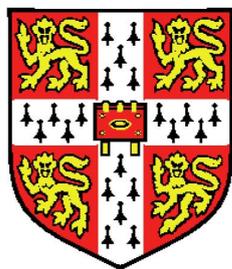


Integration and analysis of protein evolutionary relationships and small molecule bioactivity data



Felix A Krüger
Fitzwilliam College
University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

7th October 2013

To Predocs.

Declaration

INTEGRATION AND ANALYSIS OF PROTEIN EVOLUTIONARY RELATIONSHIPS AND SMALL MOLECULE BIOACTIVITY DATA

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

This thesis has been typeset in 12pt font using Latex according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Acknowledgements

I would like to thank John Overington, who has been a great source of inspiration and a caring supervisor. I owe him thanks for sharing his vast knowledge of the field and providing both vision, and an abundance of ideas. John's trust gave me the confidence to develop the project in my own style. I commend John for being both, cunning fleet admiral set on conquests in computational chemical biology as well as gate-keeper to the eclectic web of cat-videos and chemistry jokes.

I would further like to thank the members of my TAC committee, Tom Blundell, Julio Saez-Rodriguez and Peer Bork, who provided valuable input for my project and more than often took the time to meet up, discuss, and give advice well beyond what could be expected of them. I also thank Mikhail Spivakov and Andreas Bender for their friendship and honest mentoring advice.

The members of the ChEMBL group all deserve my gratitude, having supported me throughout, in technical questions and those that can only be described as 'meta'. Anna Gaulton, Jon Chambers, Patricia Bento, Rita Santos, Louisa Bellis, Shaun McGlinchey and Yvonne Light for helping me navigate the complexities of the database schema and John's itineraries. Anne Hersey, especially, for being a reliable first point of call and scheduler of group meetings. George Papadatos, Michal Nowotka, Francis Atkinson, Ruth Akhtar, Nathan Dedman, and Mark Davies for exposing me to Knime-evangelism, Pipeline Pilot orthodoxy, SQL and no-SQL wizardry, worship of the RDKit and irrational fear of busses. Gerard van Westen and Grace Mugumbate for adopting us graduate students as their own.

I would also like to thank those that worked with me on aspects of the project. Raghd Rostom who staid with us as a summer student and was a great help in implementing the mapping to Pfam domains. Albert Vilella, who introduced me to EnsemblCompara, Saqib Mir for answering all my PDBeMotif-related questions, Penny Coggill for Pfam-related questions, and Samuel Croset for exploratory text-mining work. It was also reassuring to have Anna Gaulton's support in all things related to the production-side of the ChEMBL

database.

I am grateful to Benjamin Stauch, Nenad Bartonicek, Tom Blundell, Jennifer Yen, and John Overington for reading and correcting draft chapters of my thesis.

I owe my friends a big thank you for helping me see the greater picture in life and simply for being such awesome people. Nenad Bartonicek, Anika Oellrich, Petra Schwalie, Angela Goncalves and Michele Mattioni for welcoming me to the Predoc community and staying connected throughout. Myrto Kostadima, Benedetta Baldi, Sander Timmer, Christine Seeliger, and Felipe Cadete for companionship from the day I set foot in the EMBL guest house. Benjamin Stauch and Samuel Croset as brothers in arms in the ChEMBL group. Steve, Remco, Nils, Konrad, Maria, JB and the community of EBI predocs as a whole deserve praise for general loveliness and creating a wildly positive vibe. I am also grateful to my subidon friends from Fitz, Charles Ravarani, Alexis De La Ferriere and Narseo Vallina Rodriguez, who have kept me steady company on expeditions into foreign scientific territories, both real and imagined. I thank Esther Preussler, Kilian Ströder, Julia Haseleu, Simon Rosenberg and Sina Wagner for their faithful friendship.

My housemates on Stockwell Street, Jennifer Yenn, Ricardo Milho, Laure Lam Hung, and Eve Coomber have made me feel welcome and at home and I am grateful for it. Jenn and Laure further deserve praise for generously sharing their food, cars, and entertaining theories of characteristic German behaviour.

I am deeply grateful to Siobhan Williams for her patience, and love, and the endless hours spent travelling between Cambridge and London.

Finally I want to thank my family and especially my parents Marjorie and Willi, and my sister Andrea for being there for me, and believing in me. The support and trust they gave means a lot to me.

Abstract

Interactions of small organic molecules and proteins have been studied extensively in the search of therapeutic drugs. Historically, the interaction partners have been attributed to separate scientific disciplines: small organic molecules to the domain of chemistry, proteins to the domain of biology. Likewise, chemical and biological data have been stored and maintained separately. The aim of my thesis was to integrate the ChEMBL database, a public repository of small molecule bioactivity measurements, with resources of protein evolutionary relationships, and exploit these new links to further our understanding of small molecule bioactivity.

In order to link biological assays via the evolutionary relationships of their protein targets, I established a mapping of small molecule binding to specific structural protein domains - the fundamental building blocks of protein architecture and evolution. By mapping small molecule binding to protein domains, I was able to examine links between the properties of small molecules and the evolutionary units that mediate their binding. I used domain definitions from Pfam, a database of protein domains derived from conserved sequence blocks. The mapping is now an integral part of the ChEMBL database and can be used to limit sequence-based queries to sequence partitions that are relevant to small molecule binding.

Further, I integrated information from the homology resource EnsemblCompara Genetrees with bioactivity data from ChEMBL to examine the conservation of small molecule potency between homologous proteins within and across species. Potency differences between related proteins are a useful indicator of small molecule specificity. Specificity is an early milestone for most drug discovery projects as it allows for the manipulation of a desired process in a targeted manner, with side effects reduced to a minimum. I examined pairs of closely related human proteins and found that potency differences were overall greater than the estimated background noise. Using the outlined integration approach in a cross-species comparison, I also observed that potency differences between pairs of related proteins in human and rat were overall no greater than the background noise. This is relevant to the use of model organisms for drug discovery, which relies on extrapolation from a measured response in one species to a therapeutic effect in humans.

Taken together I have integrated small molecule bioactivity and protein evolutionary data from two resources, Pfam and EnsemblCompara Genetrees. This has provided a framework for studying small molecule binding in the context of protein evolution.

Contents

Declaration	iii
Contents	ix
List of Figures	xv
List of Tables	xvii
List of symbols	xviii
1 Introduction	1
1.1 The quest for small molecules as therapeutics and research tools	2
1.1.1 Small molecules as therapeutics	5
1.1.2 Small molecules as research tools	8
1.2 Data integration in chemical biology and drug discovery	12
1.3 Bioactivity data	15
1.4 Protein evolutionary relationships	19
1.4.1 Protein Homology	20
1.4.1.1 Computational approaches to assign protein homologous relationships	21
1.4.2 Protein domains	23
1.4.2.1 Frameworks to detect and represent protein domains . .	24
1.5 Aims of the analyses	27
2 Mapping of small molecule binding to protein domains	29
2.1 Introduction	29

2.1.1	The protein domain concept in drug discovery	30
2.1.2	A heuristic mapping that relies on domain-based annotation transfer	32
2.1.3	Outline	36
2.2	Results	37
2.2.1	Small molecule binding within the boundaries of Pfam-A domains	37
2.2.2	Domain coverage of the human genome and ChEMBL target dictionary	38
2.2.3	A catalogue of protein domains with known small molecule interactions	41
2.2.4	Mapping small molecule binding by domain-based annotation transfer	43
2.3	Discussion	48
2.3.1	Prerequisites of the mapping heuristic	48
2.3.1.1	Small molecule binding within the boundaries of Pfam-A domains	48
2.3.1.2	The implications of Pfam-A model coverage	49
2.3.2	Small molecule binding to Pfam-A domains and evidence from ChEMBL	51
2.3.2.1	The HN domain	51
2.3.2.2	The Carb_anhydrase domain	52
2.3.2.3	The Pantoate_ligase domain	53
2.3.3	Limitations of a widely applicable mapping heuristic	55
2.3.3.1	Uncatalogued Pfam-A domains	56
2.3.3.2	Small molecule binding at domain interfaces	58
2.4	Conclusion	62
2.5	Methods	64
2.5.1	Retrieval of Pfam-A annotations	64
2.5.2	Retrieval of protein coding genes in human genome	64
2.5.3	Evidence of small molecule binding for single-domain proteins . .	64
2.5.4	Removal of protein fragments	65
2.5.5	Count of projected domains	67
2.5.6	Protein-ligand pairings in PDBe	67
2.5.7	Retrieval of ligand binding residues from Uniprot	67
2.5.8	Retrieval of ligand binding residues derived from PDBeMotif . . .	68

2.5.9	Mapping ChEMBL compounds to PDBe identifiers	68
2.5.10	Translation of residue numbers between PDBeMotif and Uniprot .	69
2.5.11	Small molecule binding within the boundaries of Pfam-A domains	69
2.5.12	Small molecule binding at domain interfaces	70
3	Refined mapping of small molecule binding to protein domains	73
3.1	Introduction	73
3.1.1	Proteins with multiple small molecule binding sites	74
3.1.2	Outline	78
3.2	Results	78
3.2.1	A workflow for manual refinement and integration with the ChEMBL database	78
3.2.2	Prototype of a manual curation platform	82
3.2.3	Refinement of the catalogue of domains with evidence for small molecule binding	84
3.2.3.1	Small molecule binding to the SH2 domain	87
3.2.3.2	Small molecule binding to the Lig_chan domain	87
3.2.3.3	Small molecule binding to the ANF_receptor domain	88
3.2.3.4	Small molecule binding to the 7tm_3 domain	89
3.2.3.5	Small molecule binding to the 7tm_2 domain	89
3.2.4	Manual curation of conflicting mappings	90
3.2.5	Coverage of measurements in the ChEMBL database	93
3.3	Discussion	94
3.3.1	Changes to the catalogue of domains with known small molecule interactions	94
3.3.2	Conclusions	98
3.4	Methods	98
3.4.1	Loading routine	98
3.4.2	Scope of the mapping	99
3.4.3	Export routine	99
3.4.4	Mapping tables	100
3.4.5	Catalogue of Pfam-A domains with known small molecule interactions	101
3.4.6	Prototype of a curation platform	101

3.4.7	Standard procedure of manual curation	102
3.4.8	Coverage and network view	103
4	Integration of small molecule potency measurements with the phy-	
	logeny of their protein targets	107
4.1	Introduction	107
4.1.1	Phylogenetic relationship between drug targets	108
4.1.2	Rats as model organisms in drug discovery	109
4.1.3	Outline	111
4.2	Results and Discussion	113
4.2.1	Controlling for assay variability in the ChEMBL data set	113
4.2.2	Conservation of potency between human-rat orthologs	117
4.2.3	Conservation of potency between human paralogs	120
4.2.4	Data model	123
4.2.5	Evolutionary relationship and conservation of potency	126
4.2.6	Sequence identity, ligand molecular weight and potency differences in paralogous pairs	128
4.2.7	Assessment of individual homologous pairs	133
4.2.8	Small molecules as probes of the binding site	137
4.3	Conclusion	141
4.4	Methods	143
4.4.1	Retrieval and processing of measured potencies	143
4.4.2	Data assembly for inter-assay comparison	143
4.4.3	Data assembly for homologous pairs	144
4.4.4	Sequence identity for full-length proteins	146
4.4.5	Sequence identity on a Pfam domain level	146
4.4.6	Sequence identity on a binding site level	146
4.4.6.1	Assessment of potency differences between homologous proteins	147
4.4.6.2	Data models	148
4.4.6.3	Assessment of individual homologous pairs	149
4.4.6.4	Potency differences and sequence identity	150
4.4.7	Homology model of the HRH3 receptor	150

4.4.7.1	Preparation of model templates	150
4.4.7.2	Sequence alignment	150
4.4.7.3	Model building and visualization	151
4.4.8	Cluster analysis of HRH3 ligands	153
4.4.9	Mapping small molecule binding to Pfam domains	154
5	Conclusions	155
5.1	Integration of small molecule bioactivity data and protein domain annotation	155
5.2	Integration of small molecule bioactivity data and protein homology information	157
5.3	Summary, conclusion and outlook	161
	List of Publications	163
	Appendix	165

List of Figures

1.1	Illustration of a small molecule-protein interaction: biotin and streptavidin.	4
1.2	Schematic diagram of homologous relationships.	21
1.3	Schematic illustration of domain fusion.	24
2.1	Domain poisoning of repurposing queries.	32
2.2	Schematic illustration of homology-based transfer of binding annotation.	34
2.3	Small molecule binding within Pfam-A domains.	39
2.4	Overview of Pfam-A coverage.	42
2.5	Small molecule binding within Pfam-A domains.	46
2.6	Evidence for small molecule binding to the <code>HN</code> domain.	52
2.7	Evidence for small molecule binding to the <code>Carb_anhydrase</code> domain. . .	54
2.8	Evidence for small molecule binding to the <code>Pantoate_ligase</code> domain. . .	55
2.9	Small molecule binding at domain interfaces of ‘enzyme doublet’ architectures.	61
2.10	Small molecule binding at the interface of two unusual architectures. . . .	63
3.1	Examples of proteins with multiple small molecule interactions sites . . .	75
3.2	Flow chart for manual refinement and integration with ChEMBL release cycle	80
3.3	Sitemap of the curation platform user interface	83
3.4	Binding of two small molecules at the interface of the <code>Lig_chan-Glu_bd</code> and <code>Lig_chan</code> domains	88
3.5	Network graph of domain co-occurrences in the ChEMBL target dictionary	95
4.1	Tree of evolutionary distances.	110

4.2	Overview of inter-assay variability for human and rat proteins.	116
4.3	Overview of inter-assay variability in ChEMBL.	117
4.4	Overview of variability between human and rat orthologs in ChEMBL.	119
4.5	Graph of relationships within the paralog data set.	122
4.6	Overview of variability between human paralogs in ChEMBL.	123
4.7	Model fitting	125
4.8	Evolutionary relationship and conservation of potency.	128
4.9	Genome-wide sequence identity of homologous proteins.	128
4.10	Sequence identity and absolute potency differences.	130
4.11	Ligand molecular weight and absolute potency differences.	131
4.12	Volcano plots for orthologs and paralogs.	134
4.13	Clustering of ligands of the HRH3 receptor.	139
4.14	Homology models of the HRH3 receptor.	140
4.15	Probability density function fitted to sampled data.	145
4.16	Template alignment for the HRH3 model.	152
1	Multiple binding sites of the mTOR complex	165
2	Evidence for small molecule binding to the <code>HydroLase_4</code> domain.	170
3	Evidence for small molecule binding to the <code>HSP_70</code> domain.	171
4	Evidence for small molecule binding to the <code>PEPCK</code> domain.	172
5	Evidence for small molecule binding to the <code>RrnaAD</code> domain.	173
6	Screenshot of the curation interface. Section: ‘Evidence’	174
7	Screenshot of the curation interface. Index page for ‘Evidence’ section.	175
8	Screenshot of the curation interface. Section: ‘Conflicts’	176
9	Screenshot of the curation interface. Index page for ‘Conflicts’ section.	177
10	Screenshot of the curation interface. Section: ‘Resolved’	178
11	Screenshot of the curation interface. Index page for ‘Resolved’ section	179
12	Schema sections representing the mapping of small molecule binding to Pfam-A domains in <code>chembl_15</code> and upwards	183

List of Tables

2.1	The ten most frequent Pfam domain clans in the catalogue of Pfam-A domains with known small molecule interactions.	43
2.2	Pfam-A domain types projected onto multi-domain proteins.	45
2.3	Detail view of mapping evaluation.	47
2.4	Most frequent Pfam-A domains in multi-domain proteins	58
2.5	Domain architectures with small molecule binding at domain interfaces	59
3.1	Pfam-A domains with insufficient evidence for small molecule binding	86
3.2	Overview of domains that were added manually	87
3.3	Domain configurations in conflicting mappings	91
3.4	Top 15 multi-domain architectures in the ChEMBL database	97
4.1	Activities by organism	112
4.2	Quantile estimates of 1,000 sampled distributions of inter-assay differences.	114
4.3	Target classes represented in the data set	119
4.4	Parameters of fitted models.	124
4.5	Correlation of molecular weight and absolute potency differences for individual pairs of paralogs.	132
4.6	Individual pairs of paralogs in the molecular weight analysis.	132
4.7	Human paralogs with greatest overall potency differences.	135
4.8	Human to rat orthologs with greatest overall potency differences	136
4.9	Critical binding site residue positions in the GPCR family alignment	147
4.10	Critical binding site residue positions in the kinase family alignment	148
4.11	Crystal structures of GPCRs	151
4.12	Overview of deposited model and alignment files.	153

1	Pfam-A domain types projected onto multi-domain proteins	180
2	Paralogous groups	184

Chapter 1

Introduction

In 1900, Paul Ehrlich introduced a theory of ‘receptors’ to explain the process of recognition between cells and toxins (Ehrlich and Morgenroth, 1900, reviewed in Maehle et al., 2002). Over the course of decades, this theory of receptors developed into a central concept for the scientific discipline of pharmacology. The idea that drug action relies on the interaction with highly specific receptors was first proposed by the pharmacologist Alfred Joseph Clark (Clark, 1933) and forms the basis of modern receptor theory. Today we understand that most drug receptors are proteins, and the majority of them belong to one of four functional categories, membrane receptors, ion channels, enzymes or transcription factors (Drews, 2000).

This knowledge is a valuable foundation for the development of new medicines, but it also means that drugs can be used to understand physiological processes by interfering with defined cellular components. The discipline of chemical biology can maybe be best described as making available the small molecule tools for such studies and providing insights into how the modulation of drug receptors is coupled to the physiological and pathological function of cells and whole organisms. In many cases, these tools are not suitable as therapeutic agents, but are nevertheless useful in the drug discovery process as standards of comparison, and increasingly as tools for target validation.

Besides molecular tools or drugs, both chemical biology and drug discovery generate large amounts of data that are related to the biological activity of small molecules. These data are obtained from assays of various formats, and encompass a large number of different read-outs. Growth of publicly available small molecule bioactivity data is driven

mainly by automatic and semi-automatic aggregation of published measurements.

In my thesis, I describe approaches of integrating these data with resources of protein evolutionary relationships. The promise of data integration is that new insights can be gained from previously isolated measurements. This is not only a promise, but also a great challenge: How can we compare measurements obtained in different experiments and what confidence can we place in these comparisons? What kind of questions can be answered using published bioactivity measurements, which are vast in number, but sparse or at best widely clustered in terms of their coverage of ligand and target space? During my time at the European Bioinformatics Institute, I have examined and probed these questions. Small molecule bioactivity measurements from the ChEMBL database and protein domain assignments from the Pfam database have been combined to produce a map of the target and ligand space covered by biochemical assays in the ChEMBL database. Further, integration of homology information from the EnsemblCompara Genetrees resource helped me pursue a phylogenetic analysis of small molecule activity against related protein targets.

1.1 The quest for small molecules as therapeutics and research tools

Small organic molecules form the prominent class of therapeutic drugs. In this context, small molecules are understood to be organic compounds of typically 550 Daltons molecular weight, with lower and upper limits on weight at around 200 Daltons and 1,000 Daltons. Small molecules also occur naturally, for example as hormones, metabolites, and vitamins. However, most of the therapeutic small molecules are of artificial synthetic origin, even though many of them are derived from naturally occurring molecules. Small molecules have a number of advantageous properties that make them desirable candidates for drug discovery. First among these is that many small molecule drugs can reach tissues and organs after oral ingestion. The degree of oral availability varies between small molecules and depends on a number of factors, including their capability to cross the luminal membrane of the gut as well as avoiding enzymatic degradation in the gut and liver. A famous, but rough, approximation of oral availability is compliance with the Lipinski rule-of-five, a set of rules concerning four simple molecular properties ([Lipinski](#)

et al., 2001). Many small molecules can also cross cellular membranes through integral carrier proteins and thus interfere with processes taking place inside the cell (Kell et al., 2011; Kell et al., 2012). Furthermore, small molecules are generally outside the functional remit of the immune system and do not typically provoke immune responses.¹ Crucially, small molecules can engage in highly specific and energetically favourable interactions with proteins, and through these interactions influence physiological and pathological functions on a molecular level. Linus Pauling said about the process underlying these interactions that “the secret of life is molecular recognition; the ability of one molecule to ‘recognize’ another through weak bonding interactions” (Dunn, 2010). The general principles of ligand-receptor interactions have been studied extensively and four main types of interactions that mediate small molecule binding have been identified.² These include electrostatic interactions between opposite charges on a ligand and receptor (Perutz, 1978; Hol et al., 1978), van der Waals attractive forces (Hamaker, 1937) and the formation of hydrogen bonds (Moore and Winmill, 1912; Latimer and Rodebush, 1920). A fourth and sometimes dominant contribution comes from the displacement of water molecules from the interface of the ligand and receptor binding site. This displacement of water molecules is believed to be entropy-driven and is often referred to as the hydrophobic effect (Searle et al., 1992; Dunitz, 1995; Chandler, 2005). All of the aforementioned types of interactions rely on shape complementarity between the ligand and receptor. A common analogy used to convey the high specificity of such interactions is that of a lock (the receptor) and key (the small molecule).³ Figure 1.1 illustrates an exemplary small molecule-protein interaction of biotin (Vitamin B₇) and streptavidin.

Taken together, the properties of small molecules have proven highly valuable in the development of new drugs and to date, the vast majority of approved drugs are small molecules. In the following, I give an overview of historical and current approaches to identify small molecules that can be used as therapeutic agents or research tools.

¹Immune reactions are nevertheless a frequent side effect of the administration of small molecule drugs, see e.g. Pichler, 2003.

²For a review of interactions in published crystal structures of drug-receptor complexes, see Babine and Bender, 1997.

³This analogy is helpful, but it omits the conformational flexibility of both the receptor and small molecule as well as the crucial role of water molecules and metal ions in some of these interactions.

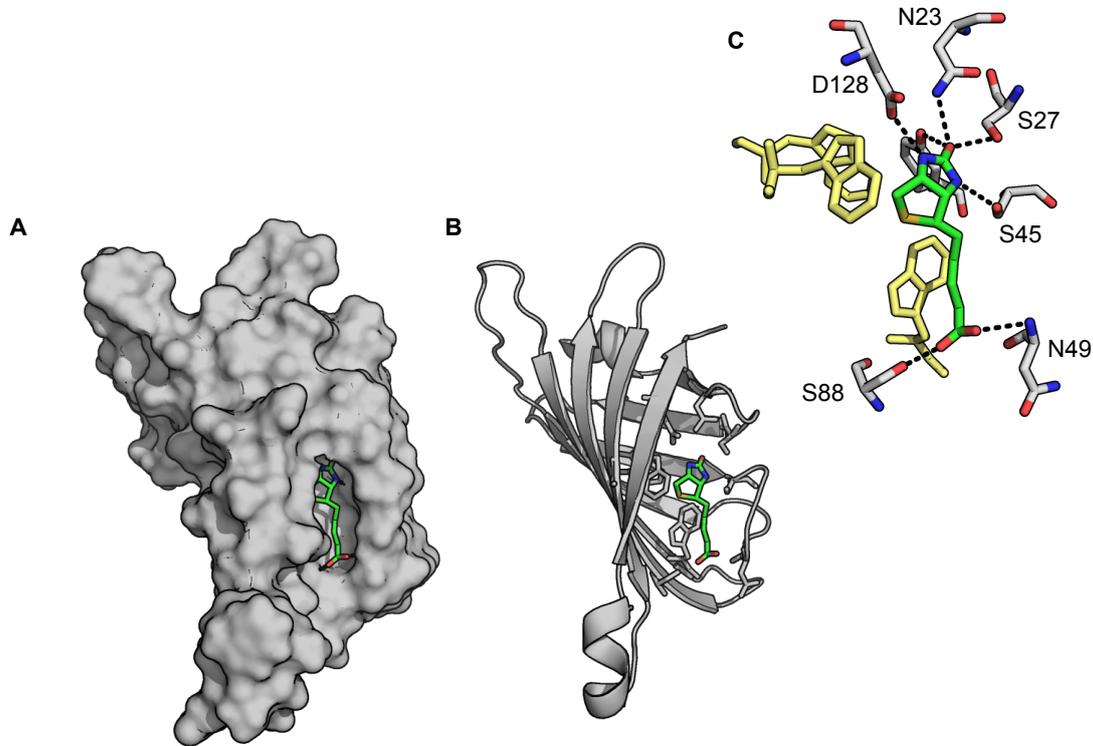


Figure 1.1: Illustration of a small molecule-protein interaction: biotin and streptavidin. Panel A and B show the streptavidin protein in surface and cartoon representation, while biotin (green) and the interacting residues in B are shown in stick representation. Panel A highlights the complimentary shapes of biotin and it's receptor streptavidine. Panel B shows the positioning of the binding site in the barrel formed by two antiparallel β -sheets. Panel C illustrates some of the interactions that contribute to the binding of biotin to streptavidin. Three tryptophan residues (Trp79, Trp92, Trp108, yellow, unlabelled) are engaged in hydrophobic interactions with the uncharged parts of the biotin molecule. Hydrogen bonds are observed between Asn23, Ser27, Tyr43, and the ureido ring oxygen of biotin. The ring nitrogens form one hydrogen bond each with Ser45 and Asp128. The carboxylate oxygens of biotin form two hydrogen bonds, one with the main chain N-H of Asn49 and the other with Ser88. Under physiological conditions, the biotin-streptavidin complex forms homo-tetramers and each biotin molecule undergoes an additional hydrophobic interaction with a tryptophan residue from an adjacent monomer (not shown). The figure was adapted from [Livnah et al., 1993](#), the pdb accession of the associated crystal structure is `1stp`.

1.1.1 Small molecules as therapeutics

The discovery of therapeutic small molecules has been driven by a number of diverse strategies and innovations that are reviewed in this chapter. The outlined strategies are supplemented with examples of drugs that were discovered using a given approach. Literature references for individual examples were in many cases obtained from [Sneader, 2005](#).

Written accounts of the use and preparation of therapeutic drugs reach back as far as to the ancient civilisations of Egypt, Mesopotamia and Greece ([Powell, 1993](#)). The basis on which drugs were used and developed was a set of mainly irrational beliefs, until in the early 19th century the technology of solvent extraction from plant materials provided pure, pharmacologically active compounds ([Sneader, 2005](#)). This kindled the recognition that plant drugs exert their effect through isolated components that were referred to as ‘active principles’. Solvent extraction paved the way for the development of drugs from alkaloids, basic substances that are extracted from plants and form crystal salts with acids. Two of the most famous among the alkaloid drugs are the analgesic morphine, isolated in 1804 by Friedrich Wilhelm Sertürner ([Schmitz, 1985](#)) and quinine, an antipyretic and analgesic drug that was isolated in 1820 by Pierre Joseph Pelletier and Joseph Bienaimé Caventou ([Pelletier and Caventou, 1820](#)). The successes of alkaloid drugs inspired pharmacists and chemists to seek synthetic derivatives of alkaloids with therapeutic value. For example, in an attempt to obtain an analog of quinine, Emil Fischer and his student Ludwig Knorr developed phenazone in 1884, an antipyretic drug that did not exhibit the severe side effects of quinine. Other synthetic derivatives of alkaloids include the morphine derivative heroine, first synthesised in 1897 by Felix Hoffmann ([Sneader, 1998](#)) and the tranquilliser haloperidol, which is a derivative of atropine and was discovered in 1958 by Paul Janssen ([Janssen et al., 1959](#)).

Over the course of the 19th century, the extraction of active compounds from plants and the synthesis of chemical derivatives of these compounds followed by animal testing had emerged as a viable strategy in drug discovery. This strategy is sometimes described as the extractive heuristic ([Nightingale, 2000](#); [Hopkins et al., 2007](#)). Soon, it was extended to extract active compounds from various human tissues and other organisms. Extraction of drugs from the hormone system prove particularly fruitful and yielded compounds

such as adrenaline⁴ and histamine (Dale, 1950). The synthesis of hormone analogues allowed pharmacologists to obtain compounds that interact with hormone receptors, but elicit a response that differs from that of the endogenous hormone. This approach was pioneered by the pharmacologist James Whyte Black, who developed propranolol, an analog of adrenalin that binds the β -adrenergic receptor without activating it (Black and Stephenson, 1962; Prichard and Gillam, 1964). Propranolol reached the market in 1964 and was widely used as an antihypertensive drug. A similar strategy was used by Black and Charon Robin Ganellin in the development of the histamine H₂ receptor antagonist cimetidine, which is an analog of histamine and became a “block-buster” drug for the treatment of stomach ulcers in 1976 (Brimblecombe et al., 2010).

Apart from the hormone system, the extractive heuristic was also applied to mimic metabolites and naturally occurring antibiotic compounds. A derivative of the vitamin folic acid (vitamin B₉), methotrexate, was developed in 1947 for the treatment of acute leukemia (Seeger et al., 1949). Other antimetabolites include modified nucleosides that inhibit the growth of tumours or DNA or to control viral infections. Among the extracted antibiotic compounds are for example chloramphenicol, which was isolated by John Ehrlich and Quentin Bartz from a bacterial culture of *streptomyces venezuelae* (Ehrlich et al., 1947) and erythromycin, isolated by Robert Bunch and James McGuire from a bacterial culture of *streptomyces erythreus* (Bunch and McGuire, 1953).

An entirely different strategy emerged early in the 20th century, the synthetic organic chemistry heuristic (Nightingale, 2000; Hopkins et al., 2007). The advent of synthetic organic chemistry had made available large collections of diverse artificial compounds that could be screened for activity in pharmacological screens. A pioneer of this approach was Paul Ehrlich, who, in 1910, together with Sacachiro Hata, discovered arsenophenyglycine (Salvarsan) from a screening of organic dyes (Ehrlich and Hata, 1910). Salvarsan was prescribed as a potent drug to treat patients with syphilis. Continuing Ehrlich’s work on dyes, Gerhard Domagk screened a large number of azo-dyes for their potential to treat bacterial infections. Eventually, the red sulfonamide dye sulfamidochrysoidine was discovered as a potent antibiotic (Domagk, 1935).

Advances in molecular biology in the 1970s and 1980s provided techniques for the cloning and functional characterisation of proteins and thus for the investigation of individual proteins as potential drug targets. Technological innovations such as automation,

⁴The discovery, isolation and synthesis of adrenalin is reviewed by Aldrich, 1905.

miniaturisation (Silverman et al., 1998) and combinatorial chemistry (Merrifield, 1963; Brenner and Lerner, 1992; Gallop et al., 1994; Fodor et al., 1991) made it possible to screen ever larger numbers of compounds in assays targeting specific proteins. By the early 1990s, systems for high throughput screening (HTS) emerged to turn over libraries in the order of one hundred thousand and more compounds in a matter of days (Pereira and Williams, 2007). Considered a breakthrough in cancer therapy, the lead structure for Imatinib, a drug for the treatment of chronic myeloid leukemia, was discovered by HTS in the mid 1990s (Buchdunger et al., 1996; Deininger and Druker, 2003). More recently, a number of limitations of HTS have been discussed in the literature, including limited diversity of combinatorial libraries (Feher and Schmidt, 2003), high false positive rates (Malo et al., 2010) and vulnerability to reporting artifactual interactions (Rishton, 1997; McGovern et al., 2002; Walters and Namchuk, 2003). A general criticism of target-based approaches is that biochemical inhibition does not reliably translate into therapeutic effects and that phenotypic screens, which assess small molecule response on an organismal or functional systems level, may be more suitable to identify effective small molecules (Swinney and Anthony, 2011).

Underlying the diverse approaches discussed above is a process that encompasses the identification of a candidate or lead structure followed by iterative cycles of modification and retesting, or lead optimisation. While traditional approaches of lead optimisation were guided by medicinal chemistry rules derived from the properties of functional groups, the increasing availability of protein crystal structures by the 1980s had given way to a methodology called structure-based design. Structure-based design relies on structural information from an observed or predicted binding site to instruct the synthesis of a suitable ligand (Greer et al., 1994; Whittle and Blundell, 1994). This methodology most commonly requires computer graphics (Goodsell et al., 1989) and methods for the calculation of electrostatic potentials (Gilson et al., 1988). The discovery of the carbonic anhydrase inhibitor, dorzolamide, a topical treatment for glaucoma, in 1995, was based largely on structure-based methods (Ponticello et al., 1998; Davis et al., 2003).

Over the course of two centuries, small molecule drugs have been established as essential components of modern medicine, provided cures for previously incurable diseases and supported the advance of evidence-based medicine. While valuable as therapeutics, small molecules also play a significant role as research tools. The next section reviews the development of such tools.

1.1.2 Small molecules as research tools

The use of small molecules to induce specific conditions in a biological system under study has a long history and early examples include the use of atropine and acetylcholine to investigate chemical transmission of nerve signals (Barger and Dale, 1910). However, it was Linus Pauling's theory of molecular recognition (Pauling and Delbrück, 1940; Pauling, 1974) combined with progress of biochemical methods for protein purification and quantitation (Cuatrecasas et al., 1968; Lowry et al., 1951) that promoted the use of small molecules to interfere with the function of specific proteins and to thus dissect and study biological processes. For example, colchicine was used to elucidate the role played by α - and β -tubulins in the homeostasis of the cytoskeleton (Weisenberg et al., 1968) and the pufferfish toxin tetrodotoxin as a tool to study neurophysiology⁵ (Narahashi et al., 1964).

Since then, many more small molecules were developed as research tools and today, hundreds of small molecule tools can be readily obtained from commercial vendors. The use of small molecules to interrogate biological processes is often called chemical genetics, in analogy to molecular genetics approaches, which use genetic perturbations to achieve similar ends (Schreiber, 1998; Stockwell, 2000). Molecular genetics approaches can be divided into forward- and reverse genetics approaches. Forward genetics seeks to establish the genetic cause of an observed phenotype. In analogy, forward chemical genetics seek to establish the mechanism by which a given small molecule induces an observed phenotype. Reverse genetics on the other hand makes use of molecular cloning techniques or crosses of mutant strains to introduce a targeted genetic perturbation and then analyse the effect of this perturbation on the resulting phenotype. This approach can deliver powerful insights into the molecular biology of cells, mechanisms of disease and development. Reverse chemical genetics introduces targeted perturbations using small molecules with known protein targets. One important advantage of traditional genetic approaches over chemical genetic approaches is their generality. Regardless of which gene is being investigated, mutations can be introduced and located using the same procedures over and over and again. Small molecules that interfere with gene function on the other hand are only available for a small fraction of all proteins encoded

⁵Tetrodotoxin acts by blocking the pores in sodium channels. This prevents the formation of action potentials, a useful condition for neurophysiological studies (Kandel et al., 2000).

by the human genome. Further, the specificity of small molecules is often limited and off-target effects have to be accounted for. Despite these shortcomings, small molecules are desirable research tools. Proteins that are targeted by small molecules are inactivated, but still present in the cell, keeping it closer to its physiological state compared to genetic knock-outs. Further, small molecules afford researchers the possibility to induce reversible and gradual perturbations of protein function. Chemical genetics approaches are also useful to cross-validate findings from traditional genetics approaches.

Successes of chemical genetic approaches, such as the discovery of FK506, a suppressor of calcineurin and immunosuppressant (Schreiber, 1991), or the characterisation of the bile acid receptor FXR using chemical probes (Kliewer et al., 1999; Downes et al., 2003) inspired calls for a generalised chemical genetics methodology. To achieve this, it was proposed that chemical probes should be identified systematically for all proteins (Zanders et al., 2002; Shokat and Velleca, 2002; Schreiber, 2003; Austin, 2003). Around the same time, a big NIH-funded screening project for small molecule research tools, called Molecular Libraries Initiative (MLI), was brought underway (Austin et al., 2004). It involved multiple high throughput screening centres and organic chemistry divisions with the aim of screening hundreds of thousands of compounds against a wide range of targets and optimising initial hits towards a given target. The chemical probes identified in this project were made available to the scientific community. The data generated in these screens was also made available, through the PubChem Bioassay repository (Wang et al., 2012a, <http://www.ncbi.nlm.nih.gov/pcassay>). The MLI did not meet the high expectations it set out with (Kaiser, 2008) and during the years following the initiation of this project, it became clear that HTS alone is not a recipe to obtain small molecule tools for a genome-scale range of targets (Lazo, 2006). Increasingly, a focus was set on HTS reporting standards (Inglese et al., 2007) and selection criteria for small molecule research tools from HTS screening hits. A widely accepted set of rules has been stated by Stephen V Frye (Frye, 2010):

- Molecular profiling. Sufficient in vitro potency and selectivity data to confidently associate its in vitro profile to its cellular or in vivo profile.
- Mechanism of action. Activity in a cell-based or cell-free assay influences a physiologic function of the target in a dose-dependent manner.
- Identity of the active species. Has sufficient chemical and physical property data

to interpret results as due to its intact structure or a well-characterized derivative.

- Proven utility as a probe. Cellular activity data available to confidently address at least one hypothesis about the role of the molecular target in a cell's response to its environment.
- Availability. Is readily available to the academic community with no restrictions on use.

Other experts have stated similar criteria (Workman and Collins, 2010; Cohen, 2010). In an alternative approach, a panel of experts was asked to evaluate a set of chemical probes produced by the MLI according to their own criteria and a consensus was determined quantitatively (Oprea et al., 2009). Intellectual property and access issues have also been identified as important factors for the success of small molecule research tools, especially for use in target validation in the early stages of drug discovery (Edwards et al., 2009).

An opportunity to improve functional annotations for small molecule research tools comes from structural biology. Efforts to structurally characterise all proteins encoded by the human genome have begun simultaneously with the MLI (Sali, 1998; Kuhn et al., 2002) and are beginning to make an impact both in term of the number of solved structures as well as the diversity of these structures in terms of covered protein families (Chandonia and Brenner, 2006; Marsden et al., 2007; Gileadi et al., 2007). Crystal structures of proteins, and especially protein inhibitor complexes provide insights into potential and observed binding sites for small molecules (Fedorov et al., 2007b; Marsden and Knapp, 2008). This information can be interpreted to understand and predict patterns of selectivity between members of protein families such as kinases (Fedorov et al., 2007a) and GPCRs (Venkatakrisnan et al., 2013). Protein structures have also fundamentally aided the design of artificial modifications to kinases that renders them sensitive to certain inhibitors (Bishop et al., 2000b; Bishop et al., 2000a; Eblen et al., 2003). Using this approach, small molecule inhibitors can be used to effect extremely specific perturbations that can be spatially and temporally controlled. Similar systems have been developed for GPCRs (Scearce-Levie et al., 2002) and the estrogen receptor (Tedesco et al., 2001).

The selectivity of chemical probes in comparison to genetic perturbations is often limited. This need not necessarily be a disadvantage, the broad-spectrum kinase inhibitor

staurosporine for example has an excellent track record as a research tool (Tamaoki et al., 1986; Tamaoki, 1991). However, to use small molecules adequately, it is absolutely vital to have an understanding of which proteins are perturbed by a given small molecule. This may explain the success of staurosporine as a chemical tool, as its excessive promiscuity is well-known and understood (Karaman et al., 2008; Tanramluk et al., 2009). Selectivity profiles of other small molecule research tools are more ambiguous and often not well appreciated by those using them. Kinases, with their highly conserved ATP-binding site, are particularly challenging in this respect. In early efforts in the group of Phillip Cohen, the selectivity of kinase inhibitors was probed by screens across a panel of about 30 kinases (Davies et al., 2000; Bain et al., 2003; Bain et al., 2007; Cohen, 2010). Efforts at much larger scale have followed that tested small molecules across a large fraction of the roughly 500 kinases⁶ in the human proteome (Fabian et al., 2005; Fedorov et al., 2007a; Anastassiadis et al., 2011; Posy et al., 2011; Metz et al., 2011; Davis et al., 2011; Gao et al., 2013). A consensus forming from these experiments is that many compounds exhibit high levels of promiscuity between related kinases, but some specific compounds can be identified for many of the 500 kinases in the human proteome (Uitdehaag et al., 2012).

Beyond profiling on a protein target level, small molecules are also profiled for their activity on different types of cell lines, for example in the anticancer drug screen of the National Cancer Institute in the US (Shoemaker, 2006) which screens approximately 3,000 compounds per year for activity across a panel of 60 cell lines. An even larger number of cell lines are being evaluated in the connectivity map screening at Broad Institute (Lamb et al., 2006) and the genomics of drug sensitivity screen at the Sanger Institute (Garnett et al., 2012). These screens are intended to assess the effect of genetic differences between cell lines on the way in which cell lines respond to small molecules. In reverse, these screens also contain information that can be used for the characterisation of small molecules involved in these screens.

To summarise, over the last decade, efforts to obtain and characterise small molecules as research tools have scaled up and increasingly rely on high-throughput technologies. The specificity and universal applicability of traditional genetic methods have not been matched, but small molecules have nevertheless been established as important research tools in molecular biology and for the validation of targets in the early stages of drug

⁶See Manning et al., 2002

discovery. Small molecule related experiments generate large amounts of data that hold value on their own. In the following section I review approaches that integrate and exploit data generated in the development of small molecule research tools and drugs.

1.2 Data integration in chemical biology and drug discovery

Chemical biology and drug discovery naturally examine biological processes on multiple levels of molecular and organismal abstraction and hence rely on data generated from a mosaic of different experiments. This requires an integrative approach beyond isolated experiments or for that matter, isolated data repositories. For example, knowledge of a binding constant of a small molecule to a protein is more useful if its role in a disease process is also known; in other words, a piece of biological data is more meaningful if a context is provided for it. This context can reveal links and relationships from previously isolated data and promote new biological theories and standards (Karp, 1996). The need for integration of data across different resources is also illustrated by the growing number of biologically relevant data repositories: in 1990, there were about one hundred biological databases (Keen et al., 1992). In 2013, the database issue of *Nucleic Acids Research* already listed 1,512 relevant repositories (Fernández-Suárez and Galperin, 2013). By combining data from these sources, data integration supports the logical interpretation and association of life science knowledge (Neumann and Thomas, 2002). More specifically to drug discovery, the application of data integration promises to help understand drug action on multiple scales, support the generation of new hypotheses, for example in target validation or mode-of-action studies, and avoid duplication of experiments (Searls, 2005; Loging et al., 2007).

Formally, data integration seeks to establish links between independent data repositories. Most frequently, data in such repositories is stored in relational databases (Codd, 1970), and in such cases the challenge for data integration is to establish links that are beyond the scope of individual relational schemas. In computer science, data integration has been an area of research for almost four decades and various technical solutions to this problem have been proposed (Ziegler and Dittrich, 2004; Louie et al., 2007). Solutions that have been applied in a biomedical context include link integration, view

integration, and data warehousing (Stein, 2003; Schneider and Jimenez, 2012). Link integration builds on the world wide web technology and refers the user between linked documents. It requires that participating data sources maintain links to other repositories and frequently check their validity. It also enforces a sequential approach to data integration in which links from one document direct to the next document and so forth. The method is nevertheless successful. SRS for example is a widely used query system that uses link integration (Zdobnov et al., 2002). View integration provides a query interface to multiple, federated datasources. The interface acts as a mediator between the databases and the user input. The cross-database resources Biomart (Smedley et al., 2009) and the distributed annotation system (DAS) (Dowell et al., 2001) are examples of view-based integration. Data warehousing seeks to integrate all relevant data sources into one consistent database schema. A critical view of data warehousing proposes that integration of drug discovery data into monolithic, fully linked systems are unsustainable and should be avoided in favour of more flexible approaches that connect multiple knowledge systems and provide mechanisms for expert reasoning and curation (Slater et al., 2008). One such approach, relying on ‘cubes’ of data that can be reformatted, for example to accommodate an additional dimension, has been proposed by Millard et al., 2011. It has also been suggested that linking of data sources by preferential attachment⁷ can improve the sustainability of data warehousing (Searls, 2003a). A relevant example of a warehousing integration approach is the STITCH database, which aggregates small molecule protein interaction data from multiple other data sources (Kuhn et al., 2012).

The use of ontologies can greatly facilitate data integration efforts for all of the approaches presented above. Ontologies provide controlled vocabulary to capture the concepts that are relevant to a scientific discipline and the terms in such a vocabulary are arranged in hierarchical relationships (Stevens et al., 2000). If two or more data sources can be represented by the same ontology, they can be integrated using links established through the hierarchy of the ontology’s controlled vocabulary. One example in this context is the BioAssay ontology (BAO), which enables cross-analysis of diverse high-throughput screening data sets (Schürer et al., 2011).

Data integration is also facilitated by the use of reporting standards for particular types of data. For example, the widely used integration platform for micro-array experiments,

⁷An integration approach using this strategy would result in a small number of densely connected data sources and many more that only have sparse links to other data sources.

ArrayExpress (Rustici et al., 2013), requires submitted experimental data to contain a minimal set of annotations defined by the MIAME standard (Brazma et al., 2001). Similarly, guidelines for submitting data of bioactive entities, including small molecules, have been introduced under the MIABE standard in 2011 (Orchard et al., 2011).

The field of chemogenomics, which is of high relevance to both drug discovery and chemical biology, relies heavily on the integration of biological and chemical databases (Bredel and Jacoby, 2004; Oprea and Tropsha, 2006). The chemogenomics approach seeks to explore the specificity of small molecule binding within protein families and makes use of both protein- and small molecule similarity metrics to predict patterns of specificity where no data is available (Harris and Stevens, 2006). Early work introduced the structure-activity relationship homology concept, which proposes that patterns of susceptibility to small molecule perturbation should, to some degree, be conserved between one given protein and its relatives in the same protein family (Frye, 1999). Extensive studies of the selectivity patterns of kinase inhibitors followed. In 2005, Vieth and colleagues published a study that explored ligand profiles of therapeutically relevant kinases and found that kinases with $\geq 60\%$ sequence identity are most likely to be inhibited by the same classes of compounds (Vieth et al., 2005). In 2007, Keiser and colleagues related 289 proteins by sets of small molecules reported to bind them (Keiser et al., 2007). Later, this approach was developed into a fully implemented method to predict small molecule protein interactions (Keiser et al., 2009). A study published in 2010, described the organisation of 102 class-A GPCRs according to sets of associated ligands, thus providing an alternative to the sequence-based organisation (Horst et al., 2010). In 2013, this approach was extended to 146 GPCRs and also evaluated overlaps with 485 non-GPCR targets (Lin et al., 2013).

Data integration studies that go beyond the chemogenomics approach have examined links between chemical-genetic and genetic interaction data and used these insights successfully to identify components of cellular pathways (Parsons et al., 2004). In a study that integrated chemical data and side effect information from package inserts, Campillos, Kuhn and colleagues have established a creative method to identify the molecular targets of marketed drugs (Campillos et al., 2008). Also with a perspective on side effects, Tatonetti and colleagues have used patient medical records of adverse side effects to predict unknown effects and interactions of therapeutic drugs (Tatonetti et al., 2012).

In a 2003 review, it was proposed that drug discovery can benefit from comprehensive analyses of the evolutionary history of proteins that are investigated as drug targets. This approach was referred to as pharmacophylogenomics (Searls, 2003b). Pharmacophylogenomics is set apart from chemogenomics by its focus on evolutionary relationships, rather than binding site similarity. In an extension of this approach, I have sought to integrate protein evolutionary relationship data with bioactivity data. My work was motivated by the prospect of learning about small molecule binding using evolutionary parameters, such as the distance from a last common ancestor or the type of event that lead to a split in the lineage. These parameters are available in increasing quantity and detail through genome sequencing projects of diverse species (Pagani et al., 2012). In the following sections, I give an overview of the data sources I have used to achieve this.

1.3 Bioactivity data

It was illustrated in the previous two sections that the development of small molecules as either therapeutics or research tools requires the testing of large numbers of small molecules. The first episode of testing in a discovery program is often a screen of some sort to identify small molecules that exhibit a desired activity. In subsequent steps, these molecules are validated, for example in assays that measure activity using a different output as well as in assays that measure off-target effects. Optimisation of a small molecule entails the synthesis and testing of derivatives. Thus, a single discovery program can generate bioactivity data from hundreds and thousands of measurements. As a byproduct of the discovery of drugs and research tools, this data is valuable in itself. It helps researchers avoid duplication of efforts and learn from the successes and failures of previous strategies. Examination of such data can also help answer questions beyond the scope of individual discovery programs. The availability of such data in the public domain has much improved over the last years and now a number of repositories exist for small molecule bioactivity data. The PubChem Bioassay resource hosts assay data generated in the Molecular Libraries program introduced in the previous section (Wang et al., 2012a). BindingDB is an academic project that focuses on bioactivity data obtained in binding assays and includes data from the scientific literature, ChEMBL (see below) and PubChem BioAssay (Liu et al., 2007). The PDSP K_i database is a data warehouse that provides inhibition constants for measurements of small molecule

binding to protein targets. These measurements derive either from the literature, or were generated within the Psychoactive Drug Screening Program (PDSP) at the University of North Carolina Chapel Hill (Roth et al., 2000). ChEMBL is a database of small molecule bioactivity data that is extracted from the scientific literature and, increasingly, from direct submissions of larger datasets generated in the pharmaceutical industry and other research organisations (Gaulton et al., 2012). The scope of journals that routinely serve as sources for data extraction into the ChEMBL database lies within the domain of medicinal chemistry. The standardised assay formats and reporting techniques established in this field facilitate both extraction and analysis of the data. As a trade-off to standardisation, the contribution of innovative assay types to the overall make-up of the data is relatively small. With regard to the composition of targets in cutting-edge drug discovery pipelines this results in a bias towards proteins that are amenable to traditional assay techniques, in particular kinases, GPCRs and proteases. This bias should be noted when interpreting outcomes of data mining studies using ChEMBL.

In the following, I introduce in more detail the types of bioactivity data and assay formats that are prevalent in these data resources, with a focus on the ChEMBL database, as this is the resource that provided bioactivity data used throughout my research. Bioactivity data hosted within the ChEMBL database is derived from a large number of diverse assay formats. A query for assay descriptions in the ChEMBL database (`chembl_15`) returns 578,979 unique terms. While many of these are variations of similar and sometimes identical formats, the diversity of assays is nevertheless staggering. One way of ordering this data is by the level of scale at which bioactivity is measured. This can be on the level of molecular interactions, cells, tissues and whole organisms. Assays on a molecular level, in the remainder of this thesis referred to as ‘biochemical’ assays, deliver outputs that are mostly independent of ‘biological’ factors such as the distribution of receptors in a tissue or the expression levels of proteins in different cell types. Comparisons between measured outputs from biochemical assays are therefore more reliable than comparisons of higher-level assays. Biochemical assays break down further into binding and functional assays. Binding assays measure the interaction or compounding of a small molecule and its protein target directly, while functional assays measure this event as a function of some downstream effect. The term used to describe the strength of binding is affinity, while the amplitude of downstream effects elicited by a small molecule are described as efficacy.

In the context of bioactivity measurements, affinity is most commonly expressed as a dissociation constant K_d , which derives from the Langmuir adsorption isotherm and can be described as

$$K_d = \frac{[L][R]}{[LR]}, \quad (1.1)$$

where $[L]$ is the concentration of free ligand, $[R]$ the concentration of unbound receptor and $[LR]$ the concentration of the ligand-receptor complex. The affinity of a small molecule for a given target is usually determined in saturation experiments. Saturation experiments measure the occupancy of receptor sites as a function of increasing concentrations of a small molecule. Receptor sites can be located on isolated proteins, membrane preparations or membrane proteins on whole cells (Bylund and Toews, 2011). Assay formats that are suitable for saturation experiments include radioligand filter binding assays (Bruns et al., 1983), scintillation proximity assays (SPA) (Alouani, 2000), as well as fluorescence polarisation assays (Banks and Harvey, 2002). Saturation experiments can also be carried out using surface plasmon resonance setups (Haes and Van Duyne, 2002; Wang et al., 2005).

Efficacy can be expressed in a range of activity types. Most frequently used within the ChEMBL database are IC50, K_i , %(inhibition) and %(potency). The latter two activity types measure the effect of a small molecule in relation to some standard of comparison, which could be either the activity of another compound or some sort of baseline activity, for example the substrate turnover of an enzyme. The activity type IC50 is determined as the concentration of small molecule at which the signal, of whatever output is being evaluated, is reduced to 50%. IC50 values can be determined in dose-response studies, where the effect of the administration of a given compound is measured in intervals of increasing concentrations. Normally, a sigmoidal curve is fitted to the measured values and the IC50 is determined as the concentration at which the slope of the curve is at a maximum. Basically, IC50 depends on both the concentration of protein and substrate⁸, while the value for K_i is an absolute constant. The K_i can be determined from an IC50 value and the assay-specific parameters through the Cheng-Prusoff equation (Cheng and Prusoff, 1973).

Efficacy can be measured using displacement binding assays, which monitor the

⁸Substrate in this context denotes additional molecules that constitute some form of competition with the small molecule that is being evaluated.

displacement of a traceable substrate, in response to increasing concentration of a given small molecule. All methods mentioned earlier for saturation experiments are also suitable formats for displacement assays. Other formats measure efficacy further downstream of the binding event. These include measuring the turnover of a fluorescent substrate or radiolabeled substrate (Windh and Manning, 2002). Assays that measure efficacy in a cellular context include those that measure the levels of second messengers such as Inositol-3-phosphate (IP₃) or Ca²⁺ in response to a stimulus either using fluorescent chelator dyes (Chambers et al., 2003) or the aequorin system (Le Poul et al., 2002). Small molecule perturbations can also lead to increased or decreased transcription of specific genes; these changes can be exploited to measure efficacy using reporter gene formats, such as β -galactosidase (Jain and Magrath, 1991), chloramphenicol acetyltransferase (Gorman et al., 1982), firefly luciferase (Gould and Subramani, 1988), and green fluorescent protein (GFP) (Zolotukhin et al., 1996).

When assaying the biological activity of small molecules, it is necessary to account for a degree of ambiguity concerning the identity of the molecular target of any given ligand. In biochemical assays, the level of certainty with which a target can be identified depends on the procedure used to obtain pure protein. Pure protein is often obtained through expression systems and subsequent purification using centrifugation, chromatography or electrophoresis. Frequently used expression systems are based on bacterial and yeast cultures or derived from insect, plant or mammalian cell lines (Baneyx, 1999; Cereghino and Cregg, 2000; Kost et al., 2005). The use of expression systems often implies that post-translational modifications do not correspond to the physiological state. Protein folding can also be affected in expression system that lack chaperone proteins specific to the target organism. In many cases, expression constructs represent only a truncated version of the original protein, containing the part that is assumed to be relevant to small molecule binding. Ambiguity around the identity of a protein target increases further when bioactivity is measured as a downstream signalling response in a cellular context. Measured endpoints from cell-based assays integrate the response to small molecule perturbation across a system of interacting components and thus obfuscate the individual contributions of on-target effects versus off-target effects. Assay formats that assess small molecule bioactivity on the levels of cell populations, tissues and whole organisms are typically referred to as phenotypic screens. In such screens, the measured output is an integrated response of networked interactions of cellular components as

well as interactions between cells, tissues or whole organisms. The mechanism through which a small molecule elicits a measured response in phenotypic screens is thus not well defined.

For the purposes of my thesis work it was desirable to minimise the ambiguity around the identity of a protein target in any given assay. Therefore, I used only information from assays that measured efficacy immediately downstream of the target protein. Within the ChEMBL database, such assays carry a flag B, for binding. Queries to the database can be further restricted using the confidence score that is assigned manually by curators. This score evaluates the relationship between the assigned identifier and the actual identity of the target. For example, an identifier could directly map through to the actual target (confidence score 9) or to a homologue of the target (confidence score 8). Queries made in the context of this thesis are presented in the methods section for each chapter and generally are restricted to bioactivities of type B, with target identifiers mapping directly to the actual protein used in the assay. Moving on from the small molecule bioactivity aspect of my work, in the following I present an introduction to protein evolutionary relationships.

1.4 Protein evolutionary relationships

In 1859, Charles Darwin published his seminal theory of the origin of species and shaped our perception of evolution as a process of variation and natural selection of fit individuals in a population (Darwin, 1859). On a molecular genetic level, variation is present in the form of sequence mutations. Until the 1960s, the predominant view was that genetic differences between species are based on mutations that improve fitness; however, this view was challenged by the increasing availability of genomic sequence data, which showed that synonymous mutations are much more frequent than mutations that change the amino acid composition of a sequence. In 1968, Motoo Kimura proposed that most of the mutations observed between species have no impact on fitness, but rather follow a stochastic process (Kimura, 1968). The prevailing view today is that mutations between species are fixed through both natural selection and neutral mutations, but there is still debate about the relative contributions (Bromham and Penny, 2003). In the context of this thesis, it is important to note that proteins, over the course of generations, undergo sequence changes. Sequence differences between two related proteins thus reflect the

generational distance between these proteins and their last common ancestor⁹. In other words, sequence dis-similarity indicates evolutionary divergence of related proteins. In the following sections, I introduce two main concepts that were developed on this principle and are used to describe evolution on a molecular level, protein homology, and protein domains.

1.4.1 Protein Homology

According to the central dogma of molecular biology, proteins are the product of transcription into messenger RNA and subsequently translation into polypeptides. Genes coding for functional proteins in any organism do not arise spontaneously, but rather develop their function and patterns of expression in a long succession of subtle or sometimes dramatic changes across many many cycles of replication (Fitch, 1970). Thus, each gene has a line of ancestors and, through ancestral duplication, homologous relatives evolving in parallel lineages. Gene duplication is driven by a number of mechanisms, for example through erroneous recombination, transposable element insertion or whole genome duplication. Genes evolving independently within the genome of one species are called paralogs if they can be traced back to a common ancestor. Another mechanism by which the evolution of a gene can split into independent lineages is speciation. Genes that derive from a common ancestor, but evolve independently in two separate species are called orthologs. Figure 1.2 provides a graphical summary of these relationships. It is a longstanding hypothesis that the formation of functionally redundant paralogs through duplication within a genome can catalyse the evolution of new functionality (neofunctionalisation) or the partitioning of the the original function (sub-functionalisation) (Hughes, 1994; Lynch, 2000; Hanada et al., 2009). Orthologs on the other hand continue to share similar function after speciation according to this hypothesis, also known as the ‘ortholog conjecture’ (Conant and Wolfe, 2008; Studer and Robinson-Rechavi, 2009).

A famous study on functional conservation between orthologous proteins was carried out by Max Perutz who studied the structural and functional impact of mutations of hemoglobin in different classes of vertebrates (Perutz, 1983). In this thesis, I examine

⁹Emiel Zuckerkandl and Linus Pauling predicted this relationship (Zuckerkandl and Pauling, 1965) and Kimura incorporated it into his theory of neutral evolution (Kimura and Ohta, 1971), coining the term of a molecular clock. It was later shown by Dickerson that the ‘speed’ of this clock differs between species. This promoted the model of a ‘relaxed’ molecular clock (Dickerson, 1971; Sanderson, 1997).

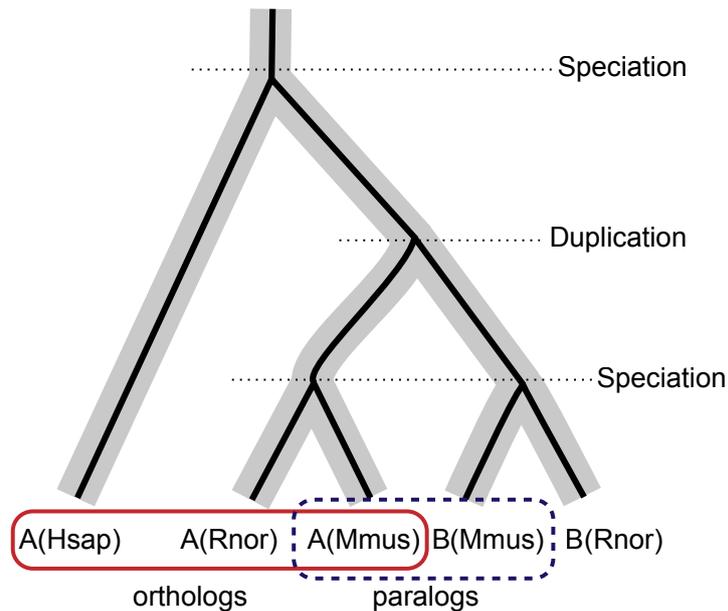


Figure 1.2: Schematic diagram of homologous relationships. This figure illustrates a hypothetical phylogenetic tree delineating the evolution of an exemplary gene A. Branchpoints in this tree represent events where gene A is duplicated either through speciation or within a genome. The first speciation event separates the evolution of A into a human lineage and a lineage representing the precursor of mice and rat. In this lineage, A is duplicated within the genome, leading to the rise of a paralog B. In this diagram, A(hsap), A(Rnor) and A(Mmus) are orthologs. In addition, A(hsap), B(Rnor) and B(Mmus) are also orthologs, as are B(Rnor) and B(Mmus). A(Rnor) and B(Rnor) are paralogs, as are A(Mmus) and B(Mmus).

functional conservation between orthologs and paralogs in terms of susceptibility to small molecule binding.

1.4.1.1 Computational approaches to assign protein homologous relationships

The bioinformatics task of assigning orthologous and paralogous relationships between genomic sequences is an important and challenging problem. The enormous growth of genomic sequencing data over the last decade requires fully automated and efficient methods. At the same time, arrangements where speciation precedes duplication, leaving for example a single gene in one species as an ortholog to two genes in another species

make such assignments more complicated.

Early implementations of ortholog mapping approaches relied on manual curation for this task, as seen in the COG database of orthologous clusters in prokaryote genes and later addition of eukaryote genes (Tatusov et al., 2001; Tatusov et al., 2003, <http://www.ncbi.nlm.nih.gov/COG/>). The COG ortholog assignments are still widely used, but the manual approach to assigning orthologs does not scale up with the exponential growth of genome sequencing data. Automated solutions that build orthologous relationships from reciprocal pairwise sequence alignments of the protein coding sequences in two genomes were developed towards the end of the millenium (Overbeek et al., 1999). These methods are now referred to as bidirectional best-hit approaches (BBH) and are still frequently used (Overbeek et al., 1999; Wolf and Koonin, 2012). The Inparanoid resource of orthologous groups (Remm et al., 2001, <http://inparanoid.sbc.su.se/cgi-bin/index.cgi>) aligns two proteomes using the BLASTP algorithm (Altschul et al., 1990) and uses resulting BBHs as seeds for orthologous groups. These groups are populated with all sequences most similar to the original seed sequences. The OrthoMCL resource determines BBHs from all-versus-all alignments of any given number of genomes and obtains orthologous groups using a Markov chain algorithm (Dongen, 2000; Li et al., 2003, <http://orthomcl.org/orthomcl/>). A different approach is taken by the EnsemblCompara Genetrees resource (ECG, Vilella et al., 2009, http://www.ensembl.org/info/genome/compara/homology_method.html). The algorithm relies on multiple sequence alignments (MSAs) to infer a tree structure of protein evolutionary relationships. In the ECG implementation, BLASTP is used for all-versus-all alignments and clusters are formed as in previous approaches. For each cluster, M-coffee (Wallace et al., 2006) is used to obtain a consensus MSA. The tree-building algorithm TreeBest (<http://treesoft.sourceforge.net/treebest.shtml>) is used to infer a gene-tree structure from the MSA. The algorithm also incorporates a preconfigured species tree and thus reconciles gene- and organism-level trees (Dufayard et al., 2005).

The resources described above are approaches to assign orthologous relationships between proteins of two or more species. However, as a byproduct of the clustering or tree inference procedures, which associate related sequences between and within genomes, they also provide mappings of paralogous relationships. In my thesis project, I worked with the ECG resource. Through discussions with Albert Vilella I learnt that paralogous mappings in ECG are limited to close relatives. For my purposes, this was an advantage

when comparing orthologs and paralogs as described in chapter 4 (see also section 4.2.3). In a benchmarking study, ECG was found to deliver decent performance compared to other approaches (Altenhoff and Dessimoz, 2009).

1.4.2 Protein domains

Protein domains are independent folding units that form the basic evolutionary and architectural ‘building blocks’ of proteins (Blake et al., 1967; Wetlaufer, 1973). Within a protein sequence, domains are subsets of consecutive residues that stabilise each other in a defined arrangement of secondary structure elements. In most cases this is achieved through the formation of a hydrophobic core as well as electrostatic interactions on the domain surface (Chothia, 1984; Garbuzynskiy et al., 2013). In contrast to simple protein folding units, protein domains are by definition structurally self-sufficient, meaning they would retain their three dimensional geometry and often their function if cleaved from the rest of the protein backbone (Levitt and Chothia, 1976). While there can be large sequence differences between members of a domain family, the fold of the peptide backbone is generally conserved (Chothia, 1984), even though exceptional cases of homologous proteins with differing folds have been identified and discussed (Grishin, 2001).

Most proteins in eukaryotes have two or more domains (Apic et al., 2001; Chothia et al., 2003) and it has been observed that increasing complexity on an organismal level also confers an increase of more complex multi-domain architectures (Koonin et al., 2000). In some eukaryotic proteins, the underlying intron-exon structure of the gene reflects the pattern of domain segmentation, suggesting that proteins can acquire new domains through intronic recombination (Patthy, 1996). However, this mechanism accounts for only a fraction of protein architectures, while in most cases, domains are gained through non-allelic homologous recombination of nearby genes (Buljan et al., 2010).

A number of evolutionary constraints that favour multi-domain architectures have been proposed. These include the rate of protein folding and protection from mis-folding (Han et al., 2007; Garbuzynskiy et al., 2013), increased efficiency of cellular processes through ‘forced’ spatial proximity of inter-dependent functional modules (Burns et al., 1990; Marcotte et al., 1999; Enright et al., 1999) and the ‘syntactical’ re-use of existing domains to serve new functions in reconfigured domain assemblies (Vogel et al., 2004).

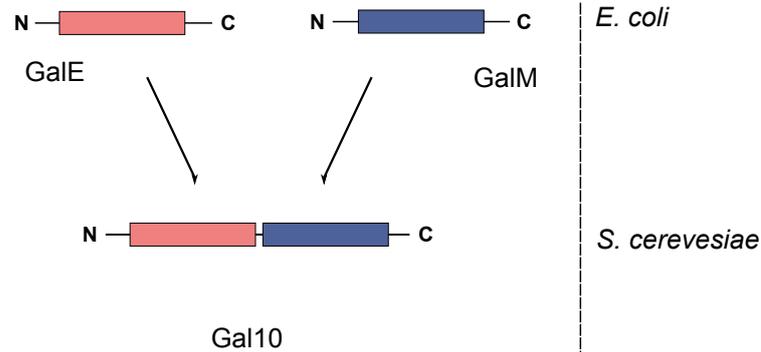


Figure 1.3: Schematic illustration of domain fusion. GalE, consisting of a Epimerase and Epimerase_Csub Pfam-A domain (red, for simplification shown as one domain), and GalM, consisting of a Aldose_epim (blue) domain are individual proteins in *E.coli*. In *S. cerevesiae*, these two genes are fused to form Gal10, a protein consisting of a Epimerase, Epimerase_Csub (red, shown as one), and Aldose_epim (blue) domain. This figure is adapted from [Enright et al., 1999](#).

Figure 1.3 illustrates a simple example of a domain fusion of GalE and GalM to form the bifunctional Gal10 protein in *S. cerevesiae*.

1.4.2.1 Frameworks to detect and represent protein domains

A number of frameworks exist for the detection and representation of domain architectures in protein sequences. In the following I will give an overview of three widely used frameworks, including the Pfam collection of protein families used in the mapping of small molecule binding to protein domains presented in this chapter. The three frameworks, SCOP, CATH, and Pfam can be grouped into structure- and sequence-based approaches. SCOP ([Murzin et al., 1995](#)) and CATH ([Orengo et al., 1997](#)) are prominent implementations that define protein architecture based on hierarchical definitions of three-dimensional structural domains, while the remaining frameworks rely on sequence-based methods for the detection of protein domains.

In SCOP, domains in proteins with known structure are classified according to a hierarchy that includes a species level representing individual domains, a protein level that groups orthologous domains from different species, a family level that establishes a grouping based on sequence similarity, a superfamily level that aggregates these groups further based on shared structural and functional characteristics, and finally a fold level,

that arranges superfamilies based on shared order and topology of secondary structure elements. All domain models in SCOP are assigned in a manual curation process, but since 2008 this process is supplemented by sequence-based clustering methods (Andreeva et al., 2008). When first published in 1995, SCOP listed 498 domain families, 366 superfamilies and 274 folds (Murzin et al., 1995). In its latest release in the year 2009 (version 1.75), SCOP listed 3,902 domain families, 1,195 superfamilies and 1,962 folds. Major updates to the SCOP database have introduced stable domain identifiers, improved definitions of family and superfamily models, and addressed challenges imposed by the rapid growth of structural data in recent years (Lo Conte et al., 2002; Andreeva et al., 2004; Andreeva et al., 2008). In an extension to the manual approach of assigning SCOP classifications, SCOPe is a framework for automated curation that achieves higher coverage of structures in the protein data bank (PDB, Berman et al., 2007) compared to SCOP (Fox et al., 2014).

CATH is another structure-based framework that uses a hierarchy of structural abstraction levels for the classification of proteins with known structure. Unlike SCOP, which relies on manual curation, the classification of proteins with known structure in CATH is carried out using a semi-automated procedure that operates on atomic coordinates obtained from entries in the PDB. (Orengo et al., 1997). The hierarchy in CATH is defined by homologous superfamilies at the most distributed level, where domains are aggregated by sequence, much like the family level in SCOP. In the next higher order level, termed topology, or T-level, domains are grouped by the arrangement and connectivity of secondary structure elements. The architecture level further aggregates domains by shape and orientation of structural elements with no constraints on connectivity. The most general level in CATH is the class level, that abstracts protein domains to their secondary structure content without further constraint (eg. all-alpha, alpha-beta, etc.). The latest version of CATH contains 2,626 homologous superfamilies, 1,313 topologies, 40 architectures and 4 classes (Sillitoe et al., 2013). Recently, a genome annotation approach has been launched that uses a combination of SCOP and CATH domains and derived structure-prediction tools to provide protein-function predictions for proteins with unknown structure (Lewis et al., 2012).

The concept of protein domains can be thought of as an extension of the homology concept: While in most cases, protein sequences belonging to the same domain family can not be related to each other through conventional sequence similarity methods,

they are nevertheless expected to share a common ancestor. Structure-based domain annotation methods such as CATH and SCOP make use of the greater conservation of structure over sequence to assign protein sequences with known function to structural domains. However, this is not an option for proteins with unknown structure. About two decades ago, Krogh et al. reported the application of hidden Markov models (HMMs) to address this protein classification problem (Krogh et al., 1994). HMMs have since been successfully applied to discover relationships between remotely related sequences with greater performance than traditional methods such as pairwise or multiple sequence alignments (Park et al., 1998; Eddy, 1998; Lindahl and Elofsson, 2000). HMMs can also be applied to other problems, such as the prediction of membrane-spanning sequences (Krogh et al., 2001) Two popular implementations of HMM-based sequence homology detection are SAM (Karplus et al., 1997; Karchin and Hughey, 1998) and HMMER3 (Eddy, 2008).

A number of methods to detect structural domains and functional motifs in sequences are aggregated in Interpro, a service often used as a first point of reference in the annotation of protein sequences with unknown function (Hunter et al., 2012). The mapping of small molecule binding to protein domains described in this chapter uses Pfam-A domains. Pfam-A domains are a set of manually curated HMM models of protein domains that can be obtained from the Pfam database¹⁰ (Sonnhammer et al., 1997; Sonnhammer et al., 1998). Since 2010, Pfam uses the HMMER3 algorithm to build HMMs and query sequence databases (Finn et al., 2010). Functional annotations and HMM seed alignments for Pfam-A domains are curated manually, and HMM models are subsequently used to query the UniProtKB database (Magrane and Consortium, 2011) for additional homologs of a given Pfam-A domain. Pfam-A domain models provide excellent annotation quality that has recently been integrated with the online encyclopedia Wikipedia (Punta et al., 2012). The coverage of most sequenced genomes is incomplete, but good in comparison with other annotation tools (Rekapalli et al., 2012; Mistry et al., 2013). A further benefit of Pfam-A is the manual revision process of seed alignments and the hidden markov models that ensures that Pfam-A domain annotations are to the greatest extent non-overlapping (Finn et al., 2006). The non-overlapping architectures described by Pfam-A models along with rich functional annotations for each model made Pfam-A an ideal choice for the mapping presented in this Chapter.

¹⁰In addition to Pfam-A models, Pfam also provides Pfam-B models, which are not manually curated.

1.5 Aims of the analyses

In this thesis I describe an integration approach of small molecule bioactivity data and protein evolutionary information. The analyses were carried out with the aim to explore the linkage between variation in the chemical and biological world through systematic indexing and organisation of data from both domains. Two use cases, one revolving around domain annotations from the Pfam database and the other around homology relationship types extracted from the EnsemblCompara Genetrees resource demonstrate that integration of bioactivity data can facilitate chemogenomics queries and answer biologically relevant questions.

Chapter 2

Mapping of small molecule binding to protein domains

2.1 Introduction

With the beginning of the molecular biology revolution, researchers acknowledged the value of studying drug mechanisms as the interaction of a small molecule with specific parts of a protein, often referred to as a binding site or binding pocket. An array of techniques had become available through innovations in gene cloning and artificial expression systems to study these interactions. They include site-directed mutagenesis, alanine scanning, and increasingly, X-ray crystallography, and nuclear magnetic resonance (NMR) techniques. Diverse as they are, the common objective of studies using these techniques is to determine the binding mode or in other words the location and relative orientation that a drug assumes when forming a bioactive complex with a protein. Studies of binding modes have yielded detailed insights into the mechanisms of drug action and have laid out the basis for virtual drug screening techniques such as molecular docking and pharmacophore modelling. However, compared to biochemical and cell-based assays of drug potency, technologies to determine the binding mode are relatively low in throughput and thus for many pairings of small molecules and proteins, there is information about the potency of the small molecule, but not its binding mode. This is also reflected in the data available in ChEMBL, where most measured potencies have no complement in the world's largest repository of protein structures, the protein data bank (PDB, [Berman](#)

et al., 2007). Knowledge of the binding mode is arguably the pinnacle of many drug discovery programs, but for other applications less stringent definitions of the binding site may be sufficient.

One such definition is provided by the concept of protein domains, which constitute smaller subunits of proteins that fold independently. A view of small molecule binding on the level of protein domains has lower resolution compared to a residue-level binding-mode, but it provides an evolutionary context together with the sequence segment that is most relevant to binding. An example where a domain-based annotation of small molecule binding can be useful is the screening of newly sequenced pathogen genomes for potential drug targets by sequence similarity searching (Tsai et al., 2013). Here, knowledge of where approximately a small molecule binds a protein sequence helps limit the query sequence to relevant protein domains. This reduces the query space and reduces false sequence matches. Machine learning applications in drug discovery often seek to exploit sequence information from proteins, genomes, and sometimes whole populations of genomes to predict drug-protein interactions; these applications could likely also benefit from using refined regions, such as the binding domain, as descriptors in their respective modelling approaches.

For this chapter of my thesis I have implemented a mapping of the small molecule binding events recorded in ChEMBL to protein domains as described by the Pfam collection of protein families. Protein domains and their representation in the Pfam database are introduced in more detail in section 1.4.2. In the following, I will discuss the significance of the protein domain concept to drug discovery (see section 2.1.1 below) and provide an overview of the mapping heuristic proposed in this chapter (see section 2.1.2).

2.1.1 The protein domain concept in drug discovery

Protein domain annotations have long been used to assign categories to drug discovery projects and as a valuable resource in the assessment of novel drug targets (Breinbauer et al., 2002). The interplay between molecular structure and function is generally appreciated in the field of drug discovery and frequently applied in the prioritisation of screening compounds and optimisation of lead structures (Johnson and Maggiora, 1990; Martin et al., 2002; Eckert and Bajorath, 2007). Drug discovery approaches are not limited to

the exploitation of similarities between ligand structures; similarities in protein structure can equally be exploited to infer characteristics of potential drug targets, often also with regard to small molecule binding properties. In this context, protein domains provide a more systematic view of small molecule bioactivity that is underpinned by the theory of molecular evolution. This is useful when inferring function and shared ligands beyond the limits of pairwise similarity detection (Martin et al., 1998; Orengo et al., 1999). A prominent example in this context is the unification of urease, phosphotriesterase, aminodeaminase and other enzymes in a novel amidohydrolase superfamily with shared active site characteristics (Orning et al., 1991; Jabri et al., 1995; Holm and Sander, 1997). The domain perspective also provides examples of evolved structure-activity relationships, as observed between the lactonizing enzyme, the mandelate racemase and enolase and their respective ligands (Hasson et al., 1998).

With the advent of the genomics era, domain annotations have also been used to quantify the content of the human genome in terms of ‘druggable’ domains (Hopkins and Groom, 2002; Russ and Lampel, 2005). Hopkins and colleagues estimated the number of genes encoding druggable proteins at around 3,000. Russ and colleagues proposed 2,200 druggable genes in a conservative estimate, and 3,000 in a more permissive estimate, albeit with different composition of domain types compared to Hopkins’ study. These studies, while often limited by availability of data, have also shown clearly that the bulk of current drug discovery efforts is not uniformly distributed across domain types encountered in the human genome, but rather, that there is a privileged subset of domains that are more relevant in drug discovery. The mapping presented in this chapter capitalises on, and significantly extends this finding, as described in the next section.

The protein domain concept also finds applications in the early stages of drug discovery. Newly sequenced pathogen genomes are frequently scanned for genes encoding ‘druggable’ domain types (Berriman et al., 2009; Tsai et al., 2013) in order to identify potential drug targets, or to identify proteins that can be targeted with existing drugs. In this context, domain-based queries can help avoid a phenomenon I will in the following refer to as ‘domain poisoning’, which occurs when the presence of a common ‘spectator domain’ links together targets on the basis of sequence searches, but the ligand-binding domain is absent from the identified homologue. In this case, a database search for potential ligands might result in a deluge of irrelevant compounds and thus ‘poison’ the query and complicate subsequent analysis. Figure 2.1 illustrates an example of domain

poisoning. The occurrence of domain poisoning when using whole sequences to query can be overcome by very conservative cut-offs for sequence similarity (Berriman et al., 2009). Another solution would be a domain based query, however, this requires knowledge of the binding domain for all proteins that are used to query a new genome (Tsai et al., 2013). The mapping heuristic presented in this chapter was motivated in part by this requirement.

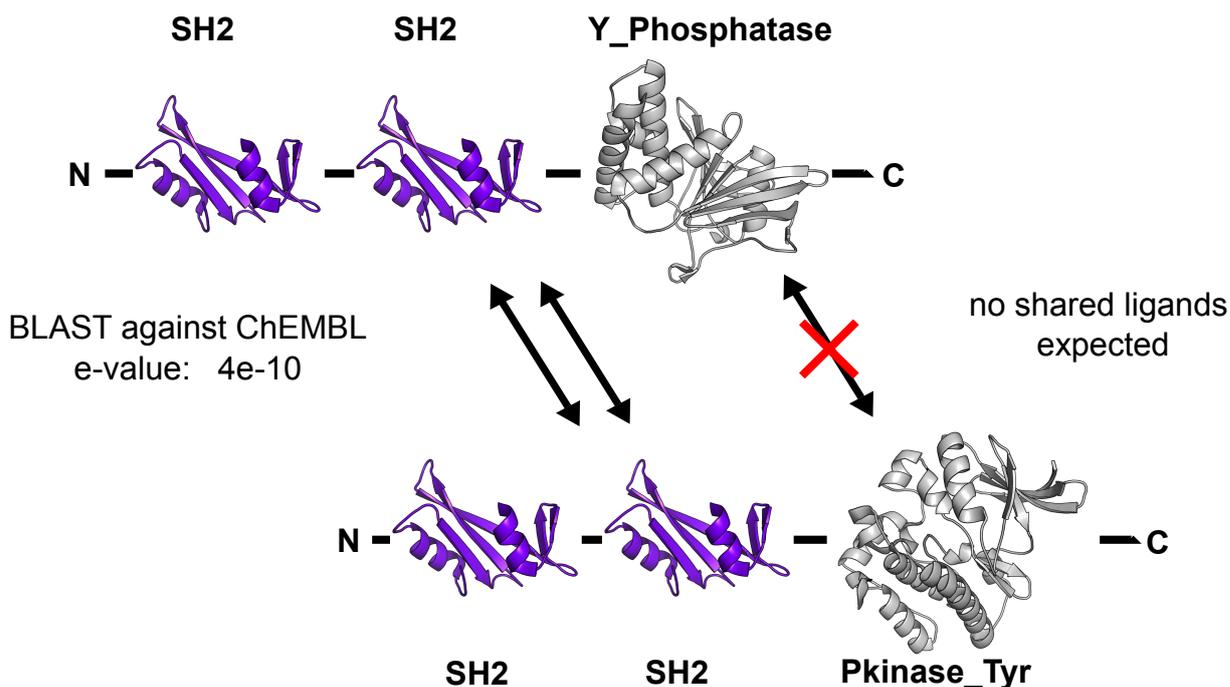


Figure 2.1: Domain poisoning of sequence-based queries. The schematic shows the domain structure of a protein in a hypothetical query - the rat Tyrosine-protein phosphatase Syp (P35235) - and one of the hits, retrieved from a BLAST query against the ChEMBL target dictionary - the rat Tyrosine-protein kinase SYK (Q64725). The significant e-value for this query results from high scoring alignments of the SH2 domains. At the same time, the overlap between small molecules binding both proteins is expected to be low.

2.1.2 A heuristic mapping that relies on domain-based annotation transfer

The domain concept introduced in the previous section provides numerous benefits when applied in a drug discovery context. One of them is a more systematic view of

small-molecule bioactivities, and another the possibility of linking potential drug targets beyond pairwise sequence-similarity. To realise these benefits, it would be desirable to annotate small molecule bioactivities with information about the domain or domains that mediate small molecule binding. In a typical drug discovery project that focusses on one or few targets, this task is often so trivial that knowledge of the binding domain remains implicit. On the other hand, in data-mining approaches, this lack of explicit knowledge of the binding domain is often insurmountable due to the sheer number of interactions that need to be examined. One of the goals of my work is therefore to provide this annotation for bioactivities stored in the ChEMBL database using a generalisable mapping procedure.

The mapping heuristic presented here is based on a corpus of protein domains with known small molecule interactions. This corpus is derived from single-domain proteins with measured small molecule interactions in the ChEMBL database. In theory, it could however be replaced by any other catalogue of domains. This catalogue can then be projected onto small-molecule protein interactions, in this case, measured bioactivities stored in the ChEMBL database. Figure 2.2 provides a schematic illustration of this process and highlights constellations that may cause issues when this mapping is applied to measured activities. In one of these constellations, the mapping is not applicable to a protein sequence because there is no overlap between domains in the sequence and domains in the catalogue. In the other constellation, more than one domain in the protein sequence of interest is present in the catalogue, resulting in a conflicting mapping. In sections 2.2.4 and 2.3.3, the impact of these constellations on the mapping process will be presented in detail.

The design and implementation of the mapping was inspired by homology-based annotation transfer, a methodology that is frequently used in bioinformatics to provide functional annotations for new protein- or nucleotide sequences. Homology-based annotation transfer can result in inaccurate or incorrect annotations, but, owing to its high coverage, often remains as the only feasible solution for large data sets (Hegyí and Gerstein, 1999; Devos and Valencia, 2000). In analogy to the transfer of functional annotation between homologous proteins, the mapping heuristic was envisaged to transfer binding site annotations between protein domains. I deemed this transfer justified because it is known from molecular dynamics simulations that conformational variability between homologous proteins is often within the bounds of the dynamic conformations

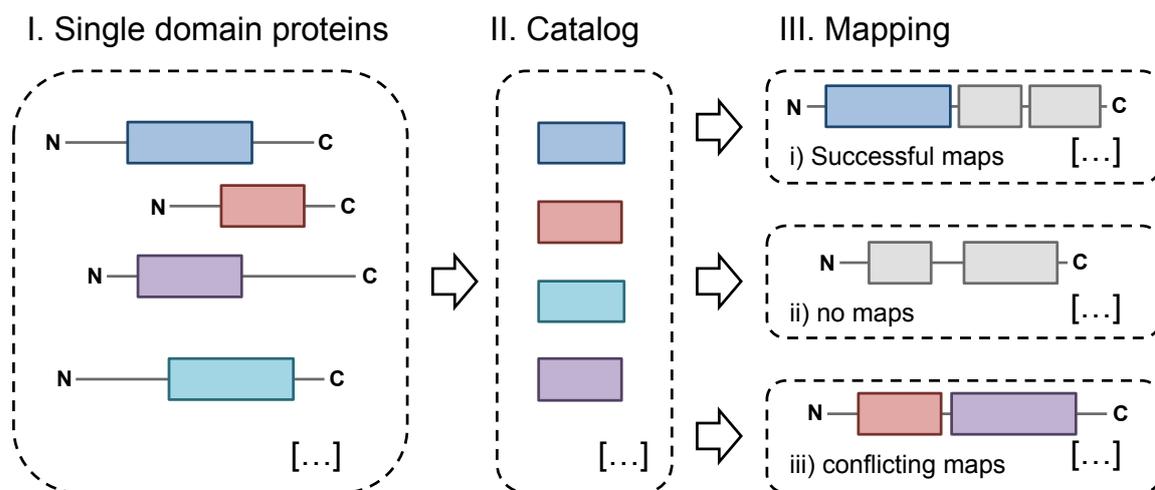


Figure 2.2: Schematic illustration of homology-based transfer of binding annotation. The schematic shows how a catalogue of Pfam-A domains with known small molecule interactions was created from the corpus of single-domain proteins within the ChEMBL database. In a second step, this catalogue was projected onto protein sequences matching more than one Pfam-A models. Three possible outcomes are: i) A successful mapping if exactly one of the Pfam-A domain models from the catalogue matches the sequence; ii) No mapping if none of the Pfam-A domain models from the catalogue match the sequence; iii) A conflicting mapping if multiple domain models from the catalogue match the sequence.

assumed by a protein in solution (Elber and Karplus, 1987). Further, from studies of functional conservation within domain families, it is known that the presence of a binding site is conserved in most cases, even though conservation of ligand specificity may vary (Martin et al., 1998; Tanramluk et al., 2009). It has further been shown that protein surfaces as well as small molecule interaction sites are conserved within domain families (Zhang et al., 2010; Davis and Sali, 2010; Kufareva et al., 2012) and an application of domain-based annotation transfer has previously demonstrated that this method can be used to make proteome-wide predictions of small molecule and protein binding sites (Davis, 2011). Thus, I present here a novel and fully implemented mapping heuristic inspired by homology-based transfer of functional annotations.

Previous computational approaches have linked small molecule binding to protein domains. In 2006, Snyder and colleagues used crystallographic structures from PDB and domain annotations derived from the conserved domain database (CDD) to obtain a set of ligand-domain interactions (Snyder et al., 2006). These interactions were used to build profiles of ‘position specific’ binding sites that map small molecule binding to a domain rather than protein sequence. Remarkably, Snyder et al. demonstrate that this approach can be useful for drug repurposing approaches. The mapping produced in this chapter differs from Snyder’s work in that it relies on potency measurements retrieved from the ChEMBL database to assign small molecule-domain interactions.

In 2009, Bender and colleagues described a method to predict protein targets for active compounds that was based on statistical associations of Interpro domain annotations and small molecule binding events recorded in the WOMBAT database (Bender et al., 2009). The representation of small molecule-domain interactions in this approach is more abstract than the direct mapping presented here and includes weighted contributions of domains that do not physically interact with a given small molecule. In 2011, Davis published the binding site prediction method discussed above (Davis, 2011). The same year, Yamanishi and colleagues proposed the use of sparse canonical correspondence analysis, a statistical approach that extracts combinations of chemical substructures and Pfam domains as features to model and predict small molecule-domain interactions (Yamanishi et al., 2011). In 2012, Wang, Nacher, and colleagues published a method that predicts drug targets based on drug-protein interactions recorded in DrugBank (Wang et al., 2012b). Later that year, an initial implementation of the mapping presented here was published (Kruger et al., 2012) and Li and colleagues published a mapping

of small molecule binding to protein domains that was based on interactions obtained from the inferred biomolecular interaction server (IBIS, see [Shoemaker et al., 2010](#)) ([Li et al., 2012](#)). Li and colleagues use this mapping to analyse a network of protein domains that are joined through shared ligands. In an analysis similar to Li and Nacher's approach, Moya-García et al. used interactions from DrugBank to construct a network of drug-domain interactions in a analysis published in 2013 ([Moya-García and Ranea, 2013](#)). Thus, there has recently been a large amount of development in the area of small molecule-protein domain interactions and it gives reason to hope that the foundation-work presented in many of these studies will find use in future drug discovery applications.

2.1.3 Outline

In this chapter of my thesis, I present an approach to map small molecule binding recorded in the form of potency assays within the ChEMBL database to confined regions within a protein. These regions stem from annotations provided through the Pfam-A collection of protein domain models (see section 1.4.2.1). Pfam domain annotation is automatic and based on protein amino acid sequence; it does not require knowledge of the three-dimensional structure of a protein. In a wider sense the aim was to obtain a catalogue of protein domains with observed small molecule interactions that are meaningful in a drug discovery context. The performance of the mapping was evaluated against binding site annotations extracted from PDBeMotif ([Golovin and Henrick, 2008](#)), a protein structure resource, and against UniprotKB ([Magrane and Consortium, 2011](#)), a protein sequence and annotation repository. The work presented in this chapter was also an effort to provide annotations of the binding region for activities stored in the ChEMBL database to the research community. The implementation and devised delivery of this mapping is described in chapter 3.

2.2 Results

2.2.1 Small molecule binding within the boundaries of Pfam-A domains

In this section, I present an analysis of the occurrence of small molecule binding sites and Pfam-A domain models. This analysis was carried out to verify a prerequisite for my mapping heuristic: The notion that small molecule binding takes place within structured regions in a protein that are covered by Pfam-A models. To probe this hypothesis I used binding site information available from two bioinformatics resources, PDBeMotif and the Uniprot database. I retrieved binding site information from PDBeMotif where interactions corresponding to a given complex of protein and small molecule were also recorded in ChEMBL (release chembl_13, for details see Methods section 2.5.6 and 2.5.8). In total I obtained binding site information for 496 complexes. Some of these complexes are represented in the PDB by multiple entries, so this corresponds to 3,341 individual PDB entries. For each complex I assessed a measure k defined as:

$$k = \frac{n(BSR_{\text{Pfam-A}})}{n(BSR)}, \quad (2.1)$$

where $n(BSR_{\text{Pfam-A}})$ is the number of binding site residues that fall within the boundaries of any Pfam-A domain and $n(BSR)$ is the number of all binding site residues (see Methods section 2.5.11). Thus, for each complex, k indicates to which degree small molecule binding is tied to specific Pfam-A domains.

To compare residue positions obtained from PDBeMotif with domain boundary positions from Pfam, the SIFTS mapping (Velankar et al., 2013) was used to translate residue positions between the two resources (see Methods section 2.5.10). As a threshold for a strong association between a small molecule binding site and a domain described by a Pfam-A model, I chose 0.5. This choice is arbitrary, but I am confident that it gives a good indication that the binding of a small molecule is mediated through this domain, even if not all residues interacting with the molecule are part of the Pfam-A model. I found that at this threshold, small molecule binding was associated with Pfam-A domains for 477 small molecule-protein complexes, or 96.2% of all examined complexes from PDBeMotif. When shifting the threshold up or down to 0.25 or 0.75, the corresponding

numbers of entries are 487 (98.2%) and 447 (90.1%), respectively. Figure 2.3A provides a graphical summary of the coincidence of Pfam-A domains and residues involved in small molecule binding.

I also retrieved binding site information from the Uniprot database (see Methods section 2.5.7). Binding site annotation in Uniprot is mostly focussed on key residues interacting with natural ligands, such as catalytic centres, and residues that leverage conformational changes upon ligand binding. Uniprot is a protein-centric database and entries for some of the proteins provide binding site annotation for multiple ligands simultaneously. In some cases, the ligand is not explicitly specified and more often is referred to by a trivial name or an acronym. It is difficult to relate these identifiers to small molecules in the ChEMBL database. For this reason, I limited my survey to protein sequences from the ChEMBL target dictionary, but did not further limit this set to small molecule-protein complexes with corresponding entries in the ChEMBL database. In total, I obtained binding site annotations for 1,460 Uniprot entries. Again, for each entry I assessed a measure k as defined in equation 2.1. The number of binding sites with $k \geq 0.5$ was 1,418, corresponding to 97.1% of all examined entries. At $k = 0.25$ and $k = 0.75$, the corresponding numbers were 1,424 (97.5%) and 1,330 (91.1%). Figure 2.3B provides a graphical overview of all possible threshold values for k . Both the coordinate data from the crystallographic structures accessed through PDBeMotif as well as the manual annotations accessed through Uniprot indicate that small molecule binding is limited in the vast majority of cases to those regions in protein sequences that are covered by Pfam-A models. Thus, a first requirement for the mapping presented in this chapter withstood a test constructed from publicly available binding site data. In the following section I present an investigation of the coverage on Pfam-A models for proteins in ChEMBL to probe a second requirement for the mapping.

2.2.2 Domain coverage of the human genome and ChEMBL target dictionary

One other prerequisite of the mapping heuristic that I present in this chapter is sufficient coverage of target proteins by the conserved regions that constitute Pfam-A families. To probe this, I investigated Pfam-A coverage both for the human proteome as well as proteins in the ChEMBL target dictionary (release chembl_13). Uniprot accessions for

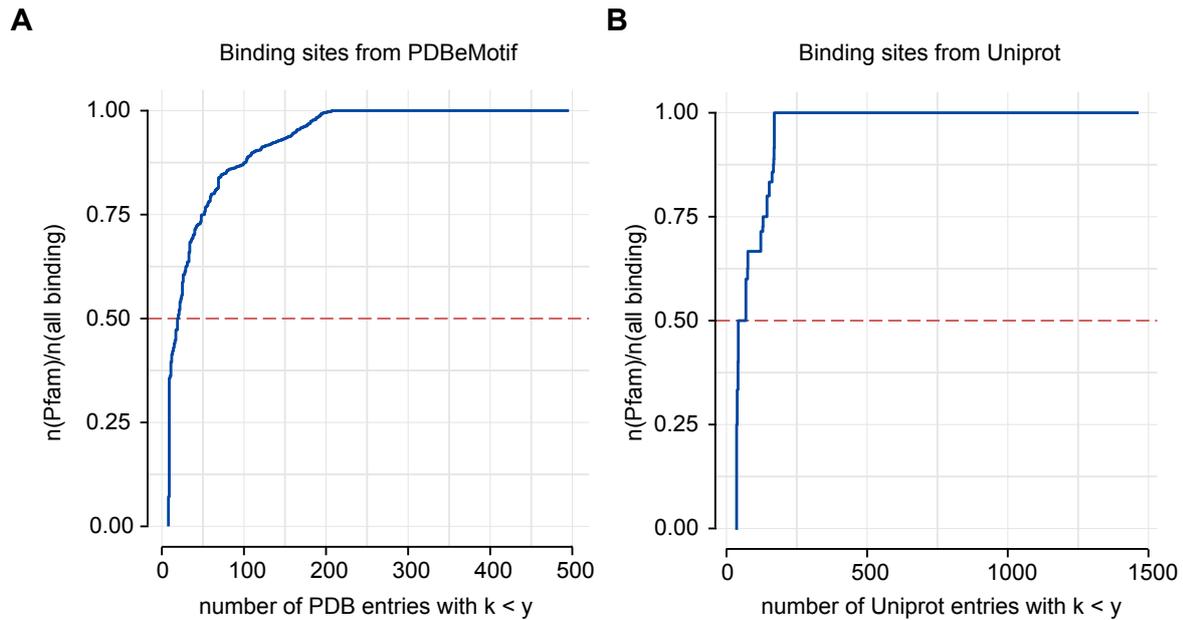


Figure 2.3: Small molecule binding within Pfam-A domains. Panel A depicts binding of small molecules within Pfam-A domains for protein-ligand complexes obtained from PDBeMotif. The measure k shown on the y-axis describes the fraction of binding-site residues within a Pfam-A domain over all binding site residues. The plot shows k as a cumulative distribution function of the number of entries having a value of at most k as indicated on the y-axis. Panel B illustrates the cumulative distribution function of the same measure for binding site annotations retrieved from Uniprot. Dashed horizontal lines represent the threshold value for k that indicates strong association of a domain and a site for small molecule binding.

the human proteome were retrieved from the Ensembl Biomart as described in Section 2.5.2. In total, I obtained 18,932 Uniprot accessions, representing all recorded proteins and protein isoforms encoded by the human genome. I retrieved Pfam-A annotations for all protein sequences associated with these Uniprot identifiers from the Pfam database (release Pfam 26.0, see Methods section 2.5.1). Further, I obtained Pfam-A annotations for 3,075 non-human proteins in the ChEMBL database target dictionary. A number of Uniprot identifiers had to be discarded because requests to the Pfam API returned empty documents for these identifiers. The errors I encountered were due to the asynchronous release cycles of the Ensembl Biomart and Pfam resources. Uniprot identifiers that have been generated after the release of the latest Pfam version do not have corresponding entries in Pfam and hence requests for these identifiers return an empty document. Of the initial 22,020 submitted Uniprot identifiers, 277 returned empty documents, leaving 21,743 proteins for the analysis. I then determined the number of Pfam-A domains for each protein sequence. 12.3% of all human proteins have no mapped Pfam-A domain, corresponding to a number of 2,305 proteins. By contrast, in the ChEMBL target dictionary, this fraction is only 1.0%, which corresponds to 58 sequences with no Pfam-A domain. For the subset of ChEMBL target proteins of human origin, this fraction is 1.2%, corresponding to 31 sequences with no Pfam-A domain. Figure 2.4A gives an overview of the distribution of domain numbers for each data set. I also determined for each protein the coverage by Pfam-A domain models. To this end, I determined a measure ρ of the coverage for each protein sequence:

$$\rho = \frac{n_{\text{Pfam-A}}}{n_{\text{all}}}. \quad (2.2)$$

Here, $n_{\text{Pfam-A}}$ is the number of residues covered by Pfam-A models and n_{all} the number of all residues in a sequence. Figure 2.4B summarises values of the coverage ρ for all sets of proteins. For the entire set of human-derived proteins, the coverage varied highly between different proteins, with a median ρ of 0.53. The outer limits of the second and third quartile were located at $\rho = 0.25$ and $\rho = 0.77$. The coverage for ChEMBL targets is again significantly higher, with median $\rho = 0.74$ and $\rho = 0.71$ for all ChEMBL proteins and the subset of human proteins, respectively. Outer limits of the second and third quartile are located at 0.58 and 0.87 for all ChEMBL proteins and 0.53 and 0.86 for the subset of human proteins in ChEMBL. Thus, Pfam-A coverage of proteins from the

ChEMBL target dictionary is significantly higher than coverage of the human genome. This finding is discussed in section [2.3.1.2](#).

2.2.3 A catalogue of protein domains with known small molecule interactions

Having established that small molecule binding as reported in ChEMBL is well covered by Pfam-A annotations, I devised a simple heuristic to infer a catalogue of Pfam-A domains with known small molecule interactions. To achieve this I examined single-domain proteins and assessed whether Pfam-A domains in these proteins could be added to the catalogue through homologous transfer. Pfam-A domains from protein sequences with only a single Pfam-A domain model were added to the catalogue if evidence of small molecule binding was recorded in the ChEMBL database. Evidence was collected from the ChEMBL database (release `chembl_13`) by screening for measured activities from binding assays that could be mapped to a protein sequence without ambiguities. Methods section [2.5.3](#) describes the process of obtaining evidence for small molecule binding in detail. It should be pointed out that in a manual curation step, a number of single-domain proteins that constitute fragments of longer sequences were flagged to be ignored when compiling the list of Pfam-A domains. This was done to ensure only Pfam-A domains that mediate the interaction with small molecules are incorporated into the catalog. Methods section [2.5.4](#) describes this step in detail.

The initial catalogue was assembled from 1,161 single-domain proteins with small molecule interactions measured in cell-free assays with unambiguously defined protein targets (for details see Methods section [2.5.3](#)). This process yielded a list of 274 Pfam-A domain types. This list did not include the `Pkinase_Tyr` domain because this type of domain does not occur in single-domain proteins, but always in combination with at least one other domain. Given the ample evidence for small molecule binding for this type of domain, I added it manually to the catalogue (for details refer to section [2.3.3.1](#)). The complete catalogue from the first implementation of the mapping thus consisted of 275 Pfam-A domain types. A full listing of identifiers is provided in the Appendix section [5.3](#). Six exemplary Pfam-A domains from this set are discussed in section [2.3](#) together with the associated evidence of small molecule binding.

In addition to Pfam domain models, the Pfam database provides a grouping of

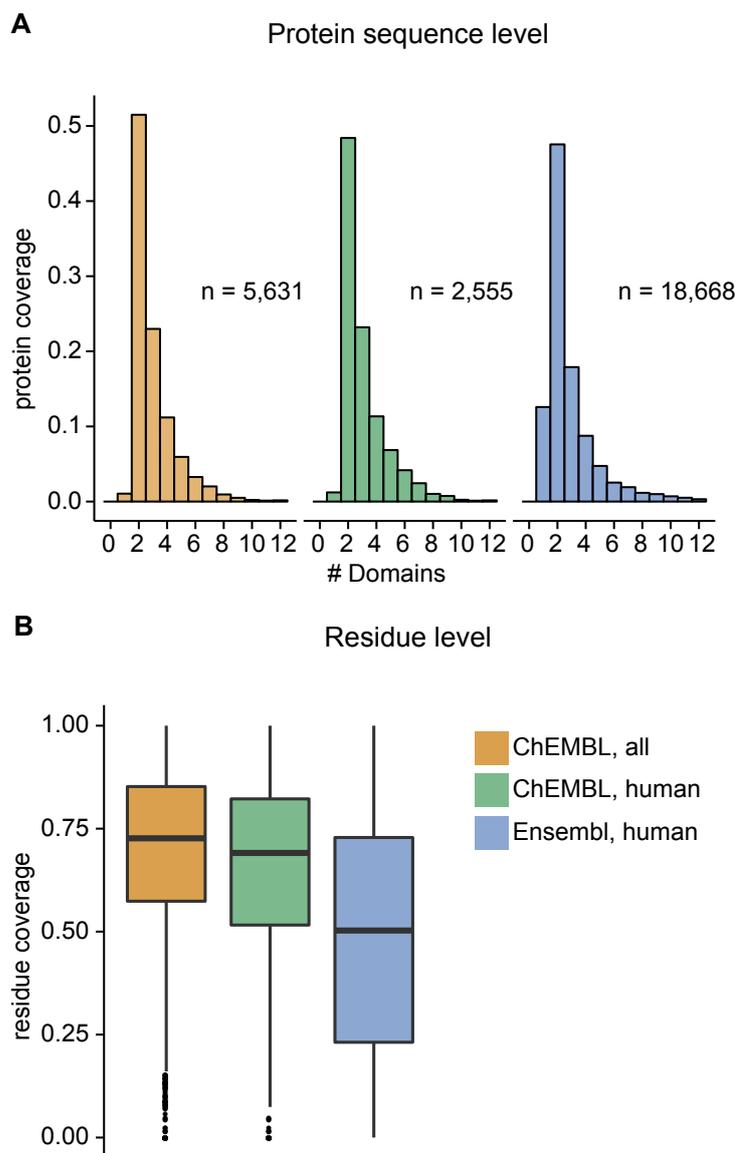


Figure 2.4: Overview of Pfam-A coverage. Panel A illustrates Pfam-A coverage on a per protein sequence level. The histogram bars represent fractions of protein sequences with n Pfam-A models, where n is the position of a bar on the x-axis. Pfam-A coverage on a protein sequence level is shown for all proteins in ChEMBL (orange), the subset of human proteins in ChEMBL (green) and the human proteome as retrieved from the Ensembl biomart (blue). Panel B shows Pfam-A coverage on a residue level, indicating for each protein the fraction of residues that are covered by Pfam-A models. The data for each set are summarised in box plots where the mid-line indicates the median, lower- and upper hinge the 25th and 75th percentile, and whiskers extend to 1.5 multiples of the interquartile range, with data points outside of these ranges plotted as outliers. Colour coding is the same as in panel A.

Table 2.1: The ten most frequent Pfam domain clans in the catalogue of Pfam-A domains with known small molecule interactions. This table lists the ten most frequent Pfam domain clans in the catalogue of Pfam-A domains with known small molecule interactions. The clan names in the first column were obtained from the `clan_id` field in Pfam. Counts of individual Pfam-A domains assigned to each clan are provided in the second column and a short description obtained from Pfam is presented in the third column.

Clan name	count	description
NADP_Rossmann	14	FAD/NAD(P)-binding Rossmann fold superfamily
Glyco_hydro_tim	11	Tim barrel glycosyl hydrolase superfamily
AB_hydrolase	9	α/β hydrolase fold
P-loop_NTPase	8	P-loop containing nucleoside triphosphate hydrolase superfamily
TIM_barrel	8	Common phosphate binding-site TIM barrel superfamily
MFS	5	Major Facilitator superfamily
PLP_aminotran	5	PLP dependent aminotransferase superfamily
Beta-lactamase	4	Serine β -lactamase-like superfamily
HUP	4	HIGH-signature proteins, UspA, and PP-ATPase.
Peptidase_CA	4	Peptidase clan CA

Pfam-A domains into clans, or collections of similar Pfam-A domains. I assessed the clan associations among the 275 Pfam-A domain types and found that most of the 102 identified clans are occupied by one or two Pfam-A domain types. There was only a small number of clans with more than two representatives in the catalogue. The ten most frequent clans are listed in table 2.1. It is important to note that this rather uniform distribution across clans does not reflect the distribution of measured activities - as described in Chapter 3, the distribution of activities over domain types follows a distribution that is much more polarised.

2.2.4 Mapping small molecule binding by domain-based annotation transfer

In the previous section I described how I established a preliminary catalogue of Pfam-A domains with known small molecule interactions. In this section I describe how I used this catalogue to infer regions mediating small molecule binding in protein sequences matching one or multiple Pfam-A domain models. The latter will in the following be

referred to as multi-domain sequences, even though it is expected that a number of those protein sequences that match only a single Pfam-A domain model are also multi-domain proteins. Mapping was attempted for protein sequences that were unambiguously defined as targets of a binding assay with at least one active compound (using the same criteria as described in Methods section 2.5.3). I inferred regions of small molecule binding for all eligible protein sequences by domain-based annotation transfer. In practice, this entailed a projection of the catalogue of validated domains onto eligible sequences. Figure 2.2 shows a schematic representation of this process. For each protein sequence subject to the mapping, three outcomes are possible:

- (i) One of the catalogue domains could be matched to a domain in the protein sequence and small molecule binding was mapped to this Pfam-A domain.
- (ii) None of the Pfam-A domains in the catalogue matched the protein sequence and no mapping was made.
- (iii) More than one Pfam-A domain from the catalogue matched the sequence and the sequence was annotated with a conflicting mapping.

The third outcome was treated as if no mapping had been assigned, but the protein sequence would be flagged for manual curation. Manual curation would be carried out using the curation interface described in chapter 3. The mapping was applied to 2,436 proteins with measured activities in the ChEMBL database (version chembl.13) and yielded mappings for 1,740 protein sequences (or 71.4%), of which 579 were multi-domain proteins. From the catalogue of 274 Pfam-A domain models with known small molecule interactions, only 59 were projected onto multi-domain proteins. Kinase domains were among the most frequently projected, followed by the `ANF_receptor` and `Ion_trans` domains. Table 2.2 lists the ten most frequently projected domains and summarises the number of targets and activities covered by this projection. Appendix table 5.3 provides a full list of the projected domains. (See methods section 2.5.5)

To validate the mappings, I again used binding site annotations obtained from PDBeMotif and Uniprot. For each sequence with an assigned mapping, I compared the overlap of annotated binding site residues and the mapped Pfam-A domain (see Methods section 2.5.11). A measure k_{pred} of the prediction coverage was determined for each

Table 2.2: Pfam-A domain types projected onto multi-domain proteins. The table summarises all Pfam-A domain types from the catalogue of Pfam-A domains with known small molecule interactions that were projected onto multi-domain proteins. The column headed ‘mapped proteins’ provides the number of multi-domain proteins a given Pfam-A domain was projected onto, the column headed ‘mapped activities’ the number of corresponding measured activities.

Pfam-A	mapped proteins	mapped activities
Pkinase_Tyr	90	9,877
Pkinase	75	4,027
ANF_receptor	48	2,358
Ion_trans	46	3,889
Hormone_recep	45	7,386
Neur_chan_LBD	33	1,163
Peptidase_C1	30	4,069
Trypsin	26	3,808
Peptidase_M10	21	4,053
PDEase_I	17	724

protein sequence:

$$k_{\text{pred}} = \frac{n(BSR_{\text{pred}})}{n(BSR)}, \quad (2.3)$$

where $n(BSR_{\text{pred}})$ is the number of binding site residues that fall within the boundaries of the predicted Pfam-A domain and $n(BSR)$ is the number of all binding site residues. For each complex, the measure k_{pred} thus indicates the degree to which small molecule binding is associated with the predicted domain. I once again argued that at $k_{\text{pred}} = 0.5$, a mapping can be considered correct because small molecule binding is clearly associated with a domain. The analysis of the prediction coverage was split according to whether a protein sequence matched one or multiple Pfam-A domain models. For small molecule-protein complexes with binding site annotations from PDBeMotif, 96.9% of single-domain complexes and 88.8% of multi-domain complexes had correctly assigned binding sites at this threshold. For proteins with Uniprot binding site annotations, 97.5% of single- and 87.2% of multi-domain proteins had correctly assigned binding sites. As expected, the mapping results are thus better for single-domain proteins over multi-domain proteins, but overall satisfactory for either category. A graphical summary of the mapping’s performance under different thresholds is given in Figure 2.5.

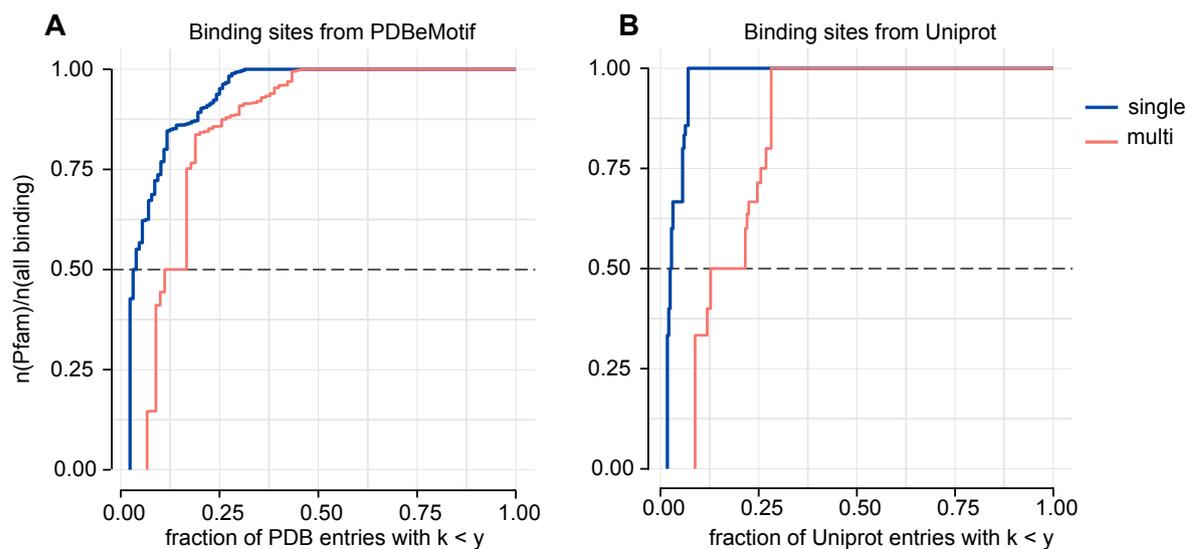


Figure 2.5: Small molecule binding within Pfam-A domains. Panel A depicts binding of small molecules within Pfam-A domains for protein-ligand complexes obtained from PDBeMotif. The measure k_{pred} shown on the y-axis describes the fraction of binding-site residues within a Pfam-A domain over all binding site residues. The plot shows k_{pred} for single- (blue line) and multi-domain proteins (red line) as a cumulative distribution function of the fraction of entries having a value of at most k as indicated on the y-axis. Panel B illustrates the cumulative distribution function of the same measure for binding site annotations retrieved from Uniprot. Dashed horizontal lines represent the value for k_{pred} that indicates strong association of a domain and a site for small molecule binding.

Table 2.3: Detail view of mapping evaluation. The table summarises the outcome of the mapping evaluation for individual target classes. Results are split by single- and multi-domain proteins. The column ' k_{pred} ' indicates the average across all assessed complexes in a protein class. The column 'count' indicates the number of assessed complexes in a class. 'LGIC' - ligand gated ion channel; 'None' - proteins without assigned target class.

protein class	PDBeMotif				Uniprot			
	single		multi		single		multi	
	k_{pred}	count	k_{pred}	count	k_{pred}	count	k_{pred}	count
7TM1	0.50	2	0.41	1	1.00	18	1.00	1
7TM3	-	-	-	-	-	-	0.80	3
Cytochrome P450	1.00	1	-	-	1.00	2	-	-
Kinase	1.00	16	0.89	29	1.00	72	0.94	156
LGIC	-	-	0.14	7	-	-	0.00	16
None	0.92	97	0.79	11	0.94	177	0.69	23
Nuclear Receptor	-	-	0.92	20	-	-	0.84	13
Phosphatase	0.97	1	1.00	1	1.00	5	0.88	4
Phosphodiesterase	1.00	4	1.00	2	1.00	5	0.61	7
Protease	0.96	7	0.98	19	1.00	5	0.62	4

For a more detailed view of the mapping performance, results of this evaluation were split into a human and non-human set and further categorised by the 'class' of the protein, as described by the L2 field in the `target_class` table of the ChEMBL database. Results are summarised in Table 2.3. Across a range of protein classes, the mapping performed well, with average values for k_{pred} well above the threshold of 0.5. However, I observed poor mapping performance for the class of ligand-gated ion channels. Factors contributing to poor performance for this class may be the complex architecture of these channels and the presence of multiple ortho- and allosteric binding sites. In section 2.3.3, I discuss some configurations that impose obstacles to accurate mapping and may in part explain the poor performance of the mapping on the class of ligand-gated ion channels. Overall, the mapping had acceptable performance for multi-domain proteins and good performance for single-domain proteins.

2.3 Discussion

2.3.1 Prerequisites of the mapping heuristic

The implementation of the mapping heuristic presented in this chapter is straight-forward. A catalogue of Pfam-A domains with known small molecule interactions was constructed from a list of single-domain proteins in the ChEMBL target dictionary. In a second step, this catalogue was projected onto multi-domain proteins. The mapping heuristic relied on three prerequisites. These were:

- (i) Small molecule binding takes place within those regions in a protein sequence that are covered by Pfam-A domains.
- (ii) Pfam-A annotations have good coverage on protein sequences that are subject to the mapping.
- (iii) Pfam-A annotations do not overlap.

The first requirement is easily explained: If small molecules routinely bind regions not covered by Pfam-A models, there is no point in mapping small molecule binding to Pfam-A domains. The second prerequisite is an extension of the first. In the first step of the mapping, a catalogue of Pfam-A domains with proven capability of small molecule binding was constructed from single-domain proteins. If structured regions in these proteins remain uncovered by Pfam-A models, then the possibility that small molecules could bind through these regions reduces confidence in the mapping. The third prerequisite ensures that there are no constellations of overlapping domains that confound the mapping. It also makes it easier to understand and interpret the mapping. Pfam-A models are curated to be non-overlapping ([Sonnhammer et al., 1998](#)) and within the ChEMBL dataset I discovered no overlaps apart from a few instances of overlap at the very margins of two Pfam-A models (data not shown). In the following, I discuss the first and second prerequisite in the light of my findings from sections [2.2.1](#) and [2.2.2](#).

2.3.1.1 Small molecule binding within the boundaries of Pfam-A domains

The first prerequisite for mapping small molecule binding to Pfam-A domains is that small molecules actually bind proteins through regions that are covered by Pfam-A domains.

There are a number of arguments to support this notion. Small molecule binding to proteins has been studied mostly at sites that are evolutionary conserved (Lichtarge et al., 1996) and therefore more likely to be covered by Pfam-A models. Whether this is due to a bias in drug discovery for ‘easier’ targets or whether it reflects an underlying propensity for small molecule binding to take place in structured regions in a protein is an important consideration (Metallo, 2010). In the context of this mapping this point is less important because the data I used to carry out the mapping is almost entirely derived from traditional drug discovery projects. To support my argument further, I tested it against binding site annotations from two public resources, Uniprot and the PDBe as described in section 2.2.1. The results of this analysis demonstrated a high coincidence of sites for small molecule binding and Pfam-A domains. 96.2% of protein-ligand complexes from PDBe had at least half of all binding site residues within a region covered by a Pfam-A model. For binding site residues obtained from Uniprot, the same measure was 97.1%. Binding site annotations from PDBeMotif are derived from structure coordinate data and typically comprise between 8 and 15 residues. Binding site annotations from Uniprot are assigned manually and mostly focussed on a small number of key residues, such as catalytic centres of enzymes or attachment sites for covalent ligands (personal communication Ursula Hinz, SIB, Geneva). Annotations from both data sources are therefore somewhat complementary. These differences are also reflected in the different shapes of the cumulative distribution functions shown in Figure 2.3. Taken together, considerations from first principles and the results presented in section 2.2.1 support strongly the prerequisite that small molecule binding (as encountered in the ChEMBL database) is confined to those regions in protein sequences that are covered by Pfam-A models.

2.3.1.2 The implications of Pfam-A model coverage

Of crucial importance to the mapping presented in this chapter is the coverage of proteins in the ChEMBL target dictionary with Pfam-A domains. This is important because evidence of small molecule binding for individual Pfam-A domains was collected under the premise that proteins with only a single Pfam-A annotation do indeed not contain other domains through which small molecule binding could take place. In section 2.2.2, I have presented an assessment of the coverage on Pfam-A models. I found that on a protein

level, about 99% of proteins in the ChEMBL target dictionary have been annotated with at least one Pfam-A domain. This is in contrast to the entire set of human proteins, where less than 88% of proteins have Pfam-A annotations. Taken together, these findings indicate that proteins in the ChEMBL target dictionary generally have good annotation with Pfam-A domains. The findings also suggest that with regard to Pfam-A annotations, proteins in ChEMBL constitute a privileged set compared to the entirety of the human proteome. On a residue level, the median coverage of the human proteome is 0.5, but with considerable variation for individual proteins, with an interquartile range of 0.52. For proteins higher in the spectrum of Pfam-A coverage, it appears likely that uncovered residues constitute N- and C-terminal extensions of Pfam-A domains beyond the borders of the HMM. In addition, these regions can represent flexible loops linking separate domains. For those proteins lower in the spectrum of Pfam-A coverage, it appears more likely that some structured regions have missing Pfam-A domain annotation, especially if one assumes that most proteins are fully structured, as previously proposed (Apic et al., 2001). A recently published study reviews the coverage of the human genome by Pfam-A and Pfam-B models and the results with regard to residue-level coverage are very similar to the results presented in this chapter (Mistry et al., 2013). The authors of the study suggest that un-annotated regions might represent a highly diverse set of families as well as unstructured regions. With regard to the mapping heuristic, neither of these types of regions are likely to mediate small molecule binding as reported in ChEMBL. Remarkably, coverage of proteins in the ChEMBL target dictionary is significantly higher compared to the human proteome and seems near-complete for the majority of proteins, with a median coverage of 0.75 and a much lower interquartile range of 0.29. It further is likely that Pfam-A domains that are relevant to small molecule binding have been prioritised during the manual process of model creation and annotation. The prioritisation of Pfam-A domain types occurring in proteins of high scientific interest, such as known and potential drug targets is stated as an explicit goal of the Pfam-A curation effort (Mistry et al., 2013; Punta et al., 2012). I am therefore confident that for most single-domain proteins, which are so crucial to the mapping, the coverage with Pfam-A models is near-complete with few omissions, that are mainly structured or unstructured regions not involved in small molecule binding.

2.3.2 Small molecule binding to Pfam-A domains and evidence from ChEMBL

The initial implementation of the mapping heuristic yielded a catalogue of 274 Pfam-A domains with evidence for small molecule binding. This catalogue was described in a conference supplement (Kruger et al., 2012) and used as the starting point for curation and analysis described in this thesis chapter. Given that the human proteome counts 5,494 individual domains and ChEMBL has 1,771 individual domains, the relatively small count of 275 suggests that there is a privileged subset of domains that are targeted in drug discovery programs. The number of domains in the catalogue is too large to discuss each individually, but in the following I discuss three exemplary Pfam-A domains and the associated evidence for small molecule binding. These three domains were selected to convey a glimpse at the underlying data set and to show that the body of evidence for small molecule binding differs between domain types. Evidence for small molecule binding to four additional Pfam-A domains is presented in the Appendix figures 2, 3, 4 and 5. Evidence for small molecule binding was collected from the chembl_15 release of the ChEMBL database.

2.3.2.1 The HN domain

The HN Pfam-A model represents a subfamily of glycoside hydrolases termed family 83. This family belongs to the clan of Sialidases, which are enzymes that catalyse the the removal of sialic acid from glycoproteins. They are found in many species, but are especially well known as virulence factors produced by bacteria and viruses. The glycoside hydrolase family 83 is found across a wide range of viruses and its members function in the virus' attachment and fusion to the host cell (Taylor, 1996). A number of crystal structures of proteins containing this domain type have been determined. The earliest structure is that of the hemagglutinin-neuraminidase of the Newcastle disease virus in the year 2000 (Takimoto et al., 2000). Multiple drugs have been developed that exert an antiviral effect though inhibition of proteins containing the related Neur Pfam-A domain, including oseltamivir, zanamivir, laninamivir and peramvir.

Evidence for small molecule binding to the HN domain is provided in Figure 2.6 and comes from two publications recorded in the ChEMBL database . In total, there were 16 compounds with 16 measured interactions for this domain. The protein targets for

these compounds are the hemagglutinin-neuraminidase from the human parainfluenza virus 1 and glycoprotein G from the Nipah virus. Evidence shown here comes from two publications (Nishino et al., 2011; Niedermeier et al., 2009). Only 6 of 16 tested compounds have a measured IC₅₀ or comparable activity type of 1 μ M or more potent. These 6 compounds demonstrate that small molecule binding to the HN domain is possible, but the evidence is scarcer than for other domain types, as for example the Carb_anhydrase domain discussed below.

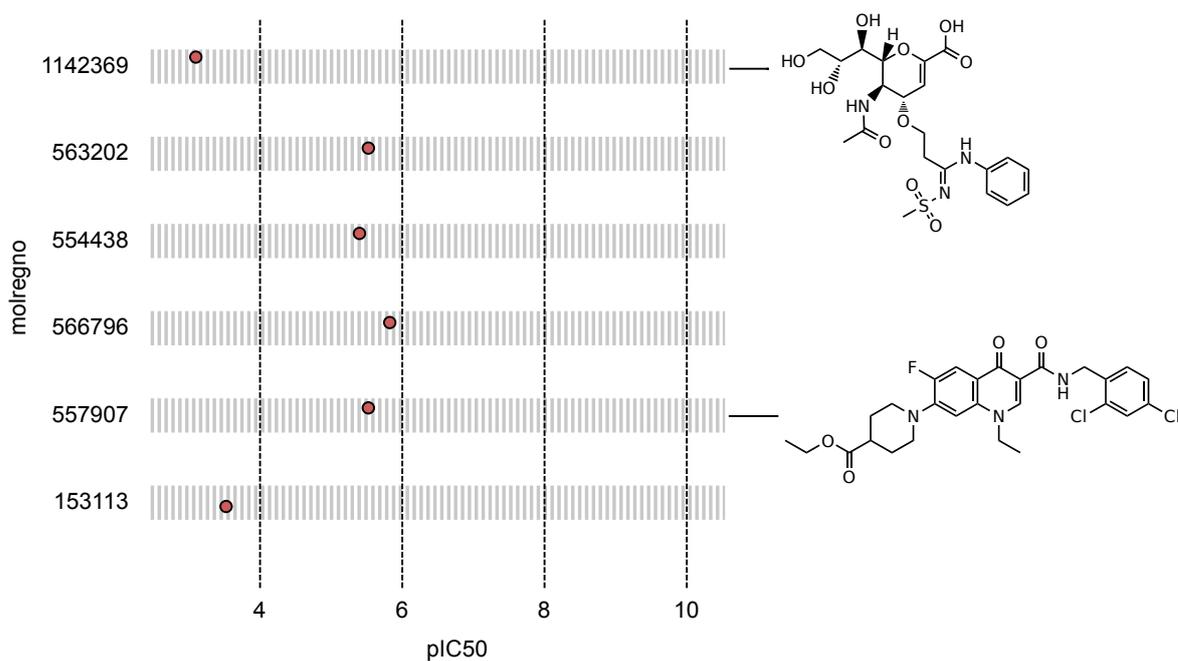


Figure 2.6: Evidence for small molecule binding to the HN domain. Measured interactions are shown for six molecules. Molecules are identified as in the `molregno` database field and measured potencies expressed as pIC₅₀ values. Two exemplary ligand structures are provided in the right plot margin.

2.3.2.2 The Carb_anhydrase domain

The Carb_anhydrase Pfam-A model represents the catalytic domain of a large family of the enzyme family of carbonic anhydrases. The Carb_anhydrase is not part of a higher order clan, but comprises more than 3,000 protein sequences across almost one

thousand different species according to the Pfam website. It is found in species from all branches of the tree of life with wide-spread occurrence in prokaryotes and virtually all metazoans (Smith et al., 1999). The enzyme catalyses the hydration of carbon dioxide to form bicarbonate, as well as the reverse reaction, and plays a crucial role in acid-base homeostasis. An abundance of crystal structures of proteins with Carb_anhydrase domains have been determined (in excess of 500) starting with the structure of the human erythrocyte carbonic anhydrase C in 1972 (Kannan et al., 1972). The human carbonic anhydrase has been exploited as a drug target for a number of indications including glaucoma, hypertension and epilepsy. Approved inhibitors of carbonic anhydrase on market are acetazolamide, methazolamide, dorzolamide and topiramate.

Evidence for small molecule binding to this domain is provided in Figure 2.7 and comes from 200 publications recorded in the ChEMBL database. In total, there were 1,882 compounds with 12,008 measured interactions for this domain. Protein targets for these compounds are 21 carbonic anhydrases from a number of species including *H. sapiens*, *M. musculus*, *B. taurus*, *S. pistillata* (smooth cauliflower coral) and others. One additional target is the human receptor-type tyrosine-protein phosphatase γ , a receptor-like protein that contains both a Carb_anhydrase as well as regions that are matched by Y_phosphatase Pfam-A domain model. The Carb_anhydrase domain is one of the highest ranking domains in terms of ligand counts and there is substantial evidence for small molecule binding to this domain for ligands of various chemotypes.

2.3.2.3 The Pantoate_ligase domain

The Pantoate_ligase Pfam-A model describes a catalytic domain that is found in the the Pantoate-beta-alanine ligase enzyme (PL) in bacteria, plants and fungi. It catalyses a condensation reaction of β -alanine and pantoate which results in the formation of pantothenate or Vitamin B₅ (Maas, 1952). Mammals do not have a functional PL enzyme and obtain Vitamin B₅ through dietary uptake. The PL of *M. tuberculosis* has been described as a potential drug target for the treatment of tuberculosis (Sambandamurthy et al., 2002) and a number of drug discovery programs have been launched to investigate lead compounds for the PL (Velaparthi et al., 2008; Ciulli et al., 2008; Hung et al., 2009). Crystal structures of the PL have been obtained for multiple species, including *M. tuberculosis* (Wang and Eisenberg, 2003), *E. coli* (Delft et al., 2001) and *S. aureus* (Satoh

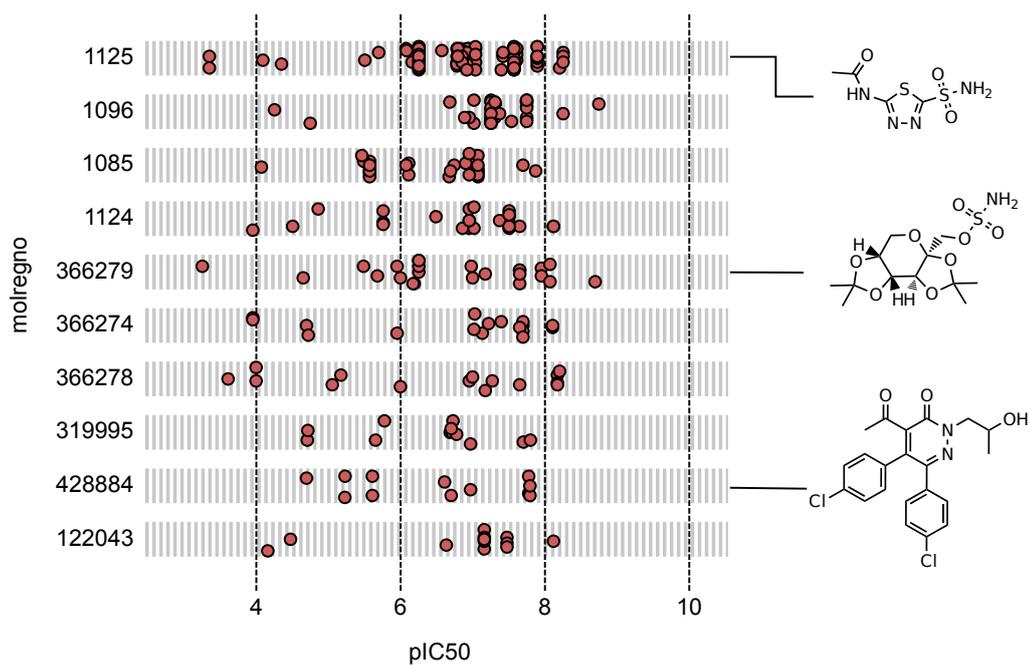


Figure 2.7: Evidence for small molecule binding to the Carb_anhydrase domain. Measured interactions are shown for ten molecules. Molecules are identified as in the molregno database field and measured potencies expressed as pIC50 values. Three exemplary ligand structures are provided in the right plot margin.

et al., 2010). Evidence for small molecule binding to this domain is provided in Figure 2.8 and comes from 2 publications recorded in the ChEMBL database. In total, there were 18 compounds with 20 measured interactions for this domain. The protein target for these compounds is the PL from *M. tuberculosis*. The evidence for small molecule binding to this domain comes from two independent publications (Velaparthi et al., 2008; Yang et al., 2011) and measured potencies are largely within an acceptable range.

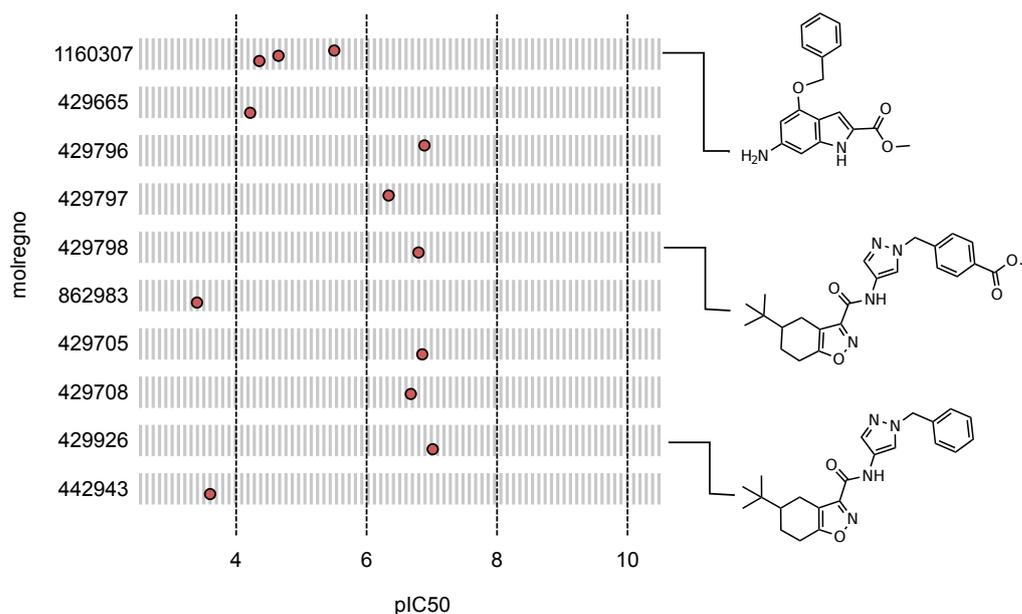


Figure 2.8: Evidence for small molecule binding to the `Pantoate_ligase` domain. Measured interactions are shown for ten molecules. Molecules are identified as in the `molregno` database field and measured potencies expressed as `pIC50` values. Three exemplary ligand structures are provided in the right plot margin.

2.3.3 Limitations of a widely applicable mapping heuristic

I presented the results of an evaluation of the mapping in section 2.2.4. Generally, the performance of the mapping was good for single-domain proteins and, with a few exceptions, at least acceptable for multi-domain proteins. It should be noted that, due to limited availability of binding site annotations, the evaluation was carried out on a small subset of the protein-ligand pairs to which the mapping was applied. Further,

the scope of the mapping is very wide, encompassing virtually all proteins that are of current or past relevance to drug discovery. The advantage of the proposed mapping is that it is simple and almost universally applicable. However, the simplicity of the process also results in generalisations that may limit the validity of the mapping for some constellations. In the following I will discuss two of these problematic constellations and their impact on the mapping presented in this chapter.

2.3.3.1 Uncatalogued Pfam-A domains

In section 2.2.3, I describe how I compiled a catalogue of Pfam-A domains with known small molecule interactions from single-domain proteins. In a second step, described in section 2.2.4, small molecule binding was projected onto proteins in ChEMBL using homologous transfer. I observed that homologous transfer was not applicable to 16.5% of proteins with measured binding activities in the ChEMBL. This was in one part due to conflicting mappings, and in the other part due to multi-domain architectures that consist entirely of domains not listed in the catalogue of Pfam-A domains with known small molecule interactions. Domains forming part of such architectures would not be catalogued because they do not occur in single-domain proteins. In this part of the discussion, I present the most frequently encountered domains in Pfam-A architectures and evaluate whether they have been adequately processed in the mapping heuristic.

As a basis for this discussion, I extracted Pfam-A domains from proteins with measured activities in ChEMBL. Table 2.4 lists the 15 most frequently encountered Pfam-A domains from multi-domain proteins in this set. The full list of domains occurring in multi-domain proteins comprises 813 domains. Most Pfam-A domains occur only once or twice (64% of the total), but the most frequent domain, **Pkinase**, occurs 122 times in multi-domain proteins with measured activities in ChEMBL. To make most efficient use of the space available, this discussion will assess only the 15 most frequently encountered domains, which cover a large fraction of the multi-domain proteins in ChEMBL.

The most frequent Pfam-A domain in multi-domain proteins is the **Pkinase** domain, a domain that was correctly included in the catalogue by homologous transfer from single-domain proteins. With a count of 109, the **Pkinase_Tyr** domain was initially not catalogued among Pfam-A domains with known small molecule interactions. However, given the ample evidence for small molecule binding at this domain from other sources,

it was included manually. This manual amendment was somewhat inconsistent with the mapping, but brought great practical improvements. It was the only amendment made to the initial version of the catalogue because the cost of violating consistency was significantly offset by the benefit of mapping activities to this very relevant Pfam-A domain. Like the `Pkinase_Tyr` domain, the `SH2` domain does not occur in single-domain proteins in the ChEMBL database. The `SH2` domain functions as a universal adapter in cell signalling and frequently co-occurs with the `Pkinase_Tyr` domain. Most drug discovery programs in this area would aim to inhibit the kinase domain, but a few small molecule inhibitors of the `SH2` domain exist (Machida and Mayer, 2005; Song et al., 2005; Siddiquee et al., 2007; Page et al., 2012). Despite this evidence, I decided not to include the `SH2` domain in the initial catalogue. This was to avoid conflicting mappings. Because of its heavy co-occurrence with the `Pkinase_Tyr` domain, all activities against proteins containing both domains would be lost, while, in practice, the vast majority of activities would be measured at the kinase domain. In chapter 3, I present a curation interface that is suitable to resolve this issue in a more coherent way, but for this part of the analysis, no further manual amendments were made. A number of other domains were not catalogued, namely the `Pkinase_C`, `zf-C4`, `SH3_1`, `fn3` and `C1_1` domains. These are frequently occurring adapter domains that are not expected to mediate small molecule binding. There was no evidence for small molecule binding from single-domain proteins and the inference that there is no small molecule binding in multi-domain proteins is adequate. The `Neur_chan_memb` domain is also not expected to mediate small molecule binding. Ligand-gated ion channels, the family of proteins where `Neur_chan_memb` occurs, bind small molecules mainly through the `Neur_chan_LBD` domain. The `Lig_chan` domain on the other hand is a good example of a domain that was ignored by the mapping heuristic, but is known to bind small molecules. The inference that, by lack of evidence from single-domain proteins, there should also not be interactions for this domain-type in multi-domain proteins is incorrect.

This overview could be continued beyond the first 15 Pfam-A domains, but for the purpose of this discussion it suffices to show that many, but not all constellations are processed adequately in the mapping heuristic. Arbitrary adjustments, such as the manual inclusion of the `Pkinase_Tyr` domain can alleviate the problem of uncatalogued Pfam-A domains superficially, but have the potential to give rise to conflicting mappings for example through clashes with the `SH2` domain. It is clear that for some constellations,

Table 2.4: Most frequently encountered Pfam-A domains in multi-domain proteins from the ChEMBL target dictionary. The column ‘count’ provides the number of time an architecture occurred in ChEMBL. The column ‘mapped’ indicates whether a domain was included in the initial catalogue of Pfam-A domains, which was assembled using homology transfer from single-domain proteins.

architecture	count	mapped
Pkinase	122	Yes
Pkinase_Tyr	109	No
SH2	52	No
Pkinase_C	46	No
zf-C4	42	No
Hormone_recep	42	Yes
ANF_receptor	41	Yes
SH3_1	31	No
Ion_trans	29	Yes
fn3	27	No
Neur_chan_LBD	26	Yes
Peptidase_C1	26	Yes
Neur_chan_memb	26	No
Lig_chan	25	No
C1_1	23	No

mappings to more than one domain are possible, depending on the ligand. A solution to this problem is presented in chapter 3 in form of a manual curation platform, where each activity can be mapped individually. Other ways of overcoming this problem can be envisaged, maybe most suitably machine learning approaches. However, even a successful implementation would require at least a limited amount of manual curation for validation purposes.

2.3.3.2 Small molecule binding at domain interfaces

The aim of the mapping was to assign small molecule binding to a Pfam-A domain for each measured activity in the ChEMBL database. One of the underlying assumptions here was that small molecule binding is tied to one domain, rather than several. This is a simplification that facilitates the mapping and also its interpretation, for example for similarity estimations of protein targets, or for the detection of potential drug targets in newly sequenced pathogen genomes. There are however known cases of small molecule binding at the interface of protein domains where contributions from both domains are

Table 2.5: Domain architectures with small molecule binding at domain interfaces. The architectures are derived from PDB entries with correlates in the ChEMBL database. The column ‘Architecture’ designates the Pfam-A domains involved in small molecule binding, the column ‘ligand’ the three letter code of the ligand. If structures with other ligands are present in the PDB this is indicated with [...]. The column ‘PDB IDs’ summarises up to three corresponding structures and ‘n’ indicates the number of proteins in the ChEMBL database that have the same architecture.

architecture	ligand	PDB	ratios	n
ADH_N, ADH_zinc_N	CCB [...]	1u3t	0.53, 0.33	21
GST_C, GST_N	D27 [...]	2vd0	0.59, 0.35	23
Guanylate_cyc, DUF1053	FOK	1ab8, 1cjk, 1cul	0.63, 0.33	18
Hexokinase_1, Hexokinase_2	LOI [...]	3goi	0.56, 0.39	14
Mur_ligase_C, Mur_ligase	1LG [...]	2am1	0.36, 0.36	6
NMT, NMT_C	MIM	2nmt	0.43, 0.40	6
OTCace, OTCace_N	PAO	1oth	0.50, 0.50	3
PH, Pkinase	IQO	3o96	0.29, 0.71	16
Peptidase_M4, Peptidase_M4_C	TIO [...]	1zdp	0.50, 0.50	2
Peptidase_S9, DPPIV_N	605 [...]	3d4l	0.53, 0.42	10
Prenyltrans_2, Prenyltrans	FPP	1mzc, 1jcq, 1sa4	0.46, 0.33	9
S-AdoMet_synt_N, S-AdoMet_synt_C	ATP	1o93, 1o9t	0.40, 0.44	2

crucial for the interaction. These interactions cannot adequately be represented in the mapping presented in this chapter. In the following I give an overview of architectures that one should be aware of when interpreting analyses derived from the mapping presented in this chapter. Architectures were systematically queried using PDBeMotif as described in Methods section 2.5.12. Ligand binding at multiple domains was detected using two thresholds, $n_{\min} \geq 5$ for the minimum number of residues involved in ligand binding and $r_{\min} \geq 0.3$ for the proportional contribution of each domain to ligand binding. These values were chosen arbitrarily, but to my judgement they give a good indication whether a domain contributes significantly to ligand binding or not. Table 2.5 gives a summary of all identified architectures. To the largest extent, the architectures I identified represent enzymes. In many cases, the regions described by different Pfam-A models represented a single functional unit, such as for example the N- and C-terminal domain of the Glutathione-S-transferase (GST_N and GST_C). This means that while these units may have different sequence profiles and be structurally independent, they still co-occur in most cases because they rely on each other to exert their function. One

striking example is the architecture of the PDB entry 3goi, representing a complex of the human glucokinase with an allosteric inhibitor (LOI) (Mitsuya et al., 2009). It consists of two Pfam-A domains, `Hexokinase_1` and `Hexokinase_2`, which are structurally similar and form a complimentary and interlocking interface. It is this interface where both glucose and the allosteric inhibitor bind the protein. The catalytic site lies in the cleft between the two Pfam-A domains, an arrangement that is also observed in other enzymes. In the following, I refer to arrangements of this type as ‘enzyme doublets’. Figure 2.9 illustrates four examples of small molecule binding at domain interfaces that fall into this category of ‘enzyme doublets’. In addition to these four, I attributed six other architectures to this category. All of the architectures in this category occur in multiple proteins with measured activities in the ChEMBL database.

The functional dependence of Pfam-A domains in enzymatic doublets imposes an obstacle to independent evolution of these Pfam-A domains and the expectation is that they occur only very rarely without their counterpart. Instead, they are expected to occur together in most if not all instances. This renders them a special case of the mandatory multi-domain architectures discussed in the previous section. This idea is supported by the fact that, exempting the `Prenyltrans` domain, none of these domains were detected when compiling the catalogue of Pfam-A domains with known small-molecule interactions (which were derived from single-domain proteins). A practical way of accommodating these architectures in the mapping presented in this chapter could be to add one domain from an enzyme doublet as a proxy for both domains.

I detected two architectures that did not assort with the ‘enzyme doublet’ category. The first of these was a combination of a protein kinase domain `Pkinase` and the pleckstrin-homology domain `PH`. The ligand is a recently discovered allosteric kinase inhibitor. The contributions of the `PH` domain are smaller than those of the `Pkinase` domain, but several literature sources state that the `PH` domain is required for inhibition (Calleja et al., 2007; Calleja et al., 2009; Wu et al., 2010). In the current implementation of the mapping heuristic, inhibition would be mapped to the `Pkinase` domain alone. It is thus a clear limitation of the mapping heuristic that it cannot represent this type of kinase inhibition adequately. Figure 2.10A and B illustrate the layout of this architecture.

Another architecture, consisting of a combination of the adenylate and guanylate cyclase catalytic domain (`Guanylate_cyc`) and a domain of unknown function (`DUF1053`), appeared to mediate small molecule binding at least in part through the `DUF1053` domain.

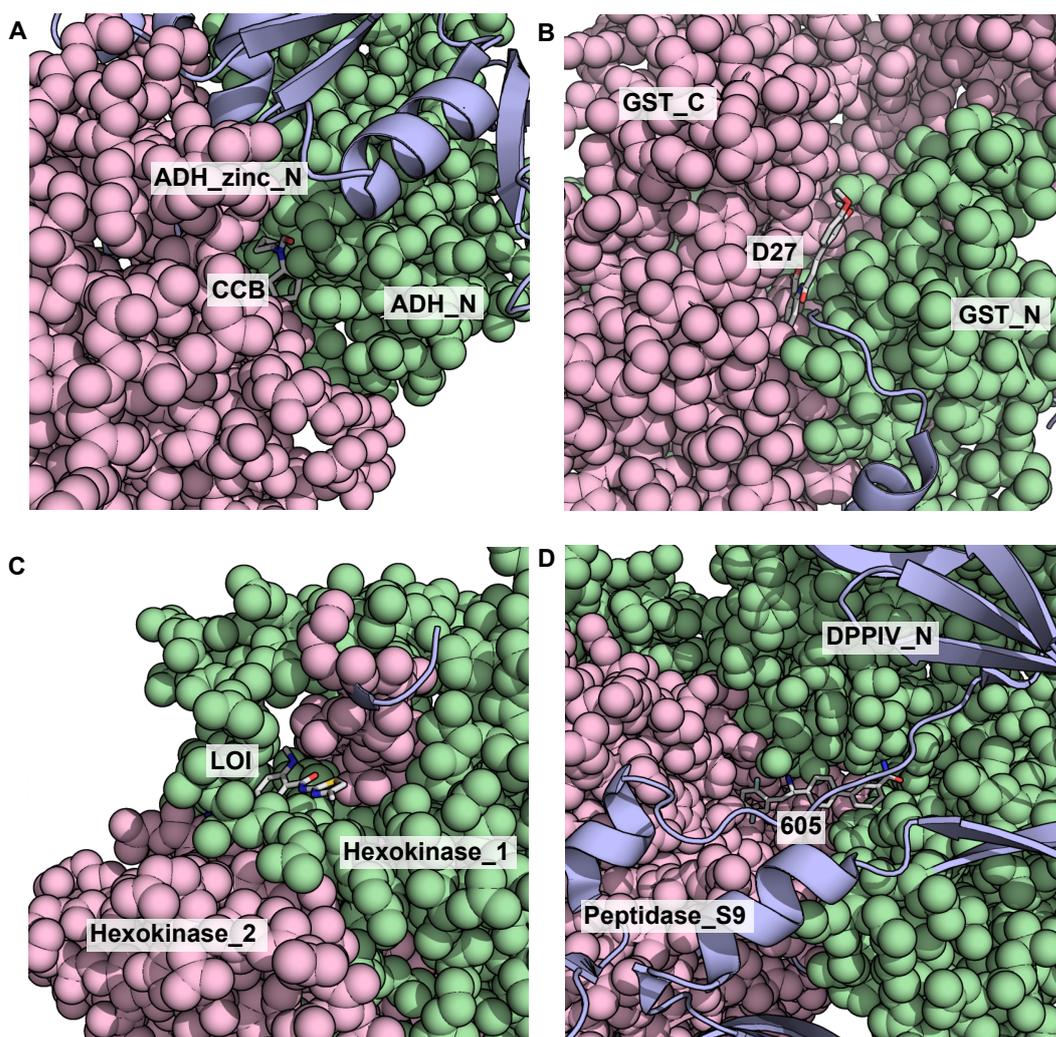


Figure 2.9: Four examples of small molecule binding at domain interfaces of ‘enzyme doublet’ architectures. Residues within binding domains are shown in space-filling representation, residues outside of these domains are shown in cartoon representation. Panel A captures the binding of the artificial inhibitor N-cyclobutyl-N-cyclopentyl-formamide (CCB) to the active site of human alcohol dehydrogenase β -1- β -1 isoform (pdb: 1u3t). Binding takes place at the interface of the alcohol dehydrogenase GroES-like domain (ADH_N, green) and the zinc-binding dehydrogenase domain (ADH_zinc_N, red). Panel B depicts binding of the inhibitor tranilast (D27) at the interface of the glutathione S-transferase N- and C-terminal domains (GST_N and GST_C) of the human glucosylceramidase (pdb: 2vd0). Panel C captures binding of the allosteric inhibitor N-(4-methyl-1,3-thiazol-2-yl)-5-[(4-methyl-4H-1,2,4-triazol-3-yl)sulfonyl]-2-(methylamino)benzamide (LOI) at the interface between the Hexokinase_1 and Hexokinase_2 domains of the human glucokinase (pdb: 3goi). Panel D captures binding of the inhibitor 4'-[(1R)-1-Amino-2-(2,5-difluorophenyl)ethyl]-3-biphenylcarboxamide (605) at the interface of the dipeptidyl-peptidase IV domain (DPPIV_N) and the prolyl oligopeptidase domain (Peptidase_S9) of the human dipeptidyl peptidase IV (pdb: 3d41).

This information was mined from a number of PDB entries including 1cjk and 1cul, which represent the enzyme adenylyl cyclase. This was unexpected, because the adenylate cyclase function depends on the presence of two `Guanylate_cyc` domains that form an active site at their interface, much like the enzyme doublets discussed in the previous section (Tesmer, 1999). A closer analysis of the structure revealed that the protein used in the underlying crystal structures was a fusion construct of a dog and a rat `Guanylate_cyc` domain (see Figure 2.10C, D). When translating between Uniprot and PDB residue coordinates, this was not taken into account by the automated script used for the analysis. Pfam annotations were obtained for the sequence of the dog protein, but parts of it should have been matched to the rat protein. The correct architecture for this entry is a combination of two `Guanylate_cyc` domains. This architecture assort with the ‘enzyme doublet’ category. As for other Pfam-A models from this category, the `Guanylate_cyc` domain was not picked up when compiling the catalogue of domains with known small molecule interactions. In this discussion, the seemingly unexpected architecture also serves as a reminder that it is challenging to anticipate all edge cases that may occur in large-scale analyses and underlines the importance of cross-checking results.

2.4 Conclusion

In this chapter, I have laid out a mapping heuristic that can be used to annotate measured activities in the ChEMBL database with the Pfam-A domain that likely mediates the interaction. The procedure is based on the transfer of binding site annotations from Pfam-A domains with known small molecule interactions onto protein sequences containing a given Pfam-A domain. To this end, a catalogue of Pfam-A domains with evidence for small molecule binding has been collated using ChEMBL potency measurements. A validation of this procedure against binding-site information from Uniprot and the PDB showed that this procedure performs well for most domain architectures. A number of domain architectures have been identified that can not be adequately processed using the mapping. Such architectures generally represent constellations where either no domain in a protein or more than one domain result in hits from the catalogue. As discussed in the following Chapter 3, the performance of the mapping for such instances could be improved by adding ‘missing’ domains to the catalogue and by resolving conflicting

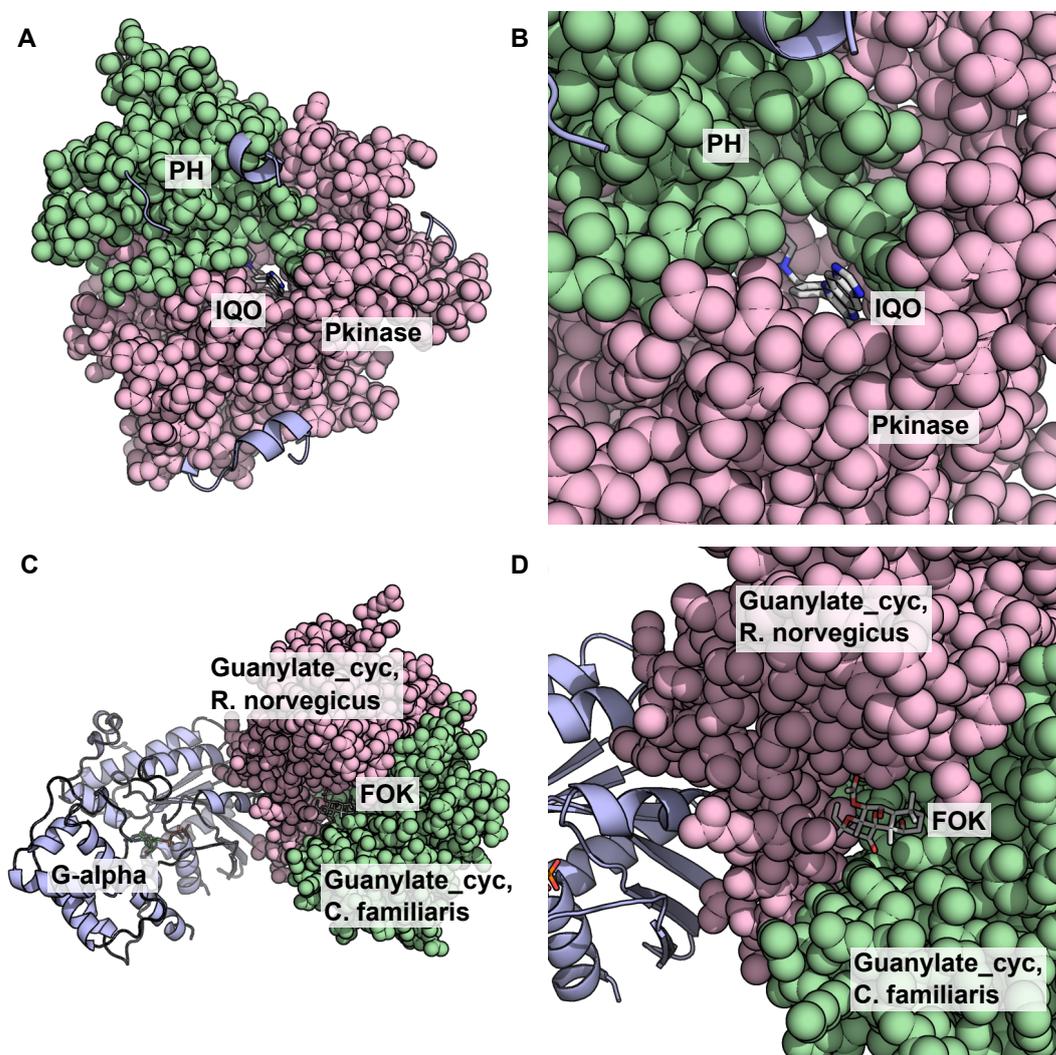


Figure 2.10: Small molecule binding at the interface of two unusual architectures. Residues within binding domains are shown in space-filling representation, residues outside of these domains are shown in cartoon representation. Panel A and B depict the binding of the allosteric inhibitor Akt Inhibitor VIII (IQO) at the interface of the Pkinase and PH domain of the protein kinase AKT1 (pdb: 3o96). Panel C and D depict binding of the inhibitor Forskolin (FOK) at the interface of two guanylate_cyc domains, which form a chimeric version of the mammalian adenylyl cyclase. The guanylate_cyc domains in this PDB entry (1cjk) were derived from two species, *C. familiaris*, and *R. norvegicus*.

mappings. The mapping will be useful for data-mining applications that have no access to implicit knowledge about binding domains and will also help disambiguate queries that seek to relate small molecule ligands for an established protein target to protein targets of unknown function, for example from newly sequenced genomes.

2.5 Methods

2.5.1 Retrieval of Pfam-A annotations

I wrote a python script to retrieve Pfam-A annotations using the RESTful API provided by the pfam web servers at <http://pfam.sanger.ac.uk>. Uniprot accessions of query proteins were used to generate a request of the form: http://pfam.sanger.ac.uk/protein/*/Uniprot/*?output=xml, where `*/Uniprot/*` corresponds to the query protein. Queries were then submitted using the python module `urllib`. The XML document returned by the Pfam API was parsed to extract all Pfam-A domains and associated start- and end positions (as defined by the `start` and `end` tags in the XML query output) using the python module `xml`. At the time this study was conducted, the Pfam API was exposing the Pfam release 26.0. The script is available at <https://github.com/fak/mapChEMBLPfam/blob/master/getPfamDomains.py>.

2.5.2 Retrieval of protein coding genes in human genome

An R script was used to obtain Uniprot accessions and amino acid sequences of protein coding genes in the human genome. I used the R package `biomaRt` to query the Ensembl dataset `hsapiens_gene_ensembl` and filtered for entries of the type `protein_coding` with the feature `with_uniprotswissprot`. The query was carried out once and results written to a tab-separated text file. The script is available at <https://github.com/fak/mapChEMBLPfam/blob/master/queryBioMaRt.R>.

2.5.3 Evidence of small molecule binding for single-domain proteins

Evidence of small molecule binding was collected for proteins that have only a single-domain mapped to their Uniprot sequence. Measured potencies and ligands were retrieved

for each single-domain target (excluding five protein fragments, see section 2.5.4) using a SQL statement and filtered to include only activities of the type `Ki`, `Kd`, `IC50`, `EC50`, `-Log Ki`, `pKd`, `pA2`, `pI`, `pKa`. To obtain only results from biochemical assays where the value of the `assay_type` field equals `B`. To ensure that the assay maps to a protein target directly and without ambiguity, I included a requirement for the `multi` and `complex` flags to hold the value `0` and the `relationship_type` attribute to hold the value `D`, which signifies that the assay target is mapped unambiguously to a Uniprot identifier. Measured activities themselves were required to meet a given potency threshold, which was set to 50 μM in this study. Evidence for individual protein was then grouped according to the Pfam-A domains. Pfam-A domains with evidence for small molecule binding were assembled into a list of ‘validated’ domains. A script used to carry out these steps is available at <https://github.com/fak/mapChEMBLPfam/blob/master/singleDomain.py>.

2.5.4 Removal of protein fragments

In some cases, small molecule bioactivities reported in ChEMBL are mapped to Uniprot identifiers that represent fragments of a protein. This might be due to annotation errors, or the lack of a Uniprot entry representing the full-length protein. Pfam-A domains from such fragments might not bind small molecules, yet they would enter the catalogue of small molecule-binding Pfam-A domains through association with a recorded activity. When projected onto multi-domain proteins, those Pfam-A domains without evidence of small molecule binding can lead to false mappings and therefore need to be excluded. I identified five critical protein fragments in the ChEMBL target dictionary. Activities associated with these fragments were excluded during the assembly of the catalogue of small molecule-binding Pfam domains outlined in Section 2.5.3. The list below provides the Uniprot names of the individual fragments and explanations to why they have been excluded.

- Fourteen activities extracted from an article about cytotoxic analogs of etoposide (Pubmed: [9804687](#)) map to [O46399](#), a fragment of sheep Tubulin that contains only a `Tubulin_C` domain. Nothing indicates that only a fragment of the protein was used in the assays specified and I attributed small molecule binding to the `Tubulin` domain, not the `Tubulin_C` domain, as would be expected for analogs of Etoposide.

- 237 activities extracted from eleven articles on phosphodiesterase inhibitors (PubMed: [8388468](#), [8709099](#), [10891111](#), [8201604](#), [9719589](#), [12570368](#), [8120866](#), [8027992](#), [8254606](#), [15780616](#)) map to the Uniprot identifier [Q864F1](#). This identifier represents an N-terminal fragment of the pig phosphodiesterase 5, containing only the `GAF` domain and, crucially, missing the `PDEase_I` domain. Thus, I excluded the `GAF` domain from the catalogue.
- In `chembl_13`, eight activities extracted from an article on excitatory amino acid receptor ligands (Pubmed: [9526567](#)) were mapped to [Q80T35](#), a fragment of the mouse metabotropic glutamate receptor 6, which contains only a `7tm_3` domain. Nothing in this article indicates that only a fragment of the protein was used and I attributed small molecule binding to the `ANF_receptor` domain and excluded the `7tm_3` domain from the catalogue of small molecule-binding Pfam-A domains.
- Six activities extracted from an article on GluR5 ionotropic glutamate receptor agonists (Pubmed: [12672235](#)) map to [Q91755](#), a fragment of the frog ionotropic glutamate receptor, which contains only a `Lig_chan` domain. Nothing in this article indicates that only a fragment of the protein was used in the assays specified and I attributed small molecule binding to either the `ANF_receptor` or `Lig_chan_Glu_bd` domain and not the `Lig_chan` domain.
- In `chembl_13`, 49 activities extracted from one of three articles on cancer therapeutics (Pubmed: [19610618](#), [20188579](#), [16415863](#)) were mapped to [A1Z199](#) a fragment of the human BCR/ABL p210 fusion protein which contains only a `SH3_1` domain. Nothing in the articles indicates that only a fragment of the protein was used in the assays specified and I attributed small molecule binding to the `Pkinase_Tyr` domain, not the `SH3_1` domain.

2.5.5 Count of projected domains

The count of projected domains and activities that are summarised in table 2.2 and appendix table 5.3 were obtained using a SQL statement against the ChEMBL database (version chembl_13):

```
SELECT domain, COUNT(DISTINCT protein_accession), COUNT(DISTINCT activity_id)
FROM map_pfam
WHERE mapType = 'multi'
GROUP BY domain;
```

2.5.6 Protein-ligand pairings in PDBe

To obtain all protein-ligands pairings from ChEMBL that could potentially be indexed by PDBeMotif, I used a SQL statement to query the database for all activities that are linked to a target with a Uniprot identifier and linked to a molecule with molecular weight lower or equal to 1,000 Da. In addition, the query was limited to molecules whose ChEMBL identifiers mapped to PDBe three-letter codes. The obtained pairings were stored in a python hash-table and used to query PDBeMotif. The script used to obtain protein-ligand pairings is available at <https://github.com/fak/mapChEMBLPfam/blob/master/getIntactDict.py>.

2.5.7 Retrieval of ligand binding residues from Uniprot

To assess the proportion of binding residues that fall into Pfam-A boundaries, I also retrieved binding site residues provided by the 'Sites' feature of the sequence annotation section provided on the Uniprot servers. Uniprot accessions were retrieved for all ligand pairings identified as described in Section 2.5.6. In many cases, a single Uniprot identifier would occur multiple times because the corresponding protein was used in assays of different small molecules. Therefore, RESTful queries were constructed for a non-redundant list of Uniprot identifiers and took the form: http://www.uniprot.org/uniprot/*/Uniprot/*.xml, where `*/Uniprot/*` corresponds to the Uniprot identifier. They were submitted using the python module `urllib` and the resulting XML documents were parsed using the python module `xml` to extract the position numbers and ligand information. The script used to send and retrieve queries is available at <https://github.com>.

[com/fak/mapChEMBLPfam/blob/master/queryUniprot.py](https://github.com/fak/mapChEMBLPfam/blob/master/queryUniprot.py).

2.5.8 Retrieval of ligand binding residues derived from PDBeMotif

To assess the proportion of binding residues that fall into Pfam-A boundaries, I retrieved binding site residues from PDBeMotif, a search tool for PDBe. Queries were constructed for each pairing of protein and small molecule that were specified in 2.5.6, using the Uniprot accession to represent the protein and the PDBe three-letter code to identify the ligand. Resulting queries were of the form:

```
requestXML=  
<!DOCTYPE query SYSTEM "http://www.ebi.ac.uk/pdbe-site/pdbemotif/query.dtd">  
<query>  
  <declaration>  
    <uniprot name="p">*/Uniprot/*</uniprot>  
    <ligand name="l1">*/TLC/*</ligand><aminoacid name="a1">X</aminoacid>  
  </declaration>  
  <bond name="b1" a="l1" b="a1"/>  
</query>
```

where `*/Uniprot/*` corresponds to the protein and `*/TLC/*` to the ligand for each pairing. The queries were submitted to <http://www.ebi.ac.uk/pdbe-site/pdbemotif/hitlist.xml> using the python module `urllib2` and the XML documents returned by the server were parsed using the python module `xml` to extract the positions of the amino acid residues that are involved in the binding of the specified small molecule. In a later step, residue positions provided by PDBeMotif were translated to match position numbers of the Uniprot protein sequence repository (see Methods section 2.5.10). The script used to send and retrieve queries is available at <https://github.com/fak/mapChEMBLPfam/blob/master/queryPDB.py>.

2.5.9 Mapping ChEMBL compounds to PDBe identifiers

PDBe identifiers were obtained for compounds in the pairings generated as described in Section 2.5.6 using mappings provided by the Unichem resource (Chambers et al.,

2013). A tab-separated table of all mappings between small molecules in the PDB and compounds in the ChEMBL database and can now be downloaded from <https://www.ebi.ac.uk/unichem/wholesourcemap>. At the time I carried out this study, the mapping was made available on request by Jon Chambers (EBI, Hinxton, UK) in a slightly different format. The mappings were parsed with a custom script and used to identify protein-ligand complexes that are indexed by PDBeMotif and correspond to measured activities in the ChEMBL database as described in Section 2.5.6. The script for parsing the mappings is available at <https://github.com/fak/mapChEMBLPfam/blob/master/parseUniChem.py>.

2.5.10 Translation of residue numbers between PDBeMotif and Uniprot

I used the Structure Integration with Function, Taxonomy and Sequences resource (SIFTS) (Velankar et al., 2013) to translate between residue numbers obtained from PDBeMotif and residue numbers in Uniprot that were used as coordinates for the mapping presented in this chapter. A look-up table was downloaded from ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/csv/pdb_chain_uniprot.csv on August 2nd, 2012 and the offset between numbering in the PDB sequence and corresponding Uniprot sequence was calculated as the relative difference between the start position in the Uniprot sequence and the start position in the PDB sequence. This offset was later added to residue numbers extracted from PDBeMotif as described in Section 2.5.9. The script is available at <https://github.com/fak/mapChEMBLPfam/blob/master/coordMap.py>.

2.5.11 Small molecule binding within the boundaries of Pfam-A domains

To assess how frequently small molecules bind within the boundaries of Pfam-A domains rather than outside of these annotated regions, I evaluated the overlap of binding residues extracted from Uniprot and PDBe with the domain boundary data obtained from the Pfam web services. The procedure was the same for Uniprot and PDBe and can be outlined as follows: For each residue, I assessed whether it lies within the boundary of any Pfam-A domain and a boolean `True` was assigned if that was the case, or else the

boolean was set to `False`. The fraction k of residues that are involved in small molecule binding and lie within a Pfam-A domain was expressed as

$$k = \frac{n(BSR_{\text{Pfam-A}})}{n(BSR_{\text{all}})},$$

where $n(BSR_{\text{Pfam-A}})$ is the number of binding site residues that fall within the boundaries of any Pfam-A domain and $n(BSR_{\text{all}})$ is the number of all binding site residues. In cases where multiple entries representing a small molecule-protein pair were retrieved from PDBe, k was calculated for a hypothetical binding site accumulating residues from all models. While $n(BSR_{\text{Pfam-A}})$ and $n(BSR_{\text{all}})$ thus increased as multiples of the number of retrieved entries, k would remain unaffected by the number of models representing a small molecule-protein pair. Fractions for individual protein-small molecule pairs were then summarised in an empirical cumulative distribution plot produced using the R library `ggplot2`. The corresponding python and R functions are available at <https://github.com/fak/mapChEMBLPfam/blob/master/matchData.py>, <https://github.com/fak/mapChEMBLPfam/blob/master/evaluatePred.py> and <https://github.com/fak/mapChEMBLPfam/blob/master/ecdf.R>.

2.5.12 Small molecule binding at domain interfaces

To assess the degree of small molecule binding at domain interfaces, I compared the ligand binding residues extracted from PDBe as described in Section 2.5.8 with the domain boundary data obtained from Pfam as described in section 2.5.1. Residue coordinates were translated between the Uniprot system used by Pfam and PDBe as described in section 2.5.10. For each domain i , the contribution to ligand binding k_i was assessed as:

$$k_i = \frac{n(BSR_{\text{Pfam}_i})}{n(BSR_{\text{all}})},$$

where $n(BSR_{\text{Pfam}_i})$ is the count of ligand-binding residues for a domain i and $n(BSR_{\text{all}})$ the total number of residues involved in ligand binding. For any domain i to be considered as involved in ligand binding, k_i was required to be greater or equal 0.3 and $n(BSR_{\text{all}})$ was required to be at least four. These thresholds were used to filter the results and exclude cases where a domain contributes only marginally to ligand binding. The script used to execute this analysis is available at <https://github.com/fak/mapChEMBLPfam/>

[blob/master/arch.py](#).

Chapter 3

Refined mapping of small molecule binding to protein domains

3.1 Introduction

In the previous chapter, a simple heuristic was presented to associate small molecule binding with Pfam-A protein domains. The mapping is a generic procedure that relies on inferences made using domains with known small molecule interactions. It does not require any additional information beyond the inferred catalogue of small molecule binding domains. The mapping was applied to a great number of small molecule-protein interactions with little need for computational or human resources. The heuristic presented in the previous chapter delivered accurate mappings for about 88% of the assessed interactions in the ChEMBL database. However, I also identified a number of shortcomings inherent to the mapping procedure. One of them was the ineptitude to process proteins containing more than one domain from the catalogue of small molecule binding domains. Interactions of small molecules with such proteins would produce conflicting mappings and were therefore left unprocessed. Secondly, the initial catalogue of small molecule binding domains was constructed from known interactions of single domain proteins. Thus, the catalogue is ‘blind’ to protein domains that have known interactions with small molecules that do not occur in single domain proteins. Both these issues are related to the more complex protein architectures. To address them, I implemented a platform that facilitates the manual refinement of individual mappings

and devised a workflow that allows for the integration of manual mappings with releases of the ChEMBL database. This chapter provides a summary of the implementation of the workflow and curation interface designed to resolve inconsistencies in the mapping of proteins that have multiple binding sites for small molecules.

3.1.1 Proteins with multiple small molecule binding sites

A significant proportion of the multi-domain architectures in the ChEMBL database could not be adequately processed using the mapping heuristic described in chapter 2 alone. One formidable challenge was imposed by proteins with multiple binding sites for small molecules. An analysis of proteins with conflicting methods (see section 3.2.4) revealed three predominant configurations associated with multiple small molecule binding sites. These are:

- phosphotyrosine signalling adapters
- multifunctional enzymes
- proteins with allosteric binding sites.

In the remainder of this section I introduce these configurations in more detail and provide examples that are relevant to the mapping heuristic. Figure 3.1 illustrates exemplary crystal structures for each type of configuration.

Phosphotyrosine signalling (PTS) is a ubiquitous process in metazoan cells that mediates external and internal signals of cell proliferation, differentiation, immune response and others. PTS relies on three components. The components are protein tyrosine kinases (PTKs), protein tyrosine phosphatases (PTPs) and modular adapters that recognise specific phosphorylated sites and recruit a downstream signalling component to the phosphorylated element (Hunter, 2009). One of the most common adapters in PTS is the SH2 domain that co-occurs with both, PTKs and PTPs (Pawson, 2004; Liu et al., 2006). SH2 domains are a prevalent feature in the genome of multicellular organisms, but absent from most unicellular eukaryotes. It is likely that the first phosphotyrosine-sensing SH2 domain occurred at around the same time as the divide of lineages into unikonts and bikonts (Liu et al., 2011a). The human genome encodes around 50 proteins with both a PTK and SH2 domain. Different members of this large family of SH2 domains

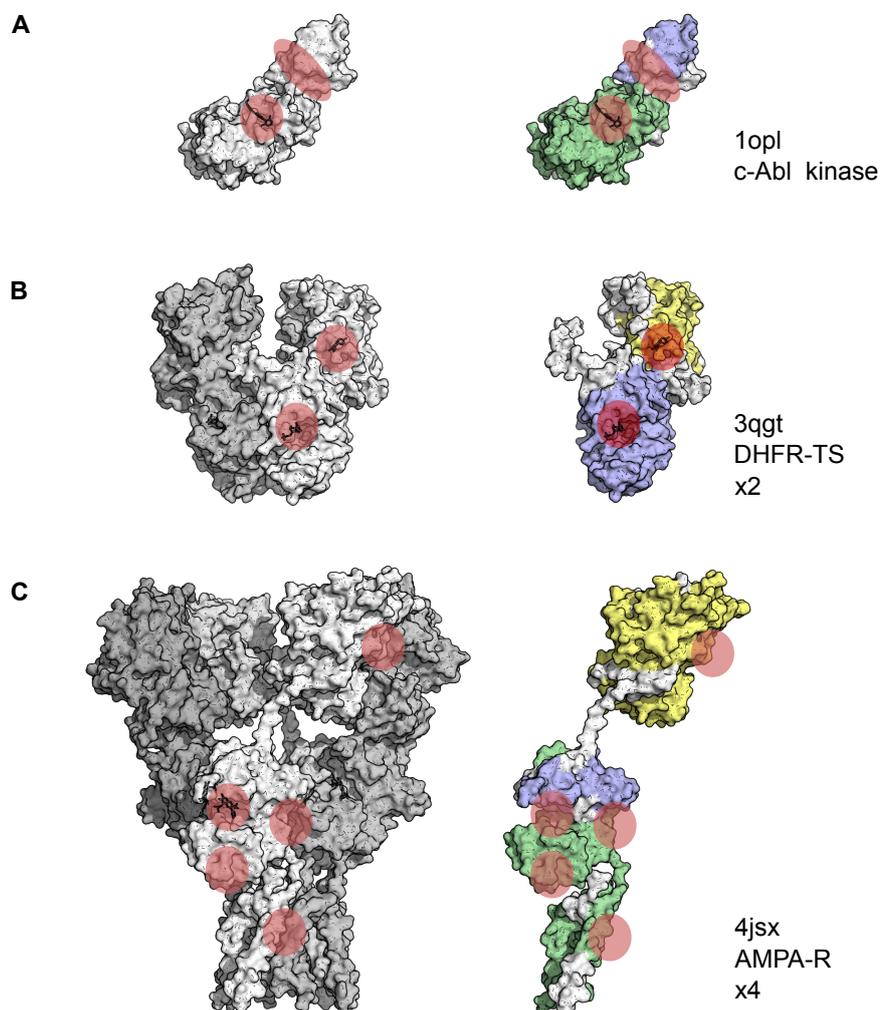


Figure 3.1: Examples of proteins with multiple small molecule interactions sites. The left hand side shows structures of whole protein complexes, the right-hand side shows individual chains. Small molecule binding sites are shown in approximation as red circles overlaid with ligands from the crystal structure. Panel A: 1opl - c-Abl tyrosine kinase. The protein contains a Pkinase_Tyr domain (green, shown with bound inhibitor PD166326), and an SH2 domain (blue, approximate binding site for substrates and inhibitors is indicated by red ellipse) Panel B: 3qgt - Bifunctional DHFR-TS as a homo-dimeric complex. Each chain contains a DHFR_1 domain (yellow, with Pyrimethamine) and a Thymidylat_synt domain (blue, with Deoxyuridine monophosphate). Panel C: 4jsx - full length homo-tetrameric complex of the AMPA-subtype glutamate receptor. The orthosteric site is located in a cleft between the Lig_chan-Glu_bd (blue) and Lig_chan (green) domains. Two additional sites are located in this region of the receptor. The N-terminal domain of the AMPA-R harbours a binding site for lectins. Philanthotoxins block the central pore of the receptor.

recognise specific substrate proteins through a number of diverse binding modes across the central β -sheet of the domain (Liu et al., 2006). The first known inhibitors of the SH2 domain were peptides (Domchek et al., 1992; Burke et al., 1994), but today a number of small molecule inhibitors are known for this domain type (Machida and Mayer, 2005; Kraskouskaya et al., 2013). PTKs on the other hand are well-established targets in drug discovery and thus any approach to mapping small molecule binding to proteins of this type must account for the possibility of small molecule binding to either domain. A crystal structure of an exemplary protein in this category (tyrosine-protein kinase ABL1) is illustrated in Figure 3.1A and approximate binding sites are indicated. Modular adapters are also found in proteins that are not involved in PTS. For example, the protein complex mTOR relies on a Rapamycin_bind domain as an adapter to recruit substrates. This process can be inhibited by a complex of the macrolide rapamycin and the endogenous protein FKBP12. The mTOR complex can also be targeted through its PI3_PI4_kinase domain and multiple inhibitors of this domain have been described (Thoreen et al., 2009; Liu et al., 2011b). A crystal structure and outlines of respective binding sites are displayed in Appendix figure 1.

The second configuration that frequently exhibits multiple small molecule binding sites are multifunctional enzymes. Multifunctional enzymes are in many instances the product of gene fusion events (Bashton and Chothia, 2007). Early work suggested gene fusion as a mechanism for the formation of bifunctional enzymes in the histidine biosynthesis pathway of *S. typhimurium* (Youno et al., 1970). Systematic analyses of fused protein domains in *E. coli* found that domain fusion frequently occurs among functionally related proteins (Marcotte, 1999; Enright et al., 1999) and in particular among enzymes in shared metabolic pathways (Tsoka and Ouzounis, 2000). The bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS) in bikonts is a product of such a fusion event. This multifunctional enzyme catalyses two subsequent steps in the biosynthesis of thymidylate monophosphate (Ivanetich and Santi, 1990). Both TS and DHFR have been studied as targets for small molecule cancer therapeutics (Touroutoglou and Pazdur, 1996; Takimoto, 1996) and DHFR-TS itself has been recognised as a viable target for anti-parasitic treatments (Knighton et al., 1994; Yuvaniyama et al., 2003; Vanichtanankul et al., 2011). DHFR-TS forms homo-dimers and each monomer consists of a DHFR_1 and Thymidylate_synth domain, corresponding to the two fused enzymatic units. Mapping of small molecule binding to domains of the DHFR-TS thus depends on the specificity

of any given small molecule ligand. Figure 3.1B illustrates a crystal structure of the DHFR-TS homo-dimer and indicates small molecule binding sites. Other proteins in this category of multi-functional enzymes are for example the fatty acid synthase (Maier et al., 2008), the 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (Hasemann et al., 1996; Rider et al., 2004) or the orotidine-5'-monophosphate decarboxylase (Wittmann et al., 2008).

A classical configuration in proteins with multiple sites for small molecule binding is the presence of allosteric sites. The term 'allosteric' signifies that binding takes place at an additional site that is spatially separate from, for example, the main binding site of a receptor, or the active site of an enzyme¹. Allosteric binding sites are known for both endogenous as well as artificial ligands. Enzyme function is often modulated through allosteric binding of metabolites from the metabolic pathway that a given enzyme contributes to. For example, the activity of the phosphofructokinase is inhibited by allosteric binding of phosphoenolpyruvate to an allosteric site on the enzyme (Whitby et al., 2000). Allosteric modulation also occurs in different types of membrane receptors, including class C GPCRs, ionotropic glutamate receptors and other ligand-gated ion channels. AMPA-subtype glutamate receptors (AMPA-Rs) for example have five² distinct binding sites which can be targeted with small molecules to modulate the receptor's activity (Traynelis et al., 2010; Sobolevsky et al., 2009). The orthosteric site of the AMPA-R binds glutamate and is located in a cytosolic, agonist binding domain (ABD). Two additional sites exist in the ABD, one at the dimer interface, binding aniracetam and related small molecules, and another at the base of the ABD, binding small molecules from the class of 2,3-benzodiazepines. The N-terminal domain (NTD) of the AMPA-R is known to interact with lectins, a class of glycosylated proteins. A corresponding domain in the structurally related NMDA-subtype glutamate receptor (NMDA-R) binds the inhibitor ifenprodil and related small molecules. An additional site lies in the central pore of the AMPA-R, which can be blocked by philanthotoxins, a class of small molecules derived from insect venoms (Stromgaard et al., 2000). Small molecule binding to allosteric sites is often desired in a drug discovery context (Christopoulos, 2002) and it is hence expected that a significant portion of measurements in the ChEMBL database relate to

¹Such sites are referred to as orthosteric sites when contrasted to allosteric sites.

²One of these sites corresponds to five partly overlapping sites at the base of the agonist binding domain (Kumar and Mayer, 2012).

allosteric inhibitors. A mapping of small molecule binding should therefore account for allosteric mechanisms in cases where the ortho- and allosteric sites lie within different domains in a given protein.

3.1.2 Outline

Proteins with multiple domains make up more than half of the proteins in simple prokaryotic organisms and about 80% of proteins in eukaryotic genomes (Apic et al., 2001). In contrast, less than 50% of the 5,631 proteins in the ChEMBL target dictionary had more than one Pfam-A domain. In one part, this is due to incomplete coverage with Pfam-A models (Mistry et al., 2013) and in another to the preference in drug discovery to select proteins with simple domain architectures as therapeutic targets (Hopkins and Groom, 2002). The heuristic described in chapter 2 resulted in conflicting mappings when applied to proteins with more than one binding site for small molecules. In addition, the mapping was not applicable to proteins that do not contain any of the domains in the catalogue of domains with known small molecule interactions, even if small molecule interactions for such a protein were known from other sources. In this chapter, I present a manual curation approach that addresses these issues. It involves both, adjustments of the catalogue, as well as manually curated mappings for protein targets with conflicting domain architectures.

3.2 Results

3.2.1 A workflow for manual refinement and integration with the ChEMBL database

The integration of small molecule bioactivity data from the ChEMBL database with other bioinformatics resources was a principal goal of my thesis project. The mapping of small molecule binding to Pfam-A domains establishes a link between the ChEMBL database and protein family annotations from the Pfam database. Both of these databases have active and asynchronous release cycles: new releases of the ChEMBL database occur every three to four months while Pfam releases are less frequent and follow a roughly

biannual cycle.³

To achieve persistent integration of the mapping and its manual refinement presented in this chapter with future releases of the ChEMBL database, I devised a workflow consisting of a python script and a modified MySQL instance of the ChEMBL database. The modification consisted of two tables, `pfam_maps` and `valid_domains`, that were added to the default schema of the ChEMBL database.⁴ Figure 3.2 shows a schematic representation of the workflow.

The table `valid_domains` contains a list of all Pfam-A domains with evidence for small molecule binding. The initial version of this table consists of all domains identified from the mapping described in the previous chapter. The `pfam_maps` table holds mappings of all eligible bioactivity measurements to Pfam-A domains. Each entry of this mapping carries a tag to mark if it has been manually curated. Both tables are created using the loading routine (see Methods section 3.4.1). Given a list of validated domains, the script projects these domains onto assay targets and maps associated bioactivity measurements to the projected domains. The loader script also tags mappings that are in conflict with other mappings. The projection encompasses all functional and binding assays that can be mapped directly to a protein sequence. A full specification of the scope of the mapping is provided in Methods section 3.4.2.

The ensuing step in this workflow is manual curation. For manual curation, I implemented a user interface as described in section 3.2.2 and Methods section 3.4.6. All measurements that were manually mapped to Pfam-A domains are tagged as such. In response to release updates of the ChEMBL database, these manual mappings can be exported into a text file (`manual_maps`). This text file, together with the exported `valid_domains` table, constitutes the input for the loader script, which projects mappings onto the next release of the ChEMBL database. In a second step, it replaces all conflicting mappings that have been manually curated with the corresponding mappings from the `manual_maps` file. Thus, I have created a workflow that allows for renewed application of the mapping heuristic described in Chapter 2 to new releases of the ChEMBL database, while retaining the manual modifications to older releases. The workflow thus has three

³archives of all releases of the ChEMBL and Pfam databases are available through <ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/> and <ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/>

⁴A visualisation of the schema of `chembl_15` can be downloaded from ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_15/chembl_15_erd.png.

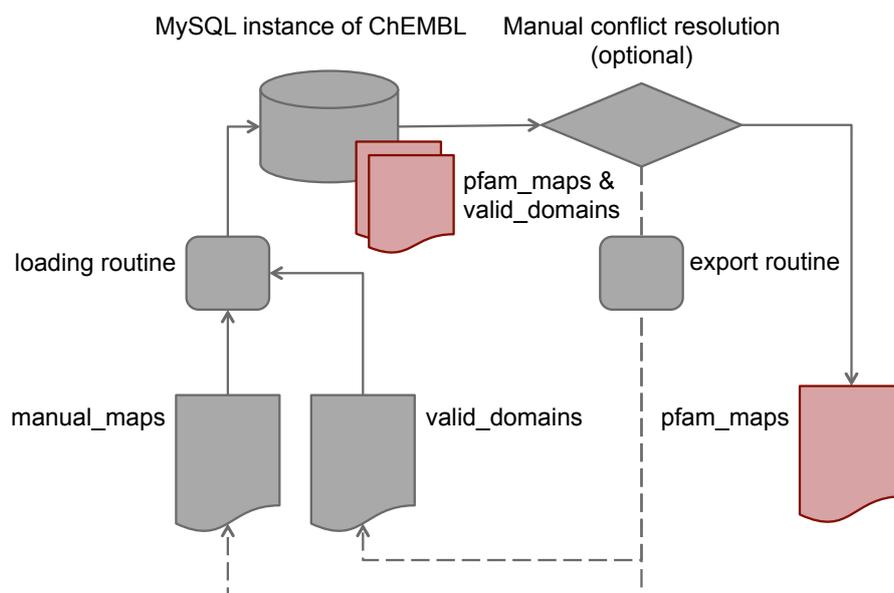


Figure 3.2: Flow chart for manual refinement and integration with ChEMBL release cycle. The schematic outlines the workflow components and their relationships. Red scrolls symbolise tables that are added to the ChEMBL schema. Beginning with the catalogue of domains in `valid_domains` and a list of previously assigned manual mappings in `manual_maps`, the loading routine scans targets in the ChEMBL target dictionary for the presence of validated domains and assigns mappings accordingly. Measurements that result in conflicting mappings are forwarded to the curation interface. At the end of a curation cycle, all manual mappings are exported to a new version of the `manual_maps` table using the export routine. Domains with known small molecule interactions can be added either directly by editing the `valid_domains` table or through the curation interface (not implemented at time of writing).

processing components, a loading routine, described in Methods section 3.4.1, an export script, described in Methods section 3.4.3, and the curation interface, which is described in section 3.2.2.

Together with Dr. Anna Gaulton, I also devised a routine to include mappings from the heuristic as well as the refined manual mappings in the ChEMBL database. Mappings for ChEMBL releases `chembl_14` and earlier were generated using the mapping heuristic described in chapter 2 only, while mappings for releases `chembl_15` and later were a combination of heuristic mappings and existing manual mappings. I prototyped the loading and export routines while Anna Gaulton implemented the schema changes necessary to accommodate the mapping. Beginning with the release `chembl_15`, the database schema contains a set of tables that describe small molecule binding sites (the corresponding schema sections in `chembl_15` are shown in Appendix figure 12). Similar to the loader script described earlier, a script is in place to project validated Pfam-A domains and existing manual mappings onto activities in the database. These mappings are represented in the aforementioned tables and can be accessed using SQL statements. As an example, Pfam-A mappings for an assay of the bifunctional dihydrofolate reductase-thymidylate synthase of *Trypanosoma cruzi* (CHEMBL1167426, from Schormann et al., 2010) can be obtained with a SQL query of the form:

```
SELECT pdb.predbind_id,  dom.domain_name
FROM assays ass
JOIN activities act
  ON act.assay_id = ass.assay_id
JOIN predicted_binding_domains pdb
  ON pdb.activity_id = act.activity_id
JOIN site_components sc
  ON pdb.site_id = sc.site_id
JOIN domains dom
  ON dom.domain_id = sc.domain_id
WHERE ass.chembl_id = 'CHEMBL1167426';
```

The mappings as represented in the ChEMBL database are static and cannot be changed between releases. Conflicting mappings curated between releases can be integrated into the next release of the ChEMBL database using the most recent export of the `manual_maps` table.

With this workflow in place, manual curation was applied as described in sections 3.2.3 and 3.2.4. The following section 3.2.2 describes my implementation of a platform for manual curation.

3.2.2 Prototype of a manual curation platform

Manual curation of measurements associated with domain configuration that have multiple binding sites requires considerable time and organisation and different domain architectures may require different strategies to resolve conflicting mappings and may thus not be achieved through the application of a generic algorithm. To accommodate these requirements I decided to implement a platform that would allow for convenient, documented and reversible assignments based on manual curation decisions.

The platform was built using the Django web-development framework (version 1.3, [Django, 2011](#)). This framework was used as a communication layer between an HTML-based user interface (UI) and a modified MySQL instance of the ChEMBL database with an additional table `pfam_maps` for the mapping information. This table is part of the workflow described in section 3.2.1. The columns of this table are laid out in the Methods section 3.4.4. The UI was composed of three sections, ‘Evidence’, ‘Conflicts’ and ‘Resolved’. A schematic representation of the sitemap is shown in Figure 3.3.

The section ‘Evidence’ was implemented to provide an overview of the catalogue of Pfam-A domains with known small molecule interactions and to list for each Pfam-A domain the evidence for which it was included in the catalogue. For most domains, this evidence consisted of activity measurements obtained from a query against the current state of the database backend. A strip-plot visualisation generated from these measurements was intended to assist the interpretation of evidence for small molecule binding. In some cases, on the basis of reports in the literature (see section 3.2.3). The ‘Evidence’ section of the UI was programmed to display evidence of small molecule binding in the form of literature references in such cases. Screenshots of the ‘Evidence’ section and an index page are shown in Appendix figures 6 and 7. The graph displayed on the ‘Evidence’ page shows some of the evidence for small molecule binding to the `COesterase` domain.

The ‘Conflicts’ section provides an overview of the domain architectures that result in conflicting mappings. The UI was programmed to display a list of assays performed

```
index
|
|-- Evidence portal
|   |
|   +-- Evidence
|
+-- Conflict portal
|   |
|   +-- Conflict
|       |
|       +-- Conflict details
|
+-- Resolved portal
    |
    +-- Resolved
        |
        +-- Resolved details
```

Figure 3.3: Sitemap of the curation platform user interface. The index page provides links to the three sections ‘Evidence’, ‘Conflicts’ and ‘Resolved’. The portal page of the ‘Evidence’ section provides a list of all Pfam-A domains in the catalogue of domains with known small molecule interactions. Each entry in this list links to a page providing evidence for small molecule binding to the individual domain. The portal page of the ‘Conflicts’ section provides a list of all domain architectures resulting in conflicting mappings and each entry links to a set of pages detailing the assays affected by the conflicting mappings. The ‘Resolved’ section mirrors the ‘Conflicts’ section, listing assays for which small molecule binding was resolved manually.

on proteins of a given architecture. For each assay, the user is presented with a view of the domain architecture, a short description of the assay as well as links to the ChEMBL and Pubmed entries corresponding to an assay and its source document. Manual curation decisions, together with comments justifying a decision, are passed back to the database and stored in the mapping table. Once resolved, an assay is moved from the ‘Conflicts’ section to the ‘Resolved’ section. Screenshots of the ‘Conflicts’ section and its index page are shown in Appendix figures 8 and 9.

The ‘Resolved’ section is a mirror of the conflicts section displaying resolved conflicts. As in the ‘Conflicts’ section, the user is presented with a view of the architecture, a short description of the assay and links to the source document. In addition, the UI displays the name of the Pfam-A domain to which small molecule binding was mapped in the manual curation process as well as the comment left by the curator. The UI can be used to revoke the mapping for a given assay and thus pass this assay back to the conflicts section. Screenshots of the ‘Resolved’ section and its index page are shown in Appendix figures 10 and 11.

In summary, I created a platform that allows to manually determine the Pfam-A domain at which small molecule binding takes place if more than one possible mapping exists. The platform enables manual mappings and their revision using the ‘Conflicts’ and ‘Resolved’ sections. The catalogue of Pfam-A domains with known small molecule interactions can be reviewed and adjusted using the ‘Evidence’ section. Changes made in either of the sections are fed back to the database tables `pfam_maps` and `valid_pfam`. In these tables, changes are logged and documented using timestamps and user comments.

3.2.3 Refinement of the catalogue of domains with evidence for small molecule binding

Evidence for small molecule binding to Pfam-A domains was initially obtained from measurements of small molecule binding to single domain proteins. As described in section 2.2.3 and the Methods section 2.5.3, a measurement was included if the measured potency was better than 50 μ M and the assay target could be unambiguously mapped to a single protein sequence. The initial catalogue was assembled using binding data from the `chembl_13` release and yielded a list of 274 Pfam-A domains.

Using the ‘Evidence’ section of the curation interface, I reviewed the evidence for

each of these 274 domains and reassessed whether a domain should be included in the catalogue of Pfam-A domains with known small molecule binding. This assessment was carried out on the basis of binding data from the `chembl_15` release. In principle, a reassessment of evidence for small molecule binding should be unnecessary as the evidence is not expected to change or disappear: an interaction once measured should persist also in new releases of the ChEMBL database. However, assay annotations in the ChEMBL database are subject to iterative improvements by biocurators, which can lead to changes in assigned target sequences or annotation flags such as the `assay_type` and `relationship_type` fields in the `assays` table of the ChEMBL database. In fact, my work related to the initial implementation of the mapping triggered a revision round of assay annotations (Yvonne Light, EBI, personal communication). I further decided to adjust the potency threshold from 50 μM to 10 μM when reviewing evidence for individual Pfam-A domains. Activities in that range often derived from highly specialised assay setups that can not reliably be compared to more standardised formats (see also section 3.3.1). In some cases multiple measurements with potency just below that threshold were observed and these were accepted as sufficient evidence for small molecule binding. In total, 34 Pfam-A domains were removed from the list of validated Pfam-A domains. In `chembl_10`, this would have reduced the total number of mapped activities by 4,580 from a total of 190,557 (2.4%). The impact on the overall coverage of the mapping is thus only minor. Of those 34 removed domains, 15 were removed because after revision of assay annotations no activities were mapped to these domains. A further 12 were removed because none of the associated measurements met the more stringent potency criteria of this refinement round and 7 were removed because associated assay annotations were incorrect. Table 3.1 lists all removed domains and provides justification for their removal. In addition, the Pfam-A domains `Telo_bind` and `Peptidase_M84` were removed and replaced with the domains `POT1` and `Reprolysin_5` to reflect name changes in the latest release of the Pfam database.

As a further manual modification of the catalogue of validated Pfam-A domains, I added a number of domains that were not initially included because they do not occur alone, but only in combination with other domains. Based on indications in the scientific literature, I added five domains to the catalogue of Pfam-A domains with known small molecule interactions. The names of these domains, together with exemplary literature references are listed in table 3.2. Below I present a short justification for each domain

Table 3.1: Pfam-A domains with insufficient evidence for small molecule binding. Table provides names of removed domains together with the number of tested ligands and the highest measured potency.

domain name	number of ligands	most potent
ANF_receptor	None	–
Aph-1	None	–
BsuBI_PstI_RE	1	19 μ M (IC50)
DNA_pol_A	3	51 μ M (IC50)
dUTPase	9	18 μ M (Ki)
Glyco_hydro_14	6	65 μ M (IC50)
Glyco_hydro_20	4	20 μ M (Ki)
Glyco_hydro_47	None	–
Glyco_transf_29	None ^a	–
HCV_NS4a	None ^b	–
HIT	3	34 μ M (Ki)
Hormone_2	None ^a	–
ICMT	None	–
IL5	None ^a	–
IL8	None ^a	–
KH_1	None	–
Lectin_legB	6	19 μ M (IC50)
Myb_DNA-binding	None	–
NAGidase	None	–
NAPRTase	2	14 μ M (IC50)
PALP	12	4.6 (pKd)
PEN-2	None	–
Peptidase_C48	None	–
Presenilin	None ^c	–
Ras	5	12 μ M (IC50)
RE_HindIII	2	17 μ M (IC50)
RHD	None	–
SHMT	1	17 μ M (Kd)
TNF	None	–
Tubulin	None	–
Urotensin_II	None	–
V-set	None	–
zf-CCCH	None	–
Lig_chan	None ^d	–

Legend: a - assay maps to incorrect target; b - high M_w peptide ligands; c - cell-based assay; d - assay maps to protein fragment

Table 3.2: Overview of domains that were added manually. Five domain types that were added manually to the catalogue of domains with known small molecule interactions are listed together with references providing evidence for small molecule binding at these domains.

domain name	references
7tm_3	Jensen and Bräuner-Osborne, 2007; Urwyler, 2011; Flor and Acher, 2012
ANF_receptor	Kunishima et al., 2000; Malitschek et al., 1999; Pin and Prézeau, 2007; Mun et al., 2004; Kew and Kemp, 2005; Mony et al., 2009
Lig_chan	Armstrong and Gouaux, 2000; Unno et al., 2011
7tm_2	Pantaloni et al., 1996; Siu et al., 2013
SH2	Machida and Mayer, 2005; Taylor et al., 2008; Kraskouskaya et al., 2013

that was inserted into the catalogue in the manual refinement step. In many instances the evidence for small molecule binding to a Pfam-A domain is ample. Here, I limit myself to providing a few references to key publications or review articles.

3.2.3.1 Small molecule binding to the SH2 domain

The SH2 domain is well known as an ‘adapter’ that mediates protein-protein interactions between cell-signalling proteins(Liu et al., 2006). While it is recognised as a potential target for cancer therapeutics, for a long time the only known SH2 inhibitors were peptides and peptidomimetics (Sawyer, 1998). Today, a range of small molecule inhibitors of the SH2 domain exist, but have not yet reached the clinic (Machida and Mayer, 2005; Taylor et al., 2008; Kraskouskaya et al., 2013). The inclusion of the SH2 domain into the catalogue triggered a number of conflicting mappings with the Pkinase_Tyr domain.

3.2.3.2 Small molecule binding to the Lig_chan domain

The Lig_chan domain occurs in ionotropic glutamate and GABA receptors. These receptors have a characteristic domain structure including a transmembrane domain (TM), an agonist binding domain (ABD) and N-terminal domain (NTD) (see section 3.1.1). The Pfam-A domains corresponding to these structural domains are the Lig_chan domain, the Lig_chan-Glu_bd and the ANF_receptor domain. However, the boundaries

of the Lig_chan Pfam-A domain extend well into the ABD in all crystal structures that I examined. Two examples that are based on a crystal structure of AMPA binding to GRIA2 (Armstrong and Gouaux, 2000) as well as Neodysiherbaine A binding to GRIK2 (Unno et al., 2011) are shown in figure 3.4. A large part of the ABD domain is made up of residues belonging to the Lig_chan domain and most contacts with the respective ligands in these structures are made through the Lig_chan domain. Within the ChEMBL target dictionary, the Lig_chan-Glu_bd domain occurs exclusively as a unit with the Lig_chan domain. This arrangement is comparable to the enzyme doublets discussed in section 2.3.3.2. Here, I used the Lig_chan domain as a proxy for the Lig_chan-Glu_bd and Lig_chan doublet. The insertion of the Lig_chan domain created a number of conflicting mappings with the ANF_receptor domain, that also occurs in ionotropic glutamate and GABA receptors.

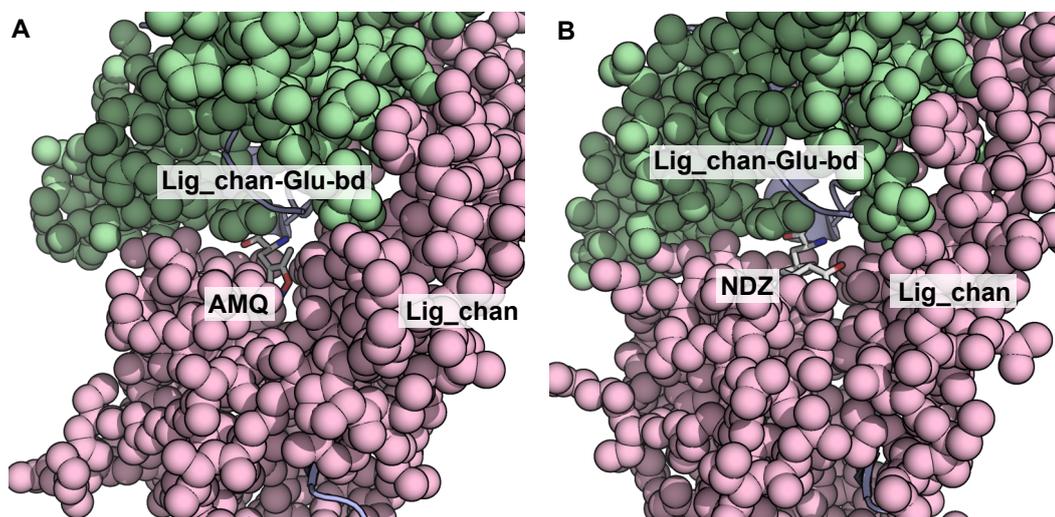


Figure 3.4: Binding of two small molecules at the interface of the Lig_chan-Glu_bd and Lig_chan domains. Residues within binding domains are shown in space-filling representation, and other residues are shown in cartoon representation. Panel A and B depict the binding of AMPA (AMQ) and Neodysiherbaine A (NDZ) at the interface of the Lig_chan-Glu_bd (green) and Lig_chan (pink) domain of the AMPA and kainate glutamate receptors (pdb: 1ftm, 3qxm).

3.2.3.3 Small molecule binding to the ANF_receptor domain

The ANF_receptor domain occurs in proteins with diverse domain architectures. Among the proteins containing the ANF_receptor domain are the class C GPCRs and ionotropic

glutamate receptors (iGluRs). The family of type C GPCRs comprises metabotropic glutamate receptors (mGluRs), calcium-sensing receptors (CaRs), GABA_B-Rs and taste receptors. Glutamate and γ -Aminobutyric acid, the natural ligands for the mGluR and GABA_B-R (as well as the artificial ligand baclofen and others) bind their respective receptors at the dimer interface formed by two ANF_receptor domains (Kunishima et al., 2000; Malitschek et al., 1999; Pin and Prézeau, 2007). Binding of L-amino acids to the CaR is also mediated through the ANF_receptor domain (Mun et al., 2004). The group of iGluRs includes the AMPA, NMDA and kainate receptors. Ifenprodil and a series of analogous compounds are well known inhibitor of the NMDA receptor that acts through the ANF_receptor domain (Kew and Kemp, 2005; Mony et al., 2009). The ANF_receptor domain had been removed from the catalogue because in the chembl_15 release, there was no evidence for small molecule binding to single domain proteins containing the ANF_receptor domain. However, given the ample evidence in the literature for small molecule binding to this domain, I reinserted the ANF_receptor domain back into the catalogue of domains with known small molecule interactions.

3.2.3.4 Small molecule binding to the 7tm_3 domain

Proteins in the ChEMBL target dictionary that contain a 7tm_3 domain belong to the class of type C GPCRs, which includes mGluRs, CaRs, and GABA_B. In addition to a 7tm_3 domain, receptors of this type contain a copy of the ANF_receptor domain. Class C GPCRs receptors form dimers in the cellular membrane. There are numerous synthetic modulators that bind class C GPCRs through an allosteric site located in the 7tm_3 domain (Jensen and Bräuner-Osborne, 2007; Urwyler, 2011; Flor and Acher, 2012). Given the large number of ligands available for class C GPCRs, I inserted the 7tm_3 domain into the catalogue. This resulted in 3,893 mapping conflicts with the ANF_receptor domain, which had to be resolved manually.

3.2.3.5 Small molecule binding to the 7tm_2 domain

The 7tm_2 domain occurs in class B GPCRs. Receptors in this class are hormone receptors and their endogenous ligands are usually large peptides such as glucagon or secretin. Class B GPCRs consist of a seven-transmembrane domain (7tm_2) and an extracellular domain (HRM). Studies of splice variants (Pantaloni et al., 1996), mutation

studies (see [Siu et al., 2013](#) for a summary), and the crystal structure of the glucagon receptor ([Siu et al., 2013](#)) have shown that ligand binding is mediated by both domains. The 7tm_2 and HRM domain always occur together in the ChEMBL target dictionary, thus forming a unit comparable to the enzyme doublets discussed in section 2.3.3.2. Here, the 7tm_2 domain was included as a proxy for the HRM domain.

In summary, a manual revision of the catalogue of Pfam-A domains with evidence for small molecule binding resulted in the removal of 34 Pfam-A domains with insufficient evidence for small molecule binding, and the addition of five domains that had previously not been included because they occur only in multi-domain architectures, resulting in a final total of 29 removed Pfam-A domains. Some of these added domains resulted in additional conflicting mappings.

3.2.4 Manual curation of conflicting mappings

The total number of bioactivity measurements in the chembl_15 release that were covered by the mapping procedure was 360,429. Of these, 12,322 measurements (3.4%) were associated with conflicting mappings while another 14,442 measurements (4%) were associated with architectures containing multiple instances of the same domain type. Curation was not carried out on the latter configurations since the mappings correctly associate small molecules with the correct domain type. I identified ten configurations of Pfam-A domains that resulted in conflicting mappings. An overview of these configurations is given in table 3.3.

Each of these configurations were subjected to manual curation and curation was completed for the vast majority of conflicting measurements (see last column in table 3.3). In an effort to make curation decisions transparent, I recorded my comments that justify each curation decision. In the following, I present domain configurations that caused conflicting assignments when subjected to the mapping heuristic and discuss the strategies I used to manually resolve these conflicts.

The largest number of conflicting assignments ($n = 4,931$) derived from domain architectures containing both an SH2 domain as well as a Pkinase_Tyr domain. This architecture is found mainly in non-receptor tyrosine kinases (NRTKs) ([Hubbard and Till, 2000](#)). Most NRTKs are involved in downstream integration of growth factor signalling and thus are attractive targets for cancer therapeutics, such as the ABL1 inhibitor imatinib

Table 3.3: Domain configurations in conflicting mappings. The co-occurring domains from the catalogue of Pfam-A domains with known small molecule interactions are listed in the column headed ‘configuration’. The numbers of associated measurements and assays are laid out in the columns ‘n(measurements)’ and ‘n(assays)’.

configuration	n(measurements)	n(assays)	resolved
7tm_3 vs. ANF_receptor	3,893	327	59%
ANF_receptor vs. Lig_chan	1,988	234	31%
ANF_receptor vs. Pkinase_Tyr	10	1	complete
Carb_anhydrase vs. Y_phosphatase	3	1	complete
DHFR_1 vs. Thymidylat_synt	1,163	60	complete
HCV_capsid vs. RdRP_3	35	4	complete
OMPdecase vs. Pribosyltran	7	4	complete
Pkinase_Tyr vs. SH2	4,931	491	complete
RVP vs. rve	8	1	complete
SH2 vs. Y_phosphatase	284	15	complete

for the treatment of chronic myeloid leukaemia (Druker et al., 2001). Drug discovery and medicinal chemistry projects have to the largest extent sought to inhibit NRTKs through the kinase function and thus explored small molecules binding the Pkinase_Tyr domain. More recently, inhibition of NRTKs has also been achieved using small molecules binding the SH2 domain (Machida and Mayer, 2005; Taylor et al., 2008; Kraskouskaya et al., 2013). However, the vast majority of assays in the medicinal chemistry literature measure inhibition of the Pkinase_Tyr domain. While inspecting conflicting mappings of this type in the curation interface I noticed that assay descriptions contain the string ‘SH2’ in cases where NRTK inhibition is mediated through the SH2 domain. Therefore, conflicting mappings of this type could simply be resolved by matching a simple regular expression over the associated assay descriptions.

Domain architectures containing both a 7tm_3 domain as well as an ANF_receptor domain yielded 2,329 conflicting mappings. Proteins with architectures of this type belong to the class C GPCRs (Pin et al., 2003). This class of GPCRs encompasses calcium sensing receptors, GABA_B receptors and metabotropic glutamate receptors. Unlike other GPCRs, these receptors form dimers when activated. Class C GPCRs bind their endogenous ligands through a so-called ‘venus flytrap’ module which corresponds to the ANF_receptor domain. In the medicinal chemistry literature, this site is referred to as

the orthosteric site. However, most of the artificial modulators that have been developed for class C GPCRs exert their function through an allosteric site closer to the membrane. This site falls mostly within the boundaries of the 7tm_3 domain. Manual curation of binding site assignments was based on assay descriptions. During this process, I queried assay descriptions for the terms ‘allosteric’ and ‘orthosteric’. Further I queried assay descriptions for names of compounds frequently used in radio-ligand displacement assays such as ‘MPEP’, ‘LY354740’ and ‘baclofen’. The occurrence of any of these names in an assay description provides indirect information about the binding site that is assessed by a given assay. The generic procedure I employed for manual curation is described in section 3.4.7.

The third largest number of conflicting mappings was caused by proteins containing both an ANF_receptor domain and a Lig_chan domain. In the ChEMBL target dictionary, this domain combination occurred in ionotropic glutamate receptors. Ionotropic glutamate receptors have up to five documented sites through which small molecules can modulate the receptor function. Figure 3.1 gives an overview of these sites and their location relative to the receptor’s domain architecture. Manual curation of this type of mapping conflict proved difficult, as the exact site of interaction was in many cases undocumented in both the assay description and the original publication. Assays for this type of receptor are often set up to measure displacement of a radioactive ligand. In most cases, the displaced ligands are well-documented chemical probes (Kew and Kemp, 2005; Jensen and Bräuner-Osborne, 2007). One viable strategy of manually mapping small molecule interaction sites is thus by inference from the ligand whose displacement is measured in the assay. While this strategy could be applied with some success, there still remained a number of assays for which the site of small molecule binding could not be extracted from either the assay description or the original publication. These mappings were left uncurated.

Conflicting mappings caused by proteins containing both a DHFR_1 and Thymidylat_synt domain derived from the bifunctional dihydrofolate reductase-thymidylate synthase of bikont organisms such as Trypanosoma cruzi and Plasmodium falciparum. The site of small molecule binding was extracted using the standard procedure for manual curation outlined in the Methods section 3.4.7.

The remaining architectures that resulted in conflicting mappings had only few associated measurements and could effortlessly be curated on a case-by-case basis using

the standard procedure for manual curation.

3.2.5 Coverage of measurements in the ChEMBL database

The primary aim of adding domains manually to the catalogue (see section 3.2.3) was to extend the coverage of the mapping and make it applicable to domain architectures that do not contain any of the domains contained in the initial version of the catalogue. In a first step, I evaluated the mapping coverage χ_{meas} as a fraction of measurements with domain mappings over all measurements in the ChEMBL database. Only measurements that represented proven small molecule-protein interactions were included in this calculation. The query to select these proven measurements was designed in analogy to the query used to detect evidence for small molecule binding to a single-domain protein: measurements were required to originate from binding assays with potency expressed as either of the following IC50, EC50, AC50, Ki, Kd, and a potency value of at least 50 μM (see Methods section 3.4.8). Using these criteria, I retrieved 308,548 measurements of small molecule protein interactions from the ChEMBL database (release `chembl_15`). Of these, 270,893 were mapped to at least one protein domain, resulting in a coverage χ_{meas} of 87.8%. I also evaluated coverage on a per-protein basis χ_{prot} . The number of targets covered by the mapping in `chembl_15` was 2,181 of a total of 2,800 targets with proven small molecule interactions, resulting in a coverage χ_{prot} of 78.0%.

For a more detailed view of the mapping coverage in multi-domain proteins, I examined patterns of domain co-occurrence among proteins with measured small molecule interactions. To represent these patterns, I created a network graph in which nodes represent Pfam-A domains and vertices represent proteins that contain any two connected domains in their sequence (see Figure 3.5 and Methods section 3.4.8). The entire network consists of 884 nodes which partitioned into a giant component of 317 nodes connected by 864 edges and 183 smaller components comprising a total of 567 nodes connected by 1,005 edges. Some of the nodes in the giant component, such as the `Pkinase`, `Pkinase_Tyr` and `SH2` domains formed large numbers of connections with domains that were otherwise isolated. Other nodes were embedded in intricate subnetworks, such as the `ANF_receptor`, `Lig_chan` and `7tm_3` domains. Viral polyproteins were represented by large isolated clusters of highly connected domains. The majority of network components consisted of only a few domains. Many of these represent enzymatic doublets as discussed in

section 2.3.3.2, for example the combination of `GST_C` and `GST_N` domain or `Tubulin` and `Tubulin_C` domains. The network view also revealed that the coverage of all possible domain architectures was incomplete. Only 65 domains from the catalogue were projected in the mapping process. These co-occur with 291 additional domains, resulting in a subnetwork of 356 domains connected by 373 edges. In contrast, a remaining 567 domains do not co-occur with any of the domains from the catalogue. As shown earlier, these remaining architectures account for only a minority of measured small molecule-protein interactions in the ChEMBL database.

3.3 Discussion

3.3.1 Changes to the catalogue of domains with known small molecule interactions

Manual curation of the catalogue of Pfam-A domains with known small molecule interactions resulted in the removal of 34 Pfam-A domains and insertion of 5 Pfam-A domains. Pfam-A domains were removed if evidence for small molecule binding was deemed insufficient. The initial catalogue was adapted from [Kruger et al., 2012](#). The catalogue presented in this publication was derived from an application of the mapping heuristic presented in chapter 2 to the `chembl_13` database release.

Subsequent releases of the database have seen improvements in the target annotation for a number of assays. Thus, a number of Pfam-A domains that were mapped to assays of small molecule binding in the `chembl_13` release are no longer associated with these assays in the `chembl_15` release. Such disassociation can occur when assay targets that were initially mapped to protein fragments containing only a single domain are mapped to full length proteins with multiple domains. In other cases, assay meta-data stored in database fields such as `assay_type` or `relationship_type` may have been corrected, for example to reflect that the readout from an assay relies on whole-cell response rather than a cell-free system. Consequently, Pfam-A domains with no associated measurements from binding assays in the `chembl_15` release were removed from the catalogue.

As described in section 3.2.3, I also applied a more stringent potency threshold of 10 μM to evaluate whether a measurement counts as evidence for small molecule binding. I selected this threshold over the more lenient threshold of 50 μM because experience

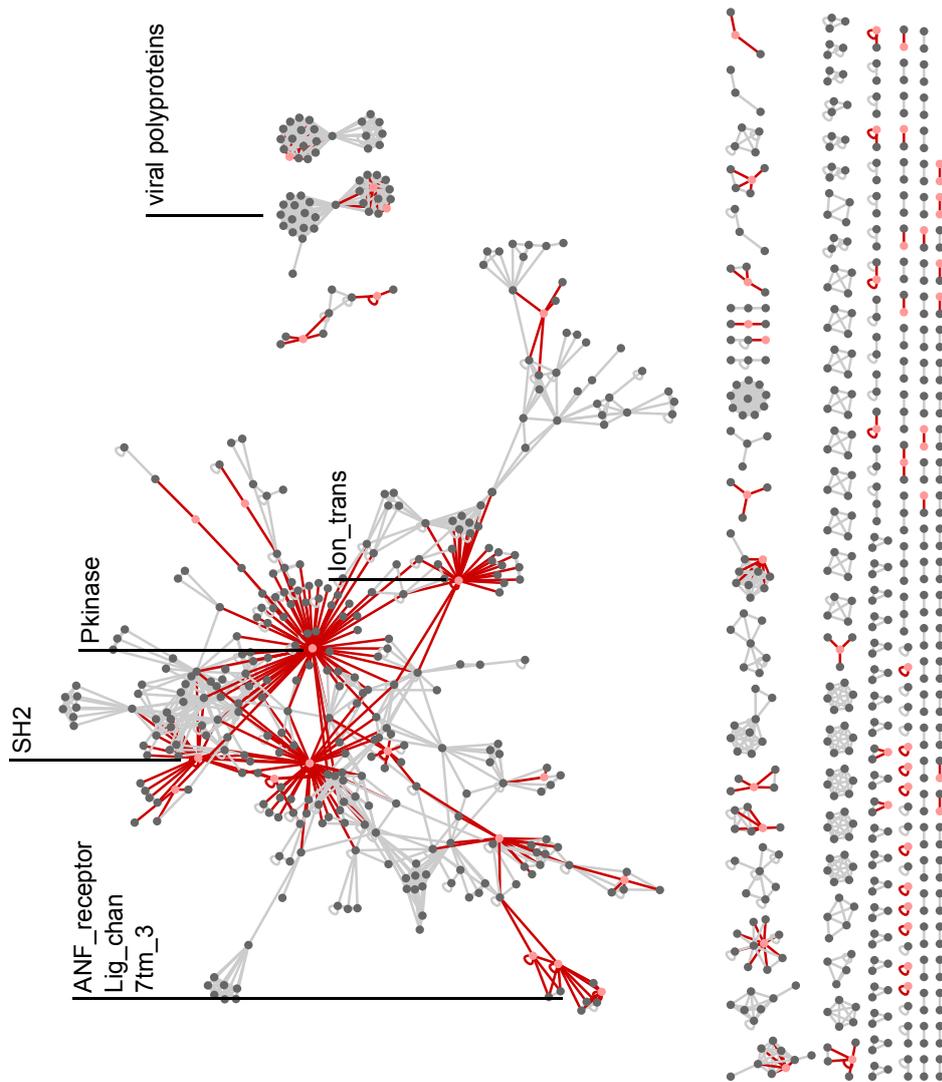


Figure 3.5: Network graph of domain co-occurrences in the ChEMBL target dictionary. Nodes represent Pfam-A domains and edges connect domains that co-occur in at least one protein with measured small molecule interactions. Nodes that form part of the catalogue of validated domains are highlighted in light-red and connecting edges in dark-red. Labels of some of the nodes mentioned in the main text are shown with vertical lines pointing towards the respective nodes or node groups.

from the initial mapping had shown that measurements with potency less than 10 μM were often associated with extraordinary assay formats such as fragment or natural product screens and protein-protein interaction data. It can also be argued that potencies in the higher μM range are often mediated by unspecific effects, such as aggregation or denaturation (McGovern et al., 2002; McGovern et al., 2003; Shoichet, 2006). The threshold of 10 μM was not strictly enforced: for example, if more than a handful of sufficiently diverse compounds exhibited activities slightly above the 10 μM threshold I retained the corresponding Pfam-A domains in the catalogue.

One of the objectives of the initial construction and later manual refinement of the catalogue was to obtain a list that can reliably be associated with small molecule binding. The exclusion of a Pfam-A domain from the catalogue does not imply that such domains cannot interact with small molecules. It rather signifies that no assays of single domain proteins in the ChEMBL database exists for this domain type. By this logic, it was inevitable that a number of Pfam-A domains that are genuinely known to interact with small molecules would be missed from the catalogue, especially such domains that do not occur in single domain proteins, as for example the `Pkinase_Tyr` domain. Owing to its significance in drug discovery, the `Pkinase_Tyr` domain had already been included in the initial mapping. To account for additional domains from this category, I assembled a list of multi-domain architectures that are frequently found in the ChEMBL target dictionary. I examined the 15 most frequent multi-domain architectures from the ChEMBL target dictionary and assessed the corresponding Pfam-A domains for their potential to bind small molecules. Table 3.4 lists the most frequent architectures and indicates components that were already included in the mapping. This list was used to prioritise Pfam-A domains for manual insertion into the catalogue of domains with known small molecule interactions.

The manual insertion of Pfam-A domains supplements the catalogue with domains that do not occur in single domain proteins. The insertion of only five domains is somewhat inconsequential, because it is certain that other domains exist that are eligible to be added to the catalogue in this way. However, this process is time-consuming and insertion of individual domains would often not contribute more than a few mapped activities per added domain. The curation interface has therefore been set up in a way that allows for incremental improvements of the catalogue.

Table 3.4: Top 15 multi-domain architectures in the ChEMBL database. Multiple occurrences of a Pfam-A domain in an architecture are indicated by (x[n]) in parentheses in the architecture column. The column ‘count’ provides the number of time an architecture occurred in ChEMBL and the column ‘mapped’ indicates whether a mapping was made using homology transfer from the catalogue of Pfam-A domains with known interactions. If no mapping was made, this column indicates ‘False’.

architecture	count	mapped
Hormone_recep, zf-C4	24	Hormone_recep
Neur_chan_LBD, Neur_chan_memb	19	Neur_chan_LBD
Pkinase_Tyr, SH2, SH3_1	18	False
Inhibitor_I29, Peptidase_C1	17	Peptidase_C1
7tm_3, ANF_receptor, NCD3G	12	ANF_receptor
ANF_receptor, Lig_chan, Lig_chan-Glu_bd	12	ANF_receptor
7tm_2, HRM	11	False
Pyr_redox, Pyr_redox_2, Pyr_redox_dim	10	False
C1_1, C1_1, C2, Pkinase, Pkinase_C	10	Pkinase
Hemopexin (x4), PG_binding_1, Peptidase_M10	9	Peptidase_M10
Pkinase (x2) Pkinase_C	8	Pkinase
AChE_tetra, COesterase	8	COesterase
ABC_membrane (x2), ABC_tran (x2)	7	False
DPPIV_N, Peptidase_S9	7	False
CARD, Peptidase_C14	7	Peptidase_C14
Neur_chan_LBD, Neur_chan_memb (x2)	6	Neur_chan_LBD
GAF (x2) PDEase_I	6	PDEase_I
Peptidase_S9, Peptidase_S9_N	6	False
Lipoxygenase, PLAT	6	False

3.3.2 Conclusions

The mapping presented in chapter 2 provided a framework for the refined mapping presented in this chapter. The purpose of the refinement was to provide a reliable resource that relates small molecule binding to Pfam-A domains. I anticipate that this resource will be useful for repurposing studies that seek to relate proteins from newly-sequenced pathogen genomes to existing drug targets by protein sequence. It may further be useful for applications in the field of proteochemometrics (PCM) that seeks to optimise predictive models of small molecule bioactivity by incorporating protein descriptors. Together with Anna Gaulton, I devised a process to make these mappings an integral part of the ChEMBL database in releases of `chembl_17` and upwards. Furthermore, the refined mapping is accessible through a stand-alone web application that provides a rich context of domain-related information on small molecule binding. This resource is decoupled from the release cycle of the ChEMBL database to prevent bottlenecks in the release preparation of either resource. The curation interface was constructed in a way that I hope will encourage participation of the chemical biology community. While I have carried out curation for the majority of conflicting mappings, the decisions I have made can be reviewed and indeed revoked by registered users. Future releases of the ChEMBL database will introduce a steady but manageable flow of additional mapping conflicts that can equally be addressed by group-internal and community curators.

3.4 Methods

3.4.1 Loading routine

The loading routine projects Pfam-A domains from the catalogue of Pfam-A domains with known small molecule interactions onto all proteins that are defined as assays targets under the scope of the mapping (see section 3.4.2). The routine parses the catalogue of Pfam-A domains with known small molecule interactions from the `valid_domains` table. It also parses all manual mappings that were made in previous curation cycles from the `manual_maps` table. The projection is applied to all assays that do not have manual mappings assigned. Measurements that already have manual mappings assigned from previous curation efforts are annotated using the `manual_maps` table. The code for

the loading routine is available at https://github.com/fak/pfam_map_loader/blob/master/loader.py

3.4.2 Scope of the mapping

The scope of the mapping was defined in a SQL query in the loader script. This query defines which measurements from the ChEMBL database are subject to the mapping. Essentially, the mapping was applied to binding and functional assays in which the target is unambiguously defined and with readouts of the types 'Ki', 'Kd', 'IC50', 'EC50', 'AC50' or logarithmic conversions thereof. A SQL query defining the scope is reproduced below:

```
SELECT DISTINCT act.activity_id
FROM activities act
JOIN assays ass
ON ass.assay_id = act.assay_id
JOIN target_dictionary td
ON ass.tid = td.tid
WHERE ass.assay_type IN('B', 'F')
AND td.target_type IN('PROTEIN COMPLEX', 'SINGLE PROTEIN')
AND act.standard_relation = '='
AND ass.relationship_type = 'D'
AND act.standard_type IN(
'Ki', 'Kd', 'IC50', 'EC50', 'AC50',
'log Ki', 'log Kd', 'log IC50', 'Log EC50', 'Log AC50'
'pKi', 'pKd', 'pIC50', 'pEC50', 'pAC50'
)
```

3.4.3 Export routine

The export routine was used to produce the two tables `manual_maps` and `valid_domains` after each round of manual curation. It is a simple script that writes the results of a SQL query for all manual mappings into a tab-separated text file. The code for this routine is available at https://github.com/fak/pfam_map_loader/blob/master/exporter.py.

3.4.4 Mapping tables

Mapping tables relate measurements of small molecule binding to Pfam-A domains. The main mapping table is called `pfam_maps` and forms part of a modified schema of the ChEMBL database. The `manual_maps` table is a subsection of the `pfam_maps` table that comprises all measurements from assays that had been manually curated. This table is produced after each curation round using the export routine.

The `pfam_maps` table contains mappings for all assays (and associated measurements) under the scope of the mapping. Mappings are defined using a combination of the `activity_id`, `component_id` and `domain_id` fields which are also found in other tables of the ChEMBL database. In addition, the `pfam_maps` table contains the following columns:

- `domain_name` - provides the domain name of the mapped domain
- `category_flag` - indicates the category of mapping
- `status_flag` - indicates the status of a mapping
- `manual_flag` - indicates whether a mapping has been manually curated
- `comment` - comment to justify curation decision
- `timestamp` - indicates the time a mapping was last changed.

The category of mapping was set to 0 if a given binding event was not mapped to more than one domain, or, in other words, if there were no conflicts with other domains. The category of mapping was set to 1 in cases where small molecule binding was mapped to multiple domains of the same type. If small molecule binding was mapped to different domain types the category of a mapping was set to 2. The `status_flag` was used as a ‘switch’ that indicates whether a mapping is valid or not. This flag was set to 1 for all conflicting mappings and to 0 for all mappings that did not conflict with other mappings or where conflicts had been manually resolved. The `manual_flag` was set to 0 by default, but changed to 1 for all mappings associated with an assay that had been manually curated. The `manual_maps` table is simply a subsection of the `pfam_maps` table including all rows where `manual_flag` is equal to 1.

3.4.5 Catalogue of Pfam-A domains with known small molecule interactions

The catalogue of Pfam-A domains with known small molecule interactions was stored in the `valid_domains` table. This table contains the names of all Pfam-A domains for which evidence for small molecule binding was found either from measurements of single domain proteins in the ChEMBL database or from the literature. Pfam-A domains that were removed from the catalogue through manual curation as described in section 3.2.3 were retained in this table, but labelled as invalid. The table consists of four columns:

- `domain_name` - provides the domain name of the mapped domain
- `removed_flag` - indicates whether a domain has been removed from the catalogue manually
- `comment` - comment provides pointer to literature evidence for small molecule binding for a given domain
- `timestamp` - indicates the time a mapping was last changed.

3.4.6 Prototype of a curation platform

The curation platform was implemented using the Django web application framework. The application consists of four parts:

- a data model that describes the schema of the database backend
- a number of views that contain functions to retrieve and arrange content from the database
- a URL management system that relates input URLs to views
- a number of HTML templates to display content generated by the views functions.

Details of the architecture for Django web applications are documented elsewhere (e.g. <http://djangobook.com>) and here, I describe only aspects that are specific to the implementation of the curation platform. The data model was created from a MySQL database instance hosting the `chembl_15` release of the ChEMBL database using the

`django-admin.py syncdb` command. The data model can be applied to future releases of the ChEMBL databases even though future schema changes might require slight adjustments. The data model is available at https://github.com/fak/pfam_maps/blob/master/chembl_15/models.py.

Views were programmed to generate dynamic content in response to requests generated at the UI. In most cases, user requests trigger the execution of a database query and subsequent processing of the query results. The processed results are forwarded to be rendered in the corresponding HTML template. The python functions for these views can be accessed at https://github.com/fak/pfam_maps/blob/master/chembl_15/views.py and some supporting functions at https://github.com/fak/pfam_maps/blob/master/chembl_15/helper.py.

The URL management system directs requests generated at the UI to functions defined in the Views section. All content generated by these functions is then routed to HTML templates via the URL management system. The code for this system is available at https://github.com/fak/pfam_maps/blob/master/chembl_15/urls.py.

The html templates are written in HTML and the Django template language. The formatting and CSS styles of the html documents are based on the EBI main website. User input is fed back to the functions defined in the views section using the POST protocol. All templates can be accessed at https://github.com/fak/pfam_maps/tree/master/chembl_15/templates/chembl_15.

Evidence for small molecule binding in the `evidence` pages is presented dynamically using the plotting library `d3.js` (Bostock et al., 2011). The javascript code used to generate these plots is available at https://github.com/fak/pfam_maps/blob/master/chembl_15/static/chembl_15/js/scatterplot.js.

Protein domain structures in the `conflict_details` section are visualised using a dedicated library `domain_graphics.js` provided by the Pfam webserver at http://pfam.sanger.ac.uk/static/javascripts/domain_graphics.js.

3.4.7 Standard procedure of manual curation

The standard procedure of manual curation consisted of up to four steps. These steps were carried out iteratively and the process was stopped as soon as a step yielded sufficient information to reach a curation decision. The items were inspected in the following

order: (1) assay description, (2) publication title, (3) publication abstract, (4) publication full-text. I searched any of these four items for keywords the ‘allosteric’, ‘orthosteric’, ‘competitive’ and ‘non-competitive’. Frequently, the site of binding could also be derived indirectly from the lead structure or comparisons of the evaluated compounds with existing, well-documented chemical probes. In cases where none of the above items contained sufficient information to reach a curation decision, no decision was made and the assay was retained for a future curation round.

3.4.8 Coverage and network view

The coverage κ_{meas} of all relevant measurements in the ChEMBL database was determined as the fraction :

$$\kappa_{\text{meas}} = \frac{n_{\text{meas}}(\textit{mapped})}{n_{\text{meas}}(\textit{total})} \quad (3.1)$$

In this fraction, $n_{\text{meas}}(\textit{total})$ corresponds to the total number of measurements from the ChEMBL database that represent small molecule-protein interactions. Of those, $n_{\text{meas}}(\textit{mapped})$ indicates the number of measurements that were mapped to one or more protein domains.

Another measure of the coverage, κ_{prot} , was determined as:

$$\kappa_{\text{prot}} = \frac{n_{\text{prot}}(\textit{mapped})}{n_{\text{prot}}(\textit{total})} \quad (3.2)$$

Here, $n_{\text{prot}}(\textit{mapped})$ corresponds to the number of proteins containing one or more domain from the catalogue of domains with known small molecule interactions and $n_{\text{prot}}(\textit{total})$ represents the total number of proteins for which small molecule binding was measured. Measurements that represent small molecule binding events were obtained using the following SQL statement below. This statements selects from all activities in ChEMBL those that have a relevant activity type and sufficient potency in binding or functional assays.

```
SELECT DISTINCT dc.tid, COUNT(DISTINCT activity_id)
  FROM assays ass
  JOIN(
    SELECT td.tid, td.target_type, COUNT(cd.domain_id) as dc
    FROM target_dictionary td
    JOIN target_components tc
      ON tc.tid = td.tid
    JOIN component_sequences cs
      ON cs.component_id = tc.component_id
    JOIN component_domains cd
      ON cd.component_id = cs.component_id
    WHERE td.target_type IN('SINGLE PROTEIN', 'PROTEIN COMPLEX')
    GROUP BY td.tid
  ) as dc
  ON dc.tid = ass.tid
  JOIN activities act
    ON act.assay_id = ass.assay_id
  WHERE act.standard_type IN('Ki', 'Kd', 'IC50', 'EC50', 'AC50')
  AND ass.relationship_type = 'D'
  AND assay_type IN('B', 'F')
  AND act.standard_relation IN('=')
  AND standard_units = 'nM'
  AND standard_value <= 50000
  GROUP BY dc.tid ORDER BY COUNT(activity_id)
```

Using the target identifiers obtained in this query, a network of domain co-occurrences was constructed by the enumeration of all pair-wise domain combinations. This list of domain pairings was imported and visualised using the Cytoscape software (version

2.8, Shannon et al., 2003). Features of the Cytoscape software were also used to assemble the network statistics presented in section 3.3.1. The script that was used to query and process the measured small molecule-protein interactions is available at <https://github.com/fak/mapChEMBLPfam/blob/master/analysis.py>.

Chapter 4

Integration of small molecule potency measurements with the phylogeny of their protein targets

4.1 Introduction

The introduction to this Chapter is split in two parts. Section [4.1.1](#) lays out the principles of phylogenetic relationships and the protein homology concept and highlights its relevance in a drug discovery context. Section [4.1.2](#) describes the rat as a model organism in drug discovery and discusses the evolutionary relationship to humans. It also presents a brief overview of physiological and molecular similarities and differences between the two species. Section [4.1.3](#) provides an outline of the integration work and analysis I have carried out in this context for the completion of my Ph.D. thesis.

4.1.1 Phylogenetic relationship between drug targets

Pharmacological intervention relies on the interaction of a chemical with molecular components of the body. To our current understanding of pharmacology, most of these molecular components are proteins, with some exceptions involving nuclear DNA, extracellular sugar-polymers and lipids. When examining a specific protein as a target of pharmacological intervention, it is useful to view the protein in its evolutionary context. In section 1.4.1, I introduced the concept of homology, which applies to genes and proteins sharing a common ancestor. Proteins that are related through gene duplication are called paralogs, while proteins that are related through a speciation event are called orthologs. Figure 1.2 provides an overview of these relationships. It was deduced from first principles that gene duplication enables fast functional divergence on the grounds that one paralog can retain the function of the ancestor. Orthologs on the other hand should have conserved function, as a loss of the function of the ancestor would likely have a detrimental effect on fitness. This is also known as the ‘ortholog conjecture’ (Conant and Wolfe, 2008; Studer and Robinson-Rechavi, 2009).

Many aspects of modern drug development rely on the principles laid out by the ortholog conjecture. The use of animal models is an extrapolation from the model species to man and in most cases implicitly assumes that the protein target of a pharmacological intervention serves a similar function in both species and is indeed susceptible to similar chemical perturbations. A number of studies show that phenotypic responses to small molecule perturbation can drastically differ between species. For example, reduced growth of yeast colonies and abnormal angiogenesis in mice would be corresponding ‘phenologous’ responses to Lovastatin administration (McGary et al., 2010; Cha et al., 2012). Both the traditional approach of direct extrapolation between species as well as this less established approach of extrapolation through modules of ‘phenologs’, rely on functional conservation of underlying molecular components. Given its importance in the drug discovery process, there is a surprising lack of systematic studies to support the ortholog conjecture in the context of small molecule perturbation. Following the molecular biology revolution that enabled cloning and over-expression of drug targets in cellular systems in the 1990s, a number of studies have been published that compare small molecule binding between orthologous proteins. These studies are of relatively small scale and focus on differences between species (mainly human and rat or mouse)

observed for individual orthologous pairs and relatively small numbers of ligands (for example [Lovenberg et al., 2000](#); [Barker et al., 1994](#); [Erion et al., 2005](#)).

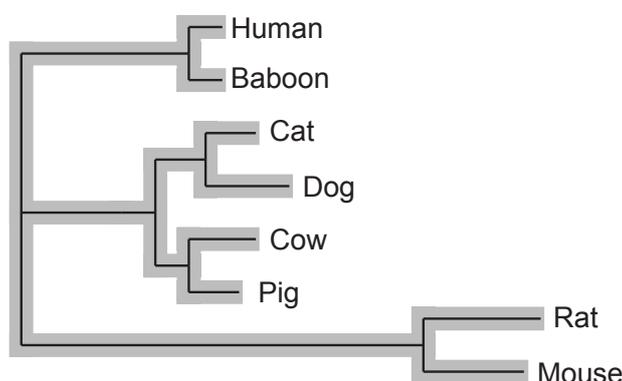
Another hallmark of modern drug discovery is the concept of selective perturbation, which seeks to reduce unwanted side effects through the development of ligands that interact exclusively with a desired receptor. Ligand selectivity for a given protein is most often probed against proteins that are linked through a common ancestor and thus form a family of paralogous proteins. Knowledge of the within-family selectivity is of vital importance for mode-of-action studies; the discovery of the vanilloid-receptor-like protein 3, for example, provided an improved model of thermal hyper-sensitivity at a time where drugs were under development for its paralog, the vanilloid receptor ([Xu et al., 2002](#)). The notion that drug action in many cases is mediated through interaction with multiple rather than one specific protein target has led to the rise of the polypharmacology concept ([Roth et al., 2004](#); [Paolini et al., 2006](#); [Yildirim et al., 2007](#)). It states that an ‘unselective’ pharmacological profile may be an advantageous property for a drug candidate. Targeted polypharmacology is difficult to attain in a practical sense ([Morphy and Rankovic, 2005](#); [Hopkins et al., 2006](#); [Morphy and Rankovic, 2007](#)); so far, it has been applied to ‘design’ inhibitors with multiple activities against (paralogous) members of the kinase family ([Knight et al., 2010](#)). Both the traditional approach of selective perturbation as well as the polypharmacology approach, rely on the notion that selectivity between paralogous proteins can be attained in a targeted manner. Large and medium scale studies have been conducted that compare small molecule potency between paralogs across the kinome ([Metz et al., 2011](#)) or, using literature data from ChEMBL, the family of G-protein coupled receptors ([Lin et al., 2013](#)).

The recent availability of public small molecule bioactivity data means it is possible to compare small molecule binding to orthologs and paralogs systematically and on a large scale. In this chapter of my thesis I explore an approach by which I examine directly the interaction of phylogenetic relation and small molecule potency.

4.1.2 Rats as model organisms in drug discovery

Rats are among the most popular organisms to study human disease. Rats, together with mice, guinea pigs, and others, form the large mammalian order of *Rodentia*. The characteristics shared by many species in this order are considered advantageous for

Figure 4.1: Tree of evolutionary distances. Figure is adapted from Cooper et al., 2003. Branch lengths approximately scale with the frequency of substitution averaged over four sites in the CFTR and BRCA1 genes and highlight the acceleration of non-synonymous substitution in the two rodent species rat and mouse.



model organisms and include unproblematic husbandry, short generation times and a relatively low cost of maintenance. Rats have long been preferred over mice in biomedical research because they are less aggressive and physiological parameters, such as heart rate and renal clearance, are generally closer to those in humans (Gill et al., 1989). Of particular relevance to pharmaceutical research, a higher blood volume facilitates dosage and administration of trial substances. Mice have traditionally been the model of choice for geneticists, owing to the relative ease of genetic manipulation and the possibility of cloning animals from embryonic stem cells (Wakayama et al., 1999; Zambrowicz and Sands, 2003). Genetic manipulation has proven more difficult in rats, where cloning is reliant on the somatic nuclear transfer methodology (Zhou et al., 2003) or lentiviral transduction (Hamra et al., 2002; Ryu et al., 2007); both methodologies lack the precision and efficiency of cloning from embryonic stem cells. The sequencing of the rat genome in 2004 (Gibbs et al., 2004), induction of embryonic stem cells and targeted embryonic knock-outs in 2009 (Li et al., 2009; Geurts et al., 2009) are likely to further the use of rats as model organisms in biomedical research.

Humans and rats share their mammalian origin and the most recent common ancestor dates back to about 91m years (Hedges et al., 2006). Since their divergence, the non-synonymous substitution rate in mice and rats has increased, while that of humans has decreased, leaving the respective rates at a roughly three-to-one ratio (Cooper et al., 2003). A phylogenetic tree representing the relationships between rats and a number of other mammalian species is shown in Figure 4.1. The number of one-to-one orthologs (of protein-coding genes) between the human and rat genome is estimated around 10,000 with an average sequence identity of 88.3% (Gibbs et al., 2004). A study that compared

the G-protein coupled receptor repertoires of the human and mouse genome found that humans and mice have 343 receptors of endogenous ligands in common. This corresponds to almost the entire complement of this type of receptor in the human and mouse genomes, where total numbers of these receptors are 367 and 392, respectively (Vassilatis et al., 2003). A study comparing the repertoire of kinases in human and mice identified 501 kinases with one-to-one orthologous relationships (Caenepeel et al., 2004). Taken together, these findings suggest that, at least with regard to some major classes of drug targets, orthologous relationships between the human and rat genome are relatively straightforward. Naturally, the extrapolation from pharmacological studies in rats to treatment outcomes in humans requires comparison on multiple levels that go well beyond the conservation of orthologous relationships. Gene expression in response to drug administration was found to be at least partly conserved for acetaminophen at liver-toxic doses (Kienhuis et al., 2009). A more recent study of transcriptional modules responding to multiple stimuli found only 15% of transcriptional modules were conserved between humans and rats (Iskar et al., 2013) - even though studies in unperturbed organisms had found higher conservation (Su et al., 2002; Liao and Zhang, 2006; Brawand et al., 2011). Besides well-documented differences in cytochrome P450-mediated metabolism (Guengerich et al., 1986; Kobayashi et al., 2003) and pharmacokinetic parameters such as renal clearance and the volume of distribution (Ward and Smith, 2004a; Ward and Smith, 2004b), there are also concerns that the rearing of rats in laboratory conditions may alter their natural metabolic state (Martin et al., 2010). Thus, with increasing complexity of the level of comparison it appears that extrapolation between rats and humans becomes more problematic. In my thesis work, I evaluated extrapolation from rats to humans at the lower intermediate level of *in vitro* drug response. The work has important implications for the use of rats in pharmaceutical research as it helps pinpoint where, on the scale of increasingly complex inter-species comparisons, we should start factoring in differences between humans and rats.

4.1.3 Outline

This chapter describes an approach of integrating small molecule bioactivity data from the ChEMBL database and homology information from the EnsemblCompara Genetrees pipeline. The approach was used to examine the link between phylogenetic relationship

and the conservation of compound potencies between pairs of proteins retrieved from the ChEMBL target dictionary. The scope of the study was not limited to a particular family of proteins, but rather ranged across all protein families. The relationships that were examined in this study are orthology between human and rat proteins and paralogy between human proteins. I chose to examine orthologs between proteins from humans and rats because of the importance of rats as model organisms in drug discovery. The ChEMBL database is rich in potency data measured against proteins from rats, reflecting their prominent role in preclinical development and ADMET¹ testing. While in theory the approach presented here could be extended, the availability of potency data for other species is limited and would not allow to draw conclusions with the same confidence. Table 4.1 summarises activity counts for the top six organisms. Potency data for human proteins is abundant in the ChEMBL database, as would be expected from a database that extracts its core content from the medicinal chemistry literature. The comparison of small molecule potencies across paralogous proteins is important in drug discovery because its pharmacological profile across related proteins gives insights into the selectivity of a compound and potential off-target effects.

Table 4.1: Summarized above are unique activities per species. Only activities of type ‘B’ that could be mapped unambiguously to a Uniprot accession were counted.

Organism	binding activities
<i>H. sapiens</i>	157,145
<i>R. norvegicus</i>	23,278
<i>M. musculus</i>	6,567
HIV 1	6,361
<i>B. taurus</i>	2,750
<i>C. porcellus</i>	2,190
<i>C. elegans</i>	1,772
<i>S. scrofa</i>	1,535
HCV	1,017
<i>O. cuniculus</i>	968

¹ADMET is an acronym for absorption, distribution, metabolism, excretion and toxicity.

4.2 Results and Discussion

4.2.1 Controlling for assay variability in the ChEMBL data set

ChEMBL is a collection of bioactivity measurements extracted from the relevant literature published during the last 20 years. Hence, I started this project with the expectation that differences in experimental set-up as well as measurement and reporting errors would result in considerable noise of potencies recorded in ChEMBL. It was therefore desirable to obtain a means to control for this noise when analysing potency differences between human-rat orthologs and human paralogs. As an estimate of the inter-assay variability, I extracted potency measurements from the ChEMBL database that had been reported for the same compound and target, but in different assays. I retrieved measurements for both human and rat proteins. As described in the Methods sections 4.4.1 and 4.4.2, the query was limited to activities of the type IC50, EC50, K_i and the respective logarithmic conversions of these activity types. In the original implementation of this project (Kruger and Overington, 2012), values obtained from all measurement types were compared directly, without additional conversion between IC50 and K_i . This was a simplification, as the relationship of IC50 and K_i depends on assay parameters. Frequently, it can be evaluated as

$$K_i = \frac{IC50}{1 + \frac{[S]}{K_m}}, \quad (4.1)$$

where K_m is the enzyme specific Michaelis-Menten constant (Michaelis and Menten, 1913; Michaelis et al., 2011) and $[S]$ the substrate concentration. However, for potency measurements in ChEMBL, the assay parameters K_m and $[S]$ are rarely available and thus no attempt was made to convert between the two activity types. Recently, Kalliokoski and colleagues published a study that suggests a conversion factor of 2.3 applied to K_i measurements (corresponding to a subtraction of 0.355 log units) when mixing data of the types K_i and IC50 (Kalliokoski et al., 2013). I followed this guideline for the preparation of my thesis by adjusting all K_i by the offset of 0.355 log units. After processing, I constructed a distribution of inter-assay differences from pairwise comparisons. To prevent comparison of measurements that were duplicates in the database, I excluded all paired measurements that were exactly equal.

First, I examined the resulting distributions for each species separately, as shown in Figure 4.2A. This comparison showed that the fraction of highly similar measurements was higher for rat proteins compared to human proteins. One factor that contributes to this apparent difference is the composition of the data in terms of target classes. As shown in Figure 4.2C, different target classes exhibit different degrees of between-assay variability. One possible explanation for this observation lies in the specific assay formats used for targets in different classes, some of which have higher fidelity than others. Thus, the composition in terms of target class influences the overall shape of the distribution for each species. For example, the proportion of transporter data, which is in general more similar between measurements, is higher for rat compared to human. However, some differences persisted even after target classes had been taken into account. To compensate for the observed differences when estimating the noise between assays, 1,500 paired measurements were picked at random from each species. Thus, the number of data points in the ortholog analysis was approximately matched (see section 4.2.2). Samples were picked only once, but to make sure the sample was representative, 1,000 additional samples were taken as described in the Methods section 4.4.2 and 50 of them visualised in Figure 4.15. This procedure allowed to determine standard errors of distribution parameters. As representative parameters, I examined quantiles that would correspond to the intervals covered by one and two standard deviations in a normal distribution (68.4% and 95.5% of all data). They are reported in Table 4.2. The low amplitude of associated standard errors shows that, at a sample size of 3,000 paired measurements, resulting distributions are representative of the underlying data.

A graphical summary of the distribution that was sampled for the remaining analysis

Table 4.2: Quantile estimates of 1,000 sampled distributions of inter-assay differences. The column ‘quantile’ indicates the distribution quantile, the column ‘mean value’ the quantile value averaged over 1,000 sampled distributions and the column ‘std. error’ the standard error of the parameter value across 1,000 sampled distributions.

quantile	mean value	std. error
2.3%	-2.28	0.057
15.9%	-0.84	0.023
median (50%)	-0.01	0.007
84.1%	0.84	0.025
97.7%	2.14	0.054

is provided in Figure 4.3. The distribution of differences between measurements is symmetric around the origin. Notably, the shape of the distribution peak suggests that the distribution can be described using the model of a Laplace distribution. A good fit of a Laplace distribution was obtained with the parameters $\mu = 0.0$ and $b = 0.75$ for location and scale, respectively (see also section 4.2.4 and Figure 4.7). The selection and filtering of potency measurements was designed to be consistent with human-rat orthologs. Indeed, the distribution of observed potency differences between orthologs can also be modelled using a Laplace distribution (see section 4.2.7). This is unexpected as under the central limit theorem, the mean of a large number of independently distributed error variables should tend towards a normal distribution, under the provision that the variance of the error is finite. However, a normal distribution might underestimate the variance in cases of unexpected error (Shlyakhter, 1994) and in such cases a double-exponential (Laplace) distribution is a better model. In the given data, unexpected errors might be due to the influence of unit transcription errors² and in cases where comparisons of IC50 and K_i data are particularly unfavourable. Unit transcription errors are visible in the provided probability density distribution as small peaks where the difference equals ± 3 pIC50 units. These errors have a strong effect on the variance of the distribution because they frequently change the observed value by three or more orders of magnitude.

Importantly, the non-normal distribution of the data required the use of non-parametric methods in subsequent analysis steps. As a numeric measure of the correlation between potencies measured for combinations of identical compounds and targets in different assays, the Spearman rank correlation coefficient was determined as $\rho = 0.74$. More informative than Spearman's correlation is a difference plot as shown in Figure 4.3B, following specifications by Bland and Altman (Bland and Altman, 1986). The quantile intervals that would correspond to the intervals covered by one and two standard deviations in a normal distribution (68.4% and 95.5% of all data) were determined by bootstrapping with one thousand iterations and reported together with the standard error as $-0.88 \pm 0.03 / 0.81 \pm 0.03$ pIC50 units and $-2.20 \pm 0.8 / 2.13 \pm 0.08$ pIC50 units, respectively. These values are global averages that may not reflect accurately the measurement error for individual pairs of values. However, they demonstrate that deviations between

²Unit transcription errors occur when, at any point in the process of transferring measured potencies into the database, a measurement unit is incorrectly reported, for example μM instead of nM . These errors often lead to characteristic offsets between measured and reported values in three orders of magnitude.

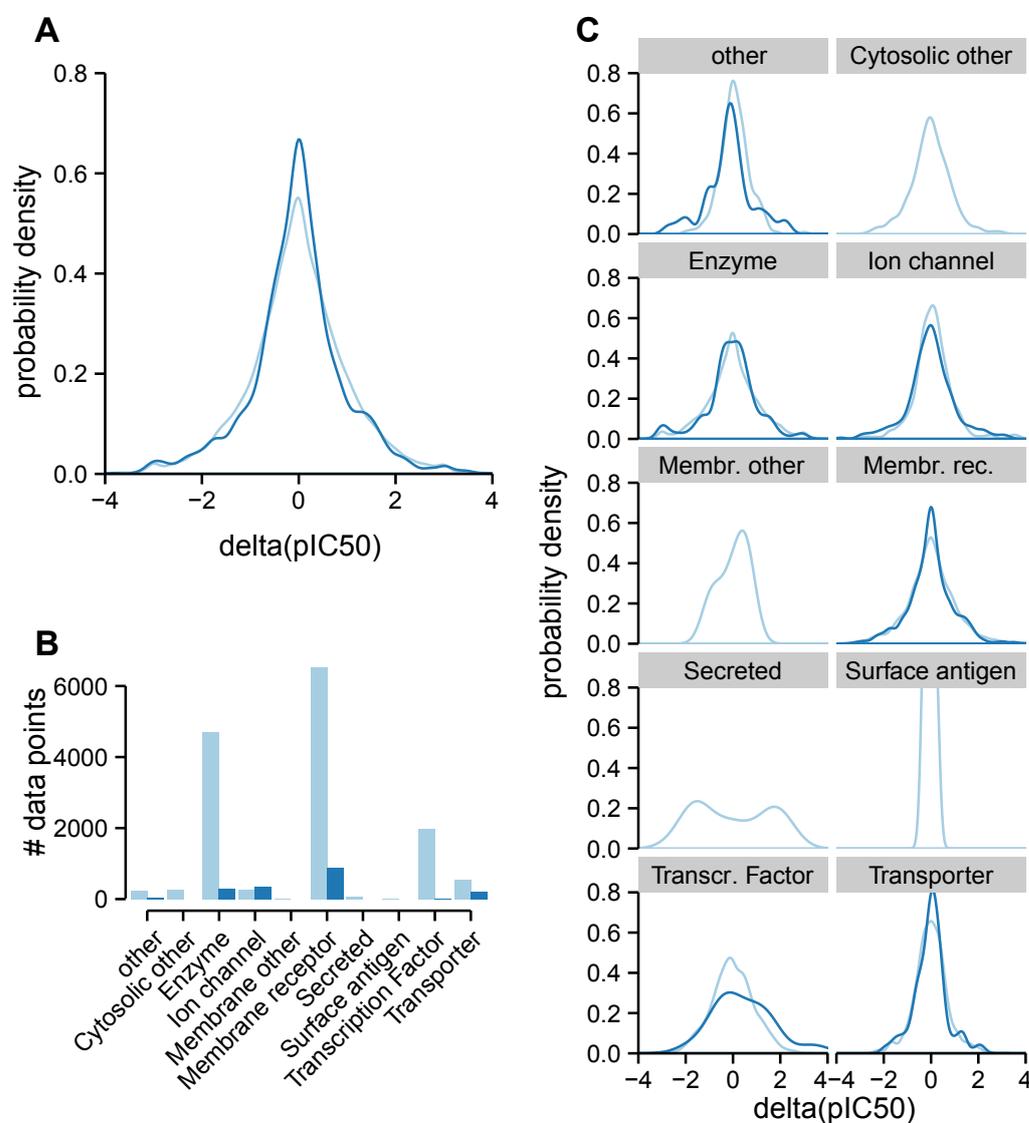


Figure 4.2: Overview of inter-assay variability for human and rat proteins. Panel A shows the combined data for potency measurements against human (light-blue) and rat targets (dark-blue). Panel B summarises the number of paired measurements obtained for each species in different target categories. Panel C summarises inter-assay variability for individual classes of drug targets. Distributions for human proteins are outlined in light-blue and for rat proteins in dark-blue. Differences between target classes may be caused by different assay formats (e.g. transcription factors versus transporters). Distributions for some classes are also skewed by small sample numbers, e.g. secreted proteins.

reported measurements in the ChEMBL database can be considerable, and in about a third of the cases assessed in this study exceed one order of magnitude.

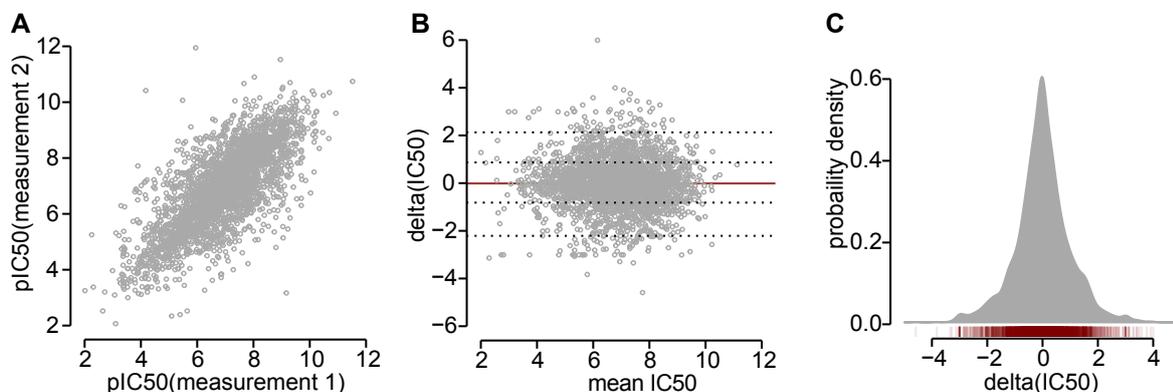


Figure 4.3: Overview of inter-assay variability in ChEMBL. Panel A is a scatter plot of the affinities observed when comparing results from different assays of the same compound and target (1,500 human targets and 1,500 rat targets) and the probability density distribution of differences between potencies observed in different assays assessing identical compounds and proteins. Panel B shows the same data transformed in a Bland-Altman plot. The x-axis represents the mean of two measurements, the y-axis the difference between corresponding measurements. Panel C shows the probability density distribution of differences between potencies measured against human and rat proteins. Contributions from individual compounds are represented as red tick marks at the bottom of the panel.

4.2.2 Conservation of potency between human-rat orthologs

With a baseline established for experimental and procedural noise, I wanted to assess potency differences between human-rat orthologs. To this end, pairs of human-rat orthologs were determined using annotation from the EnsemblCompara Genetrees pipeline (Vilella et al., 2009), version 62. The Uniprot identifiers for these pairs were then used to query the ChEMBL database for potency measurements for compounds with reported activities against both a human protein and its ortholog in rat (see Methods section 4.4.1 and 4.4.3). In total, 2,782 pairs of measured potencies were retrieved from the ChEMBL database. To be consistent with the processing of differences between assays, all potencies that were exactly equal were excluded from the analysis, leaving 2,669 paired potencies. These broke down into combinations of 2,359 compounds and 151

pairs of human-rat orthologs. The mapping between orthologs in this data set generally followed a one-to-one pattern, meaning that for each human protein there was only one rat protein in the data set and vice versa. Most of the proteins in this set are membrane receptors or kinases, but a spectrum of other target classes is also present. A summary of target classes represented in this set is presented in Table 4.3. With the number of drug targets estimated at 324 (Overington et al., 2006), this study does not cover the complete set of drug targets, but a significant subset with a relatively unbiased distribution across target classes. In analogy to the data for inter-assay variability, a scatterplot of the raw data, a difference plot and a probability density distribution of the observed differences are shown in Figure 4.4. The observed bioactivities expressed as log transformed potency values pIC₅₀ range from 4 to 10 (i.e. across a broad range from single digit μ M to high pM). The scatter of paired potencies roughly follows a straight line through the origin with a unit slope. This finding was in line with expectations that potencies in the two species should be correlated directly without any scaling effects. The distribution of differences is roughly symmetric and centred around zero. The probability density distribution displays a slight increase at a potency difference of about 0.5 pIC₅₀ units between human and rat proteins. This increase reflects mainly potency measurements for about 200 compounds against orthologs of the Histamine H₃ receptor (HRH₃). This aspect is discussed in more detail in section 4.2.7. The median of the observed differences was determined by bootstrapping with one thousand resampling iterations as 0.08 ± 0.02 (std. error). The subtle shift towards higher potencies for human targets may be caused by random fluctuations, but more likely reflects the human-centric nature of the drug discovery process. This is discussed further in section 4.2.7. In either case, the species bias within the data set is small if compared to the underlying noise. The quantile intervals that would correspond to one and two standard deviations in a normal distribution are $-0.65 \pm 0.02 / 0.88 \pm 0.02$ pIC₅₀ units and $-2.30 \pm 0.11 / 1.96 \pm 0.09$ pIC₅₀ units, respectively. In absolute terms, the differences observed between these two species are thus substantial and in about a third of all cases exceed one order of magnitude. However, as described in section 4.2.1, differences between assays are in a similar range and differences between the two species should be set in relation to this. Section 4.2.5 provides a detailed comparison of the distributions of potency differences between assays, orthologs and paralogs.

Table 4.3: Target classes represented in the ortholog data set.

Target class	count	compounds
Membrane receptor	76	1,504
Enzyme	48	501
Ion Channel	12	250
Transporter	7	366
Transcription factor	4	34
Undefined	3	8
Membrane other	1	5
Cytosolic other	1	1

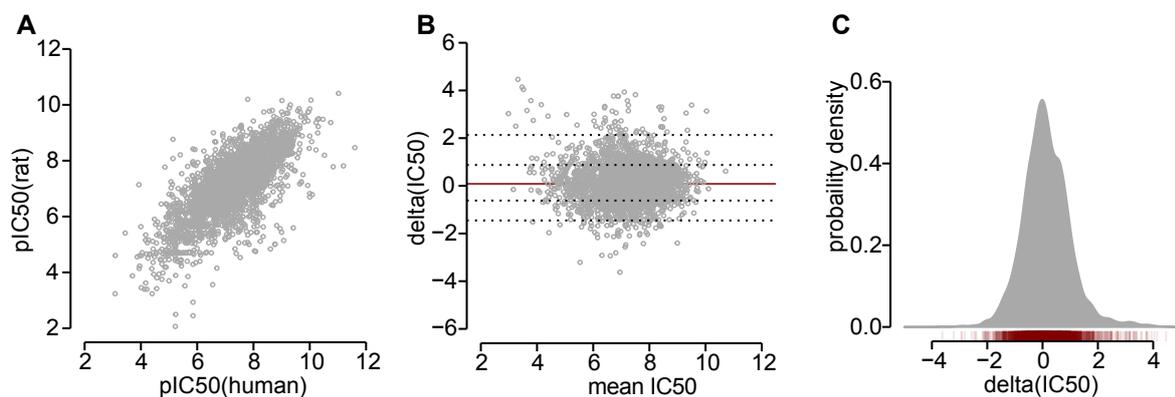


Figure 4.4: Overview of variability between human and rat orthologs in ChEMBL. Panel A shows a scatter plot of the affinities observed when comparing results from measurements against human proteins (x-axis) against rat proteins (y-axis). Panel B shows the same data transformed in a Bland-Altman plot. The x-axis represents the mean of two measurements, the y-axis the difference between corresponding measurements. Panel C shows the probability density distribution of differences between potencies measured against human and rat proteins. Contributions from individual compounds are represented as red tick marks at the bottom of the panel.

4.2.3 Conservation of potency between human paralogs

As for human-rat orthologs, I wanted to assess potency differences between proteins whose lineage traces back to a duplication event and which are thus related through paralogy. To this end, I aimed to selected pairs of measured potencies and carry out the analysis according to the same criteria as done for human-rat orthologs.

As discussed in section 4.1.3, the number of measurements available for rat proteins is only a fraction of the number of measurements available for human targets. The analysis of human-rat orthologous pairs was viable because the rarer measurements of rat proteins were combined with abundant measurements against human proteins. Due to concerns about availability of data for the rat, this part of the analysis focuses on proteins related through paralogy within the human genome.

To proceed with the analysis of paralogs in analogy to the analysis of human-rat orthologs, I obtained paralogous relationship assignments from the EnsemblCompara Genetrees pipeline (version 62, 2011). In absence of an intrinsic rule to determine which of two paralogous proteins to use as a reference and which as its paralog, proteins from each pair were randomly assigned to either of the two arbitrarily named groups ‘reference’ or ‘paralog’. Potency differences were then calculated as for orthologs. Through this procedure, the arithmetic signs of potency differences were preserved for all activities measured against a specific paralogous pair. Using the query and filtering procedures described in the Methods section 4.4.1 and 4.4.3, I retrieved 41,733 pairs of measurements from ChEMBL. After filtering out potencies that are exactly equal or differ by more than 20 pIC₅₀ units, 41,454 paired measurements remained for analysis. Thus, data available for human paralogs was much more abundant compared human-rat orthologs. The data separated into combinations of 648 pairs of human paralogs with 20,219 compounds. These 648 paralogous pairs were made up of combinations of 516 unique proteins.

I obtained a much larger number of paired measurements for human paralogs than for human-to-rat orthologs. This was due both to the human-centric nature of the ChEMBL database, as well as a combinatorial effect: Protein coding genes often map to multiple paralogs (one-to-many relationship) while orthologs typically only map to one or very few orthologs in the rat genome. This could increase drastically the number of possible permutations between paralogs. The theoretical upper limit θ_{\max} of one-to-many relations in a hypothetical data set where all proteins are paralogs depends on the number of

unique genes n_u ,

$$\theta_{\max} = \frac{n_u!}{2(n_u - 2)!}, \quad (4.2)$$

while the lower limit θ_{\min} , which would be observed in a data set with only one-to-one relationships, is simply half the number of all unique genes $\theta_{\min} = \frac{n_u}{2}$. In the set of human-rat orthologs with shared small molecule potency measurements, the observed number of combinations was found to be exactly equal to the lower limit and thus all relationships in this set are one-to-one mappings. For human paralogs, the number of observed combinations was 648 and thus greater than the lower limit of 258 for a set of 516 unique genes, but only a small fraction of the theoretical upper limit of 132,870.

To probe in more detail the homology constraints imposed by EnsemblCompara as well as the contributions of one-to-one and one-to-many relationships in the paralog set, I constructed a network of these relations. The graph of homologous relations between proteins in the paralog set was visualised using the Cytoscape software (Shannon et al., 2003) and is shown in Figure 4.5. Of the 166 connected components in this network, 78 represent one-to-one relationships and 39 represent triplets. The 50 remaining components represent one-to-many relations of between four and eight paralogous proteins. Thus, the graph is characterised by small isolated clusters. In comparison, classic polypharmacology networks (Paolini et al., 2006; Chen et al., 2009), where vertices are assigned for shared compounds without the additional homology constraint, graphs are typically dominated by a densely connected giant component. This is an indication that the EnsemblCompara GeneTrees pipeline works with a conservative definition of paralogy and assigns relationships only between close members of a gene family. The implication for results presented in this section is that the potency differences presented here are limited to closely related paralogs and do not include differences between more widely related proteins.

The distribution of observed potency differences is summarised in Figure 4.6. As for orthologs and the between-assay comparison, the distribution of differences is approximately symmetric and centred around zero. The exact median was calculated as -0.07 ± 0.006 (std. error). The deviation in symmetry between the ‘reference’ and ‘paralog’ groups was likely the result of random fluctuation. The quantile intervals that would correspond to one and two standard deviations in a normal distribution were $1.26 \pm 0.01 / -1.48 \pm 0.01$ pIC50 units and $2.82 \pm 0.02 / -3.01 \pm 0.02$ pIC50 units, respectively.

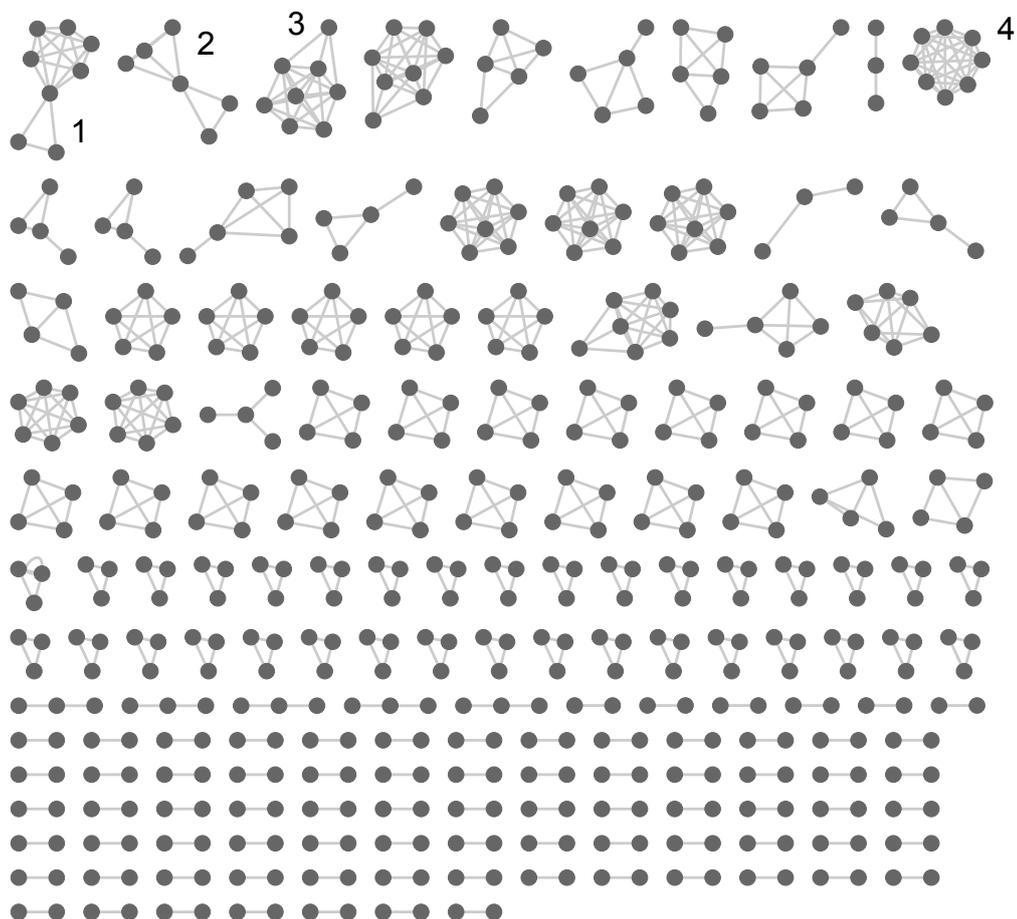


Figure 4.5: Graph of relationships within the paralog data set. Nodes represent proteins that are connected by edges if they are related through paralogy and have measured potencies for at least one compound in common. (1) denotes a group of muscarinic ACh receptors and histamine H receptors. The most highly connected node represents the the HRH1. (2) denotes a group of tyrosine kinases and the central node represent tyrosine-protein kinase BTK. (3) denotes a group of heterogeneous aminergic GPCRs, including adrenoceptors, dopamine receptors and serotonin receptors. (4) is a group of eicosanoid receptors. Table 2 in the appendix section lists individual components for each highlighted group.

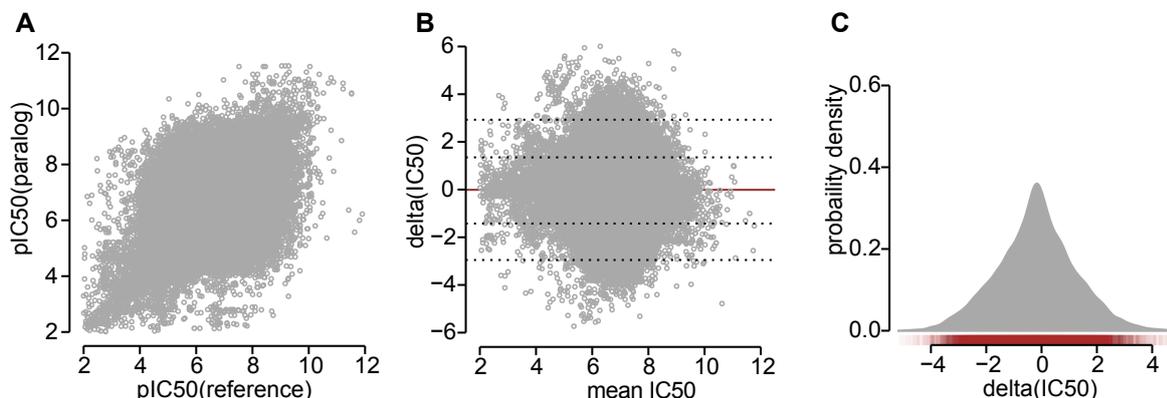


Figure 4.6: Overview of potency differences between human paralogs. Panel A shows a scatterplot where each dot represents potency measured against a ‘reference’ protein on the x-axis and potency measured against respective paralogs are on the y-axis. Panel B shows the same data transformed in a Bland-Altman plot. The x-axis represents the mean of two measurements, the y-axis the difference between corresponding measurements. Panel C shows the probability density distribution of differences between potencies measured against pairs of human paralogs. Contributions from individual compounds are represented as red tick marks at the bottom of the panel.

The deviation from symmetry of these intervals around zero could be approximately corrected for by centring the intervals around the calculated mean of -0.07 . Thus, compared to human-rat orthologs, the proportion of compounds with equal potencies against both proteins in a homologous pair is smaller for human paralogs. The observed differences in measured potencies are on average greater for paralogs. The following two sections provide a more detailed analysis of this finding.

4.2.4 Data model

The previous sections describe distributions of potency differences that were observed for identical small molecules between assays of the same target, assays of orthologous targets and finally assays of proteins that are paralogs. In this section I explore possibilities to model the observed data. As seen in previous sections, the distribution shapes suggested that the data were better approximated by models following a double-exponential distribution rather than the default model of a normal distribution.

Table 4.4: Fitted parameters for a model derived from the Laplace (\mathcal{L}) and normal distribution (\mathcal{N}).

set	\mathcal{L} , location	\mathcal{L} , scale	\mathcal{N} , location	\mathcal{N} , scale
between assays	-0.01	0.75	0.00	1.01
orthologs	0.08	0.77	0.14	0.86
paralogs	-0.01	1.30	-0.02	1.43

I used q-q plots³ to compare the observed distributions of potency differences with theoretical models based on either a double-exponential or normal distribution. As shown in Figure 4.7 the double exponential was an appropriate fit for data in the ortholog and between-assay categories. For data in the paralog category, the double-exponential distribution fit well to the distribution peak, but distribution tails were better approximated by a normal distribution. A fit of double exponential and normal models to histograms of the distribution data is shown in Figure 4.7. Parameters for the Laplace model as well as the rejected normal model were fitted as described in section 4.4.7 and are summarized in Table 4.4.

As briefly mentioned in section 4.2.1, the non-normal distribution of observed differences may have been caused by non-experimental errors such as unit transcription errors and mis-annotation of protein targets. A further contributing factor was likely residual contamination with dependent values. For example, in the unfiltered data of in-between assay measurements, I had observed a disproportionate amount of paired potencies that were exactly equal (data not shown). For the majority of these pairs it could be assumed that one of the measured potencies was reported in reference to a previously published measurement. To eliminate this source of dependent measurements, I excluded pairs of potency measurements that report exactly the same value for a given combination of ligand and protein target. Unavoidably, this would have removed some pairs that were genuinely equal. Therefore, to keep processing consistent, paired potencies that were exactly equal were also removed from the data sets for orthologs and paralogs. It is however possible that this step alone was not sufficient to eliminate all dependent measurements from the data. Kramer et al (Kramer et al., 2012) proposed that previously published values might be reported in subsequent publications as rounded values and/or

³In q-q plots or quantile plots, observed quantiles are plotted against the quantiles of a theoretical distribution proposed to model the observed data.

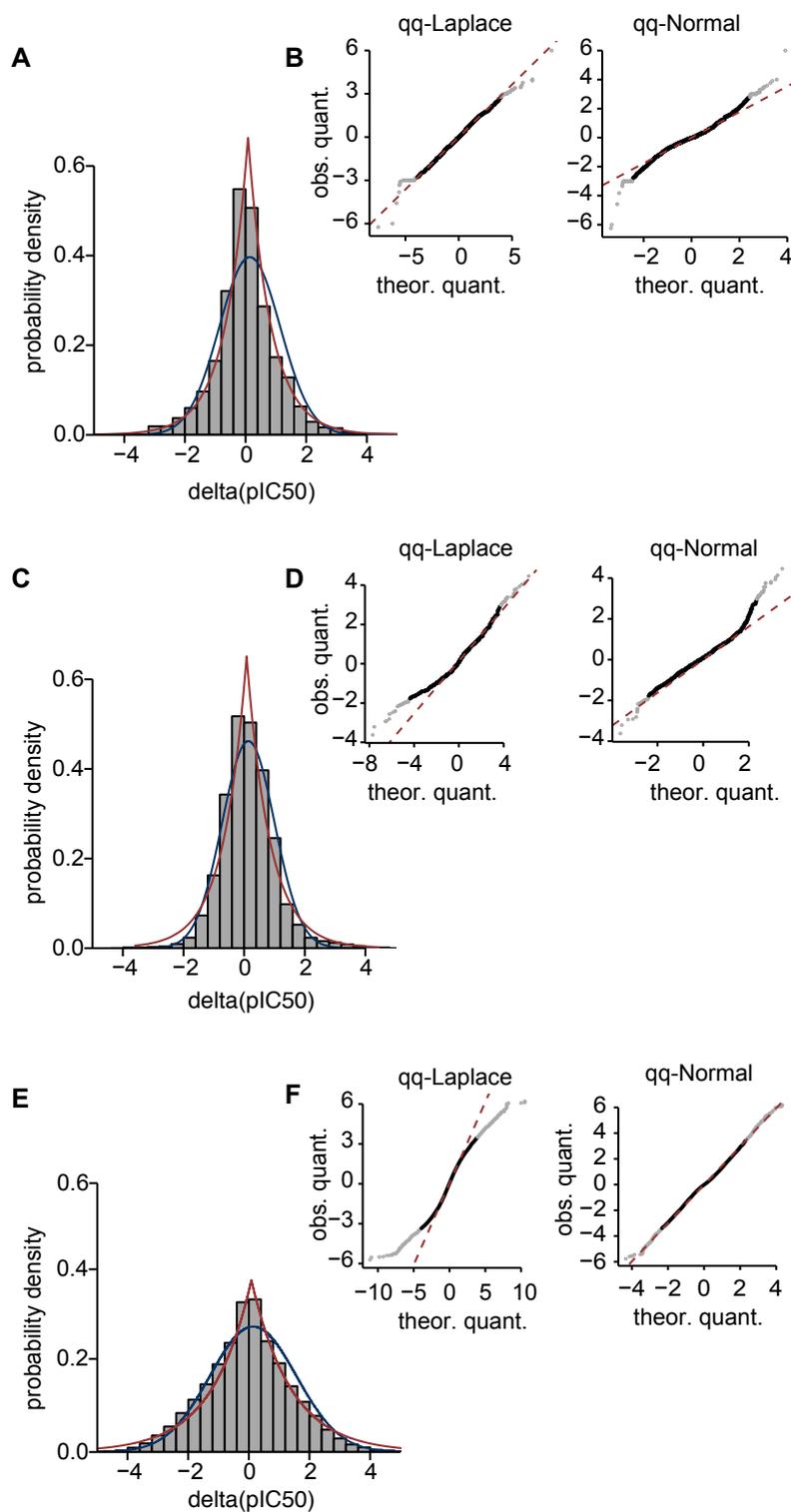


Figure 4.7: Fit of theoretical distributions to the observed data. Histograms of the data with fitted theoretical models of a normal (blue) and Laplace distribution (red) are shown for measurement differences between assays (A), between human-rat orthologs (C) and between human paralogs (E). Panels B, D and F show q-q plots for measurement differences between assays (B), between human-rat orthologs (D) and between human paralogs (F). The theoretical quantiles for plots titled ‘qq-Normal’ were derived from a normal distribution, quantiles for plots titled ‘qq-Laplace’ were derived from a Laplace distribution.

lead to a confirmation bias according to which experimentalists would ‘tweak’ their assays to obtain measurements that are in agreement with previously published values. This would result in paired potencies that are not exactly equal, but still more similar than entirely independent measurements. With regard to the results presented here it would be surprising if rounding and confirmation bias also applied on a large scale to potency measurements for orthologous or indeed paralogous proteins, yet both populations are not accurately described by a normal distribution.

On the whole it seemed more practicable to keep the filtering and processing of data consistent across all sets and accept that the data are not normally distributed, possibly owing to unexpected error and residual contamination with dependent measurements. Instead, the data were modelled using a Laplace distribution. The scale of a Laplace distribution is estimated by the average absolute deviation (AAD), which is more robust to fluctuations in the distribution tails compared to the variance, the measure of scale for a normal distribution. The mean absolute deviation is determined as

$$AAD = \frac{1}{N} \sum_{i=1}^N |Y_i - \bar{Y}|. \quad (4.3)$$

The implication of accepting the Laplace distribution as a model of the data was that statistical methods reliant on estimations of the variance or standard deviation would not be suitable and instead more robust metrics were required. Therefore, I used Spearman’s rank correlation method to determine correlation coefficients and the non-parametric Mann-Whitney U-test instead of Student’s t-test. The Laplace distribution has previously been applied to model error in gene expression data (Purdom and Holmes, 2005).

4.2.5 Evolutionary relationship and conservation of potency

In this section I describe an analysis of links between evolutionary relationship and measured differences in small molecule potencies. The ortholog conjecture states that orthologs, which share a common ancestor, but are separated through speciation have conserved function while paralogs arise from gene duplication events within a genome and, owing to the redundancy thus acquired, can evolve to develop divergent functions (Fitch, 1970; Chervitz et al., 1998; Tatusov et al., 1997). The purpose of the analysis was to examine if the contrast of functional conservation in orthologs versus functional

diversification in paralogs is also reflected in the response properties towards small molecule perturbation.

As a basis for the comparison I examined the distributions of potency differences for each category, orthologs, paralogs and between assays. In line with the expectation that there should not be a systematic bias between human and rat orthologs or the random groupings assigned when comparing potencies between assays or between human paralogs, the distributions for all categories had their mean and median around zero. With regard to conservation of response to small molecule perturbation, the most informative measure is distribution spread. The distribution spread decreases with the proportion of small molecules perturbing two homologs with the same or very similar potency. On the contrary, the spread increases with the proportion of small molecules that have substantially different potencies for two homologous proteins. For the distribution of potency differences between assays, the conservation of small molecule response properties is complete as proteins are identical. This distribution was used as a proxy to describe noise in the data. A comparison of the between-assay distribution with those for orthologs and paralogs then allowed me to infer the degree to which small molecule response properties are conserved relative to data noise on one hand and paralogs or orthologs on the other hand. Three measures of dispersion were calculated for each distribution to quantify robustly the spread for each distribution. As described in the previous sections (see 4.2.1, 4.2.2, and 4.2.3), the quantile intervals that would correspond to the interval covered by plus or minus one standard deviation and plus or minus two standard-deviations. In addition, distributions were assessed visually as shown in Figure 4.8.

The results indicate that there are only small differences between the baseline and human-rat orthologs, potency differences between paralogs are more disperse. The lack of distinction between the baseline and human-rat orthologs led me to conclude that from this global perspective, differences in susceptibility to small molecule perturbation are no greater between human and rat than between assays against identical proteins from the same species. These results are in line with expectation that protein function is largely conserved between two species that are, in evolutionary terms, closely related. On the contrary, this was not the case for human paralogs, for which potency differences were more disperse than both the baseline of differences between assays as well as differences between human-rat orthologs.

Figure 4.8: An overlay of the distributions of potency differences between assays (grey), between human-rat orthologs (red) and human paralogs (blue) is shown in this Figure.

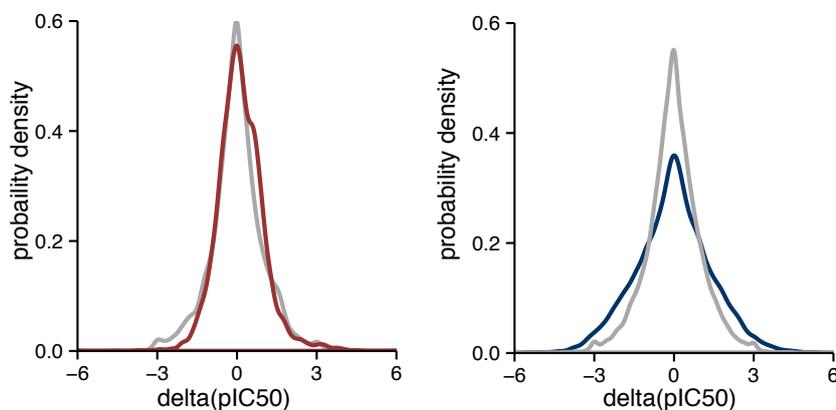
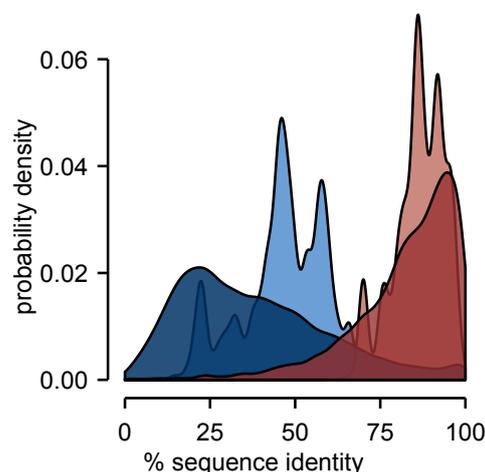


Figure 4.9: Probability density distributions of sequence identities for paralogs across the human genome (dark blue) and within the data set retrieved from ChEMBL (light blue). A density distribution of sequence identities is also shown for human-rat-orthologs across the entire genome of the two species (dark red) as well as those pairs that were part of the analysis (light red).



4.2.6 Sequence identity, ligand molecular weight and potency differences in paralogous pairs

The last common ancestor of humans and rats dates back approximately 92.3 Million years (number adopted from <http://www.timetree.org>, (Hedges et al., 2006)). All pairs of human-rat orthologs arose at this one point in time. In comparison, pairs of human paralogs have arisen at many different points spanning a vast time interval that extends to the present day. Most human paralogs are more ancient than the speciation event that separates humans and rats (Gibbs et al., 2004). Sequence identity is often viewed as a proxy of the point in time at which two genes diverged. Figure 4.9 shows

sequence identities between all pairs of human-rat orthologs and human paralogs that were retrieved from the EnsemblCompara Genetree Pipeline. Most human-rat orthologs have highly conserved sequences in the range of 90% to 100% identity, while most paralogs had much lower sequence identity, frequently around 30%. Given the pronounced contrast between orthologs and paralogs, I examined differences in susceptibility to small molecule perturbation as a function of sequence identity. To this end, sequence identity was determined on three levels, i) the full protein sequence, ii) the sequence of the Pfam domain that presumably mediates small molecule binding and, where possible, iii) the site of small molecule binding. Definitions of small molecule binding sites were based on an evaluation by Surgand et al. (Surgand et al., 2006) for GPCRs and the ‘canonical’ site definition from Kinase Sarfari for kinases. In the Methods sections 4.4.4, 4.4.5, 4.4.6, I give a detailed account of how sequence identity was determined on the three levels. From this, I calculated the Spearman’s correlation coefficient ρ of the absolute difference in small molecule binding and the sequence identity on the full protein level, both for orthologs and paralogs. For orthologs, there was no detectable correlation of sequence identity and the absolute difference in small molecule binding. In this case, ρ was -0.020 and the associated p-value 0.30. For paralogs on the other hand, ρ was -0.080 with a p-value of 1.2×10^{-59} . Hence, there was a statistically significant negative correlation that implied that human paralogs with more divergent sequences are also more likely to show differences in their susceptibility to small molecule perturbation. However, due to its small amplitude, it was difficult to decide if it reflected a meaningful trend or merely a small deviation from the exact null hypothesis that there is no relationship at all between sequence identity and susceptibility to small molecule perturbation. Further, on the level of Pfam domains and binding residues, the observed coefficients ρ were -0.084 and -0.202 respectively, and associated p-values 1.7×10^{-63} and 5.8×10^{-201} . A graphical summary of all examined relationships is provided in Figure 4.10. As stated, ρ was consistently negative for pairs of human paralogs and increased with the level of resolution of the sequence comparison. Taken together, these findings corroborate a meaningful relationship of sequence divergence and differences in small molecule binding. The increase of ρ with the level of resolution observed here further suggests that the impact of residue substitutions is generally much larger for positions in close proximity to the site of ligand binding. This suggests that mutations in or near a binding site have the greatest effect on differences in susceptibility to small molecule perturbation.

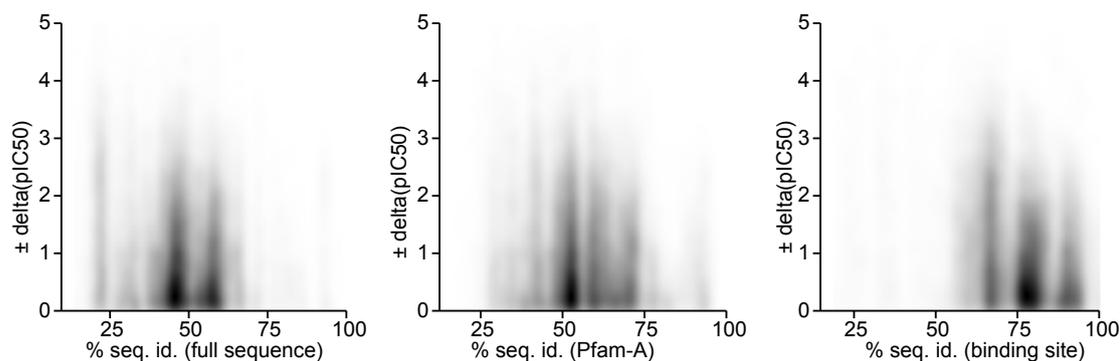


Figure 4.10: Absolute potency differences for each pair and compound are plotted against sequence identity for the full Uniprot sequence (left), the sequence of the Pfam domain predicted to mediate small molecule binding (centre) and binding site residues of GPCRs and kinases (right).

I also examined the relationship of the molecular weight of the ligand and differences in small molecule binding between paralogous pairs. Ligands that undergo interactions with larger numbers of residues should have a higher likelihood of exhibiting different potencies towards paralogous proteins as they are more likely to ‘sense’ residue substitutions, given there are any in or near the binding site. The number of interactions that a small molecule undergoes with any given protein depends to some extent on a property termed molecular complexity (Hann et al., 2001). This property cannot be accurately quantified without knowing the binding mode of a small molecule as it depends on the number and quality of protein-ligand interactions. Here, I used molecular weight of the ligand as a proxy for molecular complexity. My analysis showed a very weak positive correlation of molecular weight and potency differences (Spearman’s correlation coefficient: 0.062, p-value: 3.1×10^{-36}). By traditional standards, the very low p-value should indicate significance but the high number of data points in this analysis (41,334) might warrant a more stringent threshold. Figure 4.11A shows a density representation of the data. The possibility that only a subset of the paralogous pairs in the analysis has substitutions near the binding site that are ‘sensed’ by some, but not all ligands was identified as one factor that could contribute to a very low effect size. To investigate this, I examined the correlation of molecular weight and potency differences for individual pairs of homologous targets. This reduced the numbers of data points examined at one time, bringing it closer

to traditional experimental setups. A summary of this analysis is shown in Figure 4.11B. I found that the number of pairs that have a positive correlation is higher compared to pairs with a negative correlation and that this difference becomes more pronounced with increasingly stringent p-values. Table 4.5 summarises this observation for five p-value thresholds. Table 4.6 lists paralogous pairs left in the analysis under the most stringent p-value thresholds. Note that eighteen of those have a positive correlation of absolute potency differences with molecular weight, while only three have a negative correlation. Taken together, I interpreted the above observations as an indication that mutations in or near the binding site have important implications in the functional diversification of paralogs.

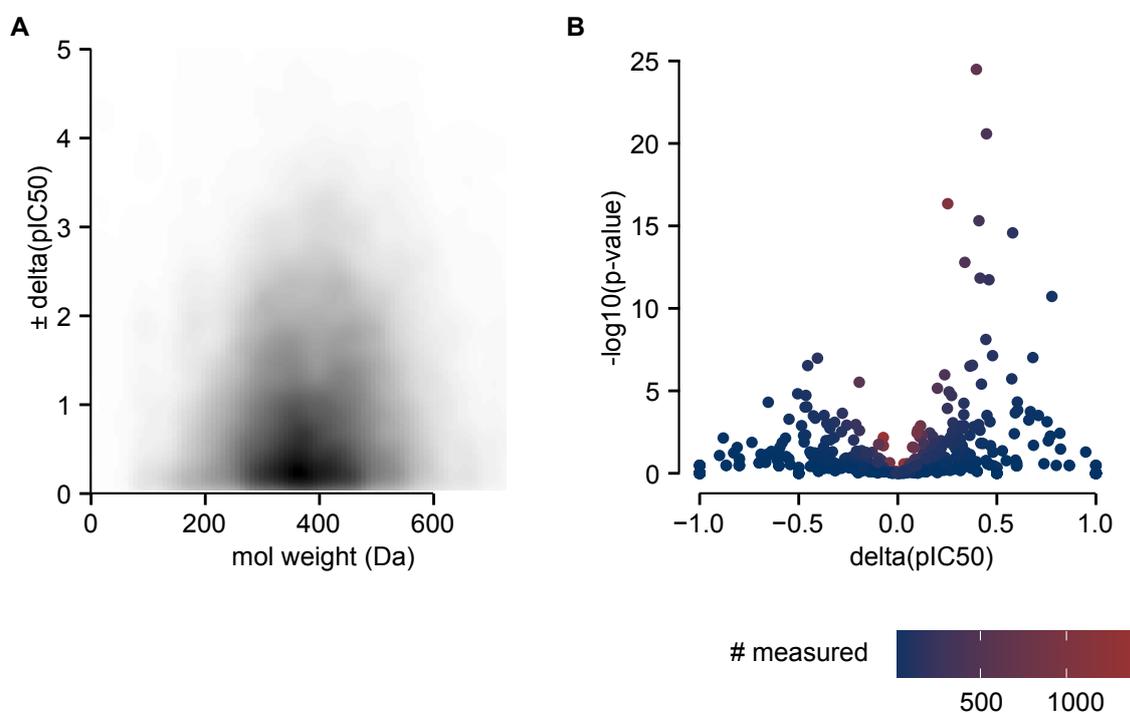


Figure 4.11: Ligand molecular weight and absolute potency differences. Panel A shows absolute potency differences plotted against molecular weight. To avoid overplotting, a density function was used to represent the distribution of 41,334 data points. Panel B shows a plot of correlations calculated for individual pairs of paralogs. Each dot represents a pair of paralogs. The position on the x-axis represents the Spearman correlation coefficient, and the position on the y-axis the negative logarithm of the associated p-value.

Table 4.5: Correlation of molecular weight and absolute potency differences for individual pairs of paralogs. Counts of pairs with positive (+) and negative (-) correlations are shown at different p-value thresholds. The ratio of pairs with positive correlation over the total number of pairs consistently increases with more stringent thresholds.

p-value	(+)	(-)	% (+)
total	294	245	0.55
< 0.01	55	29	0.65
< 0.001	33	15	0.68
< 0.0001	22	8	0.73
< 0.00001	18	3	0.86

Table 4.6: Individual pairs of paralogs in the molecular weight analysis. Listed below are pairs of paralogs where p-values associated with the correlation of ligand molecular weight and absolute potency differences are $p < 0.00001$. The columns ‘reference’ and ‘paralog’ define each pair, ‘ $-\log(p\text{-value})$ ’ the negative base 10 logarithm of the p-value and ρ the Spearman correlation coefficient. The bottom three pairs show a negative correlation.

reference	paralog	$-\log(p\text{-value})$	ρ
Muscarinic ACh-R M5	Muscarinic ACh-R M3	10.7	0.78
Muscarinic ACh-R M4	Muscarinic ACh-R M5	7.0	0.68
Nitric-oxide synthase, brain	Nitric oxide synthase, inducible	14.6	0.58
Muscarinic ACh-R M3	Muscarinic ACh-R M4	5.7	0.58
β -secretase 1	β -secretase 2	7.1	0.48
S1P-R Edg-1	S1P-R Edg-3	11.7	0.46
Cathepsin S	Cathepsin K	20.6	0.45
Matrix metalloproteinase 2	Matrix metalloproteinase 13	8.1	0.44
Thyroid hormone-R α	Thyroid hormone-R β -1	5.4	0.42
Cathepsin S	Cathepsin L	11.8	0.42
Melanocortin-R 5	Melanocortin-R 4	15.3	0.41
Dopamine D3-R	Dopamine D2-R	24.5	0.40
Caspase-7	Caspase-3	6.6	0.38
A1b adrenergic-R	A1a adrenergic-R	6.5	0.36
Adenosine A3-R	Adenosine A2b-R	12.8	0.34
Cannabinoid CB1-R	Cannabinoid CB2-R	16.3	0.25
Butyrylcholinesterase	Acetylcholinesterase	6.0	0.24
Melanocortin-R 4	Melanocortin-R 3	5.2	0.20
Serotonin 2c-R	Serotonin 2a-R	5.5	-0.20
Vasopressin V2-R	Vasopressin V1a-R	7.0	-0.41
Somatostatin-R 3	Somatostatin-R 5	6.5	-0.45

4.2.7 Assessment of individual homologous pairs

The previous parts of this analysis are based on a global approach that looks at averages over a wide range of compounds and proteins. In this part, I present an analysis of individual targets. To this end I implemented a testing procedure that compared the distribution of differences for individual orthologous pairs to the distribution of differences observed between assays for human and rat proteins (as described in section 4.2.1). In an analogous procedure, I also examined distributions for individual pairs of paralogs and how they compared to the distribution of differences observed between assays against human proteins. Non-parametric Mann-Whitney U tests were carried out to assess if the measured differences for each pair of orthologs or paralogs were consistent with the distribution of differences between assays. The resulting p-values should not be considered indicative of significance, but rather as a ranking criterion that evaluates both the effect size and the number of tested compounds. This procedure conceptually follows the selection of differentially expressed genes (Chen et al., 2007). Here, the aim of p-values is not to attach significance, but to prioritise differentially expressed genes for further analysis, e.g. manual inspection or experimental validation. Similarly, p-values calculated in this analysis are intended as a criterion for prioritisation and not a hard threshold of significance. For this reason, I did not adjust p-values for multiple comparison. Volcano plots for both orthologs and paralogs are shown in Figure 4.12A, B. The median absolute effect size was 0.57 pIC₅₀ units for paralogs and 0.43 pIC₅₀ units for orthologs. P-values were generally greater for orthologs, a consequence of both smaller effect size as well as fewer tested compounds. As shown in Figures 4.12C and D, I found there is a larger fraction of human proteins that are on average more susceptible to small molecule perturbation than would be expected under a random model, such as the one for paralogs. As discussed previously, this may point towards a bias caused by the human-centric nature of the data in ChEMBL.

The ten paralogous pairs with the greatest systematic deviation from the control distribution are summarised in Table 4.7. A prominent difference emerged between the nociceptin receptor and other members of the opioid receptor family. The nociceptin receptor is known to have a substantially different profile of ligands compared to its relatives, the opioid receptors (Thompson et al., 2012).

A table of the ten top-ranking orthologous pairs is provided in Table 4.8, summarising

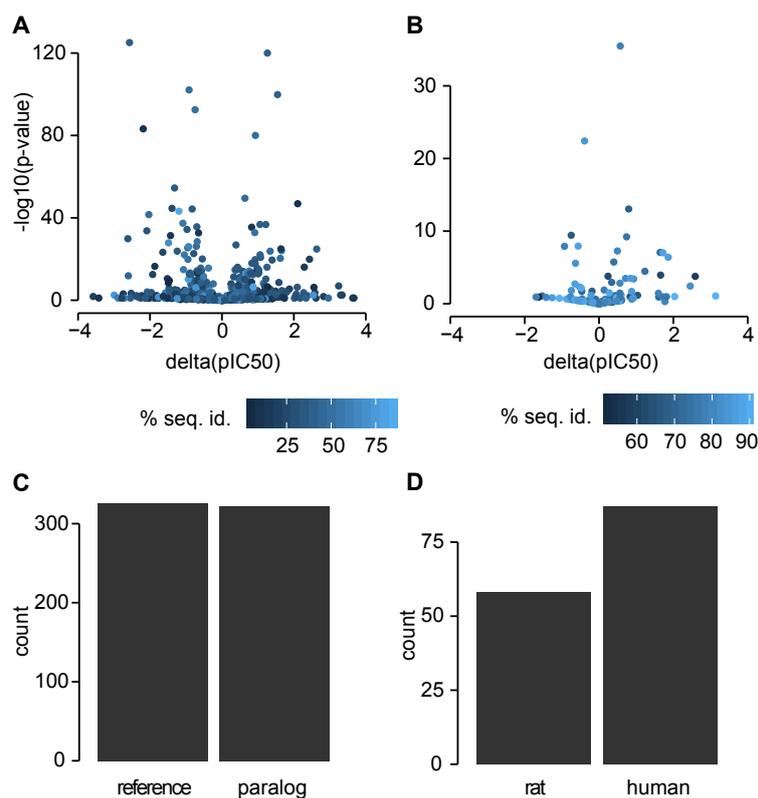


Figure 4.12: Volcano plots for orthologs and paralogs summarise the systematic deviations from the control distribution for individual homologous pairs. Each point represents a reference protein and its paralog (A) or a human protein and its rat ortholog (B) and its position on the x-axis the potency difference between the two, averaged over all tested compounds. The position on the y-axis indicates the negative logarithm of the associated p-value. Points are coloured according to the sequence identity between two proteins in a pair. The count of pairs on either side of the ordinate is summarised in barplots for orthologs (C) and paralogs (D).

Table 4.7: Human paralogs with greatest overall potency differences. The columns ‘reference’ and ‘paralog’ provide names of paralogous pairs, ‘ Δ ’ the observed average potency difference between measurements of the protein indicated in the ‘reference’ column and its paralog. The column ‘n’ indicates the number of measurements that were compared for each pair of paralogs.

reference	paralog	p-value	Δ	n
Nociceptin receptor	Delta opioid receptor	3.2×10^{-124}	2.55	232
Nociceptin receptor	Mu opioid receptor	1.9×10^{-114}	1.24	520
Nociceptin receptor	Kappa opioid receptor	1.8×10^{-98}	1.54	280
Delta opioid receptor	Mu opioid receptor	5.7×10^{-97}	-0.89	921
Carbonic anhydrase I	Carbonic anhydrase II	4.2×10^{-91}	-0.74	993
Dipeptidyl peptidase VIII	Dipeptidyl peptidase IV	9.6×10^{-83}	-2.18	252
Melanocortin receptor 3	Melanocortin receptor 4	1.7×10^{-79}	-0.93	506
α -adrenoceptor 1b	α -adrenoceptor 1a	1.0×10^{-53}	-1.31	186
Delta opioid receptor	Kappa opioid receptor	1.9×10^{-48}	-0.63	901
Dipeptidyl peptidase IX	Dipeptidyl peptidase IV	3.5×10^{-47}	-2.11	165

effect sizes and p-values. These values must be interpreted with caution, especially in cases where the number of associated publications is small. In such cases, the observed difference may be founded in systematic bias between two assay-setups and not be biologically relevant. This was definitely the case for the neuronal acetylcholine receptor subunit α -7, where, due to mis-annotation in the database, measurements from a cell-based assay were compared with measurements from a biochemical assay. Orthologs of the HRH3 were identified as the highest ranking pair using my method. Given the large number of ligands tested, the estimated bias of 0.54 log-units on average for the human receptor seemed robust and indeed, the species-specific pharmacology of the HRH3 had been noted in some of the articles the underlying data was derived from (Black et al., 2007; Nersesian et al., 2008; Zhao et al., 2009) and had also been described in independent research articles (Lovenberg et al., 2000; Ligneau et al., 2000). I also found evidence in the literature that the serotonin transporter (SERT), ranked second using my method, exhibits species-specific pharmacology for a number of chemotypes (Barker et al., 1994). In my analysis, 309 compounds had on average a bias of 0.47 log-units for the rat SERT. Across fifty compounds, the estimate of the average bias for the human Fructose-1,6-Bisphosphatase (1,6FBP) over its rat ortholog was 0.77 log-units, a bias that had also been reported in the literature (Erion et al., 2005). The last column in Table

4.8 lists publications reporting species-specific pharmacology for some of the remaining proteins in this analysis. It is important to underline that the observed differences are reported as averages across multiple different compounds and that true differences might apply only to a subset of these compounds. This was for example observed with the HRH3 (see section 4.2.8) and explains the somewhat paradoxical finding that a considerable number of proteins appear to have species-specific pharmacology while from a global perspective the potency patterns appear conserved across human-rat orthologs. In addition I noticed that, when potencies of all compounds measured against a pair were averaged, the fraction of pairs where the human ortholog was more susceptible to small molecule perturbation was disproportionately higher. The explanation for this might lie in the human-centric nature of drug discovery. The goal of drug discovery is ultimately to obtain compounds that are potent against human proteins - and potency optimisation of lead structures would typically be carried out against human proteins. It is thus likely that there is a bias for compounds with higher potencies against human proteins in the underlying data set.

Table 4.8: Human to rat orthologs with greatest overall potency differences. The column ‘ Δ ’ provides the observed average potency difference between measurements of the human protein and its rat ortholog. The column ‘n’ indicates the number of measurements that were compared for each orthologous pair. The column ‘citation’ provides references to publications that report significant differences in small molecule response between the human and rat ortholog.

name	p-value	Δ	n	citation
Histamine H3 receptor	2.01×10^{-35}	0.56	325	Lovenberg et al., 2000
Serotonin transporter	3.29×10^{-26}	-0.45	309	Barker et al., 1994
Neur ACh-R α -7 subunit	2.86×10^{-24}	-2.36	49	None
Fructose-1,6-bisphosphatase	1.37×10^{-13}	0.80	50	Erion et al., 2005
NaV type X α -subunit	9.69×10^{-10}	0.74	45	None
Adenosine A1 receptor	4.87×10^{-8}	-0.55	78	Maemoto et al., 1997
GnRH receptor	5.69×10^{-8}	0.50	118	Arora et al., 1999
D-amino-acid oxidase	1.10×10^{-7}	1.65	14	None
Neurokinin 1 receptor	1.43×10^{-7}	1.72	20	Fong et al., 1992
Adenosine A2a receptor	4.07×10^{-6}	0.38	112	None

4.2.8 Small molecules as probes of the binding site

The perturbation of protein function by small molecules is ultimately mediated by interactions between specific residues and the small molecule acting as a ligand. Medicinal chemistry experience shows that small changes in the chemistry of the ligand can have a great impact on the interaction. For example, it has been reported that the addition of a single methyl group to the structure of a lead compound can change binding affinity by one order of magnitude (Shamovsky et al., 2008; Davis and Teague, 1999; Zhi et al., 1999; Hopkins et al., 2004; Hann et al., 2001). In the context of evolutionary relationships, I wanted to investigate if the same concept applies to amino acid substitutions between related protein targets.

As I described in chapter 4.2.6, I observed a meaningful inverse correlation between sequence identity and differences in small molecule binding between pairs of human paralogs. From this protein-centric perspective, it seems plausible that amino acid substitutions in or near the site of ligand interaction have more profound effects than substitutions elsewhere. The integration of multiple experiments also permitted to examine this question from a ligand-centric perspective. By using the ensemble of tested compounds for a pair of related proteins, the impact of changes in the binding site could be measured indirectly. In the simplest case, potency differences between two proteins would be observed consistently for all tested compounds. Such cases would be indicative of an amino acid substitution at a position that is involved in interactions with all tested ligands. A more subtle scenario would involve only a subgroup of small molecules showing consistent differences in potency across two homologous proteins. Such pairs could be viewed as having an amino acid substitution in a position that mediates interactions with some, but not all tested small molecules. Clearly, this indirect effect depends not only on substitutions between the proteins, but also the set of tested molecules. From a practical point of view, it is impossible to identify all active compounds let alone their potencies against a homologous pair. I wanted to investigate if, given the limited number of known potencies, any inferences could be made about substitutions promoting selectivity between homologous targets.

Orthologs of the HRH3 had emerged as the orthologous pair with the most distinct differences in susceptibility to small molecule perturbation between the two species human and rat. Of 325 tested ligands, 201 had at least 0.5 pIC₅₀ units higher potency

against the human protein over the rat protein. I carried out a simple clustering of the 325 ligands to assess if any chemical features of the ligands were associated with higher affinity for the human ortholog of the HRH3. The clustering was based on the simple LINGO fingerprint descriptors (see Methods section 4.4.8) and yielded one large cluster of 189 compounds and 17 additional clusters with mostly between 5 and 18 compounds as well as 15 singletons. Figure 4.13 illustrates potencies observed for compounds assigned to the different clusters. From manual inspection of the clusters, I concluded that the largest cluster contained mainly compounds characterized by a pyrrolidine moiety and a smaller group of compounds with a biphenyl scaffold. Compounds in cluster 5, which predominantly had higher potencies for the human receptor, are all based on a steroid scaffold. All compounds in cluster 32 contained a diazepine moiety; these compounds also had higher potencies at the human receptor. Compounds in cluster 10 contained a terminal imidazole moiety. Compounds in this cluster had higher potency at the rat receptor. Cluster 17 contained exclusively compounds with an indole scaffolds. Compounds in this cluster mostly had equal potencies at the human and rat receptor. The fingerprints and clustering method used here are basic methods that are not very sensitive. However, as I confirmed by manual inspection of the clusters, these methods were suitable to separate compounds from different chemical series evaluated against the HRH3 receptor.

A sequence alignment showed that there were only few substituted positions in the two receptors. The overall sequence identity was 94.2%, an alignment is shown in the Methods section 4.16. Small molecule binding in Type I GPCRs is normally coupled to the cavity in the transmembrane region of the receptor. To tighten this estimate further, I constructed homology models for the human and rat HRH3, based on all structures of Type I GPCRs available at this time, including a crystal structure of the human histamine H1 receptor (HRH1) (Shimamura et al., 2011). The ligand doxepin of 3rzc was used as a surrogate ligand for modelling the binding site of the human and rat ortholog of the HRH3. The models were constructed using the Modeller software (see Methods section 4.4.7.3) and aligned using functionality of the QtMG package. Figure 4.14 shows an overlay of the two models. Datasets containing the PDB and alignment files describing the model were made available as described in table 4.12 in Methods section 4.4.7.3.

The peptide backbones of the transmembrane helices of the two receptor models aligned well. The overall root mean square deviation of the two models was 0.21 Å

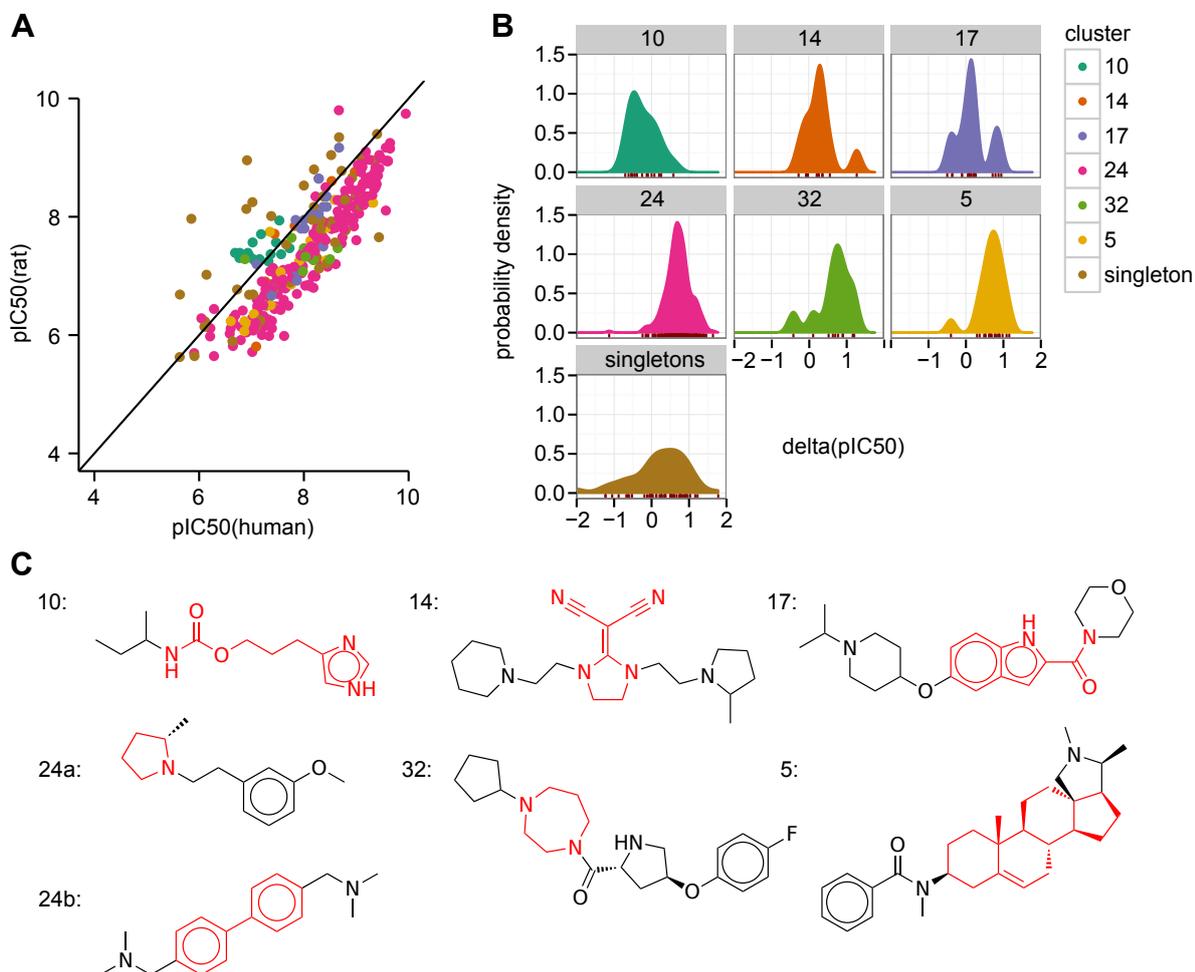


Figure 4.13: Clustering of ligands of the HRH3 receptor. Panel A summarises measured potencies. Dots represent potencies measured against the human receptor (x-axis) and rat receptor (y-axis). Dots are coloured according to the cluster they were assigned to. The diagonal line represents exactly equal measurements and is drawn as a standard of comparison. Panel B summarises probability density functions of the potency differences observed between the human and rat HRH3 for each cluster separately. Contributions from individual compounds are represented as red tick marks at the bottom of each plotting facet. Panel C shows exemplary structures from each cluster. Highlighted in red are shared features among compounds in the same cluster. Cluster 24 contains multiple series.

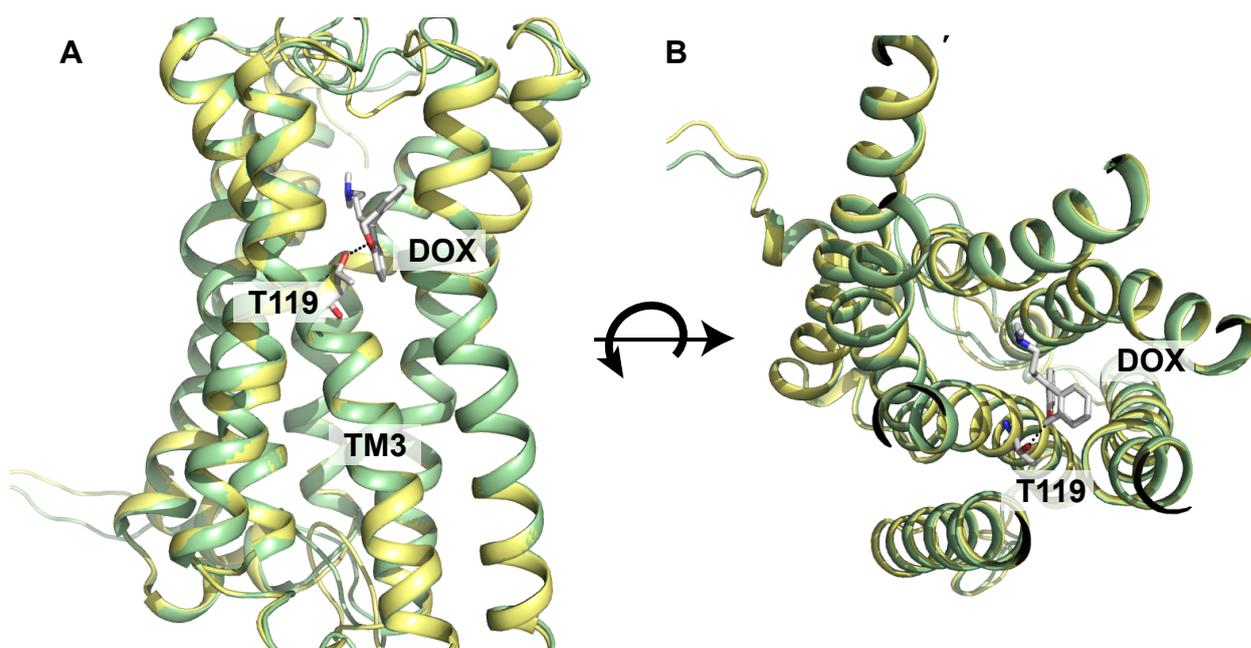


Figure 4.14: Homology models of the HRH3 receptor. In this overlay, model coordinates are represented as cartoon-structures, the human HRH3 in green, and the rat HRH3 in yellow. Panel A shows a lateral view of the aligned receptors, panel B an approximately 90° rotated view onto the top of the receptor. Parts of the extracellular loop region have been clipped from this view. The substituted residue T119A between human and rat is shown in stick representation. A hypothesised hydrogen bond between the hydroxyl group of T119 and doxepin is shown as a dashed line. The ligand Doxepin is adopted from the published crystal of the HRH1, 3rze.

over 1,253 aligned backbone atoms. One of the few substituted positions in the aligned receptors was in close proximity to the modelled ligand doxepin. The substitution in question is from threonine in human to alanine in rat at position 119 in the Uniprot sequence. The distance between the oxygen in the hydroxy-group of threonine and the oxygen-atom in the oxazepine moiety of the ligand is 2.7 Å. In their analysis of the structure, Shimamura and colleagues note that the hydroxy-group of T119 undergoes hydrogen-bonding with the oxygen of the E-, but not the Z-isomer of doxepin. They further note that it does not contribute significantly to binding affinity in predicted complexes of the HRH1 with olopatadine, cetirizine and fexofenadine. It is therefore plausible that the substitution Thr119Ala is the cause of the observed potency offset between the human and rat HRH3 for a subset of compounds. This pair of HRH3 receptor orthologs is a striking example of how subtle differences in the local context of ligand binding can modulate susceptibility to small molecule binding and override overall sequence similarity.

4.3 Conclusion

In this part of my thesis work I integrated small molecule potency measurements and homology information. I created an assembly of pairwise comparisons for human-rat orthologs and human paralogs. My analysis of these comparisons showed that, across measurements of human-rat orthologs, potency differences were no greater than the variability that was observed between assays of identical targets. These findings indicate that susceptibility to small molecule perturbation is largely conserved between humans and rats on a molecular level. From the perspective of small molecule-protein interactions represented in the ChEMBL data base, my work thus confirms the use of rats as model organisms in pharmaceutical research. However, on the level of more complex interactions, such as transcription networks and metabolism, species differences may limit the predictive power of extrapolations, even between closely related species (Odom et al., 2007; Kutter et al., 2012; Martin et al., 2010). In cases where the underlying functional modules are conserved, phenotypic outputs can be related and used to predict the effect of a specific perturbation in one species from the phenotype observed in another species (McGary et al., 2010). A systematic search for relationships between phenotypic responses in different species allows for more precise interpretations of phenotypic observations and

can thus facilitate the use of phenotypic screens in drug discovery ([Kapitzky et al., 2010](#); [Hoehndorf et al., 2013](#)).

In this chapter, I also highlighted a number of proteins that may have altered susceptibility to small molecule perturbation across the two species, most prominently the HRH3. I further observed that for a majority of paired measurements, potency was greater at the human receptor, albeit insignificantly. This observation is likely an artefact of the human-centric nature of drug discovery that favours compounds with potency in men, not rats. In contrast, potency differences between human paralogs exceeded assay variability. This finding is in line with the ortholog conjecture that states that gene duplication within a genome enables neo- and sub-functionalisation. My thesis work thus adds a new perspective to a recent debate around this hypothesis ([Nehrt et al., 2011](#); [Altenhoff et al., 2012](#); [Chen and Zhang, 2012](#)) using the indirect measure of susceptibility to perturbation by small molecules. This is also interesting from the point of view that most of the small molecules in this analysis are artificial ligands that can not have shaped the evolution of their receptor, but rather have been ‘retrofitted’ to emulate natural ligands. The use of potency data as a functional output in phylogenetic studies opens venues that have traditionally been reserved for sequence- and annotation-based approaches. For example, findings from a recent study that examines mutation rates of drug targets in a number of mammalian species as a function of subcellular localisation ([Wang et al., 2013](#)) might be tested using small molecule potency data.

Perhaps surprising was that the observed correlation of sequence divergence and measured potency differences was so weak, even when taking into account the vast amount of noise inherent to the potency data. Changes in overall protein structure correlate with sequence identity for homologous pairs ([Chothia and Lesk, 1986](#)). Apparently, structural differences between homologous pairs do not translate directly to differences in susceptibility to small molecule perturbation. Reconciliation of these outwardly contradictory findings can be found in the hypothesis that functional sites in proteins are subject to purifying selection ([Lichtarge et al., 1996](#); [Ma et al., 2003](#)) and are thus more conserved than other residues on the protein surface. By increasing the focus on residues involved in small molecule binding, I was able to obtain a clearer signal for the correlation of sequence divergence and differences in small molecule binding and could thus demonstrate that the conservation of ligand binding depends to a large degree on few, but crucial mutations in the binding site more than overall sequence identity.

4.4 Methods

4.4.1 Retrieval and processing of measured potencies

The Python library MySQLdb was used to interface with a MySQL instance of ChEMBL (version 10). A SQL statement was written to obtain the potency values and associated parameters for further analysis. The requirements within the query were set to obtain only potencies from binding assays against protein targets that are neither multimers nor complexes (`assay_type = 'B'`, `target_type = 'PROTEIN'`, `multi = 0`, `complex = 0`). A further requirement in the query was for the sequence of the assay target to match its Uniprot identifier directly, rather than through homology (`relationship_type = 'D'`). Potencies of the types 'IC50', 'EC50', 'Ki', 'pA2', 'pKi' were obtained for rat and human proteins and stored in Python objects. Potencies expressed as IC50, EC50, K_i were converted to the corresponding logarithmic scale. Following guidelines published by Kalliokoski et al. (Kalliokoski et al., 2013), an offset of 0.35 was subtracted from all pK_i measurements to avoid loss of data quality when mixing IC50 and K_i data. The queries that were used to obtain the measurement data from the ChEMBL database can be inspected at <https://github.com/fak/globalAnalysis/blob/master/queries.py>. The script that was used to filter and process the data is deposited at <https://github.com/fak/globalAnalysis/blob/master/mkDict.py>.

4.4.2 Data assembly for inter-assay comparison

The Python objects described in 4.4.1 were queried for compounds tested against identical targets in different assay systems. Potencies were retrieved and processed as described in Methods 4.4.1. If multiple measurements were found for a combination of a compound and protein target, two measured potencies were selected at random and kept for comparison. For each of the two species *H. sapiens* and *R. norvegicus* all pairs obtained in the inter-assay comparison were written to a tab-separated text file. For comparison against the data obtained for human-rat orthologs, a comparable number of inter-assay comparisons were selected at random, one half from the table generated for *H. sapiens*, one half from the table generated for *R. norvegicus*. The functions that I used to sample pairs of measurements are `interAssaySampled`, `orthologSampled`, and `paralogSampled` and can be inspected at <https://github.com/fak/globalAnalysis/>

[blob/master/writePairs.py](#).

Differences between measurements were calculated and those exactly equal to zero or with an absolute value greater than twenty were discarded. This step was performed to exclude the most implausible values and to filter out values that are duplicated in the database or reported as references to previous papers. The sample of 3,000 measurement pairs was taken only once and at random and written to a tab-separated file. The analysis is documented at <https://github.com/fak/globalAnalysis/wiki/summary>. Figure 4.15 shows the probability density function that was calculated from the first sample, and for comparison, one hundred additional probability density functions that were sampled in the same manner. This part of the analysis is also documented at <https://github.com/fak/globalAnalysis/wiki/sampling>.

4.4.3 Data assembly for homologous pairs

The assembly of both pairs of human-rat orthologs as well as pairs of human paralogs, was based on mappings provided by the EnsemblCompara GeneTrees pipeline (Vilella et al., 2009), accessed through the BioMart website (Ensembl version 62, <http://apr2011.archive.ensembl.org/biomart/martview/>). Queries against this website return tables listing all combinations of homologous genes identified within a genome (human paralogs) or between the genomes of two species (human-rat orthologs). The homologous pairs from these tables were used to query the python objects described in Methods 4.4.1) for compounds with measured potencies against both components, matching on Uniprot accessions. For instances where a compound had more than one potency measured against an identical target, the median of all values was obtained. Pairs of homologous proteins together with the paired potency measurements were written to two tab-separated files, one for human-rat orthologs, one for human paralogs. For downstream analysis, differences between measurements were calculated, and those exactly equal to zero or with an absolute value greater than twenty were discarded. This was to ensure that both differences between assays and differences between human-rat orthologs had been processed in the same way. The function that was used to match homologous proteins is deposited at <https://github.com/fak/globalAnalysis/blob/master/mkHomologTable.py>.

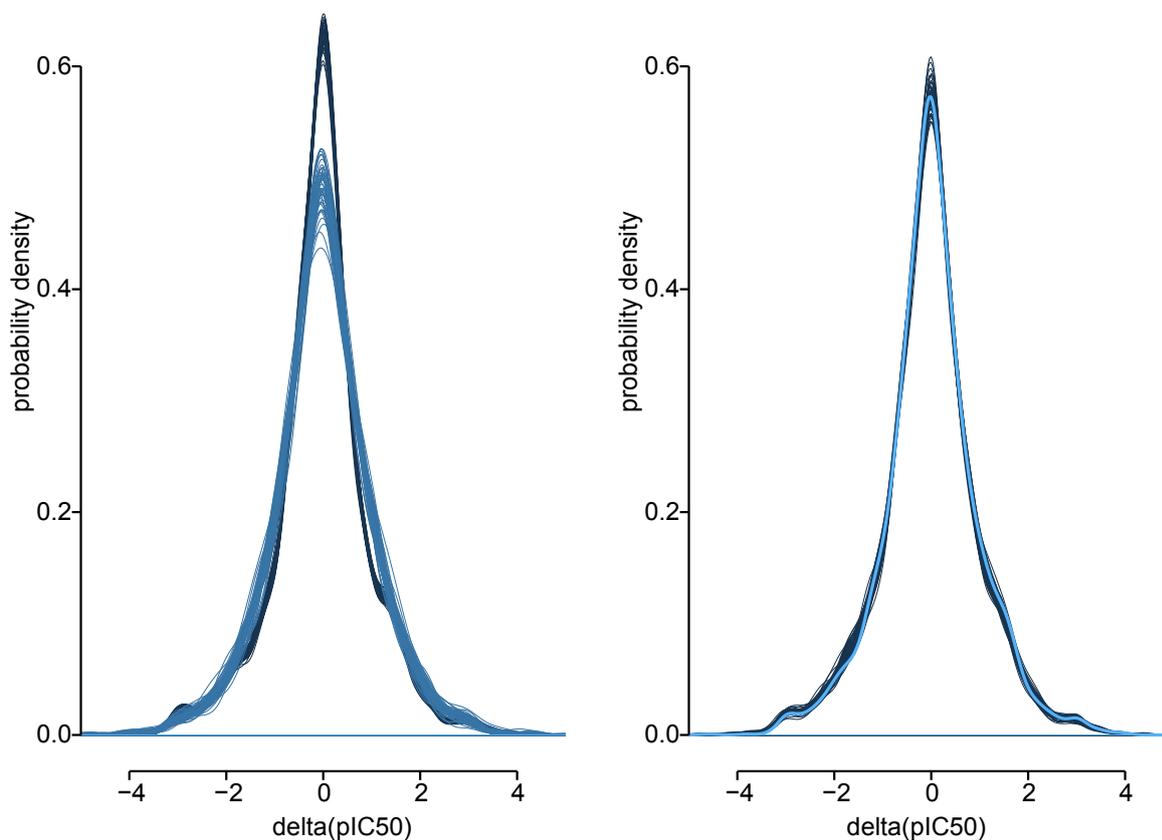


Figure 4.15: Panel A shows an overlay of 100 distributions sampled from human and rat data. Shown in light blue in panel B is the probability density function that was fitted to the data sampled from observed inter-assay differences. The grey curves represent probability density functions fitted to one hundred additional random samples.

4.4.4 Sequence identity for full-length proteins

Sequence identities between pairs of homologous proteins were obtained from alignments generated by the EnsemblCompara GeneTrees pipeline. It is based on a combination of multiple alignment algorithms, consensified by M-Coffee (Wallace et al., 2006). EnsemblCompara provides both the sequence identity of a query protein against its homolog, as well as vice versa the sequence identity of the homolog versus the query protein. From these two, the higher value was selected to focus the alignment on a region common to both proteins and most likely involved in small molecule binding.

4.4.5 Sequence identity on a Pfam domain level

Sequence identity was also determined on a protein domain level. Here, the aim was to compare sequences of the protein domains that mediate small molecule binding. To this end, a heuristic mapping described in Methods 4.4.9 were used as an estimate of the true Pfam domain mediating small molecule binding. For each protein in a homologous pair, the Pfam domain identifier as well as its start and end positions within the protein sequence were obtained from database tables generated as described in sections 2.5.1 and 2.5.3. Start and end positions were then used to slice full length sequences obtained from the `protein_sequence` field in `target_dictionary` to yield the Pfam domain sequence only. The EMBOSS implementation of the Needleman-Wunsch algorithm `needle` (EMBOSS version 6.4.0, Needleman and Wunsch, 1970; Rice et al., 2000) was used to produce pair-wise global sequence alignments for each pair of Pfam domains. Following recommendations for closely related protein sequences, `needle` was executed using a gap-open penalty of 10, gap-extend penalty of 0.5 and the EBLOSUM62 substitution matrix. Output was generated in the ‘pair’ format and parsed from `stdout` using the python regular expression module `re`. The function that was used to carry out the alignment is `pfam_a` from <https://github.com/fak/globalAnalysis/blob/master/align.py>.

4.4.6 Sequence identity on a binding site level

Sequence identity between homologous pairs was also determined on a binding site level, where possible. Thirty critical binding site residues for members of the family of G-protein coupled receptors were specified in a publication by Surgand et

al. (Surgand et al., 2006). I applied the residue positions specified to a family-wide multiple sequence alignment obtained from GPCR Sarfari <https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari>. The positions corresponding to the residues highlighted by Surgand et al are listed in Table 4.9. Similarly, known binding site residues are listed for kinases in the chemogenomics resource Kinase Sarfari. Kinase Sarfari operates with five distinct binding site definitions, ‘Canonical’, ‘Gleevec’, ‘PKA’, ‘MEK2’, and ‘P38’. In this analysis, I applied the ‘Canonical’ binding site definition to all kinases. Site definitions adopted from Kinase Sarfari are outlined in Table 4.10.

For both kinases and GPCRs, the residues corresponding to the binding site definitions were obtained from the family-wide multiple sequence alignments provided by GPCR Sarfari and Kinase Sarfari. For each homologous pair, sequence identities were calculated using the EMBOSS implementation of the Needleman-Wunsch algorithm `needle` (EMBOSS version 6.4.0, Needleman and Wunsch, 1970; Rice et al., 2000). Residue positions in this comparison are fixed and hence the alignment should be gap-free. To minimise the chances of introducing gaps into the alignment by chance, a gap-open penalty of 100 was used. Output was generated in the ‘pair’ format and parsed from `stdout` using the python regular expression module `re`. The function that was used to carry out the alignment is `bSite` from <https://github.com/fak/globalAnalysis/blob/master/align.py>.

Table 4.9: Critical binding site residue positions in the GPCR family alignment.

Helix	Ballesteros - Weinstein	alignment residue
TM1	1.35, 1.39, 1.42, 1.46	65, 69, 72, 76
TM2	2.57, 2.58, 2.61, 2.65	132, 133, 136, 140
TM3	3.28, 3.29, 3.32, 3.33, 3.36, 3.40	164, 165, 168, 169, 172, 176
TM4	4.56, 4.60	239, 243
TM5	5.38, 5.39, 5.42, 5.43, 5.46	350, 351, 354, 355, 358
TM6	6.44, 6.48, 6.51, 6.52, 6.55	485, 489, 492, 493, 496
TM7	7.35, 7.39, 7.43, 7.45	532, 536, 540, 542

4.4.6.1 Assessment of potency differences between homologous proteins

For each combination of a compound and pair of homologous proteins, measured pairwise potencies were plotted in scatterplots detailing the observed potency against a protein on the x-axis and the observed potency against its homolog on the y-axis. The Spearman

Table 4.10: Critical binding site residue positions for the ‘Canonical’ site definition for kinases.

Kinase binding model	residue positions
‘Canonical’ site definition	73, 77, 78, 87,
...	159,161,
...	227, 231, 254, 256,
...	318, 319, 320, 321, 322, 323,
...	354, 355, 356, 359,
...	494, 495, 497,
...	536,537

rank correlation coefficient ρ was calculated using the R function `cor.test`. For each combination, potency differences were calculated and plotted as probability density distributions. Probability density functions were calculated using the kernel density estimation function `density` with a gaussian kernel and the bandwidth selector `nrd0`. Bland-Altman plots were constructed following specifications from [Bland and Altman, 1986](#). Interval-lines indicate the 2.3% - 15.9% quantiles 84.1% - 97.7% quantiles, respectively. If the data were normally distributed, these quantiles would coincide with \pm one or \pm two standard deviations. Quantiles and the distribution median with associated error estimates were determined from one thousand iterations of a case resampling bootstrap. The bootstrap was implemented using the R library `boot`.

4.4.6.2 Data models

Using the observed potency differences as an input, probability densities for a Laplace distribution were simulated using the function `dlaplace()` from the R package `VGAM`. The general probability density function of the Laplace distribution is

$$P(x) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}, \quad (4.4)$$

where μ is a location parameter and b a scale parameter. For fitting to the sampled differences between assays as well as human-rat orthologs and human paralogs, the parameters μ and b were estimated with the analogic method, i.e. by applying the same function that determines parameters in the theoretical distribution to the actual data. The Laplace distribution relies on the median as an estimate of location and the mean

absolute deviation as an estimate of variance. Hence, I calculated these values for each data set and used them as parameters for the Laplacian model. The alternative model - a normal distribution - was constructed using the probability density distribution function

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, \quad (4.5)$$

where σ , the standard deviation, is the parameter of scale and μ the location parameter. As for the Laplacian distribution, I estimated the parameters using the analogic method. Hence, I determined μ as the mean of the observed distribution and σ as the standard deviation. To compare the observed distributions with the theoretical models, I generated q-q plots of the observed quantiles against quantiles that would be observed if the data were drawn from the theoretical distribution. Theoretical quantiles were calculated using the R functions `rnorm` and `rlaplace`. This part of the analysis is documented at https://github.com/fak/globalAnalysis/wiki/data_model.

4.4.6.3 Assessment of individual homologous pairs

The testing of individual homologous pairs was based on a two-sided Mann-Whitney U test. This non-parametric test was used to compare the distribution of observed potency differences for individual pairs with a control distribution sampled from inter-assay comparisons for identical targets. When analysing pairs of paralogs, the control distribution was sampled from inter-assay comparisons of human targets, and when analysing human-rat orthologs, the control distribution was sampled from a mixture of human and rat inter-assay comparisons. The distributions of potency differences for individual pairs were subjected to Mann-Whitney U tests (using the R function `wilcox.test`). The estimate for the difference of means between control and query distribution, as well as the associated p-value, were recorded for each homologous pair. Because p-values were used as a ranking criterion rather than as a means to attach significance, p-values were not corrected for multiple testing. The difference of means was plotted as effect size against the negative logarithm of the corresponding p-value in a volcano plot using the R package `ggplot2`. This part of the analysis is documented here https://github.com/fak/globalAnalysis/wiki/individual_homologs.

4.4.6.4 Potency differences and sequence identity

The relationship of sequence identity between homologous pairs and the potency differences observed between them for identical compounds was analysed on three levels: (i) Full sequence (ii) predicted Pfam domain mediating the interaction and (iii) binding site residues. Sequence identity measures were obtained as described in sections 4.4.4, 4.4.5, 4.4.6, respectively. The dependence between sequence identity and the absolute difference in compound binding was evaluated using Spearman's rank correlation coefficient ρ calculated with the R function `cor.test`. To visualise the relationship of sequence identity and the absolute difference in small molecule binding, density plots were generated detailing the sequence identity between pairs of human paralogs on the x-axis and the absolute difference in small molecule binding on the y-axis. This analysis is documented here <https://github.com/fak/globalAnalysis/wiki/sequence>.

4.4.7 Homology model of the HRH3 receptor

4.4.7.1 Preparation of model templates

A number of GPCR structures were accessible through PDBe at the time the study was conducted. Table 4.11 lists the highest resolution structure for each protein. From the available structures, we selected all GPCRs with aminergic ligands as described by Gloriam et al (Gloriam et al., 2009): The human β -adrenoceptor 2 (2rh1, 3nya, 3ny8, 3d4s), the turkey β -adrenoceptor 1 (2vt4), the human dopamine D3 receptor 3pbl and the human HRH1 (3rze). Even though additional structures were available for the β -adrenoceptor 2, these were not included as templates because they represent the receptor in an interaction with agonistic rather than antagonistic ligands.

4.4.7.2 Sequence alignment

Sequences for the model templates were obtained from PDBe. Some of the structures contain a part of the bacteriophage T4 lysozyme to stabilise the structure for crystallisation. Prior to the alignment, these residues were removed from the template sequences by hand. The initial alignment of model templates was extracted from a JOY alignment encompassing the entire range of known GPCR crystal structures. The sequences of the human and rat HRH3 were added and the alignment was refined manually. The refined alignment

Table 4.11: Crystall structures of G-protein coupled receptors available in PDBe at the time the study was conducted.

Id	Name	Ligand	Species	Resolution (Å)
1u19	Rhodopsin	RET	<i>B. taurus</i>	2.2
2z73	Rhodopsin	RET	<i>T. pacificus</i>	2.5
2rh1	β -adrenoceptor 2	CAU	<i>H.sapiens</i>	2.4
3nya	β -adrenoceptor 2	JTZ	<i>H.sapiens</i>	3.16
3ny8	β -adrenoceptor 2	JRZ	<i>H.sapiens</i>	2.8
3d4s	β -adrenoceptor 2	TIM	<i>H.sapiens</i>	2.8
2vt4	β -adrenoceptor 1	P32	<i>M. gallopavo</i>	2.7
3eml	adenosine A2a receptor	ZMA	<i>H.sapiens</i>	2.6
3odu	CXCR4 chemokine receptor	ITD	<i>H.sapiens</i>	2.5
3pbl	dopamine D3 receptor	ETQ	<i>H.sapiens</i>	2.9
3rze	histamine H1 receptor	5EH/D7V	<i>H.sapiens</i>	3.1

used to build the homology model is shown in Figure 4.16. The JOY alignment was generated by John Overington and Kazuyoshi Ikeda, an online version is available at <http://chembl.blogspot.co.uk/2012/12/gpcr-structure-human-par1-receptor.html>.

4.4.7.3 Model building and visualization

The models of the human and rat HRH3 were build using the MODELLER program (version mod9v8) (Sali and Blundell, 1993). The pdb files of all template structures were curated to remove coordinates for any residue that was not part of the alignment described in section 4.4.7.2. The parameters for the MODELLER program were set to include doxepin, the ligand from the HRH1 receptor template structure, using the `automodel` module. For both the human and rat HRH3, five models were generated and the model with the lowest `molpdf` score selected for visualisation and analysis. Models were aligned in two steps using functionality of the CCP4 molecular graphics program (Potterton et al., 2002). First, a rough alignment was obtained based on a secondary structure matching algorithm (Krissinel and Henrick, 2004) and in second step this alignment was refined using the ‘match close residues’ routine of the CCP4 molecular

graphics program. The root mean square difference between two models is defined as

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n ((p_{ix} - q_{ix})^2 + (p_{iy} - q_{iy})^2 + (p_{iz} - q_{iz})^2)}, \quad (4.6)$$

where p_i and q_i are vectors describing the position of the i -th backbone atom in three dimensions (x, y, z) and was determined using inbuilt functionality of the program. Distances between the modelled ligand doxepin and the mutated residue Ala119Thr were calculated using the geometry toolbox that is part of the program. The generated models and alignment files were made available through figshare (www.figshare.com) and assigned individual digital object identifiers (DOI) as summarised in table 4.12.

Table 4.12: Overview of deposited model and alignment files.

doi	description
10.6084/m9.figshare.915370	Homology model of human HRH3.
10.6084/m9.figshare.915371	Homology model of rat HRH3.
10.6084/m9.figshare.915368	Alignment of template sequences and human HRH3.
10.6084/m9.figshare.915369	Alignment of template sequences and rat HRH3.

4.4.8 Cluster analysis of HRH3 ligands

The clustering of HRH3 ligands was based on LINGO fingerprints. LINGO fingerprints have been described as a simple and efficient method to capture and represent molecular properties directly from a SMILES string representation (Vidal et al., 2005; Grant et al., 2006). I used the Open Eye toolkit OEChem TK (OEChem, 2006) to calculate LINGO fingerprints. The term I used to evaluate pairwise distances between compounds is the Tanimoto coefficient, defined as

$$T_c = \frac{f_{i \wedge j}}{f_i + f_j + f_{i \wedge j}}, \quad (4.7)$$

where $f_{i \wedge j}$ is the number of features two compounds i, j have in common, while f_i and f_j represent the number of features found exclusively in one or the other compound (Jaccard, 1901, see also <http://www.daylight.com/dayhtml/doc/theory/theory.finger>).

[html](#)). This term was calculated for all pairwise combinations using the OEChem TK function `OEAnimoto`. An implementation of the single linkage algorithm from the Python module `hcluster` (Eads, 2008) was used to calculate a dendrogram from the input distance matrix. To assign compounds to clusters, a T_c similarity threshold of 0.5 was set. This value was chosen arbitrarily upon inspection of the dendrogram, with the aim of obtaining between five and ten major clusters.

4.4.9 Mapping small molecule binding to Pfam domains

To determine the Pfam domain most likely to mediate small molecule binding, I used an adaptation of the mapping described in Chapter 2. In contrast to the heuristic presented there, if a protein was found to contain more than one domain from the dictionary of validated domains, the binding site was mapped to the domain with the highest count of occurrence among single-domain targets.

Chapter 5

Conclusions

During my Ph.D. project at the European Bioinformatics Institute, I have developed approaches to integrate an important source of small molecule bioactivity data, the ChEMBL database, with sources of protein evolutionary relationship data. To achieve this, I have used computational methodologies that would typically be attributed to the disciplines of data mining and statistical analysis. Where required, specialised bioinformatics methods and programs were used to process biological and chemical information. In the following, I summarise and discuss the findings of my studies and provide an outlook onto future avenues of research.

5.1 Integration of small molecule bioactivity data and protein domain annotation

Chapters 2 and 3 describe the implementation of a heuristic mapping of small molecule binding to protein domains. In an evaluation against crystallographic data extracted from PDBe, the heuristic accurately predicted the correct domain of ligand binding in 97% of cases for single domain proteins and 89% of cases for multi-domain proteins. I was able to demonstrate that most measured small molecule interactions are limited to a small number of domains that are associated with the great target classes in drug discovery: GPCRs, kinases, ion channels and a number of enzymes. The analysis also provided statistics for small molecule binding at domain interfaces and delivered detailed insights into the patterns of co-occurrence of domains in potential drug targets. The refined mapping of small molecule binding to protein domains is now an integral part of

the ChEMBL database, where it provides the community an improved index for sequence- and protein family-based queries.

The mapping heuristic and its manual refinement presented in chapter 2 and chapter 3 constitute an approach of annotation transfer between related protein sequences. Annotation transfer was proposed as a method to annotate genes in newly sequenced genomes (Shah and Hunter, 1997; Bork et al., 1998). In these approaches, annotation that is present for one gene is transferred onto a newly sequenced gene if the two genes are deemed to be sufficiently related. In the case of my proposed mapping heuristic, the information being transferred is the presence or absence of a small molecule binding site and its relationship established through Pfam-A domains. Like other implementations of annotation transfer, this approach is subject to a number of limitations. Functional divergence between related protein sequences can lead to the loss or reconfiguration of a specific function. It has been shown that variation in the configuration of small molecule binding sites increments with sequence divergence (Devos and Valencia, 2000) and it is well appreciated that protein structure, in this mapping approximated by Pfam-A domains, is more conserved than protein function (Hegyí and Gerstein, 1999). Given that the annotation transferred in this instance is simply the presence or absence of a binding site, the mapping heuristic is expected to be relatively robust to functional divergence, which is typically more subtle. Another limitation to annotation transfer approaches is the complexity introduced by multi-domain architectures (Hegyí and Gerstein, 2001; Rost, 2002). This was clearly reflected in the difference in performance between single- and multi-domain proteins in the case of the mapping heuristic presented here. A third limitation is of a more speculative nature: the dictionary of small molecule binding domains is constructed based on measured interactions with synthetic small molecules. If it turned out that potent small molecule ligands could be found for all domains that are presently not part of the dictionary of validated domains, this would make the mapping pointless as each multi-domain protein would be put forward for curation. Experience from drug discovery suggests that this is not a likely scenario (Hopkins and Groom, 2002; Russ and Lampel, 2005). However, the progress that has been made towards the development of ligands for the SH2 domain (Kraskouskaya et al., 2013) is a cautionary call to monitor future developments.

Most other approaches of mapping small molecule binding to protein domains rely on protein structural information and do not infer binding sites by annotation transfer

(Snyder et al., 2006; Bashton and Thornton, 2010; Wang et al., 2012b; Moya-García and Ranea, 2013). The advantage here is that mappings do not rely on inference and can be expected to be more accurate. However, the annotation transfer-based approach provides the benefit that the mapping can be applied to any given protein, including those for which no crystal structure exists to date. One approach of mapping small molecule binding sites that uses annotation transfer was published by Davies in 2011 (Davis, 2011). This approach assigns potential small molecule binding sites to sequence regions that match precalculated profiles of existing small molecule binding sites. This method is more sophisticated than the method I propose because it takes into account the local context of each binding site. However, the precalculated profiles of binding sites require at least one crystal structure describing this site. The only requirement of the mapping heuristic presented in this thesis is that an interaction between a small molecule and a protein domain has been measured at least once and that protein domains are detectable using Pfam-A domain models. It is thus well suited to map small molecule binding to the heterogeneous set of protein targets with measured interactions in the ChEMBL database.

In its findings of frequently targeted domain types, the mapping heuristic agrees to the largest extent with previously published studies (Hopkins and Groom, 2002; Paolini et al., 2006; Overington et al., 2006). However, owing to its wide scope, it also maps small molecule binding to domain types that were not considered in previous studies of this type. Such domains contribute only a tiny fraction of measured interactions individually, but in the sum they actually form a significant portion of measured interactions. This long tail of domains can help generate hypotheses for novel drug targets.

5.2 Integration of small molecule bioactivity data and protein homology information

In chapter 4, I have presented an approach that integrates small molecule bioactivity data from the ChEMBL database with protein homology information obtained from EnsemblCompara Genetrees. The findings of this approach were that ChEMBL contains a wealth of small molecule potency data measured against human proteins that could be compared with potency data measured against orthologous proteins in other species. I found that the overlap of small molecules tested against both human and rat proteins

was sufficient to carry out a large scale analysis of the conservation of small molecule binding between the two species. In general, susceptibility to small molecule binding between the two species is conserved, as the overall differences are no greater than what would be expected from measurements of the experimental error. I also identified pairs of proteins for which small molecule binding differed between the two species, most prominently the HRH3. For this receptor, I proposed a probable mutation explaining the observed difference between the two species. In analogy to the analysis of orthologs, I used annotations from EnsemblCompara Genetrees to identify pairs of closely related paralogs and found that overall differences in susceptibility were greater than the experimental error noise and greater than differences observed between orthologs. I observed a weak relationship between sequence identity and differences in susceptibility to small molecule binding. The strength of this relationship increased with the level of resolution of the comparison, but in general terms remained weak, even on a binding site level.

Broadly speaking, this study follows a long tradition of molecular level evaluations of functional conservation between species. In 1983, Max Perutz laid the foundation for this type of work with a study on hemoglobin, which compared functional properties of this protein between mammalian, amphibian and fish species and identified the structural correlates of these differences (Perutz, 1983). The analysis presented here differs from Perutz's work and its successors in its scale, as differences were studied on about 150 orthologous pairs and 650 paralogous pairs. An added contrast is the nature of the measured entity in this study. It is not a physiological function, but rather the susceptibility of a protein to perturbation by artificial small molecules that played no role in the evolution of a protein. This somewhat limits the authority of my analysis to answer evolutionary questions, but also gives rise to some speculative questions concerning the nature of (artificial) small molecule-protein interactions. Do they mostly emulate interactions with natural binding partners, and, if so, what is the probability of impacting protein function through a mechanism that is not part of the evolved repertoire of mechanistic responses of a protein? The analysis presented here cannot answer these questions, but they are within the wider remit of pharmacophylogenetic approaches (Searls, 2003b).

A recent debate quite close to the subject matter of this study concerns the validity of the ortholog conjecture. The ortholog conjecture states that paralogs, which arise from gene duplication, are free to evolve divergent functions, while the function of orthologs

diverges at a much slower rate and is largely conserved between closely related species. The conjecture was derived from first principles stated by Fitch in 1970 (Fitch, 1970). A study seeking to trace functional divergence in duplicated genes found that functional divergence is a rare event and that silencing of a duplicated gene is a much more frequent outcome (Lynch, 2000). In 2009, Studer and colleagues pointed out that there was little systematic evidence to support the ortholog conjecture, despite its compelling theory (Studer and Robinson-Rechavi, 2009). A study by Nehrt and colleagues challenged the ortholog conjecture, with its finding that paralogs share more gene ontology (GO) annotation terms than orthologs, on average. In contrast, the findings from my analysis showed greater conservation between orthologs in terms of susceptibility to small molecule perturbations. In their study, Nehrt and colleagues had corrected for different levels of sequence identity before comparing overlap of GO terms. For lack of orthologous pairs with low levels of sequence identity, this was not possible in my analysis. That, and the artificial nature of the interactions that were the subject of my analysis, made it impossible to compare our findings directly. Later in 2012, Altenhoff and colleagues published a study that identified authorship bias as a confounding factor in the analysis of Nehrt and colleagues and concluded that functional conservation, measured as overlap of GO terms, is slightly greater in orthologs. To say I contributed to this debate would overstate my role, but from enjoyable and insightful conversations with the authors of both studies it appears that integration of small molecule bioactivity data can be applied to examine fundamental evolutionary processes. The finding from this study, that response to small molecule perturbation is more conserved in orthologs, is also in agreement with an experimental study that evaluated differences in response to small molecule binding to odorant receptors (Adipietro et al., 2012). Having stated that differences between human paralogs were greater than the overall assay-noise, it is nevertheless remarkable that comparable potencies were observed for many combinations of ligands and paralogous pairs. This observation might indicate a pronounced degree of functional promiscuity between paralogous proteins, which has previously been described as an essential feature of biological systems (Nobeli et al., 2009).

Global analyses of large-scale chemical and biological experiments were proposed as early as 2002 (Root et al., 2002). The authors of this proposal identified the capacity to detect subtle, non-obvious relationships in such data as one of the strengths of such analyses. To some extent this applies to my study, which is not an analysis of one large-scale

experiment, but rather a large-scale aggregation of multiple, heterogeneous experiments collected from the scientific literature and aggregated in the ChEMBL database. To account for the background noise created by this heterogeneity, I constructed a distribution of inter-assay differences. In a similar approach, Kramer and colleagues published a study that examined the variation in heterogeneous measurements of inhibition constants K_i (Kramer et al., 2012). Kramer and I had been corresponding prior to the publication of this study and we agreed that it would be important to exclude duplicate measurements from such evaluations as they would artificially reduce variability. In my study, I excluded all values that are exactly equal, as it is very unlikely that these would be independent measurements. Kramer and colleagues took this filtering process further and excluded also values that are only slightly different and are likely rounded duplicates of other measurements. To fully rule out dependence of measurements, Kramer further excluded measurements from publications with overlapping authors. Thus, Kramer’s analysis provides a true estimate of inter-assay noise, while my analysis provides a useful standard of comparison for differences observed between orthologs and paralogs.

Assay noise in itself is a clear limitation to the analytic capabilities of this study and it is likely that the correlation between sequence identity and conservation of susceptibility to small molecule perturbation would have been stronger in a dataset with less inherent noise. In the initial implementation of the analysis, this was further aggravated by the uncorrected comparison of IC50 and K_i -values (see section 4.2.1), but in the analysis presented here, K_i -values have been adjusted using a correction factor for ChEMBL bioactivity data published by Kalliokoski in 2013 (Kalliokoski et al., 2013).

This study constitutes a comparison of small molecule response in two species that are highly relevant to drug discovery. The finding that small molecule response is largely conserved between these two species on a molecular level does not come as a surprise, but had not been demonstrated systematically before. It should however not be interpreted to signify pharmacological equivalence between the two species. Comparisons on a cellular and organismal level are far more complex. For example, a comparison of human and mouse tissues demonstrated that transcriptional regulation is species-specific for the majority of transcribed modules (Odom et al., 2007) and there are serious concerns about the use of laboratory rodents as controls for disease models (Martin et al., 2010).

5.3 Summary, conclusion and outlook

In my thesis project, I have applied a data integration approach to bridge the gap between bioactivity data stored in the ChEMBL database and protein evolutionary information from two resources, Pfam and EnsemblCompara GeneTrees. I have demonstrated that this approach can improve indexing and organisation of bioactivity data and facilitate analyses at the interface of chemical biology, drug discovery and protein evolution.

Future studies can build on this approach to further integration, for example with genomic resources such as the 1,000 Genomes Project (Abecasis et al., 2010) or the clinical trials resource ClinicalTrials.gov to extend the bioactivity data obtained mainly from early stage drug discovery projects into the clinical stage.

Now that it is accessible through the ChEMBL schema, the mapping of small molecule binding to protein domains warrants further analysis, for example of the distribution of measured interactions per domain type. Eventually, the development of these distributions over time can reveal trends in the development of candidate drug targets. Ligand sets associated through the mapping with individual domain types can be analysed for their chemical properties and utilised to train models for target prediction as previously demonstrated (Bender et al., 2009). The manual curation interface will be made accessible through a public webserver which hopefully can engage the scientific community to participate in the curation effort and ensure a legacy for this mapping. These efforts can take inspiration from the wikipedia integration of the Pfam database and others (Mons et al., 2008; Huss et al., 2008; Punta et al., 2012).

Having shown that integration of small molecule bioactivity data can answer questions concerned with protein evolution it would be desirable to obtain the data needed to extend this study to proteins from other species. This would allow to trace differences in susceptibility to small molecule along a phylogenetic tree and could provide a trajectory of mutations that influence small molecule binding. This in turn could instruct both target validation efforts as well as efforts to optimise drug discovery lead structures.

A study making use of patient record data to predict off-target effects of small molecule drugs shows that data integration approaches can solve questions relevant to clinical practice (Tatonetti et al., 2012). To bridge the gap between the fundamental science of small molecule-protein interactions and clinical applications is maybe the most challenging, but also most alluring step forward.

List of Publications

- Bento A.P., Gaulton A., Anne Hersey A., Louisa J. Bellis L.J., Chambers J., Davies M., Kruger F.A., Light Y., Mak L., McGlinchey S., Nowotka M., Papadatos G., Santos R., Overington J.P. The ChEMBL Bioactivity database: an update. *under Review at NAR Database Issue*
- Kruger F.A., Rostom R., Overington J. P. Mapping small molecule binding data to structural domains. *BMC Bioinf* 13(Suppl 17): S11, 2012.
- Kruger F.A., Overington J. P. Global Analysis of Small Molecule Binding to Related Protein Targets. *PloS Comp Biol* 8(1): e1002333, 2012.
- Bellis L.J., Akhtar R., ALLazikani B, Atkinson F, Bento A.P., Chambers J., Davies M., Gaulton A., Hersey A., Ikeda K., Kruger F.A., Light Y., McGlinchey S., Santos R., Stauch B., Overington J.P. Collation and data-mining of literature bioactivity data for drug discovery *Biochem Soc Trans* 39(5): 1365-70 2011.

Appendix

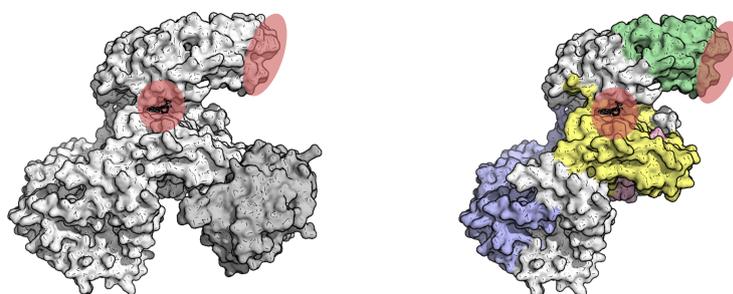


Figure 1: Multiple binding sites of the mTOR complex. Left hand side shows structures of whole protein complexes, right-hand side shows individual chains. Small molecule binding sites are shown in approximation as red circles overlaid with ligands from the crystal structure. PDB 3kg2 - complex of truncated human mTOR (light-gray) and mammalian lethal with SEC13 protein 8 (gray). mTOR contains a Rapamycin_bind domain (green, binding site for rapamycin/FKBP12 represented by red ellipse), a PI3_PI4_kinase (yellow, with torin-2) and a FAT (blue) and FATC domain (pink).

Complete catalogue of Pfam-A domains obtained from initial mapping:

- | | | |
|-------------------|-------------------|-------------------|
| 1. 2OG-FeII_Oxy_3 | 5. Abhydrolase_1 | 9. Acetyltransf_2 |
| 2. 3Beta_HSD | 6. Abhydrolase_6 | 10. ACPS |
| 3. 7tm_1 | 7. Abi | 11. Actin |
| 4. AA_permease_2 | 8. Acetyltransf_1 | 12. A_deaminase |

13. Adhes-Ig_like	35. Asp	57. DAO
14. adh_short	36. Asp_Glu_race	58. dCMP_cyt_deam_1
15. adh_short_C2	37. ATP-synt_C	59. DHFR_1
16. Aldedh	38. Bcl-2	60. DHO_dh
17. Aldo_ket_red	39. Bcl-2_BAD	61. DHquinase_II
18. Alk_phosphatase	40. Beta-lactamase	62. DMRL_synthase
19. Alpha-amylase	41. Beta-lactamase2	63. DNA_methylase
20. Alpha_L_fucos	42. BNR_2	64. DNA_pol_A
21. Amidase	43. BsuBI_PstI_RE	65. DNA_pol_lambd_f
22. Amidinotransf	44. Calc_CGRP_IAPP	66. dNK
23. Amino_oxidase	45. Carb_anhydrase	67. DSPc
24. Aminotran_1_2	46. CBAH	68. dUTPase
25. Aminotran_3	47. CD36	69. dUTPase_2
26. Aminotran_4	48. Choline_kinase	70. EBP
27. Aminotran_5	49. Clat_adaptor_s	71. ELO
28. AMP-binding	50. CM_2	72. EPSP_synthase
29. AMPKBI	51. COesterase	73. ERG2_Sigma1R
30. ANF_receptor	52. COX1	74. ERO1
31. An_peroxidase	53. CTP_transf_2	75. Esterase
32. Aph-1	54. Cupin_8	76. Exo_endo_phos
33. Arginase	55. DAGK_cat	77. FabA
34. ASC	56. DAHP_synth_1	78. FAD_binding_3
		79. FA_desaturase

80. FBPase	102. Glyco_transf_29	124. IDO
81. FKBP_C	103. Glyco_transf_6	125. IF4E
82. Flavodoxin_2	104. Glyoxalase	126. IGPD
83. Flu_M2	105. HCV_capsid	127. IL5
84. Folate_carrier	106. HCV_NS4a	128. IL8
85. Folate_rec	107. Hemagglutinin	129. IMPDH
86. G-alpha	108. Heme_oxygenase	130. Ion_trans
87. GDA1_CD39	109. Herpes_TK	131. IU_nuc_hydro
88. GDE_C	110. His_Phos_2	132. KH_1
89. Glucan_synthase	111. Hist_deacetyl	133. Lactamase_B
90. Glyco_hydro_1	112. Histidinol_dh	134. Laminin_G_1
91. Glyco_hydro_14	113. HIT	135. Lectin_legB
92. Glyco_hydro_15	114. HN	136. Lipase_3
93. Glyco_hydro_18	115. Hormone_2	137. Lipocalin
94. Glyco_hydro_20	116. Hormone_recep	138. LMWPc
95. Glyco_hydro_30	117. H_PPase	139. LpxC
96. Glyco_hydro_35	118. HSP20	140. LuxS
97. Glyco_hydro_47	119. HSP70	141. MAPEG
98. Glyco_hydro_79n	120. HTH_18	142. MBOAT
99. Glyco_hydro_85	121. Hydrolase_4	143. MCD
100. Glycolytic	122. ICL	144. Melibiase
101. Glyco_transf_21	123. ICMT	145. Metallophos
		146. Metallothio

147. Methyltransf_3	169. PALP	191. Peptidase_S28
148. Methyltransf_31	170. Pantoate_ligase	192. Peptidase_S8
149. MetJ	171. PAP2	193. Peripla_BP_4
150. MFS_1	172. PDEase_I	194. PFK
151. MIF	173. PDGF	195. PfkB
152. Multi_Drug_Res	174. PEN-2	196. PGI
153. Myb_DNA-binding	175. PEPCK	197. Phospholip_A2_1
154. NAD_kinase	176. Pep_deformylase	198. Phospholip_A2_2
155. NAD_synthase	177. Peptidase_C1	199. Phosphorylase
156. NAGidase	178. Peptidase_C12	200. FIG-L
157. Na_H_Exchanger	179. Peptidase_C13	201. PKD_channel
158. NAPRTase	180. Peptidase_C14	202. Pkinase
159. Neur	181. Peptidase_C30	203. Pkinase_Tyr
160. Neur_chan_LBD	182. Peptidase_C48	204. PMI_typeI
161. NMO	183. Peptidase_M10	205. PMM
162. NNMT_PNMT_TEMT	184. Peptidase_M2	206. PNP_UDP_1
163. Nramp	185. Peptidase_M24	207. polyprenyl_synt
164. Nucleoside_tran	186. Peptidase_M28	208. Poxvirus
165. OMPdecase	187. Peptidase_M48	209. Prenyltransf
166. P2X_receptor	188. Peptidase_M84	210. Presenilin
167. p450	189. Peptidase_S10	211. Pribosyltran
168. PAF-AH_p_II	190. Peptidase_S13	212. Pro_CA
		213. Pro_isomerase

214. Pterin_bind	235. Serpin	256. TIM
215. PTR2	236. SHMT	257. TK
216. Put_Phosphatase	237. SIR2	258. TNF
217. Pyridoxal_deC	238. SKI	259. Transpeptidase
218. Ras	239. S-methyl_trans	260. Transthyretin
219. RdRP_3	240. SNF	261. tRNA-synt_1e
220. RE_HindIII	241. Sortase	262. Trp_dioxygenase
221. RE_ScaI	242. Spermine_synth	263. Trypsin
222. RHD	243. SQS_PSY	264. TspO_MBR
223. Rib_hydrolyase	244. SSF	265. Tubulin
224. RmlD_sub_bind	245. Steroid_dh	266. Tyr-DNA_phospho
225. RnaseA	246. Sugar_tr	267. Tyrosinase
226. RRM_1	247. Sulfatase	268. UDG
227. RrnaAD	248. Sulfotransfer_1	269. UDPGT
228. rve	249. Telo_bind	270. Urotensin_II
229. RVP	250. Tetraspanin	271. V_ATPase_I
230. SAM_decarbox	251. TetR_N	272. Vitellogenin_N
231. SBF	252. TGT	273. V-set
232. Scytalone_dh	253. Thy1	274. Y_phosphatase
233. SDF	254. Thymidylate_kin	275. zf-CCCH
234. SE	255. Thymidylat_synt	

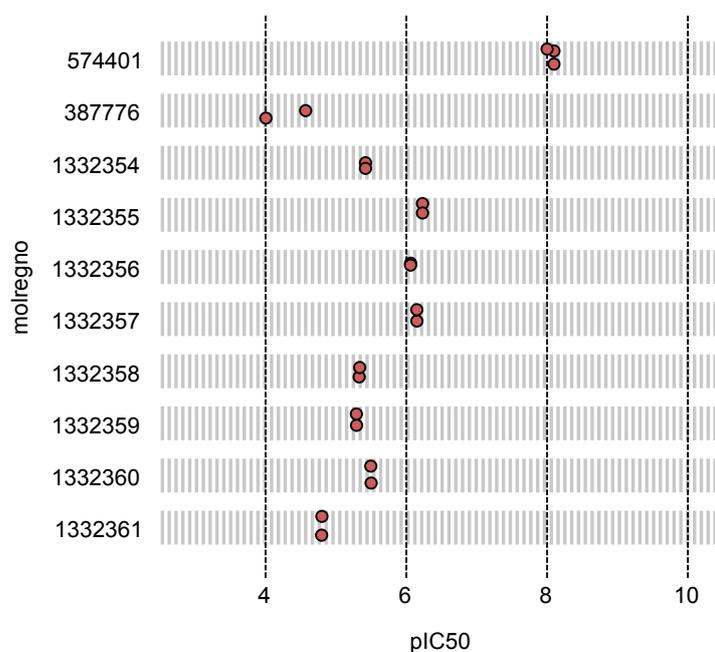


Figure 2: Evidence for small molecule binding to the *Hydrolase_4* domain. Measured interactions are shown for ten molecules. Molecules are identified as in the *molregno* database field and measured potencies expressed as pIC₅₀ values. Evidence for small molecule binding to this domain comes from 13 publications recorded in the ChEMBL database. In total, there were 124 compounds with 168 measured interactions for this domain. The protein target for these compounds is the Monoglyceride lipase from *M. musculus* and *R. norvegicus*. Interestingly, the human ortholog of this protein is assigned with the *Hydrolase_6* Pfam-A domain, rather than the *Hydrolase_4* domain. The *Hydrolase_4* Pfam-A model represents a putative lysophospholipase domain and is found in bacteria and eukaryotes. It shares the α - β hydrolase fold with a great number of other domains in the *AB_hydrolase* clan, members of which are thought to derive from a common ancestor.

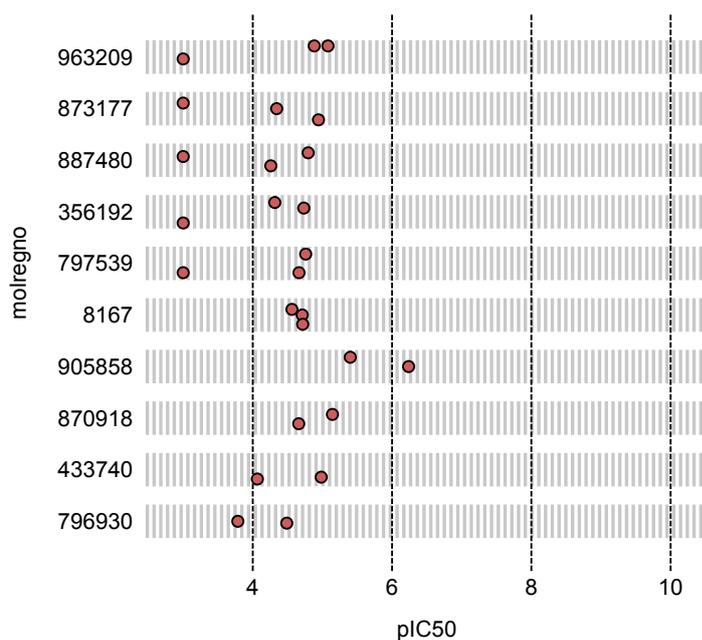


Figure 3: Evidence for small molecule binding to the HSP70 domain. Measured interactions are shown for ten molecules. Molecules are identified as in the molregno database field and measured potencies expressed as pIC50 values. Evidence for small molecule binding to this domain comes from two publications recorded in the ChEMBL database. In total, there were 79 compounds with 117 measured interactions for this domain. The protein target for these compounds are various isoforms of the Heat shock 70 kDa protein *H. sapiens*, *G. max* (soy bean) and *E.coli*.

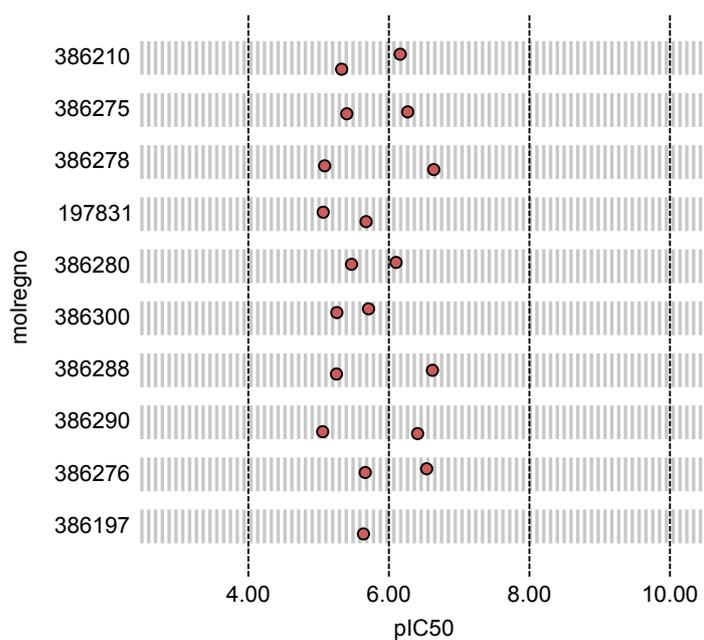


Figure 4: Evidence for small molecule binding to the PEPCK domain. Measured interactions are shown for ten molecules. Molecules are identified as in the `molregno` database field and measured potencies expressed as `pIC50` values. Evidence for small molecule binding to this domain comes from one publications recorded in the ChEMBL database. In total, there were 21 compounds with 30 measured interactions for this domain. The protein target for these compounds is the cytosolic phosphoenolpyruvate carboxykinase from *H. sapiens* and *R. norvegicus*.

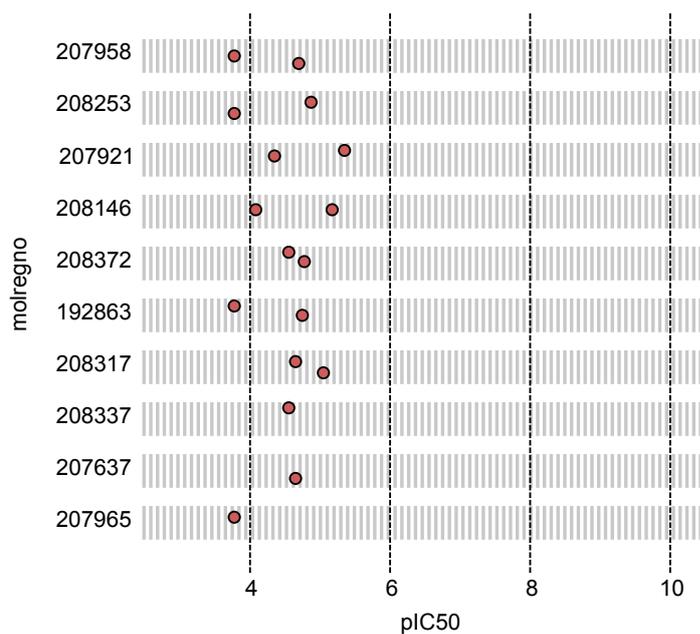


Figure 5: Evidence for small molecule binding to the RrnaAD domain. Measured interactions are shown for ten molecules. Molecules are identified as in the molregno database field and measured potencies expressed as pIC50 values. Evidence for small molecule binding to this domain comes from a single publication recorded in the ChEMBL database. In total, there were 16 compounds with 23 measured interactions for this domain. The protein target for these compounds is the rRNA adenine N-6-methyltransferase from *S. pneumoniae* and *B.subtilis*.

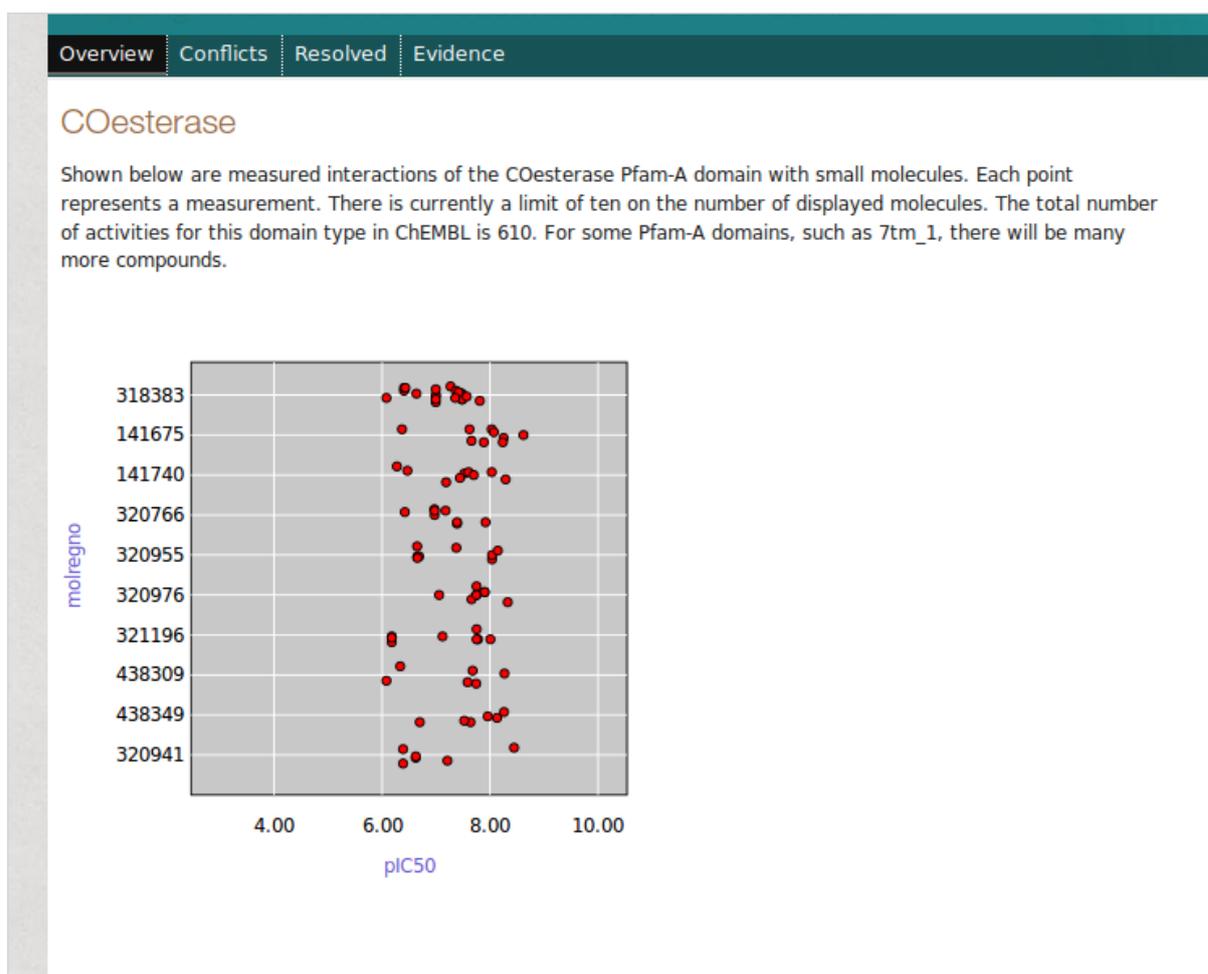


Figure 6: Screenshot of the curation interface. Section: 'Evidence'.

Overview Conflicts Resolved Evidence

Evidence for small molecule binding to Pfam-A domains

We query the ChEMBL database for measured interactions of small molecules with single domain proteins containing a given domain. The requirements for a measurement to be counted as evidence are then the following:

- o Activity type must be one of 'Ki', 'Kd', 'IC50', 'EC50', 'AC50' or logarithmic conversion.
- o The relationship in the database must be '='
- o The assay type must be assay_type = 'B'
- o The relationship type in the assay to target mapping must be relationship_type = 'D'
- o The potency threshold was set at 1 μ M.

Evidence for small molecule binding to 249 Pfam-A domains is listed below.

- o [2OG-Fell Oxy 3](#)
- o [3Beta HSD](#)
- o [7tm 1](#)
- o [7tm 2](#)
- o [7tm 3](#)
- o [AA permease 2](#)
- o [ABC membrane](#)
- o [ACPS](#)
- o [AMP-binding](#)
- o [AMPKBI](#)
- o [ANF receptor](#)
- o [ASC](#)
- o [ATP-synt C](#)
- o [A deaminase](#)
- o [Abhydrolase 1](#)
- o [Abhydrolase 6](#)
- o [Abi](#)
- o [Acetyltransf 1](#)
- o [Acetyltransf 2](#)

Figure 7: Screenshot of the curation interface. Index page for 'Evidence' section.

Overview Conflicts Resolved Evidence

Page 1 of 143. [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#) [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#) [56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [104](#) [105](#) [106](#) [107](#) [108](#) [109](#) [110](#) [111](#) [112](#) [113](#) [114](#) [115](#) [116](#) [117](#) [118](#) [119](#) [120](#) [121](#) [122](#) [123](#) [124](#) [125](#) [126](#) [127](#) [128](#) [129](#) [130](#) [131](#) [132](#) [133](#) [134](#) [135](#) [136](#) [137](#) [138](#) [139](#) [140](#) [141](#) [142](#) [143](#) [next](#)

Conflicting architecture: 7tm_3 vs. ANF_receptor

Assay-id: CHEMBL826438

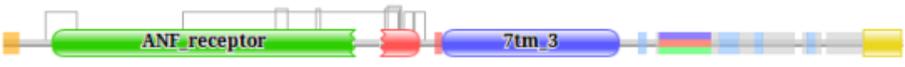
Pubmed-ID: [15482907](#) .

Description: Antagonistic activity against Metabotropic glutamate receptor 5 in [Ca²⁺] flux assay using glutamate (10 uM) as agonist

Last edit on 06 Aug 2013 14:04:55: *originates from initial implementation: <http://www.biomedcentral.com/1471-2105/13/S17/S11/>*

Assigned pref_name of the assay target - **Metabotropic glutamate receptor 5**
Metabotropic glutamate receptor 5 is composed of the following chain(s):

- [P31424](#)



7tm_3 ANF_receptor

Comment:

i Questions

If you would like to know more, please contact [Felix](#)

Figure 8: Screenshot of the curation interface. Section: 'Conflicts'

Overview
Conflicts
Resolved
Evidence

There are 333665 activities that could be mapped without conflicts. These are activities from binding assays of the type 'Ki', 'Kd', 'IC50', 'EC50' or logarithmic conversions thereof. In theory, the mapping could be extended to other activity types, as long as the mapping to the target is a direct mapping in the assay.

There are 12322 activities where the mapping defaulted to more than one domain type. Of these, 9378 have been manually resolved.

- [7tm_3 vs. ANF_receptor \(1564 activities\)](#)
- [ANF_receptor vs. Lig_chan \(1380 activities\)](#)

There are also 14442 activities where the mapping defaulted to multiple instances of the same domain type.

 **Questions**

If you would like to know more, please contact [Felix](#)

 <p>News Brochures Contact us Intranet</p>	<p>Services</p> <hr style="border: 0.5px solid #006666;"/> <p>By topic By name (A-Z) Help & Support</p>	<p>Research</p> <hr style="border: 0.5px solid #006666;"/> <p>Overview Publications Research groups Postdocs & PhDs</p>	<p>Training</p> <hr style="border: 0.5px solid #FFD700;"/> <p>Overview Train at EBI Train outside EBI Train online Contact organisers</p>
-------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44

Figure 9: Screenshot of the curation interface. Index page for ‘Conflicts’ section.

Overview Conflicts Resolved Evidence

previous Page 117 of 184. [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#) [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#) [56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [104](#) [105](#) [106](#) [107](#) [108](#) [109](#) [110](#) [111](#) [112](#) [113](#) [114](#) [115](#) [116](#) [117](#) [118](#) [119](#) [120](#) [121](#) [122](#) [123](#) [124](#) [125](#) [126](#) [127](#) [128](#) [129](#) [130](#) [131](#) [132](#) [133](#) [134](#) [135](#) [136](#) [137](#) [138](#) [139](#) [140](#) [141](#) [142](#) [143](#) [144](#) [145](#) [146](#) [147](#) [148](#) [149](#) [150](#) [151](#) [152](#) [153](#) [154](#) [155](#) [156](#) [157](#) [158](#) [159](#) [160](#) [161](#) [162](#) [163](#) [164](#) [165](#) [166](#) [167](#) [168](#) [169](#) [170](#) [171](#) [172](#) [173](#) [174](#) [175](#) [176](#) [177](#) [178](#) [179](#) [180](#) [181](#) [182](#) [183](#) [184](#) next

Conflicting architecture: 7tm_3 vs. ANF_receptor

Assay-id: CHEMBL1645999

Pubmed-ID: [21126874](#) .

Description: Antagonist activity at rat mGluR5 expressed in HEK293 cells assessed as inhibition of L-glutamate-induced calcium mobilization by FLIPR assay

Last edit on 12 August 2013 15:22:16: *paper describes allosteric modulators*

Assigned pref_name of the assay target - **Metabotropic glutamate receptor 5**
Metabotropic glutamate receptor 5 is composed of the following chain(s):

- o [P31424](#)



Comment:

i Questions

If you would like to know more, please contact [Felix](#)

Figure 10: Screenshot of the curation interface. Section: 'Resolved'

Overview Conflicts Resolved Evidence

A number of architectures caused conflicting mappings. Of these, 9378 mappings were resolved manually. Associated activities are summarised here, broken down by architecture type:

- o [ANF_receptor vs. Pkinase_Tyr](#) (10 activities)
- o [RVP vs. rve](#) (8 activities)
- o [Pkinase_Tyr vs. SH2](#) (4931 activities)
- o [ANF_receptor vs. Lig_chan](#) (608 activities)
- o [7tm_3 vs. ANF_receptor](#) (2329 activities)
- o [HCV_capsid vs. RdRP_3](#) (35 activities)
- o [SH2 vs. Y_phosphatase](#) (284 activities)
- o [Carb_anhydrase vs. Y_phosphatase](#) (3 activities)
- o [DHFR_1 vs. Thymidylat_synt](#) (1163 activities)
- o [OMPdecase vs. Pribosyltran](#) (7 activities)

i Questions

If you would like to know more, please contact [Felix](#)

EMBL-EBI 

News
Brochures
Contact us
Intranet

Services

By topic
By name (A-Z)
Help & Support

Research

Overview
Publications
Research groups
Postdocs & PhDs

Training

Overview
Train at EBI
Train outside EBI
Train online
Contact organisers

Figure 11: Screenshot of the curation interface. Index page for ‘Resolved’ section.

Table 1: Pfam-A domain types projected onto multi-domain proteins. The table summarises all Pfam-A domain types from the catalogue of Pfam-A domains with known small molecule interactions that were projected onto multi-domain proteins. The column headed ‘n(proteins)’ provides the number of multi-domain proteins a given Pfam-A domain was projected onto, the column headed ‘n(activities)’ the number of corresponding measured activities.

domain id	proteins	activities
Pkinase_Tyr	90	9,877
Pkinase	75	4,027
ANF_receptor	48	2,358
Ion_trans	46	3,889
Hormone_recep	45	7,386
Neur_chan_LBD	33	1,163
Peptidase_C1	30	4,069
Trypsin	26	3,808
Peptidase_M10	21	4,053
PDEase_I	17	724
7tm_1	13	1,184
Asp	10	1,603
COesterase	10	4,240
Peptidase_S8	10	208
Peptidase_C14	9	374
An_peroxidase	7	630
Y_phosphatase	7	207
Hist_deacetyl	6	594
Tubulin	5	54
Transpeptidase	4	7
Glyco_hydro_18	4	41
SNF	4	4,486
DSPc	3	92
Abhydrolase_1	3	183
Bcl-2	3	163
Metallophos	3	87

Continued on next page

Table 1 – *Continued from previous page*

domain id	proteins	activities
adh_short	2	47
Alpha-amylase	2	10
DNA_methylase	2	68
FA_desaturase	2	2
Glyoxalase	2	11
Hormone_2	2	8
Peptidase_M84	2	443
PKD_channel	2	47
adh_short_C2	2	193
FKBP_C	2	5
IMPDH	2	269
MFS_1	2	56
Peptidase_C30	2	113
Peptidase_M28	2	105
Amino_oxidase	1	14
Carb_anhydrase	1	41
Glucan_synthase	1	10
Glyco_hydro_15	1	2
ICL	1	10
Methyltransf_31	1	16
Pro_CA	1	99
RRM_1	1	112
SIR2	1	2
2OG-FeII_Oxy_3	1	88
Aminotran_1_2	1	10
AMP-binding	1	48
DNA_pol_A	1	47
Exo_endo_phos	1	6
Glyco_hydro_1	1	9
p450	1	52

Continued on next page

Table 1 – *Continued from previous page*

domain id	proteins	activities
polyprenyl_synt	1	21
RHD	1	5
rve	1	761

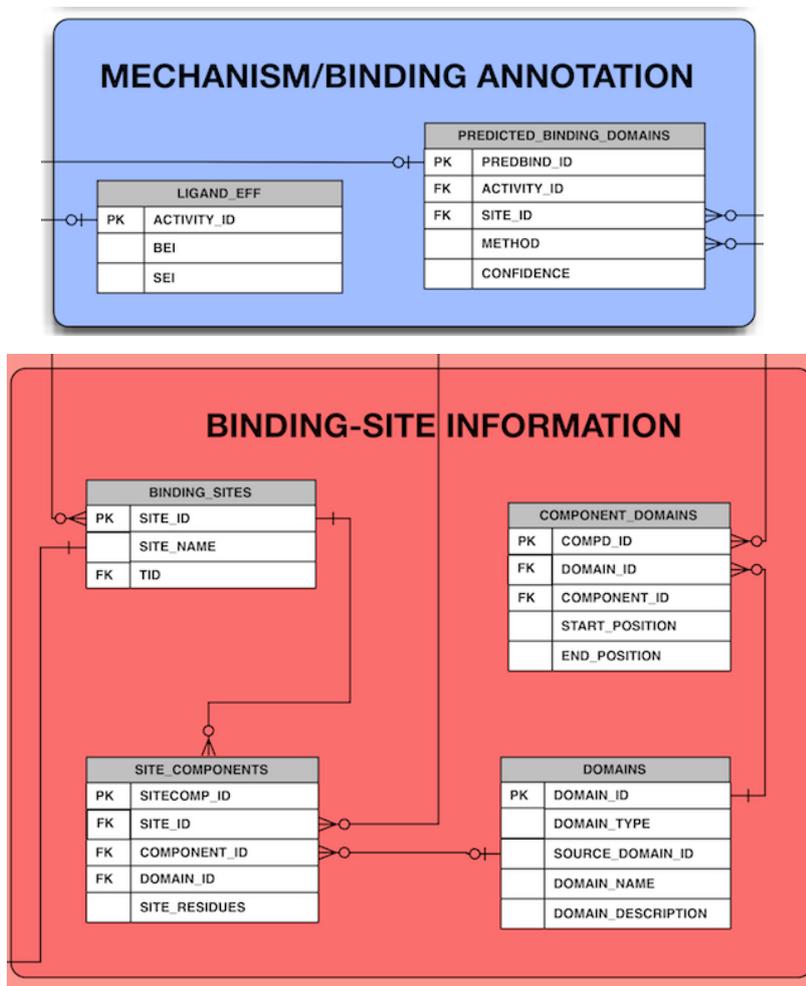


Figure 12: Schema sections representing the mapping of small molecule binding to Pfam-A domains in chembl_15 and upwards. Measured activities are linked to binding sites through the `activity_id` and `site_id` fields in the `predicted_binding_domains` table. Binding site can then further be linked to domain identifiers using the `site_id` and `domain_id` fields in the `site_components` table. Sites can also be mapped to protein targets using the `site_id` and `tid` fields in the `binding_sites` table.

Table 2: Paralogous groups. This table provides a summary of the groups of paralogs highlighted in Figure 4.5. The column groups are assigned according to the figure labels.

group	Uniprot ID	Name
1	Q9Y5N1	Histamine H3 receptor
1	P08173	Muscarinic acetylcholine receptor M4
1	P08172	Muscarinic acetylcholine receptor M2
1	P11229	Muscarinic acetylcholine receptor M1
1	P08912	Muscarinic acetylcholine receptor M5
1	Q9H3N8	Histamine H4 receptor
1	P35367	Histamine H1 receptor
1	P20309	Muscarinic acetylcholine receptor M3
2	P42681	Tyrosine-protein kinase TXK
2	P51813	Cytoplasmic tyrosine-protein kinase BMX
2	Q06187	Tyrosine-protein kinase BTK
2	P42680	Tyrosine-protein kinase Tec
2	Q13882	Protein-tyrosine kinase 6
2	Q08881	Tyrosine-protein kinase ITK/TSK
3	P25100	Alpha-1D adrenergic receptor
3	P35368	Alpha-1B adrenergic receptor
3	P21918	D(1B) dopamine receptor
3	P50406	5-hydroxytryptamine receptor 6
3	P35348	Alpha-1A adrenergic receptor
3	P08588	Beta-1 adrenergic receptor
3	P21728	D(1A) dopamine receptor
3	P07550	Beta-2 adrenergic receptor
4	P43088	Prostaglandin F2-alpha receptor
4	P34995	Prostaglandin E2 receptor EP1 subtype
4	P35408	Prostaglandin E2 receptor EP4 subtype
4	P43115	Prostaglandin E2 receptor EP3 subtype
4	P21731	Thromboxane A2 receptor
4	P43116	Prostaglandin E2 receptor EP2 subtype
4	Q13258	Prostaglandin D2 receptor
4	P43119	Prostacyclin receptor

References

- Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). “A map of human genome variation from population-scale sequencing.” *Nature* 467 (2010), pp. 1061–73.
- Adipietro, K. A., Mainland, J. D., and Matsunami, H. (2012). “Functional evolution of mammalian odorant receptors.” *PLoS Genet* 8 (2012). Ed. by J. Zhang, e1002821.
- Aldrich, T. B. (1905). “Adrenalin the active principle of the suprarenal glands”. *J Am Chem Soc* 27 (1905), pp. 1074–1091.
- Alouani, S. (2000). “Scintillation proximity binding assay.” *Mol. Biol.* 138 (2000), pp. 135–41.
- Altenhoff, A. M. and Dessimoz, C. (2009). “Phylogenetic and functional assessment of orthologs inference projects and methods.” *PLoS Comput Biol* 5 (2009), e1000262.
- Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). “Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs.” *PLoS Comput Biol* 8 (2012), e1002514.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). “Basic local alignment search tool.” *J Mol Biol* 215 (1990), pp. 403–10.
- Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H., and Peterson, J. R. (2011). “Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity.” *Nat Biotechnol* 29 (2011), pp. 1039–45.
- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). “SCOP database in 2004: refinements integrate structure and sequence family data.” *Nucleic Acids Res* 32 (2004), pp. D226–9.
- Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2008). “Data growth and its impact on the SCOP database: new developments.” *Nucleic Acids Res* 36 (2008), pp. D419–25.
- Apic, G., Gough, J., and Teichmann, S. A. (2001). “Domain combinations in archaeal, eubacterial and eukaryotic proteomes.” *J Mol Biol* 310 (2001), pp. 311–25.
- Armstrong, N. and Gouaux, E. (2000). “Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core.” *Neuron* 28 (2000), pp. 165–81.

- Arora, K. K., Chung, H. O., and Catt, K. J. (1999). "Influence of a species-specific extracellular amino acid on expression and function of the human gonadotropin-releasing hormone receptor." *Mol Endocrinol* 13 (1999), pp. 890–6.
- Austin, C. P. (2003). "The completed human genome: implications for chemical biology". *Curr Opin Chem Biol* 7 (2003), pp. 511–515.
- Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004). "NIH Molecular Libraries Initiative." *Science* 306 (2004), pp. 1138–9.
- Babine, R. E. and Bender, S. L. (1997). "Molecular Recognition of Protein - Ligand Complexes: Applications to Drug Design." *Chem Rev* 97 (1997), pp. 1359–1472.
- Bain, J., McLauchlan, H., Elliott, M., and Cohen, P. (2003). "The specificities of protein kinase inhibitors: an update." *Biochem J* 371 (2003), pp. 199–204.
- Bain, J., Plater, L., Elliott, M., Shpiro, N., Hastie, C. J., McLauchlan, H., Klevernic, I., Arthur, J. S. C., Alessi, D. R., and Cohen, P. (2007). "The selectivity of protein kinase inhibitors: a further update." *Biochem J* 408 (2007), pp. 297–315.
- Baneyx, F. (1999). "Recombinant protein expression in Escherichia coli." *Curr Opin Biotechnol* 10 (1999), pp. 411–21.
- Banks, P. and Harvey, M. (2002). "Considerations for using fluorescence polarization in the screening of G protein-coupled receptors." *J Biomol Screen* 7 (2002), pp. 111–7.
- Barger, G. and Dale, H. H. (1910). "Chemical structure and sympathomimetic action of amines." *J Physiol* 41 (1910), pp. 19–59.
- Barker, E. L., Kimmel, H. L., and Blakely, R. D. (1994). "Chimeric human and rat serotonin transporters reveal domains involved in recognition of transporter ligands." *Mol Pharmacol* 46 (1994), pp. 799–807.
- Bashton, M. and Chothia, C. (2007). "The generation of new protein functions by the combination of domains." *Structure* 15 (2007), pp. 85–99.
- Bashton, M. and Thornton, J. M. (2010). "Domain-ligand mapping for enzymes." *J Mol Recognit* 23 (2010), pp. 194–208.
- Bender, A., Mikhailov, D., Glick, M., Scheiber, J., Davies, J. W., Cleaver, S., Marshall, S., Tallarico, J. A., Harrington, E., Cornella-Taracido, I., and Jenkins, J. L. (2009). "Use of Ligand Based Models for Protein Domains To Predict Novel Molecular Targets and Applications To Triage Affinity Chromatography Data". *J Proteome Res* 8 (2009), pp. 2575–2585.
- Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." *Nucleic Acids Res* 35 (2007), pp. D301–3.
- Berriman, M. et al. (2009). "The genome of the blood fluke *Schistosoma mansoni*." *Nature* 460 (2009), pp. 352–8.
- Bishop, A. C., Ubersax, J. A., Petsch, D. T., Matheos, D. P., Gray, N. S., Blethrow, J., Shimizu, E., Tsien, J. Z., Schultz, P. G., Rose, M. D., Wood, J. L., Morgan, D. O.,

- and Shokat, K. M. (2000a). "A chemical switch for inhibitor-sensitive alleles of any protein kinase." *Nature* 407 (2000), pp. 395–401.
- Bishop, A., Buzko, O., Heyeck-Dumas, S., Jung, I., Kraybill, B., Liu, Y., Shah, K., Ulrich, S., Witucki, L., Yang, F., Zhang, C., and Shokat, K. M. (2000b). "Unnatural ligands for engineered proteins: new tools for chemical genetics." *Annu Rev Biophys Biomol Struct* 29 (2000), pp. 577–606.
- Black, J. W. and Stephenson, J. S. (1962). "Pharmacology of a new adrenergic beta-receptor-blocking compound (Nethalide)." *Lancet* 2 (1962), pp. 311–4.
- Black, L. A., Nersesian, D. L., Sharma, P., Ku, Y.-Y., Bennani, Y. L., Marsh, K. C., Miller, T. R., Esbenshade, T. A., Hancock, A. A., and Cowart, M. (2007). "4-[6-(2-Aminoethyl)naphthalen-2-yl]benzonitriles are potent histamine H3 receptor antagonists with high CNS penetration." *Bioorg Med Chem Lett* 17 (2007), pp. 1443–6.
- Blake, C. C., Mair, G. A., North, A. C., Phillips, D. C., and Sarma, V. R. (1967). "On the conformation of the hen egg-white lysozyme molecule." *Proc R Soc Lond B Biol Sci* 167 (1967), pp. 365–77.
- Bland, J. M. and Altman, D. G. (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." *Lancet* 1 (1986), pp. 307–10.
- Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). "Predicting function: from genes to genomes and back." *J Mol Biol* 283 (1998), pp. 707–25.
- Bostock, M., Ogievetsky, V., and Heer, J. (2011). "D3: Data-Driven Documents." *IEEE Trans Vis Comput Graph* 17 (2011), pp. 2301–9.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). "The evolution of gene expression levels in mammalian organs." *Nature* 478 (2011), pp. 343–8.
- Brazma, A. et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." *Nat Genet* 29 (2001), pp. 365–71.
- Bredel, M. and Jacoby, E. (2004). "Chemogenomics: an emerging strategy for rapid target and drug discovery." *Nat Rev Genet* 5 (2004), pp. 262–75.
- Breinbauer, R., Vetter, I. R., and Waldmann, H. (2002). "From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries." *Angew Chem, Int Ed Engl* 41 (2002), pp. 2879–90.
- Brenner, S. and Lerner, R. A. (1992). "Encoded combinatorial chemistry." *Proc Natl Acad Sci U S A* 89 (1992), pp. 5381–3.
- Brimblecombe, R. W., Duncan, W. A. M., Durant, G. J., Ganellin, C. R., Parsons, M. E., and Black, J. W. (2010). "The pharmacology of cimetidine, a new histamine H2-receptor antagonist. 1975." *Br J Pharmacol* 160 Suppl (2010), S52–3.

- Bromham, L. and Penny, D. (2003). “The modern molecular clock.” *Nat Rev Genet* 4 (2003), pp. 216–24.
- Bruns, R. F., Lawson-Wendling, K., and Pugsley, T. A. (1983). “A rapid filtration assay for soluble receptors using polyethylenimine-treated filters.” *Anal Biochem* 132 (1983), pp. 74–81.
- Buchdunger, E., Zimmermann, J., Mett, H., Meyer, T., Müller, M., Druker, B. J., and Lydon, N. B. (1996). “Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative.” *Cancer Res* 56 (1996), pp. 100–4.
- Buljan, M., Frankish, A., and Bateman, A. (2010). “Quantifying the mechanisms of domain gain in animal proteins.” *Genome Biol* 11 (2010), R74.
- Bunch, R. L. and McGuire, J. M. (1953). *Erythromycin, its salts, and method of preparation*. 1953.
- Burke, T. R., Smyth, M. S., Otaka, A., Nomizu, M., Roller, P. P., Wolf, G., Case, R., and Shoelson, S. E. (1994). “Nonhydrolyzable phosphotyrosyl mimetics for the preparation of phosphatase-resistant SH2 domain inhibitors.” *Biochemistry* 33 (1994), pp. 6490–4.
- Burns, D. M., Horn, V., Paluh, J., and Yanofsky, C. (1990). “Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains.” *J Biol Chem* 265 (1990), pp. 2060–9.
- Bylund, D. B. and Toews, M. L. (2011). *Receptor Signal Transduction Protocols*. Ed. by G. B. Willars and R. J. Challiss. Vol. 746. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2011, pp. 135–164.
- Caenepeel, S., Charyczak, G., Sudarsanam, S., Hunter, T., and Manning, G. (2004). “The mouse kinome: discovery and comparative genomics of all mouse protein kinases.” *Proc Natl Acad Sci U S A* 101 (2004), pp. 11707–12.
- Calleja, V., Alcor, D., Laguerre, M., Park, J., Vojnovic, B., Hemmings, B. A., Downward, J., Parker, P. J., and Larijani, B. (2007). “Intramolecular and intermolecular interactions of protein kinase B define its activation in vivo.” *PLoS Biol* 5 (2007), e95.
- Calleja, V., Laguerre, M., Parker, P. J., and Larijani, B. (2009). “Role of a novel PH-kinase domain interface in PKB/Akt regulation: structural mechanism for allosteric inhibition.” *PLoS Biol* 7 (2009), e17.
- Campillos, M., Kuhn, M., Gavin, A.-C. C., Jensen, L. J., and Bork, P. (2008). “Drug target identification using side-effect similarity”. *Science* 321 (2008), pp. 263–266.
- Cereghino, J. L. and Cregg, J. M. (2000). “Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*.” *FEMS Microbiol Rev* 24 (2000), pp. 45–66.
- Cha, H. J., Byrom, M., Mead, P. E., Ellington, A. D., Wallingford, J. B., and Marcotte, E. M. (2012). “Evolutionarily repurposed networks reveal the well-known antifungal drug thiabendazole to be a novel vascular disrupting agent.” *PLoS Biol* 10 (2012). Ed. by C. Khosla, e1001379.

- Chambers, C., Smith, F., Williams, C., Marcos, S., Liu, Z. H., Hayter, P., Ciaramella, G., Keighley, W., Gribbon, P., and Sewing, A. (2003). “Measuring intracellular calcium fluxes in high throughput mode.” *Comb Chem High Throughput Screening* 6 (2003), pp. 355–62.
- Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., and Overington, J. P. (2013). “UniChem: a unified chemical structure cross-referencing and identifier tracking system.” *J Cheminform* 5 (2013), p. 3.
- Chandler, D. (2005). “Interfaces and the driving force of hydrophobic assembly.” *Nature* 437 (2005), pp. 640–7.
- Chandonia, J.-M. and Brenner, S. E. (2006). “The impact of structural genomics: expectations and outcomes.” *Science* 311 (2006), pp. 347–51.
- Chen, B., Wild, D., and Guha, R. (2009). “PubChem as a source of polypharmacology.” *J Chem Inf Model* 49 (2009), pp. 2044–2055.
- Chen, J. J., Wang, S.-J., Tsai, C.-A., and Lin, C.-J. (2007). “Selection of differentially expressed genes in microarray data analysis.” *Pharmacogenomics J* 7 (2007), pp. 212–20.
- Chen, X. and Zhang, J. (2012). “The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data.” *PLoS Comput Biol* 8 (2012), e1002784.
- Cheng, Y. and Prusoff, W. H. (1973). “Relationship between the inhibition constant (K₁) and the concentration of inhibitor which causes 50 per cent inhibition (I₅₀) of an enzymatic reaction.” *Biochem Pharmacol* 22 (1973), pp. 3099–108.
- Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J. M., and Botstein, D. (1998). “Comparison of the complete protein sets of worm and yeast: orthology and divergence.” *Science* 282 (1998), pp. 2022–8.
- Chothia, C. (1984). “Principles that determine the structure of proteins”. *Annu Rev Biochem* (1984).
- Chothia, C. and Lesk, A. M. (1986). “The relation between the divergence of sequence and structure in proteins.” *EMBO J* 5 (1986), pp. 823–6.
- Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003). “Evolution of the protein repertoire.” *Science* 300 (2003), pp. 1701–3.
- Christopoulos, A. (2002). “Allosteric binding sites on cell-surface receptors: novel targets for drug discovery.” *Nat Rev Drug Discovery* 1 (2002), pp. 198–210.
- Ciulli, A., Scott, D. E., Ando, M., Reyes, F., Saldanha, S. A., Tuck, K. L., Chirgadze, D. Y., Blundell, T. L., and Abell, C. (2008). “Inhibition of Mycobacterium tuberculosis pantothenate synthetase by analogues of the reaction intermediate.” *ChemBioChem* 9 (2008), pp. 2606–11.
- Clark, A. J. (1933). *Mode of action of drugs on cells*. London: E. Arnold & Co, 1933.

- Codd, E. F. (1970). "A relational model of data for large shared data banks. 1970." *M D Comput* 15 (1970), pp. 162–6.
- Cohen, P. (2010). "Guidelines for the effective use of chemical inhibitors of protein function to understand their roles in cell regulation." *Biochem J* 425 (2010), pp. 53–4.
- Conant, G. C. and Wolfe, K. H. (2008). "Turning a hobby into a job: how duplicated genes find new functions." *Nat Rev Genet* 9 (2008), pp. 938–50.
- Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S., and Sidow, A. (2003). "Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes." *Genome Res* 13 (2003), pp. 813–20.
- Cuatrecasas, P., Wilchek, M., and Anfinsen, C. B. (1968). "Selective enzyme purification by affinity chromatography." *Proc Natl Acad Sci U S A* 61 (1968), pp. 636–43.
- Dale, H. (1950). "The pharmacology of histamine: with a brief survey of evidence for its occurrence, liberation, and participation in natural reactions." *Ann N Y Acad Sci* 50 (1950), pp. 1017–28.
- Darwin, C. (1859). *On the Origin of the Species by Means of Natural Selection: Or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.
- Davies, S. P., Reddy, H., Caivano, M., and Cohen, P. (2000). "Specificity and mechanism of action of some commonly used protein kinase inhibitors." *Biochem J* 351 (2000), pp. 95–105.
- Davis, A. M. and Teague, S. J. (1999). "Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis". *Angew Chem, Int Ed* 38 (1999), pp. 736–749.
- Davis, A. M., Teague, S. J., and Kleywegt, G. J. (2003). "Application and limitations of X-ray crystallographic data in structure-based ligand and drug design." *Angew Chem, Int Ed Engl* 42 (2003), pp. 2718–36.
- Davis, F. P. (2011). "Proteome-wide prediction of overlapping small molecule and protein binding sites using structure." *Mol BioSyst* 7 (2011), pp. 545–57.
- Davis, F. P. and Sali, A. (2010). "The overlap of small molecule and protein binding sites within families of protein structures." *PLoS Comput Biol* 6 (2010), e1000668.
- Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., and Zarrinkar, P. P. (2011). "Comprehensive analysis of kinase inhibitor selectivity." *Nat Biotechnol* 29 (2011), pp. 1046–51.
- Deininger, M. W. N. and Druker, B. J. (2003). "Specific targeted therapy of chronic myelogenous leukemia with imatinib." *Pharmacol Rev* 55 (2003), pp. 401–23.
- Delft, F. von, Lewendon, A., Dhanaraj, V., Blundell, T. L., Abell, C., and Smith, A. G. (2001). "The Crystal Structure of E. coli Pantothenate Synthetase Confirms It as a Member of the Cytidylyltransferase Superfamily". *Structure* 9 (2001), pp. 439–450.
- Devos, D. and Valencia, A. (2000). "Practical limits of function prediction." *Proteins* 41 (2000), pp. 98–107.

- Dickerson, R. E. (1971). "The structures of cytochrome c and the rates of molecular evolution." *J Mol Evol* 1 (1971), pp. 26–45.
- Django, (2011). *Django (Version 1.3)*. Lawrence, Kansas, 2011.
- Domagk, G. (1935). "Ein Beitrag zur Chemotherapie der bakteriellen Infektionen". *Deut Med Wochenschr* 61 (1935), pp. 250–258.
- Domchek, S. M., Auger, K. R., Chatterjee, S., Burke, T. R., and Shoelson, S. E. (1992). "Inhibition of SH2 domain/phosphoprotein association by a nonhydrolyzable phosphopeptide." *Biochemistry* 31 (1992), pp. 9865–70.
- Dongen, S. van (2000). "Graph Clustering by Flow Simulation". PhD thesis. 2000.
- Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., and Stein, L. (2001). "The distributed annotation system." *BMC Bioinf* 2 (2001), p. 7.
- Downes, M., Verdecia, M. A., Roecker, A. J., Hughes, R., Hogenesch, J. B., Kast-Woelbern, H. R., Bowman, M. E., Ferrer, J.-L., Anisfeld, A. M., Edwards, P. A., Rosenfeld, J. M., Alvarez, J. G. A., Noel, J. P., Nicolaou, K. C., and Evans, R. M. (2003). "A chemical, genetic, and structural analysis of the nuclear bile acid receptor FXR." *Mol Cell* 11 (2003), pp. 1079–92.
- Drews, J. (2000). "Drug Discovery: A Historical Perspective". *Science* 287 (2000), pp. 1960–1964.
- Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., and Sawyers, C. L. (2001). "Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia." *New Eng J Med* 344 (2001), pp. 1031–7.
- Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière, G. (2005). "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases." *Bioinformatics* 21 (2005), pp. 2596–603.
- Dunitz, J. D. (1995). "Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions." *Chem Biol* 2 (1995), pp. 709–12.
- Dunn, M. F. (2010). "Protein Ligand Interactions : General Description". *Electr Libr Sci* (2010), pp. 1–12.
- Eads, D. (2008). *scipy-cluster*. 2008.
- Eblen, S. T., Kumar, N. V., Shah, K., Henderson, M. J., Watts, C. K. W., Shokat, K. M., and Weber, M. J. (2003). "Identification of novel ERK2 substrates through use of an engineered kinase and ATP analogs." *J Biol Chem* 278 (2003), pp. 14926–35.
- Eckert, H. and Bajorath, J. (2007). "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches." *Drug Discov Today* 12 (2007), pp. 225–33.
- Eddy, S. R. (1998). "Profile hidden Markov models." *Bioinformatics* 14 (1998), pp. 755–63.
- Eddy, S. R. (2008). "A probabilistic model of local sequence alignment that simplifies statistical significance estimation." *PLoS Comput Biol* 4 (2008), e1000069.

- Edwards, A. M., Bountra, C., Kerr, D. J., and Willson, T. M. (2009). “Open access chemical and clinical probes to support drug discovery.” *Nat Chem Biol* 5 (2009), pp. 436–40.
- Ehrlich, J., Bartz, Q. R., Smith, R. M., Joslyn, D. A., and Burkholder, P. R. (1947). “Chloromycetin, a New Antibiotic From a Soil Actinomycete.” *Science* 106 (1947), p. 417.
- Ehrlich, P. and Hata, S. (1910). *Die experimentelle Chemotherapie der Sprillosen*. Berlin: Julius Springer, 1910, pp 1–164.
- Ehrlich, P. and Morgenroth, J. (1900). “Ueber Haemolysine. Dritte Mitteilung”. *Berl Klin Wochenschr* 37 (1900), pp. 453–8.
- Elber, R. and Karplus, M. (1987). “Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin.” *Science* 235 (1987), pp. 318–21.
- Enright, A. J., Iliopoulos, I., Kyripides, N. C., and Ouzounis, C. A. (1999). “Protein interaction maps for complete genomes based on gene fusion events.” *Nature* 402 (1999), pp. 86–90.
- Erion, M. D., Poelje, P. D. van, Dang, Q., Kasibhatla, S. R., Potter, S. C., Reddy, M. R., Reddy, K. R., Jiang, T., and Lipscomb, W. N. (2005). “MB06322 (CS-917): A potent and selective inhibitor of fructose 1,6-bisphosphatase for controlling gluconeogenesis in type 2 diabetes.” *Proc Natl Acad Sci U S A* 102 (2005), pp. 7970–5.
- Fabian, M. A. et al. (2005). “A small molecule-kinase interaction map for clinical kinase inhibitors”. *Nat Biotechnol* 23 (2005), pp. 329–336.
- Fedorov, O., Marsden, B., Pogacic, V., Rellos, P., Müller, S., Bullock, A. N., Schwaller, J., Sundström, M., and Knapp, S. (2007a). “A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases.” *Proc Natl Acad Sci U S A* 104 (2007), pp. 20523–8.
- Fedorov, O., Sundström, M., Marsden, B., and Knapp, S. (2007b). “Insights for the development of specific kinase inhibitors by targeted structural genomics.” *Drug Discov Today* 12 (2007), pp. 365–72.
- Feher, M. and Schmidt, J. M. (2003). “Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry.” *J Chem Inf Comput Sci* 43 (2003), pp. 218–27.
- Fernández-Suárez, X. M. and Galperin, M. Y. (2013). “The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection.” *Nucleic Acids Res* 41 (2013), pp. D1–7.
- Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2006). “Pfam: clans, web tools and services.” *Nucleic Acids Res* 34 (2006), pp. D247–51.
- Finn, R. D., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy,

- S. R., and Bateman, A. (2010). "The Pfam protein families database." *Nucleic Acids Res* 38 (2010), pp. D211–22.
- Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." *Syst Zool* 19 (1970), pp. 99–113.
- Flor, P. J. and Acher, F. C. (2012). "Orthosteric versus allosteric GPCR activation: the great challenge of group-III mGluRs." *Biochem Pharmacol* 84 (2012), pp. 414–24.
- Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). "Light-directed, spatially addressable parallel chemical synthesis." *Science* 251 (1991), pp. 767–73.
- Fong, T. M., Yu, H., and Strader, C. D. (1992). "Molecular basis for the species selectivity of the neurokinin-1 receptor antagonists CP-96,345 and RP67580." *J Biol Chem* 267 (1992), pp. 25668–71.
- Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). "SCOPE: Structural Classification of Proteins - extended, integrating SCOP and ASTRAL data and classification of new structures." *Nucleic Acids Res* 42 (2014), pp. D304–9.
- Frye, S. V. (1999). "Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era." *Chem Biol* 6 (1999), R3–7.
- Frye, S. V. (2010). "The art of the chemical probe". *Nat Chem Biol* 6 (2010), pp. 159–161.
- Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. A., and Gordon, E. M. (1994). "Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries". *J Med Chem* 37 (1994), pp. 1233–1251.
- Gao, Y., Davies, S. P., Augustin, M., Woodward, A., Patel, U. A., Kovelman, R., and Harvey, K. J. (2013). "A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery." *Biochem J* 451 (2013), pp. 313–28.
- Garbuzynskiy, S. O., Ivankov, D. N., Bogatyreva, N. S., and Finkelstein, A. V. (2013). "Golden triangle for folding rates of globular proteins." *Proc Natl Acad Sci U S A* 110 (2013), pp. 147–50.
- Garnett, M. J. et al. (2012). "Systematic identification of genomic markers of drug sensitivity in cancer cells." *Nature* 483 (2012), pp. 570–5.
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic Acids Res* 40 (2012), pp. D1100–7.
- Geurts, A. M. et al. (2009). "Knockout rats via embryo microinjection of zinc-finger nucleases." *Science* 325 (2009), p. 433.
- Gibbs, R. A. et al. (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." *Nature* 428 (2004), pp. 493–521.
- Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Müller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., Delft, F. von, Doyle, D. A., Oppermann, U. C. T., and Sundström, M. (2007). "The scientific impact of the Structural Genomics Consortium: a protein

- family and ligand-centered approach to medically-relevant human proteins.” *J Struct Funct Genomics* 8 (2007), pp. 107–19.
- Gill, T. J., Smith, G. J., Wissler, R. W., and Kunz, H. W. (1989). “The rat as an experimental animal.” *Science* 245 (1989), pp. 269–76.
- Gilson, M. K., Sharp, K. A., and Honig, B. H. (1988). “Calculating the electrostatic potential of molecules in solution: Method and error assessment”. *J Comput Chem* 9 (1988), pp. 327–335.
- Gloriam, D. E., Foord, S. M., Blaney, F. E., and Garland, S. L. (2009). “Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design”. *J Med Chem* 52 (2009), pp. 4429–4442.
- Golovin, A. and Henrick, K. (2008). “MSDmotif: exploring protein sites and motifs.” *BMC Bioinf* 9 (2008), p. 312.
- Goodsell, D. S., Mian, I., and Olson, A. J. (1989). “Rendering volumetric data in molecular systems”. *J Mol Graph* 7 (1989), pp. 41–47.
- Gorman, C. M., Moffat, L. F., and Howard, B. H. (1982). “Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells.” *Mol Cell Biol* 2 (1982), pp. 1044–51.
- Gould, S. J. and Subramani, S. (1988). “Firefly luciferase as a tool in molecular and cell biology.” *Anal Biochem* 175 (1988), pp. 5–13.
- Grant, J. A., Haigh, J. A., Pickup, B. T., Nicholls, A., and Sayle, R. A. (2006). “Lingos, finite state machines, and fast similarity searching.” *J Chem Inf Model* 46 (2006), pp. 1912–8.
- Greer, J., Erickson, J. W., Baldwin, J. J., and Varney, M. D. (1994). “Application of the three-dimensional structures of protein target molecules in structure-based drug design.” *J Med Chem* 37 (1994), pp. 1035–54.
- Grishin, N. V. (2001). “Fold change in evolution of protein structures.” *J Struct Biol* 134 (2001), pp. 167–85.
- Guengerich, F. P., Martin, M. V., Beaune, P. H., Kremers, P., Wolff, T., and Waxman, D. J. (1986). “Characterization of rat and human liver microsomal cytochrome P-450 forms involved in nifedipine oxidation, a prototype for genetic polymorphism in oxidative drug metabolism.” *J Biol Chem* 261 (1986), pp. 5051–60.
- Haes, A. J. and Van Duyne, R. P. (2002). “A nanoscale optical biosensor: sensitivity and selectivity of an approach based on the localized surface plasmon resonance spectroscopy of triangular silver nanoparticles.” *J Am Chem Soc* 124 (2002), pp. 10596–604.
- Hamaker, H. C. (1937). “The London-van der Waals attraction between spherical particles”. *Physica IV* 4 (1937), pp. 1058–1072.
- Hamra, F. K., Gatlin, J., Chapman, K. M., Grellhesl, D. M., Garcia, J. V., Hammer, R. E., and Garbers, D. L. (2002). “Production of transgenic rats by lentiviral transduction of male germ-line stem cells.” *Proc Natl Acad Sci U S A* 99 (2002), pp. 14931–6.

- Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., and Clarke, J. (2007). “The folding and evolution of multidomain proteins.” *Nat Rev Mol Cell Biol* 8 (2007), pp. 319–30.
- Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., and Shinozaki, K. (2009). “Increased expression and protein divergence in duplicate genes is associated with morphological diversification.” *PLoS Genet* 5 (2009), e1000781.
- Hann, M. M., Leach, A. R., and Harper, G. (2001). “Molecular complexity and its impact on the probability of finding leads for drug discovery.” *J Chem Inf Comput Sci* 41 (2001), pp. 856–64.
- Harris, C. J. and Stevens, A. P. (2006). “Chemogenomics: structuring the drug discovery process to gene families”. *Drug Discov Today* 11 (2006), pp. 880–888.
- Hasemann, C. A., Istvan, E. S., Uyeda, K., and Deisenhofer, J. (1996). “The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies.” *Structure* 4 (1996), pp. 1017–29.
- Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. L., Babbitt, P. C., Gerlt, J. A., Petsko, G. A., and Ringe, D. (1998). “Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase.” *Proc Natl Acad Sci U S A* 95 (1998), pp. 10396–401.
- Hedges, S. B., Dudley, J., and Kumar, S. (2006). “TimeTree: a public knowledge-base of divergence times among organisms.” *Bioinformatics* 22 (2006), pp. 2971–2.
- Hegyí, H. and Gerstein, M. (1999). “The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.” *J Mol Biol* 288 (1999), pp. 147–64.
- Hegyí, H. and Gerstein, M. (2001). “Annotation transfer for genomics: measuring functional divergence in multi-domain proteins.” *Genome Res* 11 (2001), pp. 1632–40.
- Hoehndorf, R., Hiebert, T., Hardy, N. W., Schofield, P. N., Gkoutos, G. V., and Dumontier, M. (2013). “Mouse model phenotypes provide information about human drug targets.” *Bioinformatics* (2013), pp. 1–7.
- Hol, W. G., Duijnen, P. T. van, and Berendsen, H. J. (1978). “The alpha-helix dipole and the properties of proteins.” *Nature* 273 (1978), pp. 443–6.
- Holm, L. and Sander, C. (1997). “An evolutionary treasure: unification of a broad set of amidohydrolases related to urease.” *Proteins* 28 (1997), pp. 72–82.
- Hopkins, A. L. and Groom, C. R. (2002). “The druggable genome.” *Nat Rev Drug Discovery* 1 (2002), pp. 727–30.
- Hopkins, A. L., Groom, C. R., and Alex, A. (2004). “Ligand efficiency: a useful metric for lead selection”. *Drug Discov Today* 9 (2004), pp. 430–431.
- Hopkins, A. L., Mason, J. S., and Overington, J. P. (2006). “Can we rationally design promiscuous drugs?": *Curr Opin Struct Biol* 16 (2006), pp. 127–36.

- Hopkins, M. M., Martin, P. A., Nightingale, P., Kraft, A., and Mahdi, S. (2007). “The myth of the biotech revolution: An assessment of technological, clinical and organisational change”. *Res Policy* 36 (2007), pp. 566–589.
- Horst, E. van der, Peironcelly, J. E., Ijzerman, A. P., Beukers, M. W., Lane, J. R., Vlijmen, H. W. T. van, Emmerich, M. T. M., Okuno, Y., and Bender, A. (2010). “A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization.” *BMC Bioinf* 11 (2010), p. 316.
- Hubbard, S. R. and Till, J. H. (2000). “Protein tyrosine kinase structure and function.” *Annu Rev Biochem* 69 (2000), pp. 373–98.
- Hughes, A. L. (1994). “The evolution of functionally novel proteins after gene duplication.” *Proc R Soc Lond B Biol Sci* 256 (1994), pp. 119–24.
- Hung, A. W., Silvestre, H. L., Wen, S., Ciulli, A., Blundell, T. L., and Abell, C. (2009). “Application of fragment growing and fragment linking to the discovery of inhibitors of Mycobacterium tuberculosis pantothenate synthetase.” *Angew Chem, Int Ed Engl* 48 (2009), pp. 8452–6.
- Hunter, S. et al. (2012). “InterPro in 2011: new developments in the family and domain prediction database.” *Nucleic Acids Res* 40 (2012), pp. D306–12.
- Hunter, T. (2009). “Tyrosine phosphorylation: thirty years and counting.” *Curr Opin Cell Biol* 21 (2009), pp. 140–6.
- Huss, J. W., Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T. J., Valafar, F., and Su, A. I. (2008). “A gene wiki for community annotation of gene function.” *PLoS Biol* 6 (2008), e175.
- Inglese, J., Shamu, C. E., and Guy, R. K. (2007). “Reporting data from high-throughput screening of small-molecule libraries.” *Nat Chem Biol* 3 (2007), pp. 438–41.
- Iskar, M., Zeller, G., Blattmann, P., Campillos, M., Kuhn, M., Kaminska, K. H., Runz, H., Gavin, A.-C., Pepperkok, R., Noort, V. van, and Bork, P. (2013). “Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding”. *Mol Sys Biol* 9 (2013).
- Ivanetich, K. M. and Santi, D. V. (1990). “Bifunctional thymidylate synthase-dihydrofolate reductase in protozoa.” *FASEB J* 4 (1990), pp. 1591–7.
- Jabri, E., Carr, M. B., Hausinger, R. P., and Karplus, P. A. (1995). “The crystal structure of urease from *Klebsiella aerogenes*”. *Science* 268 (1995), pp. 998–1004.
- Jaccard, P. (1901). “Étude comparative de la distribution florale dans une portion des Alpes et des Jura”. *B Soc Vaud Sci Nat* 37 (1901), pp. 547–579.
- Jain, V. K. and Magrath, I. T. (1991). “A chemiluminescent assay for quantitation of beta-galactosidase in the femtogram range: application to quantitation of beta-galactosidase in lacZ-transfected cells.” *Anal Biochem* 199 (1991), pp. 119–24.
- Janssen, P. A., Westeringh, C. van de, Jagenau, A. H., Demoen, P. J., Hermans, B. K., Daele, G. H. van, Schellekens, K. H., and Eycken, C. A. van der (1959). “Chemistry and

- pharmacology of CNS depressants related to 4-(4-hydroxy-phenylpiperidino)butyrophenone. I. Synthesis and screening data in mice.” *J Med Chem* 1 (1959), pp. 281–97.
- Jensen, A. A. and Bräuner-Osborne, H. (2007). “Allosteric modulation of the calcium-sensing receptor.” *Curr Neuropharmacol* 5 (2007), pp. 180–6.
- Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley New York, 1990.
- Kaiser, J. (2008). “Molecular biology. Industrial-style screening meets academic biology.” *Science* 321 (2008), pp. 764–6.
- Kalliokoski, T., Kramer, C., Vulpetti, A., and Gedeck, P. (2013). “Comparability of Mixed IC50 Data - A Statistical Analysis.” *PLoS One* 8 (2013), e61007.
- Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*. 4th ed. McGraw-Hill Medical, 2000.
- Kannan, K. K., Liljas, A., Waara, I., Bergstén, P. C., Lövgren, S., Strandberg, B., Bengtsson, U., Carlbom, U., Fridborg, K., Järup, L., and Petef, M. (1972). “Crystal structure of human erythrocyte carbonic anhydrase C. VI. The three-dimensional structure at high resolution in relation to other mammalian carbonic anhydrases.” *Cold Spring Harbor Symp Quant Biol* 36 (1972), pp. 221–31.
- Kapitzky, L., Beltrao, P., Berens, T. J., Gassner, N., Zhou, C., Wüster, A., Wu, J., Babu, M. M., Elledge, S. J., Toczyski, D., Lokey, R. S., and Krogan, N. J. (2010). “Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action.” *Mol Sys Biol* 6 (2010), p. 451.
- Karaman, M. W. et al. (2008). “A quantitative analysis of kinase inhibitor selectivity” . *Nat Biotechnol* 26 (2008), pp. 127–132.
- Karchin, R. and Hughey, R. (1998). “Weighting hidden Markov models for maximum discrimination.” *Bioinformatics* 14 (1998), pp. 772–82.
- Karp, P. D. (1996). “Database links are a foundation for interoperability.” *Trends Biotechnol* 14 (1996), pp. 273–9.
- Karplus, K., Sjölander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). “Predicting protein structure using hidden Markov models.” *Proteins Suppl* 1 (1997), pp. 134–9.
- Keen, G., Redgrave, G., Lawton, J., Cinkosky, M., Mishra, S., Fickett, J., and Burks, C. (1992). “Access to molecular biology databases” . *Math Comput Model* 16 (1992), pp. 93–101.
- Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007). “Relating protein pharmacology by ligand chemistry.” *Nat Biotechnol* 25 (2007), pp. 197–206.
- Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijjer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L. H., Edwards, D. D., Shoichet, B. K., and Roth, B. L. (2009). “Predicting new molecular targets for known drugs.” *Nature* 462 (2009), pp. 175–81.

- Kell, D. B., Dobson, P. D., and Oliver, S. G. (2011). "Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only." *Drug Discov Today* 16 (2011), pp. 704–14.
- Kell, D. B., Dobson, P. D., Bilsland, E., and Oliver, S. G. (2012). "The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so". *Drug Discov Today* 00 (2012).
- Kew, J. N. C. and Kemp, J. A. (2005). "Ionotropic and metabotropic glutamate receptor structure and pharmacology." *Psychopharmacol* 179 (2005), pp. 4–29.
- Kienhuis, A. S., Poll, M. C. G. van de, Wortelboer, H., Herwijnen, M. van, Gottschalk, R., Dejong, C. H. C., Boorsma, A., Paules, R. S., Kleinjans, J. C. S., Stierum, R. H., and Delft, J. H. M. van (2009). "Parallelogram approach using rat-human in vitro and rat in vivo toxicogenomics predicts acetaminophen-induced hepatotoxicity in humans." *Toxicol Sci* 107 (2009), pp. 544–52.
- Kimura, M. (1968). "Evolutionary rate at the molecular level." *Nature* 217 (1968), pp. 624–6.
- Kimura, M. and Ohta, T. (1971). "On the Rate of Molecular Evolution". *J Mol Evol* 1 (1971), pp. 1–17.
- Kliwer, S. A., Lehmann, J. M., and Willson, T. M. (1999). "Orphan nuclear receptors: shifting endocrinology into reverse." *Science* 284 (1999), pp. 757–60.
- Knight, Z. A., Lin, H., and Shokat, K. M. (2010). "Targeting the cancer kinome through polypharmacology." *Nat Rev Cancer* 10 (2010), pp. 130–7.
- Knighton, D. R., Kan, C. C., Howland, E., Janson, C. A., Hostomska, Z., Welsh, K. M., and Matthews, D. A. (1994). "Structure of and kinetic channelling in bifunctional dihydrofolate reductase-thymidylate synthase." *Nat Struct Mol Biol* 1 (1994), pp. 186–94.
- Kobayashi, K., Urashima, K., Shimada, N., and Chiba, K. (2003). "Selectivities of human cytochrome P450 inhibitors toward rat P450 isoforms: study with cDNA-expressed systems of the rat." *Drug Metab. Dispos.* 31 (2003), pp. 833–6.
- Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000). "The impact of comparative genomics on our understanding of evolution." *Cell* 101 (2000), pp. 573–6.
- Kost, T. a., Condreay, J. P., and Jarvis, D. L. (2005). "Baculovirus as versatile vectors for protein expression in insect and mammalian cells." *Nat Biotechnol* 23 (2005), pp. 567–75.
- Kramer, C., Kalliokoski, T., Gedeck, P., and Vulpetti, A. (2012). "The experimental uncertainty of heterogeneous public K(i) data." *J Med Chem* 55 (2012), pp. 5165–73.
- Kraskouskaya, D., Duodu, E., Arpin, C. C., and Gunning, P. T. (2013). "Progress towards the development of SH2 domain inhibitors." *Chem Soc Rev* 42 (2013), pp. 3337–70.
- Krissinel, E. and Henrick, K. (2004). "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." *Acta Crystallogr, Sect D: Biol Crystallogr* 60 (2004), pp. 2256–68.

- Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). “Hidden Markov models in computational biology. Applications to protein modeling.” *J Mol Biol* 235 (1994), pp. 1501–31.
- Krogh, A., Larsson, B., Heijne, G. von, and Sonnhammer, E. L. (2001). “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.” *J Mol Biol* 305 (2001), pp. 567–80.
- Kruger, F. A. and Overington, J. P. (2012). “Global analysis of small molecule binding to related protein targets.” *PLoS Comput Biol* 8 (2012), e1002333.
- Kruger, F. A., Rostom, R., and Overington, J. P. (2012). “Mapping small molecule binding data to structural domains.” *BMC Bioinf* 13 Suppl 1 (2012), S11.
- Kufareva, I., Ilatovskiy, A. V., and Abagyan, R. (2012). “Pocketome: an encyclopedia of small-molecule binding sites in 4D.” *Nucleic Acids Res* 40 (2012), pp. D535–40.
- Kuhn, M., Szklarczyk, D., Franceschini, A., Mering, C. von, Jensen, L. J., and Bork, P. (2012). “STITCH 3: zooming in on protein-chemical interactions.” *Nucleic Acids Res* 40 (2012), pp. D876–80.
- Kuhn, P., Wilson, K., Patch, M. G., and Stevens, R. C. (2002). “The genesis of high-throughput structure-based drug discovery using protein crystallography.” *Curr Opin Chem Biol* 6 (2002), pp. 704–10.
- Kumar, J. and Mayer, M. L. (2012). “Functional Insights from Glutamate Receptor Ion Channel Structures.” *Annu Rev Physiol* (2012), pp. 1–25.
- Kunishima, N., Shimada, Y., Tsuji, Y., Sato, T., Yamamoto, M., Kumasaka, T., Nakanishi, S., Jingami, H., and Morikawa, K. (2000). “Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor.” *Nature* 407 (2000), pp. 971–7.
- Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. (2012). “Rapid turnover of long noncoding RNAs and the evolution of gene expression.” *PLoS Genet* 8 (2012), e1002841.
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease”. *Science* 313 (2006), pp. 1929–1935.
- Latimer, W. M. and Rodebush, W. H. (1920). “Polarity and ionization from the standpoint of the lewis theory of valence”. *J Am Chem Soc* 42 (1920), pp. 1419–1433.
- Lazo, J. S. (2006). “Roadmap or roadkill: a pharmacologist’s analysis of the NIH Molecular Libraries Initiative.” *Mol Interventions* 6 (2006), pp. 240–3.
- Le Poul, E., Hisada, S., Mizuguchi, Y., Dupriez, V. J., Burgeon, E., and Detheux, M. (2002). “Adaptation of aequorin functional assay to high throughput screening.” *J Biomol Screen* 7 (2002), pp. 57–65.
- Levitt, M. and Chothia, C. (1976). “Structural patterns in globular proteins”. *Nature* 261 (1976), pp. 552–558.

- Lewis, T. E. et al. (2012). “Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains.” *Nucleic Acids Res* (2012), pp. 1–9.
- Li, L., Stoeckert, C. J., and Roos, D. S. (2003). “OrthoMCL: identification of ortholog groups for eukaryotic genomes.” *Genome Res* 13 (2003), pp. 2178–89.
- Li, Q., Cheng, T., Wang, Y., and Bryant, S. H. (2012). “Characterizing protein domain associations by Small-molecule ligand binding.” *J Proteome Sci Comput Biol* 1 (2012), p. 6.
- Li, W., Wei, W., Zhu, S., Zhu, J., Shi, Y., Lin, T., Hao, E., Hayek, A., Deng, H., and Ding, S. (2009). “Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors.” *Cell Stem Cell* 4 (2009), pp. 16–9.
- Liao, B.-Y. and Zhang, J. (2006). “Evolutionary conservation of expression profiles between human and mouse orthologous genes.” *Mol Biol Evol* 23 (2006), pp. 530–40.
- Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). “An evolutionary trace method defines binding surfaces common to protein families.” *J Mol Biol* 257 (1996), pp. 342–58.
- Ligneau, X., Morisset, S., Tardivel-Lacombe, J., Gbahou, F., Ganellin, C. R., Stark, H., Schunack, W., Schwartz, J. C., and Arrang, J. M. (2000). “Distinct pharmacology of rat and human histamine H(3) receptors: role of two amino acids in the third transmembrane domain.” *Br J Pharmacol* 131 (2000), pp. 1247–50.
- Lin, H., Sassano, M. F., Roth, B. L., and Shoichet, B. K. (2013). “A pharmacological organization of G protein-coupled receptors.” *Nat Methods* 10 (2013).
- Lindahl, E. and Elofsson, A. (2000). “Identification of related proteins on family, superfamily and fold level.” *J Mol Biol* 295 (2000), pp. 613–25.
- Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.” *Adv Drug Delivery Rev* 46 (2001), pp. 3–26.
- Liu, B. A., Jablonowski, K., Raina, M., Arcé, M., Pawson, T., and Nash, P. D. (2006). “The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling.” *Mol Cell* 22 (2006), pp. 851–68.
- Liu, B. A., Shah, E., Jablonowski, K., Stergachis, A., Engelmann, B., and Nash, P. D. (2011a). “The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes.” *Sci Signal* 4 (2011), ra83.
- Liu, Q., Wang, J., Kang, S. A., Thoreen, C. C., Hur, W., Ahmed, T., Sabatini, D. M., and Gray, N. S. (2011b). “Discovery of 9-(6-aminopyridin-3-yl)-1-(3-(trifluoromethyl)phenyl)benzo[h][1,6]naphthyridin-2(1H)-one (Torin2) as a potent, selective, and orally available mammalian target of rapamycin (mTOR) inhibitor for treatment of cancer.” *J Med Chem* 54 (2011), pp. 1473–80.

- Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." *Nucleic Acids Res* 35 (2007), pp. D198–201.
- Livnah, O., Bayer, E. A., Wilchek, M., and Sussman, J. L. (1993). "Three-dimensional structures of avidin and the avidin-biotin complex." *Proc Natl Acad Sci U S A* 90 (1993), pp. 5076–80.
- Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2002). "SCOP database in 2002: refinements accommodate structural genomics." *Nucleic Acids Res* 30 (2002), pp. 264–7.
- Loging, W., Harland, L., and Williams-Jones, B. (2007). "High-throughput electronic biology: mining information for drug discovery." *Nat Rev Drug Discovery* 6 (2007), pp. 220–30.
- Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., and Tarczy-Hornoch, P. (2007). "Data integration and genomic medicine." *J Biomed Inform* 40 (2007), pp. 5–16.
- Lovenberg, T. W., Pyati, J., Chang, H., Wilson, S. J., and Erlander, M. G. (2000). "Cloning of rat histamine H(3) receptor reveals distinct species pharmacological profiles." *J Pharmacol Exp Ther* 293 (2000), pp. 771–8.
- Lowry, O. H., Roseborough, N. J., Farr, A. L., and Randall, R. J. (1951). "Protein measurement with the Folin phenol reagent." *J Biol Chem* 193 (1951), pp. 265–75.
- Lynch, M. (2000). "The Evolutionary Fate and Consequences of Duplicate Genes". *Science* 290 (2000), pp. 1151–1155.
- Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces." *Proc Natl Acad Sci U S A* 100 (2003), pp. 5772–7.
- Maas, W. K. (1952). "Pantothenate studies. III. Description of the extracted pantothenate-synthesizing enzyme of *Escherichia coli*." *J Biol Chem* 198 (1952), pp. 23–32.
- Machida, K. and Mayer, B. J. (2005). "The SH2 domain: versatile signaling module and pharmaceutical target." *Biochim Biophys Acta* 1747 (2005), pp. 1–25.
- Maehle, A.-H., Prüll, C.-R., and Halliwell, R. F. (2002). "The emergence of the drug receptor theory." *Nat Rev Drug Discovery* 1 (2002), pp. 637–41.
- Maemoto, T., Finlayson, K., Olverman, H. J., Akahane, A., Horton, R. W., and Butcher, S. P. (1997). "Species differences in brain adenosine A1 receptor pharmacology revealed by use of xanthine and pyrazolopyridine based antagonists." *Br J Pharmacol* 122 (1997), pp. 1202–8.
- Magrane, M. and Consortium, T. U. (2011). "UniProt Knowledgebase: a hub of integrated protein data." *Database* 2011 (2011), bar009.
- Maier, T., Leibundgut, M., and Ban, N. (2008). "The crystal structure of a mammalian fatty acid synthase." *Science* 321 (2008), pp. 1315–22.
- Malitschek, B., Schweizer, C., Keir, M., Heid, J., Froestl, W., Mosbacher, J., Kuhn, R., Henley, J., Joly, C., Pin, J. P., Kaupmann, K., and Bettler, B. (1999). "The N-terminal

- domain of gamma-aminobutyric Acid(B) receptors is sufficient to specify agonist and antagonist binding.” *Mol Pharmacol* 56 (1999), pp. 448–54.
- Malo, N., Hanley, J. A., Carlile, G., Liu, J., Pelletier, J., Thomas, D., and Nadon, R. (2010). “Experimental design and statistical methods for improved hit detection in high-throughput screening.” *J Biomol Screen* 15 (2010), pp. 990–1000.
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). “The protein kinase complement of the human genome.” *Science* 298 (2002), pp. 1912–34.
- Marcotte, E. M. (1999). “Detecting Protein Function and Protein-Protein Interactions from Genome Sequences”. *Science* 285 (1999), pp. 751–753.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). “Detecting protein function and protein-protein interactions from genome sequences.” *Science* 285 (1999), pp. 751–3.
- Marsden, B. D. and Knapp, S. (2008). “Doing more than just the structure-structural genomics in kinase drug discovery.” *Curr Opin Chem Biol* 12 (2008), pp. 40–5.
- Marsden, R. L., Lewis, T. A., and Orengo, C. A. (2007). “Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint.” *BMC Bioinf* 8 (2007), p. 86.
- Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C., and Thornton, J. M. (1998). “Protein folds and functions.” *Structure* 6 (1998), pp. 875–84.
- Martin, B., Ji, S., Maudsley, S., and Mattson, M. P. (2010). ““Control” laboratory rodents are metabolically morbid: why it matters”. *Proc Natl Acad Sci U S A* 107 (2010), pp. 6127–6133.
- Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). “Do structurally similar molecules have similar biological activity?”: *J Med Chem* 45 (2002), pp. 4350–8.
- McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). “Systematic discovery of nonobvious human disease models through orthologous phenotypes.” *Proc Natl Acad Sci U S A* 107 (2010), pp. 6544–9.
- McGovern, S. L., Caselli, E., Grigorieff, N., and Shoichet, B. K. (2002). “A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening.” *J Med Chem* 45 (2002), pp. 1712–22.
- McGovern, S. L., Helfand, B. T., Feng, B., and Shoichet, B. K. (2003). “A specific mechanism of nonspecific inhibition.” *J Med Chem* 46 (2003), pp. 4265–72.
- Merrifield, R. B. (1963). “Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide”. *J Am Chem Soc* 85 (1963), pp. 2149–2154.
- Metallo, S. J. (2010). “Intrinsically disordered proteins are potential drug targets.” *Curr Opin Chem Biol* 14 (2010), pp. 481–8.
- Metz, J. T., Johnson, E. F., Soni, N. B., Merta, P. J., Kifle, L., and Hajduk, P. J. (2011). “Navigating the kinome”. *Nat Chem Biol* 7 (2011), pp. 1–3.

- Michaelis, L. and Menten, M. (1913). “Die Kinetik der Invertinwirkung”. *Biochem Zeitschr* 49 (1913), pp. 333–369.
- Michaelis, L., Menten, M. L., Johnson, K. A., and Goody, R. S. (2011). “The original Michaelis constant: translation of the 1913 Michaelis-Menten paper.” *Biochemistry* 50 (2011), pp. 8264–9.
- Millard, B. L., Niepel, M., Menden, M. P., Muhlich, J. L., and Sorger, P. K. (2011). “Adaptive informatics for multifactorial and high-content biological data.” *Nat Methods* 8 (2011), pp. 487–93.
- Mistry, J., Coghill, P., Eberhardt, R. Y., Deiana, A., Giansanti, A., Finn, R. D., Bateman, A., and Punta, M. (2013). “The challenge of increasing Pfam coverage of the human proteome.” *Database* 2013 (2013), bat023.
- Mitsuya, M., Kamata, K., Bamba, M., Watanabe, H., Sasaki, Y., Sasaki, K., Ohyama, S., Hosaka, H., Nagata, Y., Eiki, J.-I., and Nishimura, T. (2009). “Discovery of novel 3,6-disubstituted 2-pyridinecarboxamide derivatives as GK activators.” *Bioorg Med Chem Lett* 19 (2009), pp. 2718–21.
- Mons, B. et al. (2008). “Calling on a million minds for community annotation in WikiProteins.” *Genome Biol* 9 (2008), R89.
- Mony, L., Krzaczkowski, L., Leonetti, M., Le Goff, A., Alarcon, K., Neyton, J., Bertrand, H.-O., Acher, F., and Paoletti, P. (2009). “Structural basis of NR2B-selective antagonist recognition by N-methyl-D-aspartate receptors.” *Mol Pharmacol* 75 (2009), pp. 60–74.
- Moore, T. S. and Winmill, T. F. (1912). “The state of amines in aqueous solution”. *J Chem Soc* 101 (1912), p. 1635.
- Morphy, R. and Rankovic, Z. (2005). “Designed multiple ligands. An emerging drug discovery paradigm.” *J Med Chem* 48 (2005), pp. 6523–43.
- Morphy, R. and Rankovic, Z. (2007). “Fragments, network biology and designing multiple ligands.” *Drug Discov Today* 12 (2007), pp. 156–60.
- Moya-García, A. A. and Ranea, J. A. G. (2013). “Insights into polypharmacology from drug-domain associations.” *Bioinformatics* 29 (2013), pp. 1934–7.
- Mun, H.-C., Franks, A. H., Culverston, E. L., Krapcho, K., Nemeth, E. F., and Conigrave, A. D. (2004). “The Venus Fly Trap domain of the extracellular Ca²⁺-sensing receptor is required for L-amino acid sensing.” *J Biol Chem* 279 (2004), pp. 51739–44.
- Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). “SCOP: a structural classification of proteins database for the investigation of sequences and structures.” *J Mol Biol* 247 (1995), pp. 536–540.
- Narahashi, T., Moore, J. W., and Scott, W. R. (1964). “Tetrodotoxin Blockage of Sodium Conductance Increase in Lobster Giant Axons.” *J Gen Physiol* 47 (1964), pp. 965–74.
- Needleman, S. B. and Wunsch, C. D. (1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” *J Mol Biol* 48 (1970), pp. 443–53.

- Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. (2011). "Testing the ortholog conjecture with comparative functional genomic data from mammals." *PLoS Comput Biol* 7 (2011). Ed. by A. Rzhetsky, e1002073.
- Nersesian, D. L., Black, L. A., Miller, T. R., Vortherms, T. A., Esbenshade, T. A., Hancock, A. A., and Cowart, M. D. (2008). "In vitro SAR of pyrrolidine-containing histamine H3 receptor antagonists: trends across multiple chemical series." *Bioorg Med Chem Lett* 18 (2008), pp. 355–9.
- Neumann, E. and Thomas, J. (2002). "Knowledge assembly for the life sciences." *Drug Discov Today* 7 (2002), S160–2.
- Niedermeier, S., Singethan, K., Rohrer, S. G., Matz, M., Kossner, M., Diederich, S., Maisner, A., Schmitz, J., Hiltensperger, G., Baumann, K., Holzgrabe, U., and Schneider-Schaulies, J. (2009). "A small-molecule inhibitor of Nipah virus envelope protein-mediated membrane fusion." *J Med Chem* 52 (2009), pp. 4257–65.
- Nightingale, P. (2000). "Economies of scale in experimentation: knowledge and technology in pharmaceutical R&D". *Ind Corp Change* 9 (2000), pp. 315–359.
- Nishino, R., Ikeda, K., Hayakawa, T., Takahashi, T., Suzuki, T., and Sato, M. (2011). "Syntheses of 2-deoxy-2,3-didehydro-N-acetylneuraminic acid analogues modified by N-sulfonylamidino groups at the C-4 position and biological evaluation as inhibitors of human parainfluenza virus type 1." *Bioorg Med Chem* 19 (2011), pp. 2418–27.
- Nobeli, I., Favia, A. D., and Thornton, J. M. (2009). "Protein promiscuity and its implications for biotechnology." *Nat Biotechnol* 27 (2009), pp. 157–67.
- OEChem, (2006). *OEChem V*. Inc.: Sante Fe, 2006.
- Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). "Tissue-specific transcriptional regulation has diverged significantly between human and mouse." *Nat Genet* 39 (2007), pp. 730–2.
- Oprea, T. I. and Tropsha, A. (2006). "Target, chemical and bioactivity databases integration is key". *Drug Discov Today* 3 (2006), pp. 357–365.
- Oprea, T. I., Bologa, C. G., Boyer, S., Curpan, R. F., Glen, R. C., Hopkins, A. L., Lipinski, C. A., Marshall, G. R., Martin, Y. C., Ostopovici-Halip, L., Rishton, G., Ursu, O., Vaz, R. J., Waller, C., Waldmann, H., and Sklar, L. A. (2009). "A crowdsourcing evaluation of the NIH chemical probes." *Nat Chem Biol* 5 (2009), pp. 441–7.
- Orchard, S. et al. (2011). "Minimum information about a bioactive entity (MIABE)." *Nat Rev Drug Discovery* 10 (2011), pp. 661–9.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). "CATH—a hierarchic classification of protein domain structures." *Structure* 5 (1997), pp. 1093–108.
- Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., and Thornton, J. M. (1999). "The CATH Database provides insights into protein structure/function relationships." *Nucleic Acids Res* 27 (1999), pp. 275–9.

- Orning, L., Krivi, G., and Fitzpatrick, F. A. (1991). "Leukotriene A4 hydrolase. Inhibition by bestatin and intrinsic aminopeptidase activity establish its functional resemblance to metallohydrolase enzymes." *J Biol Chem* 266 (1991), pp. 1375–8.
- 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 (1999), pp. 2896–901.
- Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). "How many drug targets are there?": *Nat Rev Drug Discovery* 5 (2006), pp. 993–6.
- Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. a., Smirnova, T., Nosrat, B., Markowitz, V. M., and Kyrpides, N. C. (2012). "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata." *Nucleic Acids Res* 40 (2012), pp. D571–9.
- Page, B. D. G., Khoury, H., Laister, R. C., Fletcher, S., Vellozo, M., Manzoli, A., Yue, P., Turkson, J., Minden, M. D., and Gunning, P. T. (2012). "Small molecule STAT5-SH2 domain inhibitors exhibit potent antileukemia activity." *J Med Chem* 55 (2012), pp. 1047–55.
- Pantaloni, C., Brabet, P., Bilanges, B., Dumuis, A., Houssami, S., Spengler, D., Bockaert, J., and Journot, L. (1996). "Alternative splicing in the N-terminal extracellular domain of the pituitary adenylate cyclase-activating polypeptide (PACAP) receptor modulates receptor selectivity and relative potencies of PACAP-27 and PACAP-38 in phospholipase C activation." *J Biol Chem* 271 (1996), pp. 22146–51.
- Paolini, G. V., Shapland, R. H. B., Hoorn, W. P. van, Mason, J. S., and Hopkins, A. L. (2006). "Global mapping of pharmacological space." *Nat Biotechnol* 24 (2006), pp. 805–15.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods." *J Mol Biol* 284 (1998), pp. 1201–10.
- Parsons, A. B., Brost, R. L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G. W., Kane, P. M., Hughes, T. R., and Boone, C. (2004). "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." *Nat Biotechnol* 22 (2004), pp. 62–9.
- Patthy, L. (1996). "Exon shuffling and other ways of module exchange." *Matrix Biol* 15 (1996), 301–10; discussion 311–2.
- Pauling, L. and Delbrück, M. (1940). "The Nature of the Intermolecular Forces Operative in Biological Processes". *Science* 92 (1940), pp. 77–9.
- Pauling, L. (1974). "Molecular basis of biological specificity". *Nature* 248 (1974), pp. 769–771.
- Pawson, T. (2004). "Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems." *Cell* 116 (2004), pp. 191–203.

- Pelletier, J. and Caventou, J. B. (1820). *Suite: Des recherches chimiques sur les quinquinas*. Vol. 15. Landmarks II, scientific journals. Masson., 1820, pp. 337–65.
- Pereira, D. A. and Williams, J. A. (2007). “Origin and evolution of high throughput screening.” *Br J Pharmacol* 152 (2007), pp. 53–61.
- Perutz, M. F. (1978). “Electrostatic effects in proteins.” *Science* 201 (1978), pp. 1187–91.
- Perutz, M. F. (1983). “Species adaptation in a protein molecule.” *Mol Biol Evol* 1 (1983), pp. 1–28.
- Pichler, W. J. (2003). “Delayed drug hypersensitivity reactions.” *Ann Intern Med* 139 (2003), pp. 683–93.
- Pin, J.-P. and Prézeau, L. (2007). “Allosteric modulators of GABA(B) receptors: mechanism of action and therapeutic perspective.” *Curr Neuropharmacol* 5 (2007), pp. 195–201.
- Pin, J.-P., Galvez, T., and Prézeau, L. (2003). “Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors.” *Pharmacol Ther* 98 (2003), pp. 325–54.
- Ponticello, G. S., Sugrue, M. F., Plazonnet, B., and Durand-Cavagna, G. (1998). “Dorzolamide, a 40-year wait. From an oral to a topical carbonic anhydrase inhibitor for the treatment of glaucoma.” *Pharm Biotechnol* 11 (1998), pp. 555–74.
- Posy, S. L., Hermsmeier, M. A., Vaccaro, W., Ott, K. H., Todderud, G., Lippy, J. S., Trainor, G. L., Loughney, D. A., and Johnson, S. R. (2011). “Trends in kinase selectivity: insights for target class-focused library screening”. *J Med Chem* 54 (2011), pp. 54–66.
- Potterton, E., McNicholas, S., Krissinel, E., Cowtan, K., and Noble, M. (2002). “The CCP4 molecular-graphics project.” *Acta Crystallogr, Sect D: Biol Crystallogr* 58 (2002), pp. 1955–7.
- Powell, M. A. (1993). *Drugs and pharmaceuticals in ancient Mesopotamia*. Ed. by I Jacob and W Jacob. Leiden: E. J. Brill, 1993, pp. 54.
- Prichard, B. N. and Gillam, P. M. (1964). “Use of Propranolol (Inderal) in Treatment of Hypertension.” *BMJ* 2 (1964), pp. 725–7.
- Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). “The Pfam protein families database.” *Nucleic Acids Res* 40 (2012), pp. D290–301.
- Purdom, E. and Holmes, S. P. (2005). “Error distribution for gene expression data.” *Stat Appl Genet Mol Biol* 4 (2005), Article16.
- Rekapalli, B., Wuichet, K., Peterson, G. D., and Zhulin, I. B. (2012). “Dynamics of domain coverage of the protein sequence universe.” *BMC Genomics* 13 (2012), p. 634.
- Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). “Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.” *J Mol Biol* 314 (2001), pp. 1041–52.

- Rice, P., Longden, I., and Bleasby, A. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." *Trends Genet* 16 (2000), pp. 276–7.
- Rider, M. H., Bertrand, L., Vertommen, D., Michels, P. A., Rousseau, G. G., and Hue, L. (2004). "6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase: head-to-head with a bifunctional enzyme that controls glycolysis." *Biochem J* 381 (2004), pp. 561–79.
- Rishton, G. M. (1997). "Reactive compounds and in vitro false positives in HTS". *Drug Discov Today* 2 (1997), pp. 382–384.
- Root, D. E., Kelley, B. P., and Stockwell, B. R. (2002). "Global analysis of large-scale chemical and biological experiments." *Curr Opin Drug Discovery Dev* 5 (2002), pp. 355–60.
- Rost, B. (2002). "Enzyme function less conserved than anticipated." *J Mol Biol* 318 (2002), pp. 595–608.
- Roth, B. L., Lopez, E., Patel, S., and Kroeze, W. K. (2000). "The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?": *Neuroscientist* 6 (2000), pp. 252–262.
- Roth, B. L., Sheffler, D. J., and Kroeze, W. K. (2004). "Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia." *Nat Rev Drug Discovery* 3 (2004), pp. 353–9.
- Russ, A. P. and Lampel, S. (2005). "The druggable genome: an update." *Drug Discov Today* 10 (2005), pp. 1607–10.
- Rustici, G. et al. (2013). "ArrayExpress update—trends in database growth and links to data analysis tools." *Nucleic Acids Res* 41 (2013), pp. D987–90.
- Ryu, B.-Y., Orwig, K. E., Oatley, J. M., Lin, C.-C., Chang, L.-J., Avarbock, M. R., and Brinster, R. L. (2007). "Efficient generation of transgenic rats through the male germline using lentiviral transduction and transplantation of spermatogonial stem cells." *J Androl* 28 (2007), pp. 353–60.
- Sali, A. (1998). "100,000 Protein Structures for the Biologist." *Nat Struct Mol Biol* 5 (1998), pp. 1029–32.
- Sali, A. and Blundell, T. L. (1993). "Comparative protein modelling by satisfaction of spatial restraints." *J Mol Biol* 234 (1993), pp. 779–815.
- Sambandamurthy, V. K., Wang, X., Chen, B., Russell, R. G., Derrick, S., Collins, F. M., Morris, S. L., and Jacobs, W. R. (2002). "A pantothenate auxotroph of *Mycobacterium tuberculosis* is highly attenuated and protects mice against tuberculosis." *Nat Med* 8 (2002), pp. 1171–4.
- Sanderson, M. J. (1997). "Nonparametric approach to estimating divergence times in the absence of rate constancy". *Mol Biol Evol* (1997), pp. 1218–1231.
- Satoh, A., Konishi, S., Tamura, H., Stickland, H. G., Whitney, H. M., Smith, A. G., Matsumura, H., and Inoue, T. (2010). "Substrate-induced closing of the active site revealed by the crystal structure of pantothenate synthetase from *Staphylococcus aureus*." *Biochemistry* 49 (2010), pp. 6400–10.

- Sawyer, T. K. (1998). "Src homology-2 domains: structure, mechanisms, and drug discovery." *Biopolymers* 47 (1998), pp. 243–61.
- Scearce-Levie, K., Coward, P., Redfern, C. H., and Conklin, B. R. (2002). "Tools for dissecting signaling pathways in vivo: receptors activated solely by synthetic ligands." *Methods Enzymol* 343 (2002), pp. 232–48.
- Schmitz, R. (1985). "Friedrich Wilhelm Sertürner and the discovery of morphine." *Pharm Hist* 27 (1985), pp. 61–74.
- Schneider, M. V. and Jimenez, R. C. (2012). "Teaching the Fundamentals of Biological Data Integration Using Classroom Games". *PLoS Comput Biol* 8 (2012). Ed. by F. Lewitter, e1002789.
- Schormann, N., Velu, S. E., Murugesan, S., Senkovich, O., Walker, K., Chenna, B. C., Shinkre, B., Desai, A., and Chattopadhyay, D. (2010). "Synthesis and characterization of potent inhibitors of *Trypanosoma cruzi* dihydrofolate reductase." *Bioorg Med Chem* 18 (2010), pp. 4056–66.
- Schreiber, S. L. (1991). "Chemistry and biology of the immunophilins and their immunosuppressive ligands." *Science* 251 (1991), pp. 283–7.
- Schreiber, S. L. (1998). "Chemical genetics resulting from a passion for synthetic organic chemistry." *Bioorg Med Chem* 6 (1998), pp. 1127–52.
- Schreiber, S. L. (2003). "The small-molecule approach to biology". *Chem Eng News* 81 (2003), pp. 51–60.
- Schürer, S. C., Vempati, U., Smith, R., Southern, M., and Lemmon, V. (2011). "BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets." *J Biomol Screen* 16 (2011), pp. 415–26.
- Searle, M. S., Williams, D. H., and Gerhard, U. (1992). "Partitioning of free energy contributions in the estimation of binding constants: residual motions and consequences for amide-amide hydrogen bond strengths". *J Am Chem Soc* 114 (1992), pp. 10697–10704.
- Searls, D. B. (2003a). "Data integration—connecting the dots." *Nat Biotechnol* 21 (2003), pp. 844–5.
- Searls, D. B. (2003b). "Pharmacophylogenomics: genes, evolution and drug targets." *Nat Rev Drug Discovery* 2 (2003), pp. 613–23.
- Searls, D. B. (2005). "Data integration: challenges for drug discovery." *Nat Rev Drug Discovery* 4 (2005), pp. 45–58.
- Seeger, D. R., Cosulich, D. B., Smith, J. M., and Hultquist, M. E. (1949). "Analogues of Pteroylglutamic Acid. III. 4-Amino Derivatives". *J Am Chem Soc* 71 (1949), pp. 1753–58.
- Shah, I. and Hunter, L. (1997). "Predicting enzyme function from sequence: a systematic appraisal." *International Conference on Intelligent Systems for Molecular Biology*. Vol. 5. 1997, pp. 276–83.

- Shamovsky, I., Connolly, S., David, L., Ivanova, S., Nordén, B., Springthorpe, B., and Urbahns, K. (2008). “Overcoming Undesirable hERG Potency of Chemokine Receptor Antagonists Using Baseline Lipophilicity Relationships”. *J Med Chem* 51 (2008), pp. 1162–1178.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” *Genome Res* 13 (2003), pp. 2498–504.
- Shimamura, T., Shiroishi, M., Weyand, S., Tsujimoto, H., Winter, G., Katritch, V., Abagyan, R., Cherezov, V., Liu, W., Han, G. W., Kobayashi, T., Stevens, R. C., and Iwata, S. (2011). “Structure of the human histamine H1 receptor complex with doxepin.” *Nature* 475 (2011), pp. 65–70.
- Shlyakhter, A. I. (1994). “An Improved Framework for Uncertainty Analysis: Accounting for Unsuspected Errors”. *Risk Anal* 14 (1994), pp. 441–447.
- Shoemaker, B. A., Zhang, D., Thangudu, R. R., Tyagi, M., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T., and Panchenko, A. R. (2010). “Inferred Biomolecular Interaction Server—a web server to analyze and predict protein interacting partners and binding sites.” *Nucleic Acids Res* 38 (2010), pp. D518–24.
- Shoemaker, R. H. (2006). “The NCI60 human tumour cell line anticancer drug screen.” *Nat Rev Cancer* 6 (2006), pp. 813–23.
- Shoichet, B. K. (2006). “Screening in a spirit haunted world.” *Drug Discov Today* 11 (2006), pp. 607–15.
- Shokat, K. and Velleca, M. (2002). “Novel chemical genetic approaches to the discovery of signal transduction inhibitors”. *Drug Discov Today* 7 (2002), pp. 872–879.
- Siddiquee, K., Zhang, S., Guida, W. C., Blaskovich, M. A., Greedy, B., Lawrence, H. R., Yip, M. L. R., Jove, R., McLaughlin, M. M., Lawrence, N. J., Sebt, S. M., and Turkson, J. (2007). “Selective chemical probe inhibitor of Stat3, identified through structure-based virtual screening, induces antitumor activity.” *Proc Natl Acad Sci U S A* 104 (2007), pp. 7391–6.
- Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M., and Orengo, C. A. (2013). “New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.” *Nucleic Acids Res* 41 (2013), pp. D490–8.
- Silverman, L., Campbell, R., and Broach, J. R. (1998). “New assay technologies for high-throughput screening.” *Curr Opin Chem Biol* 2 (1998), pp. 397–403.
- Siu, F. Y., He, M., Graaf, C. de, Han, G. W., Yang, D., Zhang, Z., Zhou, C., Xu, Q., Wacker, D., Joseph, J. S., Liu, W., Lau, J., Cherezov, V., Katritch, V., Wang, M.-W., and Stevens, R. C. (2013). “Structure of the human glucagon class B G-protein-coupled receptor.” *Nature* 499 (2013), pp. 444–9.

- Slater, T., Bouton, C., and Huang, E. S. (2008). “Beyond data integration.” *Drug Discov Today* 13 (2008), pp. 584–9.
- Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). “BioMart—biological queries made easy.” *BMC Genomics* 10 (2009), p. 22.
- Smith, K. S., Jakubzick, C., Whittam, T. S., and Ferry, J. G. (1999). “Carbonic anhydrase is an ancient enzyme widespread in prokaryotes.” *Proc Natl Acad Sci U S A* 96 (1999), pp. 15184–9.
- Sneider, W. (1998). “The discovery of heroin.” *Lancet* 352 (1998), pp. 1697–9.
- Sneider, W. (2005). *Drug discovery: a history*. Chichester, UK: John Wiley, 2005.
- Snyder, K. A., Feldman, H. J., Dumontier, M., Salama, J. J., and Hogue, C. W. V. (2006). “Domain-based small molecule binding site annotation.” *BMC Bioinf* 7 (2006), p. 152.
- Sobolevsky, A. I., Rosconi, M. P., and Gouaux, E. (2009). “X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor”. *Nature* 462 (2009), pp. 745–756.
- Song, H., Wang, R., Wang, S., and Lin, J. (2005). “A low-molecular-weight compound discovered through virtual database screening inhibits Stat3 function in breast cancer cells.” *Proc Natl Acad Sci U S A* 102 (2005), pp. 4700–5.
- Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). “Pfam: a comprehensive database of protein domain families based on seed alignments.” *Proteins* 28 (1997), pp. 405–20.
- Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). “Pfam: multiple sequence alignments and HMM-profiles of protein domains.” *Nucleic Acids Res* 26 (1998), pp. 320–2.
- Stein, L. D. (2003). “Integrating biological databases.” *Nat Rev Genet* 4 (2003), pp. 337–45.
- Stevens, R., Goble, C. A., and Bechhofer, S. (2000). “Ontology-based knowledge representation for bioinformatics.” *Brief Bioinform* 1 (2000), pp. 398–414.
- Stockwell, B. R. (2000). “Chemical genetics: ligand-based discovery of gene function”. *Nat Rev Genet* 1 (2000), pp. 116–125.
- Stromgaard, K., Brier, T. J., Andersen, K., Mellor, I. R., Saghyan, A., Tikhonov, D., Usherwood, P. N., Krogsgaard-Larsen, P., and Jaroszewski, J. W. (2000). “Solid-phase synthesis and biological evaluation of a combinatorial library of philanthotoxin analogues.” *J Med Chem* 43 (2000), pp. 4526–33.
- Studer, R. A. and Robinson-Rechavi, M. (2009). “How confident can we be that orthologs are similar, but paralogs differ?": *Trends Genet* 25 (2009), pp. 210–6.
- Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., and Hogenesch, J. B. (2002). “Large-scale analysis of the human and mouse transcriptomes.” *Proc Natl Acad Sci U S A* 99 (2002), pp. 4465–70.

- Surgand, J.-S., Rodrigo, J., Kellenberger, E., and Rognan, D. (2006). "A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors." *Proteins* 62 (2006), pp. 509–38.
- Swinney, D. C. and Anthony, J. (2011). "How were new medicines discovered?": *Nat Rev Drug Discovery* 10 (2011), pp. 507–519.
- Takimoto, C. (1996). "New Antifolates: Pharmacology and Clinical Applications." *Oncologist* 1 (1996), pp. 68–81.
- Takimoto, T., Taylor, G. L., Crennell, S. J., Scroggs, R. A., and Portner, A. (2000). "Crystallization of Newcastle disease virus hemagglutinin-neuraminidase glycoprotein." *Virology* 270 (2000), pp. 208–14.
- Tamaoki, T. (1991). "Use and specificity of staurosporine, UCN-01, and calphostin C as protein kinase inhibitors." *Methods Enzymol* 201 (1991), pp. 340–7.
- Tamaoki, T., Nomoto, H., Takahashi, I., Kato, Y., Morimoto, M., and Tomita, F. (1986). "Staurosporine, a potent inhibitor of phospholipid/Ca⁺⁺dependent protein kinase." *Biochem Biophys Res Commun* 135 (1986), pp. 397–402.
- Tanramluk, D., Schreyer, A., Pitt, W. R., and Blundell, T. L. (2009). "On the origins of enzyme inhibitor selectivity and promiscuity: a case study of protein kinase binding to staurosporine." *Chem Biol Drug Des* 74 (2009), pp. 16–24.
- Tatonetti, N. P., Ye, P. P., Daneshjou, R., and Altman, R. B. (2012). "Data-driven prediction of drug effects and interactions." *Sci Transl Med* 4 (2012), 125ra31.
- Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). "A genomic perspective on protein families." *Science* 278 (1997), pp. 631–7.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." *Nucleic Acids Res* 29 (2001), pp. 22–8.
- 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., and Natale, D. A. (2003). "The COG database: an updated version includes eukaryotes." *BMC Bioinf* 4 (2003), p. 41.
- Taylor, G. (1996). "Sialidases: structures, biological significance and therapeutic potential." *Curr Opin Struct Biol* 6 (1996), pp. 830–7.
- Taylor, J. D., Ababou, A., Fawaz, R. R., Hobbs, C. J., Williams, M. A., and Ladbury, J. E. (2008). "Structure, dynamics, and binding thermodynamics of the v-Src SH2 domain: implications for drug design." *Proteins* 73 (2008), pp. 929–40.
- Tedesco, R., Thomas, J. A., Katzenellenbogen, B. S., and Katzenellenbogen, J. A. (2001). "The estrogen receptor: a structure-based approach to the design of new specific hormone-receptor combinations." *Chem Biol* 8 (2001), pp. 277–87.

- Tesmer, J. J. (1999). “Two-Metal-Ion Catalysis in Adenylyl Cyclase”. *Science* 285 (1999), pp. 756–760.
- Thompson, A. A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X.-P., Trapella, C., Guerrini, R., Calo, G., Roth, B. L., Cherezov, V., and Stevens, R. C. (2012). “Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic.” *Nature* 485 (2012), pp. 395–9.
- Thoreen, C. C., Kang, S. A., Chang, J. W., Liu, Q., Zhang, J., Gao, Y., Reichling, L. J., Sim, T., Sabatini, D. M., and Gray, N. S. (2009). “An ATP-competitive mammalian target of rapamycin inhibitor reveals rapamycin-resistant functions of mTORC1.” *J Biol Chem* 284 (2009), pp. 8023–32.
- Touroutoglou, N. and Pazdur, R. (1996). “Thymidylate synthase inhibitors.” *Clin Cancer Res* 2 (1996), pp. 227–43.
- Traynelis, S. F., Wollmuth, L. P., McBain, C. J., Menniti, F. S., Vance, K. M., Ogden, K. K., Hansen, K. B., Yuan, H., Myers, S. J., and Dingledine, R. (2010). “Glutamate receptor ion channels: structure, regulation, and function.” *Pharmacol Rev* 62 (2010), pp. 405–96.
- Tsai, I. J. et al. (2013). “The genomes of four tapeworm species reveal adaptations to parasitism.” *Nature* 496 (2013), pp. 57–63.
- Tsoka, S. and Ouzounis, C. A. (2000). “Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion.” *Nat Genet* 26 (2000), pp. 141–2.
- Uitdehaag, J. C. M., Verkaar, F., Alwan, H., Man, J. de, Buijsman, R. C., and Zaman, G. J. R. (2012). “A guide to picking the most selective kinase inhibitor tool compounds for pharmacological validation of drug targets.” *Br J Pharmacol* 166 (2012), pp. 858–76.
- Unno, M., Shinohara, M., Takayama, K., Tanaka, H., Teruya, K., Doh-ura, K., Sakai, R., Sasaki, M., and Ikeda-Saito, M. (2011). “Binding and selectivity of the marine toxin neodysiherbaine A and its synthetic analogues to GluK1 and GluK2 kainate receptors.” *J Mol Biol* 413 (2011), pp. 667–83.
- Urwyler, S. (2011). “Allosteric modulation of family C G-protein-coupled receptors: from molecular insights to therapeutic perspectives.” *Pharmacol Rev* 63 (2011), pp. 59–126.
- Vanichtanankul, J., Taweechai, S., Yuvaniyama, J., Vilaivan, T., Chitnumsub, P., Kamchonwongpaisan, S., and Yuthavong, Y. (2011). “Trypanosomal dihydrofolate reductase reveals natural antifolate resistance.” *ACS Chem Biol* 6 (2011), pp. 905–11.
- Vassilatis, D. K., Hohmann, J. G., Zeng, H., Li, F., Ranchalis, J. E., Mortrud, M. T., Brown, A., Rodriguez, S. S., Weller, J. R., Wright, A. C., Bergmann, J. E., and Gaitanaris, G. A. (2003). “The G protein-coupled receptor repertoires of human and mouse.” *Proc Natl Acad Sci U S A* 100 (2003), pp. 4903–8.
- Velankar, S., Dana, J. M., Jacobsen, J., Ginkel, G. van, Gane, P. J., Luo, J., Oldfield, T. J., O’Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). “SIFTS: Structure

- Integration with Function, Taxonomy and Sequences resource.” *Nucleic Acids Res* 41 (2013), pp. D483–9.
- Velaparthi, S., Brunsteiner, M., Uddin, R., Wan, B., Franzblau, S. G., and Petukhov, P. A. (2008). “5-tert-butyl-N-pyrazol-4-yl-4,5,6,7-tetrahydrobenzo[d]isoxazole-3-carboxamide derivatives as novel potent inhibitors of Mycobacterium tuberculosis pantothenate synthetase: initiating a quest for new antitubercular drugs.” *J Med Chem* 51 (2008), pp. 1999–2002.
- Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013). “Molecular signatures of G-protein-coupled receptors.” *Nature* 494 (2013), pp. 185–94.
- Vidal, D., Thormann, M., and Pons, M. (2005). “LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities.” *J Chem Inf Model* 45 (2005), pp. 386–93.
- Vieth, M., Sutherland, J. J., Robertson, D. H., and Campbell, R. M. (2005). “Kinomics: characterizing the therapeutically validated kinase space”. *Drug Discov Today* 10 (2005), pp. 839–846.
- Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). “EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.” *Genome Res* 19 (2009), pp. 327–35.
- Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). “Structure, function and evolution of multidomain proteins.” *Curr Opin Struct Biol* 14 (2004), pp. 208–16.
- Wakayama, T., Rodriguez, I., Perry, A. C., Yanagimachi, R., and Mombaerts, P. (1999). “Mice cloned from embryonic stem cells.” *Proc Natl Acad Sci U S A* 96 (1999), pp. 14984–9.
- Wallace, I. M., O’Sullivan, O., Higgins, D. G., and Notredame, C. (2006). “M-Coffee: combining multiple sequence alignment methods with T-Coffee.” *Nucleic Acids Res* 34 (2006), pp. 1692–9.
- Walters, W. P. and Namchuk, M. (2003). “Designing screens: how to make your hits a hit.” *Nat Rev Drug Discovery* 2 (2003), pp. 259–66.
- Wang, S. and Eisenberg, D. (2003). “Crystal structures of a pantothenate synthetase from *M. tuberculosis* and its complexes with substrates and a reaction intermediate.” *Protein Sci* 12 (2003), pp. 1097–108.
- Wang, W. U., Chen, C., Lin, K.-H., Fang, Y., and Lieber, C. M. (2005). “Label-free detection of small-molecule-protein interactions by using nanowire nanosensors.” *Proc Natl Acad Sci U S A* 102 (2005), pp. 3208–12.
- Wang, X., Wang, R., Zhang, Y., and Zhang, H. (2013). “Evolutionary survey of druggable protein targets with respect to their subcellular localizations.” *Genome Biol Evol* 5 (2013), pp. 1291–7.

- Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Zhou, Z., Han, L., Karapetyan, K., Dracheva, S., Shoemaker, B. A., Bolton, E., Gindulyte, A., and Bryant, S. H. (2012a). “PubChem’s BioAssay Database.” *Nucleic Acids Res* 40 (2012), pp. D400–12.
- Wang, Y.-Y., Nacher, J. C., and Zhao, X.-M. (2012b). “Predicting drug targets based on protein domains.” *Mol BioSyst* 8 (2012), pp. 1528–34.
- Ward, K. W. and Smith, B. R. (2004a). “A comprehensive quantitative and qualitative evaluation of extrapolation of intravenous pharmacokinetic parameters from rat, dog, and monkey to humans. I. Clearance.” *Drug Metab. Dispos.* 32 (2004), pp. 603–11.
- Ward, K. W. and Smith, B. R. (2004b). “A comprehensive quantitative and qualitative evaluation of extrapolation of intravenous pharmacokinetic parameters from rat, dog, and monkey to humans. II. Volume of distribution and mean residence time.” *Drug Metab. Dispos.* 32 (2004), pp. 612–9.
- Weisenberg, R. C., Borisy, G. G., and Taylor, E. W. (1968). “The colchicine-binding protein of mammalian brain and its relation to microtubules.” *Biochemistry* 7 (1968), pp. 4466–79.
- Wetlaufer, D. B. (1973). “Nucleation, rapid folding, and globular intrachain regions in proteins.” *Proc Natl Acad Sci U S A* 70 (1973), pp. 697–701.
- Whitby, F. G., Masters, E. I., Kramer, L., Knowlton, J. R., Yao, Y., Wang, C. C., and Hill, C. P. (2000). “Structural basis of the allosteric behaviour of phosphofructokinase.” *Nature* 408 (2000), pp. 140–145.
- Whittle, P. J. and Blundell, T. L. (1994). “Protein structure-based drug design.” *Annu Rev Biophys Biomol Struct* 23 (1994), pp. 349–75.
- Windh, R. T. and Manning, D. R. (2002). “Analysis of G protein activation in Sf9 and mammalian cells by agonist-promoted [³⁵S]GTP gamma S binding.” *Methods Enzymol* 344 (2002), pp. 3–14.
- Wittmann, J. G., Heinrich, D., Gasow, K., Frey, A., Diederichsen, U., and Rudolph, M. G. (2008). “Structures of the human orotidine-5’-monophosphate decarboxylase support a covalent mechanism and provide a framework for drug design.” *Structure* 16 (2008), pp. 82–92.
- Wolf, Y. I. and Koonin, E. V. (2012). “A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes.” *Genome Biol Evol* 4 (2012), pp. 1286–94.
- Workman, P. and Collins, I. (2010). “Probing the probes: fitness factors for small molecule tools.” *Chem Biol* 17 (2010), pp. 561–77.
- Wu, W.-I., Voegtli, W. C., Sturgis, H. L., Dizon, F. P., Vigers, G. P. A., and Brandhuber, B. J. (2010). “Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition.” *PLoS One* 5 (2010), e12913.
- Xu, H., Ramsey, I. S., Kotecha, S. A., Moran, M. M., Chong, J. A., Lawson, D., Ge, P., Lilly, J., Silos-Santiago, I., Xie, Y., DiStefano, P. S., Curtis, R., and Clapham, D. E. (2002). “TRPV3 is a calcium-permeable temperature-sensitive cation channel.” *Nature* 418 (2002), pp. 181–6.

- Yamanishi, Y., Pauwels, E., Saigo, H., and Stoven, V. (2011). "Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions". *J Chem Inf Model* (2011), p. 110505071700060.
- Yang, Y., Gao, P., Liu, Y., Ji, X., Gan, M., Guan, Y., Hao, X., Li, Z., and Xiao, C. (2011). "A discovery of novel Mycobacterium tuberculosis pantothenate synthetase inhibitors based on the molecular mechanism of actinomycin D inhibition." *Bioorg Med Chem Lett* 21 (2011), pp. 3943–6.
- Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. (2007). "Drug-target network." *Nat Biotechnol* 25 (2007), pp. 1119–26.
- Yourno, J., Kohno, T., and Roth, J. R. (1970). "Enzyme evolution: generation of a bifunctional enzyme by fusion of adjacent genes." *Nature* 228 (1970), pp. 820–4.
- Yuvaniyama, J., Chitnumsub, P., Kamchonwongpaisan, S., Vanichtanankul, J., Sirawaraporn, W., Taylor, P., Walkinshaw, M. D., and Yuthavong, Y. (2003). "Insights into antifolate resistance from malarial DHFR-TS structures." *Nat Struct Mol Biol* 10 (2003), pp. 357–65.
- Zambrowicz, B. P. and Sands, A. T. (2003). "Knockouts model the 100 best-selling drugs—will they model the next 100?": *Nat Rev Drug Discovery* 2 (2003), pp. 38–51.
- Zanders, E. D., Bailey, D. S., and Dean, P. M. (2002). "Probes for chemical genomics by design". *Drug Discov Today* 7 (2002), pp. 711–718.
- Zdobnov, E. M., Lopez, R., Apweiler, R., and Etzold, T. (2002). "The EBI SRS server—new features." *Bioinformatics* 18 (2002), pp. 1149–50.
- Zhang, Q. C., Petrey, D., Norel, R., and Honig, B. H. (2010). "Protein interface conservation across structure space." *Proc Natl Acad Sci U S A* 107 (2010), pp. 10896–901.
- Zhao, C., Sun, M., Bennani, Y. L., Miller, T. R., Witte, D. G., Esbenshade, T. A., Wetter, J., Marsh, K. C., Hancock, A. A., Brioni, J. D., and Cowart, M. D. (2009). "Design of a new histamine H3 receptor antagonist chemotype: (3aR,6aR)-5-alkyl-1-aryl-octahydropyrrolo[3,4-b]pyrroles, synthesis, and structure-activity relationships." *J Med Chem* 52 (2009), pp. 4640–9.
- Zhi, L., Tegley, C. M., Marschke, K. B., and Jones, T. K. (1999). "Switching androgen receptor antagonists to agonists by modifying C-ring substituents on piperidino[3,2-g]quinolinone". *Bioorg Med Chem Lett* 9 (1999), pp. 1009–1012.
- Zhou, Q., Renard, J.-P., Le Fric, G., Brochard, V., Beaujean, N., Cherifi, Y., Fraichard, A., and Cozzi, J. (2003). "Generation of fertile cloned rats by regulating oocyte activation." *Science* 302 (2003), p. 1179.
- Ziegler, P. and Dittrich, K. R. (2004). "Three Decades of Data Integration - All Problems Solved?": *In 18th IFIP World Computer Congress (WCC 2004), Volume 12, Building the Information Society*. 2004, pp. 3–12.

- Zolotukhin, S., Potter, M., Hauswirth, W. W., Guy, J., and Muzyczka, N. (1996). "A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells." *J Virol* 70 (1996), pp. 4646–54.
- Zuckerkindl, E. and Pauling, L. (1965). "Evolutionary divergence and convergence in proteins". *Evolving Genes and Proteins*. Ed. by V Bryson and H. J. Vogel. New York: Academic Press, 1965, pp. 97–165.