbriae-related orphans (as in panel F); significant differences (T-test) in cell adhesion (absorbance) between mutant and WT strains are denoted by asterisks (single, $p < 0.01$; double , $p < 0.0001$); LB, Luria Bertani medium; CFA, colonization factor antigen medium. (**I**) Metabolic modules mentioned in main text. (**J**) Mutants auxotrophic for shikimic-acid and aromatic amino acids; growth on minimal 'drop-out' media is indicated.

The reliability of our unified functional association network was independently corroborated based on the high correlations of expression among putatively interacting gene pairs (Figure 4-4C), which was comparable to that observed for components of the same curated EcoCyc pathway even after eliminating all pairs of genes belonging to an experimentally-characterized operon or those which form contiguous gene pairs in *E. coli* (Figure 4-4C). We also observed a marked enrichment for interactions among proteins annotated to the same curated Clusters of Orthologous Group (COG) functional categories (Figure 4-4D), implicating by extension any associated orphans in these same processes.

4.2.2.7 Defining the participation of orphans as the components of functional modules

Groups of functionally interacting genes form functional modules centered on a common process or biochemical pathway(s). To define orphan participation as components of such modules, we partitioned the high-confidence GC network using MCL, generating a total of 507 putative functional modules consisting of two or more components (Figure 4-4E). Examination of the functional homogeneity of these predicted modules (see Materials and Methods) indicated, as for our putative multiprotein complexes, that they were highly-enriched (p <0.0001 compared with null random models) for concerted annotated biological processes, again implicating the associated orphans in these same roles. Module membership followed a characteristic power law distribution with most modules having between 2 and 10 components.

Two hundred and eighty nine (57%) of the modules had at least one of a total of 1,189 different orphans. One notable example is shown in Figure 4-4F. Diverse lines of experimental and bioinformatic evidence support the involvement of this putative module in the biogenesis and/or activity of fimbriae, appendages or pili that are shorter than the characteristic flagellum of gram-negative bacteria, which mediate cell adhesion, biofilm formation, motility and host invasion (Fronzes et al., 2008; Hahn et al., 2002). For instance, 12 of the 13 orphan components possess sequence characteristics of bacterial adhesins and chaperone/Usher pili protein families (Madera et al., 2004; Nuccio and Baumler, 2007). Gene expression profiling studies (Domka et al., 2007; Domka et al., 2006) have previously established that most of these orphans are also coordinately induced during biofilm formation. Perhaps most compellingly, we found that single gene *E. coli* knock-out mutants of 6 of the 13 orphans display markedly reduced swarming capabilities in semi solid agar (Figure 4-4G), while 11 out of 13 mutants were significantly impaired for biofilm formation *in vitro* as compared with a wild type control (Figure

4-4H). Taken together, these observations strongly implicate this set of orphans in the formation and/or proper function of fimbriae.

Several other prominent modules are shown in Figure 4-4I. These comprise the orphans YdiN, YdiL and YdiM predicted (based on operon rearrangements) to functionally interact with several members of the Aro- operon known to participate in the metabolism of shikimate, a precursor of aromatic amino acids. Consistent with this, *ydiN*, *aroD* and *ydiB* are reportedly over-expressed when *E. coli* is grown in media containing shikimate as the sole carbon source (Johansson and Liden, 2006). Moreover, we found that deletion of either *ydiN* or *ydiB* resulted in phenotypic auxotrophy for shikimatic- and aromatic amino acids, comparable to that observed after loss of known aromatic amino biosynthetic genes (e.g., *aroA* and *aroD*).

Other functional modules include *frlA* / *frlB*, part of the Frl operon of *E. coli* responsible for the import and metabolism of the alternative carbon source fructoselysine, together with the orphan YifK, which has sequence characteristics of a transporter (Diaz-Mejia et al., 2009), implicating it in electrochemical potential-driven uptake of this sugar. Conversely, two orphans, YecC and YecS, had functional associations consistent with linkages to amino acid biosynthesis and nucleotide metabolism, four (YagU, YqeG, YhaO and YhaM) were linked to a putative module involved in transport and metabolism of threonine and serine, while three others (YjjI, YeiM, and YjjJ) were found in a module enriched for factors involved in nucleotide transport and degradation of deoxyribonucleosides. Taken as a whole, these results suggest discrete functional relationships for many previously unannotated proteins, even implicating certain orphans within specific pathways.

4.2.2.8 Improved functional inference within an integrated network framework

Examination of the extent of overlap between our physical and functional networks, both in terms of common binary interactions and shared components among the derived complexes (from PI) and modules (from GC), indicated that they are largely complementary. Since a similar trend was also evident comparing other existing curated *E. coli* physical interaction datasets (derived from either low- throughput or other high-throughput studies) with independent functional predictions (e.g. GC inferences from STRING;), this presumably stems in part from the incomplete coverage obtained by these different approaches. Regardless, these observations imply that the union of PI and GC networks is necessary to capture the widest spectrum of biologically-relevant interactions. Indeed, it has been shown previously that combination of physical interactions with functional genomic inferences, each statistically-

weighted according to dataset quality, can markedly improve both functional coverage and accuracy (Beyer et al., 2007; Ideker and Sharan, 2008; Lee et al., 2004; Myers and Troyanskaya, 2007; von Mering et al., 2005). We therefore merged our experimental and predicted associations with the same method used to generate the unified GC network (Figure 4-2C; see Materials and Methods).

The resulting combined probabilistic network consisted of 80,370 high-confidence (probability ≥ 75%) putative pair-wise interactions encompassing virtually the entire proteome of *E. coli*, including 2,769 (99%) annotated proteins and 1,375 (96%) functional orphans. Graph analysis of this final integrated network indicated that the orphans tended to have a lower overall connectivity and betweenness centrality, measured as the number of shortest paths going through a given node, relative to annotated components, suggesting more peripheral positions in the integrated networks. However, the orphans also exhibited lower average closeness, defined as the average length of shortest paths between any two nodes, and had similar overall clustering coefficients, indicating that in general the orphans are functionally connected rather than isolated from the annotated gene products. These observations implied that consideration of both the individual associations and overall placement of the orphans within the integrated interaction network would facilitate functional deduction.

We therefore devised a new network-based function prediction method (termed StepPLR; see Figure 2C and Materials and Methods) to exploit the global topological similarity among all the protein pairs and their corresponding functional annotations in the integrated network. Our method assigns functions to unannotated orphans based on the functional information from their first-order (direct) and second-order (indirect) annotated neighbors in the integrated functional association network using penalized logistic regression models and a stepwise variable selection procedure to deduce optimal functional profiles (see supplementary methods accompanying the manuscript for a detailed protocol). We based our classifications on the discrete COG functional categories and on the hierarchical, multifunctional terms of the Gene Ontology (GO)(Ashburner et al., 2000; Camon et al., 2004) and MultiFun classification schemas (Serres et al., 2004). To avoid potential sources of false predictions, we removed any proteins labeled with the evidence codes IPI (for ‘inferred from protein interaction’) and IGC (for ‘inferred from genomic context method’) when generating the GO reference set, as well as proteins in poorly characterized categories in COGs and MultiFun.

We found that StepPLR had better precision and recall compared to several other widely used guilt-by-association procedures tested, such as majority-counting and chi-squared-based methods. Although the performance achieved for the different functional categories varied, our approach generated AUC values of 0.8 or higher for most of the COG (83%), GO (67%) and MultiFun (53%) categories and was relatively insensitive to the number of annotated proteins per function. Moreover, since our method exploited the correlation among the different categories, most orphans had multiple biologically-consistent predicted functions.

4.2.2.9 Functional neighborhoods

As displayed graphically in Figure 4-5A, our prediction procedure ultimately linked many of the orphans to specific, functionally-related protein 'neighborhoods'. We again made use of the MCL algorithm to objectively delimit functionally highly homogeneous ($p < 0.0001$) protein groupings based on the profile similarity of annotations and predictions shown in this figure. One notable example is the protein translation machinery (Figure 4-5B), which has 23 associated orphans. To independently verify the functional relevance of these assignments, we examined the effects of deleting the corresponding genes in terms of conferring sensitivity to drugs that inhibit protein synthesis. Consistent with expectation of a direct role in protein synthesis, and similar to loss of *bona fide* annotated translation factors and tRNA synthetases, the mutant strains exhibited statistically significant ($p < 0.05$) differential sensitivity as compared to wild type and unrelated gene mutants to a variety of antibiotics that selectively block protein translation (Figure 4-5C).

We also examined an alternate group of orphans (YafP, YiaD and YbcM) associated with the flagellar biogenesis and motility apparatus (Figure 4-5D). Single-gene knockout mutants annotated components in this neighborhood exhibit decreased motility in semi-solid agar as compared to wild-type *E. coli* strains (Rajagopala et al., 2007). Consistent with our functional predictions, we likewise found that deletion of *yafP* ablated cell motility *in vitro* (Figure 4-5E), similar to mutants lacking core flagellum motor proteins (e.g. FliH, FliM), while loss of *yiaD* and *ybcM* reduced swarming (i.e. decreased halo formation) to an extent comparable to perturbation of other established flagellar components (e.g. *flgJ* and *fliR*). A previous study (Bresolin et al., 2006) using phenotypic complementation analysis had suggested that a *ybcM*-ortholog in *Yersinia enterocolitica* is likely an AraC-type regulatory protein involved in controlling bacterial motility. These results suggest that, akin to several other recently discovered novel motility components (Girgis et al., 2007; Rajagopala et al., 2007), these orphans are required for the proper assembly and/or subsequent locomotion of the *E. coli* flagella.

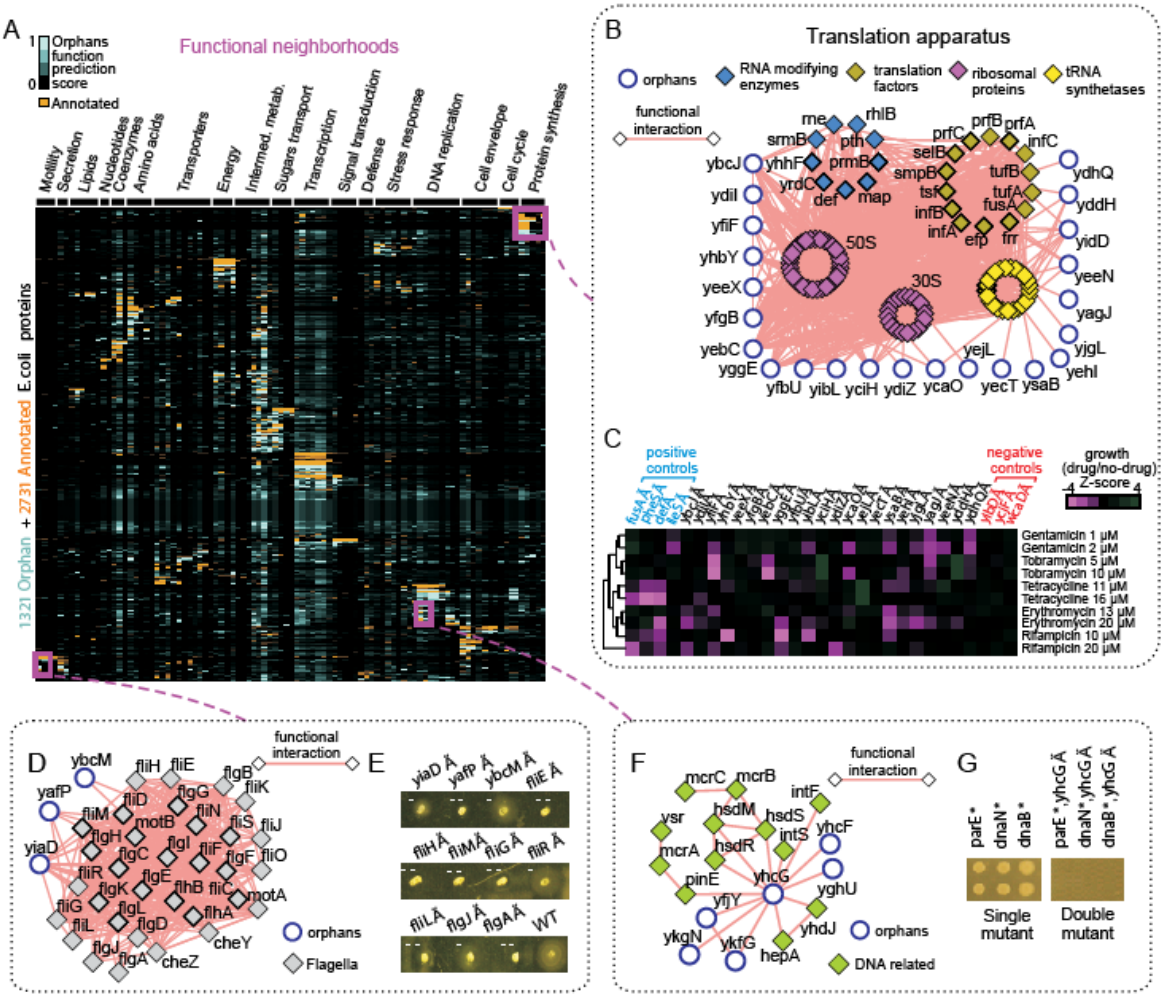


Figure 4-5. The functional neighborhoods of *E. coli*

(**A**) A 'clustergram' displaying existing annotations (*orange*) and the predicted functions (this study; *blue*) for all the protein-coding genes of *E. coli* (*x-axis*) and their associated biological processes (*y-axis*)(descriptions from different functional schemas i.e, COG, Multifun and GO not shown due to lack of space). Proteins were clustered using MCL based on the paired similarity of the functional annotations and predictions in this matrix to delimit 'functional neighborhoods'. (**B**) Putative functional neighborhood showing high-confidence integrated (combined PI and GC networks) interactions of select orphans with the protein synthesis machinery. For clarity, individual names of ribosomal proteins and tRNA synthetases are not shown. (**C**) Heatmap showing the differential sensitivity of orphan deletion strains to antibiotics targeting protein synthesis relative to the colony size in the absence of drug. Mutants deleted for annotated proteins from this neighborhood are shown as positive controls , while deletion mutants lacking genes not contained within this neighborhood are shown as negative controls; . (**D**) Neighborhood with three orphans putatively involved in flagellum assembly and motility. (**E**) Deletions of the corresponding components reduce swarming capability; *single dash*, moderately impaired motility; *double dash*, strong repression. (**F**) Sub-network of orphans associated with DNA enzymes. (**G**) Deletion of the orphan *yhcG* results in synthetic lethality when combined with hypomorphic alleles (*) of three essential DNA replication factors (*parE, dnaN, dnaB*).

Many other orphans were predicted to have roles in other conserved biological systems, such as DNA replication. For example, as shown in Figure 4-5F, we identified the orphan YhcG in association with DNA processing enzymes, including the restriction complexes HsdMRS and McrABC, the integrases IntF and IntS, and the recombinase PinE. YhcG has sequence characteristics of the PD-(D/E)XK superfamily of nucleases involved in DNA recombination and repair (Kosinski et al., 2005). Consistent with these observations, we found that deletion of *yhcG* results in a synthetic-lethal phenotype (Figure 4-5G) when combined with hypomorphic alleles of the replicative primosome (*dnaB*), DNA polymerase III (*dnaN*), and DNA topoisomerase IV (*parE*), consistent with a direct role in DNA replication or the resolution of critical intermediates.

4.3 DISCUSSION & CONCLUSION

Defining the precise biological roles and relationships of bacterial gene products in an often dynamically changing physiological context is a challenging proposition. Historically, systematic assessments of protein function in bacteria have tended to rely on molecular inferences based on sequence alignments and domain architectures, while experimental characterization has traditionally been driven by specific scientific interests rather than with the aim of providing the broader community with unbiased collections of functionally-related proteins and phenotypes. Since the biological role of a protein is not necessarily reflected in its primary sequence, the elucidation of molecular interaction networks can provide an alternate perspective even in the absence of detailed phenotypic data (Ideker and Sharan, 2008; Lee et al., 2008). Here, we have opted to view a model microbial cell mechanistically as a series of modular molecular interaction networks that underlie the major biochemical processes that mediate cell homeostasis and proliferation, wherein the functional attributes of particular gene products are reflected in their overall patterns of associations.

To this end, we have generated an extensive compendium of physical and functional linkages covering almost the entire protein-coding complement of *E. coli*. This led to the elucidation of hundreds of putative soluble multiprotein complexes and functional modules encompassing virtually all the many gene products currently lacking public annotations. While existing integrative probabilistic interaction databases like STRING (von Mering et al., 2005) and EcID (Andres-Leon et al., 2009) provide valuable additional binary interactions that are potentially useful for protein function prediction or as complementary evidence to those reported in this study, our machine learning strategy goes beyond describing binary interactions by explicitly describing the most probable biological functions of the orphans. Of particular noteworthiness, our functional predictions and phylogenetic projections associate a sizeable

fraction of the functional orphans with core bacterial processes, suggesting they may have previously eluded detection in part due to prior analytical biases.

Since the various methods used in this study discover different types of molecular relationships and each has its own intrinsic bias, complementary information was obtained through data integration. The limited overlap between the high-confidence physical and functional interaction networks presumably stems in part from to the incomplete coverage typically achieved by high-throughput experiments and their methodological differences (Rajagopala et al., 2007; Yu et al., 2008). For example, certain orphans were difficult to evaluate by GC methods due to a lack of apparent orthologs at medium-to-high evolutionary distances, which hinders comparative genomic inferences. Likewise, although we performed large-scale tandem affinity tagging and purification under near-native physiological conditions to generate highly purified preparations of stable, endogenous multiprotein complexes, we did not achieve complete coverage of the proteome. We did not attempt to purify a large number of membrane-associated proteins, which require specialized solubilization procedures, while the soluble proteins that we failed to tag or detect by mass spectrometry were presumably either of very low abundance or not expressed in our growth conditions.

Comparison of our physical interaction network with analogous public datasets produced for other model species, such as worm, fly, yeast and even the bacterium *H. pylori,* revealed very limited (<1%) overlap. These observations are congruent with recent findings by Uetz and colleagues (Rajagopala et al., 2007) showing that only a third (49) of the 173 experimentally-derived PI in the cell motility network of the spirochete *Treponema pallidum* are predicted to occur in the ε-proteobacteria *Campylobacter jejuni* on the basis of orthology could subsequently be confirmed by targeted two-hybrid testing. The limited overlap between proteomic datasets presumably reflects a combination of incomplete coverage by various experimental assays, methodological differences and imperfect conservation.

The observation that the intersection of functional genomics inferences with low-throughput curated physical interaction data is somewhat higher might be explained by two non-mutually exclusive ways: first, protein-protein interactions reported in the literature based on traditional biochemical methods might be biased towards the most evolutionarily conserved multiprotein complexes, which tend to be enriched for essential components with broadly distributed phylogenetic profiles that are more easily and accurately predicted by GC methods; second, the relatively high sensitivity of the two complementary forms of protein mass spectrometry used in this study may have resulted in the detection of lower abundance orphan proteins that have previously not been studied in depth.

The last point is consistent with the notion that different proteomic methods capture different physical interaction types (Yu et al., 2008). Hence, alternate proteomic methods, such as two-hybrid screens (Parrish et al., 2007; Rain et al., 2001; Rajagopala et al., 2007; Titz et al., 2008) or *in vivo* protein-fragment complementation assays (Tarassov et al., 2008), may be better suited for detecting certain physical interactions currently underrepresented in our dataset. In a similar vein, additional functional relationships will undoubtedly be uncovered by different experimental and computational procedures, such as high-throughput comparative analysis of mutant cellular phenotypes (Baba et al., 2006), genome-wide genetic interaction screens (Butland et al., 2008; Typas et al., 2008), and automated text mining (Hoffmann and Valencia, 2004; Rzhetsky et al., 2008).

The topological properties inherent to biological networks (e.g. their hierarchical organization and degree distributions) combined with incomplete interactome coverage make establishing definitive functional groupings difficult (Sharan et al., 2007). Our approach was to take into account both the correlations among functional categories and the overall topological structure of the integrated network to generate a more balanced probabilistic model. While alternate methods may provide enhanced interpretations of the organizational properties of the PI and GC networks, the functional enrichment and experimental validations established here suggest that our network-based computational inferences provide a reasonable perspective for exploring bacterial protein function. Similar strategies have resulted in powerful predictors of protein function in Eukaryotes (Marcotte et al., 1999a; McDermott et al., 2005; Murali et al., 2006; Myers and Troyanskaya, 2007; Schwikowski et al., 2000). The potential trade-off is that additional error or uncertainty may have occasionally been introduced by assuming functional similarity among more loosely connected proteins. Moreover, the probabilities associated with particular functional terms may not be directly comparable. Functional orphans associated with very well-characterized biological processes are more likely to be correctly assigned by computational methods (Myers and Troyanskaya, 2007) while those associated with relatively poorly studied proteins will tend to remain obscure. Nonetheless, they can be grouped together on the basis of specific PI, GC or even other functional associations and hence serve as functional groupings rather than isolated entities.

In general, the high confidence functional relationships we inferred for *E. coli* could be validated by independent experimental tests, and can be extrapolated to other bacterial species, including pathogens. Over 35% of the orphans find orthologs as far away as Archaea, and hence are likely associated with the same basic housekeeping processes we predict for *E. coli*, such as formation of the cell wall and protein synthesis. Conversely, our systematic

comparisons also revealed some unique aspects of the orphans in the evolutionary history of *E. coli,* such as the potential fimbriael factors that appear to be restricted to Enterobacteriaceae. One interpretation is that orphans with limited phylogenetic distributions contribute to fine tuning of adaptive physiological responses upon changing environmental conditions, as previously suggested for peripheral metabolic genes acquired by horizontal transfer (Pal et al., 2005). Alternatively, some orphans might belong to the well conserved biological systems which still need to be characterized for their functional role.

4.4 MATERIALS AND METHODS

4.4.1 PI network generation

Large-scale SPA tagging and purifications were performed essentially as previously described (Butland et al., 2005; Zeghouf et al., 2004). Briefly, a DNA cassette encoding the SPA-tag and a selectable marker flanked by gene-specific targeting sequences was amplified by PCR using primers with homology to a selected locus. The cassette was then transformed and integrated using homologous recombination in the lysogenic *E. coli* strain DY330 (W3110 background), which harbors the highly efficient λ-phage-encoded homologous recombination enzymes *exo, bet*, and *gam* under the control of the temperature-sensitive CI857 repressor (the “Red” system)(2), to create a C-terminal fusion with the protein of interest. Strains in which the PCR product has integrated were subjected to antibiotic selection, and tagged protein expression was confirmed by Western blotting.

Two complementary mass spectrometry techniques (gel-based MALDI peptide mass fingerprinting and gel-free LCMS shotgun sequencing) were used to detect physically interacting proteins. Details about the large-scale strain culture, protein extraction and purification, and protein identification procedures are provided as supplementary protocols accompanying the published manuscript. Scoring of tentative PI from the LCMS and MALDI assays was conducted using a logistic regression model using reference PI obtained by low-throughput experiments curated in the DIP, BIND and IntAct databases (Bader et al., 2003; Kerrien et al., 2007; Xenarios et al., 2000) as a positive training set. Our negative training set consisted of pairs of proteins in which one component was experimentally determined or predicted with high confidence to be cytoplasmic and the other residing in the outer membrane or the periplasm (Diaz-Mejia et al., 2009); inner membrane proteins were discarded from this negative dataset since they are in physical proximity (and hence could potentially physically interact) to cytoplasmic and periplasmic proteins. Our logistic regression procedure also took into account

the degree of consistency of co-purifying protein pairs, balancing the tradeoff between "spoke" and "matrix" representation models of interactions within co-purified groups of proteins to decrease the false discovery rate. We then combined the scores derived from LCMS and MALDI into a a single PI network using a previously established procedure for integrating probabilistic networks (von Mering et al., 2005), which assumes the reliabilities of associations generated by these methods are independent. To facilitate independent critical evaluation, all our processed interaction data is available through the website in HUPO-PSI molecular interaction reporting format (standard level 2.5).

4.4.2 GC network generation

The four GC methods used to predict functional interactions among *E. coli* proteins were based on: (i) functional linkages among genes which fuse to form a single open reading frame in at least one other genome i.e. Gene Fusion (Enright et al., 1999); (ii) the mutual information of the coordinated presence or absence of pairs of genes across a set of 440 non-redundant genomes i.e. Phylogenetic Profiles (Moreno-Hagelsieb and Janga, 2008; Pellegrini et al., 1999); and (iii) the natural chromosomal association of bacterial genes in operons as detected by two alternative methods, namely (a) the tendency of genes forming operons to show small Intergenic Distances (Moreno-Hagelsieb and Collado-Vides, 2002; Salgado et al., 2000), and (b) the conservation of Gene Order, in which a confidence value for each adjacent pair of genes present in the same strand was used as indicator that those genes likely form an operon as compared with the conservation of adjacent genes found in opposite strands (Janga and Moreno-Hagelsieb, 2004). For the last two methods, subsequent Operon Rearrangements were also detected by genomic mapping of orthologs across 440 non-redundant bacterial genomes (Janga et al., 2005).

For all four GC methods, we used the BLAST BDBHs as an operational definition of orthology. To avoid circularity, the prediction scores of the four GC methods were benchmarked separately using as positive reference set proteins belonging to the same metabolic pathway according to EcoCyc (Keseler et al., 2005), and as negatives proteins in different pathways. A single, unified high-confidence functional association network was then constructed by integrating the interaction predictions generated by the four genomic context methods using a the same scoring model (von Mering et al., 2005) used to integrate the MALDI and LCMS data.

(Space left for an enhanced layout of the text)

4.4.3 Clustering

Protein clusters were generated from three different networks using MCL (Enright et al., 2002): (i) the PI network (generating protein complexes), (ii) the unified GC network (generating functional modules); and (iii) the function prediction/annotation profiles derived from the integration of PI and GC networks (generating functional neighborhoods). The core idea of MCL is to simulate random walks (i.e. flux) among the proteins (nodes) within each network to delimit regions with high flux, taking into account the connectivity and weight of interaction edges. In this work, edge weights correspond to the likelihood of pairwise protein interactions in each network. In each case, the global MCL inflation parameter, which tunes the granularity of the delimited clusters, was optimized by balancing the mass fraction of clusters and efficiency of partitions. The resulting clusters were individually assessed for functional homogeneity in terms of COG annotations as described previously (Loganantharaj et al., 2006).

4.4.4 Network-based function prediction and benchmarking

Our algorithm (StepPLR) for assigning biological functions is essentially a network topology-based method in which the functions of the orphans are predicted based on the functions of their associated annotated proteins in the immediate (direct) and adjacent (indirect) network vicinity. Briefly, a single network integrating the high-confidence PI and GC probabilistic networks was first created using the same scoring model (von Mering et al., 2005) used to integrate the PI data and the four GC networks. Then the weighted topological overlap (Zhang and Horvath, 2005) between each pair of protein nodes in the integrated network was calculated to determine the correlated functional profiles based on a penalized logistic regression model. Finally, a stepwise variable selection procedure to optimize function profiles in the final logistic regression was used. Only functional categories with at least 15 annotated *E. coli* proteins were used in our integrated functional association network: 18 COG classes, corresponding to widespread bacterial protein functions; 19 biological classes from MultiFun, in which the proteins can have multiple annotations based on different classification criteria; and 51 biological process classes in GO. Other guilt-by-association representative methods (e.g. majority-counting and chi-squared-based) were also evaluated (results not shown).

REFERENCES

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. and Kanaya, S. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* **7**, 207.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

Andres-Leon, E. A., Ezkurdia, I., Garcia, B., Valencia, A. and Juan, D. (2009). EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res* **37**, D629-D635.

Aoki, K., Ogata, Y. and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol* **48**, 381-90.

Apweiler, R. (2001). Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Brief Bioinform* **2**, 9-18.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-9.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J. et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-31.

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H. C., Hirai, A. et al. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* **16**, 686-91.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008.

Bader, G. D., Betel, D. and Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50.

Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2.

Bandyopadhyay, S., Kelley, R., Krogan, N. J. and Ideker, T. (2008). Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* **4**, e1000065.

Barkan, A., Klipcan, L., Ostersetzer, O., Kawamura, T., Asakura, Y. and Watkins, K. P. (2007). The CRM domain: an RNA binding module derived from an ancient ribosome-associated protein. *Rna* **13**, 55-64.

Barker, D. and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* **1**, e3.

Barrett, C. L., Herring, C. D., Reed, J. L. and Palsson, B. O. (2005). The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. *Proc Natl Acad Sci U S A* **102**, 19103-8.

Beyer, A., Bandyopadhyay, S. and Ideker, T. (2007). Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* **8**, 699-710.

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O. and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**, R35.

Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V. et al. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40.

Bresolin, G., Neuhaus, K., Scherer, S. and Fuchs, T. M. (2006). Transcriptional analysis of long-term adaptation of Yersinia enterocolitica to low-temperature growth. *J Bacteriol* **188**, 2945-58.

Brohee, S. and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488.

Brouwer, R. W., Kuipers, O. P. and van Hijum, S. A. (2008). The relative value of operon predictions. *Brief Bioinform* **9**, 367-75.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A. and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol* **5**, R6.

Butland, G., Babu, M., Diaz-Mejia, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O. et al. (2008). eSGA: E. coli synthetic genetic array analysis. *Nat Methods.* **5**, 789-95.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N. et al. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531-7.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32**, D262-6.

Campillos, M., von Mering, C., Jensen, L. J. and Bork, P. (2006). Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res* **16**, 374-82.

Chua, H. N., Sung, W. K. and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* **22**, 1623-30.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S. et al. The genetic landscape of a cell. *Science* **327**, 425-31.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-8.

Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-62.

Deana, A., Celesnik, H. and Belasco, J. G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature* **451**, 355-8.

Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (2003). Prediction of protein function using protein-protein interaction data. *J Comput Biol* **10**, 947-60.

Diaz-Mejia, J. J., Babu, M. and Emili, A. (2009). Computational and experimental approaches to chart the Escherichia coli cell-envelope-associated proteome and interactome. *FEMS Microbiol Rev* **33**, 66-97.

Domka, J., Lee, J., Bansal, T. and Wood, T. K. (2007). Temporal gene-expression in Escherichia coli K-12 biofilms. *Environ Microbiol* **9**, 332-46.

Domka, J., Lee, J. and Wood, T. K. (2006). YliH (BssR) and YceP (BssS) regulate Escherichia coli K-12 biofilm formation by influencing cell signaling. *Appl Environ Microbiol* **72**, 2449-59.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. and Palsson, B. O. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3**, 121.

Fronzes, R., Remaut, H. and Waksman, G. (2008). Architectures and biogenesis of non-flagellar protein appendages in Gram-negative bacteria. *Embo J*.

Gaasterland, T. and Ragan, M. A. (1998). Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics* **3**, 199-217.

Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I., Salgado, H. et al. (2008). RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**, D120-4.

Girgis, H. S., Liu, Y., Ryu, W. S. and Tavazoie, S. (2007). A comprehensive genetic characterization of bacterial motility. *PLoS Genet* **3**, 1644-60.

Godzik, A., Jambon, M. and Friedberg, I. (2007). Computational protein function prediction: are we making progress? *Cell Mol Life Sci* **64**, 2505-11.

Gotoh, O. (1999). Multiple sequence alignment: algorithms and applications. *Adv Biophys* **36**, 159-206.

Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J. D., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L. S. et al. (2005). Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* **436**, 861-5.

Hahn, E., Wild, P., Hermanns, U., Sebbel, P., Glockshuber, R., Haner, M., Taschner, N., Burkhard, P., Aebi, U. and Muller, S. A. (2002). Exploring the 3D molecular architecture of Escherichia coli type 1 pili. *J Mol Biol* **323**, 845-57.

Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y. and Chen, Y. (2006). Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **6**, 4023-37.

Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.

Hawkins, T. and Kihara, D. (2007). Function prediction of uncharacterized proteins. *J Bioinform Comput Biol* **5**, 1-30.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein--protein interaction data. *Yeast* **18**, 523-31.

Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nat Genet* **36**, 664.

Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P. et al. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol* **7**, e96.

Huber, W., Carey, V. J., Long, L., Falcon, S. and Gentleman, R. (2007). Graphs in molecular biology. *BMC Bioinformatics* **8 Suppl 6**, S8.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res* **18**, 644-52.

Janga, S. C., Collado-Vides, J. and Moreno-Hagelsieb, G. (2005). Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res* **33**, 2521-30.

Janga, S. C. and Moreno-Hagelsieb, G. (2004). Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res* **32**, 5392-7.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412-6.

Jiang, M., Datta, K., Walker, A., Strahler, J., Bagamasbad, P., Andrews, P. C. and Maddock, J. R. (2006). The Escherichia coli GTPase CgtAE is involved in late steps of large ribosome assembly. *J Bacteriol* **188**, 6757-70.

Jiang, M., Sullivan, S. M., Walker, A. K., Strahler, J. R., Andrews, P. C. and Maddock, J. R. (2007). Identification of novel Escherichia coli ribosome-associated proteins using isobaric tags and multidimensional protein identification techniques. *J Bacteriol* **189**, 3434-44.

Johansson, L. and Liden, G. (2006). Transcriptome analysis of a shikimic acid producing strain of Escherichia coli W3110 grown under carbon- and phosphate-limited conditions. *J Biotechnol* **126**, 528-45.

Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S. A., Palsson, B. O. and Agarwalla, S. (2006). Experimental and computational assessment of conditionally essential genes in Escherichia coli. *J Bacteriol* **188**, 8259-71.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30.

Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R. and Kasif, S. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A* **101**, 2888-93.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R. et al. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Res* **35**, D561-5.

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M. and Karp, P. D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* **33**, D334-7.

King, A. D., Przulj, N. and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013-20.

Kosinski, J., Feder, M. and Bujnicki, J. M. (2005). The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics* **6**, 172.

Kretschmann, E., Fleischmann, W. and Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* **17**, 920-6.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P. et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637-43.

Lasko, P. (2000). The drosophila melanogaster genome: translation factors and RNA binding proteins. *J Cell Biol* **150**, F51-6.

Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555-8.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A. G. and Marcotte, E. M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181-8.

Letovsky, S. and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19 Suppl 1**, i197-204.

Linghu, B., Snitkin, E. S., Holloway, D. T., Gustafson, A. M., Xia, Y. and DeLisi, C. (2008). High-precision high-coverage functional inference from integrated data sources. *BMC Bioinformatics* **9**, 119.

Loganantharaj, R., Cheepala, S. and Clifford, J. (2006). Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression. *BMC Bioinformatics* **7 Suppl 2**, S5.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K. and Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* **8**, 299.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* **32**, D235-9.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-3.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6.

Massjouni, N., Rivera, C. G. and Murali, T. M. (2006). VIRGO: computational prediction of gene functions. *Nucleic Acids Res* **34**, W340-4.

McDermott, J., Bumgarner, R. and Samudrala, R. (2005). Functional annotation from predicted protein interaction networks. *Bioinformatics* **21**, 3217-26.

Moreno-Hagelsieb, G. and Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl 1**, S329-36.

Moreno-Hagelsieb, G. and Janga, S. C. (2008). Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins* **70**, 344-52.

Murali, T. M., Wu, C. J. and Kasif, S. (2006). The art of gene function prediction. *Nat Biotechnol* **24**, 1474-5; author reply 1475-6.

Myers, C. L., Robson, D., Wible, A., Hibbs, M. A., Chiriac, C., Theesfeld, C. L., Dolinski, K. and Troyanskaya, O. G. (2005). Discovery of biological networks from diverse functional genomic data. *Genome Biol* **6**, R114.

Myers, C. L. and Troyanskaya, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **23**, 2322-30.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21 Suppl 1**, i302-10.

Nuccio, S. P. and Baumler, A. J. (2007). Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol Mol Biol Rev* **71**, 551-75.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-901.

Pal, C., Papp, B. and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**, 1372-5.

Parrish, J. R., Yu, J., Liu, G., Hines, J. A., Chan, J. E., Mangiola, B. A., Zhang, H., Pacifico, S., Fotouhi, F., DiRita, V. J. et al. (2007). A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol* **8**, R130.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci* **4**, 1145-60.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-8.

Pereira-Leal, J. B., Enright, A. J. and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins* **54**, 49-57.

Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F. and Barton, G. J. Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods* **7**, S16-25.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33 Database Issue**, D501-4.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V. et al. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211-5.

Rajagopala, S. V., Titz, B., Goll, J., Parrish, J. R., Wohlbold, K., McKevitt, M. T., Palzkill, T., Mori, H., Finley, R. L., Jr. and Uetz, P. (2007). The protein network of bacterial motility. *Mol Syst Biol* **3**, 128.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A. et al. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* **5**, 11.

Rentzsch, R. and Orengo, C. A. (2009). Protein function prediction--the power of multiplicity. *Trends Biotechnol* **27**, 210-9.

Riley, M. (1993). Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**, 862-952.

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T. et al. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* **34**, 1-9.

Rison, S. C., Hodgman, T. C. and Thornton, J. M. (2000). Comparison of functional annotation schemes for genomes. *Funct Integr Genomics* **1**, 56-69.

Rives, A. W. and Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A* **100**, 1128-33.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A. and Koonin, E. V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30**, 2212-23.

Ruan, J., Dean, A. K. and Zhang, W. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Syst Biol* **4**, 8.

Rudd, K. E. (1998). Linkage map of Escherichia coli K-12, edition 10: the physical map. *Microbiol Mol Biol Rev* **62**, 985-1019.

Rzhetsky, A., Seringhaus, M. and Gerstein, M. (2008). Seeking a new biology through text mining. *Cell* **134**, 9-13.

Sabina, J., Dover, N., Templeton, L. J., Smulski, D. R., Soll, D. and LaRossa, R. A. (2003). Interfering with different steps of protein synthesis explored by transcriptional profiling of Escherichia coli K-12. *J Bacteriol* **185**, 6158-70.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. and Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97**, 6652-7.

Samanta, M. P. and Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A* **100**, 12579-83.

Schwikowski, B., Uetz, P. and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257-61.

Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. and Rosenow, C. (2003). Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res* **13**, 216-23.

Serres, M. H., Goswami, S. and Riley, M. (2004). GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Res* **32**, D300-2.

Serres, M. H. and Riley, M. (2000). MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics* **5**, 205-22.

Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol* **24**, 427-33.

Sharan, R., Ulitsky, I. and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol* **3**, 88.

Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* **3**, e42.

Shoemaker, B. A. and Panchenko, A. R. (2007b). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* **3**, e43.

Slonim, N., Elemento, O. and Tavazoie, S. (2006). Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol* **2**, 2006 0005.

Snel, B., Bork, P. and Huynen, M. A. (2002). The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A* **99**, 5890-5.

Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-8.

Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. and Ideker, T. (2006). A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7**, 360.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H. and Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science* **320**, 1465-70.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631-7.

Tipton, K. F. (1994). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *Eur J Biochem* **223**, 1-5.

Titz, B., Rajagopala, S. V., Goll, J., Hauser, R., McKevitt, M. T., Palzkill, T. and Uetz, P. (2008). The binary protein interactome of Treponema pallidum--the syphilis spirochete. *PLoS ONE* **3**, e2292.

Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L. et al. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. *Nat Methods.* **5**, 781.

Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* **21**, 697-700.

Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C. and Wodak, S. J. (2006). GenePro: a Cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics* **22**, 2178-9.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33**, D433-7.

Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E. E. and Zhu, J. (2009). Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput Biol* **5**, e1000616.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-40.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91.

Yao, Z. and Ruzzo, W. L. (2006). A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* **7 Suppl 1**, S11.

Yellaboina, S., Goyal, K. and Mande, S. C. (2007). Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res* **17**, 527-35.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. et al. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science.* **322**, 104-110.

Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D., Beattie, B., Emili, A. and Greenblatt, J. F. (2004). Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *J Proteome Res* **3**, 463-8.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17.

Zhang, B., Park, B. H., Karpinets, T. and Samatova, N. F. (2008). From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* **24**, 979-86.

Zhao, X. M., Chen, L. and Aihara, K. (2008a). Protein function prediction with high-throughput data. *Amino Acids* **35**, 517-30.

Zhao, X. M., Chen, L. and Aihara, K. (2008b). Protein function prediction with the shortest path in functional linkage graph and boosting. *Int J Bioinform Res Appl* **4**, 375-84.

5 Structure and dynamics of post-transcriptional regulatory networks directed by RNA-binding proteins

CONTENTS OF CHAPTER 5

OUTLINE

Gene expression is a highly controlled process which is known to occur at several levels in eukaryotic organisms. Although traditionally messenger RNAs have been viewed as passive molecules in the pathway from transcription to translation there is increasing evidence that their metabolism is controlled by a class of proteins called RNA-binding proteins (RBPs). In this chapter, I provide an overview of the recent developments in our understanding of the repertoire of RBPs across diverse model systems and discuss the approaches currently available for the construction of post-transcriptional networks governed by them. I also present the first analysis of the network properties of a post-transcriptional system in a model eukaryote using currently available data and discuss the implications of understanding the dynamic properties of this important class of regulatory molecules as more data detailing their dynamic, spatial and tissue-specific maps across diverse model systems accumulates. I argue that such developments would not only allow us to gain a deeper understanding of regulation at a level which has been under-appreciated over the past decades but would also us to use the newly developed high-throughput approaches to interrogate the prevalence of these phenomena in different states and thereby study their relevance to physiology and disease across organisms.

CONTRIBUTION TO THE WORK IN THIS CHAPTER

Please note that the work presented in this chapter is the result of the following two publications and my contribution to the work excludes the organization of the post-transcriptional network in yeast and the calculations performed on understanding the expression dynamics of RBPs, which were all performed by Nitish Mittal. I performed all other analyses.

1) Structure and dynamics of post-transcriptional network directed by RNA-binding proteins
Sarath Chandra Janga and Nitish Mittal
Invited book chapter for *Landes Bioscience Press* for an edited book on "*RNA infrastructure: RNA processing and regulatory networks*"

2) Dissecting the expression dynamics of RNA-binding proteins in post-transcriptional regulatory networks
Nitish Mittal, Nilanjan Roy, M. Madan Babu and Sarath Chandra Janga
Proc. Natl. Acad. Sci. U S A. 106(48): 20300-05, 2009

5.1 INTRODUCTION

Gene expression is a highly regulated process and is controlled at several levels. In eukaryotes, control of gene expression first occurs at the level of transcription, where transcription factors regulate the synthesis of RNA of specific gene in response to different internal and external stimuli. On the other hand, at the protein level, several post-translational modifications, such as phosphorylation by kinases and ubiquitin ligases, are known to spatially and temporally control the availability of functional protein products within the cell. However, a much less understood level of gene expression regulation, which occurs between these two layers, is due to the post-transcriptional control of RNAs. It is now increasingly known that this level is controlled by numerous factors with major players being the RNA-binding proteins (RBPs) (Glisovic et al., 2008; Keene, 2007; Mata et al., 2005). Therefore, intricate co-ordination of regulation from these three different layers is important for finely controlling the flow of genetic information from genes to proteins in different conditions. Indeed, changes in gene expression due to aberrations at any of these three levels have been shown to be responsible for the cause of a number of disorders (Cookson et al., 2009; Cooper et al., 2009; Feinberg and Tycko, 2004; Lukong et al., 2008; Nica and Dermitzakis, 2008).

Development of DNA microarray technology has made it possible to measure the expression of each annotated gene at the transcript level. Indeed, this technique has been the high-throughput approach of choice to efficiently characterize the transcriptomes of several model organisms. One common assumption in DNA microarray experiments is that the level of mRNA of particular gene reflects the amount of protein and there is little regulation at the post-transcriptional level. Recent studies comparing the high-throughput data for mRNA and protein-abundances indicate that there is a very weak correlation between the number of transcripts and protein products of a gene, challenging this notion (Gygi et al., 1999; Washburn et al., 2003). This suggests that the regulation of gene expression at the post-transcriptional level is predominant. For instance, in eukaryotic pathogen, *Trypanosoma cruzi*, it is well known that gene expression is primarily controlled at post-transcriptional level through RNA binding proteins (RBPs) (Noe et al., 2008). These studies suggest the extensive role of post-transcriptional regulation in controlling gene expression in eukaryotes (Campbell et al., 2003; Foth et al., 2008).

In eukaryotes, transcription and translation occur in different compartments. This allows for a plethora of options to control RNA at the post-transcriptional level, including their splicing, polyadenylation, transport, mRNA stability, localization and translational control (Glisovic et al.,

2008; Keene, 2007). Although some early studies revealed the involvement RBPs in the transport of mRNA from nucleus to the site of their translation, increasing evidence now suggests that RBPs regulate almost all of the post-transcriptional steps shown in Figure 5-1A. For example, in humans, Nova protein is associated with splicing (Ule et al., 2003), PUF family proteins have been shown to play an important role during *Caenorhabditis elegans* oogenesis (Lublin and Evans, 2007), Tap protein, like its yeast homolog Mex67, was reported as a bona fide mRNA nuclear export factor (Gruter et al., 1998), Puf3p in yeast was shown to be responsible for localization of mitochondrial transcripts (Saint-Georges et al., 2008) and Pab1 was reported to regulate the initiation of translation (Kessler and Sachs, 1998). While the extensive role of RBPs in post-transcriptional control of cellular processes has been reviewed by several groups (Glisovic et al., 2008; Keene, 2007; Lukong et al., 2008; Mata et al., 2005), in yeast alone I found that the known RBPs (see Methods) are involved in multiple cellular processes and components based on Gene Ontology analysis. All these aspects highlight the importance of RBPs in regulating gene expression at post-transcriptional level.

Due to their central role in controlling gene expression at post-transcriptional level, alteration in expression or mutations in either RBPs or their RNA targets (i.e., the transcripts which physically associate with the RBP) have been reported to be the cause of several human diseases such as muscular atrophies, neurological disorders and cancer (Cooper et al., 2009; Kim et al., 2009; Lukong et al., 2008; Musunuru, 2003). In particular, disorders such as myotonic dystrophy (DM) and oculopharyngeal muscular dystrophy (OPMD) have been attributed with RNA's gain-of-function - CUG repeat expansion in the case of myotonic dystrophy protein kinase (DMPK) (Musunuru, 2003) and GCG repeat expansion in exon 1 of the RBP, PABPN1 in the case of OPMD (Lukong et al., 2008) respectively. On the other hand, diseases like opsoclonus-myoclonus ataxia (POMA) and spinal muscular atrophy (SMA) have been reported to be due to the RBPs loss of function (Lukong et al., 2008), suggesting that mutations in either RBP or any of its interacting RNA target sequences can lead to extensive variations in their expression patterns and result in a number of diseases. In addition to the fitness defects that variations in RBPs can bring about in cells, it has been recently shown in yeast that RBPs form an important class of prionogenic proteins (Alberti et al., 2009).

(Space left for an enhanced layout of the figure)

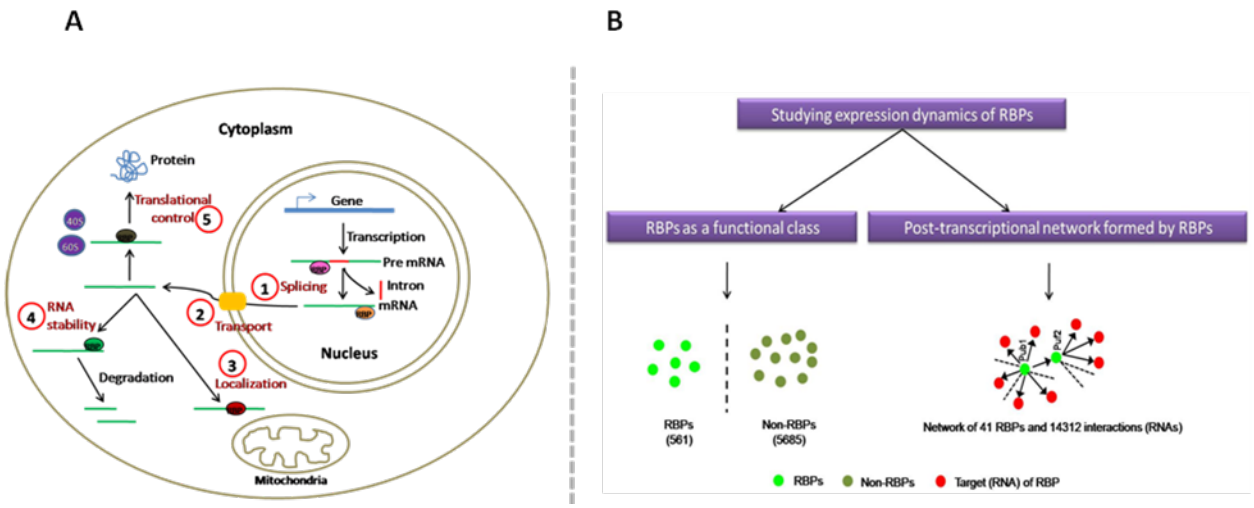


Figure 5-1: A) Schematic diagram showing the extensive role of RBPs in various post-transcriptional processes at different locations in eukaryotic cells. Circled number indicates the process in which RBPs are involved. RBPs are major players in splicing pre-mRNA into mature mRNA in the nucleus which are then exported into the cytoplasm by various other RBPs. In addition, RBPs are responsible for the localization of mRNAs to distinct sub-cellular compartments such as the mitochondria. In the cytoplasm, RBPs are also involved in governing the stability of transcripts by binding the substrate RNAs and in controlling the translation of mRNAs into corresponding protein products. For this reason, RBPs have been found to be key players either directly or indirectly responsible for the cause of several disorders due to changes in regulation they bring about at post-transcriptional level. B) In this study, an analysis of the sequence properties of the RBPs and the structure of the post-transcriptional network formed by RBPs followed by a detailed analysis on the expression dynamics of RBPs at two distinct levels is presented. First involved, RBPs as a functional class, where we compared the properties of RBPs with rest of the protein coding genes in the entire genome of *Saccharomyces cerevisiae*. This involved comparison of 561 RBPs against 5685 non-RBPs in the whole genome. Second, we studied the relationship between the RBP's connectivity, defined as the number of target mRNAs which are bound by a given RBP and their transcript and protein stability, transcript and protein expression, rate of translation and expression noise.

All these observations raise the questions: are RBPs finely controlled in terms of their expression patterns and are there constraints on their expression patterns depending on the number of distinct RNA targets they control? To address this, in what follows I present an overview of the analysis that was performed on the post-transcriptional network formed by RBPs in yeast, *S. cerevisiae* at two distinct levels shown in Figure 5-1B. The first involved asking whether RBPs as a group show distinct dynamic properties in comparison to non-RBPs in the whole genome. The second comprised of understanding the constraints placed on dynamic properties of RBPs in relation to the number of distinct transcripts controlled by them. Our analysis at the first level revealed that RBPs, as a functional class, are rapidly turned over (i.e., less stable) at the transcript level and are tightly controlled at the protein level. Analysis of the post-transcriptional network formed by RBPs indicated that highly connected RBPs are more abundant and ubiquitously present within the cell.

In this chapter, I attempt to provide a comprehensive overview and preliminary insights on this quickly developing area of post-transcriptional regulatory networks formed by RBPs by organizing the work done into three major sections, namely sequence attributes and functional processes associated with RBPs, methods used for the construction of the networks formed by them and finally discuss the structure and dynamics of these post-transcriptional networks based on recent publicly available data.

5.2 RESULTS

5.2.1 RNA binding proteins and post-transcriptional regulation

RNA binding proteins (RBPs) are key regulators of different steps in the metabolism of RNA in eukaryotes. As shown in Figure 5-1A, they participate in the processing of pre-mRNA which includes splicing, poly-adenylation and capping to get mature mRNA. Following which, they are responsible for mediating the transport of mRNA from nucleus to cytoplasm. RBPs are also found to facilitate and control the localization, translation, stability and degradation of mRNA. To regulate the different steps of RNA metabolism, RBPs bind to RNA and form ribonucleoprotein complexes (RNP). Depending upon whether RBPs are bound to pre-mRNA or mRNA, RNPs are classified as hnRNP or mRNP respectively. RNPs are inherently highly dynamic complexes due to their ability to associate and dissociate with various RBPs to mediate different steps of RNA metabolism. Some RBPs associated with RNP complexes are known to remain bound to their target RNA during all the steps of the RNA processing, from splicing to translation. For instance, SF2/ASF, a member of the SR class of RBPs in mammals, is found to facilitate splicing, export and translation initiation of its target RNA (Sanford et al., 2004; Zhong et al., 2009) . Similarly Npl3, a yeast SR protein, has also been shown to interact with pre-mRNA and regulate the events from splicing to translational elongation (Gross et al., 1998). Similarly, neuronal ELAV protein also regulates the fate of its target RNA by mediating the events from poly-adenylation to translation (Pascale et al., 2008). On the other hand, several RBPs are also responsible for participating in specific steps of RNA metabolism such as the Nova protein, which is associated with splicing in neuronal cells (Ule et al., 2003; Ule et al., 2006). Tap protein, like its yeast homolog Mex67, was reported to be a bona fide mRNA nuclear export factor (Gruter et al., 1998). All these examples highlight 1) the role of RBPs in regulating the expression of genes in multiple steps at post-transcriptional level and 2) the complex combinatorial interplay of different RBPs to integrate various post-transcriptional events to fine tune the availability of transcripts both spatially and temporally.

Table 5-1. Common RNA binding domains in putative RBPs of the yeast *S. cerevisiae*, their frequency in RBPs and domains most often associated with these RNA binding domains according to the Pfam (Finn et al.) domain database.

Domain	Pfam accession	Description	Protein frequency	Frequent Occurrence of other domain
RRM_1	PF00076	RNA recognition motif (RRM). Many eukaryotic proteins containing one or more copies of a putative RNA-binding domain of about 90 amino acids are known to bind single-stranded RNAs	0.105	RRM_1, Lsm_interact
DEAD	PF00270	DEAD/DEAH box helicase. Members of this family include the DEAD and DEAH box helicases	0.042	Helicase C,
KH_1	PF00013	K homology (KH) domain is a doamain of 70 amino acid and present in diverse RBPs.	0.015	KH_1
PUF	PF00806	Pumilio-family RNA binding repeat. Puf domain usually occurs as a tandem repeat of eight domains	0.013	PUF, RRM_1
WD40	PF00400	WD-40 repeats (also known as WD or beta-transducin repeats) are short ~40 amino acid motifs, often terminating in a Trp-Asp (W-D) dipeptide	0.013	WD40

RBPs bind to their RNA targets with the help of several domains having different specificity and affinity. Some of the most common domains are RRM (RNA recognization motif), KH (K homology domain), SR (serine arginine domain), Zn-finger, Pumilio/FBF (PUF domain) and Sm (Glisovic et al., 2008) . Table 5-1 shows the most frequently occurring RNA binding domains in the yeast, *S. cerevisiae*, along with the commonly appearing partner domains in the conventional list of 560 RBPs reported recently by Hogan and co-workers (Hogan et al., 2008) (see Materials and Methods). A large number of proteins have been predicted as RBPs in several model organisms including humans on the basis of the presence of these commonly occurring domains. A list of approximate number of RBPs identified in different model organisms is shown in Table 5-2 along with a reference to the study reporting it. For instance, in *C. elegans* approximately 500 proteins are annotated as RBPs on the basis of the presence of one or more RNA binding domains. In the yeast, *S. cerevisiae* about 560 proteins have been reported as putative RBPs till date. In human, more than 1000 proteins are considered as RBPs of which there are 497 that contain at least one RRM domain (Maris et al., 2005). Other than these putative RBPs (on the basis of previously known RNA binding domains), several metabolic

enzymes have also been shown to bind to RNA molecules (Ciesla, 2006). For example Aco1, TCA cycle enzyme, in yeast *S. cerevisiae* binds to several RNAs encoded by the mitochondrial genome (Hogan et al., 2008). Likewise, recent studies have also shown the ability of RBPs to bind to DNA suggesting that some of the known RBPs might act as unconventional DNA-binding proteins (Hu et al., 2009).These examples indicate the potential for the existence of novel classes of RBPs in eukaryotes with yet to be discovered functional roles.

Table 5-2. Putative number of RBPs reported in different organisms.

Organism	putative RBPs	Approximate number of genes	Reference
S. cerevisiae	561	7000	(Hogan et al., 2008)
C. elegans	500	20000	(Lee and Schedl, 2006)
D. Melanogaster	300	13290	(Lasko, 2000)
MusMusculus	380	28287	(McKee et al., 2005)
Human	800	30000	(Sanchez-Diaz and Penalva, 2006)

5.2.2 Methods to Identify RBPs and their targets

Although, several RBPs have been identified on the basis of conservation of domains in different organisms, targets of these RBPs are poorly understood. Therefore, several methods have been employed to identify the targets of RBPs, both in vitro and in vivo. The list of some commonly used methods for identification of RBP targets have been described in Table 5-3. Traditionally, RNA targets for known RBPs have been identified in vitro by using cross-linking immunoprecipitation followed by electromobility shift assays (Pinero et al., 2000; Thomson et al., 1999). More recently, one hybrid (Wilhelm and Vale, 1996) and three hybrid assays (SenGupta et al., 1996) have been used to identify in vivo interaction of a RBP and RNA molecule. But these traditional methods have limitations in their ability to identify new targets. Therefore, other in vivo assays have been developed to identify the novel targets of a RBP such as ultraviolet (UV) cross-linking and immunoprecipitation (CLIP) and RNP immunoprecipitation-microarray (RIP-CHIP). These assays usually work on a similar concept where in (i) the complex of RBP and its target RNAs is first extracted and (ii) the target RNA identified. However, they differ in the procedure used for extracting RBP-RNA complexes and identification of target RNAs. For example, in ultraviolet (UV) cross-linking and immunoprecipitation (CLIP)

method, cells are exposed to ultraviolet light to crosslink RBP-RNA molecules inside the cells. Then cells are lysed and cross-linked RBP-RNA complexes are immunoprecipitated using antibody against the RBP of interest. Further, RNA is isolated from the complexes and identified by RT-PCR. For instance, in a study to discover the targets of the splicing factor Nova, thirty four transcripts were identified by using the CLIP method (Ule et al., 2003).

In RNP immunoprecipitation-microarray (RIP-Chip) method, cells are not treated with UV light to crosslink RBP-RNA complex but cells are lysed directly and native RBP-RNA complexes for RBP of interest are purified from the cell lysate using immunoprecipitation method. Following which RNA is isolated from the complexes and identified by using high-density oligonucleotide microarrays. The targets of Puf family of RBPs and other RBPs in yeast *S. cerevisiae* have been identified by using modified RIP-Chip method, where tandem affinity tagged (TAP) RBPs are used to facilitate the immunoprecipitation (Gerber et al., 2004; Hogan et al., 2008). These studies showed that the RNA targets vary from 1-1300 approximately for the studied RBPs in yeast *S. cerevisiae*. For instance Nop13, responsible for pre-18s rRNA processing, has 2 RNA targets whereas Npl3 and Mex67, both involved in mRNA export, have 1266 and 1150 RNA targets respectively (Hieronymus and Silver, 2003; Hogan et al., 2008).

Another fundamental area of exploration in elucidating post-transcriptional networks is the identification of the repertoire of RBPs across organisms and several approaches both computational and experimental have been developed in recent years. Computational approaches involve the identification of the set of protein-coding genes which contain the bonafide RNA-binding domains, following which manual curation of the collected set is undertaken to identify a high confidence set of RBPs (Galante et al., 2009; Hogan et al., 2008). Experimental techniques comprise of employing the protein chip of an organism of interest to probe for the potential binding of the cellular RNA molecules and is analogous to the attempts to characterize the repertoire of DNA-binding proteins (Fasolo and Snyder, 2009; Hall et al., 2004; Hu et al., 2009; Zhu et al., 2001). Another strategy which has been developed to identify the RBPs attached to known RNA molecule is the PNA-assisted identification of RBPs (PAIR) (Zeng et al., 2006). This assay utilizes specific mRNA binding probe (PNA) that has ability to cross the cell membrane and can bind to RNA of interest. This probe also contains photoactivable amino acid adduct p-benzophenylalaline (Bpa) which can covalently crosslinked to adjacent RBP on photoactivation. After delivery of PNA, cells are exposed to ultra violet light for crosslinking of PNA to RBPs associated with RNA of interest. Cells are then lysed, treated with RNase and PNA-RBP adducts are isolated by using sense oligo (bind to PNA) coupled magnetic beads. Following which RBPs are identified by mass spectrometry. This method has

been used to identify the RBPs associated with ankylosis (ank) RNA, a panneuronal dendritically localized RNA (Zielinski et al., 2006).

Table 5-3. Different methods to identify novel RBPs, their targets or RBP_RNA interactions.

Method	Description	Reference
Three hybrid	in vivo yeast genetic method to detect and analyze the RNA-RBP interaction of known RNA and RBP. This methods is based on the binding of bifunctional RNA to both the two hybrid protein which activates the expression of reporter gene.	(SenGupta et al., 1996)
RNAcompete	in vitro identification RNA binding specificity of RBP. High concentration of RNA pool are used and incubated with tagged RBP. High concentration of RNA provide the competition for bindng and hence technique gets its name. RBP-RNA complexes are purified and microarray is used to identify the specific binding sites of RBP.	(Ray et al., 2009)
RIP-ChIP	in vivo identification of RNA targets for RBP of interest. Cells are lysed and RBP-RNA complexes are immunoprecipitated in native state. Target RNA are extrated from the RBP-RNA complexes. Target RNAs are identified by microarray method where control RNAs are total RNA of the cell.	(Tenenbaum et al., 2000)
CLIP	in vivo identification of RNA targets for RBP of interest. Cells are treated with ultraviolet light to covalently crosslink RBP-RNA complex. Cells are lysed and RBP-RNA complexes are immunoprecipitated and RNA are identified by RTPCR.	(Ule et al., 2003)
PAIR	in vivo identification of novel RBPs. mRNA binding PNA probe is delivered to cell. Cells are exposed to ultraviolet light that enable PNA to bind with RBP. Cells are lysed and PNA-RNA-RBP complexes are immunoprecipitated and RBPs are identified by mass spectrometry.	(Zielinski et al., 2006)
SERF	in vitro selection of RNA fragments that bind to RBP. Random pool of fragmented RNA is generated. RNA pool is incubated with RBP in test tube. RBP-RNA complex is extracted by filtration on nitrocellulose membrane. Selection cycle is repeated several time and selected RNA fragment are cloned and identified the consensus sequences binding to RBP	(Stelzl and Nierhaus, 2001)
TRAP	in vivo system for identification of RNA-RBP interaction in yeast. Transformation of reporter mRNA encoding GFP protein and expression of RBP of interest. Fluorescence intensity of GFP is measured to know the binding of RBP of interest. Higher the interaction leads to lower expression and low fluorescence intensity.	(Paraskeva et al., 1998)
SNAAP	in vitro method used to identify mRNAs bind to specific RBP. Purified tagged RBP is treated with cell lysate. Immunoprecipitation of mRNP using antibody against tag. Target mRNA are identified by differential display method	(Rodgers et al., 2002)
Quantitative proteomics	in vitro method to identify RBPs bind to specific RNA sequence. RNA aptamer tagged RNA sequence is incubated with cell lysate. RNA aptamer-RNA-RBPs complex is purified. RBPs are identified by using mass spectrometer.	(Butter et al., 2009)

5.2.3 RBPs and post-transcriptional operons

In prokaryotes, it has been long known that the genes involved in similar processes tend to cluster on chromosomes and are transcribed together using the same promoter thus forming DNA operons such as the well studied, Gal and Lac operons. On the other hand, in eukaryotes, DNA operons are rare. However, following this notion recently the concept of post-transcriptional operons has been proposed in eukaryotes (Keene and Tenenbaum, 2002) which has become possible due to the availability of the wealth of information on RBP-RNA interactions. According to this concept, diverse RNAs related to a common biological process are regulated by similar RBPs. For instance, in yeast *S. cerevisiae*, study of the RBP-RNA interactions by modified RIP-Chip method has revealed that each member of Puf family RBPs bind with functionally and cytotopically related RNAs (Gerber et al., 2004). Puf1 and Puf2 have been shown to bind to mRNAs of membrane associated proteins. Similarly, Puf3 binds to cytoplasmic mRNAs of mitochondrial proteins. Likewise, Nova protein was found to regulate splicing of pre-mRNA encoding components of inhibitory synapses and a stem loop binding protein (SLBP) was involved solely in splicing and translation of replication dependent histone RNAs (Townley-Tilson et al., 2006). Further examples in support of post-transcriptional operons have been reviewed extensively elsewhere (Keene, 2007; Keene and Lager, 2005). These examples demonstrate the role of RBPs in view of post-transcriptional operons for coordinating the expression of functionally related genes in eukaryotes.

5.2.4 Post-transcriptional network formed by RBPs

Development of several high throughput approaches has increased the amount of data for targets of RBPs in diverse organisms. This data of RBPs and their targets could be utilized to construct RBP-RNA interaction network which is also typically referred to as post-transcriptional regulatory network (see Figure 5-1B). This post-transcriptional network is represented in the form of a directional network with each edge corresponding to a regulatory link between the nodes as shown in Figure 5-2A. In this directed network, one set of nodes are RBPs forming the regulatory proteins while the other set of nodes are RNAs encoded by either protein-coding or non-protein coding genes referred to as the target nodes. These two nodes (regulator node and target node) are joined by an arrow starting from regulator node and directing towards target node. The target RNA may belong to diverse functional proteins including other RBPs. This network can also contain loops as a link starting from RBP and targeting itself, typically referred to as autoregulation of an RBP (Figure 5-2B). This loop structure suggests that RBP can bind to

its own RNA and control its metabolism at transcript level. There are several examples suggesting the auto-regulation of RBPs at post-transcriptional level. For instance, in humans, RBPs such as AUF1, HuR, KSRP, NF90, TIA-1 and TIAR were reported to associate with their own mRNA and other RBPs (Pullmann et al., 2007).

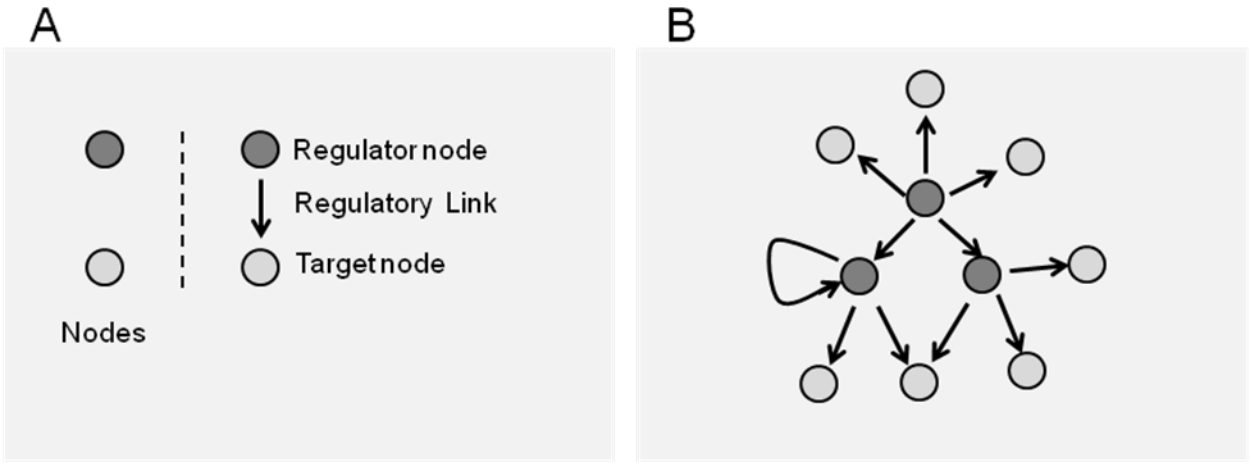


Figure 5-2: Concept figure showing the RBP mediated post-transcriptional regulatory network. A) Dark (Regulator) and light (Target) grey circles denote nodes in the network. These nodes are linked to each other via a directional arrow starting from regulator (which is RBP in the network) and pointing towards target (which may be RNA or miRNA) in the directional network. These linked nodes simply indicate that RBP (Dark grey circle) binds to RNA/miRNA of target gene (Light grey circle) and regulate its metabolism. B) Shows a toy network representing a dense set of RBP-RNA interactions with different RBPs having diverse targets. The targets of one RBP in the network may be RNA of other genes or miRNA (dark and light circle linked by arrow), the RNA of the RBP itself (loop from dark circle) and the RNA of other RBPs (two dark circles linked by an arrow).

Due to the availability of the network of post-transcriptional interactions for a considerable fraction of RBPs in model systems such as *S. cerevisiae* (see Materials and Methods), it has become possible to address several questions concerning the structure and organization of post-transcriptional networks directed by RBPs. Table 5-4 summarizes some of the properties which govern the structure of this network obtained as described in the Materials and Methods section. It is evident from this table that majority of the mRNA transcriptome encoded by about ~ 70% of the genes had significant associations with at least one of the RBPs screened for RNA interactions, and on average, each distinct yeast mRNA was found to interact with three of the RBPs, suggesting the potential for a combinatorial and multidimensional network of regulation. Indeed, it was found that the average connectivity of a node in this network was ~7 indicating that most nodes in this network have more number of targets and/or more the number of RBPs controlling them.

Table 5-4. Properties defining the structure of the post-transcriptional network of RBPs and their target RNAs in the model eukaryote, *S. cerevisiae*. Dataset employed for characterizing the network structure was obtained from Hogan et. al.(Hogan et al., 2008) and all the network properties are calculated using igraph, a publicly available R package for analyzing graphs [http://cneurocvs.rmki.kfki.hu/igraph/ & http://www.r-project.org].

Property	Definition	Value*
No. of edges	Each edge corresponds to a single RBP-RNA interaction. Hence, total edges represent all the interactions in the post-transcriptional network	19396
No. of vertices/nodes	Total number of nodes, which comprise of both the RBPs as well as the RNAs, encoding for both protein coding and non-coding genes. This network comprises of 41 RBPs which are screened for their RNA targets.	5398
Degree or Connectivity	Degree or connectivity refers to the number of interactions a protein or RNA has in this network – the higher the connectivity (i.e., hub nodes) the more the number of targets and/or more the number of RBPs controlling it.	7.18
Clustering coefficient	Clustering coefficient of a node reflects the extent to which the neighbors of a given node are interconnected among themselves to what is expected theoretically and indicates the cohesiveness or local modularity of the network. Average value taken over all nodes reflects the modularity of the network.	0.37
Betweenness	Betweenness centrality of a node measures the number of shortest paths between all pairs of nodes in the network that pass through a node of interest – the higher the number of paths that pass through a node, the more important it is.	43.11
Average path length	Average length of the shortest paths between all pairs of nodes in the network.	2.65
Closeness	Closeness centrality is defined as the inverse of the average length of all the shortest paths from a node of interest to all other nodes in the network - note that closeness centrality defined this way implies that higher the closeness value, the higher the importance (centrality) of a node.	0.38
Diameter	The diameter of a network is the length of the longest path among all the shortest paths defined between two nodes. It gives an estimation of the distance between nodes in the network.	6
Graph density	The density of a network is the ratio of the number of edges to the number of total possible edges.	1.33x10-3
Power law fit (exponent-alpha)	Fitting a power-law distribution function to the degree distribution of the network to study whether the network is likely to exhibit a scale-free network structure.	1.77

* Note that average values for the entire network are reported for properties which are defined for specific node or edge.

Other measures of centrality like betweenness and closeness which provide a measure of the importance of a node in a network, shown in this table, also reflect this trend (see (Junker et al., 2006) and references there in for comprehensive definitions). For instance, the average length of the shortest path between two nodes in this network which gives an indication of the

distance between nodes suggests that most nodes are separated by no more than 3 edges - a measure reflecting the dense networking in this network. Similarly, diameter of a network which refers to the longest of all the shortest paths between a pair of nodes is about 6 indicating that two nodes in this network are separated by no more than 6 edges. Likewise, clustering coefficient which is a proxy for the modularity of the network shows that neighbors of most nodes tend to be highly interconnected among themselves forming a dense and cohesive network of regulatory linkages at this level of regulation. Finally, although incomplete in size, scaling exponent of this network is about 1.8 which suggests that the network might obey a scale-free topology with a power-law degree distribution.

5.2.5 Expression dynamics of RBPs in post-transcriptional networks

5.2.5.1 RBPs show high abundance and tight regulation at the protein level

To compare and understand the differences in the gene expression dynamics of RBPs with other protein coding genes in *S. cerevisiae,* we first compiled the set of RBPs and non-RBPs as described in Materials and Methods (also see Figure 5-1B). This allowed us to define a set of 561 proteins in yeast as those that encode for RNA-binding proteins and the remaining 5685 proteins (from the complete set of protein coding genes) as non-RNA-binding proteins. We also collected high-throughput data documenting various dynamic properties of messenger RNA transcripts and their translated protein products in yeast from different sources as described in Materials and Methods. These properties included the mRNA stability, mRNA copy number, ribosome occupancy, protein stability and abundance. In addition to these attributes of mRNAs and proteins, we also obtained the data describing the cell-to-cell variation in protein expression in a genetically homogenous population of cells, typically referred to as protein expression noise.

(Space left for an enhanced layout of the figure)

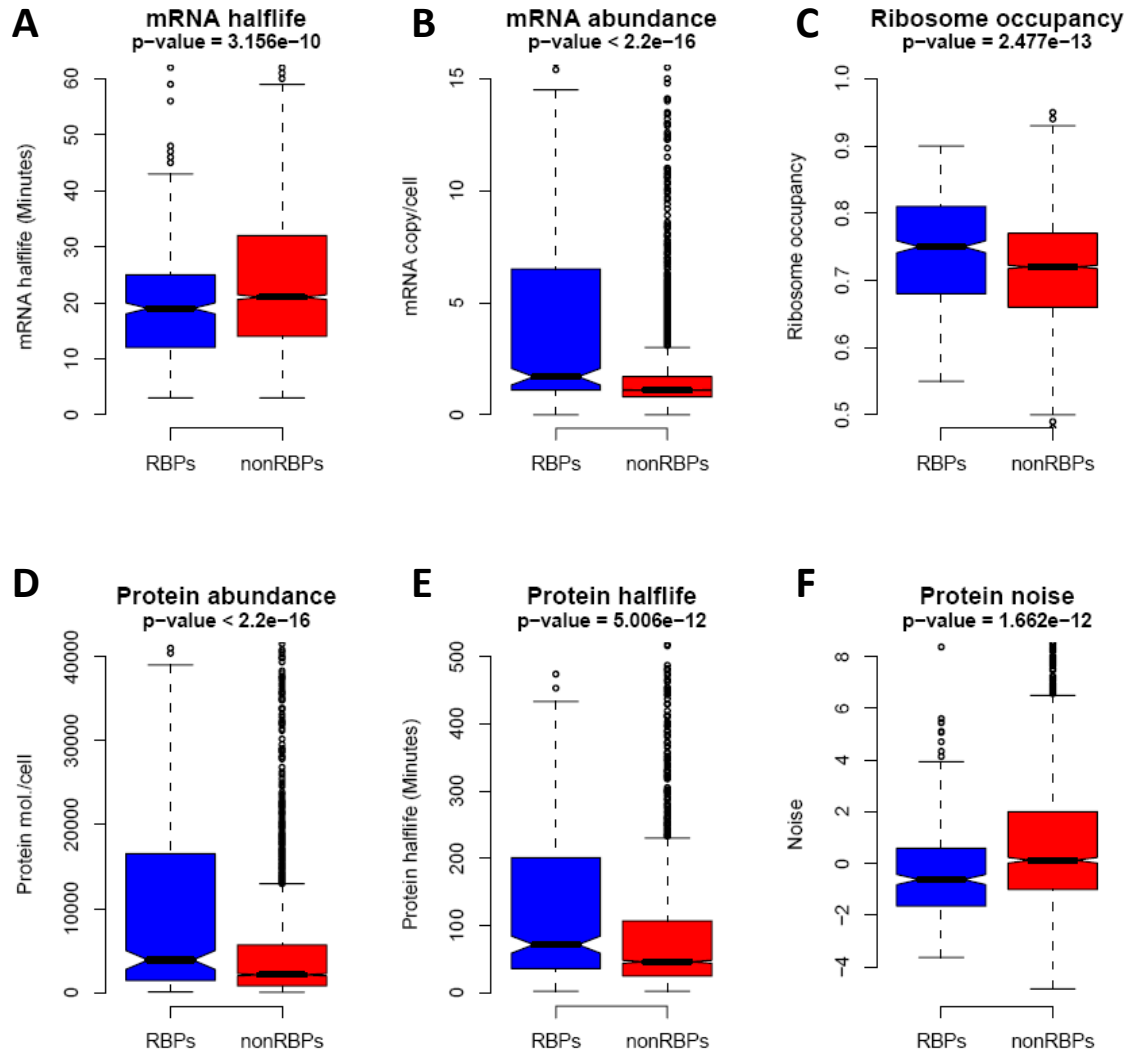


Figure 5-3: Comparing expression dynamics of RBPs with non-RBPs in the entire genome. Box-plots showing the distribution of values for various regulatory properties for the two different groups of proteins (RBPs and non-RBPs) in *S. cerevisiae*. Blue and red bars correspond to RBP and non-RBP populations respectively. Box-plot identifies the middle 50% of the data, the median, and the extreme points. The entire set of data points is divided into quartiles and the inter-quartile range (IQR) is calculated as the difference between x0.75 and x0.25. The range of the 25% of the data points above (x0.75) and below (x0.25) the median (x0.50) is displayed as a filled box. The horizontal line and the notch represent the median and confidence intervals, respectively. Data points greater or less than 1.5 IQR represent outliers and are shown as dots. The horizontal line that is connected by dashed lines above and below the filled box (whiskers) represent the largest and smallest non-outlier data points, respectively. (A) mRNA half-life (B) mRNA copy number (C) Ribosome occupancy (D) Protein abundance (E) Protein half-life (F) Protein noise. In each case, P-values shown correspond to the significance estimated based on Wilcoxon test comparing the RBP and non-RBP group of proteins. RBPs were found to show significantly lower transcript stability, higher mRNA copy number, ribosome occupancy, protein stability and abundance. However protein noise which reflects the extent of cell-to-cell variation in protein levels, was found to be significantly lower for RBPs compared to non-RBPs suggesting that most RBPs are uniformly expressed across a homogenous population of cells.

Messenger RNA half-life is a measure of transcript stability in the cell, while mRNA copy number reflects its abundance. We first asked whether RBPs as a functional class show a different tendency in comparison to non-RBPs in these properties. As a result of this analysis, we found that mRNAs encoding RBPs are significantly less stable (*i.e.*, short half-life) at the transcript level compared to those genes that do not encoded RBPs ($p = 3.1 \times 10^{-10}$, Wilcoxon test) (Figure 5-3A). In yeast it has been shown that, in general, mRNAs of central physiological pathways have longer half-life and mRNAs encoding regulatory and signaling proteins have shorter half-life (Pombo et al., 1999). In line with these observations, the observed lower half-life of RBPs in our analysis is consistent with their regulatory function and quick turn over at transcript level. However, a comparison of the mRNA copy number of the two groups of genes, which is a proxy for mRNA abundance in the cell, indicated that RBPs are encoded by genes which exhibit much higher mRNA copy number ($p < 2.2 \times 10^{-16}$, Wilcoxon test) (Figure 5-3B). Exclusion of translation and ribosome associated genes which form a significant fraction of the total repertoire of RBPs and are known to be highly expressed, did not change our results. These observations suggest that RBPs tend to be less stable but more abundant at transcript level suggesting that abundance is a more prominent factor than their stability. Both mRNA half-life and mRNA abundance data indicate that RBP's expression at mRNA level is likely to be transient but whenever they are transcribed they are produced at high concentrations.

Ribosome occupancy has been shown to be a measure of translational efficiency of mRNA. Higher ribosome occupancy relates to higher protein synthesis and lower ribosome occupancy indicates low translation rate of mRNA. We next asked whether the ribosome occupancy i.e, rate of translation, of RBPs is higher than those for non-RBPs and if their protein levels are higher within the cell. This analysis clearly revealed that RBPs have high ribosome occupancy ($p = 2.5 \times 10^{-13}$, Wilcoxon test) (Figure 5-3C) and are also present in much higher concentrations ($p < 2.2 \times 10^{-16}$, Wilcoxon test) (Figure 5-3D) with median abundances of RBPs being roughly double that observed for non-RBPs (3895 *versus* 2132 protein molecules/cell). These results indicate that RBPs are abundant and are translated rapidly, supporting the versatile nature of their involvement in multiple post-transcriptional control mechanisms at different cellular locations. Exclusion of ribosome and translation associated factors from RBPs to compare non-ribosomal RBPs against non-RBPs indicated that ribosomal RBPs contribute significantly to the observed differences in the rate of translation and protein abundance of RBPs. Comparing the protein concentrations of non-ribosomal RBPs with non-RBPs indicated that the former are still significantly more abundant ($p = 2.2 \times 10^{-2}$).

Stability of a protein measured as its half-life can be considered as a proxy for the life time of a protein in a cell. Therefore, to understand the degradation rates of RBPs and to compare them against non-RBPs we analyzed their protein half-lives (see Materials and Methods). This analysis revealed that RBPs are significantly more stable than non-RBPs, with RBPs exhibiting a median half-life of 71 min as against non-RBPs with 46 min ($p = 5 \times 10^{-12}$, Wilcoxon test) (Figure 5-3E). Repeating the analyses with non-ribosomal RBPs showed a consistent trend despite their exclusion ($p = 4.8 \times 10^{-2}$). Our observations on the increased protein stability and concentration of the RBPs compared to other proteins in the cell suggests that RBPs, whose main functional role is in the processing and localization of their mRNA targets, might be required at multiple sub-cellular locations and be used throughout the cell cycle. This may likely warrant their higher abundance and stability at the protein level. It is important to note that although RBPs exhibit high protein stability, they also show low transcript stability which indicates that most RBPs which are stable at the protein level, might be avoiding cellular crowding of their transcripts by quick turnover at the transcript level. Indeed, it has been shown in yeast that most RBPs auto-regulate their own activity at the transcript level (Hogan et al., 2008).

In order to understand how these properties vary with different processes in which RBPs are involved, we divided RBPs in to four major categories: translation, transport, RNA localization and processing using GO annotations and compared them with non-RBPs. This analysis revealed that the general trends observed for different categories are similar to those seen for RBPs as a whole although certain categories comprised of relatively few RBPs.

Several RBPs have been shown to be post-translationally modified, which adds a layer of flexibility to their function. Many of these post-translational modifications have been shown to modify their RNA-binding properties or their sub-cellular localization. Indeed, at least four types of post-translational modifications namely phosphorylation, ubiquitination, methylation and SUMOylation have been reported for RBPs (Glisovic et al., 2008). High stability of RBPs indicates the potential that post-translational modifications can offer in the diversification of their function. Infact, analysis of the number of kinase substrates in RBP and non-RBP populations using the currently available protein phosphorylation map for yeast (Ptacek et al., 2005), suggests that some kinases not only target higher number of RBPs compared to non-RBPs ($p = 2.7 \times 10^{-2}$) but also more number of kinases are associated with RBPs ($p < 2.2 \times 10^{-16}$).

Gene expression is a highly dynamic process and because of its dynamic nature there is a large variation in a protein's abundance among different cells in a population. This variation is termed as biological noise. Genes whose expression varies to a large extent show more noise

and these are typically involved in stress response, amino acid biosynthesis and heat shock. On the other hand, genes which show consistent expression during the cell cycle such as those involved in protein degradation and ribosomal proteins tend to show low noise (Newman et al., 2006). Here, we have explored this noise data, to address whether RBPs show significant difference from non-RBPs in terms of biological noise. As shown in Figure 5-3F, RBPs were found to show significantly lower noise levels in comparison to non-RBPs ($p = 1.7 \times 10^{-12}$, Wilcoxon test). Re-analyzing the data by excluding ribosomal proteins still clearly indicated that RBPs exhibit much lower noise compared to other protein coding genes ($p = 6.3 \times 10^{-6}$, Wilcoxon test). This analysis unambiguously reveals that low noise is an inherent property of all RBPs and suggests that RBPs are tightly regulated at the protein level with little variation in their expression from cell to cell.

5.2.5.2 The number of distinct targets bound by a RBP is correlated with its cellular abundance

RBPs are the key elements responsible for the post-transcriptional control of gene expression and when combined with their RNA targets, this information can be represented as a RBP-RNA network. Although, on a genomic scale, RBPs are believed to control diverse range of functions with some eukaryotic systems predominantly using post-transcriptional mechanisms for gene expression control (Foth et al., 2008; Noe et al., 2008), large-scale elucidation of post-transcriptional networks is limited to few model organisms for a select set of RBPs. In yeast, few recent genome-wide studies identified the targets for several RBPs using RIP-chip technology (Gerber et al., 2004; Hogan et al., 2008). These studies revealed the important roles played by different families of RBPs and the structure of the post-transcriptional network formed by them. These high-throughput studies showed that the number of targets of a RBP can vary widely, from fewer than ten to more than thousands. In this study we obtained this network discussed above, where nodes represent RBPs or their targets and links represent a distinct physical association between the RBP and the target RNA. We then systematically investigated the relationship between different dynamic properties of RBPs and the number of distinct RNA targets they control.

(Space left for an enhanced layout of the figure)

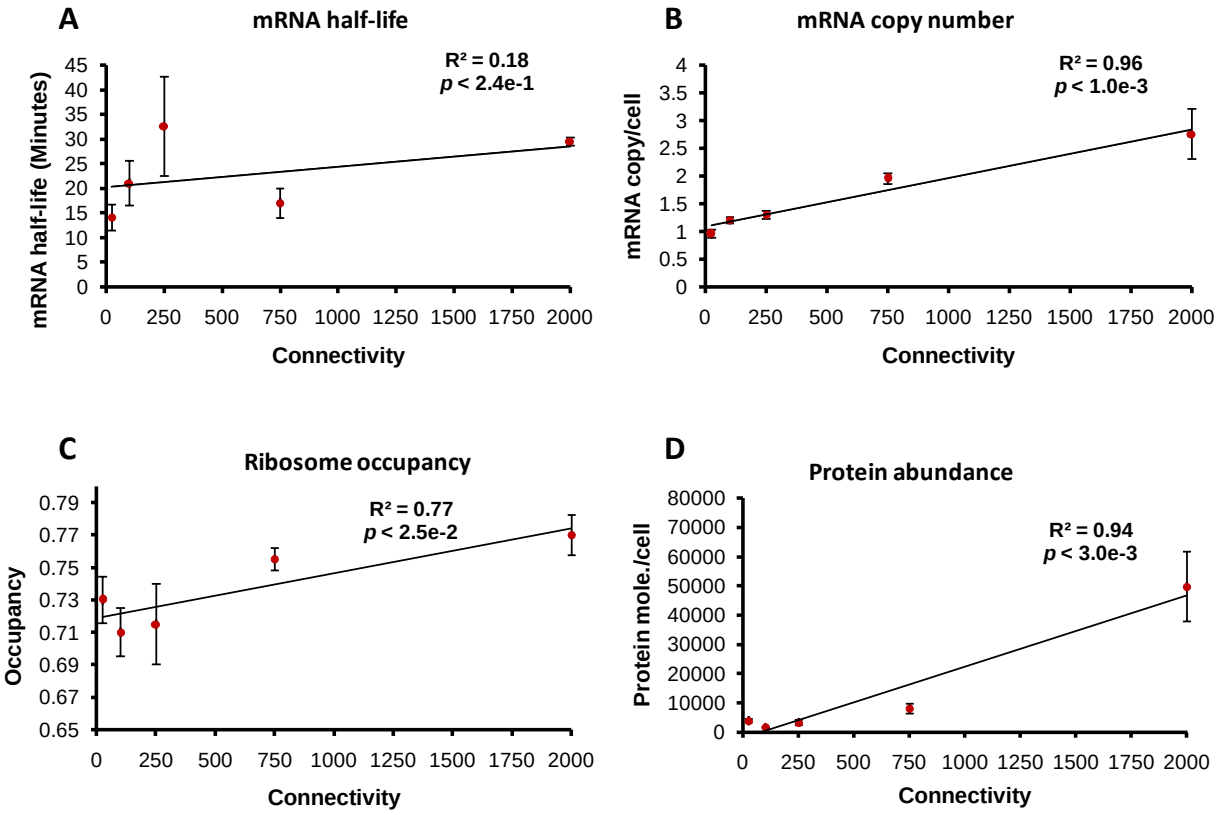


Figure 5-4: Relationship between the number of targets of a RBP and it's A) transcript turn over B) estimated mRNA copy number per cell C) extent of ribosome occupancy and D) protein abundance. In each case, except for transcript stability, we found a strong correlation between the connectivity of a RBP and the regulatory property studied, suggesting that RBPs which regulate high number of targets are present at higher levels at the protein level. RBPs are divided into 5 bins, with approximately equal number of RBPs, based on their connectivity. Points correspond to the median values in the respective bins while the error bars show the normalized median deviation calculated as the ratio between the Median Absolute Deviation (MAD) and the square-root of the number of values in the bin.

We first asked whether the number of targets of a RBP is correlated with its transcript stability by grouping the RBPs into different connectivity bins i.e., groups of RBPs comprising of number of distinct RNA targets (see Methods). As a result of this analysis, we found that there was a weak but positive correlation between them suggesting that transcript turnover of RBPs may not be dependent on their number of targets (R^2=0.18, p < 0.24) (Figure 5-4A). On the other hand, a comparison of the mRNA copy number of a RBP and its number of targets revealed a strong positive correlation between them suggesting that RBPs with high number of targets are likely to be more highly expressed at the mRNA level (R^2=0.96, p < 1×10^{-3}) (Figure 5-4B). For instance, PAB1 is a highly connected essential RBP which can bind to the poly (A) tail of an mRNA to regulate its translational initiation through its binding with eIF4G protein (Kessler and Sachs, 1998; Sachs et al., 1987). Indeed, it was reported to bind to 1,994 distinct

RNA targets and was among the genes with very high mRNA copy number (7.1 mRNA copies/cell). These observations point to a direct link between the number of distinct targets of a RBP and its available number of copies of mRNA in the cell. To test the existence of a correlation between the connectivity and the rate of translation or the absolute protein abundance profile of RBPs, we further explored the relationship between them (Figure 5-4C and 5-4D). This comparison uncovered a more general link between translational efficiency of a RBP and its degree. For instance, Pub1p is another poly (A) binding protein (Matunis et al., 1993) which binds to diverse sets of transcripts involved in ribosome biogenesis, cellular metabolism and transport (Duttagupta et al., 2005). This protein was reported to be localized to both nucleus and cytoplasm (Anderson et al., 1993). Hence to be present at different locations and to bind to a large number of transcripts it has to be translated more often and should be present in more number of copies. Consistent with this, we find that it's transcript exhibits high ribosome occupancy. Indeed, Hogan et. al (Hogan et al., 2008) demonstrated that RNA targets of highly connected RBPs were enriched for multiple processes and sub-cellular localizations. These results clearly unveil the strong relationship between the concentration of a RBP and the number of distinct RNA targets bound by them, indicating that RBPs responsible for controlling a wide range of targets must occur in more number of copies at the protein level. It is important to note that although RBPs as a group of genes are significantly higher expressed at the transcript and protein levels compared to non-RBP population, relative abundance of the RBPs is correlated to the hierarchy of a RBP, defined as the number of distinct RNA targets. It is also noteworthy to mention that the RBPs analyzed for connectivity in this section did not comprise of core ribosomal proteins, strengthening the generality of these observations.

5.2.5.3 RBPs bound to many RNA targets are less frequently degraded and tightly controlled at protein level

Although RBPs with more number of distinct targets are expressed at a higher level compared to those which control fewer targets, it is not evident if their protein turnover rates would hold a similar trend. Therefore, to understand whether there is any dependence between the stability of a RBP and the number of transcripts it controls, we employed a similar approach as above. This analysis clearly showed that RBPs which regulate many targets are highly stable at the protein level (R^2=0.95, $p < 3 \times 10^{-3}$) (Figure 5-5A). The link between protein stability and RBP's degree indicates that RBPs controlling several targets are less frequently degraded at the protein level and might be present throughout the cell cycle. Taken together, these observations raise the question: If highly connected RBPs are consistently expressed in large concentrations and are

less frequently degraded, would their regulation be tightly controlled at the protein level. The fact that RBPs as a group show significantly lower noise in comparison to non-RBPs and that previous studies reported that regulatory proteins generally exhibit low noise (Newman et al., 2006) suggests that highly connected RBPs can be expected to show less noise in comparison to those which are poorly connected. Hence, we compared the connectivity of RBPs with their noise value. As shown in Figure 5-5B, we found a strong correlation between the number of targets of a RBP and its protein noise. In particular, highly connected RBPs showed minimal variation in their protein expression across a population of cells (R^2=0.93, $p < 4 \times 10^{-3}$). This suggests that RBPs controlling many targets are very tightly regulated with little cell-to-cell variation in their protein expression. These observations indicate that any significant change in their availability or regulation may result in an imbalance in cellular homeostasis as it may affect a vast number of transcripts. Indeed, a comparison of the number of essential genes in RBPs showed a two-fold enrichment compared to the whole genome, suggesting their central role in maintaining cellular homeostasis. These lines of evidence reveal that RBPs act as an important class of regulatory molecules in the cell whose expression is tightly controlled despite their occurrence in large cellular concentrations and in multiple sub-cellular locations.

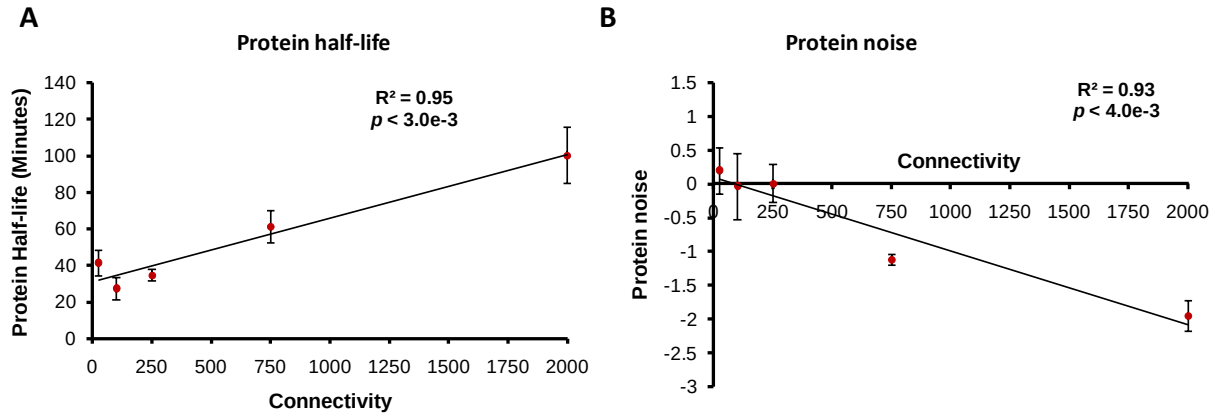


Figure 5-5: Relationship between RBP's connectivity versus it's A) protein stability and B) noise. RBPs controlling more number of targets showed an increasing tendency to be stable at the protein level and deceasing tendency in protein noise. RBPs are divided into 5 bins, with approximately equal number of RBPs, based on their connectivity. Points correspond to the median values while the error bars show the normalized median deviation calculated as the ratio between the Median Absolute Deviation (MAD) and the square-root of the number of values in the bin.

5.3 DISCUSSION & CONCLUSION

RBPs form an important class of evolutionarily conserved proteins (Anantharaman et al., 2002) and are known to be involved in a wide range of cellular processes. In addition to their functional roles in diverse processes as shown in Figure 5-1A, RBPs are also known to be implicated in a number of disorders due to their mis-expression or mutations in the sequences that are employed to recognize their cognate target RNAs. For instance, in humans, malfunctioning of RBPs like NOVA, which is a neuron specific protein responsible for the alternative splicing of a subset of pre-mRNAs, is known to be involved in the pathogenesis of the neurodegenerative syndrome Paraneoplastic Opsoclonus-Myoclonus Ataxia (POMA) (Ule et al., 2003). In line with this and other observations on the impact of changes in the expression levels of RBPs being associated with diseases and fitness defects (Cooper et al., 2009; Lukong et al., 2008) results reported here reveal that RBPs as a functional class show very little variation in their expression across cells suggesting the importance in tightly controlling them. In addition, it was found that RBPs which regulate multiple transcripts show a significantly reduced noise indicating that variations in the expression levels of these key post-transcriptional regulators can have significant impact on the functioning of the cell thereby leading to a disease phenotype.

The fact that RBPs are generally less stable at the transcript level but exhibit higher stability and abundance at the protein level demonstrates that they form a group of proteins which follow the theoretically proposed time averaging effect on noise propagation (Paulsson, 2004), which suggests that if the protein has long half life compared to its mRNA then it averages over the noisy fluctuations in the mRNA decreasing the protein expression noise. These results also indicate that regulation of RBPs is predominantly controlled at the protein level through the use a number of post-translational modifications (PTMs) like phosphorylation, arginine methylation and sumoylation which have been reported to occur in several well-studied RBPs (Schullery et al., 1999; Vassileva and Matunis, 2004; Yu et al., 2004). Indeed, a comparison of the number of phosphorylated targets in RBPs and non-RBPs revealed the predominance of post-translational control in RBPs. Therefore, it is possible to suggest that a wide variety of these PTMs might be responsible for their ability to spatially and temporally regulate transcripts in eukaryotic systems. It is possible to speculate from these observations that the low noise levels of RBPs together with extensive regulatory flexibility at the protein level might give them an advantage to control gene regulation at a finer level compared to transcriptional control by transcription factors. This might thereby provide a quick and extensive framework for controlling gene expression of a wide range of genes. This is also supported

based on the observation that RBPs which are central to the cell are not only required in large quantities but are also found to be present for a longer time in the cell. All these observations suggest the importance of a post-transcriptional network of interactions in higher eukaryotes and raise several open questions in the regulation of gene expression beyond transcription. It should be possible to address such questions in the near future as more data from different levels of regulation becomes available (Halbeisen et al., 2008; Hieronymus and Silver, 2004; Lackner et al., 2007).

While the post-genomic era has introduced the genomic complement of hundreds of genomes, it has also left us with several unanswered questions regarding the functional relevance of the genes an organism encodes or principles that govern the regulation of the genes encoded on them. It is noteworthy to mention that even in a model organism like *S. cerevisiae*, regulation of gene expression at the post-transcriptional level is rather poorly understood. Nevertheless with recent improvements in and availability of high-throughput approaches such as RNA-sequencing and immunoprecipitation protocols, future years can expect to see a wealth of data detailing the dynamic, spatial and tissue-specific nature of the interactions governed by these exciting class of regulatory molecules, which would undoubtedly allow us to gain a deeper understanding of regulation at a level which has been under-appreciated over the past decades. Given the unprecedented detail at which these high-throughput technologies can reveal the link between the regulatory elements on the target genes and the RNA-binding proteins specific to environmental conditions, it is possible to use these approaches to interrogate the prevalence of these phenomena in different states and thereby study their relevance to physiology and disease in diverse model systems.

5.4 MATERIALS AND METHODS

5.4.1 Data on RNA-binding proteins in *S. cerevisiae* and their interactions

The complete list of annotated RBPs and the data for well studied RBPs in *S. cerevisiae* was obtained from Hogan et al (Hogan et al., 2008). The total number of annotated RBPs in yeast reported in this study was 561 and mRNA targets for 41 RBPs have been systematically identified on a whole genome scale by employing the RIP-chip technology. This approach essentially consists of two steps. The first involves generation of two RNA samples, isolation of RBP bound mRNA by immunoprecipitation of messenger-ribonucleoproteins using affinity purification and isolation of cellular RNA representing the whole set of transcripts in the cell. The

second step involves hybridization of the two isolated RNA samples using dual-color microarrays and are analyzed for enriched transcripts, to detect the bound targets of a RBP (Sanchez et al., 2007). A total of 14,312 interactions comprising of 41 RBPs and 5025 genes in the entire genome of *S. cerevisiae*, which forms a network of post-transcriptional interactions between RBPs and the target RNAs encoding for proteins obtained using this approach was used for studying the expression dynamics, while the network properties have been studied using the entire network of 19396 interactions reported in the original study (Hogan et al., 2008).

5.4.2 Analysis of the structure and properties of post-transcriptional regulatory network

We used igraph, a publicly available R package [see http://cneurocvs.rmki.kfki.hu/igraph/ and http://www.r-project.org] to study the properties of this network and to calculate the centrality of the nodes in this framework. In particular, since the network analyzed in this study was considered as undirected for the sake of simplicity, we used the corresponding versions of the functions: degree, transitivity, betweenness and closeness for calculating the degree, clustering coefficient, betweenness and closeness centralities of a node. Betweenness centrality, which is the number of shortest paths going through a node was calculated using the brandes algorithm (Brandes, 2001) implemented in R. Similarly, closeness, measured as average length of the shortest paths to all the other vertices in the graph, was obtained using the implementation in R. Since the centrality measures, betweenness and closeness use the shortest path lengths between all pairs of nodes in a graph, for cases where no path exists between a particular pair of nodes, shortest path length was taken as one less than the maximum number of nodes in the graph. Note that this is also the default assumption for calculating centrality measures in *igraph*. The Clustering coefficient is a property of a node which tells how connected are the neighbors of a given node to what is expected when all the neighbors are completely connected. An extension of this metric to the complete network defined as the average clustering coefficient tells whether the network is modular or is sparsely connected. Other network properties were calculated using the default implementations in *igraph* or as discussed in the main text.

5.4.3 Data for comparative analysis of expression dynamics

To study the expression dynamics of RBPs in comparison to other protein coding genes in the genome and to analyze its relationship with the number of RNAs controlled by RBPs, we have employed a variety of datasets. These include the transcript stability, mRNA copy number, ribosome occupancy, protein half-life, protein abundance and protein noise. Transcript stability

which is measured as the RNA half-life of a transcript was obtained from Wang et. al (Wang et al., 2002) and contained mRNA half-lives for 4687 genes in the entire genome. A key parameter describing the translational status of a gene is the fraction of its transcripts engaged in translation which is defined by the ribosome occupancy (Arava et al., 2003). Likewise, the number of mRNA copies of a gene can be best described by the parameter mRNA copy number per cell. Both these parameters for genes in *S. cerevisiae* were obtained from Arava et. al (Arava et al., 2003) where the authors employed velocity sedimentation to separate mRNAs bound to ribosomes and quantified them using microarray analysis. mRNA copy number could be obtained for 5643 genes while ribosome occupancy could be mapped for 5700 genes, allowing us to study the extent of transcript abundance and translation rates of the genes and transcripts. Stability of a protein which is an estimate of the duration it occurs with in the cell is measured as the half-life of the protein. In yeast, protein half-lives have been estimated by Belle and co-workers for about 3750 proteins by inhibiting translation (Belle et al., 2006). In this study we used this data by excluding proteins whose half-lives have been obtained by extrapolation. Protein abundance which reveals the absolute number of protein molecules per cell was obtained from Ghaemmaghami et. al (Ghaemmaghami et al., 2003). We could obtain abundance values for 3868 proteins in the entire genome. Biological noise which is typically defined as the variation in the expression of a protein between different cells in a homogenous population of cells was obtained from Newman et. al (Newman et al., 2006). We could obtain noise data for 2213 genes for cells grown on rich media. The authors in this study employed two distinct measures for calculating protein noise, coefficient of variation (CV), which is the ratio of the standard deviation in the expression of a protein and it's mean expression and distance from median (DM), which was calculated as the difference between the CV value of a protein and a running median of all CV values. In this study we have used DM as a measure of protein noise as it was indicated to be a more robust measure compared to CV to understand protein to protein variations in noise levels (Newman et al., 2006). Since DM is the distance between the CV and median value of all CVs, negative values correspond to relatively less noise while positive values reflect higher levels of noise in the protein expression.

5.4.4 Comparison of the regulatory properties of RBPs with other protein coding genes

To study whether RBPs show differences in dynamic properties when compared to other protein coding genes, we defined non-RBP set of proteins. This set essentially comprised of proteins in the whole genome after excluding the list of 561 RBPs defined above. To assess whether RBPs

exhibit a different trend compared to non-RBPs for each of the properties studied, we used Wilcoxon rank-sum test or Mann-Whitney U test available in the R statistical package to calculate the significance. Wilcoxon test enables the comparison of two samples to assess whether they come from the same distribution or not. Since this test is non-parametric and does not assume any inherent distribution of the samples it is ideal to compare different samples. Box plots were used to represent the distribution of values for each property. Since the RBP set comprised of a number of ribosome associated proteins we also excluded them from this list and repeated the analysis to test the robustness of the tendencies observed, in the absence of ribosomal proteins.

5.4.5 Analysis of the relationship between the number of targets of a RBP and its dynamic properties

To understand the link between the number of targets of a RBP and its dynamic properties, RBPs were first grouped on the basis of their number of distinct RNA targets to which they were bound. This grouping was done in such a way that each bin of RBPs contained roughly equal number of RBPs. This resulted in five different bins corresponding to varying degrees of RBPs, with some RBPs controlling as many as 2000 mRNAs in the RBP-RNA network. To nullify the effect of outliers in each bin, median values were calculated for different dynamic properties and correlation was estimated between median values and connectivity of RBPs. P-values were calculated using the coefficient of correlation and the number of data points, based on a linear fit.

REFERENCES

Alberti, S., Halfmann, R., King, O., Kapila, A. and Lindquist, S. (2009). A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* **137**, 146-58.

Anantharaman, V., Koonin, E. V. and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* **30**, 1427-64.

Anderson, J. T., Paddy, M. R. and Swanson, M. S. (1993). PUB1 is a major nuclear and cytoplasmic polyadenylated RNA-binding protein in Saccharomyces cerevisiae. *Mol Cell Biol* **13**, 6102-13.

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O. and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **100**, 3889-94.

Belle, A., Tanay, A., Bitincka, L., Shamir, R. and O'Shea, E. K. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A* **103**, 13004-9.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* **25**, 163-177.

Butter, F., Scheibe, M., Morl, M. and Mann, M. (2009). Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A* **106**, 10626-31.

Campbell, D. A., Thomas, S. and Sturm, N. R. (2003). Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect* **5**, 1231-40.

Ciesla, J. (2006). Metabolic enzymes that bind RNA: yet another level of cellular regulatory network? *Acta Biochim Pol* **53**, 11-32.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184-94.

Cooper, T. A., Wan, L. and Dreyfuss, G. (2009). RNA and disease. *Cell* **136**, 777-93.

Duttagupta, R., Tian, B., Wilusz, C. J., Khounh, D. T., Soteropoulos, P., Ouyang, M., Dougherty, J. P. and Peltz, S. W. (2005). Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. *Mol Cell Biol* **25**, 5499-513.

Fasolo, J. and Snyder, M. (2009). Protein microarrays. *Methods Mol Biol* **548**, 209-22.

Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nat Rev Cancer* **4**, 143-53.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K. et al. The Pfam protein families database. *Nucleic Acids Res* **38**, D211-22.

Foth, B. J., Zhang, N., Mok, S., Preiser, P. R. and Bozdech, Z. (2008). Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. *Genome Biol* **9**, R177.

Galante, P. A., Sandhu, D., de Sousa Abreu, R., Gradassi, M., Slager, N., Vogel, C., de Souza, S. J. and Penalva, L. O. (2009). A comprehensive in silico expression analysis of RNA binding proteins in normal and tumor tissue: Identification of potential players in tumor formation. *RNA Biol* **6**, 426-33.

Gerber, A. P., Herschlag, D. and Brown, P. O. (2004). Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* **2**, E79.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. and Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature* **425**, 737-41.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-86.

Gross, T., Richert, K., Mierke, C., Lutzelberger, M. and Kaufer, N. F. (1998). Identification and characterization of srp1, a gene of fission yeast encoding a RNA binding domain and a RS domain typical of SR splicing factors. *Nucleic Acids Res* **26**, 505-11.

Gruter, P., Tabernero, C., von Kobbe, C., Schmitt, C., Saavedra, C., Bachi, A., Wilm, M., Felber, B. K. and Izaurralde, E. (1998). TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Mol Cell* **1**, 649-59.

Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720-30.

Halbeisen, R. E., Galgano, A., Scherrer, T. and Gerber, A. P. (2008). Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell Mol Life Sci* **65**, 798-813.

Hall, D. A., Zhu, H., Zhu, X., Royce, T., Gerstein, M. and Snyder, M. (2004). Regulation of gene expression by a metabolic enzyme. *Science* **306**, 482-4.

Hieronymus, H. and Silver, P. A. (2003). Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat Genet* **33**, 155-61.

Hieronymus, H. and Silver, P. A. (2004). A systems view of mRNP biology. *Genes Dev* **18**, 2845-60.

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. and Brown, P. O. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6**, e255.

Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H. S., Woodard, C., Wang, H., Jeong, J. S. et al. (2009). Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**, 610-22.

Junker, B. H., Koschutzki, D. and Schreiber, F. (2006). Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **7**, 219.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-43.

Keene, J. D. and Lager, P. J. (2005). Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res* **13**, 327-37.

Keene, J. D. and Tenenbaum, S. A. (2002). Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell* **9**, 1161-7.

Kessler, S. H. and Sachs, A. B. (1998). RNA recognition motif 2 of yeast Pab1p is required for its functional interaction with eukaryotic translation initiation factor 4G. *Mol Cell Biol* **18**, 51-7.

Kim, M. Y., Hur, J. and Jeong, S. (2009). Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB Rep* **42**, 125-30.

Lackner, D. H., Beilharz, T. H., Marguerat, S., Mata, J., Watt, S., Schubert, F., Preiss, T. and Bahler, J. (2007). A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol Cell* **26**, 145-55.

Lasko, P. (2000). The drosophila melanogaster genome: translation factors and RNA binding proteins. *J Cell Biol* **150**, F51-6.

Lee, M. H. and Schedl, T. (2006). RNA-binding proteins. *WormBook*, 1-13.

Lublin, A. L. and Evans, T. C. (2007). The RNA-binding proteins PUF-5, PUF-6, and PUF-7 reveal multiple systems for maternal mRNA regulation during C. elegans oogenesis. *Dev Biol* **303**, 635-49.

Lukong, K. E., Chang, K. W., Khandjian, E. W. and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet* **24**, 416-25.

Maris, C., Dominguez, C. and Allain, F. H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **272**, 2118-31.

Mata, J., Marguerat, S. and Bahler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* **30**, 506-14.

Matunis, M. J., Matunis, E. L. and Dreyfuss, G. (1993). PUB1: a major yeast poly(A)+ RNA-binding protein. *Mol Cell Biol* **13**, 6114-23.

McKee, A. E., Minet, E., Stern, C., Riahi, S., Stiles, C. D. and Silver, P. A. (2005). A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev Biol* **5**, 14.

Musunuru, K. (2003). Cell-specific RNA-binding proteins in human disease. *Trends Cardiovasc Med* **13**, 188-95.

Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. and Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840-6.

Nica, A. C. and Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet* **17**, R129-34.

Noe, G., De Gaudenzi, J. G. and Frasch, A. C. (2008). Functionally related transcripts have common RNA motifs for specific RNA-binding proteins in trypanosomes. *BMC Mol Biol* **9**, 107.

Paraskeva, E., Atzberger, A. and Hentze, M. W. (1998). A translational repression assay procedure (TRAP) for RNA-protein interactions in vivo. *Proc Natl Acad Sci U S A* **95**, 951-6.

Pascale, A., Amadio, M. and Quattrone, A. (2008). Defining a neuron: neuronal ELAV proteins. *Cell Mol Life Sci* **65**, 128-40.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature* **427**, 415-8.

Pinero, D. J., Hu, J. and Connor, J. R. (2000). Alterations in the interaction between iron regulatory proteins and their iron responsive element in normal and Alzheimer's diseased brains. *Cell Mol Biol (Noisy-le-grand)* **46**, 761-76.

Pombo, A., Jackson, D. A., Hollinshead, M., Wang, Z., Roeder, R. G. and Cook, P. R. (1999). Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *Embo J* **18**, 2241-53.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R. et al. (2005). Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679-84.

Pullmann, R., Jr., Kim, H. H., Abdelmohsen, K., Lal, A., Martindale, J. L., Yang, X. and Gorospe, M. (2007). Analysis of turnover and translation regulatory RNA-binding protein expression through binding to cognate mRNAs. *Mol Cell Biol* **27**, 6265-78.

Ray, D., Kazan, H., Chan, E. T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B. J., Morris, Q. and Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27**, 667-70.

Rodgers, N. D., Jiao, X. and Kiledjian, M. (2002). Identifying mRNAs bound by RNA-binding proteins using affinity purification and differential display. *Methods* **26**, 115-22.

Sachs, A. B., Davis, R. W. and Kornberg, R. D. (1987). A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Mol Cell Biol* **7**, 3268-76.

Saint-Georges, Y., Garcia, M., Delaveau, T., Jourdren, L., Le Crom, S., Lemoine, S., Tanty, V., Devaux, F. and Jacq, C. (2008). Yeast mitochondrial biogenesis: a role for the PUF RNA-binding protein Puf3p in mRNA localization. *PLoS ONE* **3**, e2293.

Sanchez-Diaz, P. and Penalva, L. O. (2006). Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. *RNA Biol* **3**, 101-9.

Sanchez, M., Galy, B., Hentze, M. W. and Muckenthaler, M. U. (2007). Identification of target mRNAs of regulatory RNA-binding proteins using mRNP immunopurification and microarrays. *Nat Protoc* **2**, 2033-42.

Sanford, J. R., Gray, N. K., Beckmann, K. and Caceres, J. F. (2004). A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* **18**, 755-68.

Schullery, D. S., Ostrowski, J., Denisenko, O. N., Stempka, L., Shnyreva, M., Suzuki, H., Gschwendt, M. and Bomsztyk, K. (1999). Regulated interaction of protein kinase Cdelta with the heterogeneous nuclear ribonucleoprotein K protein. *J Biol Chem* **274**, 15101-9.

SenGupta, D. J., Zhang, B., Kraemer, B., Pochart, P., Fields, S. and Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. *Proc Natl Acad Sci U S A* **93**, 8496-501.

Stelzl, U. and Nierhaus, K. H. (2001). SERF: in vitro election of random RNA fragments to identify protein binding sites within large RNAs. *Methods* **25**, 351-7.

Tenenbaum, S. A., Carson, C. C., Lager, P. J. and Keene, J. D. (2000). Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A* **97**, 14085-90.

Thomson, A. M., Rogers, J. T., Walker, C. E., Staton, J. M. and Leedman, P. J. (1999). Optimized RNA gel-shift and UV cross-linking assays for characterization of cytoplasmic RNA-protein interactions. *Biotechniques* **27**, 1032-9, 1042.

Townley-Tilson, W. H., Pendergrass, S. A., Marzluff, W. F. and Whitfield, M. L. (2006). Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. *RNA* **12**, 1853-67.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A. and Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212-5.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J. and Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580-6.

Vassileva, M. T. and Matunis, M. J. (2004). SUMO modification of heterogeneous nuclear ribonucleoproteins. *Mol Cell Biol* **24**, 3623-32.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. and Brown, P. O. (2002). Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* **99**, 5860-5.

Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E. and Yates, J. R., 3rd. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **100**, 3107-12.

Wilhelm, J. E. and Vale, R. D. (1996). A one-hybrid system for detecting RNA-protein interactions. *Genes Cells* **1**, 317-23.

Yu, M. C., Bachand, F., McBride, A. E., Komili, S., Casolari, J. M. and Silver, P. A. (2004). Arginine methyltransferase affects interactions and recruitment of mRNA processing and export factors. *Genes Dev* **18**, 2024-35.

Zeng, F., Peritz, T., Kannanayakal, T. J., Kilk, K., Eiriksdottir, E., Langel, U. and Eberwine, J. (2006). A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. *Nat Protoc* **1**, 920-7.

Zhong, X. Y., Wang, P., Han, J., Rosenfeld, M. G. and Fu, X. D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell* **35**, 1-10.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T. et al. (2001). Global analysis of protein activities using proteome chips. *Science* **293**, 2101-5.

Zielinski, J., Kilk, K., Peritz, T., Kannanayakal, T., Miyashiro, K. Y., Eiriksdottir, E., Jochems, J., Langel, U. and Eberwine, J. (2006). In vivo identification of ribonucleoprotein-RNA interactions. *Proc Natl Acad Sci U S A* **103**, 1557-62.

6 Conclusions and Perspectives

CONTENTS OF CHAPTER 6

6.1 Outline

An important notion that is emerging in post-genomic biology is that cellular components can be visualized as a network of associations between different molecules like proteins, DNA, RNA and metabolites. This has led to the application of network theory and network-based approaches to a wide range of biological problems from understanding regulation of gene expression to prediction of gene's function and phenotype to drug discovery settings. In Chapter 1, I introduced the notion of networks and the basic principles of network biology together with an overview of different kinds of networks that are being widely studied in biological sciences at the systems level. For instance, while in transcriptional and post-transcriptional networks, typically trans-acting elements like TFs, RBPs and sigma factors form one set of nodes and their target genes or RNAs, of which they control the activity, form the other set of nodes. The links between them which have directionality from the trans-acting elements to their target genes, controlled by their cis-regulatory elements, form a complex and directional network of interactions. In contrast, functional linkage networks constructed in function prediction pipelines typically comprise of undirected networks where all the nodes are treated essentially the same and there is no directionality between nodes. These networks aim to uncover the broad functional role of the uncharacterized genes using the annotations of already characterized members to which they are connected to. I then give a brief overview of small-molecule protein interaction networks which are also referred to as the drug-target networks to extend the generality and applicability of the network-guided approaches in understanding biological systems.

Gene expression is a highly regulated process and is controlled at several levels. In prokaryotes, control of gene expression predominantly occurs at the level of transcription and TFs play important role in this process. In Chapter 2, I address the questions, how and why are genes organized on a particular fashion on bacterial genomes and what are the constraints bacterial transcriptional regulatory networks impose on their genomic organization. I extend this one step further to unravel the constraints imposed on the network of TF-TF interactions and relate it to the numerous phenotypes they can impart to growing bacterial populations.

In contrast to prokaryotes, regulation of gene expression in eukaryotes is much more complex and is known to occur at many different levels even at the stage of transcription. In Chapter 3, I first present an overview of our current understanding of eukaryotic gene regulation at different levels and then present evidence for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation. These

results demonstrate that specific organization of genes across and within chromosomes that allowed for efficient control of transcription within the nuclear space has been selected during evolution.

Determining the functions of proteins encoded by genome sequences represents a major challenge in contemporary biology. With traditional methods for annotation of a genome reaching their saturation there is an increasing need to develop alternate and complementary approaches for solving the genomic function prediction challenge. As a result, alternate computational methods for inferring the protein function such as those which exploit the context of a protein in protein association networks have come to be sought after. These network-based approaches aim to integrate diverse kinds of functional interactions as a means of boosting coverage as well as confidence level of an association. In Chapter 4, I first present an overview of different computational approaches for inferring the function of uncharacterized genes and discuss network-based approaches currently employed for predicting function. I then summarize a recent high-throughput study performed to provide a 'systems-wide' functional blueprint of the bacterial model, *Escherichia coli* K-12, with insights into the biological and evolutionary significance of previously uncharacterized proteins. Given the volume of high-throughput data that is being reported for understanding diverse model systems, the network-based approaches presented here would undoubtedly be a useful addition to unravel the functions of an increasing number of uncharacterized proteins accumulating in the genomic databases.

While control of gene expression in eukaryotes first occurs at the level of transcription, there is accumulating evidence that an often neglected set of factors called RNA-binding proteins play major roles in controlling the expression of a protein by regulating expression at post-transcriptional level. In Chapter 5, I attempt to provide a comprehensive overview and preliminary insights on this rapidly developing area of post-transcriptional regulatory networks formed by RBPs. I discuss the sequence attributes and functional processes associated with RBPs, methods used for the construction of the networks formed by them and finally discuss the structure and dynamics of these post-transcriptional networks based on recent publicly available data. The results obtained from this study show that RBPs exhibit distinct gene expression dynamics compared to other class of proteins in a eukaryotic cell and that these properties are also reflected from an analysis of the post-transcriptional networks formed by them.

In the current chapter, I first summarize the key findings of all the previous chapters and then discuss their broader implications.

6.2 Major Findings

6.2.1 Constraints imposed by transcriptional regulation on genome organization and regulatory network

In Chapter 2, using network-guided approaches for understanding the transcriptional regulatory networks of bacteria, I show that there are at least two kinds of constraints. The first is among the network of transcriptional regulatory interactions between TFs, where in I show that while the mode of regulatory interaction between transcription factors (TFs) is predominantly positive, TFs are frequently negatively auto-regulated. Furthermore, feedback loops, regulatory motifs and regulatory pathways are unevenly distributed in this network with short pathways, multiple feed-forward loops and negative auto-regulatory interactions being abundant in the sub-network controlling metabolic functions such as the use of alternative carbon sources. In contrast, long hierarchical cascades and positive auto-regulatory loops are over-represented in the sub-networks controlling developmental processes for biofilm and chemotaxis. Based on these observations, I propose that these long transcriptional cascades coupled with regulatory switches (positive loops) for sensing external conditions enable the coexistence of multiple bacterial phenotypes in growing bacterial populations (Martinez-Antonio et al., 2008). A second constraint is that of a link between the transcriptional hierarchy of regulons (TFs) and their genome organization. In particular, I show that, to drive the kinetics and concentration gradients, TFs belonging to big and small regulons (classified based on the number of genes they regulate in the transcriptional network) organize themselves differently on the genome with respect to their targets. Using data from independently reported studies in *E. coli,* I demonstrate that higher a TF is in the transcriptional hierarchy more are its detected number of mRNA and protein molecules per cell, reflecting its need to be expressed in higher concentrations to regulate target genes located dispersedly on the chromosome. In contrast to big regulons, local or dedicated TFs (lower in the network hierarchy and regulating much fewer genes) were found to be expressed in much lower concentrations explaining the reasons for their proximity on the chromosome to their target genes (Janga et al., 2009). These observations give insights into how the scale-free structure of transcriptional networks can be encoded on the chromosome to drive the kinetics and concentration gradients of TFs, depending on the number of genes they regulate and could facilitate the horizontal transfer of local environment-specific transcriptional modules. I then propose a conceptual model based on these observations to explain how the hierarchical structure of TRNs might be ultimately governed by the dynamic biophysical

requirements for targeting DNA-binding sites by transcription factors. These results suggest that the main parameters defining the position of a TF in the network hierarchy are the number and chromosomal distances of the genes they regulate and their protein concentration gradients. These observations give insights into how the hierarchical structure of transcriptional networks can be encoded on the chromosome to drive the kinetics and concentration gradients of TFs depending on the number of genes they regulate and could be a common theme valid for other prokaryotes, proposing the role of transcriptional regulation in shaping the organization of genes on a chromosome.

In Chapter 3, extending these ideas to eukaryotic systems, I first describe our current understanding of eukaryotic regulation in all the three dimensions (DNA sequence level, chromatin level and nuclear organizational levels) to reinforce the notion that regulation in higher organisms is much more complex and needs intricate co-ordination of several molecular events in space and time. I then present evidence, analyzing the currently known transcriptional regulatory network of the single-celled model eukaryote, *Saccharomyces cerevisiae,* for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation. In particular, here I reveal that the target genes (TGs) of transcription factors (TFs) for the yeast, *S. cerevisiae*, are encoded in a highly ordered manner both across and within the 16 chromosomes by showing that (i) the TGs of a majority of TFs show a strong preference to be encoded on specific chromosomes, (ii) the TGs of a significant number of TFs display a strong preference (or avoidance) to be encoded in regions containing particular chromosomal landmarks such as telomeres and centromeres, and (iii) the TGs of most TFs are positionally clustered within a chromosome (Janga et al., 2008). These results demonstrate that specific organization of genes that allowed for efficient control of transcription within the nuclear space has been selected during evolution which has lead to the constraints observed at different levels reported in this chapter. Further analysis on human and mouse TFs permitted us to also show that the constraints are more general and are not limited to yeast alone suggesting that uncovering such higher-order organization of genes in other eukaryotes will provide insights into nuclear architecture, and will have implications in genetic engineering experiments, gene therapy, and understanding disease conditions that involve chromosomal aberrations.

6.2.2 Uncovering the functional landscape of a bacterial genome

Determining the functions of proteins encoded by genome sequences represents a major challenge in contemporary biology. As of now, public databases report more than 1000

completely sequenced genomes with over 3700 genome projects underway leading to a situation where we know the location and position of the protein coding genes on the genome but we hardly have a clue on what many of these protein machines do across genomes. Add to this the sequencing of metagenomic samples which currently stand at more than 100 in number, with the venter's marine microbial community's project alone contributing more than 6,000,000 proteins to the already accumulating list of protein repertoire (Venter et al., 2004). All these point out to the slow pace at which we are able to understand the protein repertoire of the organisms at the functional level despite rapid pace at which sequencing technologies are able to generate the genome sequence data.

For instance, yet despite being the most highly studied model bacterium, a recent comprehensive community annotation effort for the fully sequenced reference K-12 laboratory strains (Riley et al., 2006) indicated that only half (~54%) of the protein-coding gene products of *E. coli* currently have experimental evidence indicative of a biological role. The remaining genes have either only generic, homology-derived functional attributes (e.g. 'predicted DNA-binding') or no discernable physiological significance. In Chapter 4, I discuss a recent study where we attempted to characterize one-third of the 4,225 protein-coding genes of *Escherichia coli* K-12 which remain functionally unannotated (functional orphans) (Hu et al., 2009). In particular, to elucidate their biological roles, we performed an extensive proteomic survey using affinity-tagged *E. coli* strains and generated comprehensive genomic context inferences to derive a high-confidence compendium for virtually the entire proteome consisting of 5,993 putative physical interactions and 74,776 putative functional associations, most of which were novel. We then clustered the respective probabilistic networks to reveal putative orphan membership into discrete multiprotein complexes and functional modules, while a machine-learning strategy based on network integration methods implicated the orphans in specific biological processes. In an attempt to uncover the functions of these orphans and to have a complementary understanding (to traditional methods) of their biological roles in *E. coli* as well as in other of its close relatives, I highlight this resource in this chapter which provides a 'systems-wide' functional blueprint of a model microbe, with insights into the biological and evolutionary significance of previously uncharacterized proteins. The network-based methods developed and the approach adopted in this study can not only be used for understanding the functions of uncharacterized genes in other prokaryotic systems but will also enable to identify novel cellular processes and the interplay between them – an fundamental goal of systems biology which at the moment is rather under-appreciated.

Defining the precise biological roles and relationships of bacterial gene products in an often dynamically changing physiological context is a challenging proposition. Historically, systematic assessments of protein function in bacteria have tended to rely on molecular inferences based on sequence alignments and domain architectures, while experimental characterization has traditionally been driven by specific scientific interests rather than with the aim of providing the broader community with unbiased collections of functionally-related proteins and phenotypes. Since the biological role of a protein is not necessarily reflected in its primary sequence, the elucidation of molecular interaction networks can provide an alternate perspective even in the absence of detailed phenotypic data (Ideker and Sharan, 2008; Lee et al., 2008). Therefore, the notion of viewing a model microbial cell mechanistically as a series of modular molecular interaction networks that underlie the major biochemical processes that mediate cell homeostasis and proliferation provides a complementary understanding of biological systems with insights into the functional roles of proteins in the context of other cellular entities.

Since the various methods used in this study discover different types of molecular relationships and each has its own intrinsic bias, complementary information was obtained through data integration. The limited overlap between the high-confidence physical and functional interaction networks presumably stems in part from to the incomplete coverage typically achieved by high-throughput experiments and their methodological differences (Rajagopala et al., 2007; Yu et al., 2008). For example, certain orphans were difficult to evaluate by GC methods due to a lack of apparent orthologs at medium-to-high evolutionary distances, which hinders comparative genomic inferences. Likewise, although large-scale tandem affinity tagging and purification was performed under near-native physiological conditions to generate highly purified preparations of stable, endogenous multiprotein complexes, complete coverage of the proteome was not achieved. For instance, a large number of membrane-associated proteins were not purified, which require specialized solubilization procedures, while the soluble proteins that we failed to tag or detect by mass spectrometry were presumably either of very low abundance or not expressed in our growth conditions.

The observation that the intersection of functional genomics inferences with low-throughput curated physical interaction data is somewhat higher might be explained by two non-mutually exclusive ways: first, protein-protein interactions reported in the literature based on traditional biochemical methods might be biased towards the most evolutionarily conserved multiprotein complexes, which tend to be enriched for essential components with broadly distributed phylogenetic profiles that are more easily and accurately predicted by GC methods

(like those of ribosomal proteins which form conserved clusters on the genome); second, the relatively high sensitivity of the two complementary forms of protein mass spectrometry used in this study may have resulted in the detection of lower abundance orphan proteins that have previously not been studied in depth.

In general, the high confidence functional relationships inferred for *E. coli* in this study can be validated by independent experimental tests, and can be extrapolated to other bacterial species, including pathogens. In fact, over 35% of the orphans find orthologs as far away as Archaea, and hence are likely associated with the same basic housekeeping processes we predict for *E. coli*, such as formation of the cell wall and protein synthesis. Conversely, our systematic comparisons also revealed some unique aspects of the orphans in the evolutionary history of *E. coli,* such as the potential fimbriael factors that appear to be restricted to Enterobacteriaceae. One interpretation is that orphans (and orphan groups) with limited phylogenetic distributions in any major phyla contribute to fine tuning of adaptive physiological responses upon changing environmental conditions and hence might be responsible for not yet characterized processes in bacterial adaption. Alternatively, some orphans might belong to the well conserved biological systems which still need to be characterized for their functional role.

6.2.3 Structure and dynamics of post-transcriptional networks controlled by RNA binding proteins

While transcription factors regulate the synthesis of RNA of specific gene in response to different internal and external stimuli at the level of transcription and several post-translational modifications, such as phosphorylation by kinases and ubiquitin ligases, are known to spatially and temporally control the availability of functional protein products within the cell, little is known about regulation at the post-transcriptional level and major players involved in it. In contrast to prokaryotes where transcription and translation are coupled, in eukaryotes transcription usually takes place in nucleus and translation in cytoplasm. This uncoupling of transcription and translation provides an additional level of gene regulation at post-transcriptional level in eukaryotes. Although ignored for a long time the presence of this post-transcriptional control has been evidenced by a number of post-genomic studies which showed that in general there is a poor correlation between the mRNA and protein pools in eukaryotic cells (Greenbaum et al., 2003; Gygi et al., 1999; Ideker et al., 2001). It is now increasingly known that this level is controlled by numerous factors with major players being the RNA-binding proteins (RBPs) (Glisovic et al., 2008; Keene, 2007; Mata et al., 2005). These observations have suggested that there is need for an intricate co-ordination of regulatory events from these three different layers

to finely control the flow of genetic information from genes to proteins in different conditions. Indeed, changes in gene expression due to aberrations at any of these three levels have been shown to be responsible for the cause of a number of disorders (Cookson et al., 2009; Cooper et al., 2009; Feinberg and Tycko, 2004; Lukong et al., 2008; Nica and Dermitzakis, 2008).

In Chapter 5, I introduce the important class of post-transcriptional regulators - RBPs and show that RBPs are key regulators of different steps in the metabolism of RNA in eukaryotes including splicing, poly-adenylation, capping to get mature mRNA, localization, translation, stability and degradation of cellular RNAs. To regulate all these different steps of RNA metabolism, RBPs bind to RNA and form ribonucleoprotein complexes (RNP). RNPs are inherently highly dynamic complexes due to their ability to associate and dissociate with various RBPs to mediate different steps of RNA metabolism. I then summarize based on current knowledge that RBPs control almost all the steps at post-transcriptional level with some RBPs having the ability to be involved in multiple steps of a post-transcriptional regulatory cascade. I also argue that the complex combinatorial interplay of different RBPs to integrate various post-transcriptional events is an inherent property of these post-transcriptional controllers as this property facilitates them to fine tune the availability of transcripts both spatially and temporally.

I then provide an overview of the recent developments in our understanding of the repertoire of RBPs across diverse model systems and discuss the approaches currently available for the construction of post-transcriptional networks governed by them. Following that I present for the first time an indepth analysis of the properties of post-transcriptional network governed by RBPs and proceed to discuss a study where in we compared the expression dynamics of RBPs with other protein coding genes in yeast (Mittal et al., 2009). The analysis on the expression dynamics showed that RBPs are generally less stable at the transcript level but exhibit higher stability and abundance at the protein level demonstrating that they form a group of proteins which follow the theoretically proposed time averaging effect on noise propagation (Paulsson, 2004), which suggests that if the protein has long half life compared to its mRNA then it averages over the noisy fluctuations in the mRNA, thereby decreasing the protein expression noise. These results also indicate that regulation of RBPs is predominantly controlled at the protein level through the use a number of post-translational modifications (PTMs) like phosphorylation, arginine methylation and sumoylation which have been reported to occur in several well-studied RBPs (Schullery et al., 1999; Vassileva and Matunis, 2004; Yu et al., 2004). Indeed, I also show that a comparison of the number of phosphorylated targets in RBPs and non-RBPs reveals the predominance of post-translational control in RBPs. Based on this I suggest that a wide variety of these PTMs might be responsible for their ability to spatially

and temporally regulate transcripts in eukaryotic systems. It is possible to speculate from these observations that the low noise levels of RBPs together with extensive regulatory flexibility at the protein level might give them an advantage to control gene regulation at a finer level compared to transcriptional control by transcription factors. This might thereby provide a quick and extensive framework for controlling gene expression of a wide range of genes. This is also supported based on the observations presented in this chapter that RBPs which are central to the cell are not only required in large quantities but are also found to be present for a longer time in the cell.

Implications and Future Directions

The observation that short regulatory pathways composing of multiple feed-forward loops with negative auto-regulatory interactions (of TFs) are abundant in the sub-network controlling metabolic functions, such as the use of alternative carbon sources in *E. coli,* indicates that free living bacteria which have the ability to uptake a wide number of sugars to adapt themselves to diverse conditions must harbor a high number of such circuits as a means of switching between different carbon sources. Likewise, organisms living in extremely fluctuating environments might comprise of a higher number of long hierarchical cascades so as to accommodate them with developmental-like pathways so that a mixed number of phenotypes can be generated to survive the variations in the conditions. Alternatively, network structure in such fluctuating environments might be complemented with longer cascades or even bifurcations or divisions in the already established circuits as a means of generating novelty to the existing developmental programs. Part of the plasticity in such extended network structure could come from the presence of multiple auto-regulatory TFs at different stages so that decisions can be made at multiple stages enhancing the number of phenotypes and hence the adaptive potential of microbes. Therefore, while variations in regulatory network topology might be expected, for instance in the case of bacteria with asymmetric cell division (mostly alpha-proteobacteria), where the offspring asymmetric cells cause a transient genetic asymmetry that triggers different developmental processes, such as the formation of stalked and swarmer cells in Caulabacter or vegetative and spore-forming cells in Bacillus (Ausmees and Jacobs-Wagner, 2003; Dworkin, 2003; Dworkin and Losick, 2001; Hilbert and Piggot, 2004; Yudkin and Clarkson, 2005), future comparisons between network topologies for different model systems should further enhance our understanding of regulatory network organization and its conservation or variations among different bacterial phyla.

It is now clear that regulatory networks are plastic with TFs evolving faster than their target genes (Borneman et al., 2007; Hogues et al., 2008; Lozada-Chavez et al., 2006; Madan Babu et al., 2006; Tuch et al., 2008). For instance, TFs, TF-families and global regulators have been shown not to be conserved in between different major groups of bacteria such as *E. coli* and *B. subtilis,* with different families being distinctly expanded in different lineages (Janga and Perez-Rueda, 2009; Lozada-Chavez et al., 2006; Madan Babu et al., 2006). This suggests that although the general topological properties such as power-law degree distribution, hierarchical organization etc of the regulatory network as well as the gene repertoire might be well-conserved across organisms, variations might be happening at the wiring of the interconnections between the TFs and their targets between different organisms. This implies that both genomic organization and architecture on one hand and TFs and their binding specificities and locations across genome on the other, play major roles in enabling a significant rewiring of the network across organisms as is evidenced from some recent studies (Borneman et al., 2007; De et al., 2009; Tuch et al., 2008). It is easy to imagine that this genomic and network rewiring together can explain the conservation of the constraints across genomes suggesting that the observed constraints might be generic principles valid for all genomes. Nevertheless, it remains to be learnt whether this rewiring is valid for high eukaryotes and if so how fast and what factors might best explain the rewiring while preserving the constraints. Recent experimental data show that regulatory networks can be plastic even among members of a population indicating that transcriptional control is much more dynamic in evolution than previously thought (Kasowski et al., ; McDaniell et al., ; Zheng et al.). Naturally, it follows that some types of TFs might be showing greater plasticity than others in closely related organisms or with in individuals and hence might be major contributors for the rewiring of regulatory programs. Therefore, it would be interesting to understand the design principles underlying these variations. In light of these recent studies, an interesting open question which still remains is whether the rewiring in regulatory network can be explained based on phylogenetic distance or if other factors like adaptation play more important roles as has been seen in bacteria (Madan Babu et al., 2006).

Observations in Chapter 4 show that genome-context and network-based methods developed here can be employed as powerful means for automating the functional prediction pipelines so that any newly sequenced genome can be studied as soon as the gene coordinates are available. With the availability of metagenomic sequences the power of these computational methods for generating functional association networks will increase not only in terms of

coverage but also in terms of quality of predicted associations- thereby increasing the quality of the function predictions. With the increasingly cheaper availability of RNA sequencing technologies it should be possible to construct expression compendiums to bacterial genomes as soon as genome sequences are available. Thus, enabling the addition of these high-throughput transcriptomic profiling data to function prediction pipelines just like what microarrays did in the past decade. Integration of all these high-throughput computational methods together with genetic, physical and small molecular perturbation experiments aimed to provide different kinds of associations in a condition specific manner, should all enable the rapid screening for the phenotypes of newly identified genes faster than it was in the previous century and in elucidating their detailed functional inter-relationships with the rest of the cellular machinery.

In a similar vein, availability of invivo crosslinking assays and cheaper sequencing should enable the identification of RNA targets of RBPs from different families at single nucleotide resolution which should enable the elucidation of genome-wide RBP-RNA maps. Such high density maps will not only permit the understanding of the mechanism of action of RBPs but also, when they are performed on a high-throughput way for many RBPs, allow the understanding of their interplay in Ribo-nucleoproteins (RNPs) to mediate different RNA processing events. In addition, such maps also allow the variations in the binding of different RBPs across conditions and between cell types so that tissue-specific variations and aberrations can be identified to further exploit them for therapeutic use. The data generated using these high-throughput techniques will also enable the improvements in our understanding of the cross-talk between different post-transcriptional events and processes. Understanding the links between different layers of regulation can also improve our global understanding of regulatory processes enabling a better modeling of eukaryotic systems.

In general, the vast amount of data that will be generated using the next generation technologies in the coming years, will form a foundation not only to test many of the hypothesis that have been generated during my doctoral work but will also improve our understanding of the interpretations of these constraints at different levels. Improved understanding of the design principles governing biological systems will improve our ability to model disease phenotypes in a larger context. Such developments will allow the development of disease treatment strategies using modern systems approaches (Janga and Tzakos, 2009).

REFERENCES

Ausmees, N. and Jacobs-Wagner, C. (2003). Spatial and temporal control of differentiation and cell cycle progression in Caulobacter crescentus. *Annu Rev Microbiol* **57**, 225-47.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M. and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815-9.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184-94.

Cooper, T. A., Wan, L. and Dreyfuss, G. (2009). RNA and disease. *Cell* **136**, 777-93.

De, S., Teichmann, S. A. and Babu, M. M. (2009). The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Res* **19**, 785-94.

Dworkin, J. (2003). Transient genetic asymmetry and cell fate in a bacterium. *Trends Genet* **19**, 107-12.

Dworkin, J. and Losick, R. (2001). Differential gene expression governed by chromosomal spatial asymmetry. *Cell* **107**, 339-46.

Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nat Rev Cancer* **4**, 143-53.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-86.

Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117.

Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720-30.

Hilbert, D. W. and Piggot, P. J. (2004). Compartmentalization of gene expression during Bacillus subtilis spore formation. *Microbiol Mol Biol Rev* **68**, 234-62.

Hogues, H., Lavoie, H., Sellam, A., Mangos, M., Roemer, T., Purisima, E., Nantel, A. and Whiteway, M. (2008). Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell* **29**, 552-62.

Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P. et al. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol* **7**, e96.

Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome Res* **18**, 644-52.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-34.

Janga, S. C., Collado-Vides, J. and Babu, M. M. (2008). Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc Natl Acad Sci U S A* **105**, 15761-6.

Janga, S. C. and Perez-Rueda, E. (2009). Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. *Comput Biol Chem* **33**, 261-8.

Janga, S. C., Salgado, H. and Martinez-Antonio, A. (2009). Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Res* **37**, 3680-8.

Janga, S. C. and Tzakos, A. (2009). Structure and organization of drug-target networks: insights from genomic approaches for drug discovery. *Mol Biosyst* **5**, 1536-48.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E. et al. Variation in transcription factor binding among humans. *Science* **328**, 232-5.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-43.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A. G. and Marcotte, E. M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181-8.

Lozada-Chavez, I., Janga, S. C. and Collado-Vides, J. (2006). Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res* **34**, 3434-45.

Lukong, K. E., Chang, K. W., Khandjian, E. W. and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet* **24**, 416-25.

Madan Babu, M., Teichmann, S. A. and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* **358**, 614-33.

Martinez-Antonio, A., Janga, S. C. and Thieffry, D. (2008). Functional organisation of Escherichia coli transcriptional regulatory network. *J Mol Biol* **381**, 238-47.

Mata, J., Marguerat, S. and Bahler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* **30**, 506-14.

McDaniell, R., Lee, B. K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A. et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235-9.

Mittal, N., Roy, N., Babu, M. M. and Janga, S. C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc Natl Acad Sci U S A* **106**, 20300-5.

Nica, A. C. and Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet* **17**, R129-34.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature* **427**, 415-8.

Rajagopala, S. V., Titz, B., Goll, J., Parrish, J. R., Wohlbold, K., McKevitt, M. T., Palzkill, T., Mori, H., Finley, R. L., Jr. and Uetz, P. (2007). The protein network of bacterial motility. *Mol Syst Biol* **3**, 128.

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T. et al. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* **34**, 1-9.

Schullery, D. S., Ostrowski, J., Denisenko, O. N., Stempka, L., Shnyreva, M., Suzuki, H., Gschwendt, M. and Bomsztyk, K. (1999). Regulated interaction of protein kinase Cdelta with the heterogeneous nuclear ribonucleoprotein K protein. *J Biol Chem* **274**, 15101-9.

Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H. and Johnson, A. D. (2008). The evolution of combinatorial gene regulation in fungi. *PLoS Biol* **6**, e38.

Vassileva, M. T. and Matunis, M. J. (2004). SUMO modification of heterogeneous nuclear ribonucleoproteins. *Mol Cell Biol* **24**, 3623-32.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W. et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. et al. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science.* **322**, 104-110.

Yu, M. C., Bachand, F., McBride, A. E., Komili, S., Casolari, J. M. and Silver, P. A. (2004). Arginine methyltransferase affects interactions and recruitment of mRNA processing and export factors. *Genes Dev* **18**, 2024-35.

Yudkin, M. D. and Clarkson, J. (2005). Differential gene expression in genetically identical sister cells: the initiation of sporulation in Bacillus subtilis. *Mol Microbiol* **56**, 578-89.

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. and Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature*.

APPENDIX

for

EXPLOITING NETWORK-BASED APPROACHES FOR UNDERSTANDING GENE REGULATION AND FUNCTION

SARATH CHANDRA JANGA

CONTENTS OF APPENDIX

A.1 LIST OF PUBLICATIONS

Publications during PhD (January 2008- April 2010)

*indicates corresponding author either jointly or alone

† indicates joint first author

** indicates papers which are not discussed in the thesis but are appended as reprints here

- Ten simple rules for organizing a scientific meeting **
  Manuel Corpas, Nils Gehlenborg, **Sarath Chandra Janga** and Philip E Bourne
  PLoS Comput Biol 2008 4(6):e1000080

- Highlights from the Fourth International Society for Computational Biology Student Council Symposium **
  Lucia Peixoto, Nils Gehlenborg and **Sarath Chandra Janga**
  BMC Bioinformatics, 2008

- Functional organization of *Escherichia coli* transcriptional regulatory network
  Agustino Martinez-Antonio, **Sarath Chandra Janga** and Denis Thieffry
  Journal of Molecular Biology, 2008, Vol. 381(1):238-247

- Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes
  Sarath Chandra Janga*, Julio Collado-Vides and M. Madan Babu
  Proc. Natl. Acad. Sci. U S A. 105(41): 15761-6, 2008
  +Featured on the news and highlights section of the journal Molecular Biosystems

- Eukaryotic gene regulation in three dimensions and its impact on genome evolution
  M. Madan Babu, **Sarath Chandra Janga**, Ines Santiago and Ana Pombo
  Curr. Opin. Genet. Dev., 2008, Vol. 18(6):571-582
  +Featured on the cover page of the issue with a cover image

- Network-based approaches for linking metabolism with environment **
  Sarath Chandra Janga* and M. Madan Babu
  Genome Biology, 2008, 9(11):239
  +Featured on the journal's home page

- Transcript stability in the protein interaction network of *Escherichia coli* **
  Sarath Chandra Janga* and M. Madan Babu
  Molecular Biosystems, 2009, 5(2):154-62
  +Featured on the cover page of the issue with a cover image

- Transcriptional regulation shapes the organization of genes on bacterial chromosomes
  Sarath Chandra Janga*, Heladia Salgado and Agustino Martinez-Antonio
  Nucleic Acids Research, 2009, Vol.37, No. 11, 3680-3688

- Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins
  Pingzhao Hu†, **Sarath Chandra Janga**†, Mohan Babu†, J. Javier Díaz-Mejía†, Gareth Butland†, et. al

PLoS Biology 2009, 7(4): e96
[+]Featured in the journal 'Nature methods'

- Scaling relationship in the gene content of transcriptional machinery in bacteria **
  Ernesto Perez-Rueda, **Sarath Chandra Janga*** and Agustino Martinez-Antonio
  Molecular Biosystems, 2009, 5(12):1494-501

- Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families **
  Sarath Chandra Janga* and Ernesto Perez-Rueda
  Computational Biology and Chemistry, 2009, Vol. 33, No. 4, 261-268

- Structure and organization of drug-target networks : Insights from genomic approaches for drug discovery **
  Sarath Chandra Janga* and Andreas Tzakos
  Molecular Biosystems, 2009, 5(12):1536-48

- Interfacing systems biology and synthetic biology **
  Allyson Lister, Varodom Charoensawan, Subhajyoti De, Katherine James, **Sarath Chandra Janga** and Julian Huppert
  Genome Biology, 2009, *10(6):309*

- Dissecting the expression dynamics of RNA-binding proteins in post-transcriptional regulatory networks
  Nitish Mittal, Nilanjan Roy, M. Madan Babu and **Sarath Chandra Janga***
  Proc. Natl. Acad. Sci. U S A. 106(48): 20300-05, 2009

- Protein Complexes and Functional Pathways in *S. cerevisiae* and *E. coli*
  Mohan Babu, Gareth Butland, J. Javier Díaz-Mejía, Pingzhao Hu, S Pu, Gabriel Moreno-Hagelsieb, **Sarath Chandra Janga**, Shoshana Wodak, Andrew Emili, Jack Greenblatt
  Mol Cell Proteomics, S27-27, 2009 (Meeting Abstract)

- Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin **
  Ernesto Perez-Rueda and **Sarath Chandra Janga***
  Mol Biol Evol., 2009. 27(4): 1-11, 2010

- Transcriptional regulatory networks
  Book chapter for the edited book '*Networks in Cell Biology*', 2010 for *Cambridge University Press* (Ed: Michele Vendruscolo, Department of Chemistry, University of Cambridge)
  Sarath Chandra Janga* and M. Madan Babu

- Operons and bacterial genome organization **
  Book chapter for the edited book "*Bacterial Gene Regulation and Transcriptional Networks*", 2010 for *Horizon Scientific Press* (Ed: M. Madan Babu, MRC Laboratory of Molecular Biology, University of Cambridge)
  Sarath Chandra Janga* and Gabriel Moreno-Hagelsieb

- Construction, structure and dynamics of post-transcriptional networks directed by RNA-binding proteins
  Book chapter for the edited book "RNA infrastructure: RNA processing and regulatory networks", 2010 for *Springer/Landes Bioscience Press* (Ed: Lesley Collins, Allan Wilson

Centre for Molecular Biology and Ecology & Institute of Molecular BioSciences, Massey University)
Sarath Chandra Janga* and Nitish Mittal

Publications under review, revision and in preparation

- Coordination of bacterial transcription and the role of the RNA polymerase omega subunit
  Marcel Geertz, Andrew Travers, **Sarath Chandra Janga**, Sanja Mehandziska, Nobuo Shimamoto and Georgi Muskhelishvili
  Mol. Microbiology, 2009 (Submitted)

- Network-based function prediction in post-genomic era : Metabolic enzymes as a case study
  Sarath Chandra Janga* and Gabriel Moreno-Hagelsieb
  Metabolic Engineering, 2010 (Submitted)

- Dissecting the expression patterns of transcription factors across conditions using an integrated network-based approach
  Sarath Chandra Janga* and Bruno Contreras-Moreira
  Nucleic Acids Research, 2010 (Submitted)

- Polypharmacological approaches to fight antibacterial resistance : Insights from drug-target networks
  Sarath Chandra Janga* and Andreas Tzakos
  Trends in Biotechnology (Submitted)

- Transcriptional profiling of fetal hypothalamic TRH neurons
  Magdalena Guerra-Crespo, Carlos Pérez-Monter, **Sarath Chandra Janga**, Santiago Castillo-Ramírez, Rosa Maria Gutierrez-Rios, Patricia Joseph-Bravo, Leonor Pérez-Martínez and Jean-Louis Charli
  BMC Genomics, 2010 (Submitted)

- Genome-wide analysis of RNA decay patterns during early *Drosophila* development
  Stefan Thomsen, Simon Anders, **Sarath Chandra Janga**, Wolfgang Huber and Claudio R. Alonso
  Genome Biology, 2010 (Submitted)

- Systematic identification of RNA-binding proteins in yeast suggests dual functions for enzymes
  Tanja Scherrer, Nitish Mittal, **Sarath Chandra Janga**, André P Gerber
  Nature Molecular Systems Biology (Submitted)

- Intrinsic modularity in the genomic organization of eubacterial transcription factors
  Sarath Chandra Janga* and Gabriel Moreno-Hagelsieb
  (To be Submitted)

- Dissecting the interactome of RNA-binding proteins in post-transcriptional regulatory networks
  Sarath Chandra Janga*, Nitish Mittal and M. Madan Babu
  (To be Submitted)

Publications prior to starting PhD

- Conservation of adjacency as evidence of paralogous operons
  Sarath Chandra Janga and Gabriel Moreno-Hagelsieb
  Nucleic Acids Research, 2004 Vol.32, No. 18, 5392-5397

- Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons
  Sarath Chandra Janga, Julio Collado-Vides and Gabriel Moreno-Hagelsieb
  Nucleic Acids Research, 2005 Vol.33, No. 8, 2521-2530

- The network of transcriptional interactions imposes linear constrains in the genome
  Ricardo Menchaca-Mendez[†], **Sarath Chandra Janga**[†] and Julio Collado-Vides
  OMICS: A Journal of Integrative Biology Jun 2005, Vol.9, No. 2: 139-145

- Internal sensing machinery directs the activity of the regulatory network in *Escherichia coli*
  Agustino Martínez-Antonio, **Sarath Chandra Janga**, Heladia Salgado and Julio Collado-Vides
  Trends in Microbiology, 2006 Vol.14, No. 1, 22-27

- The Partitioned *Rhizobium etli* Genome: Genetic and Metabolic Redundancy in Seven Interacting Replicons
  Víctor González, Rosa I. Santamaría, Patricia Bustos, Ismael Hernández-González, Arturo Medrano-Soto, Gabriel Moreno-Hagelsieb, **Sarath Chandra Janga**, Miguel A. Ramírez, Verónica Jiménez-Jacinto, Julio Collado-Vides and Guillermo Dávila
  Proc. Natl. Acad. Sci. U S A. 103(10): 3834-9, 2006

- The distinctive signatures of promoter regions and operon junctions across Prokaryotes
  Sarath Chandra Janga*, Warren F. Lamboy, Araceli M. Huerta and Gabriel Moreno-Hagelsieb
  Nucleic Acids Research, 2006 Vol.34, No. 14, 3980-3987

- Bacterial regulatory networks are extremely flexible in evolution
  Irma Lozada-Chávez, **Sarath Chandra Janga*** and Julio Collado-Vides
  Nucleic Acids Research, 2006 Vol.34, No. 12, 3434-3445
  [+]Featured in the list of hot research papers on NAR website

- Identification and analysis of DNA-binding Transcription Factors in Bacillus subtilis and other Firmicutes- A genomic approach
  Samadhi Moreno-Campuzano, **Sarath Chandra Janga** and Ernesto Perez-Rueda
  BMC Genomics. 2006 Jun 13;7(1):147
  [+]Accessed over 800 times in less than 5 months according to BMC report

- Prediction and evolution of transcription factors and their evolutionary families in prokaryotes
  Sarath Chandra Janga*
  BMC Systems Biology, 2007, 1(Suppl 1):P3 (tutorial presentation as part of the proceedings of the BioSysBio conference)

- Internal versus external effector and transcription factor gene pairs differ in their relative chromosomal position in *Escherichia coli.*

Sarath Chandra Janga*, Heladia Salgado, Julio Collado-Vides and Agustino Martinez-Antonio
Journal of Molecular Biology, 2007 Vol. 368(1):263-72

- Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles
  Gabriel Moreno-Hagelsieb and **Sarath Chandra Janga**
  Proteins: Structure, Function, and Bioinformatics, 2007 70(2):344-352

- Conservation of Transcriptional Sensing Systems in prokarya: A perspective from *Escherichia coli*
  Heladia Salgado, Agustino Martinez-Antonio and **Sarath Chandra Janga***
  Febs letters, 2007 Vol. 581:3499-3506

- Structure and evolution of gene regulatory networks in microbial genomes
  Sarath Chandra Janga* and Julio Collado-Vides
  Research in Microbiology, 2007 158(10):787-94
  [+]Featured on the cover page of the issue & selected as journals' representative image

- Co-ordination logic of sensing machinery in the transcriptional regulatory network of *Escherichia coli*
  Sarath Chandra Janga*, Heladia Salgado, Agustino Martinez-Antonio and Julio Collado-Vides
  Nucleic Acids Research, 2007 Vol.35, No. 20, 6963-6972

- Highlights from the Third International Society for Computational Biology Student Council Symposium
  Nils Gehlenborg, Manuel Corpas and **Sarath Chandra Janga***
  BMC Bioinformatics, 2007, 8(Suppl 8):I1

A.2 REPRINTS

(Please see next pages for publications not discussed in the thesis)

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Organizing a Scientific Meeting

Manuel Corpas[1], Nils Gehlenborg[1,2], Sarath Chandra Janga[3], Philip E. Bourne[4]*

1 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **2** Graduate School of Life Sciences, University of Cambridge, Cambridge, United Kingdom, **3** Medical Research Council–Laboratory of Molecular Biology, University of Cambridge, Cambridge, United Kingdom, **4** Skaggs School of Pharmacy and Pharmaceutical Science, University of California San Diego, La Jolla, California, United States of America

Scientific meetings come in various flavors—from one-day focused workshops of 1–20 people to large-scale multiple-day meetings of 1,000 or more delegates, including keynotes, sessions, posters, social events, and so on. These ten rules are intended to provide insights into organizing meetings across the scale.

Scientific meetings are at the heart of a scientist's professional life since they provide an invaluable opportunity for learning, networking, and exploring new ideas. In addition, meetings should be enjoyable experiences that add exciting breaks to the usual routine in the laboratory. Being involved in organizing these meetings later in your career is a community responsibility. Being involved in the organization early in your career is a valuable learning experience [1]. First, it provides visibility and gets your name and face known in the community. Second, it is useful for developing essential skills in organization, management, team work, and financial responsibility, all of which are useful in your later career. Notwithstanding, it takes a lot of time, and agreeing to help organize a meeting should be considered in the context of your need to get your research done and so is also a lesson in time management. What follows are the experiences of graduate students in organizing scientific meetings with some editorial oversight from someone more senior (PEB) who has organized a number of major meetings over the years.

The International Society for Computational Biology (ISCB) Student Council [2] is an organization within the ISCB that caters to computational biologists early in their career. The ISCB Student Council provides activities and events to its members that facilitate their scientific development. From our experience in organizing the Student Council Symposium [3,4], a meeting that so far has been held within the context of the ISMB [5,6] and ECCB conferences, we have gained knowledge that is typically not part of an academic curriculum and which is embodied in the following ten rules.

Rule 1: The Science Is the Most Important Thing

Good science, above all else, defines a good meeting; logistics are important, but secondary. Get the right people there, namely the best in the field and those who will be the best, and the rest will take care of itself. When choosing a topic for your conference, map it to the needs of your target audience. Make sure that you have a sufficiently wide range of areas, without being too general. The greater the number of topics covered, the more likely people are to come, but the less time you will have to focus on particular subject matter. Emerging areas can attract greater interest; try to include them in your program as much as possible; let your audience decide the program through the papers they submit to the general call for papers. This can be done with broad and compelling topic areas such as "Emerging Trends in …" or "New Developments in …".

Rule 2: Allow for Plenty of Planning Time

Planning time should range from nine months to more than a year ahead of the conference, depending on the size of your event. Allow plenty of time to select your meeting venue; to call for, review, and accept scientific submissions; to arrange for affordable/discounted hotel rooms; to book flights and other transportation options to the conference. Having outstanding keynote speakers at your event will also require you contact them months in advance—the bigger the name, the more time is required.

Rule 3: Study All Potential Financial Issues Affecting Your Event

Sponsors are usually your primary source of funds, next to the delegates' registration fees. To increase the chances of being sponsored by industry, write them a clear proposal stating how the money will be spent and what benefits they can expect to get in return. You may also want to reserve a few time slots for industry talks or demos as a way of attracting more sponsors, but be wary that the scientific flavor of the meeting is not impacted by blatant commercialism. Make sure you first approach the sponsors that match your interest topics the closest. If they say they are not interested this year, keep their contact information, as they might be able to sponsor you in future events. Approach them early rather than later in any case. The cost of your conference will be proportional to the capacity of the venue; therefore, a good estimation of the number of attendees will provide you with a good estimate of your costs. You will need to include meals and coffee breaks together with the actual cost of renting your venue. Be aware that audiovisual costs can be additional as well as venue staff—look out for hidden costs. Aside from venue-related costs, additional expenditures might include travel fellowships, publication costs for proceedings in a journal, and awards for outstanding contributors. All these issues will determine how much you need to charge your participants to attend. Map all this out on a spreadsheet and do the math. Allow for contingencies, such as currency fluctuations and world-changing

Citation: Corpas M, Gehlenborg N, Janga SC, Bourne PE (2008) Ten Simple Rules for Organizing a Scientific Meeting. PLoS Comput Biol 4(6): e1000080. doi:10.1371/journal.pcbi.1000080

Published June 27, 2008

Funding: The authors have received no specific funding for this article.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bourne@scsd.edu

events that will impact attendance. For large meetings, consider insurance against such events. Starting with a template that others have used for previous similar conferences can be a big help.

Rule 4: Choose the Right Date and Location

Your conference needs to be as far away as possible from established conferences and other related meetings. Alternatively, you may want to organize your event around a main conference, in the form of a satellite meeting or Special Interest Group (SIG). Teaming up with established conferences may increase the chances of attracting more people (especially if this is your first time) and also save you a great deal of administrative work. If you decide to do it on your own, you should consider how easy it is to travel to your chosen location, whether it has a strong local community in your field, and whether it has cultural or other tourist attractions. Inexpensive accommodation and airfares to your conference are always a plus.

Rule 5: Create a Balanced Agenda

A conference is a place for people wanting to share and exchange ideas. Having many well-known speakers will raise the demand for your event (and the cost) but that has to be balanced with enough time for presentation of submitted materials. A mix of senior scientists and junior scientists always works for the better. Young researchers may be more enthusiastic and inspiring for students, while top senior scientists will be able to present a more complete perspective of the field. Allow plenty of time for socializing, too; breaks, meals, and poster sessions are ideal occasions to meet potential collaborators and to foster networking among peers.

Rule 6: Carefully Select Your Key Helpers: the Organizing Committees

A single person will not have all the skills necessary to organize a large meeting, but the organizing committee collectively needs to have the required expertise. You might want to separate the areas of responsibilities between your aides depending on their interests and availability. Some potential responsibilities you might delegate are: 1) content and design of the Web site promoting the meeting; 2) promotion materials and marketing; 3) finance and fundraising; 4) paper submissions and review; 5) posters; 6) keynotes; 7) local organization; 8) program and speakers; 9) awards. Your organizing committee should be large enough to handle all the above but not too large, avoiding freeloaders and communication issues. It is invaluable to have a local organizing committee since they know local institutions, speakers, companies, and tourist attractions. Local organizations may also help you with administrative tasks; for example, dealing with registration of attendees and finding suitable accommodations around the venue.

Rule 7: Have the Members of the Organizing Committees Communicate Regularly

It is good to have planning sessions by teleconference ahead of the meeting. As far as possible, everyone should be familiar with all aspects of the meeting organization. This collective wisdom will make it less likely that important issues are forgotten. The local organizers should convince everyone that the venue will work. Use these sessions to assign responsibilities ahead of the meeting. Tasks such as manning the registration tables, carrying microphones for attendees to ask questions, introducing sessions and speakers, checking presentations ahead of time, and having poster boards, materials to attach posters, etc., are easily overlooked. In short, good communication will lead to you covering all the little things so easily forgotten.

Good communication continues throughout the meeting. All organizers should be able to contact each other throughout the meeting via mobile phone and e-mail. Distribute to all organizers the names and contact information of caterers, building managers, administrative personnel, technicians, and the main conference organizer if you are having your event as part of another conference. Onsite changes that incur additional costs, however, should require the approval of a single, key organizer rather than all organizers operating independently of one another. This will ensure there are no financial surprises in the end. It is also important that you have a designated meeting point where someone from the organizing committee is going to be available at all times to help with problems.

Rule 8: Prepare for Emergencies

Attendees need to be aware of all emergency procedures in terms of evacuation, etc. This should be discussed with the venue managers. All attendees should be reachable as far as possible during the conference. If an attendee has an emergency at home, his or her family should be able to reach them through the conference desk—mobile phones are not perfect after all.

Rule 9: Wrap Up the Conference Properly

At the end of the conference, you should give credit to everyone who helped to make the event a success. If you have awards to present, this is the right time for the awards ceremony. Dedicate some time to thank your speakers and sponsors as well as everyone involved in the organization of the conference. Also collect feedback about the event from the delegates through questionnaires. This evaluation will help you to understand the strengths and weaknesses of your conference and give you the opportunity to improve possible future events. Have a party or some other event for all those organizing the conference.

Rule 10: Make the Impact of Your Conference Last

Published proceedings are the best way to make the results of your conference last. Negotiate with journals far in advance of the conference to publish the proceedings. Make those proceedings as widely accessible as possible. Upload photos and videos of the event to the conference Web site and post the names of presenters who have received awards or travel fellowships. It is also a good idea to link the results of your evaluation to the Web site. Send one last e-mail to all delegates, including a summary of the activities since the conference and thanking them for their participation. This is particularly important if you are considering holding the conference again in future years, in which case include some information on your plans for the next event.

As always, we welcome your comments and experiences that you think would enrich these ten rules so that they might be useful to others. The comment feature now supported by this journal makes it easy to do this.

Acknowledgments

We would like to acknowledge the International Society for Computational Biology (ISCB) for their support in the organization of the Student Council Symposiums, in particular BJ Morrison-McKay and Steven Leard. Thanks to Michal Linial and Rita Casadio (our liaisons

at the ISCB Board of Directors), Burkhard Rost (the ISCB President), and all the ISCB Board of Directors for being so supportive of the work of the Student Council. We are also grateful to all the Student Council leadership and current and past Student Council members for their enthusiasm and hard (unpaid) work. You all have made the Student Council a great organization.

References

1. Tomazou EM, Powell GT (2007) Look who's talking, too: Graduates developing skills through communication. Nat Rev Genet 8: 724–726. doi:10.1038/nrg2177.
2. The International Society for Computational Biology Student Council. Available: http://www.iscbsc.org. Accessed 22 April 2008.
3. Corpas M (2005) Scientists and societies. Nature 436: 1204. doi:10.1038/nj7054–1204b.
4. Gehlenborg N, Corpas M, Janga SC (2007) Highlights from the Third International Society for Computational Biology (ISCB) Student Council Symposium at the Fifteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB). BMC Bioinformatics 8 (Supplement 8):I1.
5. Lengauer T, McKay BJM, Rost B (2007) ISMB/ECCB 2007: The premier conference on computational biology. PLoS Comput Biol 3: e96. doi:10.1371/journal.pcbi.0030096.
6. Third ISCB Student Council Symposium. Available: http://www.iscbsc.org/scs3 Accessed 22 April 2008.

Minireview

Network-based approaches for linking metabolism with environment

Sarath Chandra Janga and M Madan Babu

Address: MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK.

Correspondence: Sarath Chandra Janga. Email: sarath@mrc-lmb.cam.ac.uk

Published: 24 November2008

*Genome **Biology*** 2008, **9:**239 (doi:10.1186/gb-2008-9-11-239)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/11/239

Abstract

Progress in the reconstruction of genome-wide metabolic maps has led to the development of network-based computational approaches for linking an organism with its biochemical habitat.

The sequential nature of the reactions in metabolic pathways means that they can be modeled in the form of a graph (network) of enzymes and chemical transformations, and network theory can be used to represent and understand metabolism [1,2]. The connected collection of metabolic pathways, describing the set of all enzymatic interconversions of one small molecule into another, is defined as the metabolic network of an organism (Figure 1a).

The most commonly used network representations are 'metabolite-centric'. They consider metabolites as the nodes of the graph and two metabolites are linked if one can be converted into the other by an enzymatic reaction (Figure 1b, left). An alternative network representation is 'enzyme-centric'. It considers the enzymes as nodes and links enzymes that catalyze successive reactions (Figure 1b, right). Although several studies have provided insights into the structure and evolution of a metabolic network, very few have addressed the influence of environment on metabolic network structure in species from diverse environmental conditions. The availability of many completely sequenced genomes means that metabolic-network analysis can now be extended from a few model organisms to species from different branches of the tree of life and living in very different environments. This should enable the elucidation of general principles underlying metabolic networks.

Two recent studies, published in the *Proceedings of the National Academy of Sciences* by Eytan Ruppin and colleagues (Kreimer *et al.* [3] and Borenstein *et al.* [4]), provide important insights into links between the environment of an organism and the structure of its metabolic network. Using data from a large number of bacterial metabolic networks, Kreimer *et al.* address the question of how the topologies of the metabolic networks from different species reflect both genome size and the diversity of environmental conditions the species would encounter. Borenstein *et al.* set out to identify the 'seed set' - that set of small molecules that are absolutely needed from the external environment - of each species and how this seed set differs across species from different environments.

A network view of metabolism

Several studies have addressed a wide-range of questions using network representation of small-molecule metabolism [5-7]. For instance, at the structural level, the metabolic network of an organism has been shown to have a scale-free topology with few nodes (for example, pyruvate or coenzyme A) reacting with many other substrates [8,9]. A distinguishing feature of such scale-free networks is the existence of a few highly connected metabolites, which participate in a very large number of metabolic reactions. By definition, when a large number of links integrate several substrates into a single highly connected component, fully separated modules will not exist. This has led to the notion of hierarchical modular structures within the fully connected metabolic network, where a 'module' is defined as a group of nodes that are more connected to each other than to other nodes in the network [10].

Kreimer *et al.* [3] have carried out a comprehensive, large-scale characterization of metabolic-network modularity (defined as in [11]) using 325 prokaryotic species with

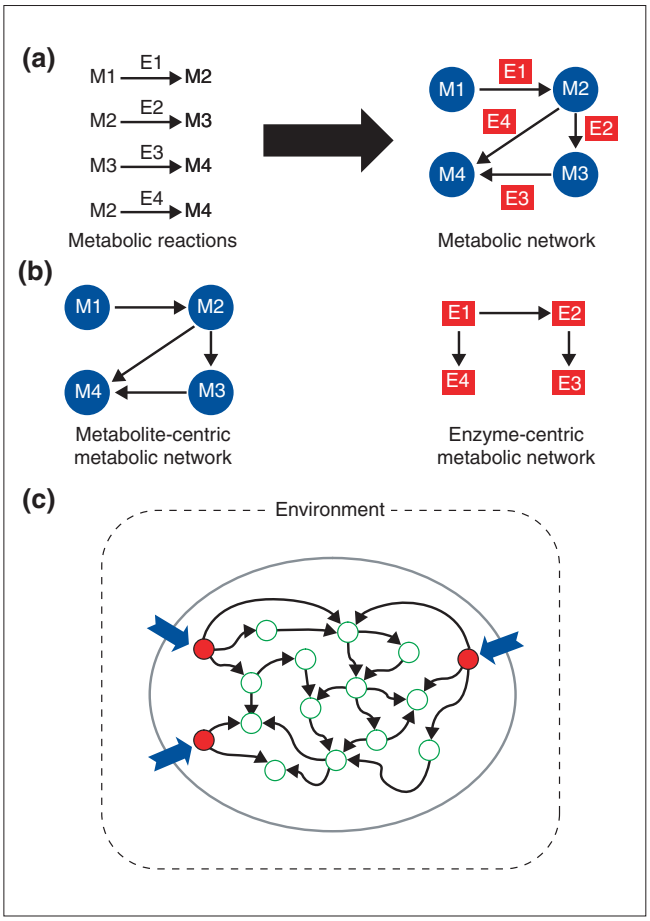


Figure 1
Metabolic networks. **(a)** A set of related metabolic reactions can be represented as a network. M1, M2, and so on are metabolites and E1, E2, and so on are the enzymes that catalyze the conversion of one metabolite into another. The arrows represent the direction of the reaction. **(b)** Different ways of representing a metabolic network: left, with the metabolites as nodes; right, with the enzymes as nodes. **(c)** Representation of seed compounds in a hypothetical metabolic network. The metabolic boundary of the organism is represented by the gray oval. Metabolites (the nodes in the network) are represented by colored circles. The set of compounds that cannot be internally synthesized but must be obtained from the environment is referred to as the seed set, and is represented here as red circles. Seed metabolites form the interface between the environment and the metabolic system and link the metabolic habitats of an organism with its core metabolic processes. In this hypothetical network, it is possible to reach any of the internal nodes (open green nodes) from any other node except those that have to be obtained from the environment (blue arrows).

sequenced genomes and metabolic networks in the KEGG pathway database [12]. They found that network size was an important topological determinant of modularity, with larger genomes exhibiting higher modularity scores (that is, a higher proportion of edges in the network forming part of modules than would be expected by chance). In addition, several environmental factors were shown to contribute to the variation in metabolic-network modularity across species. In particular, the authors found that endosymbionts and mammal-specific pathogens have lower modularity scores than bacterial species that occupy a wider range of niches. Moreover, among the pathogens, those that alternate between two distinct niches, such as insect and mammal, were found to have relatively high metabolic-network modularity. This supports the notion previously put forward by Parter *et al.* [13] that variability in the natural habitat of an organism promotes modularity in its metabolic network. Kreimer *et al.* [4] also reconstructed likely ancestral states, and found that modularity tends to decrease from ancestors to descendants; they attribute this to niche specialization and incorporation of peripheral metabolic reactions.

In line with the above effects of environmental diversity on network structure, Pal *et al.* [14] observed that bacterial metabolic networks grow by retaining horizontally acquired genes (genes acquired from other species) involved in the transport and catalysis of external nutrients, and that evolutionary changes in networks are primarily driven by adaptation to changing environments. Accordingly, horizontally transferred genes were found to be integrated at the periphery of the network, whereas the central parts remain evolutionarily stable. Indeed, genes encoding physiologically coupled reactions were often found to be transferred together, frequently in operons. This suggests that bacterial metabolic networks evolve by direct uptake of peripheral reactions in response to changing environments [14].

In this regard, a recent genome-wide study in yeast found that central and highly connected enzymes evolve more slowly than less connected ones and that duplicates of highly connected enzymes tend to have a higher likelihood of retention [15]. Enzymes carrying high metabolic fluxes under natural biological conditions were also found to experience greater evolutionary constraints. Interestingly, however, it was shown that highly connected enzymes are no more likely to be essential to survival than the less connected ones [15].

The functional and evolutionary modularity of the *Homo sapiens* metabolic network has also been investigated from a topological point of view and was shown to be organized with a highly modular, 'core and periphery' topology [16]. In such a structure, the core modules are tightly linked together and perform basic metabolic functions, whereas the peripheral modules only interact with few other modules and accomplish relatively independent and specialized functions. Interestingly, as in bacteria and yeast, peripheral modules were found to evolve more cohesively and faster than core modules [16].

Linking external environment to the metabolic circuitry

Microorganisms constantly monitor their surroundings for the availability of nutrients and other chemicals, using both

Box 1. Models of metabolic pathway evolution

The most influential models of metabolic pathway evolution have been the 'retrograde model' proposed by Horowitz in 1945 [24] and the 'patchwork model' proposed by Ycas in 1974 [25] and later improved by Jensen in 1976 [26].

The retrograde model

In the retrograde model, pathways evolve bottom-up from a key metabolite, which is assumed to be initially abundant in the ancestral condition. The model presupposes the existence of a chemical environment in which both the key metabolite and potential intermediates are available. An organism primarily dependent on molecule Z will use up environmental reserves of the metabolite to the point at which its growth is restricted; in such an environment, an organism capable of synthesizing molecule Z from environmental precursors X and Y will have a selective advantage. Any natural variant evolving an enzyme that catalyzes this synthesis will have a fitness advantage in such an environment. As a result, with the drop in environmental concentration of X or Y, the process will be repeated, with the similar recruitment of further enzymes.

The retrograde model also proposes that the simultaneous unavailability of two intermediates (say X and Y) would favor symbiotic association between two mutants, one capable of synthesizing X and the other of synthesizing Y from other environmental precursors. One of the major assumptions of this model is that the evolution of metabolic pathways occurs in an environment rich in metabolic intermediates, and it therefore cannot explain their evolution during major environmental transitions in the history of life such as, for example, the depletion of organic molecules from the environment [24,27]. The retrograde model also fails to explain the development of pathways that include labile metabolites, which could not have accumulated in the environment for long enough for retrograde recruitment to take place.

The patchwork model

In light of these limitations, Ycas [25] and Jensen [26] proposed the patchwork model of metabolic pathway evolution, in which pathway evolution depends on the initial existence of broad-specificity enzymes. In its original formulation [25], such enzymes catalyze whole classes of reactions, forming a large network of possible pathways. The broad specificities would mean that many metabolic chains, synthesizing key metabolites, may have existed, although short and incomplete compared with the pathways observed today. The duplication of genes in such pathways (advantageous because increased levels of the enzyme would generate more of the key metabolites), followed by their specialization, would account for extant pathways. Jensen [26] subsequently pointed out that the fortuitous evolution of a novel chemistry, together with the biological leakiness of such a system, could allow the production of a key metabolite from a novel intermediate, even if it is several enzymatic steps away from the original product.

external and internal sensors to respond dynamically to environmental changes [17]. Integration of the external environment with metabolism occurs through the import of compounds from the environment and results, for example, in a transcriptional response or an allosteric interaction with an enzyme [18-20]. In the second of the recent studies from Ruppin and co-workers, Borenstein *et al.* [4] propose a graph-theoretical approach to define these exogenously acquired compounds - the seed set of an organism - and have identified their repertoire across the tree of life (Figure 1b). This is one of the most comprehensive studies so far that links organisms' metabolic circuitry with their environment.

The authors represent the metabolic network of a given species as a directed graph with nodes representing metabolites and edges corresponding to the linking reactions converting substrates to products. Using this, they identify the maximal set of metabolites that can be synthesized from a particular precursor metabolite. This graph-based representation of the metabolic network then enabled them to discover the seed-set compounds for each of the 478 prokaryotic species with available metabolic networks in the KEGG database [12]. On the whole, they found that about 8-11% of the compounds in the metabolic network of an organism correspond to the seed set. Their predictive ability to correctly identify seed compounds reached a precision of 95% when benchmarked against a set of compounds experimentally characterized as being taken up from the environment by the rickettsia that cause the disease ehrlichiosis in humans and animals. Recall values (defined as the percentage of correctly identified seeds of all exogenously acquired compounds) based on the same dataset were low, suggesting that other factors might have a role in the identification of seed compounds of an organism, such as

the incompleteness of the metabolic network or ways of acquiring an exogenous compound that cannot be captured by currently available metabolic maps. The resulting compilation, which represents the overall static metabolic interface of each organism characterizing its biochemical habitat, enabled Borenstein *et al*. to trace the evolutionary history of both metabolic networks and growth environments.

When the seed sets identified in each organism were analyzed in detail, species living in variable environments were found to have more versatile seed sets, in terms of variability of size and diversity of composition. On the other hand, obligate parasites like *Buchnera aphidicola* and those microorganisms, such as archaea, that live in extreme and narrowly defined environments, were found to have much smaller seed set sizes. These results suggest that although organisms surviving in predictable environments can take up many compounds from their surroundings, this capability is still significantly smaller than in organisms that have to survive in a wide range of niches.

Borenstein *et al*. [4] carried out a phylogenetic analysis of the seed sets across different taxa, which suggested not only that an accurate tree of life can be reconstructed from them but that such a tree can provide insights into the evolutionary dynamics of seed compounds. In particular, the study revealed that novel compounds can be integrated into the metabolic network of an organism as either non-seeds or seeds, and that seed compounds are more likely to be lost during evolution than non-seed compounds. From the comparison with ancestral metabolic networks, Borenstein *et al*. [4] suggest that the transition from seed to non-seed compound occurs 2.5 times more often than the reverse. This suggested that, of the two main current hypotheses of metabolic network evolution - the 'patchwork' and 'retrograde' models (see Box 1) - the retrograde model, in which pathways evolve in a direction opposite to the metabolic flow, might best explain the observed events. However, the observations of Borenstein *et al*. [4] on the high overall rate of integration of non-seed compounds and the relatively high rate of transition of non-seed compounds into seed metabolites, suggest that some aspects of network evolution could be explained by the patchwork and other models. The results highlight the fact that these models are not mutually exclusive, but complementary, and might have contributed to pathway evolution to different extents [21,22].

It should be noted that there are limitations to studies such as those reported here, in that the incompleteness of metabolic maps, the reversibility of reactions, possible alternative mechanisms controlling metabolic import, and the ignoring of the distinction between catabolic and anabolic pathways can all potentially result in false positives in the identified seed sets. Nevertheless, it is exciting to note that seed sets obtained using the approach developed in these studies not only reflect the metabolic environments of the species themselves but also provide insight into their natural biochemical habitats - the union of all the metabolic environments an organism encounters.

Hence, such approaches can be exploited to study the interaction and association of microbes with other species thriving in similar habitats. This may help in the identification of host-parasite and symbiotic relationships between organisms and also enable the prediction and design of drugs that can precisely target an organism of interest without adversely affecting the host. With the availability of metagenomic data ranging from viromes to biomes [23], we anticipate that similar approaches can be applied to study metagenomic environments to decipher species relationships and dependencies occurring in large ecological niches, thereby providing insights into ecological imbalances or tradeoffs.

Acknowledgements

SCJ and MMB acknowledge financial support from the MRC Laboratory of Molecular Biology. SCJ acknowledges financial support from Cambridge Commonwealth Trust. MMB thanks Darwin College and Schlumberger Ltd for generous support. We thank A Wuster, R Janky, K Weber, V Espinosa-Angarica and JJ Díaz-Mejía for critically reading the manuscript and providing helpful comments.

References

1. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO: **Metabolic pathways in the post-genome era.** *Trends Biochem Sci* 2003, **28:**250-258.
2. Feist AM, Palsson BO: **The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*.** *Nat Biotechnol* 2008, **26:**659-667.
3. Kreimer A, Borenstein E, Gophna U, Ruppin E: **The evolution of modularity in bacterial metabolic networks.** *Proc Natl Acad Sci USA* 2008, **105:**6976-6981.
4. Borenstein E, Kupiec M, Feldman MW, Ruppin E: **Large-scale reconstruction and phylogenetic analysis of metabolic environments.** *Proc Natl Acad Sci USA* 2008, **105:**14482-14487.
5. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules.** *Proc Natl Acad Sci USA* 2003, **100:**15428-15433.
6. Spirin V, Gelfand MS, Mironov AA, Mirny LA: **A metabolic network in the evolutionary context: multiscale structure and modularity.** *Proc Natl Acad Sci USA* 2006, **103:**8774-8779.
7. Guimera R, Nunes Amaral LA: **Functional cartography of complex metabolic networks.** *Nature* 2005, **433:**895-900.
8. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268:**1803-1810.
9. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.
10. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
11. Newman ME: **Modularity and community structure in networks.** *Proc Natl Acad Sci USA* 2006, **103:**8577-8582.
12. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Res* 2008, **36(Web Server issue):**W423-W426.
13. Parter M, Kashtan N, Alon U: **Environmental variability and modularity of bacterial metabolic networks.** *BMC Evol Biol* 2007, **7:**169.
14. Pal C, Papp B, Lercher MJ: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.** *Nat Genet* 2005, **37:**1372-1375.
15. Zhao J, Ding GH, Tao L, Yu H, Yu ZH, Luo JH, Cao ZW, Li YX: **Modular co-evolution of metabolic networks.** *BMC Bioinformatics* 2007, **8:**311.

16. Vitkup D, Kharchenko P, Wagner A: **Influence of metabolic network structure and function on enzyme evolution.** *Genome Biol* 2006, **7:**R39.
17. Martinez-Antonio A, Janga SC, Salgado H, Collado-Vides J: **Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*.** *Trends Microbiol* 2006, **14:**22-27.
18. Seshasayee AS, Fraser GM, Babu MM, Luscombe NM: **Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*.** *Genome Res* 2008. doi: 10.1101/gr.079715.108.
19. Balaji S, Babu MM, Aravind L: **Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli*.** *J Mol Biol* 2007, **372:**1108-1122.
20. Janga SC, Salgado H, Martinez-Antonio A, Collado-Vides J: **Coordination logic of the sensing machinery in the transcriptional regulatory network of *Escherichia coli*.** *Nucleic Acids Res* 2007, **35:**6963-6972.
21. Diaz-Mejia JJ, Perez-Rueda E, Segovia L: **A network perspective on the evolution of metabolism by gene duplication.** *Genome Biol* 2007, **8:**R26.
22. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **Small-molecule metabolism: an enzyme mosaic.** *Trends Biotechnol* 2001, **19:**482-486.
23. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F: **Functional metagenomic profiling of nine biomes.** *Nature* 2008, **452:**629-632.
24. Horowitz NH: **On the evolution of biochemical syntheses.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.
25. Ycas M: **On earlier states of the biochemical system.** *J Theor Biol* 1974, **44:**145-160.
26. Jensen RA: **Enzyme recruitment in evolution of new function.** *Annu Rev Microbiol* 1976, **30:**409-425.
27. Lazcano A, Miller SL: **On the origin of metabolic pathways.** *J Mol Evol* 1999, **49:**424-431.

Transcript stability in the protein interaction network of *Escherichia coli*

Sarath Chandra Janga* and M. Madan Babu

Received 25th September 2008, Accepted 21st November 2008
First published as an Advance Article on the web 9th December 2008
DOI: 10.1039/b816845h

Gene expression is a dynamic process which can be controlled by a number of mechanisms as genetic information flows from nucleic acids to proteins. The study of gene expression in the steady state, while informative, overlooks the underlying dynamics of the processes. Steady-state transcript levels are a result of both RNA synthesis and degradation, and as such, measurements of degradation rates can be used to determine their rates of synthesis as well as reveal regulation that occurs *via* changes in RNA stability. Messenger RNA degradation plays a central role in diverse cellular processes and is controlled primarily by the activity of the degradosome in prokaryotes. In this study, we use the currently available network of protein–protein interactions (PPIs) and mRNA half-lives in *Escherichia coli* to demonstrate that centrality of a protein in the PPI network is strongly correlated with its mRNA half-life. We find that interacting proteins tend to show similar half-lives, commonly referred to as assortative behavior in networks, which is frequently found in biological and social networks. While a major fraction of the interacting proteins show significantly lower differences in mRNA stabilities, a smaller but significant number of protein pairs tend to show higher differences than expected by chance. Higher differences in transcript stabilities often involved those that encode for transcription factors and enzymes, suggesting a feedback link at the post-translational level. We also note that although essential genes, which act as a proxy for *in vivo* centrality in PPI networks, are highly expressed compared to non-essential ones, they do not encode for more stable transcripts than non-essential genes. Our results provide a direct link between mRNA stability and centrality of a protein in PPI network indicating the importance of post-transcriptional mechanisms on nascent RNAs in the cell.

Introduction

RNAs can be classified by their stability in the cell. The best-known stable RNAs are the tRNAs and rRNAs. mRNAs are unstable, with half-lives in *Escherichia coli* ranging from 2 to 25 min (see Fig. 1A). In eukaryotic cells, mRNA turnover is slower, but the half-lives are usually shorter than the generation time. The instability of mRNA is an important property permitting timely adjustments to changes in growth conditions or to genetically controlled programs of expression. Until recently, tRNAs and rRNAs were believed to be protected by their rapid folding and assembly into compact structures. This simplistic view seems unlikely because of the discovery of ribonucleolytic multienzyme complexes capable of unwinding and degrading structured RNA. Another widely held preconception was that the enzymes involved in the processing of stable RNA would be distinct from those involved in the degradation of mRNA. With the discovery in *E. coli* and *Saccharomyces cerevisiae* that ribonucleases involved in the processing of rRNA are also important in the degradation of mRNA, it is now clear that there is a close connection between processing and degradation.[1–4]

mRNA instability is an intrinsic property that permits timely changes in gene expression by limiting the lifetime of a transcript and acts as a regulator for controlling the production of a protein product at the post-transcriptional level. It is becoming increasingly clear that in eubacteria like *E. coli*, RNase E, a single-strand-specific endonuclease is involved in the processing of rRNA and the degradation of mRNA.[5–7] A nucleolytic multienzyme complex now known as the RNA degradosome was discovered during the purification and characterization of RNase E.[8,9] Two other major components of this complex include a 3′ exoribonuclease (polynucleotide phosphorylase, PNPase) and a DEAD-box RNA helicase (RNA helicase B, RhlB). RNase E is a large multidomain protein with N-terminal ribonucleolytic activity, an RNA-binding domain and a C-terminal 'scaffold' that binds PNPase, enolase and RhlB. The association of RNase E and PNPase in a complex provides a direct physical link for their co-operation in the degradation of mRNA. Other associated proteins, present in substoichiometric amounts, include polyphosphate kinase (PPK), DnaK and GroEL. Interactions with other enzymes, such as *E. coli* poly(A) polymerase and the ribosomal protein S1, have also been described, although the role of enolase, PPK and other associated proteins in the degradation of mRNA is still unknown.[7] However, a 'minimal' degradosome containing RNase E, RhlB, PNPase and enolase can be reconstituted from purified components and has been proposed to comprise the degradosome complex.

In *E. coli*, the degradation of mRNA is mediated by the combined action of endo- and exo-ribonucleases, RNase E and PNPase, respectively, which degrade RNA in a 3′–5′

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 0QH. E-mail: sarath@mrc-lmb.cam.ac.uk; Fax: +44 (0)1223 213556; Tel: +44 (0)1223 402479

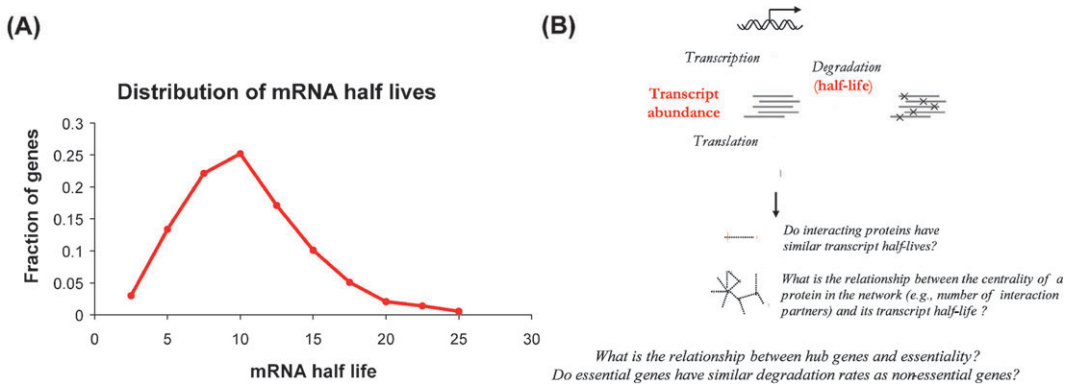


Fig. 1 Schematic showing the (A) distribution of mRNA half-lives (in minutes) for all the protein coding genes in *E. coli* analyzed in this study. (B) Concept of mRNA stability, protein–protein interaction (PPI) network and its relationship with *in silico* and *in vivo* centrality and essentiality measures addressed in this study.

pathway.[10] Enzymes related to RNase E and PNPase are widespread in both eubacteria and eukaryotes.[11] Recent studies have shown the existence of endonuclease binding proteins like RraA and RraB that can modulate the remodelling of degradosome composition in bacteria and can result in dramatic, distinct, and inhibitor-specific changes in degradosome composition. These effects have also been shown to be associated with alterations in RNA decay and global transcript abundance profiles. These profiles were found to be dissimilar to those observed during simple RNase E deficiency, and such effects have been suggested to make degradosome remodelling as a mechanism for the differential regulation of RNA cleavages in *E. coli*.[12,13] In addition, recent whole genome microarray studies have revealed the importance of the contribution from different components of the degradosome, the relation between mRNA stability and its abundance and the higher order cleavage characteristics implying the importance of mRNA stability and the role of post-transcriptional regulation in mediating cellular interactions and cross-talk.[14–17]

Two key factors for controlling the concentration of a protein in a bacterial cell include the number of transcripts per cell cycle and the stability of the transcript. Evidence points to the fact that most transcripts in bacteria and eukarya are produced only once per cell cycle suggesting that stability of a transcript during the cell cycle might play a more important role than the actual number of mRNA molecules, which is already low.[18] As the transcription rate is generally low, it follows that cells must depend on their mother's mRNAs and/or proteins for survival. Therefore, transcript half-life might enforce a constraint on transcripts that have critical roles in important cellular processes, which may take place throughout the cell cycle or longer. On the other hand, smaller cellular sizes in bacteria would force the infrequently used transcripts to be rapidly decayed. In fact, it has been shown in the *E. coli* transcriptional regulatory network that highly connected transcription factors tend to be less stable with short half-lives although they are highly expressed,[19,20] indicating that stability of a transcript might play a vital role in several cellular processes. Given these observations, it is imperative to understand how stability of a transcript can constrain the interaction of its protein product with other cellular components. With the availability of data from high-throughput technologies like affinity purification and two hybrid system, it has become possible to address such questions on large-scale PPI maps.[21–23] In this study, we use for the first time the PPIs of a bacterial model organism, *E. coli* from the database of interacting proteins (DIP)[24] and ask how the *in silico* and *in vivo* measures of centrality of a protein are related to the stability of its transcript (Fig. 1B). *In silico* centrality of a protein in a PPI network refers to its number of connected neighbors and other network based topology measures which indicate its importance while the *in vivo* centrality indicates whether a protein is essential for survival in specific experimental conditions.

Results and discussion

mRNA half-lives of proteins correlate positively with their PPI network centrality

The control of mRNA degradation plays a central role in diverse cellular processes and is regulated primarily by the activity of the degradosome in prokaryotes.[25–27] mRNA decay has been studied in a range of organisms, and much has been learned about the substrate features and ribonucleolytic enzymes that influence mRNA stability based on data from small sets of transcripts.[27,28] However, due to the availability of DNA microarrays, it has recently become possible to screen and measure the mRNA levels of transcripts of thousands of individual genes, enabling the determination of mRNA abundance and stability on a genome-wide scale. A common strategy used to determine mRNA half-lives is to block new transcription and monitor expression levels of transcripts over a period of time to obtain rates of decay of individual transcripts using a microarray. Typically, rifampicin, a drug known to prevent the initiation of new transcripts by binding to the β subunit of RNA polymerase is used.[15,16] In this study, we used the repertoire of mRNA half-lives determined by Selinger *et al.*[16] in *E. coli* (see Materials and methods). The data on PPIs were obtained from the DIP, which contains a high quality set of interactions[24] (see Materials and methods). The final PPI network consisted of 5667 interactions involving 998 proteins, encompassing about 25% of the predicted proteome of *E. coli*.

By integrating the data on mRNA half-life and protein interaction network, we first asked if there is a correlation between the mRNA half-life of the transcript and the importance of a protein (encoded by the transcript) in the protein interaction network. To obtain the importance of a protein in the PPI network, we calculated different centrality measures[29] for every protein in the network as described in the Materials and methods. In brief, three centrality measures have been described in the literature: (i) degree or connectivity, which is the number of interactions a protein has in the PPI network—the higher the connectivity (*i.e.*, hub nodes) the more important a protein is, (ii) betweenness centrality, which measures the number of shortest path lengths between all pairs of proteins in the network that pass through a protein of interest—the higher the number of paths that pass through a protein, the more important it is, (iii) closeness centrality, which provides the average length of all the shortest paths from a protein of interest to all other proteins in the network—note that closeness centrality defined this way implies that lower the closeness value, the higher the importance (centrality) of a node. We used all of these parameters measuring the importance of a protein in a PPI network and compared them against the mRNA half-lives of the encoding transcripts. As shown in Fig. 2, we found that all the centrality measures correlate positively with the mRNA half-lives, indicating that the transcripts of the highly central proteins in the PPI network tend to be more stable in the cell. It should be noted that the results presented here are insensitive to the removal of up to 10% of the interactions in the network suggesting that the findings presented here are generally robust (see Materials and methods; data not shown). This implies that proteins that are more central in the PPI network of *E. coli* (*e.g.*, hubs—those which interact with a large number of other proteins) tend to have much more stable transcripts in order to enable their availability for most of the cell cycle. This might ensure that proteins that are important to co-ordinate cellular activity by interacting with several other proteins can be synthesized from their corresponding transcripts in required concentrations at different times. It is interesting to note that the finding reported here is in contrast to what is observed for hubs in the transcriptional network of *E. coli*[19] (*i.e.*, transcription factors which regulate the expression of several genes) possibly as a result of different functional constraints and mechanisms governing the roles for hubs in different networks. In other words, since hubs in the transcriptional network have to be transcription factors only (which regulate gene expression by binding to upstream region) while hubs in the PPI network can be any protein that need not be a transcription factor, this may introduce very different constrains for transcript stability for hubs in the two networks.

In addition to these centrality measures, we also computed the clustering coefficient of a node, which reflects the extent to which the neighbors of a given node are interconnected among themselves and indicates the cohesiveness or local modularity of the network. It is interesting to note from Fig. 2D that half-lives of these very stable transcripts are inversely correlated with their clustering coefficient implying that highly stable transcripts may not form cohesive local modules in the interaction network. This result also suggests that highly connected nodes in PPI network may not form part of any particular

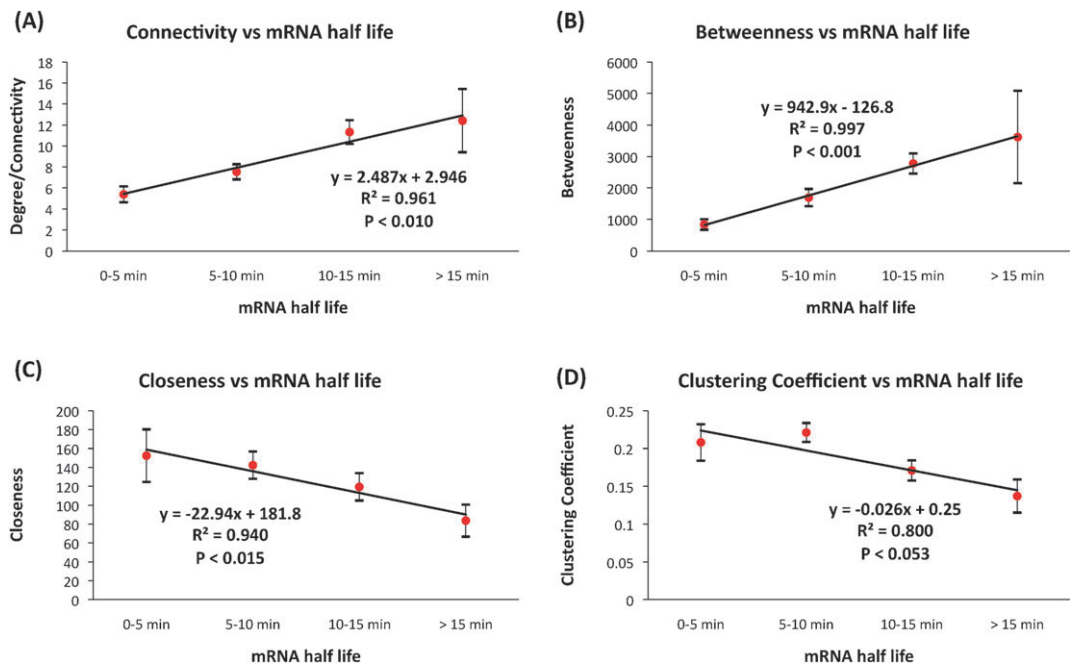


Fig. 2 Relationship between network properties of proteins in the PPI network and their corresponding mRNA stability measured as the transcripts half-life (A) degree of a node in the PPI network *versus* its mRNA half-life, (B) betweenness of a node *versus* mRNA half-life, (C) closeness of a node *versus* its mRNA half life and (D) clustering coefficient of a node *versus* its mRNA half-life. All the centrality measures indicate that proteins with high centrality tend to exhibit high mRNA half-lives. Clustering coefficient of highly stable nodes in the PPI network decreases, indicating that although central proteins are more stable, they may not form multi-protein assemblies. Error bars are shown in each case to show the extent of variation of the network property in each bin. *p*-values correspond to the significance level of the correlations.

module but rather might be involved in multiple modules, due to the hierarchical nature of biological systems previously demonstrated for metabolic networks.[30]

mRNA half-lives of interacting proteins tend to have similar degradation rates with few pairs showing significant differences in half-lives

We then investigated if the transcripts of interacting proteins in the PPI network tend to have similar or dissimilar half-lives. To investigate this question, we calculated the "assortativity value" for the PPI network based on the half-life values associated with a node by adapting the formula described by Newman[31,32] (see Materials and methods). As originally described by Newman,[31,32] assortativity value is a single global measure that tries to capture the dominant type of interaction in a network. For instance, positive assortativity value (assortative mixing) based on the degrees of a node in a network would mean that there is a preference for interacting nodes to have a similar degree, while negative assortativity value (disassortative mixing) would imply that there is a preference for interacting nodes to have dissimilar degree (*e.g.*, high-connectivity nodes interacting with low-connectivity ones). Negative assortative values have been shown to correspond to scale-free graphs like that of world-wide web, internet and protein interaction networks with values ranging from -0.06 to -0.18.[31] We calculated the assortativity value based on the mRNA half-life (rather than the degree) of a protein in the PPI network and found a low but positive value of 0.03, suggesting that proteins which interact with each other tend to have comparable half-lives.

To investigate this observation in more detail, we computed the differences in half-lives of interacting proteins and compared the distribution with that observed from randomly selected pairs of proteins as described in Materials and methods. As a result of this analysis (Fig. 3), we found that

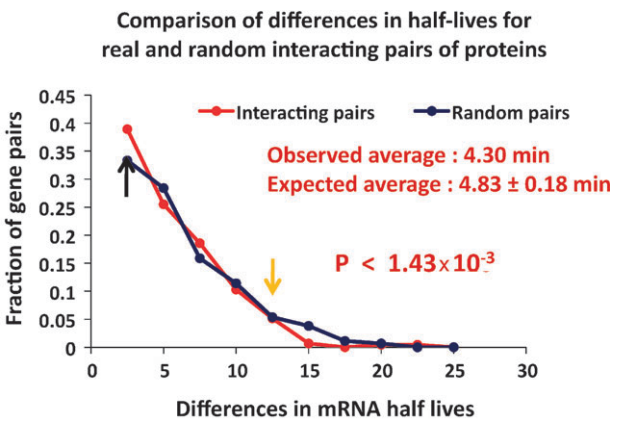


Fig. 3 Distribution of the differences in mRNA half-lives of interacting proteins compared against random pairs of proteins, indicating that interacting protein pairs exhibit a higher tendency to have lower differences in half-lives. Marked in black and yellow arrows are the half-life thresholds where interacting protein pairs show significantly lower and higher differences in half-lives, respectively. Very few interacting protein pairs showed high differences in half-lives as shown in Table 1.

in general interacting proteins tend to show lower differences in half-lives than what is expected by chance ($P \leq 1.43 \times 10^{-3}$), although a small fraction of interacting proteins did show high differences in half-lives (Table 1). This calculation shows that interacting proteins on an average have a variation in half-life of about $\sim$4 minutes with a vast majority of them falling in the difference range of $<$3 minutes (marked with a black arrow in Fig. 3). Likewise, very few interacting proteins were found to show high differences in half-lives (threshold value of $\sim$13 minutes above which a much smaller fraction of interacting pairs was found compared to random pairs, marked with a yellow arrow in Fig. 3) and they corresponded to interactions between and among regulators and enzymes involved in global regulatory processes (see Table 1). Sensitivity analysis to test the robustness of the results indicated that the results are reproducible with networks where up to 10% of the interactions are randomly removed (see Materials and methods; data not shown). An analysis of the function of interacting proteins that show large differences in half-lives reveals that some highly stable transcripts belong to the enzyme or regulator functional classes. These might be involved in PPIs as a means of linking cellular processes as diverse as metabolism, replication, repair and regulation. One possible explanation for such large differences in stabilities might be the usage of the stable transcripts (which typically encode for hubs) as feedback controllers at different stages of the cell cycle.

Since it has been known that gene duplication is an important mechanism for genome evolution which has also contributed to the growth of the PPI networks, we investigated if duplicate genes have similar half-lives. By investigating the sequences of the proteins in the network, we found that only 2% of the interacting protein pairs (108 of 5667 interactions) are composed of duplicated protein partners at a BLAST e-value threshold of $1e^{-5}$. Of these, we had half-life data for 39 pairs and they did not show any significant tendency for high or low differences in half-lives compared to overall distribution (data not shown). Since the fraction of duplicated genes and the corresponding interactions in our network is very low and does not show any inherent trends, our results were robust to removal of these duplicate proteins. Varying the BLAST e-value thresholds by 2 orders of magnitude to detect duplicate genes did not change our end results (data not shown). However, as more data on protein interaction networks become available, it should be possible to address this question in greater detail.

Hubs in PPI network tend to be essential although transcripts of essential genes are not more stable than non-essential ones

While the importance of a protein can be assessed by measuring the centrality of a node in the PPI network, it can also be inferred by experimentally testing if removal of the gene renders a cell lethal or not. To complement our understanding of the relationship between the *in silico* centrality of a node in the PPI network against the stability of a transcript from an experimental perspective, we investigated the following questions: do the experimentally determined essential genes tend to be important proteins in the PPI interaction network? Are

Table 1 Interacting protein pairs exhibiting highest differences in their mRNA half-lives. Most of these interactions are between or among regulatory factors and enzymes suggesting that these interactions contribute to the cellular integrity by linking different cellular processes and pathways with the core machinery of the cell. All interacting pairs which showed high differences in half-lives and are under-represented in proportion with increasing differences in half-lives threshold compared to random pairs, are shown below (see Fig. 3 for additional details)

Gene 1 (higher half-life)	Function	Gene 2 (lower half-life)	Function	Difference (min)
hupA|*b4000*	Factor; basic proteins-synthesis, modification; HU, DNA-binding transcriptional regulator, alpha subunit	*galR*|*b2837*	Regulator; degradation of small molecules: Carbon compounds; DNA-binding transcriptional repressor	18.1
rpoD|b3067	Factor; global regulatory functions; RNA polymerase, sigma 70 (sigma D) factor	*crp*|*b3357*	Regulator; global regulatory functions; DNA-binding transcriptional dual regulator	13.9
pflB|*b0903*	Enzyme; energy metabolism, carbon: Anaerobic respiration; pyruvate formate lyase I	*yjeE*|b4168	ATPase with strong ADP affinity	20.9
ybdN|b0602	Conserved protein	*sspB*|*b3228*	Regulator; global regulatory functions; ClpXP protease specificity-enhancing factor	14.6
recG|*b3652*	Enzyme; DNA-replication, repair, restriction/modification; ATP-dependent DNA helicase	*ssb*|b4059	Factor; DNA-replication, repair, restriction/modification; single-stranded DNA-binding protein	21.6
sgbH|b3581	Putative enzyme; central intermediary metabolism: pool, multipurpose conversions; 3-keto-L-gulonate 6-phosphate decarboxylase	*nudF*|*b3034*	ADP-ribose pyrophosphatase	13
hupA|*b4000*	Factor; basic proteins-synthesis, modification; HU, DNA-binding transcriptional regulator, alpha subunit	*nudH*|b2830	Putative factor; not classified; nucleotide hydrolase	18.2

transcripts of essential genes more stable or highly expressed compared to the non-essential genes? To address these questions, we obtained a list of essential and non-essential genes in the *E. coli* genome from a recent study conducted by Mori and co-workers.[33]

To address the first question, we integrated essentiality data with the connectivity data of proteins in the PPI network. In particular, we compared the proportion of essential genes in different connectivity bins (see Materials and methods), after dividing the proteins in the PPI network into three different groups based on their degree, D: low ($D < 9$), intermediate ($9 \leq D < 44$) and high ($D \geq 44$). Fig. 4A shows the proportion of essential genes for each of the three groups of proteins. We note that the bin with the highly connected proteins also has a higher proportion of essential genes, while the bin with lowly connected proteins shows a depletion in the fraction of essential genes, indicating that essentiality (a qualitative measure of *in vivo* centrality) of a gene correlates with its degree in the PPI network, similar to what has been observed in the yeast PPI network.[34] We found that a total of 173 essential genes (58% of all essential genes) formed part of these bins (82, 64 and 27 in the low, intermediate and high-connectivity bins, respectively) signifying that most of the essential genes in *E. coli* also form part of the PPI network. Note that although the absolute number of essential genes in highly connected bin is low, it overlaps with a significant fraction of the hubs (*i.e.*, proteins with degree ≥ 44).

To address the second question on the relationship between essentiality and transcript stability, we integrated the gene essentiality data with mRNA half-life and expression data. Though it is commonly believed that essential genes which comprise core proteins necessary for the survival of the cell need to be expressed in higher concentrations,[35,36] this has not been tested so far. Thus we investigated if essential genes would be highly expressed compared to non-essential genes and how hubs (proteins with degree ≥ 44) in the interaction network compare in their expression with respect to essential genes. It should be noted that though hubs tend to be essential genes, not all essential genes are hubs. Hence it becomes important to make this distinction and test them independently. Fig. 4B shows the expression levels of essential and non-essential genes in *E. coli* along with hubs in the interaction network using most of the publicly available expression data generated on the affymetrix platform (see Materials and methods).[37] Both hubs and essential genes were found to be significantly more highly expressed than non-essential genes (t-test, $p < 2.2 \times 10^{-16}$) and hubs were found to be more highly expressed than essential genes ($p < 8.87 \times 10^{-10}$). Given these observations that both hubs and essential genes are significantly more abundant than non-essential genes at the mRNA level, we asked whether this difference is also reflected in the stability of their transcripts. To address this, we compared their mRNA half-lives and found that although the transcripts of essential genes are not more stable than non-essential genes ($p < 0.057$), transcripts encoding hubs exhibited significantly higher stabilities compared to those encoding non-essential genes ($p < 8.9 \times 10^{-5}$) (Fig. 4C). These results suggest that although essential genes are highly expressed, they do not tend to be more stable than non-essential genes. These observations together with biological processes enriched in these classes of genes (see Table 2) suggested that the differences may stem due to the nature of proteins with distinct functions in the two groups. While hubs predominantly encode for proteins involved in translation and protein synthesis, essential gene set comprises genes belonging to various metabolic and biosynthetic processes important for cellular growth. Thus the higher abundance of essential genes compared to non-essential genes may be a result of higher transcription of essential genes rather than increased stability

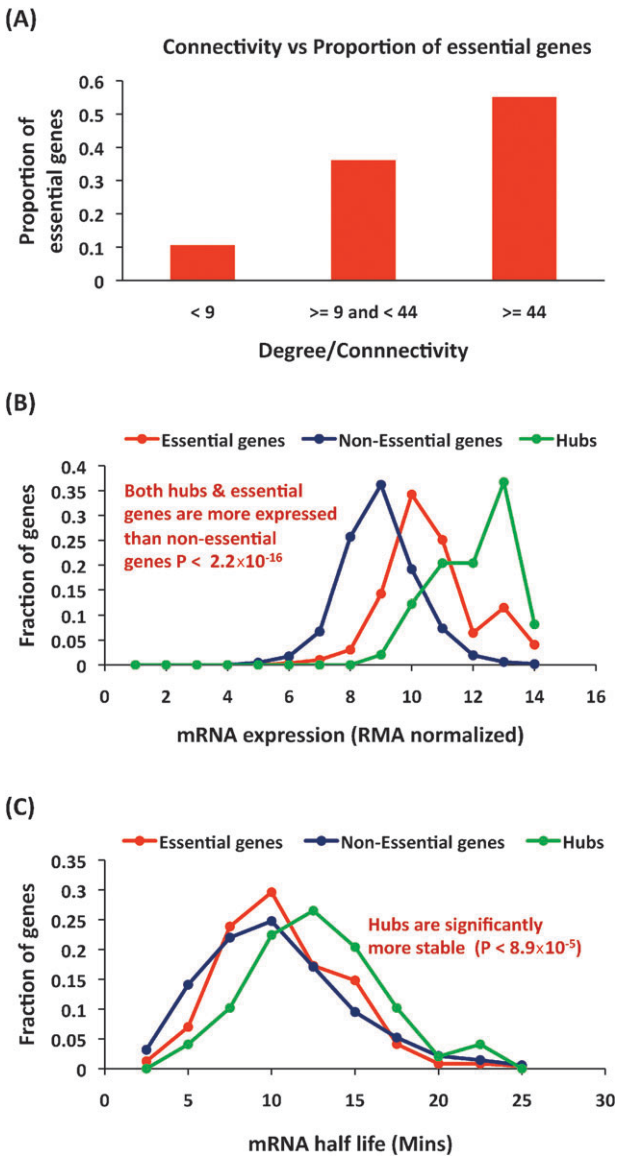


Fig. 4 (A) Hubs tend to be essential in the protein interaction network of *E. coli* similar to what has been observed in yeast.[34] Each bin corresponds to the range of the degree of a protein in the PPI network and shows the proportion of essential genes in the bin on the *Y*-axis. (B) Although both hubs and essential genes were found to be significantly more expressed than non-essential genes ($p < 2.2 \times 10^{-16}$), when their expression levels were compared using the publicly available microarray compendium for more than 400 microarray experiments performed on *E. coli*, (C) only hubs show a tendency to be more stable at the mRNA level compared to non-essential genes ($p < 8.9 \times 10^{-5}$ comparing the stability of hubs against non-essential genes *versus* $p < 0.057$ for essential genes against non-essential genes). *p*-values were calculated using the *t*-test function in the R statistical package.

of their transcripts. Since cells prefer to degrade transcripts encoding essential genes in the same way as other non-essential genes, this also suggests that increased transcript stability of essential genes, resulting in very high levels of essential genes, might be generally unfavorable or expensive for the cell to have them present for a longer time.

Materials and methods

Data on mRNA half-lives, gene essentiality and PPI network in *E. coli*

Half-life of RNA acts as a direct measure of transcript stability and is frequently used for measuring the stability of messenger RNAs. mRNA half-lives of protein coding genes in *E. coli* were obtained from a previous study where the authors analyzed the global patterns of RNA degradation on a genome-wide scale using high-density, subgenic-resolution oligonucleotide microarrays.[16] Half-life data could be obtained for a total of 2680 genes whose distribution (in minutes) is shown in Fig. 1A. Data on gene essentiality were obtained from a recent genome-wide knock out study, where the authors generated a whole genome single gene knock library.[33] We collected a total of 298 genes in *E. coli K12* which were reported to be lethal according to this study. The remaining genes were considered as non-essential. Manually curated and high quality data on PPIs in *E. coli* were obtained from the DIP,[24] which included data from traditional independent studies and high-throughput studies such as Butland *et al.*[21] Our final dataset used for this study consisted of 5667 PPIs with 49 proteins qualifying as hubs (top 5% of the nodes with highest connectivity). Since not all the genes had half-life data associated with it, only those set of interactions for which both the genes had half-life information available were considered for studying the differences in half-lives. This subset composed of 447 interacting protein pairs.

Network properties of the proteins in the PPI network

To study the properties of the PPI network and their dependence on mRNA half-life, we used *igraph*, a publicly available R package for analyzing graphs [see http://cneurocvs.rmki.kfki.hu/igraph/ and http://www.r-project.org]. In particular, since the network of PPIs analyzed in this study is undirected, we used the corresponding versions of the functions: degree, transitivity, betweenness and closeness for calculating the degree, clustering coefficient, betweenness and closeness centralities of a node. Betweenness centrality, which is the number of shortest paths going through a node, was calculated using the brandes algorithm[38] implemented in R. Similarly, closeness, measured as average length of the shortest paths to all the other vertices in the graph, was obtained using the implementation in R. Since the centrality measures, betweenness and closeness use the shortest path lengths between all pairs of nodes in a graph, for cases where no path exists between a particular pair of nodes, shortest path length was taken as one less than the maximum number of nodes in the graph. Note that this is also the default assumption for calculating centrality measures in *igraph*. Hubs were defined as the nodes with degrees greater than two standard deviations above average degree of a node in the PPI network (*i.e.*, degree $\geq$ 44), while poorly connected nodes were defined as nodes with degrees less than average degree of a node in the network (*i.e.*, degree $<$ 9). To assess the robustness of the results, we have performed sensitivity analysis by randomly removing 10% of the PPI network in 10 independent trails and calculated the network properties and other results reported in

Table 2 Gene ontology biological processes enriched in hubs and essential genes along with their significance values. *p*-values were calculated using BINGO,[46] a JAVA-based tool for calculating predominant functional categories in a collection of genes. *p*-values were corrected for multiple testing using the same package, at a false discovery rate (FDR) of 0.05 and only those less than 1×10^{-3} are shown. Hub class which comprised 49 proteins showed enrichment for only translation and protein metabolism while the essential-gene class comprising 243 genes showed enrichment for many other metabolic and biosynthetic processes

Biological process ontology	*p*-Value	Corrected *p*-value
Enriched biological processes in hubs		
Translation	1.9492×10^{-9}	2.0077×10^{-7}
Protein metabolic process	1.4225×10^{-6}	4.8841×10^{-5}
Cellular protein metabolic process	1.4225×10^{-6}	4.8841×10^{-5}
Cellular macromolecule metabolic process	8.6460×10^{-6}	2.2263×10^{-4}
Gene expression	2.1656×10^{-5}	4.4611×10^{-4}
Enriched biological processes in essentials		
Cellular biosynthetic process	5.4841×10^{-13}	1.6672×10^{-10}
Translation	5.2333×10^{-12}	7.9547×10^{-10}
Oxidoreduction coenzyme metabolic process	7.8414×10^{-9}	7.9460×10^{-7}
Coenzyme biosynthetic process	2.4621×10^{-8}	1.6639×10^{-6}
Biosynthetic process	2.7367×10^{-8}	1.6639×10^{-6}
Cellular macromolecule metabolic process	1.5135×10^{-7}	7.6686×10^{-6}
Protein metabolic process	3.4365×10^{-7}	1.3059×10^{-5}
Cellular protein metabolic process	3.4365×10^{-7}	1.3059×10^{-5}
Cofactor biosynthetic process	4.5055×10^{-7}	1.5219×10^{-5}
Coenzyme metabolic process	1.1352×10^{-6}	3.4511×10^{-5}
Ubiquinone biosynthetic process	3.1663×10^{-6}	8.0213×10^{-5}
Ubiquinone metabolic process	3.1663×10^{-6}	8.0213×10^{-5}
tRNA aminoacylation	1.0490×10^{-5}	2.1260×10^{-4}
Amino acid activation	1.0490×10^{-5}	2.1260×10^{-4}
tRNA aminoacylation for protein translation	1.0490×10^{-5}	2.1260×10^{-4}
Cofactor metabolic process	1.6198×10^{-5}	3.0776×10^{-4}
tRNA metabolic process	1.7515×10^{-5}	3.1322×10^{-4}

this study in each case. We found that the general observations were consistent in each run indicating that incompleteness of the network or the existence of false positives is unlikely to affect the findings presented here (data not shown).

Calculation of assortativity value for the mRNA half-lives associated network of PPIs

To calculate the assortativity value, r, for the PPI network with mRNA half-lives associated to nodes, we used the formula defined by Newman[31,32] as below, wherein the variables j_i and k_i were substituted for the mRNA half-lives of the interacting proteins of the ith edge with i varying from 1 to M, which stands for the total number of edges or interactions in the network. The range of r is the closed interval $[-1, 1]$ with positive values corresponding to assortative behavior while negative values suggest disassortativity of the network.

$$r = \frac{\left(\frac{1}{M}\right)\sum_i j_i k_i - \left[\frac{1}{M}\sum_i \left(\frac{1}{2}\right)(j_i + k_i)\right]^2}{\left[\left(\frac{1}{M}\right)\sum_i \left(\frac{1}{2}\right)(j_i^2 + k_i^2)\right] - \left[\left(\frac{1}{M}\right)\sum_i \left(\frac{1}{2}\right)(j_i + k_i)\right]^2}$$

Estimating the significance in the differences of observed half-lives of interacting proteins

To assess the significance of the observed differences in mRNA half-lives of interacting proteins, we compared the average value of the observed differences against the same value in a collection of randomly selected pairs of genes whose half-life values were available. In each randomization, we generated 447 pairs of genes, which is equal to the number of interactions in the real dataset and obtained their average of the differences. A total of 100 000 random networks were generated to estimate the statistical significance of the observed difference in half-lives. Statistical significance was assessed based on *p*-value estimation, defined as the fraction of the 100 000 random networks which showed a value $\geq$ what was observed in the real network. Since the *p*-value of the average of the differences in mRNA half-lives when compared against random networks was lower than $\leq 1.43 \times 10^{-3}$, the results were considered to show a significant difference in comparison to the null model described above, suggesting that interacting proteins tend to show lower differences in mRNA half-lives than what is expected by chance.

Analysis of expression data for protein coding genes in *E. coli*

To compare the expression levels of essential and non-essential genes, we obtained a large compendium composing of 445 microarray datasets available as a public resource for *E. coli*.[37] These data were available in the form of Robust Multi Array (RMA) normalized profiles thus enabling us to directly calculate the average expression value of protein coding genes across all experimental conditions tested. Therefore, averaged gene expression values were used to compare the levels of expression of essential genes, non-essential genes and genes encoding hubs.

Statistical analysis for comparing gene expression and mRNA half-lives of essential, non-essential and hub encoding genes

To test the significance for the observed higher expression of hubs and essential genes over non-essential genes and to investigate whether essential genes produce stable transcripts compared to non-essential ones, we used the Welch two-sample *t*-test as implemented in the R statistical package

[see http://www.r-project.org]. We found a significantly lower p-value ($<2.2 \times 10^{-16}$), when we compared expression values of hubs and essential genes with respect to non-essential ones suggesting increased expression of the former groups across different conditions. In contrast, essential genes did not show a significant difference in their half-lives compared to non-essential gene set while hubs did.

Conclusion

It has long been believed that in bacteria, most of the regulation of gene expression is at the transcriptional level with little involvement of post-transcriptional control. However, recent studies indicate a widespread role for several novel mechanisms and molecules in regulating expression at the RNA level.[17,39–41] In this context, the findings presented here indicate for the first time, that proteins which are central in the protein interaction network tend to encode for stable mRNA transcripts and might be constrained to possess properties such as the formation of stem–loop structures or protection of their 3′ ends. This would provide them with enhanced stability over other transcripts, so that they would be available for a longer duration in the cell. The findings presented here suggest that hubs in the protein interaction network, which may need to interact with multiple proteins possibly at different time points during the cell cycle, might have been selected during evolution to have increased transcript stability, thereby enabling them to be utilized for multiple rounds of translation and decreasing the cost of transcription. This hypothesis is supported by the observation that certain functional categories like transcription factors and enzymes involved in core regulatory roles and central metabolism show extensive differences in their stabilities so that while one transcript is available for most of the cell's life time the other is available only under appropriate conditions in order to fine tune the interaction between them and/or to prevent undesirable cross-talk between them. Such high differences in half-lives can also act as fine-tuned feedback mechanisms at appropriate conditions through the physical interaction between proteins. In light of recent studies demonstrating the impact of the change in expression level of single gene over generations,[42] our results suggest that mRNA stability might not only provide a fitness advantage but also mediate regulation at post-transcriptional level, thereby allowing an organism to adapt to changing environments.

Our analysis of the expression level of essential genes suggests that while essential genes are highly expressed compared to non-essential ones and are enriched in hub encoding genes, they do not seem to encode for more stable transcripts. This is in contrast to what we observe for hubs which were found to be highly expressed and were encoded by stable transcripts. These findings are also in contrast to what was seen in eukaryotic PPI network where transcripts encoding hubs were significantly short-lived.[43] These observations suggest that the rapid turnover of hubs in eukaryotic PPI network reported earlier[43] might be explained based on the distinct mechanisms that prokaryotic and eukaryotic cells use to compartmentalize and regulate their protein availability. For instance, microRNAs which are known to regulate the expression of a significant fraction of the genes at the post-transcriptional level in higher organisms are known to preferentially inhibit the expression of hubs in both transcriptional and protein–protein interaction networks[44,45] possibly explaining their high turn over rates. Another possible explanation for these observed differences could be due to the fact that in bacteria, cell cycle duration is often small so that sometimes the turnover time of RNAs and other molecules exceeds the lifespan of a single generation. This may thereby provide the advantage of having higher stabilities for frequently used transcripts by allowing them to carry over the transcripts to future generations.

Our results also suggest that essential genes, which are highly expressed, might compensate for their abundance by not coding for highly stable transcripts, which might otherwise cause cellular crowding. On the contrary, hubs which were found to be highly expressed and produce stable transcripts might be utilized by the cell during most of its cell cycle or be translationally regulated, so that they are readily available whenever they are needed. It is also interesting to note that essential genes are composed of two kinds of genes, a small fraction forming hubs in the PPIs and showing higher transcript stability and a majority which are not highly connected in the PPI and show lower half-lives. These contributions from essential genes to form a small but significant fraction in hubs and a majority showing lower half-lives might be the cause for the observation that essential genes are not more stable than non-essential ones. Taken together our results demonstrate for the first time that mRNA stability has a significant role in mediating PPIs in bacteria and physical interactions might be influenced by a variety of post-transcriptional mechanisms.

Acknowledgements

SCJ and MMB acknowledge financial support from the MRC Laboratory of Molecular Biology. SCJ acknowledges financial support from Cambridge Commonwealth Trust. MMB thanks Darwin College and Schlumberger Ltd for generous support. We thank Wuster A, De S, Venkatakrishnan AJ and Weber K for critically reading the manuscript and providing helpful comments.

References

1 A. Jacobson and S. W. Peltz, *Annu. Rev. Biochem.*, 1996, **65**, 693–739.
2 C. A. Beelman and R. Parker, *Cell*, 1995, **81**, 179–183.
3 T. E. LaGrandeur and R. Parker, *EMBO J.*, 1998, **17**, 1487–1496.
4 J. S. Anderson and R. P. Parker, *EMBO J.*, 1998, **17**, 1497–1506.
5 R. S. Cormack, J. L. Genereaux and G. A. Mackie, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 9006–9010.
6 K. J. McDowall and S. N. Cohen, *J. Mol. Biol.*, 1996, **255**, 349–355.
7 A. J. Carpousis, *Annu. Rev. Microbiol.*, 2007, **61**, 71–87.
8 C. P. Ehretsmann, A. J. Carpousis and H. M. Krisch, *Genes Dev.*, 1992, **6**, 149–159.
9 A. J. Carpousis, G. Van Houwe, C. Ehretsmann and H. M. Krisch, *Cell*, 1994, **76**, 889–900.
10 Y. Feng, T. A. Vickers and S. N. Cohen, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 14746–14751.
11 Y. Zuo and M. P. Deutscher, *Nucleic Acids Res.*, 2001, **29**, 1017–1026.

12 K. Lee, X. Zhan, J. Gao, J. Qiu, Y. Feng, R. Meganathan, S. N. Cohen and G. Georgiou, *Cell*, 2003, **114**, 623–634.
13 J. Gao, K. Lee, M. Zhao, J. Qiu, X. Zhan, A. Saxena, C. J. Moore, S. N. Cohen and G. Georgiou, *Mol. Microbiol.*, 2006, **61**, 394–406.
14 J. A. Bernstein, P. H. Lin, S. N. Cohen and S. Lin-Chao, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 2758–2763.
15 J. A. Bernstein, A. B. Khodursky, P. H. Lin, S. Lin-Chao and S. N. Cohen, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 9697–9702.
16 D. W. Selinger, R. M. Saxena, K. J. Cheung, G. M. Church and C. Rosenow, *Genome Res.*, 2003, **13**, 216–223.
17 A. Szalewska-Palasz, G. Wegrzyn and A. Wegrzyn, *J. Appl. Genet.*, 2007, **48**, 281–294.
18 M. Bon, S. J. McGowan and P. R. Cook, *FASEB J.*, 2006, **20**, 1721–1723.
19 E. Wang and E. Purisima, *Trends Genet.*, 2005, **21**, 492–495.
20 A. Martinez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
21 G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt and A. Emili, *Nature*, 2005, **433**, 531–537.
22 M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H. C. Huang, A. Hirai, K. Tsuzuki, S. Nakamura, M. Altaf-Ul-Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, M. Kitagawa, M. Tomita, S. Kanaya, C. Wada and H. Mori, *Genome Res.*, 2006, **16**, 686–691.
23 J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne and P. Legrain, *Nature*, 2001, **409**, 211–215.
24 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Res.*, 2004, **32**, D449–451.
25 M. J. Marcaida, M. A. DePristo, V. Chandran, A. J. Carpousis and B. F. Luisi, *Trends Biochem. Sci.*, 2006, **31**, 359–365.
26 M. Grunberg-Manago, *Annu. Rev. Genet.*, 1999, **33**, 193–227.
27 R. Rauhut and G. Klug, *FEMS Microbiol. Rev.*, 1999, **23**, 353–370.
28 D. A. Steege, *RNA*, 2000, **6**, 1079–1090.
29 A. L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
30 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, *Science*, 2002, **297**, 1551–1555.
31 M. E. Newman, *Phys. Rev. Lett.*, 2002, **89**, 208701.
32 M. E. Newman, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2003, **67**, 026126.
33 T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori, *Mol. Syst. Biol.*, 2006, **2**, 0008.
34 H. Jeong, S. P. Mason, A. L. Barabasi and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
35 X. Gong, S. Fan, A. Bilderbeck, M. Li, H. Pang and S. Tao, *Mol. Genet. Genomics*, 2008, **279**, 87–94.
36 I. K. Jordan, I. B. Rogozin, Y. I. Wolf and E. V. Koonin, *Genome Res.*, 2002, **12**, 962–968.
37 J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, *PLoS Biol.*, 2007, **5**, e8.
38 U. Brandes, *J. Math. Sociol.*, 2001, **25**, 163–177.
39 G. Storz, S. Altuvia and K. M. Wassarman, *Annu. Rev. Biochem.*, 2005, **74**, 199–217.
40 W. C. Winkler and R. R. Breaker, *Annu. Rev. Microbiol.*, 2005, **59**, 487–517.
41 A. Serganov and D. J. Patel, *Nat. Rev. Genet.*, 2007, **8**, 776–790.
42 E. Dekel and U. Alon, *Nature*, 2005, **436**, 588–592.
43 N. N. Batada, L. D. Hurst and M. Tyers, *PLoS Comput. Biol.*, 2006, **2**, e88.
44 Q. Cui, Z. Yu, Y. Pan, E. O. Purisima and E. Wang, *Biochem. Biophys. Res. Commun.*, 2007, **352**, 733–738.
45 H. Liang and W. H. Li, *RNA*, 2007, **13**, 1402–1408.
46 S. Maere, K. Heymans and M. Kuiper, *Bioinformatics*, 2005, **21**, 3448–3449.

Scaling relationship in the gene content of transcriptional machinery in bacteria†‡

Ernesto Pérez-Rueda,[a] Sarath Chandra Janga*[b] and Agustino Martínez-Antonio*[c]

Received 14th April 2009, Accepted 9th June 2009
First published as an Advance Article on the web 17th July 2009
DOI: 10.1039/b907384a

The metabolic, defensive, communicative and pathogenic capabilities of eubacteria depend on their repertoire of genes and ability to regulate the expression of them. Sigma and transcription factors have fundamental roles in controlling these processes. Here, we show that sigma, transcription factors (TFs) and the number of protein coding genes occur in different magnitudes across 291 non-redundant eubacterial genomes. We suggest that these differences can be explained based on the fact that the universe of TFs, in contrast to sigma factors, exhibits a greater flexibility for transcriptional regulation, due to their ability to sense diverse stimuli through a variety of ligand-binding domains by discriminating over longer regions on DNA, through their diverse DNA-binding domains, and by their combinatorial role with other sigmas and TFs. We also note that the diversity of extra-cytoplasmic sigma factors and TF families is constrained in larger genomes. Our results indicate that most widely distributed families across eubacteria are small in size, while large families are relatively limited in their distribution across genomes. Clustering of the distribution of transcription and sigma families across genomes suggests that functional constraints could force their co-evolution, as was observed in sigma54, IHF and EBP families. Our results also indicate that large families might be a consequence of lifestyle, as pathogens and free-living organisms were found to exhibit a major proportion of these expanded families. Our results suggest that understanding proteomes from an integrated perspective, as presented in this study, can be a general framework for uncovering the relationships between different classes of proteins.

[a] *Departamento de Ingeniería Celular y Biocatálisis, IBT-UNAM. AP. 565-A, Cuernava-ca, Morelos, 62210, México*
[b] *MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 0QH. E-mail: sarath@mrc-lmb.cam.ac.uk*
[c] *Departamento de Ingeniería Genética. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Irapuato, 36500, México. E-mail: amartinez@ira.cinvestav.mx*

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

‡ Electronic supplementary information (ESI) available: An extensive set of TFs from all 675 bacterial genomes (including redundant ones) and further supplementary material associated with this study. See DOI: http://www.ibt.unam.mx/~erueda/ScalingTranscription.htm

Introduction

Bacteria respond and adapt to diverse environmental conditions as a consequence of their gene repertoire and regulatory mechanisms, among other elements.[1–3] The availability of their genome sequences has enabled the investigation of their differences at genetic, molecular and biochemical levels. Recent studies have shown that the evolutionary events associated with regulatory gene families, such as their expansion and contraction, contribute significantly to shaping the gene repertoire and genome size of different lineages of prokaryotes.[4–7]

Ernesto Pérez-Rueda

Ernesto Perez-Rueda has been a professor at Universidad Nacional Autonoma de Mexico (UNAM) since 2004. He obtained his PhD at the Center for Genomic Sciences in UNAM and worked on the identification of functional residues in homeoproteins in his post-doctoral research at the Free University of Brussels. His research focuses on the analysis of DNA-binding transcription factors in diverse bacteria, such as E. coli and B. subtilis, to understand the evolution of TFs and predict their functional roles. He has published several international publications on these topics.

Sarath Chandra Janga

Sarath Chandra Janga is a PhD student at the MRC Laboratory of Molecular Biology and University of Cambridge. Sarath obtained his Bachelors and Masters in biochemical engineering and biotechnology at the Indian Institute of Technology, Delhi in 2003. Prior to starting his PhD, Sarath worked extensively and co-ordinated a number of research projects on transcriptional regulation, genome organization and comparative genomics in bacteria at UNAM in Mexico. He has published more than 25 research manuscripts on various aspects of prokaryotic and eukaryotic biology in the fields of computational molecular and systems biology. His current research interests include understanding the design principles and constraints imposed on post-transcriptional and post-translational gene control in prokaryotic and eukaryotic organisms.

Based on comparative genomics, it has been shown that genes associated with transcriptional regulation increase in a quadratic proportion with respect to the genome size.[8–10] These observations become pertinent given that the regulation of transcription initiation in bacteria is primarily mediated by sigma factors (σs), which provide most of the specificity for promoter recognition and DNA melting needed for transcription initiation.[11–13] In fact, sigma factors perform these functions only when bound to the RNA polymerase (RNAP). On the other hand, DNA-binding transcription factors (TFs)[14] affect gene expression by blocking or allowing the access of the RNAP to the promoter, depending on the operator context and ligand-binding status.[15–18] Usually, most gene transcription in exponentially growing bacteria is initiated by RNAP carrying a housekeeping σ, similar to *E. coli* σ^{70} or *B. subtilis* σ^{A}. Alternative σs typically redirect the RNAP towards a subset of genes required during specific conditions, such as stress response or growth transitions, among others.[11–13] TFs represent a class of proteins devoted to sense and bind signals to regulate genes, in response to specific compounds.[17,19] Although there is extensive evidence for the existence of alternative regulatory mechanisms in diverse bacterial systems from post-transcriptional regulation,[20–22] they are not considered in this study, as we focus on the specific role of TFs in mediating regulatory mechanisms in a wide range of completely sequenced bacterial genomes.

It has been previously suggested that the abundance of TFs increases with an increase in an organism's complexity[8,23–26] as a consequence of different evolutionary events, such as gene expansion, gene loss and lateral gene transfer.[24,27,28] On the other hand, the repertoire of TFs, depending on their hierarchical position in the network of transcriptional interactions, have also been shown to play an important role in shaping the organization of genes on bacterial chromosomes.[29–32] In this study, we analyze the repertoires of σs and TFs in 291 eubacterial genomes and compare their distribution in relation to the genome size to understand their contribution to gene regulation in different lineages and lifestyles. The results obtained here provide insights into the functional and evolutionary constraints imposed on different classes of regulatory factors in bacterial organisms.

Agustino Martínez-Antonio

Agustino Martínez-Antonio obtained his doctoral degree in biochemical sciences from UNAM. After postdoctoral research at UNAM and INSERM, he is currently a professor at the Research and Advanced Studies Centre of the National Polytechnic Institute (CINVESTAV-IPN). His interests include understanding the design principles governing the structure and function of regulatory networks in prokaryotes.

Results

The abundance of sigma factors and TFs correlates with genome size in bacteria

To study the abundance and diversity of regulatory proteins controlling transcription initiation, the repertoires of σs and TFs were obtained in 291 non-redundant (NR) bacterial genomes (see Materials and methods section for details). A comparison of regulatory elements across genomes suggested that they increase almost quadratically with genome size (Fig. 1). In particular, we found that the repertoire of TFs is roughly 10 times higher than σs (hundreds *vs.* tens) when we considered the general profiles in all the genomes analyzed, suggesting a proportion in the order of 1 σs : 10 TFs : 100 annotated ORFs per genome, although some genomes deviate from this trend. This observation suggests that possible functional relationships between TFs and σs, on one hand, and bacterial lifestyles, on the other, could both be influencing the observed trend. We discuss the impact of both of these scenarios in the following sections.

The variation in the extent of conservation of σs compared to TFs might be explained based on their regulatory roles at transcription initiation

Firstly, the differences in the abundance of repertoires of σs and TFs in bacteria might be attributed to the different regulatory roles associated with them. Transcription starts when a σ interacts with RNAP to recognize its specific sequence promoter (Fig. 1). This promoter recognition stage imposes the existence of at least one σ per organism, which

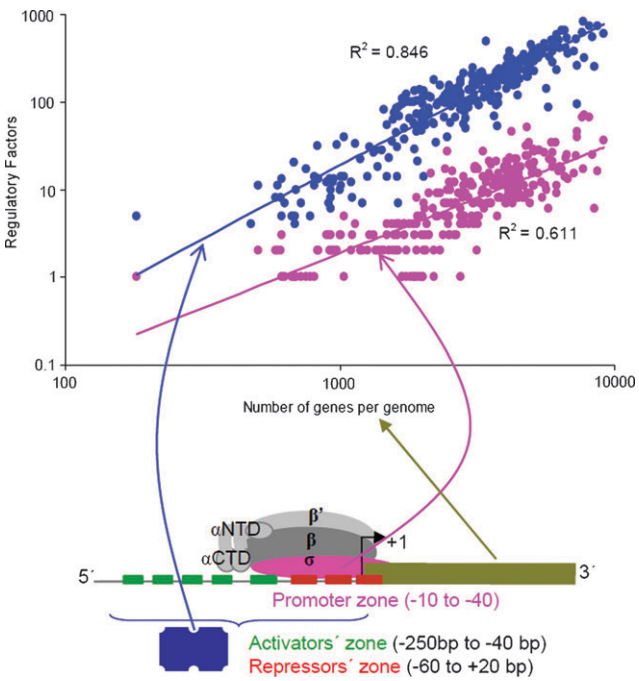


Fig. 1 The distribution of the number of TFs and σs in bacterial genomes as a function of genome size. Genomes are sorted on the x-axis by the number of ORFs. The abundance of TFs and σs in each genome is shown on the y-axis (each dot corresponds to one genome). σs are shown in pink and transcription factors in blue.

typically belongs to the σ^{70} family.[13] As a result, bacterial systems might be able to switch between different transcriptional programs based exclusively on their repertoire of σ factors. Nonetheless, the transcriptional programs mediated uniquely *via* σs would be restricted, as a result of their limited repertoire and the small collection of ligands they can recognize, such as guanosine tetraphosphate (ppGpp).[33] As a consequence, σs exhibit a limited ability to directly couple the environmental conditions with gene transcription. In addition, σs have a constrained DNA-binding region in terms of length and the diversity of sequences they recognize, as they need to be structurally-coupled to the RNAP in the promoter zone. These restricted zones of action divide the universe of σs into promoters recognized by σ^{70} and those recognized by σ^{54} (the binding zones correspond to about −10 to −35 bp for σ^{70} and −12 to −24 for σ^{54}, relative to the transcription start site).[34,35]

On the other hand, TFs define a different regulatory level compared to σs. These proteins exhibit diverse structural and functional domains, where one of them specifically binds to DNA and the other can sense and bind one or more ligand compounds from endogenous and/or exogenous sources,[17] such as the TyrR of *E. coli*, which bind to three aromatic amino acids and ATP.[36] In addition, TFs associate combinatorially, not only with σs, but also with a number of other TFs and DNA-binding sites,[37,38] thus allowing the rewiring of a transcriptional network depending on the environmental conditions; for instance, *sodA*, a gene encoding for superoxide dismutase in *E. coli*, is regulated by up to eight different TFs responsible for various cellular responses, including Fur (ferric uptake regulation protein), Arc (aerobic respiratory control) and Fnr (fumarate nitrate reduction/regulator of anaerobic respiration).[39,40] Finally, the diversity of sequences that TFs can recognize is enormous and can occur anywhere from a few bases downstream of the promoter zone to up to hundreds of bases upstream of the transcription start site (Fig. 1).[41,42] For instance, the global regulator CRP (catabolic repressor protein) in *E. coli* can regulate promoters associated with four out of the seven possible σs and co-regulate with more than 50 different TFs.[43,44] In summary, TFs constitute a class of proteins whose space of action is more flexible than that of σs, not only in sensing diverse environmental and endogenous stimuli, but also in recognizing a wide range of binding site sequences over a larger zone on the DNA around the transcription start site.

Lifestyles explain the abundance of σs and TFs in bigger genomes

The results of the previous sections suggest that regulatory complexity should increase in larger genomes and might be associated with bacterial lifestyles, as the environment should influence the bacterial genome structure and function. Thus, we analyzed the genomes in relation to the four global classes of lifestyles.[45] These included extremophiles (21 genomes), intracellular bacteria (28 genomes), pathogens (109 genomes) and free-living bacteria (133 genomes). To understand how the complexity of gene regulation depends on the number of σs and/or TFs, as a function of increasing genome size and how they are associated to lifestyle, we calculated the ratio of TFs/number of genes (T/G) and σs/number of genes (S/G), (Fig. 2). From this analysis, we found that the increase in regulatory complexity in intracellular (I) and extremophilic (E) bacteria depends almost exclusively on the TF repertoire (no correlation was observed for an increase in σ with genome size for these lifestyles). On the other hand, in pathogenic (P) bacteria, the regulatory repertoire is contributed-to by TFs and to some extent by σs. In contrast, σs and TFs contributed almost equally to the regulatory repertoire in free-living (F) bacteria. Thus, TFs contribute significantly to the regulatory complexity of bacteria belonging to different lifestyles, whereas σs contribute more significantly to the transcriptional machinery of regulation in pathogens and free-living bacteria. These results agree with previous observations, which suggest that few regulatory elements identified in small genomes would compensate the regulation of the entire genome with an increase in the number of DNA-binding sites per element, in contrast to the large number of elements identified in large genomes that control a lesser proportion of DNA-binding sites

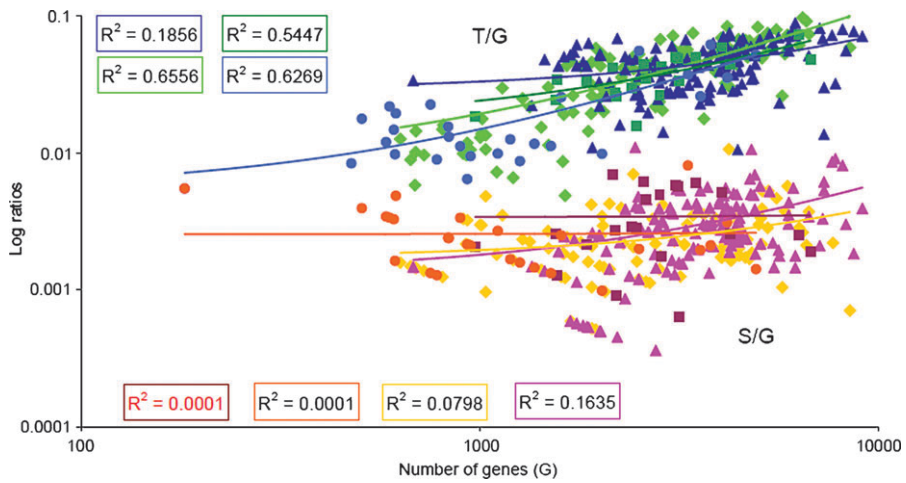


Fig. 2 The ratio of regulatory factors to the total number of ORFs per genome. The number of genes encoding for TFs and σs were normalized with respect to the total number of ORFs per genome (T/G and S/G, respectively), and these ratios are shown for bacteria belonging to four different lifestyles: free-living (F) (▲), extremophiles (E) (■), pathogens (P) (◆) and intracellular (I) (●).

on average.[10] In addition, genes in small genomes are organized into large operons, simplifying the transcriptional machinery necessary for gene expression. This is in contrast to large genomes, which have a reduced number of genes in operons, influencing the proportion of σs and TFs in those organisms,[46] suggesting that complex lifestyles would require a higher proportion of TFs and transcription units to better orchestrate a response to changing conditions.

The contribution of sigma factors to the transcriptional machinery trend

In order to assess the contribution of σs to the trends described in Fig. 1 and Fig. 2, they were divided into three main groups based on their sequence and function. As described in the previous section, we then computed the ratio of the number of σs/number of genes (S/G) in all the genomes for each group of σs, namely σ^{54}, σ^{70} and extra cytoplasmic function (ECF) sigma factors.[13] From this analysis, we found that the abundance of σs is primarily determined by the number of ECFs and σ^{70}s, as the number of σ^{54} members was found to be roughly constant and often occurred in no more than a single copy in most genomes (Fig. 3(a)). ECFs were highly abundant in free-living and pathogenic bacteria, with genomes containing more than 2000 genes, and might be the result of massive gene duplications.[47,48] The extent of conservation of different types of σs across bacteria suggests a functional role for each, depending on their distribution. For instance, σ^{70} is indispensable to the adequate maintenance of a cell and is the only sigma identified in small genomes with less than 800 genes, whereas ECFs are factors associated with the regulation of functional processes beyond the basal ones. In obligate intracellular pathogens, such as *Mycoplasma sp*, *Streptococcus* mutants or *Lactobacillus plantarum*, there is only one housekeeping σ^{70} and no alternative σs. σ^{54} factors were found to exhibit an almost constant distribution of one copy per genome, except in some pathogens and free-living eubacteria, where they were identified in two-copies (see the ESI‡). σ^{54} factors require the assistance of specialized activators of the EBP (enhancer binding protein) family of TFs, and this might have constrained the number of genes regulated by σ^{54}, *i.e.* promoters associated with σ^{54} frequently require the bending of long intergenic DNA stretches *via* IHF, resulting in a specific physical proximity between the RNAP and TFs.[49,50] Thus, evolutive mechanisms working for chromosome compactness might be working against the increased use of σ^{54} promoters in bacteria.

To analyze the specific contribution of the different families of σs to gene transcription, we computed the ratio of the number of σs/genes (S/G) in all the genomes. Fig. 3(b) shows, as expected, that σ^{70}s have a higher proportion of genes to transcribe in small genomes, but that as genome size increases, this proportion diminishes; ECF is the only family whose proportion of regulated genes increases in larger genomes. Most of the diversification of ECFs corresponds to free-living and pathogenic genomes with ~5000 ORFs.

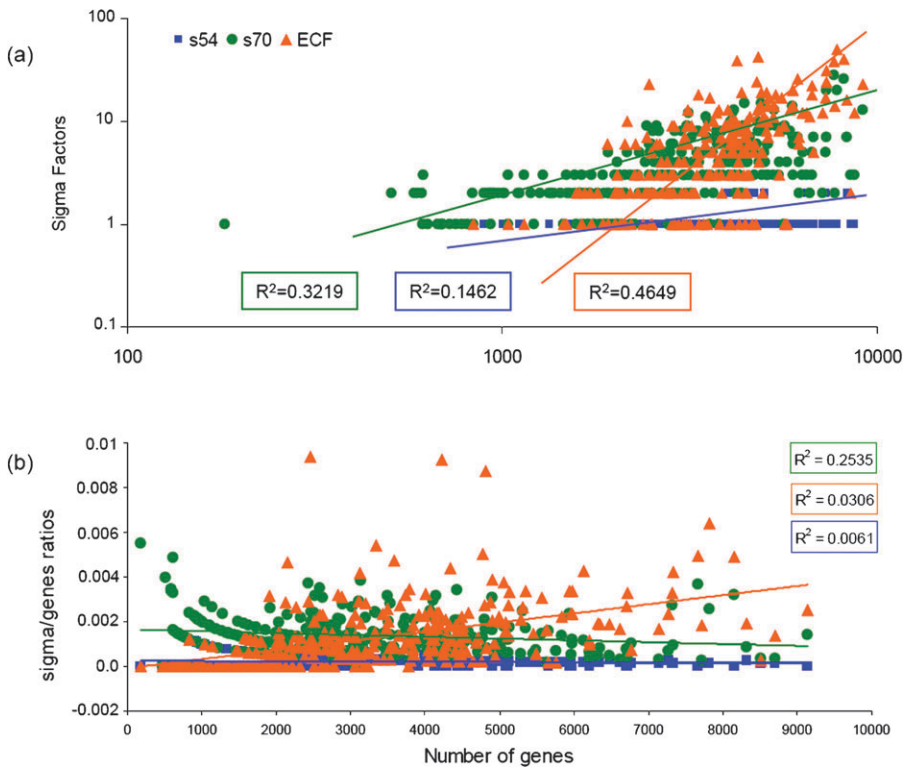


Fig. 3 The distribution of families of σs in bacterial genomes. (a) Genome size is shown on a log scale on the x-axis. The y-axis shows the number of σ factors in each family per genome. (b) The ratio of the number of sigma factors from each family to the total number of ORFs per genome; the three outliers, with a high number of ECFs, correspond (from left to right) to β-proteobacteria (*N. europaea*) and two bacteriodes (*B. fragilis* NCTC9434 and *B. thetaiotaomicron* VPI-5482).

The abundance of TFs does not correlate with the diversity of families, and large families are not the most widely distributed

An appealing hypothesis is that a high diversity of TF families would contribute more significantly to regulatory plasticity than σs. In line with this hypothesis, an analysis of 93 TF families, comprising of a total of 46 255 TFs across all the genomes analyzed in this study, showed a reduced diversity of families in small genomes, with an increasing proportion in larger ones, especially in pathogens (P) and free-living organisms (F) (Fig. 4(a)). The diversity of families reaches a maximum in genomes with around 5000 ORFs. The higher number of TFs in larger genomes does not necessarily imply the diversity of families beyond this plateau, but instead an increase in the size of some families of TFs. Congruent with this observation, Fig. 4(b) shows that the average number of TFs per family increases linearly, with a few families of TFs expanding disproportionately. These families comprise of LysR and TetR, which represent about 24% of the total set of TFs identified (11 078 of 46 255 proteins). Members of these two families increase abruptly in larger genomes, as shown in Fig. 4(c), which also shows three other most-populated families of TFs in eubacteria for the sake of comparison. The increase in the size of these two families in larger genomes coincides with the plateauing of the diversity of families in these bacterial genomes (marked by arrows in Fig. 4(a), (b), and (c)). Another feature associated with large families is that they are not widely distributed among bacteria, despite their role in controlling important processes, such as cell–cell communication (LuxR), the response to external conditions by two-component systems (OmpR), the sensing, uptake and metabolism of external food sources (GntR and LysR), or resistance to antibiotics (TetR). On the other hand, some families with an average size of a few copies per genome, such as DnaA, LexA and IHF from *E. coli*, proposed to be essential in standard growth conditions in this bacterium and in keeping its DNA and nucleoid integrity,[51,52] can be considered to be conserved across bacteria. This is because they were identified in at least 86% of the genomes, suggesting probable gene loss events in bacteria where they are absent (Fig. 5).

In summary, our results suggest that a family's abundance and distribution is associated with evolutionary events in bacteria. For instance, small families widely distributed among

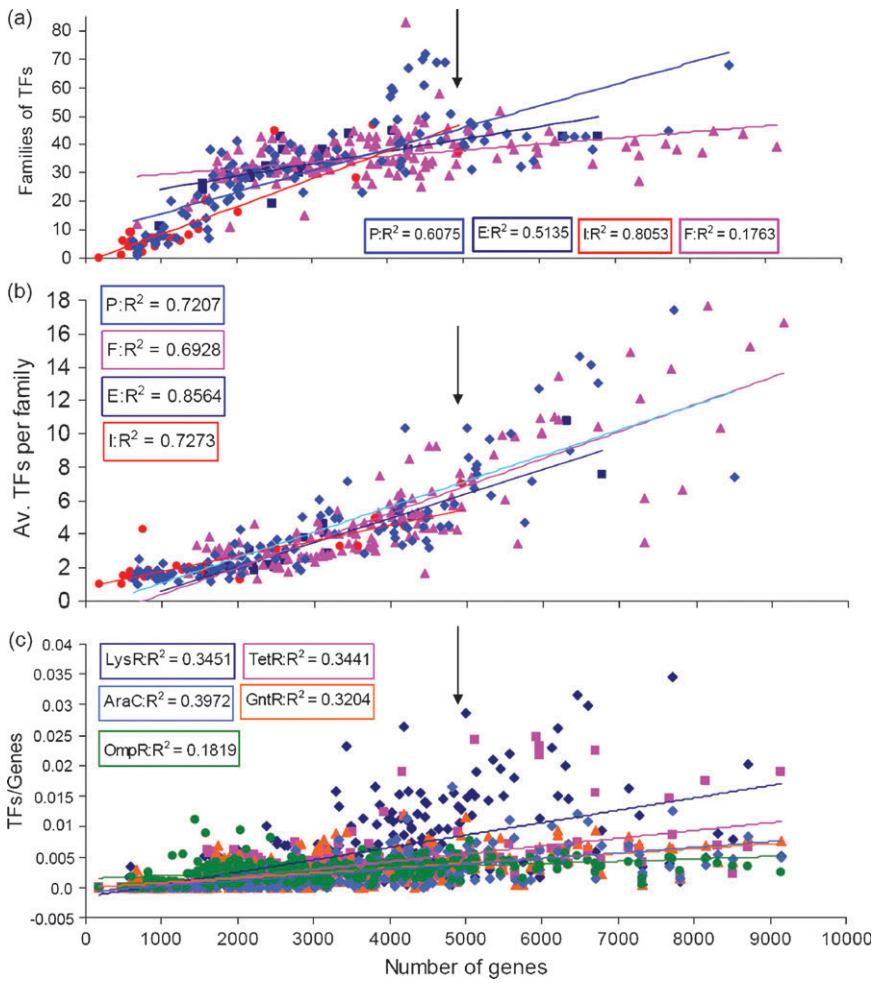


Fig. 4 Characteristics of TF families in bacterial genomes. (a) The number of TF families as a function of the number of ORFs in each genome, grouped according to the lifestyle of the organism: E (extremophiles), I (intracellular), P (pathogens) and F (free-living bacteria). (b) The average number of TFs per family as a function of the number of ORFs in each genome, grouped according to the lifestyle of the organism, as in (a). (c) The ratio of the number of TFs to ORFs per genome for the five most abundant families of TFs in bacterial genomes.

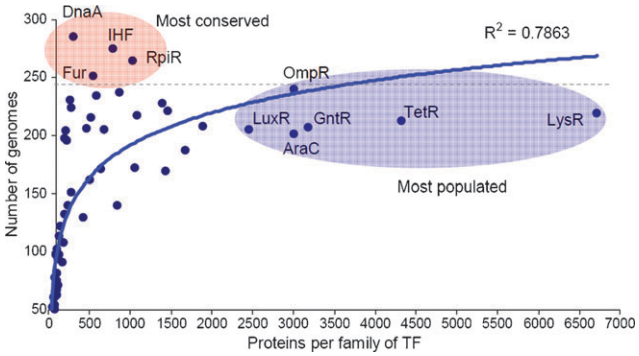


Fig. 5 The diversity and conservation of TF families in bacteria. The occurrence of a TF family across genomes as a function of the total number of TFs identified. Some families of TFs conserved in a few copies per genome are circled in pink. Note that these are also the most conserved families of TFs in the analyzed genomes. In contrast, some families (circled in blue) are the most populated, though are less conserved, in comparison to those circled in pink across genomes.

bacteria might be related to ancestral functions beyond transcriptional regulation, such as DNA organization, or nucleoid integrity or DNA salvage, whereas large families might be associated with the regulation of dispensable or emergent processes in bacterial evolution, such as quorum sensing, belonging to the members of the LuxR family, which are widely identified in bacteria. Indeed, the evolution of this mechanism in bacteria has been proposed to be one of the early steps in the development of multicellularity,[53] and may be correlated with bacterial specialization.

Functional relationships might impose evolutionary constraints

Since some proteins tend to work together in a functional context, we analyzed the distributions of different families, as this would give us an indication about the co-evolution of regulatory factors. Hence, we clustered the co-occurrence of the regulatory protein families (TFs and σs) in all 291 bacterial genomes, as shown in Fig. 6. From this analysis, we found that the distribution of σ^{54}, IHF and EBP families is correlated, supporting the functional interdependence discussed above (and inset in Fig. 6) and probable co-evolution, where members and mechanisms have been preserved along the course of evolution. A second cluster including σ^{70}, the ECF family of sigma factors and other highly abundant families (more than 15 members per genome) responsible for regulating diverse mechanisms of stress responses (MarR), antibiotic resistance (TetR), osmotic response (OmpR) and quorum sensing response (LuxR), among other processes, were also found to be clustered as a result of this analysis. This suggests a strong functional relationship among these σ and TF families. These clusters, in addition, give insights into the functional interdependence between regulatory proteins from different families, which could help in the characterization of regulators in poorly studied genomes.

Materials and methods

Genome sequences

Predicted proteomes for 291 eubacteria were obtained from the entrez genome database of the NCBI (ftp://ncbi.nlm.nih.gov/genomes/bacteria).[54]

A complete list of non-redundant genomes can be obtained at http://popolvuh.wlu.ca/Phyl_Profiles/NR_genomes/REDUNDANCY.html. In brief, two genomes are considered redundant if they share a genomic similarity score (GSS) higher than 0.95, where GSS is defined as the ratio of the sum of all the BLAST bit-scores for protein coding genes that have orthologs between two genomes being compared and reaches a maximum of one if all the proteins of one organism are identical to their corresponding orthologs of another organism. This would be the case when the proteomes are identical.[55,56] A complete list of genomes analyzed and their repertoire of TFs is provided as ESI.‡

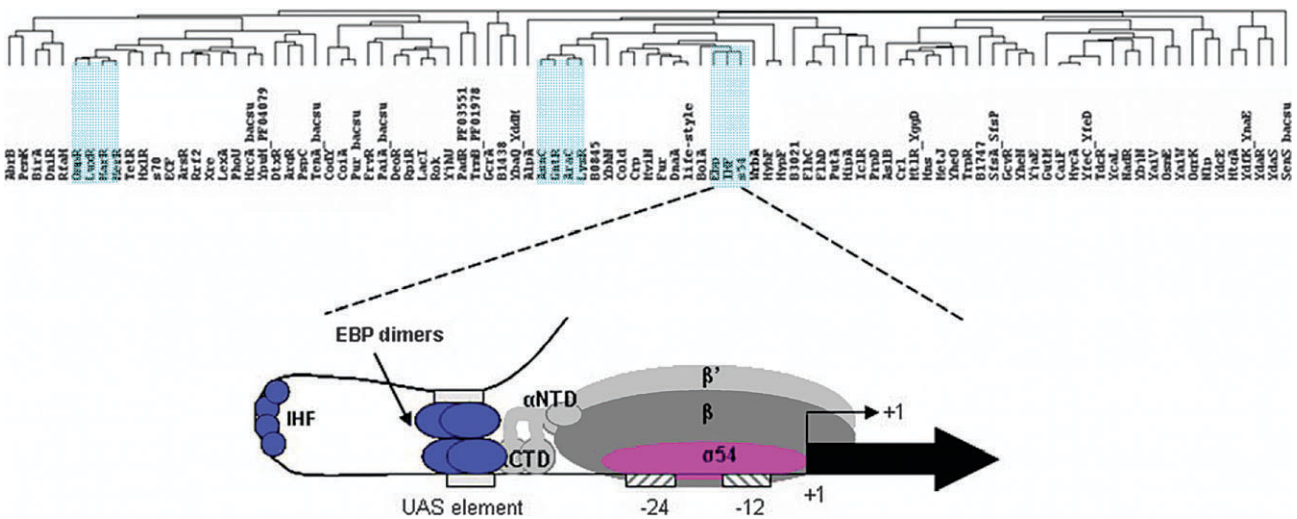


Fig. 6 The clustering of transcription and sigma factor families across bacterial genomes based on their co-occurrence profiles. A clear co-occurrence distribution is observed for IHF, EBP and σ^{54} families, suggesting a functional interdependence between them. The co-regulatory mode of action for these regulatory proteins is shown in the inset.

The identification of families of DNA-binding transcription factors (TFs)

To identify and analyze the repertoire of TFs in bacterial genomes, a combination of information from different sources and bioinformatics tools were used. Firstly, 45 088 putative TFs were collected from the transcription factor DB,[57] a database devoted to the identification and classification of DNA-binding TFs by means of the SUPERFAMILY library and PFAM hidden Markov models (HMMs). In a second phase, 90 family-specific HMMs previously reported from *E. coli* K12 and 57 family-specific HMMs from *B. subtilis*[5,58] were used to scan the complete genome sequences (E-value threshold = 10^{-3}) with the hmmsearch module of the HMMer suite program (http://hmmer.janelia.org). TF families were identified based on their DNA-binding domains: in a first step, if a protein shared more than 25% of the identity in its DNA-binding region with any member of the well-characterized TFs of *E. coli* and/or *B. subtilis*, it was included in this particular family. In order to include distant homologs and to decrease the bias associated with the over-representation of TFs from specific organisms, these families were expanded by Blast searches[59] against the SwissProt database[60] using an E-value threshold of 10^{-6}. Proteins retrieved were filtered at 100% to exclude redundancy using the program CD-hit[61] and aligned with ClustalW.[62] Proteins with less than 50% similarity against their corresponding HMM were excluded. This step is important to explore potential TFs not identified through the first approach and *vice versa*, *i.e.* the coverage of the DBD database corresponds to approximately 70% of the universe of TFs and can be complemented with family-specific HMMs.[63] Previous studies using this approach for predicting new TFs suggest that these models are successful in identifying a significant fraction of experimentally confirmed TFs in different lineages,[40,64] confirming the value of these predictions for studying genome-scale patterns. An extensive set of TFs from all 675 bacterial genomes (including redundant ones) and supplementary material associated with this study is available.‡

The identification of σ factors

Three HMMs were used to identify σ^{70}, σ^{54} and ECF-like sigma factors across genomes. σ^{70} and σ^{54} models were retrieved from the PFAM database.[65] ECFs have been considered as a separate group of σ^{70} proteins because of their significant sequence divergence from the σ^{70} family. Thus, we constructed a specific ECF HMM based on the well-known repertoire of ECF proteins in *B. subtilis*. These proteins were used to run the motif discovery and search system, MEME/MAST (using default parameters), to identify specific regions associated with this group. We selected two motifs to construct HMMs and to scan the whole repertoire of bacterial genomes. The motifs and HMMs are available in the ESI.‡

Clustering of families of regulatory factors

To analyze the distribution of σs and TF families across the 291 bacterial genomes, they were first saved as a matrix. This matrix was then loaded into the cluster 3.0 program[66] to identify groups of families that correlate in terms of their occurrence profile across all the bacterial genomes. A hierarchical complete linkage clustering algorithm was run with an uncentered correlation as a similarity measure. The clustering results were then visualized using the Treeview program.[66]

Conclusions

To understand the relationship between the expansion patterns of different regulatory factors involved in gene regulation at transcription initiation, 291 completely sequenced bacterial genomes, which represent adaptive designs for different lifestyles, were analyzed. We showed that the distribution of σs and TFs follows a trend, with a ratio of 1 σ per 10 TFs and 100 ORFs in all the genomes analyzed, coinciding with our present knowledge that σs direct RNAP to a small repertoire of binding sites in sequence and location, compared to the diversity provided by the collection of TFs at the promoters in a genome. For instance, in *E. coli*, around 95% of its genes are transcribed by σ^{70}, with the fine tuning of their expression mediated by TFs.[44] In addition, we found that, in large genomes, there is a decrease in the number of different families of TFs, *i.e.*, in the diversity of families, than would otherwise be expected. In this context, abundant families are not widely distributed across all bacteria. In contrast, some small families are the most widely distributed. This difference might be associated with different phenomena, such as evolutionary constraints by regulatory mechanisms, as discussed in the case of DnaA or LexA and EBP families. Our results also suggest that in larger genomes, regulatory complexity may possibly increase as a result of the increasing number of members from the ECF family and some TF families. However, it is unclear if this increase would correspond to an increase in complexity by means of multiple parallel switches and feed-forward loops in regulatory networks (as shown for carbon sources in *E. coli*[67]), as long regulatory cascades, or as a combination of both. Overall, the analyses presented here will not only contribute to improving our understanding of the influence of design on the regulation of gene expression, but also support the basis for a comprehensive modelling of transcriptional regulatory networks in bacteria. The observations discussed in this study should be valid for a wide-range of bacteria in most genomic studies; the analysis of over 100 genomes is reported to be sufficient and robust enough to be generalized.[68]

Abbreviations

σs	Sigma factors
TFs	Transcription factors
EBP	Enhancer binding proteins
ECF	Extra-cytoplasmic sigma factors

Acknowledgements

EP-R was financed by a grant (IN-217508) from DGAPA-UNAM and by grants given to Lorenzo Segovia. S. C. J. acknowledges financial support from the Cambridge Commonwealth Trust and the MRC-Laboratory of Molecular

Biology. A. M.-A. acknowledges financial support from CONCYTEG (young researcher) and CINVESTAV (multidisciplinary). S. C. J. thanks colleagues at MRC-LMB for critically reading the manuscript and providing helpful comments. We thank Jose Antonio Ibarra and Gabriel Moreno-Hagelsieb for their helpful comments in the preparation of the manuscript, and Derek Wilson for providing us with the collection of TFs available through the DBD database.

References

1 M. Lynch and J. S. Conery, *Science*, 2003, **302**, 1401–1404.
2 M. Lynch, *Annu. Rev. Microbiol.*, 2006, **60**, 327–349.
3 B. O. Bengtsson, *J. Theor. Biol.*, 2004, **231**, 271–278.
4 Y. Minezaki, K. Homma and K. Nishikawa, *DNA Res.*, 2005, **12**, 269–280.
5 E. Perez-Rueda, J. Collado-Vides and L. Segovia, *Comput. Biol. Chem.*, 2004, **28**, 341–350.
6 D. A. Rodionov, *Chem. Rev.*, 2007, **107**, 3467–3497.
7 J. A. Oguiza, K. Kiil and D. W. Ussery, *Trends Microbiol.*, 2005, **13**, 565–568.
8 E. van Nimwegen, *Trends Genet.*, 2003, **19**, 479–484.
9 O. X. Cordero and P. Hogeweg, *Trends Genet.*, 2007, **23**, 488–493.
10 N. Molina and E. van Nimwegen, *Genome Res.*, 2008, **18**, 148–160.
11 M. M. Wosten, *FEMS Microbiol. Rev.*, 1998, **22**, 127–150.
12 A. Ishihama, *Annu. Rev. Microbiol.*, 2000, **54**, 499–518.
13 T. M. Gruber and C. A. Gross, *Annu. Rev. Microbiol.*, 2003, **57**, 441–466.
14 D. F. Browning and S. J. Busby, *Nat. Rev. Microbiol.*, 2004, **2**, 57–65.
15 N. S. Miroslavova and S. J. Busby, *Biochem. Soc. Symp.*, 2006, 1–10.
16 M. E. Wall, W. S. Hlavacek and M. A. Savageau, *Nat. Rev. Genet.*, 2004, **5**, 34–42.
17 A. Martinez-Antonio, S. C. Janga, H. Salgado and J. Collado-Vides, *Trends Microbiol.*, 2006, **14**, 22–27.
18 S. C. Janga and J. Collado-Vides, *Res. Microbiol.*, 2007.
19 A. Goelzer, F. Bekkal Brikci, I. Martin-Verstraete, P. Noirot, P. Bessieres, S. Aymerich and V. Fromion, *BMC Syst. Biol.*, 2008, **2**, 20.
20 A. Gutierrez-Preciado, T. M. Henkin, F. J. Grundy, C. Yanofsky and E. Merino, *Microbiol. Mol. Biol. Rev.*, 2009, **73**, 36–61.
21 R. R. Breaker, *Science*, 2008, **319**, 1795–1797.
22 C. A. Wakeman, W. C. Winkler and C. E. Dann, 3rd, *Trends Biochem. Sci.*, 2007, **32**, 415–424.
23 J. H. Brown, V. K. Gupta, B. L. Li, B. T. Milne, C. Restrepo and G. B. West, *Philos. Trans. R. Soc. London, Ser. B*, 2002, **357**, 619–626.
24 M. Levine and R. Tjian, *Nature*, 2003, **424**, 147–151.
25 M. A. Changizi, *J. Theor. Biol.*, 2001, **211**, 277–295.
26 G. B. West and J. H. Brown, *J. Exp. Biol.*, 2005, **208**, 1575–1592.
27 L. Aravind, V. Anantharaman, S. Balaji, M. M. Babu and L. M. Iyer, *FEMS Microbiol. Rev.*, 2005, **29**, 231–262.
28 M. Madan Babu, S. A. Teichmann and L. Aravind, *J. Mol. Biol.*, 2006.
29 S. C. Janga, H. Salgado, J. Collado-Vides and A. Martinez-Antonio, *J. Mol. Biol.*, 2007, **368**, 263–272.
30 S. C. Janga, H. Salgado and A. Martinez-Antonio, *Nucleic Acids Res.*, 2009, **37**, 3680–3688.
31 G. Kolesov, Z. Wunderlich, O. N. Laikova, M. S. Gelfand and L. A. Mirny, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 13948–13953.
32 C. Marr, M. Geertz, M. T. Hutt and G. Muskhelishvili, *BMC Syst. Biol.*, 2008, **2**, 18.
33 L. Jores and R. Wagner, *J. Biol. Chem.*, 2003, **278**, 16834–16843.
34 J. D. Gralla, *Curr. Opin. Genet. Dev.*, 1996, **6**, 526–530.
35 G. Lloyd, P. Landini and S. Busby, *Essays Biochem.*, 2001, **37**, 17–31.
36 J. Pittard, H. Camakaris and J. Yang, *Mol. Microbiol.*, 2005, **55**, 16–26.
37 S. Adhya, *Sci. STKE*, 2003, **2003**, pe22.
38 A. Barnard, A. Wolfe and S. Busby, *Curr. Opin. Microbiol.*, 2004, **7**, 102–108.
39 I. Compan and D. Touati, *J. Bacteriol.*, 1993, **175**, 1687–1696.
40 H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio and J. Collado-Vides, *Nucleic Acids Res.*, 2006, **34**, D394–397.
41 M. Madan Babu and S. A. Teichmann, *Trends Genet.*, 2003, **19**, 75–79.
42 J. Collado-Vides, B. Magasanik and J. D. Gralla, *Microbiol. Rev.*, 1991, **55**, 371–394.
43 S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Treviño-Quintanilla and J. Collado-Vides, *Nucleic Acids Res.*, 2008, **36**, D120–124.
44 A. Martinez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.*, 2003, **6**, 482–489.
45 I. Cases, V. de Lorenzo and C. A. Ouzounis, *Trends Microbiol.*, 2003, **11**, 248–253.
46 G. Moreno-Hagelsieb, *Curr. Genomics*, 2006, **7**, 163–170.
47 D. Missiakas and S. Raina, *Mol. Microbiol.*, 1998, **28**, 1059–1066.
48 J. D. Helmann, *Adv. Microb. Physiol.*, 2002, **46**, 47–110.
49 S. V. Kuznetsov, S. Sugimura, P. Vivas, D. M. Crothers and A. Ansari, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18515–18520.
50 S. Sugimura and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18510–18514.
51 Y. Yamazaki, H. Niki and J. Kato, *Methods Mol. Biol.*, 2008, **416**, 385–389.
52 S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai and A. L. Osterman, *J. Bacteriol.*, 2003, **185**, 5673–5684.
53 M. B. Miller and B. L. Bassler, *Annu. Rev. Microbiol.*, 2001, **55**, 165–199.
54 R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin, *Nucleic Acids Res.*, 2001, **29**, 22–28.
55 G. Moreno-Hagelsieb and S. C. Janga, *Proteins*, 2008, **70**, 344–352.
56 G. Moreno-Hagelsieb and J. Collado-Vides, *Bioinformatics*, 2002, **18**(suppl. 1), S329–336.
57 S. K. Kummerfeld and S. A. Teichmann, *Nucleic Acids Res.*, 2006, **34**, D74–81.
58 S. Moreno-Campuzano, S. C. Janga and E. Perez-Rueda, *BMC Genomics*, 2006, **7**, 147.
59 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
60 A. Bairoch and R. Apweiler, *Nucleic Acids Res.*, 2000, **28**, 45–48.
61 W. Li, L. Jaroszewski and A. Godzik, *Bioinformatics*, 2002, **17**, 77–82.
62 J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.
63 E. Sonnhammer, S. Eddy and R. Durbin, *Proteins*, 1997, **28**, 405–420.
64 N. Sierro, Y. Makita, M. de Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–96.
65 J. Mistry and R. Finn, *Methods Mol. Biol.*, 2007, **396**, 43–58.
66 M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 14863–14868.
67 A. Martinez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
68 D. E. Whitworth, *Trends Microbiol.*, 2008, **16**, 512–519.

Computational Biology and Chemistry 33 (2009) 261–268

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families

Sarath Chandra Janga [a,*,1], Ernesto Pérez-Rueda [b,*,2]

[a] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK
[b] Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnologia. Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, Mexico

ARTICLE INFO

Article history:
Received 30 September 2008
Received in revised form 16 June 2009
Accepted 17 June 2009

Keywords:
Transcription factor families
Regulatory network
Transcription machinery
Prokaryotes
Evolution

ABSTRACT

Escherichia coli K12 and *Bacillus subtilis* 168 are two of the best characterized bacterial organisms with a long history in molecular biology for understanding various mechanisms in prokaryotic species. However, at the level of transcriptional regulation little is known on a comparative scale. Here we address the question of the degree to which transcription factors (TFs) and their evolutionary families are shared between them. We found that 59 proteins and 28 families are shared between these two bacteria, whereas different subsets were lineage specific. We demonstrate that majority of the common families expand in a lineage-specific manner. More specifically, we found that AraC, ColD, Ebp, LuxR and LysR families are over-represented in *E. coli*, while ArsR, AsnC, MarR, MerR and TetR families have significantly expanded in *B. subtilis*. We introduce the notion of regulatory superfamilies based on an empirical number of functional categories regulated by them and show that these families are essentially different in the two bacteria. We further show that global regulators seem to be constrained to smaller regulatory families and generally originate from lineage-specific families. We find that although TF families may be conserved across genomes their functional roles might evolve in a lineage-specific manner and need not be conserved, indicating convergence to be an important phenomenon involved in the functional evolution of TFs of the same family. Although topologically the networks of transcriptional interactions among TF families are similar in both the genomes, we found that the players are different, suggesting different evolutionary origins for the transcriptional regulatory machinery in both bacteria. This study provides evidence from complete repertoires that not only novel families originate in different lineages but conserved TF families expand/contrast in a lineage-specific manner, and suggests that part of the global regulatory mechanisms might originate independently in different lineages.

1. Introduction

The genomes of the two model organisms, *Escherichia coli* K12 (Blattner et al., 1997) and *Bacillus subtilis* 168 (Kunst et al., 1997), contain a different proportion of Transcription Units (TU's) (Moreno-Hagelsieb and Collado-Vides, 2002), sigma factors and promoters (Salgado et al., 2006; Makita et al., 2004). Despite these basic differences, it has been possible to find some conserved and unique DNA-binding transcription factors (TFs) acting over their complete gene repertoires (Makita et al., 2004; Perez-Rueda et al., 2004). Such TFs have been related to a wide diversity of functions including catabolite repression, differentiation and cellular maintenance, among others. However, it is unclear how the collection of proteins performing similar functions (DNA-binding ability) could have evolved in these two organisms with different evolutionary history and ancestry (Hedges, 2002). Understanding the evolution of the transcriptional regulatory machinery across genomes would improve our knowledge about the evolutionary constraints that play a role in the formation of regulatory networks and would also help to decipher the design principles governing these networks across bacteria (Janga et al., 2009). Although some recent works have dealt with the evolution of the components and suggested duplication of genes as the main factor contributing to the formation of the Transcriptional Regulatory Network (TRN) (Madan Babu and Teichmann, 2003; Teichmann and Babu, 2004), there has not been comparative analysis of TFs and their families between genomes to understand the evolutionary constraints, functional aspects and design principles governing their formation. Despite the fact that there has been an increasing interest to identify and understand the regulatory repertoires of entire genomes using a variety of computational approaches (Perez-Rueda

* Corresponding authors.
E-mail addresses: sarath@mrc-lmb.cam.ac.uk (S.C. Janga), erueda@ibt.unam.mx (E. Pérez-Rueda).
[1] Tel.: +44 1223 402479; fax: +44 1223 213556.
[2] Tel.: +52 56 22 76 10; fax: +52 777 3 17 23 88.

1476-9271/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2009.06.004

et al., 2004; Brune et al., 2005; Moreno-Campuzano et al., 2006; Kummerfeld and Teichmann, 2006), there has not been genome scale comparative study reported so far to our knowledge, using representative genomes from distant lineages especially in the context of regulatory networks. Here we present the first comprehensive comparative analysis of the complete repertoires of TFs from two prokaryotic model organisms, *E. coli* K12 and *B. subtilis*.

In this work, we first identify and classify the repertoire of DNA-binding TFs of *E. coli* and *B. subtilis* into families using a previously reported approach applied to *E. coli* (Perez-Rueda and Collado-Vides, 2000). We then analyze the collection of TFs and their TF families at various levels to deduce thereof the common set of regulatory genes and families and to infer specific tendencies of TFs. Our analyses were based on the collection of TFs reported and collected from two different databases: RegulonDB (Salgado et al., 2006) for *E. coli K12* and DBTBS (Makita et al., 2004) for *B. subtilis*. Additional literature look up was performed, to retrieve a more complete dataset of TFs in these organisms. Here, we demonstrate that although *E. coli* and *B. subtilis* contain a similar proportion of DNA-binding TFs, the majority of the TF families have expanded and evolved independently. The regulatory networks based on the set of well-known TFs in both genomes suggest that the functions of genes regulated by similar families could be different. These findings open diverse opportunities to understand the complex regulatory systems in different bacteria, beyond Proteobacteria and Firmicutes.

2. Materials and Methods

2.1. Identification of TFs and Construction of TF Families in B. subtilis

In order to identify the repertoire of TFs in *B. subtilis*, we used a combination of information sources and bioinformatics tools as reported earlier (Moreno-Campuzano et al., 2006). Briefly, 237 TFs were identified by an exhaustive analysis of three sources, those TFs identified from DBTBS, a database devoted to the gene regulatory mechanisms in *B. subtilis* strain 168 (Makita et al., 2004), TFs identified by the search of family-specific Hidden Markov Models (HMMs) reported previously (Perez-Rueda et al., 2004) from *E. coli* TFs (*E*-value threshold ≥10-3), and those TFs identified with the library of HMMs from the Superfamily database (*E*-value ≥10-3) (Madera et al., 2004). This HMM library is based on the sequences of domains collected in the Structural Classification of Proteins (SCOP) database (Hubbard et al., 1997) and is thus applicable for a structural classification of proteins. In summary, the final dataset included those proteins identified by HMMs, Superfamily searches, and the repertoire (manually curated) of TFs described in DBTBS. These proteins were classified into families by using HMMs deposited in the PFAM DB (Bateman et al., 2000), and aligned by using the program *hmmalign* from HMMer. Our final collection included 90 families in *E. coli* and 51 families for *B. subtilis*. Additionally, their corresponding HMMs were used to scan a collection of 234 genomes, including bacterial, archaeal and eukaryotic species, in order to determine their evolutionary emergence in different lineages (see Supplementary Material for a complete list of genomes analyzed and the number of TFs identified across genomes).

2.2. Data of Regulatory Interactions

Transcriptional regulatory interactions of *E. coli* K12 were obtained from RegulonDB (Salgado et al., 2006), which contains experimental information extracted from literature, whereas the regulatory interactions of *B. subtilis* were retrieved from DBTBS (Makita et al., 2004). Those interactions from the datasets where a sigma factor is known to control the expression of a gene were excluded. Therefore, a total of 1816 regulatory interactions were considered for *E. coli* while 745 were included from the *B. subtilis* TRN.

2.3. Identification of Orthologs

Orthologs are defined as proteins in different species that evolved from a common ancestor by speciation (Fitch, 1970) and usually have the same function. Our working definition of orthology consisted of BLASTP reciprocal best hits, which is a widely accepted notion for identifying functional orthologs and homologous genes were identified with an *E*-value cutoff of 1e-6 as described elsewhere (Janga and Moreno-Hagelsieb, 2004).

3. Results and Discussion

3.1. Conserved TFs and TF Families Between E. coli and B. subtilis Genomes

Two proteins associated to common functions might be a consequence of common origin in different genomes (orthologous) or gene duplication within a genome after speciation (paralogous). Thus, we sought to determine the fraction of the total repertoire of TFs in *E. coli* and *B. subtilis* related by orthology and how it compares with genomic conservation. We found that 59 TFs from *E. coli* which correspond to around 20% of total TFs, had orthologs in *B. subtilis*, while around 29% of their total gene products are related by orthology which is statistically significant (see Supplementary Material), as has been previously observed about their conservation patterns using only a known subset of TFs in these genomes (Madan Babu et al., 2006; Lozada-Chavez et al., 2006). This finding suggests that TFs between the two genomes are 30% less conserved than other protein classes, indicating that TFs are likely lost to a greater extent at such phylogenetic distances (Lozada-Chavez et al., 2006). These observations give rise to several questions concerning the evolutionary and functional conservation of TFs between these bacterial genomes, so in order to have an insight into the commonalities and differences in the gene regulation between the prokaryotic species from the perspective of TFs, we used the complete repertoires of TFs in *E. coli* and *B. subtilis*. Based on diverse sequence and HMM searches, a total number of 303 *E. coli* TFs and 237 *B. subtilis* TFs were identified. These repertoires were also classified into families and compared to understand their evolutionary trends. Fig. 1 evidences the different proportions of TF families identified in the genomes. However, it can be noted that ArgR, BirA, DnaA, FrvR, LexA, PrpD and WrbA families show a very similar distribution in both the genomes. The similar proportion of these groups suggests the possibility of an early evolution of these families before the split of Proteobacteria and Firmicutes and no subsequent lineage-specific expansion or loss. A closer look at the functions of these families indicates that they are mostly involved in the synthesis of amino acids, replication and DNA repair mechanisms and metabolism of sugars. On the contrary AraC, ColD, DeoR, Ebp, IclR, LacI, LuxR, RpiR, YjhU_YdeW and YeiL families are dominant in *E. coli*, whereas ArsR, AsnC, GntR, Fur, MarR, MerR, ROK, TetR and OmpR can be seen to be dominant in *B. subtilis*. It is interesting to observe that AraC, ColD, Ebp, LuxR and LysR families are roughly double in proportion in *E. coli* than in *B. subtilis*, while ArsR, AsnC, MarR, MerR and TetR show a marked over-representation in *B. subtilis*. To test the significance of this observation and to determine if these distributions are in fact very different we performed a chi-square test, with the expected distribution in each genome calculated as the product of the total TFs from the common families and proportion of the TF family as seen in other genome. We observed a *P*-value < 10^{-53} when the familial distribution in *B. subtilis* was considered as the observed

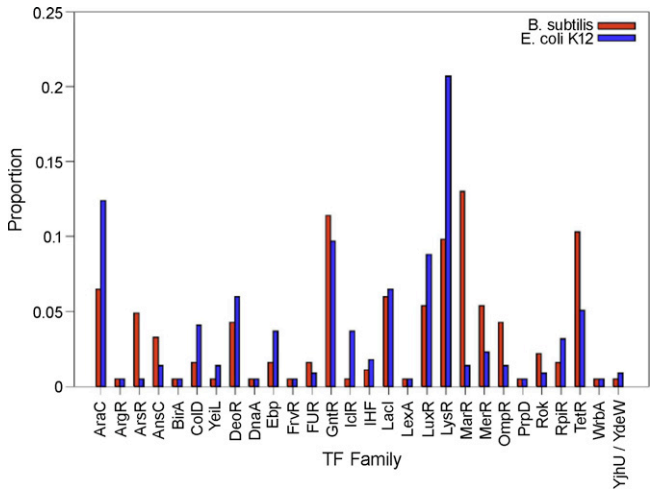


Fig. 1. Proportion of TFs in the common families identified among *E. coli* and *B. subtilis*. Proportion of TFs in each family was calculated as the fraction of TFs in a given family normalized against the total TFs that were identified in that genome in the common families. Chi-square tests against the distributions suggest that they are significantly different with *P*-values lower than 10^{-19} observed in both the genomes.

set and *P*-value < 10^{-19} when families in *E. coli* were treated as observed, indicating a significant difference in the distribution of common families. In order to test if some of the well conserved single-gene families like ArgR, BirA, DnaA, discussed above affect the *P*-values, we re-calculated them excluding these families and found them to be lower than 10^{-57} and 10^{-22} respectively, suggesting that these families indeed make the distributions more similar due to their very similar distribution in both the genomes and are probably evolutionary conserved as single copies in most prokaryotic genomes (Perez-Rueda et al., 2004; Lozada-Chavez et al., 2006; Makarova et al., 2001; Rodionov et al., 2002; Erill et al., 2004; Fujita et al., 1989). A plausible explanation for this biased lineage-specific over-representation of antibiotic related families in *B. subtilis* and metabolic and structural components in *E. coli* could be attributed to the niche in which these organisms survive. For instance, although both of them are free-living bacteria, *E. coli* has adapted to thrive inside its host and can degrade a wide variety of carbon sources thereby harboring a number of TFs responsible for degradation of carbon compounds (Martinez-Antonio et al., 2008), while *B. subtilis* has adapted to soil environments and can accept limited number of carbon sources but this is probably complemented due to its ability to form spores and starve long durations in the absence of substrates which might be driven by the excess number of sigma factors present in this bacterium.

We proceeded to analyze those DNA-binding TFs that are common to both the bacterial genomes. Table 1 shows the distribution of 59 orthologous TFs between the genomes grouped according to their association to TF families (Perez-Rueda et al., 2004). We found that the core of the TFs conserved between the genomes are involved in the regulation of amino acid biosynthesis, carbon sources assimilation, antibiotic resistance, DNA replication and repair and biosynthesis of membrane components. To identify shared and species-specific TFs quantitatively, overlooking the different evolutionary phenomenon the organisms might have undergone over time, homologous proteins were identified in the complete collection of TFs by using BLASTP matches. This comparative analysis allowed us to assess the extent of variation between the two bacteria at the level of regulatory proteins. We found that 151 TFs have homologs in both genomes, corresponding to about 52% of the TFs of *E. coli*, suggesting that the regulatory networks might be poorly conserved between the species and controlled by few homologous regulatory functions. It is also interesting to note

Table 1
Distribution of orthologous TFs between *E.coli* and *B. subtilis* into different families.

Family	Number of orthologs	Regulatory roles
LysR	7	Amino acid biosynthesis
GntR	7	NA
AraC	5	Carbon metabolism, cell wall synthesis, stress responses and pathogenesis
LuxR	3	Quorum sensing
TetR	3	Tetracycline resistance
DeoR	3	Deoxiribose assimilation
MerR	3	Mercury resistance
MarR	2	Multiple antibiotic resistance response
YjeB	2	NA
Rok	2	NA
Cold	2	Low temperature adaptation
LacI	2	Carbon sources assimilation
Fis	2	NA
YeiL	1	Global regulatory functions
PrpD	1	NA
LexA	1	DNA repair
AsnC	1	Leucine repressor protein
Fur	1	Iron assimilation
Ihf	1	Integration host factor
OmpR	1	Biosynthesis of membrane components
YjhU_YdeW	1	NA
RpiR	1	Ribose phosphate isomerase
IclR	1	Glyoxylate bypass operon
BirA	1	Biosynthesis of biotin
DnaA	1	Transcription and replication regulation
ArsR	1	Arsenic resistance
ArgR	1	Arginine biosynthesis
HipB	1	NA

Functions of the TF families were as described in a previous study on TF families in *E. coli*. TF families whose function cannot de determined either because of lack of information or due to the complexity of the members to regulate more than one function are represented as not available (NA).

that *E. coli* and *B. subtilis* contain 48% and 36% unique TFs at the BLASTP thresholds used to determine homologs, indicating that the genomes have considerably changed their repertoires of regulatory machinery.

3.2. Lineage-specific Families

Diverse families were identified exclusively in *E. coli* or *B. subtilis*. In Table 2 we present 16 families identified as specific to Firmi-

Table 2
Lineage-specific TF families identified in *B. subtilis*.

TF family	Family size	Regulatory roles
AbrB	3	Transition state genes
CodY	1	Global regulatory mechanisms
ComK	1	Late competence genes
CtsR	1	Class III stress genes
DtxR	1	Manganese transport
DUF24	3	Unknown
HxlR	2	Detoxification system regulation
LytTR	3	Regulation of autolysis
PadR	1	Phenolic acid synthesis
PRD	1	Phosphotransferase system
Psq	1	NA
PucR	1	Purine degradation
Rrf2	3	NA
RsfA	1	Prespore-specific regulation
TenA	1	Extracellular enzyme genes
Xre	17	Prophage, competence development and sporulation associated genes

TF families whose function cannot de determined because of lack of information are represented as not available (NA).

Table 3
Distribution of global TFs into different families in *E. coli* K12 and *B. subtilis* 168.

Family	Family size (*E. coli*)	Family size (*B. subtilis*)	Global TFs (*E. coli*)	Global TFs (*B. subtilis*)
YeiL	3	1	Crp/Fnr	–
Hns	2	0	Hns	–
AsnC	3	6	Lrp	–
IHF	4	2	IHF	–
Fis	8	3	Fis	–
LacI	14	11	–	CcpA
ComK	0	1	–	ComK
AbrB	0	3	–	AbrB
CodY	0	1	–	CodY
OmpR	14	8	ArcA-	Spo0A

cutes or closer lineages in terms of their taxonomic distribution. In these families there are diverse global regulators, such as ComK, AbrB and CodY, which act as switches between sporulation and free-living state in *B. subtilis*. From the perspective of *E. coli*, we found 15 characterized families which are specific to this bacterium or closely related lineages. For instance, MetJ and TrpR, the regulators of methionine and tryptophan related genes and AlpA and Crl which are known to be involved in the context of lipopolysaccharide adhesion to human gastric tissue and regulation of curly surface fibers respectively, are constrained to enterobacteria while some families like CaiF, HycA and HtgA are exclusive to *E. coli* and *Salmonella* strains. This suggests that diverse lineage-specific TFs might be involved in specific and important processes, such as sporulation in bacilli or in some specific amino acid biosynthesis routes in enterobacterial species. It is interesting to note that the absence of TFs for several important amino acid biosynthetic routes in *B. subtilis* and other Firmicutes is complemented by the invention of novel regulatory mechanisms such as transcription attenuation, despite the fact that these genomes might be responding to identical regulatory signals in the synthesis of these amino acids, suggesting the possibility for variations even in fundamental processes of the cell (Gollnick et al., 2005; Gutierrez-Preciado et al., 2005; Winkler et al., 2003; Merino and Yanofsky, 2005; Rodionov et al., 2004). In other bacteria, similar lineage-specific TFs and TF families might be expected as has been previously reported for *Streptomyces coelicolor* (Bentley et al., 2002).

3.3. Evolution of Global TFs in the Context of TF Families

Global TFs, defined as those regulatory proteins which regulate a wide variety of functional categories and have their influence on a considerable number of genes (Martinez-Antonio and Collado-Vides, 2003), provide important insights into the evolution of regulatory mechanisms in bacterial genomes. Therefore, it was our interest to understand how this class of TFs is distributed across TF families in both the genomes (see Table 3). From this table, we did not find any global TFs in common families, thus although there are common families between the two genomes, global TFs have originated from completely different TF families in different lineages. Some specific examples in this direction have been also demonstrated in other bacteria, like Crc in *Pseudomonas putida* which belongs to the endonuclease/exonuclease/phosphatase family (Morales et al., 2004) or ArlR in *Staphylococcus aureus* and PrrA in *Rhodobacter sphaeroides* which are members of two component response regulators (Liang et al., 2005; Mao et al., 2005). However, many of the global TFs occur in families identified in both the genomes except for Hns and ArcA in *E. coli* and ComK, AbrB and CodY in *B. subtilis* which occur in genome or lineage-specific families. A glance at the functions of these global TFs indicates that they are specific in their functional roles and might have evolved depending on the organism specific needs like sporulation in *B. subtilis*, essentially implying that different bacteria might have developed

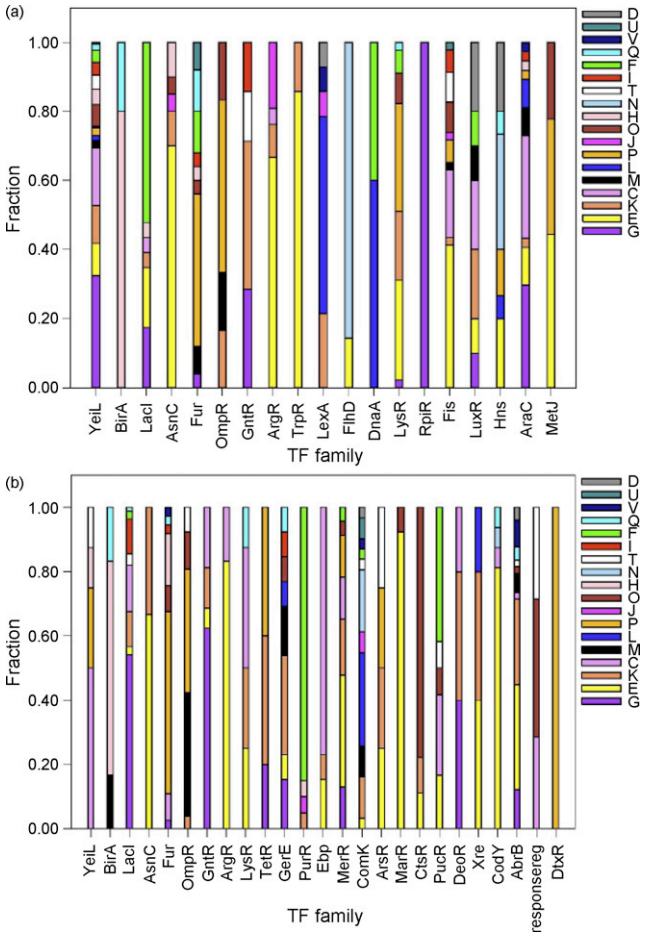


Fig. 2. Distribution of the COG categories of the regulated genes by each family of TFs in (a) *E. coli* and b) *B. subtilis*. Only those families which have more than 3 regulated genes per family are shown. The first 8 TF families correspond to the ones which exist in both the genomes. The fractions in each column are normalized against the total COG annotated genes. COG functional categories: amino acid transport and metabolism (E); carbohydrate transport and metabolism (G); energy production and conversion (C); transcription (K); cell wall/membrane/envelope biogenesis (M); replication, recombination and repair (L); inorganic ion transport and metabolism (P); translation, ribosomal structure and biogenesis (J); posttranslational modification, protein turnover, chaperones (O); signal transduction mechanisms (T); coenzyme transport and metabolism (H); cell motility (N); nucleotide transport and metabolism (F); lipid transport and metabolism (I); secondary metabolites biosynthesis, transport and catabolism (Q); defence mechanisms (V); intracellular trafficking, secretion, and vesicular transport (U); cell cycle control, cell division, chromosome partitioning (D). Correlations observed in the distribution of functional categories of the regulated genes in the common TF families: YeiL (R^2 = 0.0398), BirA (R^2 = 0.9423), LacI (R^2 = 0.0023), AsnC (R^2 = 0.8036), Fur (R^2 = 0.734), OmpR (R^2 = 0.4332), GntR (R^2 = 0.1103) and ArgR (R^2 = 0.8598).

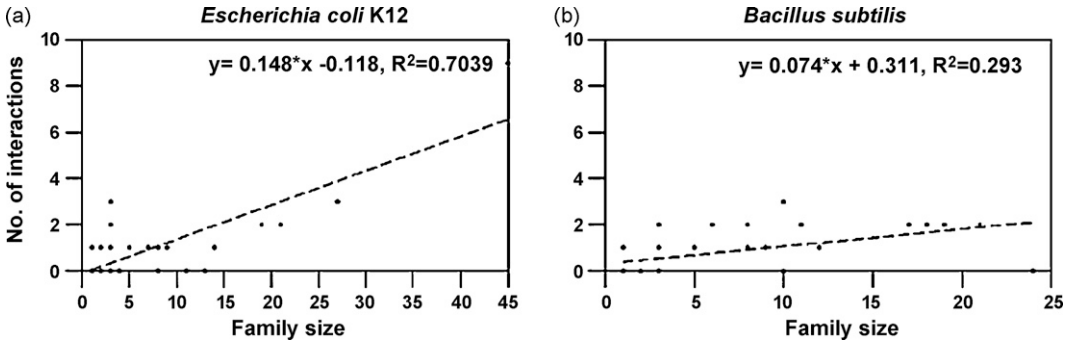


Fig. 3. TF family size versus number of regulatory interactions with in members of the same family (a) *E. coli* and (b) *B. subtilis*.

at least part of their global regulatory mechanisms independently. A second observation that can be made from the table is that most of the global TFs seem to fall into smaller TF families hinting that global TFs might avoid the cross talk over the binding sites between different members of their TF family by reducing the number of family members. Some TFs of the same family are known to bind to very similar binding sites when the sequences encoding them have significant sequence similarity as in the case of MarA, Rob and SoxS (Martin and Rosner, 2002). An alternate explanation for the observed tendency could be that large families through gene duplication could have sub-divided their regulatory functions among many TFs, thus leaving no room for global regulators in larger families. To test the significance of this observation we compared the average size of a family for a global regulator (observed to be 3.4 and 5.6 in *E. coli* and *B. subtilis* respectively) with the average family size of a general TF in 1000 randomly sampled collections each equal to the size of respective total repertoires in both the genomes. We found that the average family size of a general TF in the randomized collection followed a normal distribution and hence used *Z*-scores to calculate *P*-values. In both *E. coli* and *B. subtilis*, *P*-values $< 10^{-37}$ were observed indicating that global TFs have a strong tendency to occur in small families. Despite the reasons which can best explain the tendency, the above observation should enhance our ability to predict global TFs in other microbial genomes.

3.4. Distribution of Functional Classes in the genes Regulated by TF Families

In order to study the heterogeneity of the TFs in families in a functional context one has to compare the functions of the regulators in each TF family. However, given the poor annotations for genes encoding TFs about their specific functional roles it would be hard to use them for a comparative functional analysis. Moreover, most of the functional classification schemes for genes do not contain a detailed description for the physiological roles played by the regulators in the context of the genome being analyzed. Considering these issues we used the functional categories of the regulated genes in each family to analyze the extent of functional variation in TF families in both the organisms.

To understand the variability in the functions of the regulated genes by each family of TFs in *E. coli* and *B. subtilis* we used the COG annotations of the protein coding genes available from NCBI (Tatusov et al., 1997; Tatusov et al., 2003). In Fig. 2, we show the distribution of the COG categories of the regulated genes for TF families that are known to regulate more than 3 COG annotated genes. The families YeiL, Fis, AraC and Fur in *E. coli* and GerE, ComK, LacI, Fur and AbrB in *B. subtilis* regulate more than 7 different categories. These families can be considered as "regulatory superfamilies" in these organisms because of their ability to control diverse physiological processes. Fur is the only family which regulates a large number

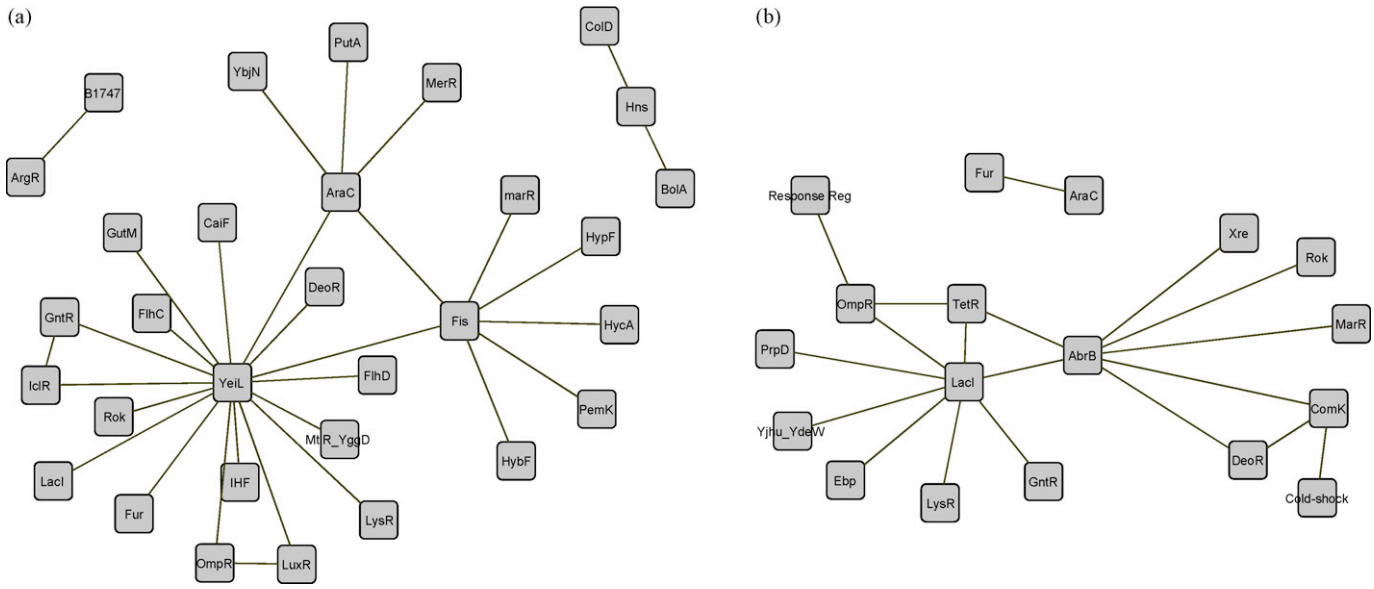


Fig. 4. Network of transcriptional interactions between different TF families identified in (a) *E. coli* and (b) *B. subtilis*. Two nodes are shown to be connected if there exists at least one regulatory interaction between the nodes.

of categories in both the bacteria, probably indicating its presence in the vital roles of the cell. Other families like GerE, ComK and AbrB suggest the evolution and expansion of their functions independently in Firmicutes. When we examined the TF families which regulate far fewer categories we found that most of them are either very restricted in their function or have an ancient origin.

A closer look into the distribution of COG categories of the 8 common families, namely YeiL, BirA, LacI, AsnC, Fur, OmpR, GntR and ArgR, between the two genomes gave further insights into the evolution of TF families in the context of functional roles. For instance, the YeiL family which composes of the Crp, Fnr and YeiL TFs in *E. coli* regulates 15 different categories and 4 in *B. subtilis*, of which the 2 categories "inorganic ion transport and metabolism" and "signal transduction mechanisms" are predominantly regulated in the second bacterium. The case of the BirA and ArgR families is interesting, because they are known to be well conserved across all the genomes (Makarova et al., 2001; Rodionov et al., 2002). Accordingly we found an appreciable overlap in the functional categories of the regulated genes in these families. TFs from the families LacI and GntR were found to preferentially regulate the functions "carbohydrate transport and metabolism" and "Transcription and energy production and conversion" in *B. subtilis* while in *E. coli* the dominantly regulated categories included "nucleotide transport", "carbohydrate transport" and "amino acid transport and metabolism" for LacI members and "transcription" and "carbohydrate transport and metabolism" for GntR members. The case of the family of Fur regulators seems to be interesting with the majority of regulated genes in both *E. coli* and *B. subtilis* belonging to "inorganic ion transport and metabolism" possibly suggesting partial conservation of the regulatory roles of its members. The family of OmpR regulators are known to be involved in the regulation of genes related to the biosynthesis of membrane components in *E. coli*, accordingly we found them to regulate the categories "inorganic ion transport and metabolism", "cell wall/membrane/envelope biogenesis", "lipid transport and metabolism" and "transcription" in both bacteria. These observations lead us to conclude that although TF families may be conserved across genomes their functional roles might evolve in an organism-specific or lineage-specific manner and are not always conserved indicating convergence to be a major phenomenon involved in the functional evolution of transcription factors of the same TF family. This finding also suggests that existence of common families between two organisms could be the result of a common ancestry initially but with speciation, functional divergence and lineage-specific expansion or contraction of TF families occurs rapidly to adapt to changing environments.

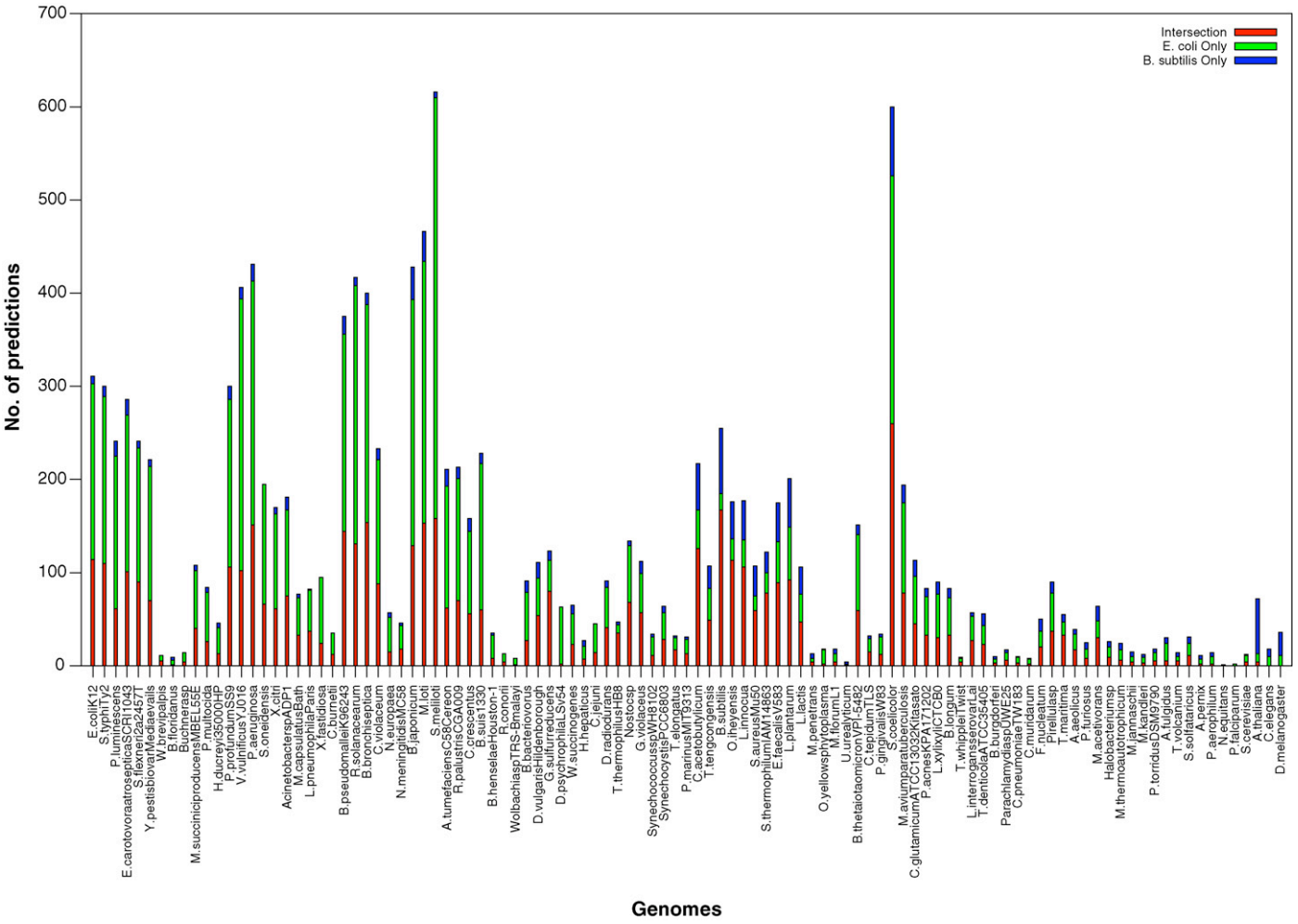


Fig. 5. Number of TFs predicted across genomes by using TF family models from *E. coli* and *B. subtilis* as the phylogenetic distance with respect to *E. coli* increases. Intersection stands for the number of TFs predicted by the models from both the genomes while the predictions identified using models in *E. coli* only are shown in green and those based on *B. subtilis* models only are shown in blue. To facilitate the display of results, we only show 105 complete genomes, obtained by filtering out strains and species of the same bacterial genus keeping the strain or species with the maximum number of genes among a given genera of organisms. The evolutionary distance from *E. coli* to all organisms was obtained according to the evolutionary branching process previously reported (Brown et al., 2001).

3.5. Interaction of TFs Within and Across Families

In order to understand if TFs in a given family interact with each other to regulate biological processes, we sought to see any relation between the number of regulatory interactions among members in a given family and its family size. In Fig. 3 we show the number of transcriptional interactions between members of the same family of TFs as the family size increases. It appears from this figure that interactions among members of a TF family increase as the family size increases, although the number of interactions is always low in both the genomes.

To study the interaction of TFs from different families we identified the transcriptional interactions between the regulators belonging to different families. As shown in Fig. 4 we observed a scale-free like topology when the interactions between TF families were modeled as a network. Some of the well-connected families are those containing the global regulators; however, it should be noted that the families which are responsible for the scale-free nature in *E. coli* and *B. subtilis* are different as has been observed in the TRNs of these genomes (Madan Babu et al., 2006). It is interesting to visualize from this figure the case when one of the well-connected hubs like Fis or AraC in *E. coli* or AbrB or ComK in *B. subtilis* but not the central node is removed from their genomes, which would lead to removal of a branch of the interactions rather than lethality of the whole network (and hence the cell) which is typically what is observed in scale-free networks and has been described as robustness (Albert et al., 2000). However, in this context, robustness might refer to the conditions of growth in which these regulatory families are no longer needed by the cell to regulate its processes. Although the data of the TRN of *B. subtilis* is smaller in size compared to *E. coli*, these observations allow us to conclude that scale-free nature in the networks of TF families is common to both the genomes and might have evolved to choose different nodes as hubs despite the existence of common families.

3.6. Effect of Lineage Specificity in Predicting TFs From Comparative Genomics

Despite the poor conservation of the TFs between the two bacteria, we wanted to determine how much comparative genomics can help to identify TFs across organisms using the family-specific HMMs developed in these bacteria and how much overlap the predictions based on the models from different organisms might have in a given genome. We therefore identified the repertoires of TFs in complete genomes using the models from both the genomes (see Section 2 and Supplementary Material). In Fig. 5 the number of predicted TFs across 105 complete non-redundant genomes from the perspective of both the genomes is shown. It is clear from the figure that although the number of TFs predicted from *B. subtilis* perspective is lower than that from *E. coli*'s, there is almost a complete overlap in the predictions between the two sets across genomes suggesting that comparative genomics approaches based on family-specific HMMs as against homology based approaches which typically search similarity across the entire length of the sequence can be very powerful to predict TFs with a high positive predictive value (calculated as True Positives/(True Positives + False Positives)). For example in *B. subtilis* we identified 185 TFs using *E. coli* based models of which 167 were a subset of 237 TFs identified in this bacterium, similarly we identified 122 TFs in *E. coli* based on *B. subtilis* models of which 114 were a subset of the collection identified earlier in *E. coli* (Perez-Rueda and Collado-Vides, 2000). It is also easy to note from the figure that as the evolutionary distance with respect to *E. coli* increases (in Archaea and Eukarya) the predictive coverage drops rapidly indicating the loss of domain level signal at such distances. A second observation to note is that *B. subtilis* models tend to predict slightly higher proportion of TFs in closer lineages like Bacillales and Lactobacillales while *E. coli* models clearly dominate the number of predictions in all proteobacterial lineages suggesting the effect of lineage or genome-specific expansion of TFs playing an important role in identifying TFs across genomes.

These observations suggest that although this approach to identify TFs can produce high quality predictions, the limiting factor can be the evolutionary distance because at large evolutionary distances it would be hard to trace the repertoires of TFs not only due to the poor conservation of domains but also due to the evolution of novel TF families as has been demonstrated in this work. However, as the experimental knowledge about the TFs from lineage-specific families increases it should be possible to expand the repertoires of TFs across prokarya beyond the few model organisms that are the focus of the study.

4. Conclusions

Based on genome analysis we defined the individual set of DNA-binding TFs in *E. coli* and *B. subtilis* genomes and deduced thereof the common repertoire of transcriptional regulators and regulatory families of these species. The set of the well-conserved TFs between the two genomes is involved in fundamental cellular processes and could have an ancient origin. We show that TF families evolve rapidly and expand in a lineage-specific manner to adapt to varying environmental needs of the organisms. Similar trends have been observed in previous comparative studies on TF families in plants versus animals and at the level of taxa (Shiu et al., 2005; Coulson et al., 2001). A more general perspective of lineage-specific expansion of protein families and its implications on the diversification of organisms has also been shown in eukaryotic species (Lespinet et al., 2002). Our results show that global TFs responsible for global regulatory mechanisms in bacteria can evolve independently in different organisms and from totally different regulatory families, suggesting that transcriptional regulatory machinery plays a very important role in the speciation of organisms. We observe that global regulators have a tendency to occur in smaller and lineage-specific families which might be of recent origin indicating a source for the innovation of novel regulatory interactions and mechanisms across different lineages while still keeping the genetic repertoire well conserved. Our findings show that larger TF families regulate disproportionately low number of genes. It is possible that these large families function as local modules of regulation while the smaller families act as major hubs of the Transcriptional Regulatory Network.

It is interesting to speculate the variation of regulatory networks across prokaryotic organisms at three different levels (a) variation of regulon composition due to the re-organization of genomic context of genes accompanied by changes in the cis-regulatory regions in closely related species, though preserving the mode of action of TFs (Espinosa et al., 2005) (b) variation at the level of repertoire of TFs due to the need for different requirements of regulatory machinery in different environments (Madan Babu et al., 2006) (c) variation at the level of the regulatory mechanisms employed to perform the same biological process, as is seen in the case of attenuation mechanisms replacing transcriptional regulation in some bacteria. While the variations at the first level can be believed to occur mostly in the same phylogenetic group/lineage and can be treated analogously to changes affecting a given valley of mountains and the variations at the second and third levels could be major reasons for differentiation of lineages/phylogenetic groups analogous to differences between valleys.

Supplementary Material

Supplementary material can be accessed at: http://tikal.ccg.unam.mx/sarath/tfevolution/.

Acknowledgements

We would like to thank Gabriel Moreno-Hagelsieb, Martin Peralta-Gil and Bruno Contreras-Moreira for helpful discussions and comments on the initial versions of this manuscript. We would also like to thank Irma Lozada-Chávez for helping us with the phylogenetic analysis. SCJ acknowledges support from Medical Research Council, Laboratory of Molecular Biology and Cambridge Commonwealth Trust. EP-R was financed by a grant (IN-217508) from DGAPA-UNAM, and by grants given to Lorenzo Segovia.

References

Albert, R., Jeong, H., Barabasi, A.L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382, PMID: 10935628.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266, PMID: 10592242.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., Hopwood, D.A., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417, 141–147, PMID: 12000953.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence data sets. Nat Genet 28, 281–285.

Blattner, F.R., Plunkett 3rd, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

Brune, I., Brinkrolf, K., Kalinowski, J., Puhler, A., Tauch, A., 2005. The individual and common repertoire of DNA-binding transcriptional regulators of Corynebacterium glutamicum, Corynebacterium efficiens, Corynebacterium diphtheriae and Corynebacterium jeikeium deduced from the complete genome sequences. BMC Genomics 6, 86.

Coulson, R.M., Enright, A.J., Ouzounis, C.A., 2001. Transcription-associated protein families are primarily taxon-specific. Bioinformatics 17, 95–97.

Erill, I., Jara, M., Salvador, N., Escribano, M., Campoy, S., Barbe, J., 2004. Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. Nucleic Acids Res 32, 6617–6626.

Espinosa, V., Gonzalez, A.D., Vasconcelos, A.T., Huerta, A.M., Collado-Vides, J., 2005. Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. J Mol Biol 354, 184–199.

Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. Syst Zool 19, 99–113.

Fujita, M.Q., Yoshikawa, H., Ogasawara, N., 1989. Structure of the dnaA region of Pseudomonas putida: conservation among three bacteria, *Bacillus subtilis*, *Escherichia coli* and *P. putida*. Mol Gen Genet 215, 381–387.

Gollnick, P., Babitzke, P., Antson, A., Yanofsky, C., 2005. Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. Annu Rev Genet 39, 47–68.

Gutierrez-Preciado, A., Jensen, R.A., Yanofsky, C., Merino, E., 2005. New insights into regulation of the tryptophan biosynthetic operon in Gram-positive bacteria. Trends Genet 21, 432–436.

Hedges, S.B., 2002. The origin and evolution of model organisms. Nat Rev Genet 3, 838–849.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Danchin, A., et al., 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

Hubbard, T.J., Murzin, A.G., Brenner, S.E., Chothia, C., 1997. SCOP: a structural classification of proteins database. Nucleic Acids Res 25, 236–239.

Janga, S.C., Moreno-Hagelsieb, G., 2004. Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res 32, 5392–5397.

Janga, S.C., Salgado, H., Martinez-Antonio, A., 2009. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. Nucleic Acids Res. 37, 3680–3688.

Kummerfeld, S.K., Teichmann, S.A., 2006. DBD: a transcription factor prediction database. Nucleic Acids Res 34, D74–81.

Lespinet, O., Wolf, Y.I., Koonin, E.V., Aravind, L., 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 12, 1048–1059.

Liang, X., Zheng, L., Landwehr, C., Lunsford, D., Holmes, D., Ji, Y., 2005. Global regulation of gene expression by ArlRS, a two-component signal transduction regulatory system of Staphylococcus aureus. J Bacteriol 187, 5486–5492.

Lozada-Chavez, I., Janga, S.C., Collado-Vides, J., 2006. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 34, 3434–3445.

Madan Babu, M., Teichmann, S.A., 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res 31, 1234–1244.

Madan Babu, M., Teichmann, S.A., Aravind, L., 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358, 614–633.

Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., Gough, J., 2004. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 32, D235–239.

Makarova, K.S., Mironov, A.A., Gelfand, M.S., 2001. Conservation of the binding site for the arginine repressor in all bacterial lineages. Genome Biol 2 (4), RESEARCH0013.

Makita, Y., Nakao, M., Ogasawara, N., Nakai, K., 2004. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Res 32, D75–D77.

Mao, L., Mackenzie, C., Roh, J.H., Eraso, J.M., Kaplan, S., Resat, H., 2005. Combining microarray and genomic data to predict DNA binding motifs. Microbiology 151, 3197–3213.

Martin, R.G., Rosner, J.L., 2002. Genomics of the marA/soxS/rob regulon of *Escherichia coli*: identification of directly activated promoters by application of molecular genetics and informatics to microarray data. Mol Microbiol 44, 1611–1624.

Martinez-Antonio, A., Collado-Vides, J., 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol 6, 482–489.

Martinez-Antonio, A., Janga, S.C., Thieffry, D., 2008. Functional organisation of *Escherichia coli* transcriptional regulatory network. J Mol Biol 381, 238–247.

Merino, E., Yanofsky, C., 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21, 260–264.

Morales, G., Linares, J.F., Beloso, A., Albar, J.P., Martinez, J.L., Rojo, F., 2004. The Pseudomonas putida Crc global regulator controls the expression of genes from several chromosomal catabolic pathways for aromatic compounds. J Bacteriol 186, 1337–1344.

Moreno-Campuzano, S., Janga, S.C., Perez-Rueda, E., 2006. Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes—a genomic approach. BMC Genomics 7, 147.

Moreno-Hagelsieb, G., Collado-Vides, J., 2002. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 (Suppl. 1), S329–S336.

Perez-Rueda, E., Collado-Vides, J., 2000. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. Nucleic Acids Res 28, 1838–1847.

Perez-Rueda, E., Collado-Vides, J., Segovia, L., 2004. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput Biol Chem 28, 341–350.

Rodionov, D.A., Mironov, A.A., Gelfand, M.S., 2002. Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Genome Res 12, 1507–1516.

Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S., 2004. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. Nucleic Acids Res 32, 3340–3353.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A., Collado-Vides, J., 2006. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 34, D394–D397.

Shiu, S.H., Shih, M.C., Li, W.H., 2005. Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol 139, 18–26.

Teichmann, S.A., Babu, M.M., 2004. Gene regulatory network growth by duplication. Nat Genet 36, 492–496.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.

Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R., 2003. An mRNA structure that controls gene expression by binding S-adenosylmethionine. Nat Struct Biol 10, 701–707.

Structure and organization of drug-target networks: insights from genomic approaches for drug discovery†

Sarath Chandra Janga*[a] and Andreas Tzakos‡*[b]

Received 23rd April 2009, Accepted 12th August 2009
First published as an Advance Article on the web 4th September 2009
DOI: 10.1039/b908147j

Recent years have seen an explosion in the amount of "omics" data and the integration of several disciplines, which has influenced all areas of life sciences including that of drug discovery. Several lines of evidence now suggest that the traditional notion of "one drug–one protein" for one disease does not hold any more and that treatment for most complex diseases can best be attempted using polypharmacological approaches. In this review, we formalize the definition of a drug-target network by decomposing it into drug, target and disease spaces and provide an overview of our understanding in recent years about its structure and organizational principles. We discuss advances made in developing promiscuous drugs following the paradigm of polypharmacology and reveal their advantages over traditional drugs for targeting diseases such as cancer. We suggest that drug-target networks can be decomposed to be studied at a variety of levels and argue that such network-based approaches have important implications in understanding disease phenotypes and in accelerating drug discovery. We also discuss the potential and scope network pharmacology promises in harnessing the vast amount of data from high-throughput approaches for therapeutic advantage.

[a] *MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 0QH. E-mail: sarath@mrc-lmb.cam.ac.uk; Fax: +44-1223-213556; Tel: +44-1223-402479*
[b] *Institut de Biologie Structurale et Microbiologie, CNRS, Marseille, France. E-mail: atzakos@cc.uoi.gr; Fax: +30-26510-97200; Tel: +30-26510-08387*
† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.
‡ Current address: Department of Chemistry, Section of Organic Chemistry and Biochemistry, University of Ioannina, Ioannina, Gr-45110, Greece

Introduction

In living organisms, viability and functionality is accomplished through a constant flow of information transmitted through interactions between the basic building blocks RNA, DNA, proteins and small molecules. This "biological cosmos", represented as a global biological network, although inherent in its complexity, is bound with stability and equilibrium. Any change that irreversibly distorts the equilibrium in this network could result in pathological conditions and hence

Sarath Chandra Janga

Sarath Chandra Janga is a PhD student at the MRC Laboratory of Molecular Biology and University of Cambridge. Sarath obtained his Bachelors and Masters in Bio-chemical engineering and Biotechnology at the Indian Institute of Technology, Delhi in 2003. Prior to starting his PhD, Sarath worked extensively and co-ordinated a number of research projects, on transcriptional regulation, genome organization and comparative genomics in bacteria, at UNAM in Mexico. He has published more than 25 research manuscripts on various aspects of prokaryotic and eukaryotic biology in the fields of computational molecular and systems biology. His current research interests include understanding the design principles and constraints imposed on post-transcriptional and post-translational gene control in prokaryotic and eukaryotic organisms.

Andreas Tzakos

Andreas Tzakos obtained his doctoral degree at the Department of Chemistry, Section of Organic Chemistry and Biochemistry, University of Ioannina, Greece. After postdoctoral research at the MRC Laboratory of Molecular Biology and CNRS, he is currently (since 2009) a Lecturer at the University of Ioannina and a member of the recently established Human Cancer Biobank Center, Molecular Oncology Lab, University of Ioannina. His current interests include decoding the molecular mechanisms, which underlie cancer development using multi-disciplinary approaches in the interfaces of chemistry, biology and medicine.

confer disease. It is increasingly becoming clear that for a drug to combat a disease effectively, it should target not a single but several building blocks in this biological network[1–4] to re-establish the equilibrium state. This emerging field is often referred to as network pharmacology.[3,4] Network pharmacology resembles the LeChatelier's principle of biological networks, *i.e.* if a system (biological network) at equilibrium (healthy state) experiences a change (disease state), the effective drug will shift the equilibrium in order to minimize that change (Fig. 1A). For over twenty years the focus of drug development has been to develop highly selective molecules. However, recent improvements in high throughput screening methods and advances in genomics, transcriptomics, proteomics and metabolomics have enabled us to gather large amounts of data on drug-target interactions in several model organisms and this picture is starting to be challenged.[5–10] This is also driven by the fact that the cost for discovery of new drugs continues to rise disproportionately in comparison to the approval rates despite the munificent investment in screening technologies and genomics.[11] The costs of discovering and developing a drug are in the order of US $900 million. The majority of this cost is spent in the later stages of the pharmaceutical pipeline,[11–13] as can be ascribed in the form of a drug development clepsydra (Fig. 1B). Unfortunately, the attrition rates are reported to be higher in the later phases (Phases IIb and III) of full clinical development, where most of the cost will have been apparently invested in the wrong direction due to molecules that failed to pass the final stages.[12,13] The most important reasons for drugs failing in development are either due to inadequate efficacy or their inability to pass the safety standards, as initial screenings done on animal models can often be unpredictive, with other factors like complexity of the disease playing an important role as well.[13] Recent surveys suggest that the success rates are less than 10%, with this figure being even more worrisome for drugs targeting novel mechanisms, since they have higher attrition rates.[13,14] Attrition rates are not equally distributed across different therapeutic areas and remarkable differences have been reported,[13] with oncology suffering from higher failure rates in Phase II and III trials.[13,15] Fortunately, these figures are not irreversible and multidisciplinary research can identify and remedy the causes of attrition. For instance, it is clear that the rate of attrition of compounds with novel mechanisms of action is higher than those with previously precedent mechanisms of action, suggesting a need to focus on already approved drugs for new therapeutic benefits.[13] Indeed, an emerging, promising and cost efficient direction of drug development surmounting the risk of toxicity and efficacy is the area of finding new therapeutic uses for approved drugs, so called drug repurposing.[16] It is anticipated that decoding the molecular pathophysiology of disease, through global understanding of disease heterogeneity and connection between diseases and targets in the biological network, will eventually lead to improved target validation. This is evident even in the therapeutic area of oncology with attrition rates in the order of 82% as a whole, wherein if one considers the clinical success of multitargeted kinase inhibitors as a subset, the attrition rate is only 53%, emphasizing the importance of polypharmacological approaches to drug discovery.[17] All these challenges that pharmaceutical industries are facing call for novel ideas, approaches and methodologies for speeding up the drug development process using our current understanding. In particular, more effective tools will be needed to critically analyze the information flow in the early stages of the drug discovery and development pipeline. Network pharmacology, which is built on the foundations of multi-target drugs *i.e.*, polypharmacology, could be a strong asset to treat complex diseases such as cancer. This paradigm shift together with the explosion of information from several multi-disciplinary areas aggregated into three dimensions: Drug space—comprising of small molecules, Target space—comprising of the cellular interactome available for small molecules and Disease space—comprising of the disease states an organism encounters, has brought great attention in drug discovery circles. In this review, we discuss recent advancements in Drug, Target and Disease spaces in the context of this paradigm and propose new research venues in light of these recent findings.

Drug space

Sampling for biologically active compounds in the vast drug space

Several advances in medicinal chemistry, including parallelization and miniaturization of synthetic compounds, have increased dramatically the synthesis and screening of thousands of compounds against a single target.[18] Combinatorial chemistry is widely used to build large libraries of many thousands of compounds both for the identification and optimization of lead compounds. The design of these libraries followed different philosophies and approaches. Initial efforts to chart the global drug space followed probabilistic approaches implementing chemical libraries of large size and diversity.[19] However, such approaches were only successful in identifying hits but not lead compounds.[20] The reason is simple: the chemical universe is just vast and may contain 10^{20}–10^{200} molecules.[20,21] Thus, "when trying to find a needle in a haystack, the best strategy might not be to increase the size of the haystack".[20] Therefore, over the years a more rational design of chemical libraries was required for more successful optimizations which included preserving the drug-like properties and generating pharmacophore mapping libraries which can possess attributes of drugs with minor variations in the backbone structure.[22–24] Due to the enormous size of the chemical space, a thorough experimental exploration is not feasible and thus novel methodologies and strategies are required to effectively and intelligently map the sub-portion of the chemical space that is of biological relevance. Current trends in the design strategy of chemical libraries include high diversity in molecular properties such as hydrophobicity and hydrogen bond donors/acceptors, variations in the backbone and scaffold of the molecules (skeletal diversity)[25–27] and diversity in the spatial placement of atoms in the 3D space (stereochemical diversity).[26–30] These efforts aim to sample in an effective manner the chemical and conformational space and increase the likelihood to discover a novel hit. Since the number of molecules is large and only a small fraction can be potential drugs which can satisfy a number of constraints such as bioavailability, cell membrane

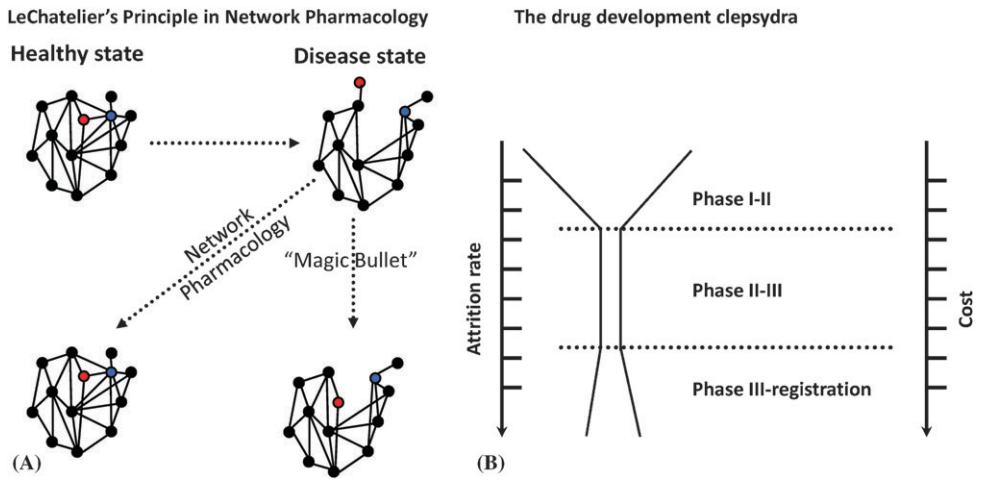


Fig. 1 (A) Network pharmacology resembles the LeChatelier's principle in biological networks, *i.e.* if a system (biological network) at equilibrium (healthy state) experiences a change (disease state), the effective drug will shift the equilibrium in order to minimize that change. (B) The drug development clepsydra over the different phases of drug development (I, II, III and registration). The sizes of the clepsydra correlate with the number of tested drug candidates, the cost incurred and attrition rate over the process of drug development. Initially, there are a huge number of molecules that enter the drug development pipeline but this shrinks over time in contrast to the attrition rate and effective cost.

permeability and non-toxicity, a number of *in silico* approaches have been developed in the past years to pre-filter drugs in the early stages of their synthesis. One of the most popular is Lipinski's rules which is a molecular property filter developed after analysis of marketed drugs and describes in a quantitative manner the cut-off or upper-limits for a number of molecular properties like hydrogen bond donor/acceptor, molecular weight, rotatable bonds, solubility, *etc*.[31,32] Recommendations in terms of bioavailability, solubility and drug-likeness have also been constructed on the basis of predicted physicochemical properties from the 2D structure. Although several properties and guidelines exist, most commercial drugs satisfy Lipinski's,[22] Veber's,[33] Bergström's[34] and Wenlock's[35] recommendations in terms of solubility, bioavailability and drug-likeness. Prediction of pharmacokinetic properties is also considered early on in the drug development pipeline and chemical libraries are screened for absorption, distribution, metabolism and excretion properties (ADME).[36,37]

There are several databases that have accumulated information on chemical molecules and drugs (see examples in Table 1). Such databases contain judicious information on the drug space currently available to be navigated. Several computational algorithms have been employed to explore this chemical space from 1D to 3D.[38–41] Artificial intelligence, machine learning and pattern recognition approaches have been recruited and gained determinant roles in rational drug design and screening of candidate molecules.[42] There is now a growing tendency to sculpt a drug for multiple targets since it can be a strong asset for the treatment of numerous disorders. A common tactic is to take as a framework a drug that is well-established for a given disease and introduce additional functionalities to enhance its efficacy and reduce its side-effects (see the section of target space for detailed discussion). The trend is to identify more promiscuous drugs, drugs that can recognize multiple targets following upon the notion of polypharmacology. All these observations indicate the importance of new approaches to efficiently and intelligently navigate the vast drug space to identify the desired multi-target drugs.

Guidelines and methods to construct functional ligand promiscuity

Ligand promiscuity is a plus according to the paradigm of network pharmacology. However, the scope is not a generic transformation of compounds affecting a single node in a disease network to compounds perturbing non-selectively several nodes.[3] On the contrary, the aim is to affect the ideal combination of nodes, which will only perturb the disease state to restore it to its natural un-diseased state, by creating functional promiscuous ligands (Fig. 1A). Given the enormous size of drug and target space, an extensive exploration seems impossible and makes rational design of functional ligand promiscuity rather difficult. Therefore, a more focused navigation towards fractions or regions of the chemical space with increased likelihood to contain biologically active and promiscuous ligands is required. Several studies have identified general qualitative physicochemical and structural principles for sculpting promiscuous molecules capable of binding to multiple targets.[43–45] Generally, it has been proposed that ligand promiscuity is favoured by molecules exhibiting specific physicochemical criteria such as: (a) low molecular weight *i.e.*, small size, (b) increased hydrophobicity—such ligands are closer to the centre of the biological charge space and are not very sensitive to differences in the shapes of targets,[44] (c) conformational flexibility—which allows for increased binding affinity to multiple partners, however, induces higher specificity for polar and charged ligands,[44] (d) asymmetric groups can also lead to increased promiscuity, (e) increased molecular complexity of a ligand reduces the probability to recognize multiple targets as it would also increase the mismatch probability between the ligand and the targets.[46]

Several methods have been suggested for the construction of ligands with promiscuity, with the predominant technique

Table 1 Representative set of databases and resources for chemical molecules, drugs and their targets

Name	Descriptions	Website
DrugBank[86]	DrugBank is a bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information in several model organisms. It contains over 4800 drug entries including >1350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and >3243 experimental drugs.	http://www.drugbank.ca/
STITCH[100]	STITCH integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug-target relationships. It contains interaction information for over 68 000 different chemicals, including 2200 drugs, and connects them to 1.5 million genes across 373 genomes.	http://stitch.embl.de/
WOMBAT	WOMBAT contains over 11 000 medicinal chemistry papers, over 260 000 records with 550 000 bioactivities on more than 2200 targets.WOMBAT-PK contains 1228 approved drugs and active metabolites, totalling over 10 000 pharmacokinetic, toxicity and drug-target related endpoints, covering more than 550 drug targets.	http://www.sunsetmolecular.com/
BindingDB[120]	BindingDB is a database containing measured binding affinities, focusing on interactions between proteins, considered to be drug-targets and small, drug-like molecules. It contains about 28 000 small molecules with activity data (55 000 experimentally determined binding affinities) for about 600 protein targets.	http://www.bindingdb.org
KEGG DRUG[121]	KEGG DRUG is a chemical structure based information resource for all approved drugs in Japan and the USA, with many also approved in Europe.	http://www.genome.jp/kegg/drug/
ChEBI[122]	Chemical Entities of Biological Interest (ChEBI) is a dictionary of molecular entities focused on small chemical compounds. The molecular entities are either natural products or synthetic products, used to intervene in the processes of living organisms. In addition to molecular entities, ChEBI contains groups (parts of molecular entities) and classes of entities.	http://www.ebi.ac.uk/chebi/init.do
ChemDB[123]	ChemDB is a chemical database containing nearly 5 million commercially available small molecules, important for their use as synthetic building blocks, probes in systems biology and as leads for the discovery of drugs and other useful compounds. The chemical data includes predicted or experimentally determined physicochemical properties.	http://cdb.ics.uci.edu
Zinc[124]	Zinc is a database of over 8 million molecules with their 3D structure and annotated with physicochemical properties. Each molecule in the library contains vendor and purchasing information and is ready for docking using a number of popular docking programs.	http://zinc.docking.org
Supertarget[125]	SuperTarget integrates drug-related information about medical indication areas, adverse drug effects, drug metabolism, pathways and Gene Ontology terms of the target proteins. Provides tools for 2D drug screening and sequence comparison of the targets. The database contains more than 2500 target proteins, which are annotated with about 7300 relations to 1500 drugs.	http://insilico.charite.de/supertarget/
MATADOR[125]	MATADOR is a resource for protein-chemical interactions (both direct and indirect).	http://matador.embl.de/
MMsINC[126]	MMsINC is a database of non-redundant, annotated, and biomedically relevant chemical structures. The current database contains about 4 million unique compounds.	http://mms.dsfarm.unipd.it/MMsINC/search/
ChemSpider	ChemSpider provides access to millions of chemical structures and integration to a multitude of other online services.	http://www.chemspider.com/
ChemBank[127]	ChemBank is a database that includes data derived from small molecules and small-molecule screens and resources for studying this data. It stores an increasingly varied set of measurements derived from cells and other biological assay systems treated with small molecules. Analysis tools are available that allow the relationships between small molecules, cell measurements, and cell states to be studied. It stores information on hundreds of thousands of small molecules (1 273 443) and hundreds of biomedically relevant assays (4742).	http://chembank.broad.harvard.edu
MDDR	MDDR contains over 150 000 biologically relevant compounds and well-defined derivatives. Updates add about 10 000 substances a year to the database. MDDR covers patent literature, journals, meetings and conference proceedings.	http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp
PubChem[128]	PubChem focuses on the chemical, structural and biological properties of small molecules, particularly their application as diagnostic and therapeutic agents. It comprises of over 19.6 million compounds with over 11 million unique structures.	http://pubchem.ncbi.nlm.nih.gov/search/search.cgi

being referred to as the pharmacophore combination approach.[47–52] In this approach, different pharmacophores from an array of selective ligands are joined together. This technique is a knowledge-based approach since it requires the knowledge of the structure-activity relationship (SAR) of the functionalities of the ligands to be joined.[48–50,52] Different approaches for connecting pharmacophores into single entities have been demonstrated in the literature.[47–51] Typically pharmacophore connections are achieved *via* incorporation of a cleavable or a non-cleavable conjugated spacer.[47–52] Merging of pharmacophores through cleavable conjugates resembles drug cocktails since the different pharmacophore-drugs are separated after administration in to two independent moieties commonly *via* plasma esterases[53,54] (Fig. 2A). The pharmacophore combination approach with linkers normally suffers by not providing compounds with good oral drug-like properties due to the inevitable increase of the molecular weight and complexity of the resulting compounds. Other approaches targeting high merging through overlapping[55] (Fig. 2B) or integration[56] (Fig. 2C) of the different pharmacophores may lead to smaller and simpler molecules surmounting problems of unfavourable physicochemical properties.

An alternative to the SAR knowledge-based approaches are the screening approaches of diverse or focused compound libraries for activity on different targets.[47,49] The later approaches are especially valuable for cases where there is lack of selective ligands for the targets of interest or a combination of unrelated receptor families is to be targeted. Once a compound is identified having a predetermined set of requirements, a heavy elaboration follows to improve binding affinity and drug-like properties. This can be either performed through "fragment evolution",[57] where a systematic incorporation of chemical functionalities to the starting core is attempted or "fragment linking"[57] that basically follows the approaches of pharmacophore connection mentioned above.

Approaches to generate effective promiscuous drugs

Promiscuous drugs are typically developed by employing one of the three approaches outlined below. Firstly, in drug-repurposing approach, knowledge about existing old drugs or other historical compounds available from literature or proprietary company sources are exploited. This involves discovering new therapeutic uses for approved drugs.[16] Drugs have been traditionally designed to have unidirectional character interacting with a single target that was relevant to the disease of interest and hence during the drug optimization process, very limited attention was given to address properly the issue of target selectivity. One of the most interesting examples in this direction is that of aspirin, which was originally developed

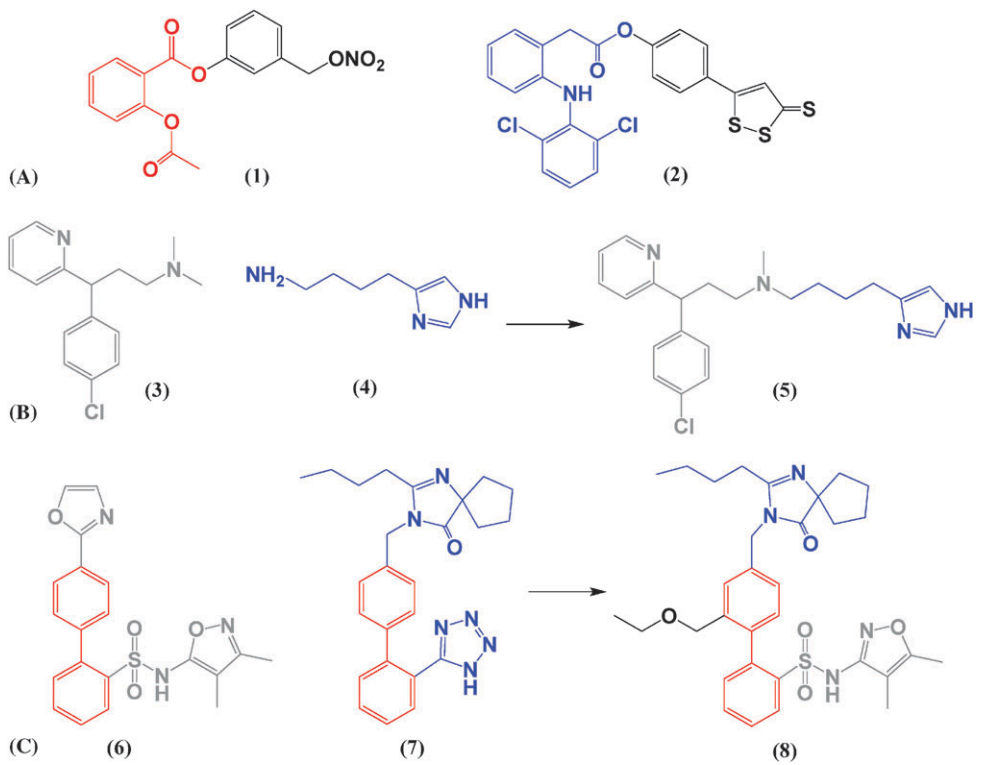


Fig. 2 Examples of different pharmacophore combination approaches to design promiscuous ligands. (A) Pharmacophore connection through cleavable ester linker for a nitric oxide-releasing derivative of aspirin (1)[53] and a hydrogen sulfide-releasing derivative of diclofenac (2).[54] The drugs aspirin and dichlofenac are coloured in red and blue respectively. (B) Pharmacophore merging through overlapping of pharmacophores from antagonists of histamine receptors H1 (3), coloured in grey, and H3 (4), coloured in blue, to construct the dual H1/H3 antagonist (5). Merging was made *via* the amine moieties that are common to both (3) and (4).[55] (C) Pharmacophore merging through integration of pharmacophores from an endothelin A receptor antagonist (6) and an angiotensin II receptor antagonist (7) to construct the dual endothelin A/angiotensin II receptor antagonist (8).[56] In grey and red are the pharmacophores used from compounds (6) and (7) respectively. This high pharmacophore integration was achieved since starting compounds shared a common biphenyl core (coloured in red).

to combat arthritis but it was found later on to have antipyretic, analgesic, anti-inflammatory, anti-platelet activities, to inhibit the synthesis of prothrombin, promote apoptosis and to have cancer preventive effects.[58] The capability of *in silico* target profiling methods to identify new targets for old drugs as also to alert for potential off-target effects has been demonstrated.[58–60]

Natural products (NPs) have a dominant role in pharmacology, since almost 60% of anticancer compounds and 75% of drugs for infectious diseases are either natural products or natural product derivatives and hence form an important source of chemicals for natively increasing promiscuity.[61–63] An important feature exhibited by them is their ability to interact with multiple targets and modulate multiple signal transduction pathways. One example is quercetin that targets cancer prevention at several levels due to its favourable anti-mutagenic and anti-proliferative effects, its role in the regulation of cell signaling, cell cycle and apoptosis.[64] NPs have been evolutionary selected after nature's combinational chemistry to have chemical diversity and interact with multiple biological target molecules.[62,65,66] In addition, natural products often resemble endogenous metabolites or biosynthetic intermediates, thus favourably operating in active transport mechanisms.[67] Therefore, the investigation of such compound collections in biochemical and biological screens should yield high hit rates at comparably small library size and will be an important source for the identification of small molecules for multi-target compounds.[65,68] From analysis of the drug-like properties of NPs that have been approved as drugs, it was found that they could be divided in two equal subsets.[67] The first subset follows Lipinski's rules and the second violates them. Interestingly, nature through its multiple combinatorial design efforts succeeded to bypass Lipinski constraints for large compounds in terms of molecular weight and number of rotatable bonds maintaining at the same time low hydrophobicity and intermolecular H-bond donating potential.[67] Unexpectedly, both subsets had identical success rate in delivering an oral drug.[67] We should therefore take lessons from nature's effort to design multi-target compounds. A comparison of the molecular property profiles in combinatorial libraries, natural products and marketed drugs indicated a broader distribution and increased diversity in natural products and drugs compared to combinatorial compounds.[69] Therefore, a good strategy to increase the chemical space charting by combinatorial library efforts will be to mimic the molecular properties and diversity found in natural products.[69]

Yet another approach to increase the promiscuity of a drug is by creating drug cocktails. Although one of the major disadvantages of a drug cocktail compared to a single multi-target agent is the risk of drug–drug interactions,[70] there is an increasing interest in developing drugs which can bypass these issues. Such an approach becomes especially appealing towards evolving targets that generate mutant forms escaping drug interaction as it requires the consideration of more than just one drug binding tightly to the existing target molecule. Due to the potential for increased toxicity with each additional drug in the cocktail and possible cross interactions, the smallest cocktail possible should be targeted that will effectively cover the different ensembles of the evolving target. To that end, focus has been given for the development of theoretical methods to design optimal drug cocktails for targeting molecular ensembles.[71]

Target space

Major components of the target space

Organisms respond to continuous variations in internal and external cellular conditions by orchestrating their responses depending on the environmental challenges they are faced with. This involves the usage of a complex network of interactions among different proteins, RNA, metabolites and several other cellular entities, which undergo rewiring when perturbed by chemicals or drugs from the Drug space (Fig. 3A). The interaction between different chemicals and cellular entities can be represented in the form of a network—so called Drug Target network. Recent years have seen the development of a number of approaches both computational and experimental for the identification and elucidation of the molecular targets of a drug on a genomic scale.[72–81] This cellular target space which contains the targets of drugs, can be considered to predominantly comprise of three components namely protein–protein, metabolic and transcriptional interactions (Fig. 3B). While the vast majority of the drugs target the protein–protein and metabolic components, limited number of targets have been identified till date for the transcriptional pool.[10,82–85] Indeed, most common therapeutic targets for established drugs belong to either protein kinase or receptor families with enzymes and ion channels forming the second most predominant class of targets.[86] This explains the reasons for the increased attention towards understanding the biophysics of protein–protein contacts in the context of drug targets as these protein classes form major players in protein–protein interactions.[87]

An interesting possibility for systematic target identification is that the structures of biological networks may actually provide valuable information in assessing targets and their combinations. In recent years, it has been appreciated that many effective drugs in therapeutic areas as diverse as oncology, psychiatry and infectious diseases act on multiple rather than single targets.[6] Indeed, this has been confirmed by network analysis of the drug target interactions where it was found that not only drugs commonly act on multiple targets but also drug targets are often involved in multiple diseases with over 40% of the drug targets that map onto disease genes involved in more than one disease.[10] This observation is further strengthened by an independent analysis to analyze the genetic origins of most diseases using OMIM database,[88] where the authors found that of 1284 disorders documented in OMIM nearly 70% share at least one gene with another disorder.[82] Taken together, studies employing network approaches reveal that in most cases exquisitely selective compounds may exhibit a lower than needed efficacy for the treatment of disorders and that compounds that selectively act on two or more targets of interest might be more efficacious than single target agents, ruling out the assumption of one drug for one target in a disease which has significantly

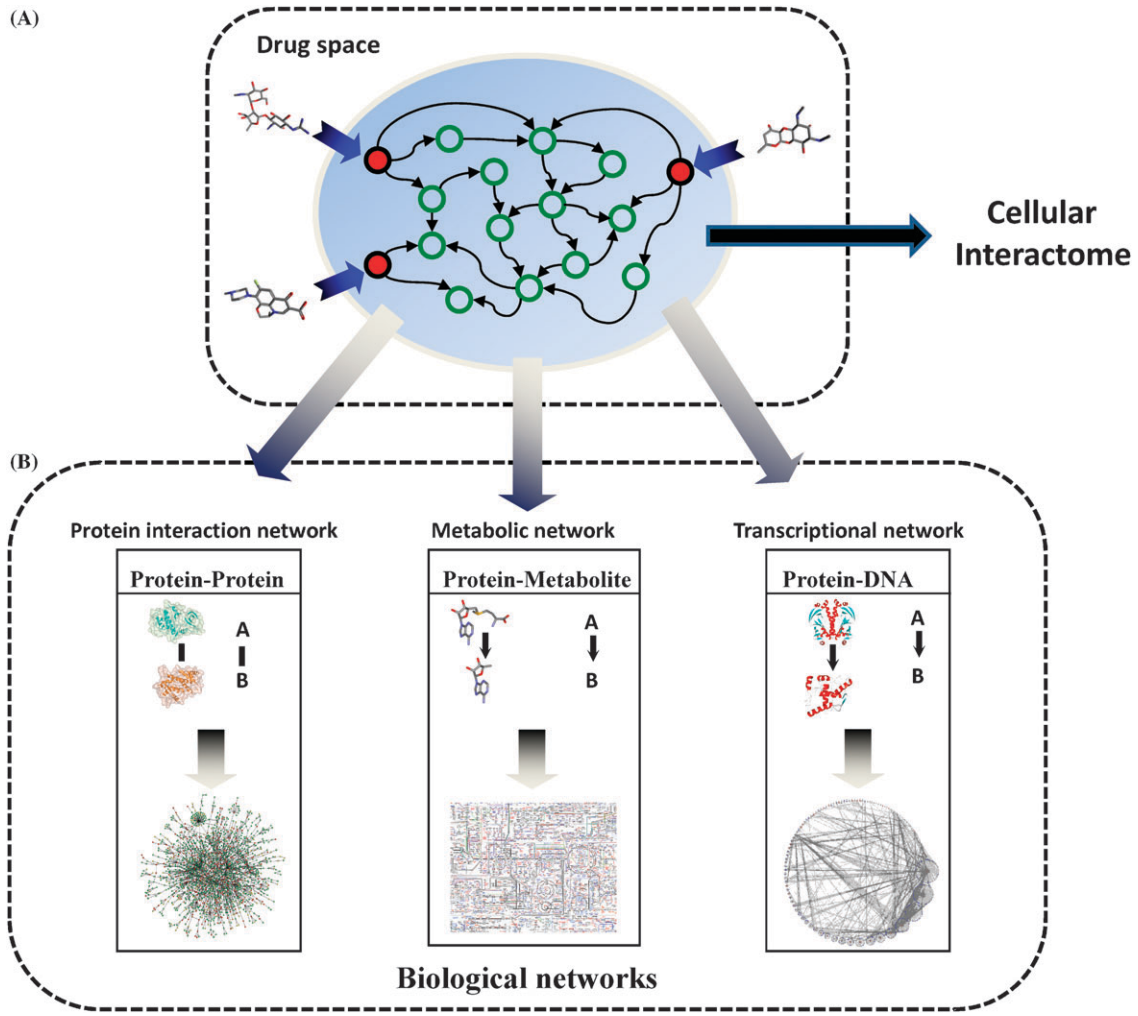


Fig. 3 Different components of the drug-target network. (A) Drug space (marked with the outer box) consists of the small-molecules which can potentially bind entities with-in the cell (marked as drug targets in red spheres). In turn cellular interactions between different components (marked with red spheres and green circles) form cellular interactome comprising the target space. (B) Target space comprises of different components namely protein–protein interactions, metabolic pathways and transcriptional circuits which together form the biological network or the cellular interactome.

influenced the drug discovery pipeline for more than a decade before the advent of genomics.

An additional insight gained from network-based approaches in pharmacology is that, although disease genes play central roles in the protein–protein interaction networks of the target space, the vast majority of disease genes are nonessential and show no tendency to encode hub proteins and their expression pattern indicates that they are localized in the functional periphery of the network. This is in contrast to essential genes which show higher likelihood to encode for hub proteins in the protein interactome, higher transcript levels and are expressed widely in most tissues.[82,89] These studies also show that genes with intermediate connectivities are likely to harbor germ-line disease mutations, suggesting that disease genes tend to occupy an intermediate niche in terms of their physiological and cellular importance.[89] Likewise, analysis of the protein interactions of the drug targets suggested that they have more interactions than expected by chance but lower than that observed for essential genes, re-enforcing the trend observed for disease genes.[10]

Different knowledge based approaches have also been employed to prioritize the drug targets based on the existing datasets of protein interactomes. For instance, Lage *et al.*[83] constructed a phenome–interactome network comprising gene products implicated in many different categories of human disease which permitted them to identify previously unknown complexes likely to be associated with disease by using a phenotype similarity score. Others exploited the topological features of the protein interactome for predicting novel disease genes.[90] Similarly, the notion of polypharmacology had its effect on computational approaches to associate targets to well-established drugs based on similarity in protein sequence space.[81] Polypharmacology also had its influence on experimental screening in a high throughput fashion wherein attempts have been made to understand the relationship between similarity of ligands and their targets when they belong to one or more different sequence clusters.[59,80]

While much of mainstream research has focused on the protein–protein interaction component of the target space few

attempts are made to exploit the therapeutic potential of metabolic and transcriptional components. Many reasons have been accounted for the slow pace in targeting these components, including the challenges involved in manipulating them with ligands and lack of experimental protocols for studying their impact when perturbed. For instance, transcription factors (TFs) play a major role in many human diseases including cancer, inflammatory and heart diseases however very few TFs such as those containing the ligand binding domains could be successfully exploited.[84] However, availability of new approaches such as those which can directly block transcription factor dimerization or those which can indirectly target specific DNA and DNA decoys are being increasingly explored.[84]

Methods for identifying drug targets

Availability of genome sequences together with technologies for high-throughput screening of chemicals on a large scale has revolutionized our ability to understand the biological activity of novel ligands and in identifying their targets in short time periods. These methods can be broadly classified into genetics-based, proteomics-based and knowledge-driven approaches (See Table 2). Genetics based approaches typically involve use of mutant libraries of large set of genes, generated either by exploiting the RNA interference pathways in mammalian systems or by knocking-out the genes, which are exposed to drugs at different concentrations to study the resistance measured as the fitness of a strain of interest with respect to the wild type.[76,77] Improvements in these approaches include bar-coding of deletion strains with unique DNA sequences which enable parallel genetic screens of a large number of drugs.[78,91] Alternatives to the mutant-based approaches involve forward chemical genetics techniques which typically involves screening of small molecule libraries for their ability to induce a particular phenotype in cells or cellular extracts. In these approaches, instead of deleting or impairing protein function at the genetic level, small molecules generally act by inhibiting (or activating) a particular protein or set of proteins directly. Tracing the inhibitor (or activator) back to its target protein can, in principle, provide a causal link between the target and its associated phenotype.[92]

Proteomic methods for drug-target identification can be classified into three major categories: (a) based on affinity chromatography, (b) based on small-molecule or protein microarrays, (c) based on active-site profiling of the proteins of interest (reviewed in ref. 93). In affinity-based methods typically a small molecule is immobilized *via* a functional group onto a solid support followed by the addition of a protein extract. This is followed by a series of washing steps to finally isolate and identify proteins which remain in the column due to the affinity for the small molecule.[94] Protein microarray based methods comprise of recombinant protein molecules or antibodies immobilized on the surface of a substrate material like glass or silicon, which is then exposed to small molecules which are labeled and the binding on the chip is monitored.[95] Antibody microarrays form a variant which can be useful to study the ligands which can bind to low abundance proteins.[96] Chemical microarrays form a promising class of methods when the goal is to screen for a large number of small-molecules against a selected set of proteins.[97] In activity-based profiling methods, active-site directed chemical probes consisting of two-components, a moiety which covalently binds to the active site of an enzyme and a reporter tag for tracking the modified proteins, are used to measure of binding.[98] More recently there is an increasing interest in using metabolic approaches on systems-scale for identifying drug targets due to improvements in mass spectrometry techniques.[99]

Knowledge-driven approaches comprise of both literature derived data sources and informatics methods which employ our current understanding of design principles of drug space, target space or an integration of them.[72,73,81,100–102] These include but are not limited to the use of structure activity relationships, network-based, genomics-based, pathway analysis and integration of data sources[72,73,80,81,103] (briefly summarized in Table 2). One of the major limitations of these approaches is that these are often only incremental and can not predict counterintuitive or unexpected outcomes.

Identifying off-target effects in the target space

One of the significant outcomes of the notion of poly-pharmacology is that the promiscuity of drugs can be of therapeutic advantage if the drug under investigation can perturb the relevant genes in the disease state to bring it to equilibrium. For instance, if the target is a rapidly mutating agent, a drug that is too specific will quickly lose its efficacy by not binding well to functional mutants. Therefore, in molecular design, it is crucial to tailor the binding specificity of a drug in such cases to all the functional mutants to improve its efficacy.[44] The ideal situation will be to design promiscuous drugs that are directed towards a desirable multi-target space, however this is not straightforward for protein families such as kinases that are structurally divergent but yet need to be targeted by a single drug.[45] Recently, Apsel *et al.*[80] reported the discovery of dual inhibitors of tyrosine and phosphoinositide kinases which form intensely pursued cancer drug target families. Although tyrosine and phosphoinositide kinases lack significant sequence similarity, they share several short motifs. Through iterative chemical synthesis, X-ray crystallography and kinome-level biochemical profiling they were able to develop molecules that adopted dual selectivity to the hydrophobic pocket conserved in both enzyme classes. Other approaches employed the phenotypic side-effects of the marketed drugs to cluster drugs which exhibited similar therapeutic indications as a means of identifying potential new drug targets for existing drugs thereby extending the applicability of existing drugs in less explored disease phenotypes.[104]

In addition to using off-target effects for therapeutic benefits recent advances have also involved the use of computational and experimental means to identify them at a first glance.[60,105] For instance, Ericson *et al.*[105] employed chemo-genomic screening to identify the off-target effects of 81 psychoactive drugs in yeast. The general consensus is that most drugs showed a propensity to affect multiple cellular functions ranging from secretion, protein folding to chromatin remodeling roles suggesting the utility of model organism pharmacogenetic studies to provide a rational foundation for studying the off-target effects of clinically important drugs.

Table 2 Different methods available for identifying drug-targets on a genomic scale. Methods can be broadly classified into proteomics-based, genetics-based and knowledge-driven

Proteomics-based methods	Description
Activity based protein profiling (ABPP)[129]	This is a functional proteomic technology that uses chemical probes that react with mechanistically related classes of enzymes. The basic unit of ABPP is a probe that typically consists of a reactive group (electrophile or a photoreactive group) that covalently binds to the active site of an enzyme (nucleophilic residue) and a tag. The tag can either be a reporter (*i.e.* fluorophore, radioactive group) or a handle (*i.e.* affinity tags such as biotin). A tag-free strategy for activity-based protein profiling has also been introduced that utilizes the copper(I)-catalyzed azide-alkyne cycloaddition reaction (click chemistry) and gives the advantage of not interfering with biological activity or binding affinities of the probes. The activity-based protein profiling and multidimensional protein identification technologies (ABPP-MudPIT) can provide profiling of inhibitor selectivity, as the potency of an inhibitor can be tested against hundreds of targets simultaneously.[130]
Affinity chromatography[94]	This is a protein separation method based on the interaction between target proteins and specific immobilized ligands. Traditionally, the ligand is tethered on a solid support *via* a spacer arm followed by the addition of a cellular lysate or tissue extract. Only target proteins binding tightly to the ligand are selectively purified, eluted off (denaturation or competition with free ligand) and subsequently identified by mass spectroscopy. To minimize the identification of nonspecifically bound proteins, the protein profile that is obtained with an inactive ligand analogue is also determined and compared with the relevant profile, determined with the desired analogue. More recently, an improved method for the identification of proteins that can bind to small-molecules and drugs has been established which uses quantitative mass spectrometry (MS)-based proteomics (utilizing stable isotope labeling with amino acids in cell culture (SILAC)) and affinity chromatography.[131]
Microarrays[95–97,132]	Microarrays in drug target discovery provide miniaturized high-throughput tools to study binding of specific molecules to immobilized proteins or small molecules. In protein microarrays, different recombinant proteins or antibodies that are immobilized on a solid substrate are exposed to a drug solution to identify the target protein(s) which can bind to the small molecule. In chemical microarrays, immobilized drug compounds can be screened for candidate drug–target interactions with purified proteins.[97] When the target protein is known, small molecule arrays can be also used to identify off-target interactions that could have implications for side-effects.
Genetics-based Methods	**Description**
Synthetic lethality/ Gene knock-out[76,78]	Single gene knock-out strains on a genomic scale or for a selected set are exposed to small molecules at different concentrations to evaluate the fitness defects and fitness levels are compared to wild-type populations exposed to the same conditions. This provides an easy means to identify targets on a large scale.[76,78]
RNAi	RNA interference pathways in mammalian systems are used for silencing genes and similar approaches as above are employed to study the fitness defects of cell lines to identify potential drug targets in higher eukaryotes[77,133]
Forward chemical genetics[92]	Unlike the use of mutants in previous approaches, small molecules are screened for their ability to induce a particular phenotype in cells or cellular extracts. Instead of deleting or impairing protein function at the genetic level, as in classical genetics, small molecules generally act by inhibiting (or activating) a particular protein or set of proteins directly. Tracing the inhibitor (or activator) back to its target protein can, in principle, provide a causal link between the target and its associated phenotype. Forward chemical genetics requires three components: one, a collection or 'library' of compounds; two, a biological assay with a quantifiable phenotypic output; and three, a strategy for identifying the target(s) of active compounds.[92,134,135]
Knowledge-driven approaches	**Description**
Literature derived interactions.[10,103,136,137]	In these approaches, manually curated set of interactions are obtained from the literature to generate high confidence set of drug-target relationships to either study their overall structure[10] or focus on specific disease of interest.[10,103,136,137]
Network-based approaches.[3,80]	In these approaches, literature derived interactions are exploited to predict new interactions based on the principles governing the structure of the networks, so that new disease targets are identified using comparative genomics or other informatics-based methods, followed possibly by experiments to improve the chemicals.[3,80]
in silico chemogenomics[41]	In predictive chemogenomics one predicts relationships between genes/proteins and compounds. *In silico* approaches that are used can be classified into ligand-based approaches (ligand comparison for target prediction), target-based approaches (target comparison for ligand prediction) or ligand-target based approaches[41]

Disease space

Decomposing drug-target associations using network-based approaches

An important notion that has emerged in post-genomic drug discovery is that the large-scale integration of genomic, proteomic, signalling and metabolomic data can allow us to construct complex networks of the cell that would provide us with a new framework for understanding the molecular basis of physiological or pathophysiological states.[106] Such an integrated view has important implications in improving our understanding of the disease phenotypes by viewing them as perturbations in a complex system rather than effects on a selective set of proteins. Using such a framework, network-based drug discovery aims to harness this knowledge to investigate and understand the impact of interventions, such as candidate drugs, on the molecular networks that define different states and therefore can significantly complement the existing drug discovery pipelines.

In such a framework at the most basic level is the Drug-Target (DT) network which composes of a directed graph connecting the set of drugs from the drug space which

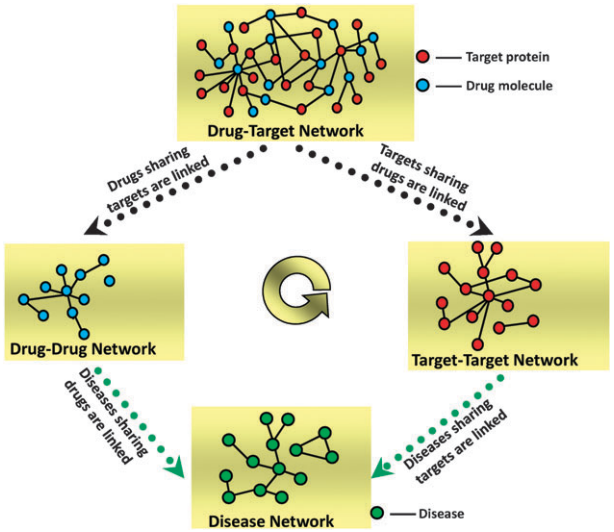


Fig. 4 Decomposition of Drug-Target (DT) network. Drug-target network composes of a directed graph with the interactions from drugs to targets. Such a directed network can be exploited to study associations between drugs or targets. The former consists of linking two drugs (DD network) when they share significant number of targets while the later involves linking target proteins which share a significant number of drugs (TT network). Both these decompositions have yielded significant insights in recent years, in particular, to understand the polypharmacological nature of drugs. One can study the disease associations by using either DD or TT networks by overlaying the phenotypic knowledge accumulated in databases for either drugs or targets. Such integrated approaches not only provide insights into relationships between different diseases but also allow the applicability of existing drugs in less explored disease phenotypes in the paradigm of network pharmacology.

can bind to the targets in the target space (Fig. 4). Such a network has been shown to form a bipartite graph with a giant connected component, suggesting the involvement of most drugs in targeting more than one target protein.[9,10] At the second level, one can visualize the decomposition of the DT network into the Drug-Drug (DD) network or Target-Target (TT) network (Fig. 4). The former is typically constructed by linking two drugs if they share a significant number of targets while the later comprises of associations or links between targets which are targeted by the same set of drugs. As is expected, increasing the level of significance in the number of shared targets[104] or drugs would improve the quality of the polypharmacological downstream network. Initial analysis by decomposing the experimentally validated DT network, for FDA-approved drugs, into DD or TT network confirmed the tendency of pharmaceutical industry to target already experimentally validated proteins, leading to an increase in the number of follow-on drugs.[10,104] The authors also found that a number of drugs in the DD network grouped into distinct clusters confirmed from their similarity according to the Anatomical Therapeutic Chemical (ATC) classification.

At the third level in this framework one can connect diseases and therapies associated with them using the already known network of DD or TT interactions. This typically involves mapping the already known disease phenotype of a drug or a target onto the DD or TT networks respectively to generate a network of disease associations (Fig. 4). Recent attempts to generate such disease association or common therapy based networks starting from drug–drug interactions revealed that the average path length between drugs is shorter than three steps reinforcing the polypharmacology notion from the perspective of drugs *i.e.*, most chemicals might be sharing their phenotypic affects to a significant extent when perturbing the target space.[107] An alternative explanation for these observations is that most drugs currently employed are a result of follow-on of existing knowledge about previously established drugs, suggesting that there is enormous potential in the drug space for identifying bioactive compounds with novel targets in the target space which yet needs to be exploited. Future studies in this direction can address questions on the link between different diseases and the role of multi-target drugs in a range of related disorders to pinpoint the basis for some of these observations.

Exploiting disease networks to study disease associations

Independently drug–drug relations can also be constructed by obtaining only the structural similarity or prior phenotypic characteristics of well-exploited drugs in contrast to the network-based approaches described above. Likewise other variants such as a Target-Target (TT) network in which proteins documented in literature to be involved in the same disease can also be constructed to study relationships between different entities in the target space or diseases. For instance, recent studies used the phenotypic information for several disease associated genes available in the OMIM database[88] to construct networks of disease associations.[82,89] These studies unambiguously demonstrated that network properties of genes in such networks influence the likelihood and phenotypic consequences of disease mutations, with genes exhibiting intermediate connectivities having the highest probability of harboring germ-line disease mutations, suggesting that disease genes tend to occupy an intermediate niche in terms of their physiological and cellular importance. In addition, disease genes were found to show significant functional clustering in the studied network suggesting the existence of disorder-specific functional modules.[89]

Conclusion

Data completeness in the drug, disease and target space is a crucial issue to our understanding of Drug-Target networks. Thus, efforts should be directed to systematically illuminate the drug interactome.[108,109] Nevertheless, advances in genomics have influenced the way we understand the action of drugs on a genomic scale.[110–112] One of the challenging aspects of drug discovery is the evolution of drug resistant strains emerging in several human diseases such as malaria, tuberculosis and cystic fibrosis. Systematic genomic screens have improved the way we understand the combinatorial drug chemistry[113,114] and would enable us to design ''hyper-antagonistic'' drugs which can fight drug-resistant strains by working as drug cocktails when the agents have acquired resistance to individual drugs.[110] Indeed, recent studies show that although synergistically acting drugs, which are commonly used in clinical settings, might favor immediate efficacy, they might also favor evolution

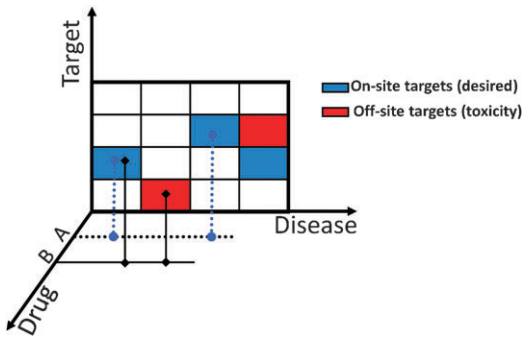


Fig. 5 The Disease-Target-Drug matrix: correlating drug, disease and target space in three dimensions. Targets are grouped into specific diseases and drugs, drugs are grouped into specific diseases and targets. Cyan boxes represent ideal targets that are related to specific diseases. Red boxes represent off targets that should not be targeted. Different diseases can share common targets and different drugs can aim different targets. Detailed knowledge of the global Disease-Target-Drug map can allow targeting different diseases with a single drug (drug A), avoiding at the same time off-target effects (drug B).

of resistance compared to antagonistic combinations, thereby indicating a need to develop antagonistic combinations, which might be effective in combating diseases in multi-drug chemotherapy.[115,116]

An emerging view of polypharmacology in the post-genomic era is that drug, target and disease spaces can be correlated to study the effect of drugs on different spaces and their interrelationships can be exploited for designing drugs or cocktails which can effectively target one or more disease states (Fig. 5). According to such a view, systems-level understanding of the cell by integration of data from different omics platforms can be of unprecedented value, as it not only improves our understanding of the pathophysiological states, but also its relationship in the context of different chemicals thereby making useful interpretation of existing data and accelerating hypothesis generation for testing in disease models.[117–119]

Acknowledgements

SCJ acknowledges financial support from the MRC Laboratory of Molecular Biology and Cambridge Commonwealth Trust. We thank Briasoulis E., De S., Lang B., Mittal N., Perica T., Venkatakrishnan A. J., Wuster A. and Michnick S. for critically reading the manuscript and providing helpful comments. We apologize to colleagues whose relevant work could not be cited due to lack of space.

References

1 P. Csermely, V. Agoston and S. Pongor, *Trends Pharmacol. Sci.*, 2005, **26**, 178–182.
2 Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, A. Leonardson, L. W. Castellini, S. Wang, M. F. Champy and B. Zhang, *et al.*, *Nature*, 2008, **452**, 429–435.
3 A. L. Hopkins, *Nat. Chem. Biol.*, 2008, **4**, 682–690.
4 A. L. Hopkins, *Nat. Biotechnol.*, 2007, **25**, 1110–1111.
5 C. G. Wermuth, *Drug Discovery Today*, 2004, **9**, 826–827.
6 B. L. Roth, D. J. Sheffler and W. K. Kroeze, *Nat. Rev. Drug Discovery*, 2004, **3**, 353–359.
7 G. R. Zimmermann, J. Lehar and C. T. Keith, *Drug Discovery Today*, 2007, **12**, 34–42.
8 A. Petrelli and S. Giordano, *Curr. Med. Chem.*, 2008, **15**, 422–432.
9 A. Ma'ayan, S. L. Jenkins, J. Goldfarb and R. Iyengar, *Mt. Sinai. J. Med.*, 2007, **74**, 27–32.
10 M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi and M. Vidal, *Nat. Biotechnol.*, 2007, **25**, 1119–1126.
11 J. A. DiMasi, R. W. Hansen and H. G. Grabowski, *J. Health Econ.*, 2003, **22**, 151–185.
12 J. A. DiMasi, *Pharmacoeconomics*, 2002, **20**(supplement 3), 1–10.
13 I. Kola and J. Landis, *Nat. Rev. Drug Discovery*, 2004, **3**, 711–715.
14 P. Ma and R. Zemmel, *Nat. Rev. Drug Discovery*, 2002, **1**, 571–572.
15 I. Walker and H. Newell, *Nat. Rev. Drug Discovery*, 2009, **8**, 15–16.
16 K. A. O'Connor and B. L. Roth, *Nat. Rev. Drug Discovery*, 2005, **4**, 1005–1014.
17 B. Apsel, J. A. Blair, B. Gonzalez, T. M. Nazif, M. E. Feldman, B. Aizenstein, R. Hoffman, R. L. Williams, K. M. Shokat and Z. A. Knight, *Nat. Chem. Biol.*, 2008, **4**, 691–699.
18 J. Inglese, R. L. Johnso n, A. Simeonov, M. H. Xia, W. Zheng, C. P. Austin and D. S. Auld, *Nat. Chem. Biol.*, 2007, **3**, 466–479.
19 J. Alper, *Science*, 1994, **264**, 1399–1401.
20 R. Lahana, *Drug Discovery Today*, 1999, **4**, 447–448.
21 P. D. Leeson and B. Springthorpe, *Nat. Rev. Drug Discovery*, 2007, **6**, 881–890.
22 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
23 J. Sadowski and H. Kubinyi, *J. Med. Chem.*, 1998, **41**, 3325–3329.
24 A. Ajay, W. P. Walters and M. A. Murcko, *J. Med. Chem.*, 1998, **41**, 3314–3324.
25 M. Soural, I. Bouillon and V. Krchnak, *J. Comb. Chem.*, 2008, **10**, 923–933.
26 W. M. Dai and J. Y. Shi, *Comb. Chem. High Throughput Screening*, 2007, **10**, 837–856.
27 M. D. Burke and S. L. Schreiber, *Angew. Chem., Int. Ed.*, 2004, **43**, 46–58.
28 M. Peuchmaur and Y. S. Wong, *Comb. Chem. High Throughput Screening*, 2008, **11**, 587–601.
29 D. A. Spiegel, F. C. Schroeder, J. R. Duvall and S. L. Schreiber, *J. Am. Chem. Soc.*, 2006, **128**, 14766–14767.
30 D. S. Tan, *Nat. Chem. Biol.*, 2005, **1**, 74–84.
31 C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235–249.
32 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.
33 D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, 2002, **45**, 2615–2623.
34 C. A. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman and P. Artursson, *J. Med. Chem.*, 2003, **46**, 558–570.
35 M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis and P. D. Leeson, *J. Med. Chem.*, 2003, **46**, 1250–1256.
36 S. D. Pickett, I. M. McLay and D. E. Clark, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 263–272.
37 T. I. Oprea and H. Matter, *Curr. Opin. Chem. Biol.*, 2004, **8**, 349–358.
38 A. Bender, J. L. Jenkins, M. Glick, Z. Deng, J. H. Nettles and J. W. Davies, *J. Chem. Inf. Model.*, 2006, **46**, 2445–2456.
39 A. Bender, D. W. Young, J. L. Jenkins, M. Serrano, D. Mikhailov, P. A. Clemons and J. W. Davies, *Comb. Chem. High Throughput Screening*, 2007, **10**, 719–731.
40 E. Gregori-Puigjane and J. Mestres, *Comb. Chem. High Throughput Screening*, 2008, **11**, 669–676.
41 D. Rognan, *Br. J. Pharmacol.*, 2007, **152**, 38–52.
42 W. Duch, K. Swaminathan and J. Meller, *Curr. Pharm. Des.*, 2007, **13**, 1497–1508.
43 S. M. Lippow, K. D. Wittrup and B. Tidor, *Nat. Biotechnol.*, 2007, **25**, 1171–1176.
44 M. L. Radhakrishnan and B. Tidor, *J. Phys. Chem. B*, 2007, **111**, 13419–13435.

45 A. L. Hopkins, J. S. Mason and J. P. Overington, *Curr. Opin. Struct. Biol.*, 2006, **16**, 127–136.
46 M. M. Hann, A. R. Leach and G. Harper, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 856–864.
47 R. Morphy and Z. Rankovic, *Drug Discovery Today*, 2007, **12**, 156–160.
48 R. Morphy and Z. Rankovic, *J. Med. Chem.*, 2006, **49**, 4961–4970.
49 R. Morphy and Z. Rankovic, *J. Med. Chem.*, 2005, **48**, 6523–6543.
50 R. Morphy, C. Kay and Z. Rankovic, *Drug Discovery Today*, 2004, **9**, 641–651.
51 R. Morphy and Z. Rankovic, *Curr. Pharm. Des.*, 2009, **15**, 587–600.
52 R. Horuk, *Expert Reviews in Molecular Medicine*, 2009, **11**, e1.
53 J. L. Wallace and P. Del Soldato, *Fundam. Clin. Pharmacol.*, 2003, **17**, 11–20.
54 J. L. Wallace, *Trends Pharmacol. Sci.*, 2007, **28**, 501–505.
55 R. Aslanian, M. Mutahi, N. Y. Shih, J. J. Piwinski, R. West, S. M. Williams, S. She, R. L. Wu and J. A. Hey, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 1959–1961.
56 N. Murugesan, Z. Gu, L. Fadnis, J. E. Tellew, R. A. Baska, Y. Yang, S. M. Beyer, H. Monshizadegan, K. E. Dickinson, M. T. Valentine, W. G. Humphreys, S. J. Lan, W. R. Ewing, K. E. Carlson and M. C. Kowala, *et al.*, *J. Med. Chem.*, 2005, **48**, 171–179.
57 D. Fattori, A. Squarcia and S. Bartoli, *Drugs R. D.*, 2008, **9**, 217–227.
58 P. C. Elwood, A. M. Gallagher, G. G. Duthie, L. A. Mur and G. Morgan, *Lancet*, 2009, **373**, 1301–1309.
59 M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nat. Biotechnol.*, 2007, **25**, 197–206.
60 A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread and J. L. Jenkins, *ChemMedChem*, 2007, **2**, 861–873.
61 G. M. Cragg and D. J. Newman, *J. Ethnopharmacol.*, 2005, **100**, 72–79.
62 J. D. McChesney, S. K. Venkataraman and J. T. Henri, *Phytochemistry*, 2007, **68**, 2015–2022.
63 D. J. Newman, G. M. Cragg and K. M. Snader, *J. Nat. Prod.*, 2003, **66**, 1022–1037.
64 A. Murakami, H. Ashida and J. Terao, *Cancer Lett.*, 2008, **269**, 315–325.
65 M. A. Koch, L. O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt and H. Waldmann, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 16721–16726.
66 J. Clardy and C. Walsh, *Nature*, 2004, **432**, 829–837.
67 A. Ganesan, *Curr. Opin. Chem. Biol.*, 2008, **12**, 306–317.
68 M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl and H. Waldmann, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17272–17277.
69 M. Feher and J. M. Schmidt, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 218–227.
70 I. R. Edwards and J. K. Aronson, *Lancet*, 2000, **356**, 1255–1259.
71 M. L. Radhakrishnan and B. Tidor, *J. Chem. Inf. Model.*, 2008, **48**, 1055–1073.
72 G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason and A. L. Hopkins, *Nat. Biotechnol.*, 2006, **24**, 805–815.
73 Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics*, 2008, **24**, i232–240.
74 L. Jacob and J. P. Vert, *Bioinformatics*, 2008, **24**, 2149–2156.
75 S. C. Brewerton, *Curr. Opin. Drug Discov. Devel.*, 2008, **11**, 356–364.
76 M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow and G. Giaever, *Science*, 2008, **320**, 362–365.
77 A. W. Whitehurst, B. O. Bodemann, J. Cardenas, D. Ferguson, L. Girard, M. Peyton, J. D. Minna, C. Michnoff, W. Hao, M. G. Roth, X. J. Xie and M. A. White, *Nature*, 2007, **446**, 815–819.
78 C. H. Ho, L. Magtanong, S. L. Barker, D. Gresham, S. Nishimura, P. Natarajan, J. L. Koh, J. Porter, C. A. Gray, R. J. Andersen, G. Giaever, C. Nislow, B. Andrews, D. Botstein and T. R. Graham, *et al.*, *Nat. Biotechnol.*, 2009, **27**, 369–377.
79 M. A. Fabian, W. H. Biggs, 3rd, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld and S. Herrgard, *et al.*, *Nat. Biotechnol.*, 2005, **23**, 329–336.
80 B. Apsel, J. A. Blair, B. Gonzalez, T. M. Nazif, M. E. Feldman, B. Aizenstein, R. Hoffman, R. L. Williams, K. M. Shokat and Z. A. Knight, *Nat. Chem. Biol.*, 2008, **4**, 691–699.
81 M. Kuhn, M. Campillos, P. Gonzalez, L. J. Jensen and P. Bork, *FEBS Lett.*, 2008, **582**, 1283–1290.
82 K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–8690.
83 K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau and S. Brunak, *Nat. Biotechnol.*, 2007, **25**, 309–316.
84 P. Brennan, R. Donev and S. Hewamana, *Mol. BioSyst.*, 2008, **4**, 909–919.
85 D. S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai and A. L. Barabasi, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 9880–9885.
86 D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Res.*, 2008, **36**, D901–906.
87 A. I. Archakov, V. M. Govorun, A. V. Dubanov, Y. D. Ivanov, A. V. Veselovsky, P. Lewi and P. Janssen, *Proteomics*, 2003, **3**, 380–391.
88 A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic Acids Res.*, 2005, **33**, D514–517.
89 I. Feldman, A. Rzhetsky and D. Vitkup, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 4323–4328.
90 J. Xu and Y. Li, *Bioinformatics*, 2006, **22**, 2800–2805.
91 Z. Yan, M. Costanzo, L. E. Heisler, J. Paw, F. Kaper, B. J. Andrews, C. Boone, G. Giaever and C. Nislow, *Nat. Methods*, 2008, **5**, 719–725.
92 R. S. Lokey, *Curr. Opin. Chem. Biol.*, 2003, **7**, 91–96.
93 L. Sleno and A. Emili, *Curr. Opin. Chem. Biol.*, 2008, **12**, 46–54.
94 H. Katayama and Y. Oda, *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.*, 2007, **855**, 21–27.
95 S. F. Kingsmore, *Nat. Rev. Drug Discovery*, 2006, **5**, 310–320.
96 C. Wingren and C. A. Borrebaeck, *OMICS*, 2006, **10**, 411–427.
97 H. Ma and K. Y. Horiuchi, *Drug Discovery Today*, 2006, **11**, 661–668.
98 H. Schmidinger, A. Hermetter and R. Birner-Gruenberger, *Amino Acids*, 2006, **30**, 333–350.
99 J. K. Nicholson and J. C. Lindon, *Nature*, 2008, **455**, 1054–1056.
100 M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen and P. Bork, *Nucleic Acids Res.*, 2008, **36**, D684–688.
101 W. Loging, L. Harland and B. Williams-Jones, *Nat. Rev. Drug Discovery*, 2007, **6**, 220–230.
102 C. G. Wermuth, *Drug Discovery Today*, 2006, **11**, 160–164.
103 I. F. Tsui, R. Chari, T. P. Buys and W. L. Lam, *Cancer Inform*, 2007, **3**, 389–407.
104 M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.
105 E. Ericson, M. Gebbia, L. E. Heisler, J. Wildenhain, M. Tyers, G. Giaever and C. Nislow, *PLoS Genet.*, 2008, **4**, e1000151.
106 E. E. Schadt, S. H. Friend and D. A. Shaywitz, *Nat. Rev. Drug Discovery*, 2009, **8**, 286–295.
107 J. C. Nacher and J. M. Schwartz, *BMC Pharmacol.*, 2008, **8**, 5.
108 M. Cases and J. Mestres, *Drug Discovery Today*, 2009, **14**, 479–485.
109 J. Mestres, E. Gregori-Puigjane, S. Valverde and R. V. Sole, *Nat. Biotechnol.*, 2008, **26**, 983–984.
110 R. Chait, A. Craney and R. Kishony, *Nature*, 2007, **446**, 668–671.
111 M. T. Holden, E. J. Feil, J. A. Lindsay, S. J. Peacock, N. P. Day, M. C. Enright, T. J. Foster, C. E. Moore, L. Hurst, R. Atkin, A. Barron, N. Bason, S. D. Bentley, C. Chillingworth and T. Chillingworth, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 9786–9791.
112 J. M. Rolain, P. Francois, D. Hernandez, F. Bittar, H. Richet, G. Fournous, Y. Mattenberger, E. Bosdure, N. Stremler,

J. C. Dubus, J. Sarles, M. Reynaud-Gaubert, S. Boniface, J. Schrenzel and D. Raoult, *Biology Direct*, 2009, **4**, 1.
113 J. Lehar, A. Krueger, G. Zimmermann and A. Borisy, *Mol. Syst. Biol.*, 2008, **4**, 215.
114 J. Lehar, A. S. Krueger, W. Avery, A. M. Heilbut, L. M. Johansen, E. R. Price, R. J. Rickles, G. F. Short, 3rd, J. E. Staunton, X. Jin, M. S. Lee, G. R. Zimmermann and A. A. Borisy, *Nat. Biotechnol.*, 2009, **27**, 659–666.
115 M. Hegreness, N. Shoresh, D. Damian, D. Hartl and R. Kishony, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 13977–13981.
116 J. B. Michel, P. J. Yeh, R. Chait, R. C. Moellering, Jr and R. Kishony, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 14918–14923.
117 F. Aguero, B. Al-Lazikani, M. Aslett, M. Berriman, F. S. Buckner, R. K. Campbell, S. Carmona, I. M. Carruthers, A. W. Chan, F. Chen, G. J. Crowther, M. A. Doyle, C. Hertz-Fowler, A. L. Hopkins and G. McAllister, *et al.*, *Nat. Rev. Drug Discovery*, 2008, **7**, 900–907.
118 E. C. Butcher, E. L. Berg and E. J. Kunkel, *Nat. Biotechnol.*, 2004, **22**, 1253–1259.
119 R. S. Faustino and A. Terzic, *Clin. Pharmacol. Ther.*, 2008, **84**, 543–545.
120 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–201.
121 S. Goto, Y. Okuno, M. Hattori, T. Nishioka and M. Kanehisa, *Nucleic Acids Res.*, 2002, **30**, 402–404.
122 K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj and M. Ashburner, *Nucleic Acids Res.*, 2008, **36**, D344–350.
123 J. H. Chen, E. Linstead, S. J. Swamidass, D. Wang and P. Baldi, *Bioinformatics*, 2007, **23**, 2348–2351.
124 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
125 S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne and P. Bork, *et al.*, *Nucleic Acids Res.*, 2008, **36**, D919–922.
126 J. Masciocchi, G. Frau, M. Fanton, M. Sturlese, M. Floris, L. Pireddu, P. Palla, F. Cedrati, P. Rodriguez-Tome and S. Moro, *Nucleic Acids Res.*, 2009, **37**, D284–290.
127 K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons, *Nucleic Acids Res.*, 2008, **36**, D351–359.
128 D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin and O. Khovayko, *et al.*, *Nucleic Acids Res.*, 2008, **36**, D13–21.
129 A. E. Speers and B. F. Cravatt, *ChemBioChem*, 2004, **5**, 41–47.
130 N. Jessani, S. Niessen, B. Q. Wei, M. Nicolau, M. Humphrey, Y. Ji, W. Han, D. Y. Noh, J. R. Yates, 3rd, S. S. Jeffrey and B. F. Cravatt, *Nat. Methods*, 2005, **2**, 691–697.
131 S. E. Ong, M. Schenone, A. A. Margolin, X. Li, K. Do, M. K. Doud, D. R. Mani, L. Kuai, X. Wang, J. L. Wood, N. J. Tolliday, A. N. Koehler, L. A. Marcaurelle, T. R. Golub and R. J. Gould, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 4617–4622.
132 M. Salcius, G. A. Michaud, B. Schweitzer and P. F. Predki, *Methods Mol. Biol.*, 2007, **382**, 239–248.
133 N. C. Turner, C. J. Lord, E. Iorns, R. Brough, S. Swift, R. Elliott, S. Rayter, A. N. Tutt and A. Ashworth, *EMBO J.*, 2008, **27**, 1368–1377.
134 S. W. Michnick, P. H. Ear, E. N. Manderson, I. Remy and E. Stefan, *Nat. Rev. Drug Discovery*, 2007, **6**, 569–582.
135 D. A. Bachovchin, S. J. Brown, H. Rosen and B. F. Cravatt, *Nat. Biotechnol.*, 2009, **27**, 387–394.
136 R. Frijters, S. Verhoeven, W. Alkema, R. van Schaik and J. Polman, *Pharmacogenomics*, 2007, **8**, 1521–1534.
137 E. S. Chen, G. Hripcsak, H. Xu, M. Markatou and C. Friedman, *J Am Med Inform Assoc*, 2008, **15**, 87–98.

Meeting report

Interfacing systems biology and synthetic biology

Allyson Lister*, Varodom Charoensawan†, Subhajyoti De†, Katherine James*, Sarath Chandra Janga† and Julian Huppert‡

Addresses: *Centre for Integrated Systems Biology of Ageing and Nutrition (CISBAN) and School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. †MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK. ‡Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, UK.

Correspondence: Julian Huppert. Email: jlh29@cam.ac.uk

Published: 26 June 2009

Genome Biology 2009, **10**:309 (doi:10.1186/gb-2009-10-6-309)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/6/309

A report of BioSysBio 2009, the IET conference on Synthetic Biology, Systems Biology and Bioinformatics, Cambridge, UK, 23-25 March 2009.

The fourth meeting in the BioSysBio conference series brought together international researchers in the interacting disciplines of synthetic biology, systems biology and bioinformatics. This conference was largely student-run, and as well as the formal talks included workshops, discussion sessions and a panel session on ethics, public engagement and biosecurity. A wide range of topics was covered at the conference, including modeling, biofuels and environmental bioremediation, metabolomics, structural and computational genomics, and software tools. Of note were the number of groups presenting improved models of metabolism, studying cellular subsystems such as cell death and circadian rhythms. Others are developing new approaches and standards for systems and synthetic biology, and significant improvements were reported for Systems Biology Markup Language (SBML) and the MIT Registry of Standard Biological Parts. A few highlights of the meeting are given here.

Synthetic biology and its standardization

Synthetic biology is a newly emerging field, where biological components are reengineered to provide new, designed functions. In a keynote lecture, Adam Arkin (University of California, Berkeley, USA) discussed the origins of synthetic biology and its scalability, as well as the engineering challenges that lie beyond the bioreactor. In his view, using synthetic biology, whether to meet an engineering or biological challenge, can be transparent, efficient, reliable, predictable and safe, unlike other human interventions such as selective breeding and the introduction of non-native species. Arkin also described ways of reducing the time and improving the reliability of biosynthesis, such as the use of standardized parts, computer-assisted design, and methods for quickly assembling parts. Evolved systems are complex and subtle, and he highlighted the fact that synthetic organisms need to deal with the same uncertainty and competition as do existing organisms.

Among the 'parts' required in synthetic biology are switches that can function, for example, as regulators of gene expression. Christina Smolke (Stanford University, USA) presented novel design strategies for constructing RNA-based molecular switches that can function as both biosensors and ligand-controlled regulators of gene expression. Binding of the appropriate ligand leads to a regulated conformational change in a designed RNA molecule, which in turn can be linked to an appropriate readout signal, enabling these molecules to act as sophisticated cellular biosensors. She also described how such riboswitches can be used as targeted or 'intelligent' therapeutic molecules for treatment of cancer, allowing them to be carefully tuned to respond as a precise set of molecular stimuli.

Given the recent explosion in the number of approaches to synthetic biology and the amount of data at the interface of genomic and systems biology, there is now an over-whelming need to organize these data efficiently in appropriate repositories. An update on current standards for DNA description by Guy Cochrane (EBI, Cambridge, UK) focused on the different raw sequencing formats available and, in particular, the work that is being done at EMBL to integrate them, via SRS. In an overview of standards and improvements in SBML language, which is the platform for most software in systems biology, Herbert Sauro (University of Washington, Seattle, USA) emphasized the need to

incorporate multi-compartment models into the existing framework of SMBL. Randy Rettberg (Massachusetts Institute of Technology, Cambridge, USA) provided an overview of the publicly available synthetic biology repository being developed at MIT [http://partsregistry.org/Main_Page] as a result of contributions from participants in iGEM - the international genetically engineered machine competition.

Systems biology and automation

Because of the complexity of biological systems, it has always been a challenge to develop predictive dynamic models that are sensitive to changes in biological inputs, but at the same time robust to technical noises. A variety of approaches were described at the meeting. Using a Bayesian framework to study the inferability of model parameters under experimental noise, Kamil Erguler (Imperial College London, UK) introduced sensitivity profiles to identify the relative impacts of changes in parameters on the global dynamics of biochemical models. This analysis revealed the degree of robustness of inferences drawn from different parts of biochemical pathways and thus provides a guide to improved data collection. Andre Ribeiro (Tampere University of Technology, Finland) has developed a delayed stochastic model to investigate the stepwise elongation motion of RNA polymerase and its pauses during transcription. He showed that transcriptional noise level was affected by the durations of the pauses, which could in turn be intrinsically encoded within the DNA sequence.

Another challenge is to store all the information being generated by all the -omic sciences. Catherine Lloyd (Auckland Bioengineering Institute, New Zealand) described the language CellML, which is written in XML and uses existing formats such as MathML and RDF to describe biological models of cellular function. The CellML model repository has over 380 models, free to download [http://www.cellml.org/]. CellML has a number of other useful features, including modularity and the sharing of components such as entities and processes. Ulrike Wittig (EML Research, Heidelberg, Germany) presented SABIO-RK, a database of information about biochemical reactions and enzyme kinetics. The reactions in the database are mainly taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the literature, and the kinetic data comes from the literature. SABIO-RK can be accessed via both a user interface and web services [http://sabio.villa-bosch.de/]. Recent improvements include a new data model for SABIO-RK that allows the storage of intermediate steps in a reaction, making SABIO-RK the first database to offer kinetic information for both biochemical reactions and their individual steps.

DNA synthesis and sequencing comprise one of the cornerstones of modern biology, and Tuval Ben Yehezkel (Weizmann Institute, Rehovot, Israel) described new strategies for synthesizing completely *de novo* DNA fragments using single-molecule PCR in a completely automated fashion. Single-molecule PCR can be readily scaled up, and will complement the highly parallel DNA sequencing technologies such 454 and Solexa sequencing in the future.

Steve Oliver (University of Cambridge, UK) and his colleagues have taken automation even further, describing an automated experimental system to study yeast metabolism. He and colleagues have designed a robot, called Adam, that uses abductive logic programming (ALP) and is capable of reasoning about hypotheses and data, designing experiments to test the hypotheses, and then carrying out those experiments and interpreting the results.

Ethics and security

Scientists in all fields have a duty to consider the public impact of their work and the conference included a lively panel discussion covering ethics, public engagement and biosecurity. Drew Endy (Stanford University, USA) asserted that while the basics of genetic engineering have not changed in more than 30 years, synthetic biology is revolutionary. He raised the question of people trying to 'hack' genomes in their garage: how should they be managed, if indeed they should be managed at all? He also described how the patent system is flawed with regard to synthetic biology; for example, patenting the BioBricks registry of DNA parts encoding basic biological function would be expensive and counterproductive.

Matthew Harvey (Royal Society, London, UK) cautioned that we should not assume that the public must be engaged: sometimes the public simply are not interested. In contrast to genetically modified organisms, there are no synthetic biology products queuing up to be sold right now. Therefore, questioning the public about synthetic biology is currently less like traditional public engagement and more like social-intelligence gathering.

Two concerns were discussed by Julian Savulescu (University of Oxford, UK): that synthetic biology may pose risks in terms of malevolent use, and that the use of synthetic biology might undermine the moral status of living things. For regulators, the challenge is to minimize the risk of malevolent use. For scientists, it is to make better predictions about how research will be used in the future. For philosophers, the challenge is to ascertain criteria for moral status, and determine how to weigh the risk of future wrongdoing against the benefits of pursuing research in synthetic biology. Piers Millet (UN Biological Weapons Convention Implementation Support Unit, Geneva, Switzerland) invited scientists to work with security people to prevent bioterrorism. He highlighted that this engagement

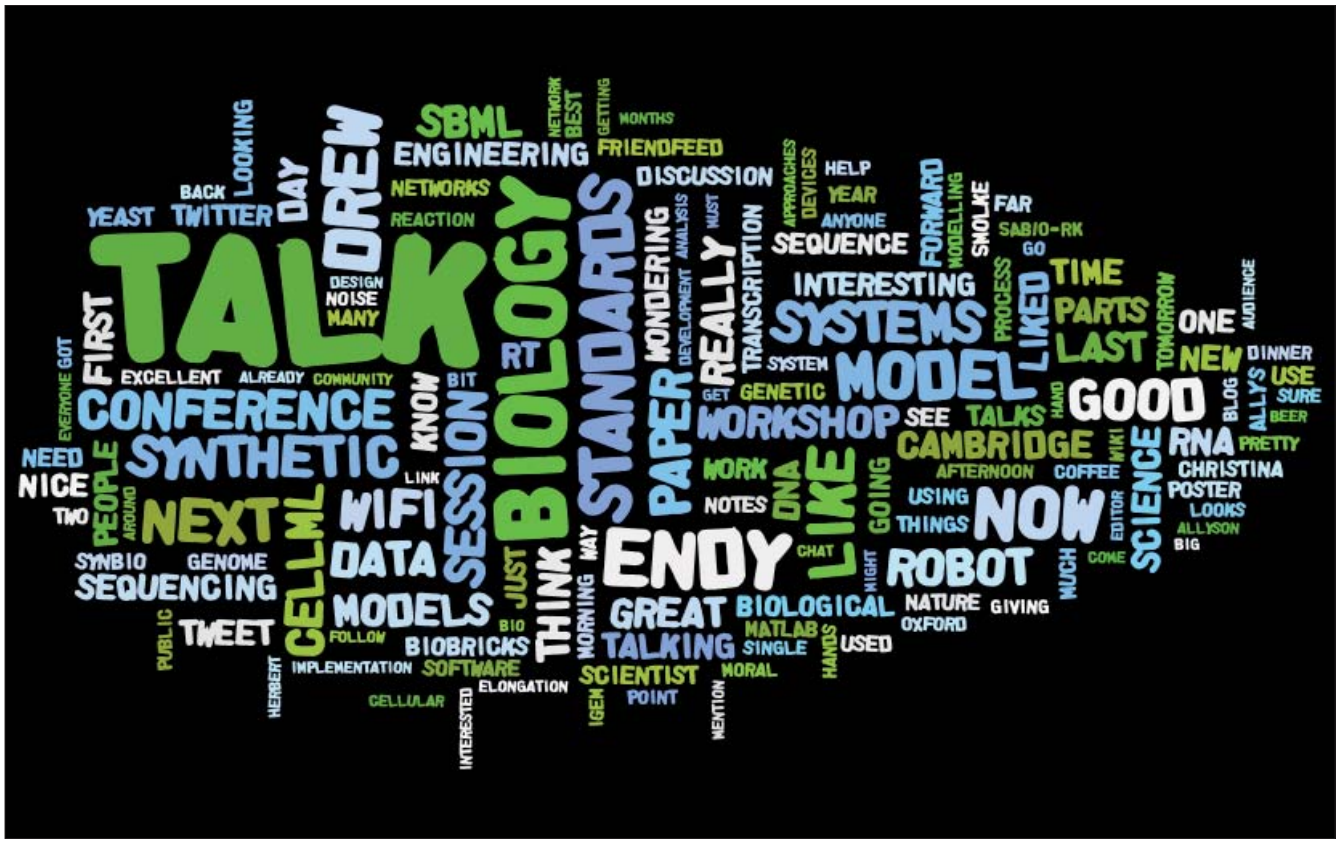


Figure 1
Word cloud of the contents of the BioSysBio Twitter feed, identified via the search term "#biosysbio". The size of each of the words corresponds to their usage frequency. Image generated using wordle.net by Simon Cockell [http://www.flickr.com/photos/sjcockell/3389493857/]. Licensed under the Attribution 2.0 Generic License [http://creativecommons.org/licenses/by/2.0/deed.en_GB].

needs to be bottom up, not top down, and that his organization could help.

A new feature for BioSysBio 2009 to extend participation in the conference was to a wider audience by communicating live content through microblogging (using FriendFeed and Twitter; Figure 1) and live blogging (providing an immediate and permanent log) [http://themindwobbles.wordpress.com/tag/biosysbio-2009/]. The fields covered by the conference are still developing. Researchers are opening up new topics, discovering that mathematical, physical and engineering concepts apply to ever more biological problems. The new generation of researchers increasingly see themselves as forming a new discipline, and while this is exciting, they must ensure that they do not cut themselves off from either of the 'parent' disciplines, the physical sciences (including engineering) and the biological sciences; in particular, more traditional biologists do have important knowledge to convey and questions to pose. However, the results reported at the meeting show that, in most cases, the best from both disciplines is being matched - and exceeded.

MBE | msq033 | RLX
Journal Name | Art. No. | CE Code

Identification and Genomic Analysis of Transcription Factors in Archaeal Genomes Exemplifies Their Functional Architecture and Evolutionary Origin

[AQ1] Ernesto Pérez-Rueda*,[1] and Sarath Chandra Janga*,[2]

[1]Departamento de Ingeniería Celular y Biocatálisis, IBT-UNAM, AP 565-A, Cuernavaca, Morelos, México

[AQ2] [2]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

***Corresponding author:** E-mail: erueda@ibt.unam.mx; sarath@mrc-lmb.cam.ac.uk.

Associate editor: Michele Vendruscolo

Abstract

Archaea, which represent a large fraction of the phylogenetic diversity of organisms, are prokaryotes with eukaryote-like basal transcriptional machinery. This organization makes the study of their DNA-binding transcription factors (TFs) and their transcriptional regulatory networks particularly interesting. In addition, there are limited experimental data regarding their TFs. In this work, 3,918 TFs were identified and exhaustively analyzed in 52 archaeal genomes. TFs represented less than 5% of the gene products in all the studied species comparable with the number of TFs identified in parasites or intracellular pathogenic bacteria, suggesting a deficit in this class of proteins. A total of 75 families were identified, of which HTH_3, AsnC, TrmB, and ArsR families were universally and abundantly identified in all the archaeal genomes. We found that archaeal TFs are significantly small compared with other protein-coding genes in archaea as well as bacterial TFs, suggesting that a large fraction of these small-sized TFs could supply the probable deficit of TFs in archaea, by possibly forming different combinations of monomers similar to that observed in eukaryotic transcriptional machinery. Our results show that although the DNA-binding domains of archaeal TFs are similar to bacteria, there is an underrepresentation of ligand-binding domains in smaller TFs, which suggests that protein–protein interactions may act as mediators of regulatory feedback, indicating a chimera of bacterial and eukaryotic TFs' functionality. The analysis presented here contributes to the understanding of the details of transcriptional apparatus in archaea and provides a framework for the analysis of regulatory networks in these organisms.

Key words: transcription factors, protein families, archaeal genomes, evolution, gene regulation.

Research article

Introduction

Regulation of gene expression at the transcriptional level is a ubiquitous and fine-tuned process observed in all cellular organisms. The ability to respond and adapt to environmental changes is defined by the cell's repertoire of DNA-binding transcription factors (TFs) through interactions between the TFs and the *cis*-regulatory regions of their target genes in the form of a transcriptional regulatory network (Babu et al. 2004; Janga and Collado-Vides 2007). These TFs bind to the promoter regions of specific genes to, either positively or negatively, regulate expression. Due to the crucial role of TFs in coordinating the gene expression kinetics of a genome, they have been studied in many aspects, including mutational analysis, sequence comparisons, and elucidation of numerous 3D structures.

The identification of the TF repertoire in a genome sequence is a prerequisite to understanding the regulation of gene expression and, on a global scale, for the elucidation of regulatory networks. In this context, the organisms with the best studied transcriptional regulatory networks, where TFs have been identified, are the eukaryote *Saccharomyces cerevisiae* (Lee et al. 2002; Janga et al. 2008) and the bacteria *Escherichia coli* K12 (Babu and Teichmann 2003; Gama-Castro et al. 2008), *Bacillus subtilis* (Moreno-Campuzano et al. 2006; Sierro et al. 2008), and more recently *Corynebacterium glutamicum* (Brune et al. 2005; Brinkrolf et al. 2006). However, relatively, little is known about TFs and the transcriptional regulatory networks controlled by them in archaeal genomes, despite the fact that they represent a large fraction of the phylogenetic diversity of organisms. Furthermore, archaea are well suited as model organisms for eukaryotes because of the similarities they share in their information transfer machinery, due to a common ancestor, as proposed by the symbiotic theory (Martin and Muller 1998; Moreira and Lopez-Garcia 1998; Lopez-Garcia 1999; Martin et al. 2001; Esser and Martin 2007).

Archaea constitute one of the three cellular domains in the universal tree of life (Woese 1998) composed of organisms highly diverse in morphology, physiology, and natural habitats (Chaban et al. 2006; Clementino et al. 2007; Nam et al. 2008; Auguet et al. 2009). Organisms included in this cellular domain possess basal transcription machinery resembling that of eukaryotes. For instance, archaea include a TATA box promoter sequence, a TATA box–binding protein (TBP), a homologue of the transcription factor TFIIB (TFB), and a RNA polymerase (RNAp) containing between 8 and 13 subunits (Goede et al. 2006) (see supplementary fig. S1, Supplementary Material online). In contrast,

archaeal messenger RNAs (mRNAs) are structurally similar to bacterial mRNAs, and, most importantly, the majority of identified TFs in archaeal organisms are homologous to bacterial activators and repressors (Kyrpides and Woese 1998; Bell 2005). Indeed, very few eukaryotic-like TFs were found to occur in archaea (Kruger et al. 1998). These observations raise different basic questions with regard to the mechanisms of transcriptional regulation and the manner by which bacterial-like TFs may interact or interfere with the components of the eukaryotic-like basal transcriptional machinery within an archaeal cell. It is for this reason that archaeal DNA-binding TFs represent an important class of proteins to explain the molecular mechanisms that underlie transcription regulation. Even though the ever-growing number of archaeal genome sequences reveals an increasing list of potential regulators (Coulson et al. 2007; Wu et al. 2008), archaeal transcriptional regulation is still poorly documented, and the most detailed and advanced studies have been performed with only a dozen TFs, mainly from the AsnC family (formerly feast/famine protein family) (see supplementary table S1, Supplementary Material online) (Napoli et al. 1999; Leonard et al. 2001; Bell 2005). Initial sequence analysis–based attempts using family-specific models from *E. coli* TFs resulted in a low proportion of bacterial-like TFs in archaea (Perez-Rueda et al. 2004; Coulson et al. 2007). One probable cause for this discrepancy could be that archaeal TF zregulatory repertoire includes additional classes of DNA-binding motifs not observed in *E. coli*, suggesting that our current knowledge on the repertoire of TFs in archaeal genomes is far from being complete. Importantly, comparative genomic analysis of archaea represents an opportunity to fill in this gap and is an indispensable step toward our understanding of gene regulation networks in prokaryotes and eukaryotes.

In the present study, an exhaustive analysis of gene sequences from 52 completely sequenced archaeal genomes to identify potential DNA-binding TFs was performed. In addition, a comparative analysis was carried out to deduce the distribution of TFs and their evolutionary families among the archaeal genome sequences. Using this repertoire of TFs, we show that 1) there is an underrepresentation of the number of TFs in these organisms compared with bacterial genomes, 2) a considerable number of TFs encode for short polypeptides with a significant fraction encoding for single-domain proteins, and 3) a high proportion of TFs are homologous between archaea and bacteria, mainly from the class clostridia of firmicutes.

Materials and Methods

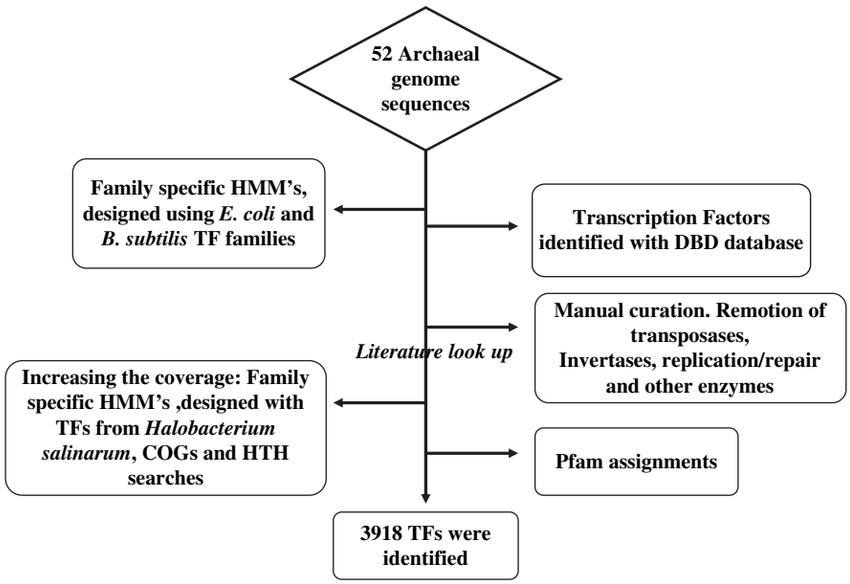
List of Archaeal Genomes Analyzed in This Study

The archaeal genomes analyzed in this work are as follows (see supplementary table S2, Supplementary Material online, for a more detailed annotation of the genomes): Crenarchaea (C): *Aeropyrum pernix* K1, *Caldivirga maquilingensis* IC-167, *Hyperthermus butylicus* DSM 5456, *Ignicoccus hospitalis* KIN4/I, *Metallosphaera sedula* DSM 5348, *Nitrosopumilus maritimus* SCM1, *Pyrobaculum aerophilum* str. IM2, *Pyrobaculum arsenaticum* DSM 13514, *Pyrobaculum calidifontis* JCM 11548, *Pyrobaculum islandicum* DSM 4184, *Staphylothermus marinus* F1, *Sulfolobus acidocaldarius* DSM 639, *Sulfolobus solfataricus* P2, *Sulfolobus tokodaii* str. 7, *Thermofilum pendens* Hrk 5, *Thermoproteus neutrophilus* V24Sta; Euryarchaea (E): *Methanocorpusculum labreanum* Z, *Methanoculleus marisnigri* JR1, *Methanopyrus kandleri* AV19, *Methanosaeta thermophila* PT, *Methanosarcina acetivorans* C2A, *Methanosarcina barkeri* str. *Fusaro*, *Methanosarcina mazei* Go1, *Methanosphaera stadtmanae* DSM 3091, *Methanospirillum hungatei* JF-1, *Methanothermobacter thermautotrophicus* str. Delta H, *Natronomonas pharaonis* DSM 2160, *Picrophilus torridus* DSM 9790, *Pyrococcus abyssi* GE5, *Pyrococcus furiosus* DSM 3638, *Pyrococcus horikoshii* OT3, *Thermococcus kodakarensis* KOD1, *Thermoplasma acidophilum* DSM 1728, *Thermoplasma volcanium* GSS1, uncultured methanogenic archaeon RC-I, *Methanocaldococcus jannaschii* DSM 2661, *Methanococcoides burtonii* DSM 6242, *Methanococcus aeolicus Nankai-3*, *Methanococcus maripaludis* C5, *Methanococcus maripaludis* C6, *Methanococcus maripaludis* C7, *Methanococcus maripaludis* S2, *Methanococcus vannielii* SB, *Archaeoglobus fulgidus* DSM 4304, *Candidatus Methanoregula boonei* 6A8, *Haloarcula marismortui* ATCC 43049, *Halobacterium salinarum* R1, *Halobacterium* sp. NRC-1, *Haloquadratum walsbyi* DSM 16790, *Methanobrevibacter smithii* ATCC 35061; Korarchaeota (K): *Candidatus Korarchaeum cryptofilum* OPF8; Nanoarchaeum (N): *Nanoarchaeum equitans* Kin4-M.

Identification of DNA-Binding TFs

To identify and analyze the repertoire of TFs in 52 archaeal genome sequences, we used a combination of information sources and bioinformatics tools. First, 1,820 putative TFs were collected from Transcription Factor DB (Kummerfeld and Teichmann 2006), a database comprising computationally derived predictions of DNA-binding TFs using the SUPERFAMILY library and PFAM hidden Markov models (HMMs). From this data set, 223 proteins, annotated as transposases, invertases, and integrases, were manually excluded. In brief, this exclusion was based on sequence comparisons against the National Center for Biotechnology Information's nonredundant (NR) protein database (*E* value $= 10^{-3}$) by using Blast search followed by the identification of protein domains with CD-search (*E* value $= 10^{-3}$) (Marchler-Bauer et al. 2007).

In the second phase, 90 family-specific HMMs previously reported for *E. coli* K12 (Perez-Rueda et al. 2004) and 57 family-specific HMMs for *B. subtilis* (Moreno-Campuzano et al. 2006) were used to scan the whole 52 archaeal genome sequences (*E* value threshold $= 10^{-3}$), with the hmmsearch module from HMMer suite of programs (http://HMMER.wustl.edu). Briefly, these HMMs were constructed by using the previously identified TF families in *E. coli* K12 and *B. subtilis* as seeds, considering every protein family's DNA-binding domain (DBD) sequences (around 60 amino acids). Proteins with less than 50% similarity in the DNA-binding region against their corresponding HMM were excluded. At this stage, 424 proteins were identified as potential TFs. This was an important step to explore potential TFs not



[AQ13] FIG. 1. Flowchart showing the different steps involved in the identification of high confidence set of archaeal TFs. Branch points on the vertical line from top to bottom correspond to the stage at which a particular step was taken in the process of obtaining a cleanerdata set.

identified in the first step and vice versa, that is, the coverage of superfamily and PFAM assignations correspond to approximately 70% of the universe of TFs, whereas the rest were complemented with these family-specific HMMs.

In the third phase, 70 new TFs were identified with HMMs constructed from 17 proteins annotated as TFs and not identified in previous searches. This step essentially involved retrieving these 17 TFs from Haloweb server (http://halo4.umbi.umd.edu/cgi-bin/haloweb/nrc1.pl?operation=nrc1), and using them as sequence seeds in Blast searches to retrieve homologous sequences from the NR database with an *E* value = 10^{-3}. Redundancy was removed using CD-hit (Li and Godzik 2006) at 90%, and the potential DBD was identified with CD-search (Marchler-Bauer et al. 2007) (varying the *E* value from 10^{-3} to 10^{-1}) in the remaining proteins. This region was then aligned using ClustalW, with parameters set to default and manually editing output. Finally, 14 HMMs were constructed with the HMMer suite of programs corresponding to the 17 proteins clustered by sequence similarity into 14 different groups. For two proteins, there was not enough information to construct a HMM as they appeared to be lineage specific and no homologues were identified.

In addition, a HMM corresponding to the helix-turn-helix (HTH) DNA-binding motif kindly provided by Yan (2006) was used to identify 686 HTH proteins in the archaeal genomes. This data set was also filtered to exclude those proteins described as transposases, ligases, synthases, synthetases, TFIIB, and TFIIE and those proteins identified in the previous phases, resulting in a total of 95 new probable TFs. Finally, COG assignations associated to TFs in archaea were also used to retrieve new potential archaeal TFs. This resulted in 491 proteins, which were filtered and compared against the whole data set of predictions, but only 2 of them were found to be novel predictions.

All data sets were finally compared and a total set of 3,918 proteins were compiled and used in this study as the final collection of TFs (see fig. 1 for a summary of the steps). This collection of proteins was classified into 75 families by using HMMs deposited in the PFAM DB (Finn et al. 2006) and searches with CD-search server (*E* value = 10^{-1}) and aligned against their corresponding models by using the program hmmalign from HMMer.

Identification of Homologous DNA-Binding TFs in Bacteria and Eukarya

In order to identify TFs, which are homologous to the archaeal set, we compared the whole repertoire against 291 NR genome sequences (Moreno-Hagelsieb and Janga 2008), which included bacterial, archaeal, and eukaryotic sequences. A protein was considered as a homologue of a TF in a given genome if the alignment covered at least $\geq$60% of the query sequence with an *E* value $\leq 10^{-6}$.

Results and Discussion

Identification of DNA-Binding TFs in Archaea

To understand the distribution of TFs in 52 archaeal genomes (34 Euryarchaea, 16 Crenarchaeota, 1 Korarchaeota, and Nanoarchaeota each), we used a HMM-based strategy in two steps. In the first step, we used a battery of family-specific HMMs (see Materials and Methods for details) and DBD assignments characteristic of TFs to scan the archaeal genomes (see fig. 1 for a complete outline). These steps allowed the detection of 3,751 TFs in 52 genomes (see Materials and Methods for a complete list of genomes analyzed), including 53 of the 72 TFs (75%) from *Halobacterium* sp. NRC-1 described so far in the Haloweb server. *Halobacterium* sp. NRC-1 is one of the few archaea whose TF repertoire has been extensively analyzed, and

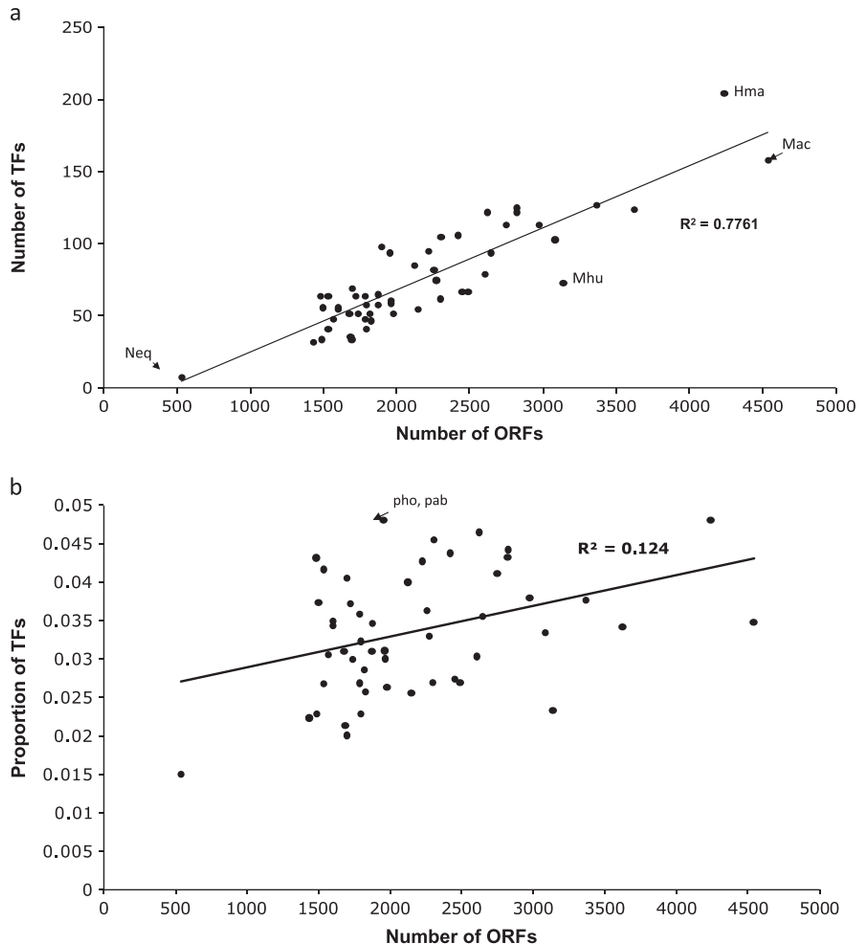


Fig. 2. a) Distribution of TFs identified in 52 archaeal genomes. *Nanoarchaeum equitans* (Neq), *Haloarcula marismortui* (Hma), *Methanospirillum hungatei* (Mhu), and *Methanosarcina acetivorans* C2A (Mac) are indicated as a reference. On x axis, genomes are sorted from smallest to largest size and on y axis the number of TFs is plotted. A linear regression was calculated using the Pearson correlation (r^2) between the number of genes and the total number of TFs. b) Proportion of TFs in all the archaeal genomes. Proportion of TFs was calculated as the fraction of ORFs encoding for TFs and plotted against the total number of ORFs for each genome. *Pyrococcus horikoshii* (pho) and *Pyrococcus abyssi* (pab) are indicated as a reference. On x axis, genomes are sorted from smallest to largest size and on y axis, the fraction of TFs is plotted.

thus, we used its TFs repertoire as a benchmark. In the second step, in order to increase the sensitivity, the 19 *Halobacterium* sp. NRC-1 TFs not identified in the first step were used as seeds for Blast searches against the NR database (*E* value cutoff = 10^{-3}), and the matched proteins were used to build new HMMs for a second round of searches, identifying 70 new TFs. Additionally, archaeal genomes were scanned to look for HTH and COG annotations to identify new potential TFs not identified previously. Because it is known that HTH is one of the most prominent structure associated with TFs in prokaryotes (Perez-Rueda and Collado-Vides 2000, 2001), with at least 80% of the TFs containing this DNA-binding structure, we employed a specific HMM, which considers amino acid residue identity and solvent accessibility, constructed from a set of heterogeneous DNA-binding proteins with standard HTH motifs (Yan 2006). After manually excluding proteins that, although can bind to DNA, are unlikely to be TFs, 97 potential TFs that escaped our HMM-based searches were identified. This composite strategy allowed the detection of additional 167 potential archaeal TFs not identified previously and included all the 72 TFs described in *Halobacterium* sp. NRC-1. In total, a set of 3,918 potential TFs in 52 archaeal genomes were finally identified.

Although extensive survey performed in this work identified a large set of TFs widely distributed in archaea, it is still possible that some potential novel TFs escaped the search criteria or are missing because of their linage-specific nature, presumably due to de novo invention of TFs whose DNA-binding models are not included in our seeddata set.

Dissecting the Repertoire of TFs

Comprehensive identification and characterization of the repertoire of TFs across archaeal genomes are the first step toward expanding the possibilities for exploration of their regulatory networks. Based on our predictions, we found that smaller archaeal genomes contain fewer TFs than larger ones, following a linear correlation (r^2 = 0.82), as has been previously reported for bacteria (Perez-Ruedaet al. 2004; fig. 2*a*). This finding might represent

either an expansion or a contraction of the repertoire of TFs in archaea, as a consequence of adaptation to particular habitats or lifestyles. Although larger genomes might be harboring ampler repertoire of TFs to exploit diverse or more complex habitats, smaller genomes containing fewer regulators might be associated with specific niches. For instance, *E. coli*, which thrives on a large number of sugars, was found to harbor a higher number of TFs compared with *B. subtilis*, which is similar in genome size (Janga and Perez-Rueda 2009). Likewise, we found that the symbiotic hyperthermophile, *N. equitans*, has both a reduced genome and a lower proportion of TFs than other archaea, whereas *Haloarcula marismortui*, a chemoheterotrophic halophilic archaea, was found to have the highest proportion of TFs and *Methanosarcina acetivorans* (an aerobic chemolitho(aceto)autotrophic methanogen, nitrogen fixing) with one of the largest genomes contained the highest the number of TFs among archaeal genomes sequenced so far. An interesting case is that of *Methanospirillum hungatei*, a methanogenic archaea reported to have an unusual filamentous structure, which was found to have the lowest proportion of TFs after *N. equitans* among the archeael genomes studied. Complex lifestyles might require a higher proportion of genes and TFs to better orchestrate responses to changing environments, as is the case of *Methanosarcina acetivorans* that can form aggregate multicellular structures when passing from anaerobiosis to aerobiosis (Oelgeschlager and Rother 2008) or the case of *Haloarcula marismortui*, a halophilic archaea, which are generally described to be surprisingly different in its nutritional demands and metabolic pathways (Falb et al. 2008). In fact, the proportion of TFs in larger genomes is consistent with the hypothesis that an increase of genome complexity and physiological functionality is generally associated with a more complex regulation of gene expression (Woese 1998).

In this context, the number of predicted TFs in archaea is variable (see supplementary table S2, Supplementary Material online), ranging from 8 in the archaeon with the smallest sequenced genome (*N. equitans*) to up to 158 TFs in the largest genome, *Methanosarcina acetivorans* C2A. A closer look into the normalized distribution of TFs calculated as the proportion of the genes coding for TFs gave further insights into the evolution of TFs in the context of their genome size and lifestyles. For instance, as shown in figure 2*b*, less than 5% of the open reading frames (ORFs) in most archaeal genomes are devoted to gene regulation in contrast to about 8–10% observed in bacterial genomes with similar number of ORFs (Perez-Rueda and Collado-Vides 2000, 2001). Indeed, larger archaeal genomes, such as *Methanosarcina acetivorans* and *Haloarcula marismortui*, with similar number of ORFs to *E. coli* K12, encode a lesser proportion of TFs (4.8%, 3.5%, and 8%, respectively). Thus, the TF repertoire observed in archaea is much more similar to bacteria associated with gene loss events, such as intracellular pathogens and endosymbionts (3.9% in average). Notable exceptions are *Pyrococcus horikoshi* and *Pyrococcus abyssi*, two small genomes containing 4.8% and 5.1% of TFs, respectively, comparable with the proportion of TFs in larger archaeal genomes. In contrast, *N. equitans*, which was found to follow the trend in figure 2*a*, exhibited a clear deviation when proportion of genes coding for TFs was compared against genome size.

Although this intriguingly low proportion of TFs in archaea compared with bacteria could be partially explained due to our inability to identify those lineage or organism-specific TFs, it is also possible to suggest that other regulatory strategies in this cellular domain might be compensating for this underrepresentation. These could involve, for example, formation of alternative TBP–TFB–RNAp complexes, with the possibility of interactions with different accessory factors (Baliga et al. 2000; Facciotti et al. 2007). However, the existence of new classes of TFs not explored here or archaeal-specific regulatory mechanisms cannot be excluded to be responsible for this trend. For instance, it has been shown recently from a global analysis of translationally regulated genes in *Halobacterium salinarum* and *Halobacterium volcanii* that 20% and 12% of all genes in these genomes show growth phase–dependent differential translational regulation (Lange et al. 2007). However, the overlap between the two sets was found to be negligible, indicating that archaeal organisms may use differential translational control for regulation of gene expression, adding a layer of regulatory complexity at post-transcriptional level (Mittal et al. 2009). Therefore, regulatory strategies such as either those that are found exclusively in archaea or those that are exploited to a greater extent in archaea compared with bacteria might be responsible for these differences.

Archaeal Genomes Encode a Large Proportion of Small TFs

Transcription regulation in archaea appears to be a chimera, with general TFs being clearly eukaryotic like and candidates for regulating specific responses being bacterial like (Aravind and Koonin 1999). We found that a large proportion (43.5%) of TFs in the archaeal genomes were small in size (100–200 amino acids). In contrast, 42% of the bacterial TFs have between 200 and 300 amino acids (vs. 26.5% of the archaeal TFs with this length). Nonetheless, 287 large TFs with amino acid length greater than 400, corresponding to about 2.3%, were identified in the archaeal repertoire (fig. 3). To determine the significance of these findings, we randomly sampled 1,000 collections of 3,918 proteins from the archaeal genome sequences and compared their lengths with those observed in TFs. As the distribution of average length of proteins in the random samples followed a normal distribution, a *Z* score was used as a test statistic. *Z* score was calculated as the number of standard deviations the observed value (average length of an archaeal TF) is away from the mean of the 1,000 random collections. This is obtained as the ratio of the difference between the observed, x, and the random expected, μ, values to the standard deviation, σ, that is, $Z = (x - \mu)/\sigma$. *P* value was defined as the

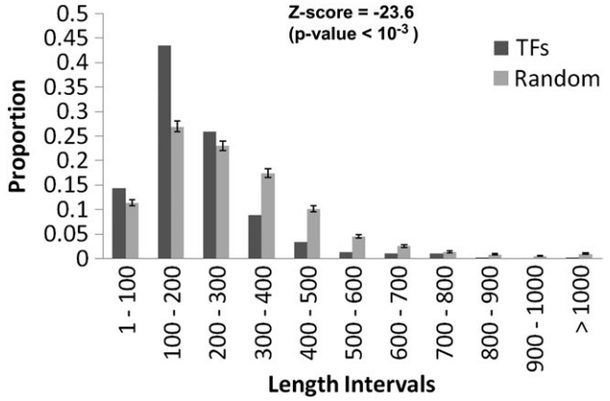


Fig. 3. Distribution of amino acid sequence lengths for TFs. On *x* axis, the intervals of protein size are shown and on *y* axis, the normalized frequency of TFs per interval is shown. Thousand groups of 3,918 protein sequences were randomly retrieved from archaeal genome sequences to compare the length distribution of TFs against other protein-coding genes. In each length internal, bars marked as random represent the proportion of proteins in an interval ± their standard deviations from the average in the random samples.

fraction of the 1,000 random collections that showed an average length greater than or equal to what was observed in the archaeal TF collection. Using this approach for the TF population, a *Z* score of −23.6 (corresponding to a *P* value $<10^{-3}$) was found, indicating that TFs in archaea tend to be significantly smaller than the overall proteome. In contrast, the repertoire of TFs in *E. coli* K12 does not exhibit such a tendency compared with the rest of the proteome (see supplementary fig. S2, Supplementary Material online). In fact, a higher proportion of TFs in *E. coli* are generally longer compared with other proteins, indicating that archaeal TFs are indeed encoded as small genes. To test whether this observation is more general, we compared the lengths of archaeal TFs against a complete set of bacterial TFs available from the DBD database (Kummerfeld and Teichmann 2006). We found that archaeal TFs showed significantly lower lengths compared with bacterial ones (median size of 179 vs. 236 amino acids, $P < 2.2 \times 10^{-16}$, Wilcoxon test; see supplementary fig. S3, Supplementary Material online). Because three of the abundant families, ArsR, AsnC, and HTH_3, were found to be composed of small proteins contributing to about 40% of the total TF repertoire (see below), to exclude the possibility that these large families are indeed responsible for this tendency, we excluded this set of TFs from the complete collection and compared their length distribution with bacterial TFs. This comparison clearly revealed that independent of these large families archaeal TFs show smaller lengths compared with bacterial ones (median size of 190 vs. 236 amino acids, $P < 2.2 \times 10^{-16}$, Wilcoxon test; see supplementary fig. S3, Supplementary Material online). These observations raise the question, if archaeal TFs are shorter than bacterial TFs, do they also encode for smaller number of domains? To address this, we compared the number of domains archaeal TFs possess in comparison with those seen for bacterial ones by obtaining all those TFs for which superfamily domain assignments were available (Madera et al. 2004). Of the 2,621 archaeal TFs for which domain assignments were available, we found that 1,963 comprised single-domain proteins (~75%), whereas single domain containing TFs in bacteria comprised 50% of the total data set analyzed. Further analysis of the distributions of the number of domains in TFs of both the major kingdoms of life unambiguously revealed that archaeal TFs encode for lesser number of domains independent of the exclusion of the large archaeal families ($P < 2.2 \times 10^{-16}$, Wilcoxon test). These results clearly unveil that archaeal TFs comprise a significant proportion of single-domain proteins. One possibility is that most of these one-domain proteins encode for a DBD and might not contain a ligand-binding domain, suggesting that although archaeal TFs contain DBDs similar to bacteria, their mechanism of action might be similar to eukaryotic TFs. In light of these observations, it is possible to hypothesize that archaeal TFs although similar in sequence recognition domains with bacteria (discussed below) might be similar to eukaryotic TFs in mechanistic sense.

The high proportion of small TFs in archaea together with the observation that most archaea have few TFs per genome also suggests a dense combinatorial interplay of TFs for mediating regulation. These data support various possible scenarios namely 1) regulation similar to bacteria, where homodimers can regulate gene expression; 2) formation of different oligomeric assemble forms affected by the interaction with metabolites associated to a particular metabolic state, that is, the formation of oligomers with different sizes, that is, dimers, tetramers, octamers, and so on, as has been observed for the members of the AsnC family (with an average length of around 160 amino acids), whose small TFs can form dimers, tetramers, or octamers with differing regulatory functions (Koike et al. 2004), such as FL11 of *Pyrococcus* sp., which can form a disc or a chromatin-like cylinder upon interaction of two peptides and TrmB of *Pyrococcus furiosus*, which is tetrameric at ambient temperature and octameric in the presence of its inducer (maltotriose or maltose) (Lee et al. 2005; Krug et al. 2006); 3) binding of the same protein to a broad spectrum of compounds or ligands, enhancing its activity under different metabolic states, such as TrmB that binds maltose, sucrose, maltotriose, and trehalose compounds in decreasing order of affinity (Koike et al. 2004; Lee et al. 2005); and 4) alternative physical interactions or co-complex memberships with TBP–TFB–RNAp can also be modulating the structure of the regulatory network in archaea similar to eukarya. In this regard, Facciotti et al. found with protein coimmunoprecipitation, ChIP-Chip, global transcriptional factor (GTF) perturbation and knockout, and measurement of transcriptional changes that global transcriptional factors can associate to nearly half of all putative promoters and show evidence for at least 7 of the 42 possible functional GTF pairs (Baliga et al. 2000; Facciotti et al. 2007).

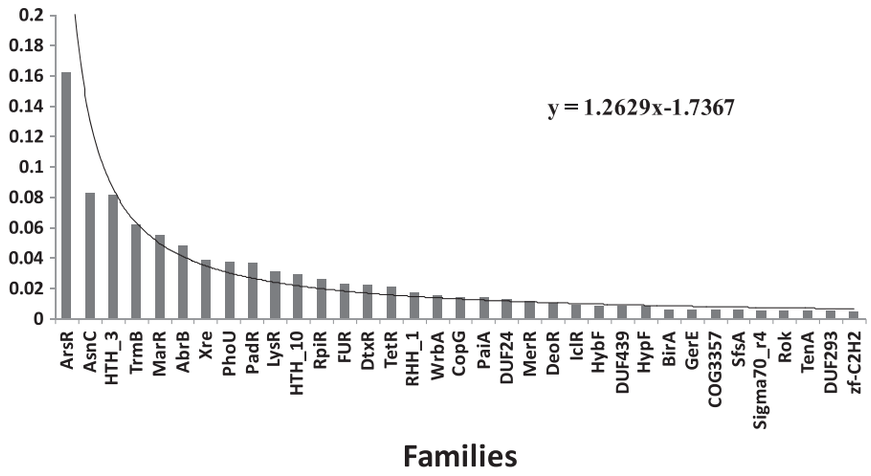


Fig. 4. Abundance of TF families in archaeal genomes. Proportion of TFs in each family was calculated as the fraction of total TFs identified that belonged to a particular family. The families are displayed from largest to smallest size. Families with less than 20 members were not displayed as they corresponded to less than 6% of the totaldata set.

Phylogenetic Distribution of TFs in Archaea

It has been previously proposed that DNA-binding TFs can be grouped into families based on their amino acid sequence similarity (Perez-Rueda and Collado-Vides 2000). In order to determine the number of TF families associated with archaeal genomes, all the 3,918 DNA-binding TFs were grouped into 75 families according to the PFAM database (Finn et al. 2006). As elaborated below, we explored the familial abundance in the archaeal genomes and the relative contribution of each family to the proteome size and overall proportion of TFs. This analysis also enabled us to determine the families that are shared between archaea, bacteria, and eukarya and the main functions of these families.

The population of TF families was found to follow a power-law distribution, with 13 families containing more than 100 members each, representing 71% of the whole TF repertoire (fig. 4). The top three most populated families are ArsR (721 TFs), the HTH_3 (361 TFs), and the AsnC (367 TFs), whereas other ten families contained between 101 and 276 TFs. About 49 families comprised less than 30 TFs, each representing in total ~11% of the TF repertoire. Previous analysis (Moreno-Campuzano et al. 2006; Janga and Perez-Rueda 2009) suggests that global regulators (GRs) in bacteria usually belong to small families; however, in Archaea apparently, this is not the case, at least for the GRs identified so far. For instance, ArsR and TrmB were found to belong to two large families with 721 and 276 members, respectively.

Figure 5 shows that four families are universally distributed across the four archaeal divisions (Crenarchaea, Euryarchaea, Nanoarchaea, and Korarchaea) namely: the HTH_3 (a family of putative activator proteins), AsnC (associated with global regulation of amino acid biosynthesis), TrmB (maltose-specific regulation), and ArsR (detoxification process). These families might belong to the ancestral core of TFs in archaea. A second group of families (PhoU and RpiR) was detected in all archaeal genomes, with the exception of the endosymbiont, *N. equitans*, and hence can also be considered as part of the archaeal TF core set. These families are mainly putative regulators of phosphate uptake (PhoU) and sugar metabolism (RpiR). Based on these findings, it is possible to suggest that archaea from new divisions might carry on TFs from these universal families, potentially regulating central metabolic processes, as might be the case with the last common ancestor of archaea. Some families such as TrpR were found exclusively in *Metallosphaera sedula*, and CopY was found in diverse *Halobacterium* strains suggesting that they might have been transferred laterally from bacteria to archaea.

It is possible to speculate from this data that abundant families like ArsR, AsnC, or HTH_3 might be a consequence of the lifestyles and a response to the deficit of TFs, that is, archaea might have expanded certain families associated with small sizes, to generate a plethora of combinatorial possibilities to regulate their gene expression. It is noteworthy to mention in this context that these three families contribute to around 40% of the total TFs with length between 100 and 200 amino acids.

In order to understand the similarity of TF repertoires per family among the archaeal genomes, a hierarchical centroid linkage-clustering algorithm (Eisen et al. 1998) was applied with uncentered correlation as the similarity measure. The clustering results were visualized using the treeview program (Saldanha 2004). From this clustering, six groups of archaea sharing a common set of TFs were identified (based on a node correlation value $\geq$0.6), whereas three organisms could not be included in any cluster and were hence considered as orphans (see fig. 5). It is evident from this analysis that these six clusters reflect the major taxonomic positions of the organisms analyzed, although some exceptions could be observed. The TF repertoire also reflects the main lifestyle of archaea, such as the

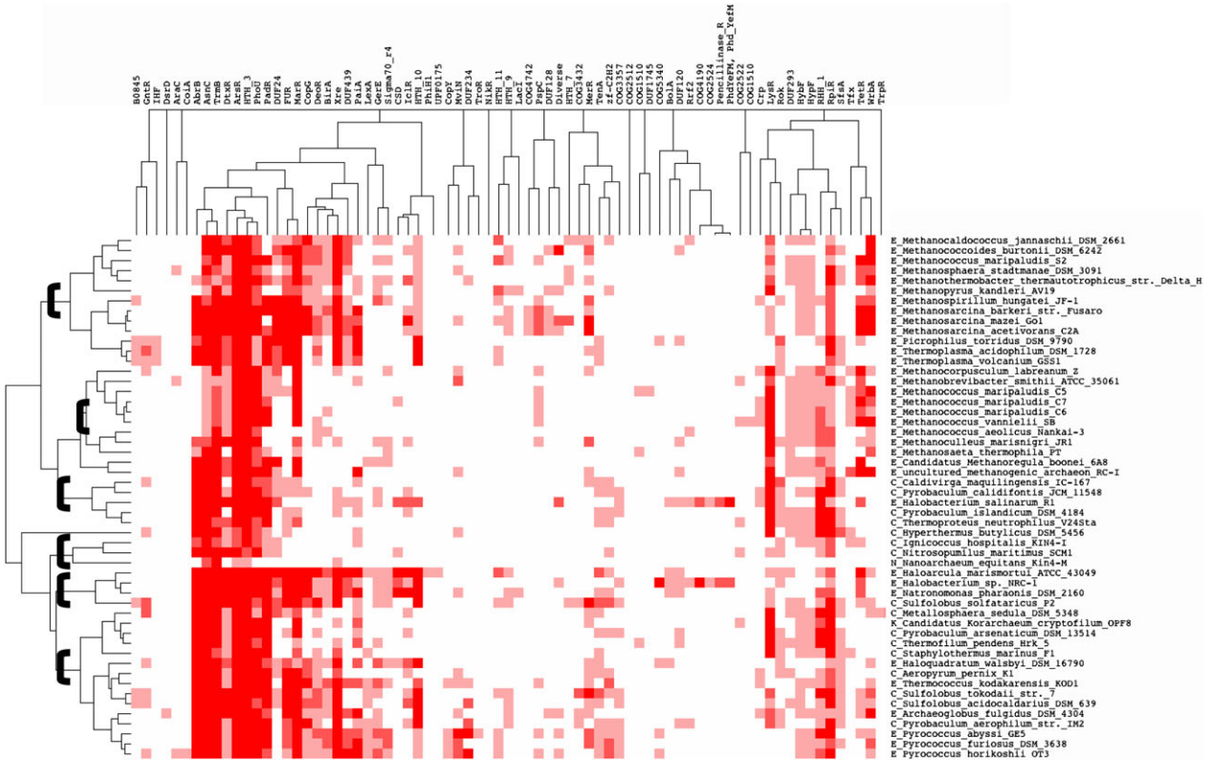


Fig. 5. Clustering of TF families and archaeal genomes. A hierarchical centroid linkage-clustering algorithm was applied with uncentered correlation as the similarity measure and complete linkage (Eisen et al. 1998). Brackets indicate the clusters identified by using a correlation value $\geq$0.6. Nomenclature is as follows: Crenarchaea (C); Euryarchaea (E); Korarchaeota (K), and Nanoarchaeum (N).

first cluster that includes mainly methanogenic archaea (such as *Methanocaldococcus jannaschii* and *Methanococcus maripaludis* S2 among others). The intermixing of organisms in some clusters might be a consequence of lateral gene transfer events, as has been suggested for archaea included in the fourth cluster, that is, *N. equitans* (Nanoarchaeum) and *I. hospitalis* (Desulfurococcales) (Podar et al. 2008).

Comparison of the TF Repertories of Bacteria and Archaea

It has been proposed that bacteria and archaea share a great similarity at gene regulatory level (Aravind and Koonin 1999), with archaeal TFs clearly being bacterial like, whereas their basal transcriptional machinery clearly associated to eukarya. Thus, to understand the degree of conservation of TFs between archaea, bacteria, and eukarya, the probable homologues of the repertoire of transcriptional regulators were identified (see Materials and Methods). From this analysis, it was found that 53% of the 3,918 archaeal TFs exhibit at least one homologue in bacterial genomes (fig. 6). In particular, archaea and clostridia share TFs from the families HTH_3, Xre, and Rrf2, whereas TFs from the families DeoR, IclR, and cold shock are shared with several actinobacteria and some gammaproteobacteria. Another 45% of the 3,918 TFs were clearly identified as archaeal specific, whereas other 6% exhibited homology with bacterial and eukaryotic TFs and about 2% exhibited homology with only eukaryotes (mainly with Ascomycetes) possibly suggesting a lateral gene transfer. This reinforces the notion that TFs of bacteria and archaea share a common ancestry and highlight a close relationship between the TFs from archaea and firmicutes, pointing evidence to drive experiments that can confirm if they share a functional relatedness as well.

Archaeal TFs Are Predominantly Comprised Bacterial DBDs

An important aspect of TFs is their ability to organize into multidomain proteins and hence understanding them in a structural context can provide important clues about how they coordinate regulation. Therefore, the repertoire of archaeal TFs was analyzed using the library of HMMs deposited in superfamily database (Madera et al. 2004). From this analysis, we found that the most abundant DBD in these TFs is the winged helix DBD, detected in 45% of the total set. Followed by the lambda repressor-like DBD (~15%). This result is similar to that previously observed for the repertoire of bacterial TFs, reinforcing the notion of common ancestry in the transcriptional regulatory machinery of prokaryotes (Aravind and Koonin 1999; Aravind et al. 2005). Alternative DBDs, such as IHF-like DBD, PhoU-like domain, nucleic acid–binding domain associated to cold shock proteins or zinc-finger domains,

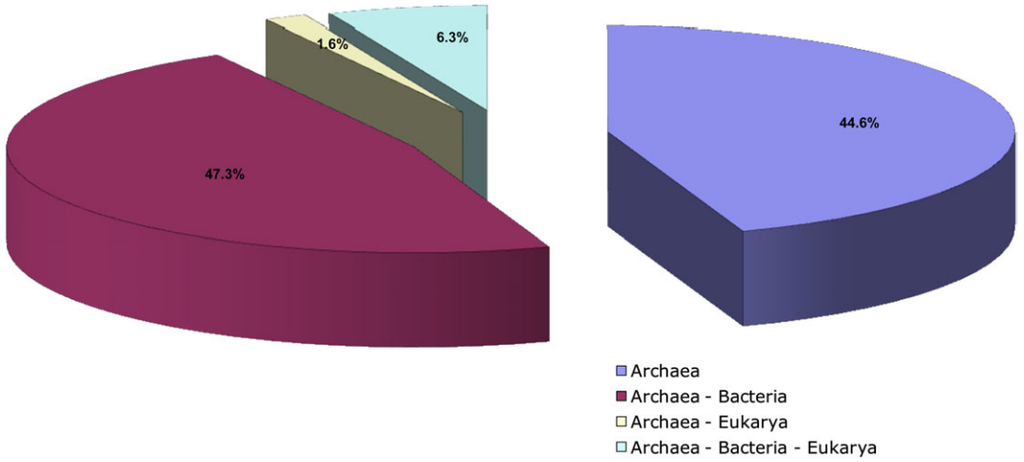


Fig. 6. Distribution of archaeal TFs shared by the three cellular domains, archaea, bacteria, and eukarya. Pie chart showing the distribution of archaeal TF homologues identified in different domains of life; Blast searches were performed between all TFs previously identified against total sequences of bacterial and eukaryotic genomes. A protein was considered as homologue if the alignment covered at least $\geq$60% of the query sequence, with an *E* value $\leq 10^{-6}$.

were also identified, although in lower proportions (corresponding to around 12% of the total TFs). Several of these domains were also identified in bacterial TFs. Zinc fingers represent an intriguing result because this class of proteins has been found exclusively in eukaryotic transcriptional proteins.

Most TF families have been found to undergo lineage-specific duplications resulting in the accumulation of particular families in some microbial species, such as LysR family in *E. coli* (45 TFs; Janga and Perez-Rueda 2009) or ArsR in *Methanosarcina acetivorans* C2A (48 TFs). Indeed, this hypothesis is consistent with the more general notion that a genome evolves from a set of precursor genes to a mature size by gene duplications and increasing modifications (Yanai et al. 2000; Koonin et al. 2002). Therefore, the domain organization and more generally the properties of the TF repertoire described for archaeal genomes in this study open diverse questions like, if the evolution of regulatory networks in archaea is different to that observed in *E. coli*, *B. subtilis*, and/or other biological systems (Aravind and Koonin 1999; Koike et al. 2004; Lee et al. 2005; Lozada-Chavez et al. 2006; Janga et al. 2008, 2009; Perez and Groisman 2009).

Conclusions

In this study, 52 archaeal genome sequences representing a plethora of lifestyles were analyzed to identify the repertoire of proteins involved in controlling the gene expression. Given the fact that there is currently no archaeal genome, which is completely characterized at the level of transcriptional regulation, the repertoire of TFs and the conclusions presented here can be a good starting point in understanding transcriptional regulatory networks in archaeal genomes. In particular, because the archaeal genomes studied here are from different taxa, the results presented here should be valid with high confidence for a wide range of archaea.

Our analysis suggests that although there is a correlation between the number of TFs and genome size, there is also a deficit for TFs in all the archaeal genomes, indicating that this deficit in TFs, and hence, regulatory plasticity is possibly supplemented by their ability to form different assembly structures by small-sized TFs found to be enriched in archaea. We also note that there is an important fraction of transcriptional regulators common to archaea and bacteria. The distribution of TF families common to prokaryotes shows an ancient evolution of transcriptional machinery in bacteria and archaea. We found that the number of TF families is distributed almost homogeneously among all archaea, although there are a small proportion of them that are overrepresented in all archaea but not in bacteria. Further research is necessary to determine the physiological function of such species-specific or shared transcriptional regulators. Nevertheless, the analysis presented here will provide a basis for understanding the organization and evolution of regulatory networks in archaea.

Supplementary Material

Supplementary figures and tables are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

E.P.R. was financed by a grant (ASTF 224-2005) from EMBO and by a grant (IN-217508) from DGAPA-UNAM. E.P.R. thanks Lorenzo Segovia, Claudia Martinez-Anaya, and Javier Diaz-Mejia for their helpful comments in the preparation of the manuscript and Rosa Maria Gutierrez in the clustering analysis. S.C.J. acknowledges financial support from MRC Laboratory of Molecular Biology and Cambridge Commonwealth Trust. We would also like to thank Nitish Mittal and Arthur Wuster for critically reading the manuscript and providing helpful comments.

References

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev.* 29:231–262.

Aravind L, Koonin EV. 1999. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* 27:4658–4670.

Auguet JC, Barberan A, Casamayor EO. 2009. Global ecological patterns in uncultured archaea. *ISME J.*

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol.* 14:283–291.

Babu MM, Teichmann SA. 2003. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* 31:1234–1244.

Baliga NS, Goo YA, Ng WV, Hood L, Daniels CJ, DasSarma S. 2000. Is gene expression in Halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors? *Mol Microbiol.* 36:1184–1185.

Bell SD. 2005. Archaeal transcriptional regulation—variation on a bacterial theme? *Trends Microbiol.* 13:262–265.

Brinkrolf K, Brune I, Tauch A. 2006. Transcriptional regulation of catabolic pathways for aromatic compounds in Corynebacterium glutamicum. *Genet Mol Res.* 5:773–789.

Brune I, Brinkrolf K, Kalinowski J, Puhler A, Tauch A. 2005. The individual and common repertoire of DNA-binding transcriptional regulators of Corynebacterium glutamicum, Corynebacterium efficiens, Corynebacterium diphtheriae and Corynebacterium jeikeium deduced from the complete genome sequences. *BMC Genomics.* 6:86.

Chaban B, Ng SY, Jarrell KF. 2006. Archaeal habitats—from the extreme to the ordinary. *Can J Microbiol.* 52:73–116.

Clementino MM, Fernandes CC, Vieira RP, Cardoso AM, Polycarpo CR, Martins OB. 2007. Archaeal diversity in naturally occurring and impacted environments from a tropical region. *J Appl Microbiol.* 103:141–151.

Coulson RM, Touboul N, Ouzounis CA. 2007. Lineage-specific partitions in archaeal transcription. *Archaea* 2:117–125.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A.* 95:14863–14868.

Esser C, Martin W. 2007. Supertrees and symbiosis in eukaryote genome evolution. *Trends Microbiol.* 15:435–437.

Facciotti MT, Reiss DJ, Pan M, et al. (11 co-authors). 2007. General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci U S A.* 104:4630–4635.

Falb M, Muller K, Konigsmaier L, Oberwinkler T, Horn P, von Gronau S, Gonzalez O, Pfeiffer F, Bornberg-Bauer E, Oesterhelt D. 2008. Metabolism of halophilic archaea. *Extremophiles* 12:177–196.

Finn RD, Mistry J, Schuster-Bockler B, et al. (13 co-authors). 2006. Pfam: clans, web tools and services. *Nucleic Acids Res.* 34:D247–D251.

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, et al. (19 co-authors). 2008. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* 36:D120–D124.

Goede B, Naji S, von Kampen O, Ilg K, Thomm M. 2006. Protein-protein interactions in the archaeal transcriptional machinery: binding studies of isolated RNA polymerase subunits and transcription factors. *J Biol Chem.* 281:30581–30592.

Janga SC, Collado-Vides J. 2007. Structure and evolution of gene regulatory networks in microbial genomes. *Res Microbiol.* 158:787–794.

Janga SC, Collado-Vides J, Babu MM. 2008. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc Natl Acad Sci U S A.* 105:15761–15766.

Janga SC, Perez-Rueda E. 2009. Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. *Comput Biol Chem.* 33:261–268.

Janga SC, Salgado H, Martinez-Antonio A. 2009. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Res.* 37:3680–3688.

Koike H, Ishijima SA, Clowney L, Suzuki M. 2004. The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription. *Proc Natl Acad Sci U S A.* 101:2840–2845.

Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–223.

Krug M, Lee SJ, Diederichs K, Boos W, Welte W. 2006. Crystal structure of the sugar binding domain of the archaeal transcriptional regulator TrmB. *J Biol Chem.* 281:10976–10982.

Kruger K, Hermann T, Armbruster V, Pfeifer F. 1998. The transcriptional activator GvpE for the halobacterial gas vesicle genes resembles a basic region leucine-zipper regulatory protein. *J Mol Biol.* 279:761–771.

Kummerfeld SK, Teichmann SA. 2006. DBD: a transcription factor prediction database. *Nucleic Acids Res.* 34:D74–D81.

Kyrpides NC, Woese CR. 1998. Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families. *Proc Natl Acad Sci U S A.* 95:3726–3730.

Lange C, Zaigler A, Hammelmann M, Twellmeyer J, Raddatz G, Schuster SC, Oesterhelt D, Soppa J. 2007. Genome-wide analysis of growth phase-dependent translational and transcriptional regulation in halophilic archaea. *BMC Genomics.* 8:415.

Lee SJ, Moulakakis C, Koning SM, Hausner W, Thomm M, Boos W. 2005. TrmB, a sugar sensing regulator of ABC transporter genes in Pyrococcus furiosus exhibits dual promoter specificity and is controlled by different inducers. *Mol Microbiol.* 57:1797–1807.

Lee TI, Rinaldi NJ, Robert F, et al. (21 co-authors). 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298:799–804.

Leonard PM, Smits SH, Sedelnikova SE, Brinkman AB, de Vos WM, van der Oost J, Rice DW, Rafferty JB. 2001. Crystal structure of the Lrp-like transcriptional regulator from the archaeon Pyrococcus furiosus. *EMBO J.* 20:990–997.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

Lopez-Garcia P. 1999. DNA supercoiling and temperature adaptation: a clue to early diversification of life? *J Mol Evol.* 49:439–452.

Lozada-Chavez I, Janga SC, Collado-Vides J. 2006. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 34:3434–3445.

Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. 2004. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* 32:D235–D239.

Marchler-Bauer A, Anderson JB, Derbyshire MK, et al. (25 co-authors). 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* 35:D237–D240.

Martin W, Hoffmeister M, Rotte C, Henze K. 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol Chem.* 382:1521–1539.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37–41.

Mittal N, Roy N, Babu MM, Janga SC. 2009. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc Natl Acad Sci U S A.* 106:20300–20305.

Moreira D, Lopez-Garcia P. 1998. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol.* 47:517–530.

Moreno-Campuzano S, Janga SC, Perez-Rueda E. 2006. Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes—a genomic approach. *BMC Genomics*. 7:147.

Moreno-Hagelsieb G, Janga SC. 2008. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins* 70:344–352.

Nam YD, Chang HW, Kim KH, Roh SW, Kim MS, Jung MJ, Lee SW, Kim JY, Yoon JH, Bae JW. 2008. Bacterial, archaeal, and eukaryal diversity in the intestines of Korean people. *J Microbiol.* 46:491–501.

Napoli A, van der Oost J, Sensen CW, Charlebois RL, Rossi M, Ciaramella M. 1999. An Lrp-like protein of the hyperthermophilic archaeon Sulfolobus solfataricus which binds to its own promoter. *J Bacteriol.* 181:1474–1480.

Oelgeschlager E, Rother M. 2008. Carbon monoxide-dependent energy metabolism in anaerobic bacteria and archaea. *Arch Microbiol.* 190:257–269.

Perez JC, Groisman EA. 2009. Evolution of transcriptional regulatory circuits in bacteria. *Cell* 138:233–244.

Perez-Rueda E, Collado-Vides J. 2000. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res.* 28:1838–1847.

Perez-Rueda E, Collado-Vides J. 2001. Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol.* 53:172–179.

Perez-Rueda E, Collado-Vides J, Segovia L. 2004. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem.* 28:341–350.

Podar M, Anderson I, Makarova KS, et al. (27 co-author). 2008. A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biol.* 9:R158.

Saldanha AJ. 2004. Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.

Sierro N, Makita Y, de Hoon M, Nakai K. 2008. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.* 36:D93–D96.

Woese C. 1998. The universal ancestor. *Proc Natl Acad Sci U S A.* 95:6854–6859.

Wu J, Wang S, Bai J, Shi L, Li D, Xu Z, Niu Y, Lu J, Bao Q. 2008. ArchaeaTF: an integrated database of putative transcription factors in archaea. *Genomics* 91:102–107.

Yan C. 2006. A hidden Markov model approach to model protein sequence and structural information: identification of helix-turn-helix DNA-binding motif In: Proceedings of IEEE International Conference on Granular Computing. p. 385–388. [**AQ12**]

Yanai I, Camacho CJ, DeLisi C. 2000. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys Rev Lett.* 85:2641–2644.

The original printed version of the thesis includes a copy of a book chapter to be printed with Horizon Press in 2011:

Janga, S.C. and Moreno-Hagelsieb, G. 'Operons and bacterial genome organization' to be published with **Horizon Scientific Press** for an edited book on 'Bacterial Gene Regulation and Transcriptional Networks' (Ed: M. Madan Babu, MRC Laboratory of Molecular Biology, Cambridge, U. K)

This chapter has been removed from the electronic file for copyright reasons.

Figure 1a

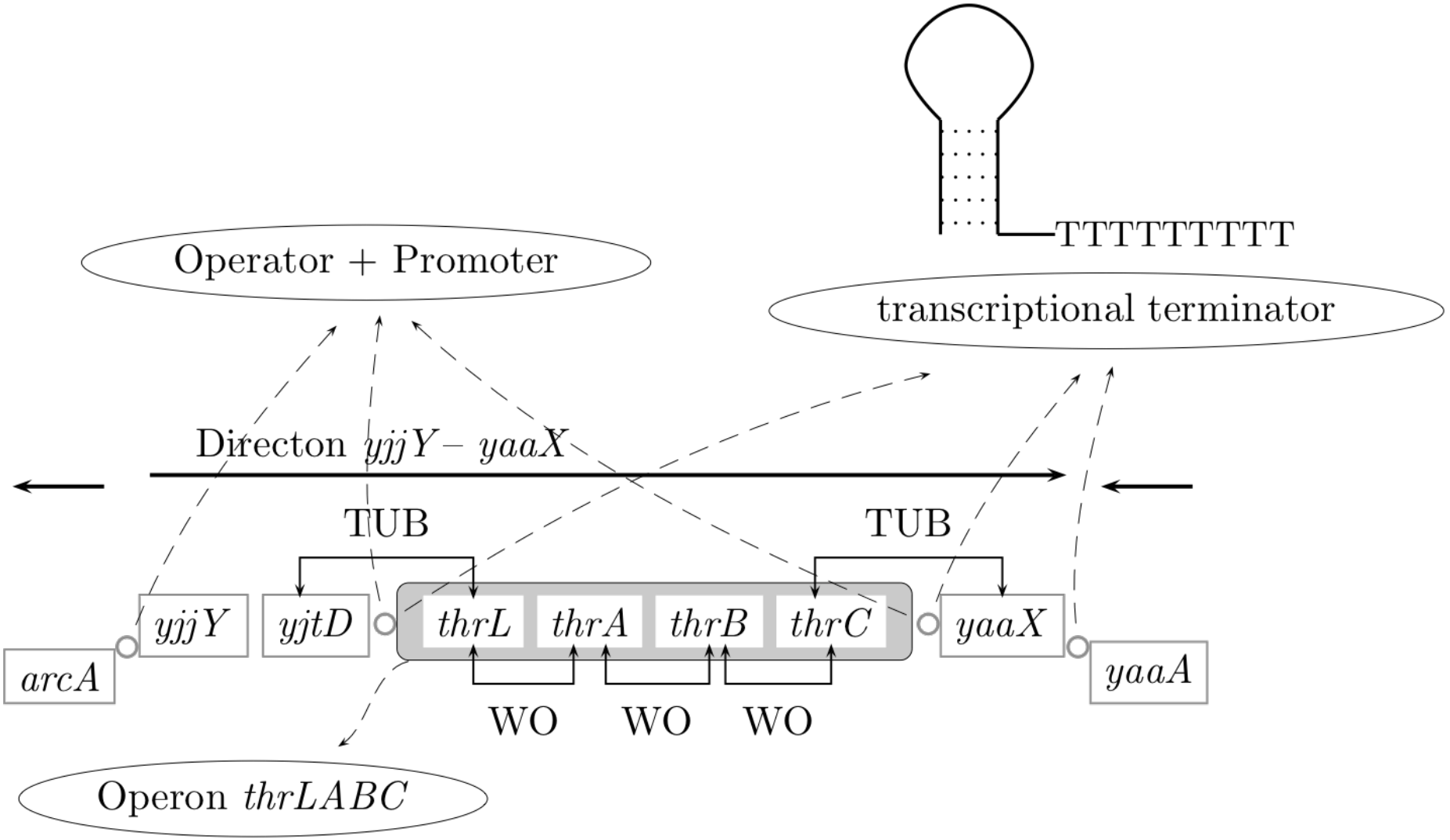

TTTTTTTTT
Operator + Promoter
transcriptional terminator
Directon yjjY – yaaX
TUB
TUB
yjjY
yjtD
thrL
thrA
thrB
thrC
yaaX
arcA
yaaA
WO
WO
WO
Operon thrLABC

Figure 1b

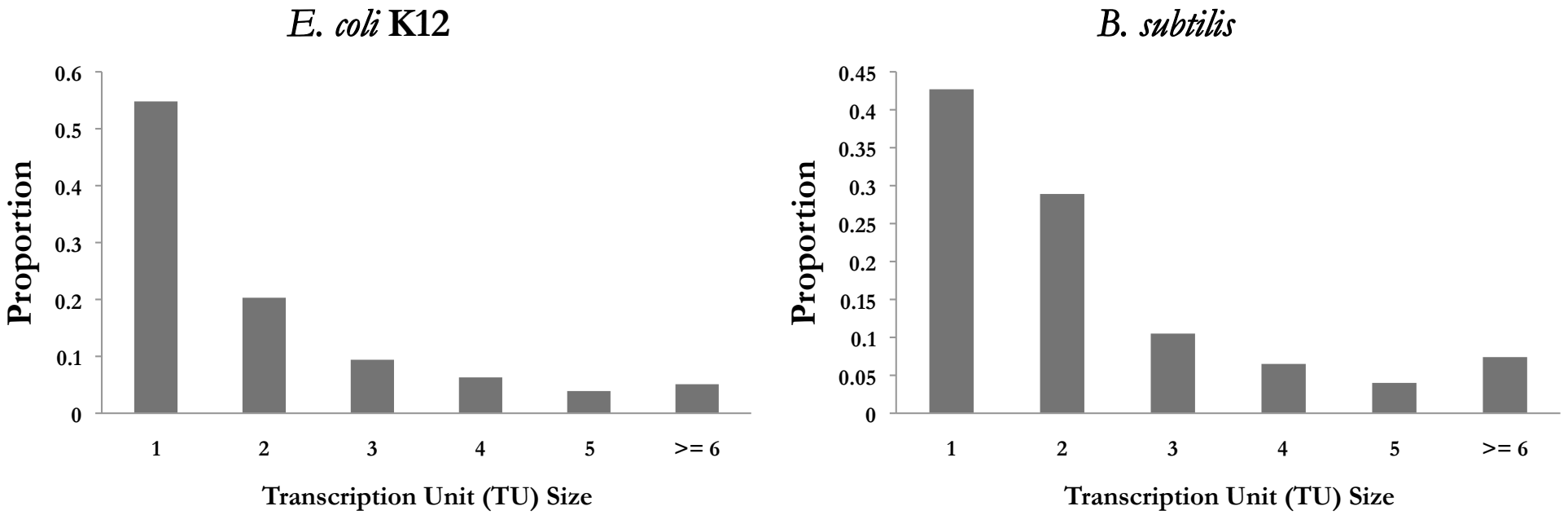

E. coli K12
Proportion
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
>= 6
Transcription Unit (TU) Size
B. subtilis
Proportion
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
>= 6
Transcription Unit (TU) Size

Figure 2

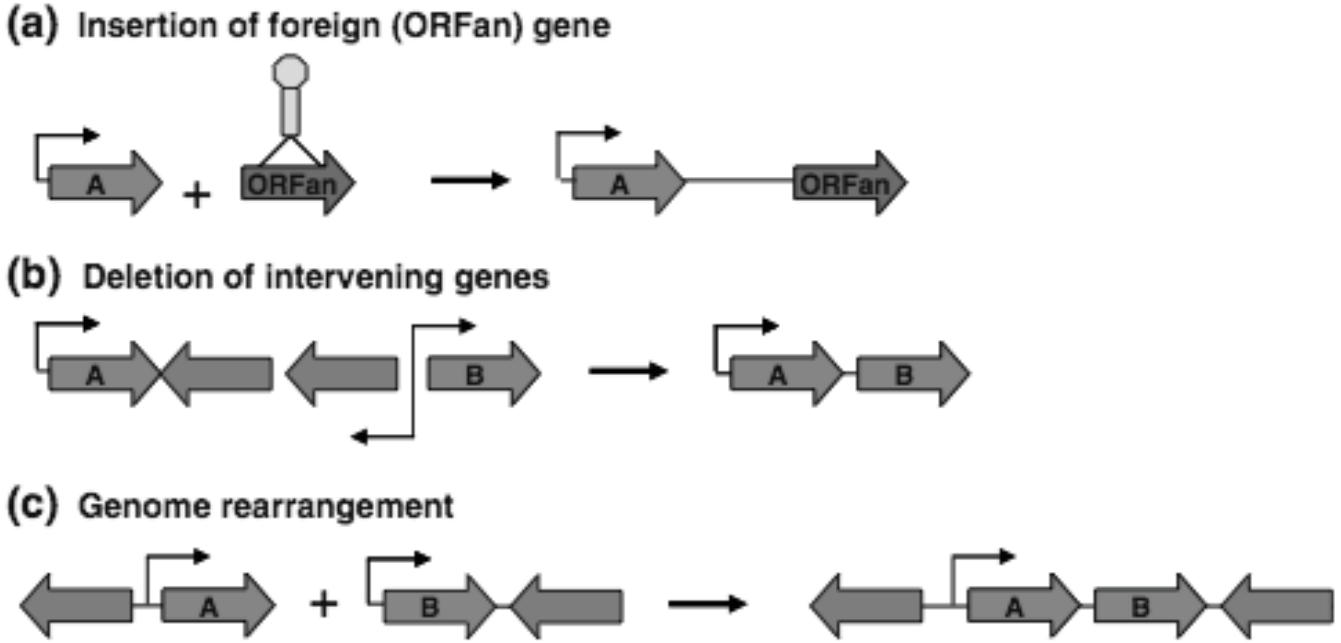

(a) Insertion of foreign (ORFan) gene
A
ORFan
A
ORFan
(b) Deletion of intervening genes
A
B
A
B
(c) Genome rearrangement
A
B
A
B

Figure 3

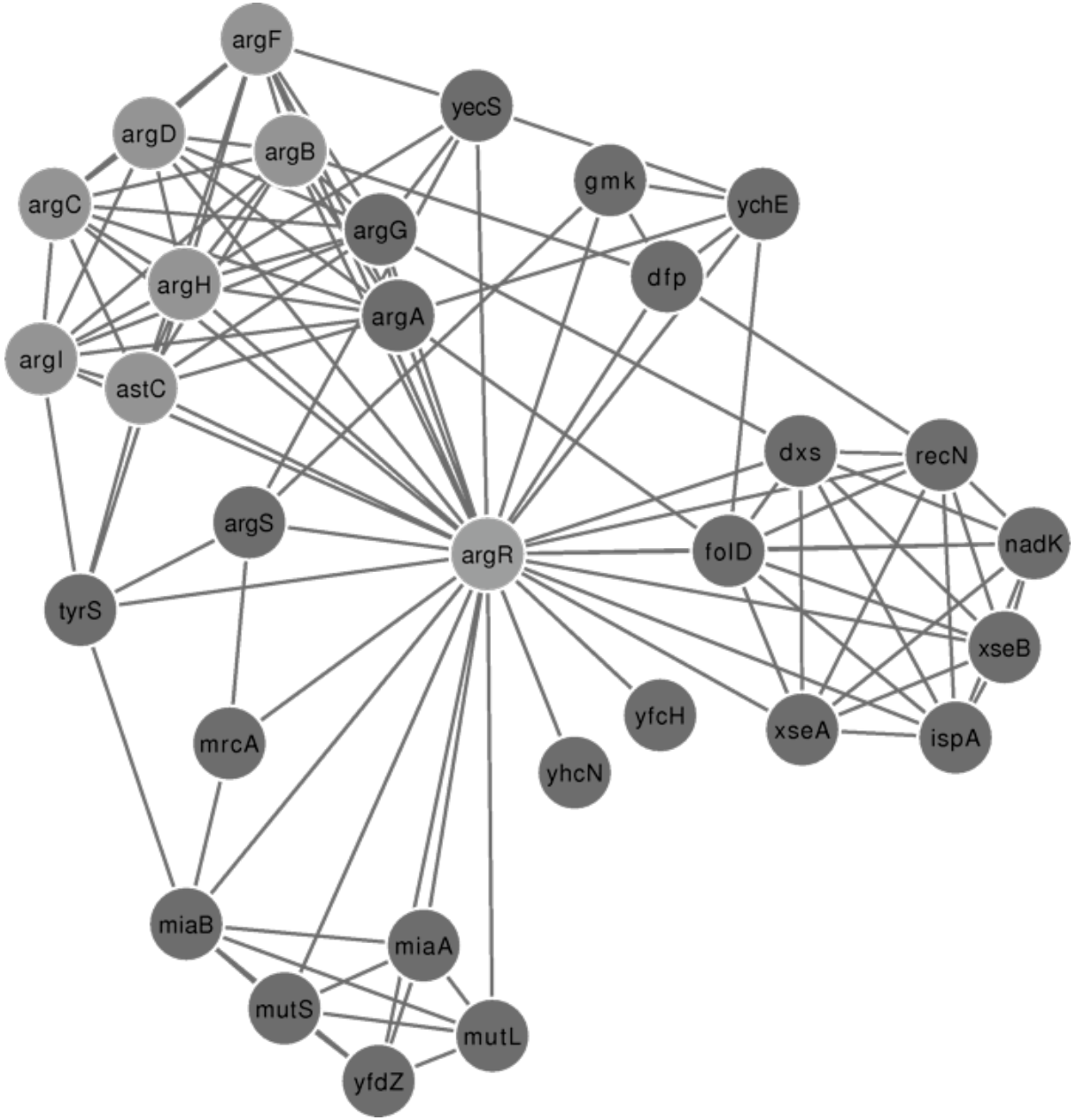

argF
yecS
argD
argB
gmk
argC
ychE
argG
dfp
argH
argA
argI
astC
dxs
recN
argS
argR
folD
nadK
tyrS
xseB
yfcH
xseA
ispA
mrcA
yhcN
miaB
miaA
mutS
mutL
yfdZ

Figure 4a

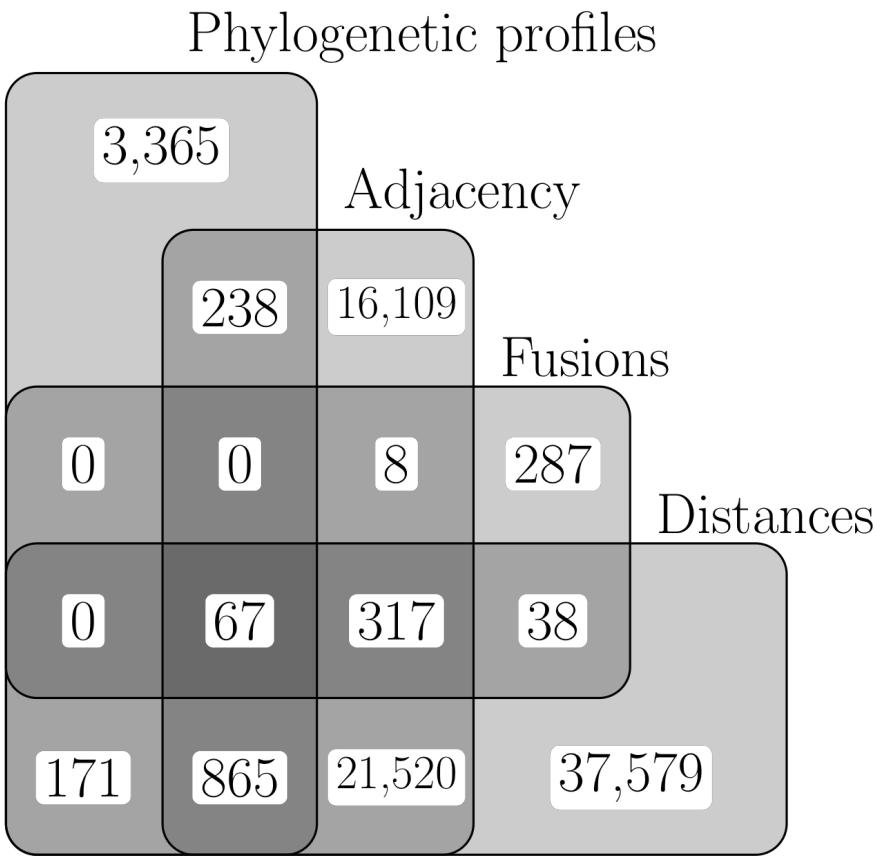
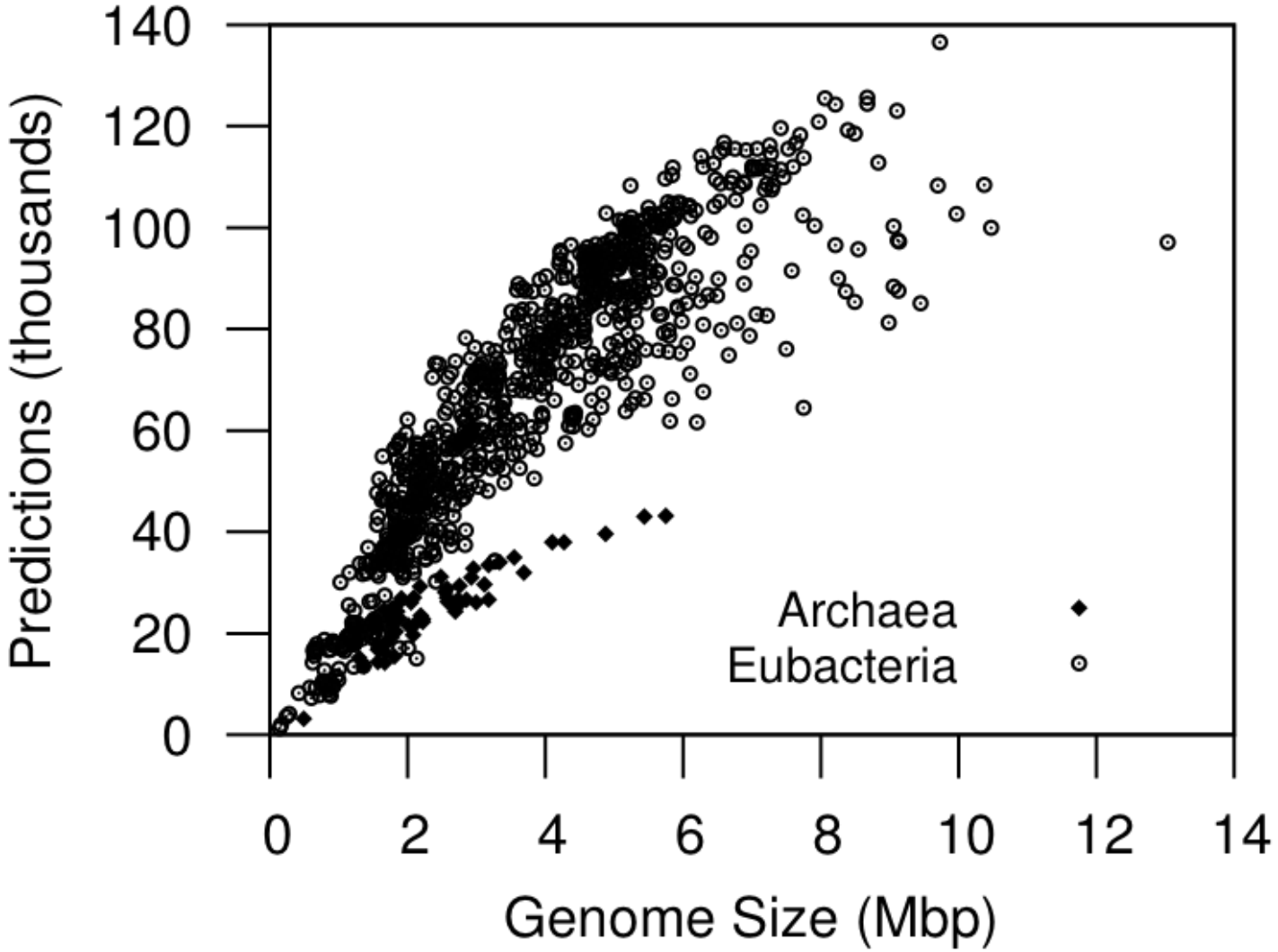

Phylogenetic profiles
Adjacency
Fusions
Distances
3,365
238
16,109
0
0
8
287
0
67
317
38
171
865
21,520
37,579

Figure 4b


Predictions (thousands)
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
Genome Size (Mbp)
Archaea
Eubacteria