Phylogenetic Signals in Protein Data

a dissertation presented
by
Chongli Qin
to
The Department of Chemistry

This dissertation is submitted
for the degree of
Doctor of Philosophy

Selwyn College, University of Cambridge
Cambridge, Cambridgeshire
December 2020

- This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.
- It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.
- It does not exceed the prescribed word limit for the Chemistry Degree Committee.

Thesis advisor: Professor Lucy J. Colwell

Chongli Qin

Phylogenetic Signals in Protein Data

Abstract

Structural biology has seen major advances over the past decade. In the area of protein structure prediction we have seen significant increase in accuracy with the discovery of coevolutionary signals in a multiple sequence alignment (MSA). Unlike methods which fold proteins using molecular dynamic (MD) simulations, these coevolutionary methods make use of correlation information to fold large protein structures orders of magnitudes faster.

Often the correlation signals in a MSA are a strong indicator that a pair of amino acids are sufficiently close together to be in contact, thus interacting with each other. It has been shown that accurate inference of amino acid pairs that are in contact in the protein gives rise to accurate prediction of protein structure itself. Hence, statistical inference of amino acid pairs in contact is an important problem for protein folding.

However, one of the major challenges of these statistical inference methods is that levels of noise significantly overwhelm the relevant signal for protein data. In this thesis, we attempt to alleviate one of the most important sources of noise which is also one that is often ignored: spurious correlations induced by phylogeny. To this end, we introduce a novel method for disentangling phylogenetic noise from the relevant structural signals. This method is grounded in an extension to a well-known theorem in Random Matrix Theory. Through extensive analysis on both synthetic and protein data, we demonstrate that it is possible to disentangle these two sources of information. Crucially, we find that the phylogenetic correlations can be largely removed by finding principal modes of the empirical correlation matrix where its corresponding eigenvalue satisfies a power-law.

Contents

This is dedicated to my family and friends.

Acknowledgments

I would like to acknowledge my parents, for being there for me during my time in Cambridge, often visiting me with food and home-comforts. I want to thank my friends – they are truly inspirational and they have definitely inspired my love for research. In particular, I want to thank Laura Mitchell, Raman Ganti, Lukas Wutschitz, Peter Wirnsberger, William Drazin and Oliver Strickson. I want to thank my sister and her partner Kevin, for always being there when I needed an escape from Cambridge. I want to thank my partner Tobi, who is always a ray of sunshine and happiness, it really helped me get through the last few weeks in writing up this PhD thesis. I would like to thank for my examiners Professor Tom McLeish and Professor Daan Frenkel for the very useful feedback. Last but definitely not least, I would like to thank my supervisor Lucy Colwell for giving me the opportunity to work on such an interesting project with her.

The internal machinery of life, the chemistry of the parts, is something beautiful. And it turns out that all life is interconnected with all other life.

Richard Feynman

1

Introduction

Proteins are indispensable to every living organism – it is still an ongoing challenge to predict their tertiary structures. Over the past several decades, significant progress has been made to folding a protein using molecular dynamics (MD) simulations. However, MD is expensive and often computationally infeasible for proteins of a larger size.

Following the technology boost over the past decade, we are able to accumulate and process enormous amounts of data, such as US stocks data, medical data or data for weather prediction. In the area of proteins, we now have access to large databases of protein sequences within the same protein family. Consequently, statistical inference methods are now accessible as competitive alternatives to solving protein folding via MD simulations such as the recent AlphaFold system[54]. Moreover, these statistical methods are capable of inferring the shape of a protein with significantly less compute. For these statistical inference methods, the key is to understand which signals to extract from the data and how they can be used to decode the physics behind the fold of a protein.

The link between the structure of a protein and how the structure gives rise to signals in the data has already been well studied. In particular, the correlation in the data is a result of coevolving pairs of amino acids which are sufficiently close together, or in contact, within the protein; a finding empirically observed by Altschuh et al. [1], Miller & Eisenberg[39]. The amino-acid pairs in contact can be viewed as a set of evolutionary constraints which conserve the shape and the function of the protein. This set of constraints can be expressed as a "contact map": a matrix of ones and zeros where the (i, j)th element is one if and only if amino acids at position i and j in the protein sequence are in contact with each other. Fig. 1.1 shows an example of such a contact map.

As also shown in Fig. 1.1, the regularity of the structure of a protein is reflected in the contact map. Thus, this map can also be used to infer the structure of a protein. However, for many families of proteins the data available is still limited and in the low-data regime, noise factors such as phylogeny become more prevalent and inference of the contact map becomes harder.

1.1 Molecular Dynamics

A classical approach to predicting a protein's tertiary structure is via molecular dynamical simulation. For a general overview see Frenkel & Smit[21]. These approaches try to solve the intrinsic physics underlying the 3D structure. They achieve this, by minimizing the Gibbs free energy of the protein structure; resolving forces such as electrostatic or Van der Waals. However, for this methodology to be accurate, it requires accurate numerical simulation at atomic resolution. Thus this approach is hard to scale and for sufficiently large proteins it is computationally infeasible.

On the other hand, statistical inference approaches can scale well with respect to the size of the protein. In the section below, we will give a general overview of how to exploit the data (multiple sequence alignment) to give us an approximation to a protein's tertiary structure.

1.2 Evolutionary Constraints From Multiple Sequence Alignment

The goal of statistical inference based approaches is to approximate an energy function of a protein which can be minimized to find the protein structure.

To find this approximate energy, we consider a dataset of sequences that are related to each other via their evolutionary history. Concretely, consider the example of haemoglobin. This is a protein which exists in rats, dogs as well as humans. Perhaps not surprisingly, the haemoglobins present in these different organisms are all very similar in both their primary sequence and tertiary structure; as they are all related to each other strongly via evolution. Importantly, although the primary sequences are similar, they do differ while their tertiary structure remains largely the same. This differ-

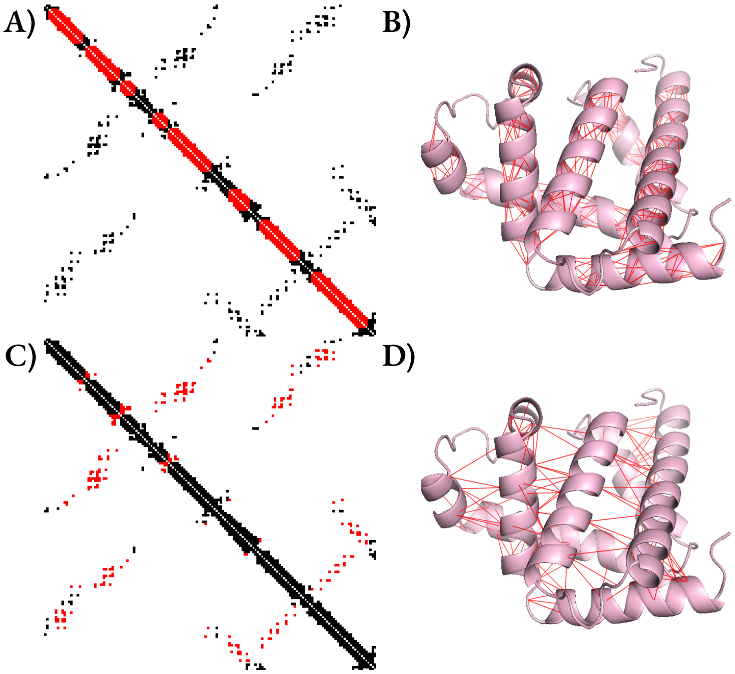


Figure 1.1: This plot shows how the regularities of the protein tertiary structure is reflected in the contact map of oxymyoglobin. A-B) shows the intra helical-contacts correspond to the diagonal of the contact map, whereas C-D) shows the contacts between these helices are on the off-diagonal of the contact map where the regularity of the structure is also reflected. We can exploit the regularities in the contact map to infer the structure of the protein.

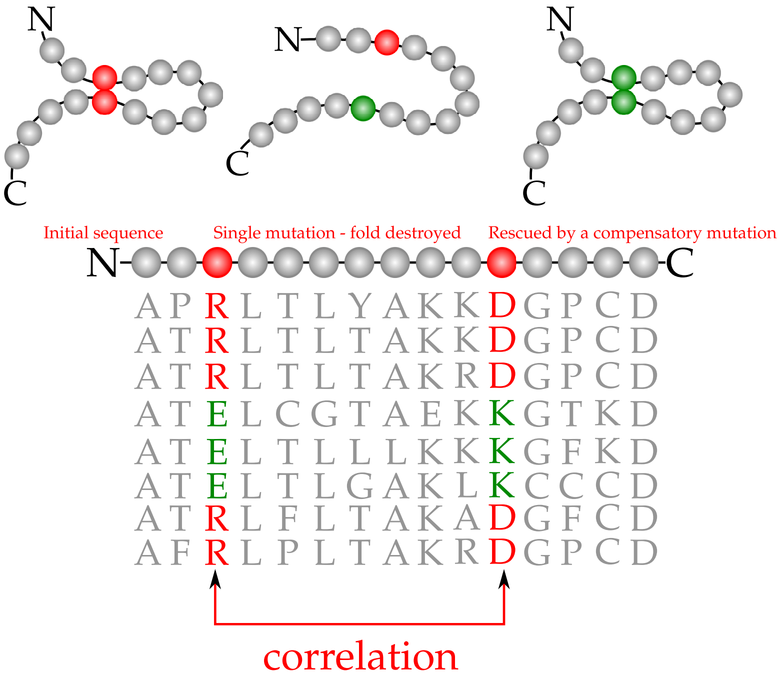
ence between the primary sequences within a set of related proteins offers crucial information about the evolutionary constraints which govern its structure. The concept of evolutionary constraints and related sequences (multiple sequence alignment) are described in the sections below.

1.2.1 Evolutionary Constraints

Consider the case of two amino acids held together by an electrostatic force. One amino acid positively charged while the other is negatively charged. If one of these amino acid mutates to another amino acid with a different charge, the two amino acids will repel rather than attract each other. This would break the protein structure. To ensure the proteins are robust to these mutations, a compensatory mutation (at the paired amino acid) is required to maintain the attractive force. This process is depicted in Fig. 1.2, a compensatory mutation prevents the protein from mis-folding. More importantly, this coevolution between amino acids in contact have been observed by many experiments [1,39], giving empirical evidence as to why our proteins are robust to mutations.

1.2.2 Multiple Sequence Alignment

Given the preliminaries above, we can now turn to an explanation of what a multiple sequence alignment (MSA) is.


N
C
N
C
N
C
Initial sequence
Single mutation - fold destroyed
Rescued by a compensatory mutation
N
C
correlation

A P R L T L Y A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K R D G P C D
A T E L C G T A E K K G T K D
A T E L T L L L K K K G F K D
A T E L T L G A K L K C C C D
A T R L F L T A K A D G F C D
A F R L P L T A K R D G P C D

Figure 1.2: A depiction of the coevolution process between two amino acids in contact. The red and green colours indicate the pairs of amino acids that are attracted to each other. The initial protein structure is depicted in the top-left. If only one residue mutates to green, the force between the pair may no longer be attractive (middle figure in the top row), therefore the contact is destroyed. To retain the contact, a compensatory mutation of the other residue is necessary (top-right figure). A corresponding multiple-sequence alignment is depicted below showing the coevolution of the pair which avoids mis-folding.

A multiple sequence alignment is a set of sequences with the same tertiary structure but slightly different primary sequences. Concretely, we consider a set of related sequences, or sequences within the same *homolog* that are related via a phylogenetic tree and are assumed to have the same tertiary structure. Ideally, these sequences would only differ by mutations, however in practise, they can also differ by their length. In this work, we assume the sequences have been pre-aligned by standard alignment algorithms (for more general overview see Durbin et al.[15]) such that they are of the same length.

In this context, we denote the MSA as $X = [x^1, \cdots, x^n]$, consisting of n related protein sequences, x^i, each of which is of length p. The sequences are aligned such that each column in X contains the amino acids which are functionally equivalent.

1.2.3 Correlations to Coevolution

The differences between these primary sequences in a multiple sequence alignment do not violate the evolutionary constraints. To see this we note when two amino acids are close enough in the native structure to interact with each other – if one of these amino acids mutate it might change the nature of the interaction, therefore, also the tertiary structure of a protein. Since the tertiary structure for related sequences are roughly the same, the differences in the primary sequences must satisfy evolutionary constraints. In particular, the mutations observed in these sequences must show signals of coevolution, as described in Section 1.2.1, between pairs of amino acids which are in contact.

To move from coevolution to signal, the central idea is to use multiple sequence alignments to infer a set of structural constraints. Coevolution implies that we should be able to use correlated amino-acid pairs, as is shown in Fig. 1.2, to infer which amino acids are interacting or in contact in its native structure. Thus, correlations within a collection of sequences with very similar structure and function can be used to infer a protein's contact map, see Fig. 1.1. As alluded to before, creating this contact map allows us to construct an energy function that respects the structural constraints of a protein structure; which can subsequently be used as a rough guide to fold the protein.

This idea, though intuitive and elegant, has come with many complications. The coevolutionary signals related to the evolutionary constraints should be given by the correlations between the columns of X. Thus, a naive way of finding the residues in contact is through the empirical correlation matrix

$$C = \frac{1}{n-1}(X-\overline{X})^T(X-\overline{X}).$$

However, many have found that this matrix is an extremely poor measure of residues (amino-acids) which are in contact[11,9,55]. This is because of the significant level of noise contained in collected sequences. In section 1.3 we will address many such possible source of noises.

Other Phenotypic Interactions In this thesis we will only be investigating the evolutionary constraints when a protein is in its native state and is treated as a "static" structure. However there are many other forms of phenotypic interactions in folding pathways of a protein which can induce correlation within the data. In these cases contact analysis is also useful. These can include allosteric interactions whereby a substrate can bind to an allosteric site[61] or other non-native interactions, for example McLeish[38] demonstrated that the folding strategies of a protein might be dependent on both native and non-native interactions - highlighting that more information can be gained on protein structures by studying it's nonnative interactions.

1.3 The Problem of Noise

Phylogenetic Noise To address why the empirical covariance matrix is a poor measure of the correlations induced by residues in contact, firstly we note that the sequences themselves are correlated by definition. Concretely, the sequences x^i in X are related to each other via a phylogenetic tree.

When the samples of the data are correlated - it is intrinsically difficult to extract features (evolutionary constraints) which are inherent in the sequences themselves. There has been a lot of work done to find ways of disentangling the phylogenetic signals from the phenotypic (residues in contact) signals, however, most existing methods are derived from heuristics drawn from intuition. We will refer to these methods in more detail in Section 2.3. The main focus of this thesis is on the novel methods which we have developed using the tools from Random Matrix Theory to disentangle the correlations induced by the similarity between the sequences and the correlations induced by coevolution.

Transitivity: Another example of a source of noise corrupting the coevolution signal is the transitivity of correlations. Transitive correlations can happen if a residue at position A is in contact with residues at positions B, and B in turn is in contact with C. See Fig. 2.2. Then if A mutates, compensatory mutations at B, C will be needed to maintain the favourable energy setup. Thus residues at positions A and C are indirectly correlated via B. There have been many developments to filter out the indirect correlations, one of which is the enforcement of a global statistical model such as by Ekeberg et al. [18], Marks et al. [36]. We will further expand on these global statistical model techniques in Section 2.2.1.

1.4 Random Matrix Theory for Phylogeny

In this thesis, our focus is on phylogenetic bias and its effects on our ability to infer the tertiary structure of proteins. Towards this end, we use Random Matrix Theory (RMT) to disseminate the signals induced by phylogeny from the signals induced by the evolutionary constraints in the protein itself.

RMT methods are rigorous statistical approaches which analyse the properties of *large* random matrices. This field has developed rapidly over the past several decades. The first physical application was found by Wigner [63]. He found that analysing the eigen-properties of random matrices corresponds to finding the energy spectra in systems of quantum mechanics, thereby reducing the complexity of the problem. Now RMT methods are widely used in analysing financial data and many other statistical applications [5,8,59,58]. Due to the steadfast progress in the collection of DNA sequences, now over 4400 full genome sequences are accessible, equivalent to 3.6×10^7 proteins sequences [12]. As random matrix theory hinges on the necessity of large matrices, many of the protein families in this dataset fall within the regime in which random matrix theory can be applied with precision.

Another motivation behind using random matrix theory is that one of its critical discoveries is that when the elements of a matrix are random, the eigenvalue distribution of its empirical covariance matrix is compactly supported. This is a property which can be used to see if the elements of the matrix are random. An important question then is whether we can discover these signatures when the matrices are correlated via a phylogenetic tree or via structural constraints. Moreover ine might wonder: can they be used to disentangle different sources of signal?

1.5 Contribution

The contribution of this thesis is two-fold*. Firstly, we develop a novel analysis technique that hinges on Random Matrix Theory to give insight into how phylogenetic dependencies between protein sequences can affect covariance matrices. Using this analysis, we find that phylogenetic correlation between the samples induces a distinct signature in the spectral distribution of the covariance matrix.

Secondly, this distinct signature is a feature that is predictive of which eigenvectors in the empirical covariance matrix are corrupted by phylogenetic noise. Using this observation, we show that not only can we remove phylogenetic corruption from Boltzmann-generated sequences that are related via a phylogenetic tree, but such an analysis also transfers directly to real protein sequences.

*Note that some of the main findings is also published in Qin & Colwell [48]

2

Background on Contact Prediction

The literature review will be split into three different sections. Section 2.1 will highlight the research showing that amino-acids which are close together (in contact) in the protein structure should co-evolve. Section 2.2 will detail the work which has been done to improve the contact prediction. Section 2.3 will outline the work which has been done to investigate how phylogeny affects our inference ability for contact prediction.

2.1 Coevolution

One of the first to make observations that the spatial coordinates of amino acids can be inferred by looking at the mutation patterns of the amino acid residues is Altschuh et al. [1]. They found that the mutation patterns of amino acids become more conservative near the binding regions of some viruses related to tobacco mosaic virus. This indicates that it might be possible to also infer the binding regions of viruses by looking at mutation patterns in proteins. Thus this paper sparked many statistical studies into looking for ways of inferring spatial coordinates of amino-acids from looking at the mutation patterns of amino-acids. This idea was taken up by Shindyalov et al. [55]. In this study, they show that it is possible to infer which amino acids are in contact with each other within the tertiary structure of a protein by looking at the correlated mutations of amino acid sequences alone. They performed this analysis for 67 protein families, demonstrating that mutagenesis analysis is a potential methodology for even protein folding itself.

2.2 Contact Prediction

Since the work done by Altschuh et al. [1], Shindyalov et al. [55], the field has made significant progress towards enhancing contact prediction methods. Most of these methods fall into two categories: local and global statistical models. We outline some of the methods below.

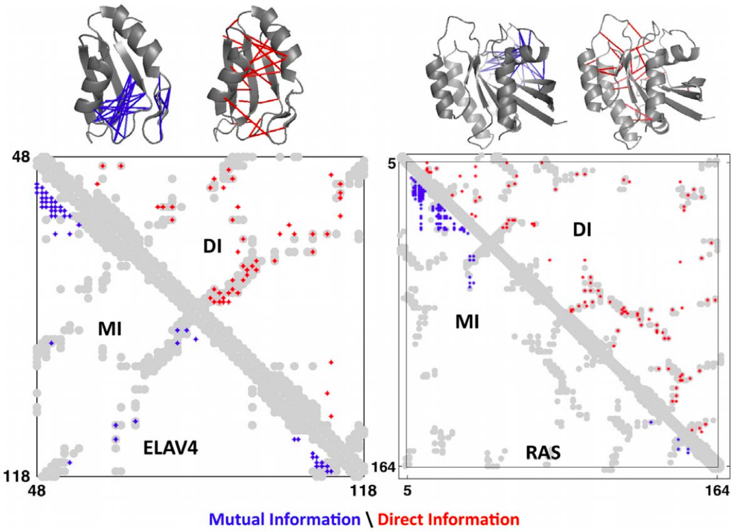


Figure 2.1: Here, we compare the difference when we use a local model (MI) vs global model (DI) to predict contacts. DI is detailed in Section 2.2.2. The predicted contacts for DI (red) accurately match the positions of the observed contacts (grey), whereas many of the contact predictions using MI (blue) are false. Figure adapted from Marks et al.[36].

2.2.1 Local Model vs Global Model

Fig. 1.2 shows amino-acid sequences within the same homolog being collected together and assembled into a multiple sequence alignment. More concretely, a set of amino-acid sequences within the same homolog is denoted by $X = (\mathbf{x}^1, \cdots, \mathbf{x}^n)^T$ throughout, where each sequence is of length p, $\mathbf{x}^j = (x^j_1, \cdots, x^j_p)$, and $x^j_i \in \{1, 2, \cdots, q\}$. For amino-acid sequences $q = 21$ which stands for the twenty amino acids and one extra for the gap.

Here,'local' and 'global' refers to whether the contact prediction between positions i and j is dependent on the local residues in question (i, j) or on the entire protein sequence $(1, \cdots p)$.

Local Model

The most common method used in this category is mutual information (MI)[14]; a way of measuring the correlation between two amino-acid positions i, j. The mathematical definition of MI is

$$(MI)_{\mathrm{ij}} = \sum_{a,b} P_{ij}(a,b) \log\left(\frac{P_{ij}(a,b)}{P_i(a)P_j(b)}\right) = \sum_{k_{\mathrm{i}},k_{\mathrm{j}}} f_{\mathrm{ij}}(k_{\mathrm{i}},k_{\mathrm{j}}) \ln\left\{\frac{f_{\mathrm{ij}}(k_{\mathrm{i}},k_{\mathrm{j}})}{f_{\mathrm{i}}(k_{\mathrm{i}})f_{\mathrm{j}}(k_{\mathrm{j}})}\right\} , \tag{2.1}$$

where f_{i} and f_{ij} are the empirically observed moments:

$$f_{\mathrm{i}}(k) = \frac{1}{n}\sum_{\mathrm{s}=1}^{n} \mathcal{I}(x_{\mathrm{i}}^{\mathrm{s}} = k) \tag{2.2}$$

$$f_{\mathrm{ij}}(k,l) = \frac{1}{n}\sum_{\mathrm{s}=1}^{n} \mathcal{I}(x_{\mathrm{i}}^{\mathrm{s}} = k)\mathcal{I}(x_{\mathrm{j}}^{\mathrm{s}} = l) , \tag{2.3}$$

$\mathcal{I}$ is the indicator function. Eq. (2.1) shows that MI measures how much information is lost if a pairwise probability model is replaced by two independent probability models, this is zero if and only if the distributions are independent, otherwise it is always positive. The calculation of this is straightforward since it only involves the local frequencies at residue positions i and j.

Local measures are limited by transitivity, which is explained in Figure 2.2, this can significantly reduce the accuracy of the contact predictions. Fodor & Aldrich[20] found that the extent to which the contact predicted differed from the native contacts is substantial. Thus to disentangle the direct and indirect correlations, a global correlation measure is crucial. Another reason why MI is insufficient is that it is significantly affected by phylogenetic bias, a detail we will elaborate on in

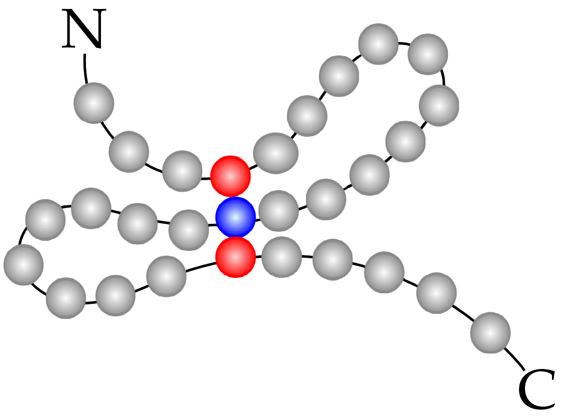


Figure 2.2: This shows one residue, shown in blue, which is in contact with two other residues, shown in red. A local correlation measure will score all three pairs highly with no way of distinguishing which ones are contacts and which ones are not. This may result in a false contact being predicted between the red pair.

Section 2.3.

Global Model

A global correlation measure must follow from a probability distribution dependent on the entire sequence, $\mathbf{P}(x_1, \cdots, x_p)$. This distribution must also agree with the observed frequencies. We enforce the first two moments of $\mathbf{P}$ to match the first two empirically observed moments

$$\mathbf{P}(x_\mathrm{i} = k) = f_\mathrm{i}(k) \quad \mathrm{i} = 1, \cdots, p \tag{2.4}$$

$$\mathbf{P}(x_\mathrm{i} = k, x_\mathrm{j} = l) = f_\mathrm{ij}(k, l) \quad \mathrm{i}, \mathrm{j} = 1, \cdots, p \quad \mathrm{i} < \mathrm{j}\,. \tag{2.5}$$

There are many probability distributions which satisfy Equations (2.4) and (2.5). So which probability should we choose? It has been observed that we should always choose the model that places no more constraints than necessary. Maximum entropy models[3] are models which maximize the entropy while enforcing the constraints shown in Eqs. (2.4), (2.5). Formally, maximizing the probability model with respect to its Shannon entropy, $S = -\sum_X \mathbf{P}(X)\ln(\mathbf{P}(X))$ while satisfying the constraints gives rise to the Boltzmann distribution. This distribution can be found using Lagrangian multipliers. For an Ising model where the sequence variables are either 1 or -1, this results in the following functional form :

$$\mathbf{P}(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\left\{ \sum_{\mathrm{i}=1}^{p} h_{\mathrm{i}} x_{\mathrm{i}} + \sum_{\mathrm{i}<\mathrm{j}} J_{\mathrm{ij}} x_{\mathrm{i}} x_{\mathrm{j}} \right\} . \tag{2.6}$$

Here h_{i} and the symmetric matrixJ_{ij} are the fields and true interactions respectively and

$$E(\mathbf{x}) = -\sum_{\mathrm{i}=1}^{p} h_{\mathrm{i}} x_{\mathrm{i}} - \sum_{\mathrm{i}<\mathrm{j}} J_{\mathrm{ij}} x_{\mathrm{i}} x_{\mathrm{j}} \tag{2.7}$$

is the energy of the system. The generalised case when there are more than 2 variables is given by

$$\mathbf{P}(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\left\{ \sum_{\mathrm{i}=1}^{p} h_{\mathrm{i}}(x_{\mathrm{i}}) + \sum_{\mathrm{i}<\mathrm{j}} J_{\mathrm{ij}}(x_{\mathrm{i}}, x_{\mathrm{j}}) \right\} , \tag{2.8}$$

known as the Potts Model. Note that Potts model is often understood as the planar Potts Model, defined in Sec 5.1. This is a more generalised definition as it does not impose functional form of$J$

nor h. For the Ising model the constraints (2.4) and (2.5) are equivalent to the average moments

$$\langle m_{\mathrm{i}} \rangle = \frac{1}{n} \sum_{\mathrm{k}=1}^{n} x_{\mathrm{i}}^{\mathrm{k}} \,, \tag{2.9}$$

$$\langle m_{\mathrm{i}} m_{\mathrm{j}} \rangle = \frac{1}{n} \sum_{\mathrm{k}=1}^{n} x_{\mathrm{i}}^{\mathrm{k}} x_{\mathrm{j}}^{\mathrm{k}} \,. \tag{2.10}$$

For the Potts model, the constraints (2.4) and (2.5) become the marginal probability distributions :

$$\mathbf{P}_{\mathrm{i}}(k_{\mathrm{i}}) = \sum_{\{k_s : s \neq \mathrm{i}\}} \mathbf{P}(k_1, \cdots, k_p) = f_{\mathrm{i}}(k_{\mathrm{i}}) \tag{2.11}$$

$$\mathbf{P}_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = \sum_{\{k_s : s \neq \mathrm{i,j}\}} \mathbf{P}(k_1, \cdots, k_p) = f_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) \,, \tag{2.12}$$

where $\mathbf{P}_{\mathrm{i}}(k_{\mathrm{i}})$ is the shorthand for $\mathbf{P}(x_{\mathrm{i}} = k_{\mathrm{i}})$.

Finding the parameters of this distribution is known as the 'Inverse Ising/Potts problem'. This problem is difficult to resolve due to the intractability of the normalisation constant $\mathcal{Z}$. The number of computations required for the evaluation of $\mathcal{Z}$ scales as q^p, where p is the number of residues and q is the number of types of residues. As a result, most people find approximations in order to avoid $\mathcal{Z}$. In addition, for the Potts model the calculations for Equations (2.11), (2.12) are also computationally expensive and hence approximations are made for these marginal probabilities.

2.2.2 Inverse Ising Problem: Finding h and J

An intuitive approach to estimating the parameters h_{i} and J_{ij} is to maximise the likelihood of the test statistic X. Even though they are connected by phylogeny, a simplification is made here to treat the sequences as independent models. The log-likelihood function is :

$$l(h_{\mathrm{i}},J_{\mathrm{ij}}) = -n\ln(\mathcal{Z}) - \sum_{\mathrm{k}=1}^{n}\sum_{\mathrm{i}=1}^{p} h_{\mathrm{i}}(x_{\mathrm{i}}^{\mathrm{k}}) - \sum_{\mathrm{k}=1}^{n}\sum_{\mathrm{i}=1}^{p-1}\sum_{\mathrm{j}=\mathrm{i}+1}^{p} J_{\mathrm{ij}}(x_{\mathrm{i}}^{\mathrm{k}}, x_{\mathrm{j}}^{\mathrm{k}})\ . \tag{2.13}$$

In addition, we can analytically determine the functional form of $J_{ij}(x,y)$ and $h_i(x)$ by noting that the following expressions are equivalent

$$\sum_{\mathrm{k}=1}^{n} h_{\mathrm{i}}(x_{\mathrm{i}}^{\mathrm{k}}) = n\sum_{\mathrm{s}=1}^{q} h_{\mathrm{i}}(\mathrm{s})f_{\mathrm{i}}(\mathrm{s})$$

$$\sum_{\mathrm{k}=1}^{n} J_{\mathrm{ij}}(x_{\mathrm{i}}^{\mathrm{k}}, x_{\mathrm{j}}^{\mathrm{k}}) = n\sum_{\mathrm{s}=1}^{q}\sum_{\mathrm{t}=1}^{q} J_{\mathrm{ij}}(\mathrm{s},\mathrm{t})f_{\mathrm{ij}}(\mathrm{s},\mathrm{t})\ .$$

To see that the above is equivalent, first we substitute the above into Equation (2.13). Differentiating with respect to the parameters h and J then gives:

$$-\frac{\partial}{\partial h_{\mathrm{i}}(\mathrm{s})}\ln\mathcal{Z} = f_{\mathrm{i}}(\mathrm{s}) \tag{2.14}$$

$$-\frac{\partial}{\partial J_{\mathrm{ij}}(\mathrm{s},\mathrm{t})}\ln\mathcal{Z} = f_{\mathrm{ij}}(\mathrm{s},\mathrm{t})\ . \tag{2.15}$$

Since $-\frac{\partial}{\partial h_{\mathrm{i}}(\mathrm{s})}\ln\mathcal{Z} = \mathbf{P}_{\mathrm{i}}(s)$ and $-\frac{\partial}{\partial J_{\mathrm{ij}}(\mathrm{s},\mathrm{t})}\ln\mathcal{Z} = \mathbf{P}_{\mathrm{ij}}(s,t)$, this reduces Equations (2.14) and (2.15) to Equations (2.4) and (2.5). Thus the above are equivalent by self-consistency.

From this, finding parameters h and J becomes the following constrained optimization problem:

$$l(h_{\mathrm{i}},J_{\mathrm{ij}}) = -n\ln(\mathcal{Z}) - n\sum_{\mathrm{i}=1}^{p}\sum_{\mathrm{s}=1}^{q} h_{\mathrm{i}}(\mathrm{s})f_{\mathrm{i}}(\mathrm{s}) - n\sum_{\mathrm{i}=1}^{p-1}\sum_{\mathrm{j}=\mathrm{i}+1}^{p}\sum_{\mathrm{s}=1}^{q}\sum_{\mathrm{t}=1}^{q} J_{\mathrm{ij}}(\mathrm{s},\mathrm{t})f_{\mathrm{ij}}(\mathrm{s},\mathrm{t}) \tag{2.16a}$$

$$\mathbf{P}(x_{\mathrm{i}} = k) = f_{\mathrm{i}}(k) \quad \mathrm{i} = 1,\cdots,p \tag{2.16b}$$

$$\mathbf{P}(x_{\mathrm{i}} = k, x_{\mathrm{j}} = l) = f_{\mathrm{ij}}(k,l) \quad \mathrm{i},\mathrm{j} = 1,\cdots,p \quad \mathrm{i} < \mathrm{j}. \tag{2.16c}$$

Here $P(x)$ is given by Eq. (2.8). To circumvent the intractability of $\mathcal{Z}$, we can make the following approximations.

Pseudolikelihood: As the Inverse Ising problem is intractable due to the intractability of the partition function $\mathcal{Z}$. Ekeberg[18] shows that we can make this tractable by approximating the Boltzmann distribution via the following

$$\mathbf{P}(x_1,\cdots x_p) = \prod_{i=1}^{p}\mathbf{P}(x_i|x_{-i}),$$

where x_{-i} denotes the vector $(x_1,\cdots,x_p)$ with the ith element taken out.

Mean Field Approximations: The mean field approach for treating the Inverse Ising problem is to approximate quadratic coupling terms with a linear term by adding an external field. This means that each spin can be treated independently, which greatly simplifies the calculation of $\mathcal{Z}$.

Morcos et al.[41] generalised the mean field approximation for the Potts model to estimate the couplings J. They used an expansion for a Gibbs potential which was first introduced by Plefka[47], a neater derivation is given by Georges and Yedidia[23]. Firstly, we introduce the perturbed energy

$$E(\mathbf{x}, \varepsilon) = -\sum_{\mathrm{i}=1}^{p} h_{\mathrm{i}}(x_{\mathrm{i}}) - \varepsilon \sum_{\mathrm{i}<\mathrm{j}} J_{\mathrm{ij}}(x_{\mathrm{i}}, x_{\mathrm{j}}) \ ,$$

where the parameter ε can be tuned to interpolate between the independent system ($\varepsilon = 0$) and the original system ($\varepsilon = 1$). The corresponding Gibbs potential is given by

$$\mathcal{G}(\varepsilon) = \sum_{\mathrm{i}=1}^{p} \sum_{k=1}^{q-1} h_{\mathrm{i}}(k) \mathbf{P}_{\mathrm{i}}(k) - \ln(\mathcal{Z}(\varepsilon)) \ , \tag{2.17}$$

where

$$\mathcal{Z}(\varepsilon) = \sum_{\mathbf{x}} \exp\{-E(\mathbf{x}, \varepsilon)\} \ .$$

This potential ensures that Equation (2.11) is met for all ε. Due to the gauge of the couplings and the normalisation of the marginals, q is not an independent variable, hence the sum of the variables is only up to $q - 1$ in Equation (2.17).

The first and second derivate of Equation (2.17) is

$$h_{\mathrm{i}}(k) = \frac{\partial \mathcal{G}}{\partial \mathbf{P}_{\mathrm{i}}(k)} \tag{2.18}$$

$$(C^{-1})_{\mathrm{ij}}(k, l) = \frac{\partial^2 \mathcal{G}}{\partial \mathbf{P}_{\mathrm{i}}(k) \mathbf{P}_{\mathrm{j}}(l)} \ , \tag{2.19}$$

where

$$C_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = f_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) - f_{\mathrm{i}}(k_{\mathrm{i}}) f_{\mathrm{j}}(k_{\mathrm{j}}) \ .$$

Expanding the Gibbs potential around the independent system where $\varepsilon = 0$ gives

$$\mathcal{G}(\varepsilon) = \mathcal{G}(0) + \varepsilon \left.\frac{\partial G}{\partial \varepsilon}\right|_{\varepsilon=0} + \mathcal{O}(\varepsilon^2) \ . \tag{2.20}$$

The first two terms in the expansion can be found explictly. The first term is the Gibbs potential for the independent system which is given by

$$\mathcal{G}(0) = \sum_{\mathrm{i}=1}^{p} \sum_{k=1}^{q} \mathbf{P}_{\mathrm{i}}(k) \ln \mathbf{P}_{\mathrm{i}}(k) \ . \tag{2.21}$$

The second term is found by differentiating Equation (2.17) evaluated at $\varepsilon = 0$, which gives

$$\left.\frac{\partial G}{\partial \varepsilon}\right|_{\varepsilon=0} = -\sum_{\mathrm{i}<\mathrm{j}} \sum_{k,l} J_{\mathrm{ij}}(k,l) \mathbf{P}_{\mathrm{i}}(k) \mathbf{P}_{\mathrm{j}}(l) \ . \tag{2.22}$$

Differentiating the left and right hand side of Equation (2.20) twice, yields

$$\frac{\partial^2 \mathcal{G}}{\partial \mathbf{P}_{\mathrm{i}}(k) \mathbf{P}_{\mathrm{j}}(l)} = \varepsilon \frac{\partial^2}{\partial \mathbf{P}_{\mathrm{i}}(k) \mathbf{P}_{\mathrm{j}}(l)} \left.\frac{\partial G}{\partial \varepsilon}\right|_{\varepsilon=0} + \mathcal{O}(\varepsilon^2)$$

$$\Rightarrow (C^{-1})_{\mathrm{ij}}(k,l) = -J_{\mathrm{ij}}(k,l)$$

Subsequently, the mean field approximation for the interaction matrix J_{ij} is dependent on the inverse covariance matrix :

$$J_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = -(C^{-1})_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{j}}) \ . \tag{2.23}$$

This approximation of the interaction matrix might be counter-intuitive - as it dampens the effects of the eigenmodes associated with the highest eigenvalues of C. We show later on through our phylogenetic analysis, that dampening of these eigenmodes corresponds to removing phylogenetic noise. This gives an alternative explaination as to the effectiveness of this approach.

Now equipped with the estimation of J_{ij}, the global probability distribution can be approximated

$$\mathbf{P}_{\mathrm{ij}}^{\mathrm{Dir}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = \frac{1}{\mathcal{Z}} \exp\left\{-(C^{-1})_{\mathrm{ij}}(k_{\mathrm{i}}, k_{\mathrm{i}}) + \tilde{h}_{\mathrm{i}}(k_{\mathrm{i}}) + \tilde{h}_{\mathrm{j}}(k_{\mathrm{j}})\right\} \ . \tag{2.24}$$

The parameters $\tilde{h}_{\mathrm{i}}$ and $\tilde{h}_{\mathrm{j}}$ are found by meeting the conditions

$$\sum_{k_{\mathrm{j}}=1}^{q} \mathbf{P}_{\mathrm{ij}}^{\mathrm{Dir}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = f_{\mathrm{i}}(k_{\mathrm{i}}) \ , \quad \sum_{k_{\mathrm{i}}=1}^{q} \mathbf{P}_{\mathrm{ij}}^{\mathrm{Dir}}(k_{\mathrm{i}}, k_{\mathrm{j}}) = f_{\mathrm{j}}(k_{\mathrm{j}}) \ .$$

Marks et al. [36] builds upon this and introduces direct information (DI) which is given by the following

$$(DI)_{\mathrm{ij}} = \sum_{k,l} \mathbf{P}_{\mathrm{ij}}^{\mathrm{Dir}}(k_{\mathrm{i}}, k_{\mathrm{j}}) \ln \left\{ \frac{\mathbf{P}_{\mathrm{ij}}^{\mathrm{Dir}}(k_{\mathrm{i}}, k_{\mathrm{j}})}{f_{\mathrm{i}}(k_{\mathrm{i}}) f_{\mathrm{j}}(k_{\mathrm{j}})} \right\} \ . \tag{2.25}$$

Note that DI is a form of a global model as $\mathbf{P}^{\mathrm{Dir}}$ uses a mean-field approximation to approximate the Boltzmann distribution, i.e. $J = -C^{-1}$, and the Boltzmann distribution is a distribution over the entire amino-acid sequence. To compute the contacts, DI is ranked by their numerical values. These numerical values are processed to achieve the best minimal set of couplings. This process involves filtering out residue pairs which are known to result in high DI score but are not in contact. Due to the nature of the amino-acid chain, residues are likely to co-vary when they are close in sequential position. A manual approach to this problem is to set all the DI scores for pairs separated by less than five residues to zero[36]. There are no correlation signals when residues are completely conserved, therefore when residues are close to being completely conserved the correlation signal comes with a significant amount of uncertainty. A heuristic way to resolve this uncertainty is to ignore the pairs which involve highly conserved residues.

Marks et al.[36] compared the contact predictions using DI and MI, see Figure 2.1. The results show that the contact predictions using MI are substantially different from the contacts in the native state. On the other hand, there is a high level of agreement between contact predictions using DI and the native contacts.

2.3 Overcoming Phylogeny

Protein sequences in the same family share a common ancestor, therefore these sequences are not independent. They are connected by a phylogenetic tree and this tree is a graphical representation of a series of duplication and mutation events which happened between the time of the ancestor and the sequences that we observe.

The restrictions (2.4) and (2.5) put on the maximum entropy formulation to produce the Ising/Potts model correspond also to the maximum-likelihood estimate for independent sequences. This estimate will be affected strongly by the phylogenetic biases for cases where the mutation rate is slow. There have been many methods developed to counteract the effects of phylogenetic bias, however, most adopt heuristic approaches as it is hard to quantify the effects of phylogeny.

Dutheil[16] summarised some of the different methodologies developed to overcome the phylogenetic biases. One approach is to analyse the sequences by ignoring phylogeny at first, then the significance of phylogenetic effects are determined by comparing the observed correlations with stochastic correlations that one would expect if no phylogeny was present. For example, methods presented by Lapedes et al.[33], Larson et al.[34] used a null-model approach, where the phylogeny is integrated into the null-hypothesis and the correlations due to phylogeny are determined by comparing the observed correlations with the expected correlations. MI was used as the correlation measure in these cases. In order to compare the correlations, they generated phylogenetic trees which created sequences that do not have any interactions and thus a threshold for the null-hypothesis is set by finding MI for this set of sequences. Another way to minimise phylogenetic biases is to reweight sequences according to their redundancy. One of the more notable efforts to attenuate phylogenetic bias for mutual information is Average Product Correction (APC)[14]. APC of MI is given by:

$$APC_{ij} = \frac{\overline{MI}_{i,:}\overline{MI}_{j,:}}{\overline{MI}} \; , \tag{2.26}$$

the overline represents the mean of the vector/matrix, and $M_{i,:}$ denotes the sum over the *i*the column but does not include the *i*th element - together this gives the following:

$$\overline{MI}_{i,:} = \frac{1}{p}\sum_{j\neq i}^{p} MI_{i,j}. \tag{2.27}$$

This is an estimation of the background MI induced by the similarity between the sequences. Deducting this from MI has been shown to significantly increase the correlation between MI and the contact map.

These methods lack precision as the treatment for all sequences are the same even when some sequences are more redundant than others. Altschul et al.[2] motivated the idea of reweighting the sequences as a function of their correlation. Given a set of observed data $\mathbf{x} = (x_1, \cdots, x_n)$ taken from a normal distribution with uniform mean μ and covariance matrix Σ. The log-likelihood function is:

$$l(\mu, \Sigma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(\mathbf{x} - \mu 1)^T \Sigma^{-1}(\mathbf{x} - \mu 1) \ ,$$

where we overload the notation $1 = (1, \cdots, 1)$. Therefore the maximum likelihood estimate of μ is given by

$$\widehat{\mu} = \frac{1\Sigma^{-1}\mathbf{x}}{1\Sigma^{-1}1} = \sum_{\mathrm{i}=1}^{n} \left(\frac{1\Sigma^{-1}}{1\Sigma^{-1}1}\right)_{\mathrm{i}} x_{\mathrm{i}} = \sum_{\mathrm{i}=1}^{n} w_{\mathrm{i}} x_{\mathrm{i}} \ .$$

The weights $\mathbf{w} = (w_1, \cdots, w_n)$ can be determined from Σ, however the inference of this parameter is challenging. For this reason, most of the weighting algorithms are based on a heuristic understanding of the phylogeny. One intuitive methodology is to weigh these sequences by how similar they are[62,41], for example, with respect to the Hamming distance. A similarity threshold is then put in place to define whether the sequences are similar or not. Let this threshold be xp, where $0 \leq x \leq 1$ and p is the length of the sequence, the number of sequences similar to the sth sequence is then given by

$$\theta_s = \sum_{\mathrm{i}=1}^{n} \mathcal{H}\left(\sum_{\mathrm{j}=1}^{p} \mathcal{I}(\mathbf{x}_{\mathrm{j}}^{s}, \mathbf{x}_{\mathrm{j}}^{\mathrm{i}}) - xp\right) \ ,$$

where $\mathcal{H}$ is the heaviside function and $\mathcal{I}$ is the indicator function. Various other weighting systems have been devised, such as using pairwise distances between sequences to cluster them into bifurcating trees[24]. This reweighting is accounted for by redefining the frequency counts in Equations (2.2) and(2.3):

$$f_{\mathrm{i}}(k) = \frac{1}{\lambda + M_{\mathrm{eff}}}\left(\frac{\lambda}{q} + \sum_{\mathrm{s}=1}^{n}\frac{1}{\theta_{\mathrm{s}}}\mathcal{I}(x_{\mathrm{i}}^{\mathrm{s}}, k)\right) \tag{2.28}$$

$$f_{\mathrm{ij}}(k, l) = \frac{1}{\lambda + M_{\mathrm{eff}}}\left(\frac{\lambda}{q^2} + \sum_{\mathrm{s}=1}^{n}\frac{1}{\theta_{\mathrm{s}}}\mathcal{I}(x_{\mathrm{i}}^{\mathrm{s}}, k)\mathcal{I}(x_{\mathrm{j}}^{\mathrm{s}}, l)\right) \tag{2.29}$$

where

$$M_{\mathrm{eff}} = \sum_{s=1}^{n}\frac{1}{\theta_s} ,$$

is the effective number of sequences after reweighting. λ, otherwise known as the pseudocount, adds randomisation which allows the covariance matrix C to be inverted.

A more mathematical approach was taken by Obermayer and Levine[44]. They formulated a new probability distribution which accounts for both the interactions and the phylogeny by adding an extra term to the energy in the Boltzmann distribution stated in Equation (2.6). This is given by

$$E(X) = \underbrace{-\sum_{a,\mathrm{i}} g_a x_{\mathrm{i}}^a - \sum_{a<b,\mathrm{i}} K_{ab} x_{\mathrm{i}}^a x_{\mathrm{i}}^b}_{\text{extra term}} - \sum_{a,\mathrm{i}} h_{\mathrm{i}} x_{\mathrm{i}}^a - \sum_{\mathrm{i}<\mathrm{j},a} J_{\mathrm{ij}} x_{\mathrm{i}}^a x_{\mathrm{j}}^a . \tag{2.30}$$

Here K_{ab} account for the interactions between the sequences due to phylogeny. To simplify the problem, the phylogeny is chosen to be a linear chain so that $K_{ab} = K\delta_{a,b-1}$. Further simplification was made by restricting the length of the chain to two, therefore the only unknown parameters are the interactions J_{12} and external fields h_1, h_2. This particular partition function can be resolved using standard transfer matrix methods where a recursive relationship can be found between the partial partition functions, which is given by

$$Z_N(K, J_{12}, h_1, h_2|\mathbf{x}^{N+1}) = \sum_{\mathbf{x}^1,\cdots,\mathbf{x}^N} \exp\left\{\sum_{a=1}^{N} h_1 x_1^a + K x_1^a x_1^{a+1} + h_2 x_2^a + K x_2^a x_2^{a+1} + J_{12} x_1^a x_2^a\right\}$$

$$= \sum_{\mathbf{x}^N} \exp\left\{h_1 x_1^N + h_2 x_2^N + K(x_1^{N+1} x_1^N + x_2^{N+1} x_2^N) + J_{12} x_1^N x_2^N\right\} Z_{N-1}(K, J_{12}, h_1, h_2|\mathbf{x}^N). \quad (2.31)$$

To gain further insight into how the linear phylogeny affects the sequences, X, a constant parameter is chosen $K_{ab} = K_0$. This phylogenetic parameter is estimated from using background data, X^0 . This data is connected by linear phylogeny but independent of observed physical interactions, thereby singling out the effects of phylogeny. They found that the estimate for K_0 is given by

$$\widehat{R} = \tanh(\widehat{K_0}) = \frac{1}{np} \sum_{a,\mathrm{i}} X^0_{a\mathrm{i}} X^0_{(a+1)\mathrm{i}} \,, \quad (2.32)$$

where $\widehat{K_0}$ is the estimate, p and n are respectively the length of the sequence and the number of sequences. The estimates $\widehat{h_i}$ and $\widehat{J_{ij}}$ satisfy equations

$$\tanh(\widehat{h_i}) = \frac{(1-\widehat{R})m_i}{\sqrt{(1+\widehat{R}^2)^2 - 4m_i^2\widehat{R}}} \quad (2.33)$$

$$\tanh(\widehat{J_{ij}}) = \frac{(1-\widehat{R})m_{ij}}{\sqrt{(1+\widehat{R}^2)^2 - 4m_{ij}^2\widehat{R}^2}} \,, \quad (2.34)$$

where

$$m_i = \frac{1}{n}\sum_a X_{ai} \,, \quad m_{ij} = \frac{1}{n}\sum_a X_{ai}X_{aj} \,. \tag{2.35}$$

The leading orders of the expected errors were calculated, which yields

$$\langle(\widehat{h_i} - h_i)^2\rangle = h_i^2 \left(e^{-2(\widehat{K_0}-K_0)} - 1\right)^2 + \frac{1}{n}e^{-2(2\widehat{K_0}-K_0)} \tag{2.36}$$

$$\langle(\widehat{J_{ij}} - J_{ij})^2\rangle = J_{ij}^2 \left(\frac{\cosh(2K_0)}{\cosh(2\widehat{K_0})} - 1\right)^2 + \frac{1}{n}\frac{\cosh(2K_0)}{\cosh^2(2\widehat{K_0})} \,. \tag{2.37}$$

Obermayer and Levine noted that the estimates (2.33) and (2.34) can be independent of the phylogenetic estimate $\widehat{K_0}$ if the moments (2.35) are rescaled :

$$\tilde{m_i} = m_i e^{-2\widehat{K_0}} \,, \quad \tilde{m_{ij}} = \frac{m_{ij}}{\cosh(2\widehat{K_0})} \,. \tag{2.38}$$

The argument is that if the estimates are independent of phylogenetic parameters, the samples can then be retreated independently with these rescaled moments. The results using rescaled moments are then compared with reweighting the moments. By using the mean field approximation, they found that the error decreased dramatically, as K_0 increases, when both rescaling and reweighting were applied, whereas the error sustained when only reweighting was applied.

These results show that this probability distribution (2.30) maybe valid for the case where the phylogeny is a linear chain, but it is still unclear whether this model can account for more convoluted phylogeny. K_0 is also a ficticious parameter to create 'interaction' between the sequences, but the dependence of this on the phylogeny in terms of the number of mutations, duplications and protein lengths is unclear.

It is intuitive that mutations will affect the smaller proteins more strongly than the larger proteins. However, on less intuitive grounds such as how the number of sequences affect the phylogeny, a more mathematical approach is needed. In this thesis, we explicitly investigate how tuning the number of mutations, duplications and protein lengths affect the covariances between samples and residues, furthermore, by using random matrix methods we hope to show how this may affect the detection of interactions.

A living organism must be studied from two distinct aspects. One of these is the causal-analytic aspect which is so fruitfully applicable to ontogeny. The other is the historical descriptive aspect which is unravelling lines of phylogeny with ever-increasing precision. Each of these aspects may make suggestions concerning the possible significance of events seen under the other, but does not explain or translate them into simpler terms.

Sir Gavin De Beer

3

Modelling Phylogeny

Phylogenetic effects are known to significantly affect the ability for us to find relevant signals for contact prediction. In the previous chapter we highlighted some techniques, mostly heuristic, researchers have developed to alleviate the effects of phylogeny. Characterizing the effects of phylogeny has proven to be challenging as the phylogenetic tree is generally unknown. Thus we often would need to estimate both the phylogenetic tree and it's effects on the amino acid sequences in conjunction. In this work, we bypass the need to infer the phylogenetic tree by generating sequences for which the ground truth information is known; both with respect to the contacts and the phylogeny. In this chapter, we will explain in detail the synthetic-model we will use to ground all of our analysis.

3.1 Phylogenetic Ising Sequences

Consider that we are given an initial amino-acid (Ising) sequence, $x^0 = (x_1^0, \cdots, x_p^0)$, which is the beginning of the evolutionary history. Now this sequence has a certain amount of energy (Gibbs energy) which is given by

$$E(x) = h^T x + \frac{1}{2} x^T J x \tag{3.1}$$

where h and J are known. This sequence will go through a series of mutation events, M, and branching events B which are predetermined.

We mimic the evolutionary mutation event with the Metropolis-Hastings algorithm which is one form of Markov Chain Monte Carlo (MCMC)[21,42]. Here we choose a random position along the sequence i to be mutated. The residue along this position will be mutated to another amino acid with the following probability

$$P(x \to \tilde{x}) = \min(1, \exp(E(x) - E(\tilde{x}))/T), \tag{3.2}$$

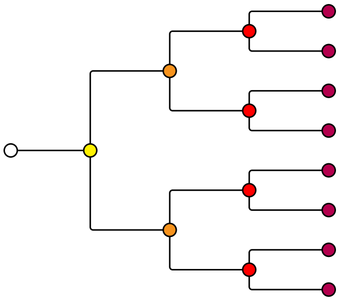


Figure 3.1: This shows a tree with three branching events and same of number of mutation events before branching.

where x is the sequence before the mutation and $\tilde{x}$ is the sequence if the mutation was to be accepted. T is the temperature of the system; the lower the temperature is, the less likely an unfavourable mutation will be accepted. Whether the mutation is accepted or not this is known as one mutation event.

3.2 Algorithm

We outline our algorithm for simulating phylogenetic Ising Sequences in Listing 1. It is worth noting that, our algorithm makes the assumption that the branching factor is always two.

3.3 High Mutation vs Low Mutation Rates

The algorithm outlined above will take in n_0 initial sequences and output $n_0 2^{|B|}$ sequences. The probability distribution of the resulting sequences will be dependent on the number of mutation events performed before each branching event. For example, if the mutation rate is high then the sequences will be less correlated with each other across the branches; while low mutation rate corresponds to sequences being almost identical to each other across the branches. We note that, proteins

Algorithm 1 Simulating Phylogenetic Ising Sequences

```
procedure PhylogeneticIsing
Require: Initial sequences $x$. Mutations events $M$, branching events $B$ and $h, J, T$.
    for $b$ in $B$ do
        for $m$ in $M_b$ do
            $\tilde{x} = x$
            Random-uniformly pick $i \in \{1, \cdots p\}, a \in \{1, \cdots, q\} \setminus x_i$
            $\tilde{x}_i = a$
            if $E(x) > E(\tilde{x})$ then
                $x \leftarrow \tilde{x}$
            else
                if $r \leq \exp(E(x) - E(\tilde{x})/T)$ then
                    $x \leftarrow \tilde{x}$
        $x = [x; x]$
```

within the same homolog are closer to the low mutation regime.

3.4 Temperature T

Another important factor that affects the output distribution of the algorithm is the temperature T. The higher the temperature is the lower the effects of the parameters h and J, thus the signal from the interactions between elements within the sequence increases when we decrease the temperature.

3.5 Boltzmann Sequences vs Protein Sequences

We have chosen to use Boltzmann distribution to generate our sequences with energy given in Eq. 3.1 as this is also a probability model which can be used to infer protein contacts[19,18]. This energy will be high when the probability of the given state **x** is low and vice versa. This is also akin to an energy in a physical system or fitness function of in a population[31]. Temperature is a parameter which can be controlled during the simulation that allows us to control for the diversity of the sequences generated (an equivalence has been made between temperature and population size before[31]). We anneal the temperature to ensure that local optimal states are avoided but we tune the temperature to ensure a level of redundancy between the sequences which is similar to the ones we observe in proteins.

4

Correlation Induced by Phylogeny

To single out the effects of phylogeny, we simulate sequences using Algorithm 1 with h and J set to zero. In other words, there are no interactions between the amino acid residues. To further simplify the model, the number of mutation events along each branch is constant and denoted by m.

Suppose sequences $\mathbf{x}^1$ and $\mathbf{x}^2$ both of length p are generated using the tree shown in Fig. 4.1. The covariance between the two sequences must be dependent on both m and p, i.e. $(\Sigma_S)_{12} = \alpha(m, p)$.

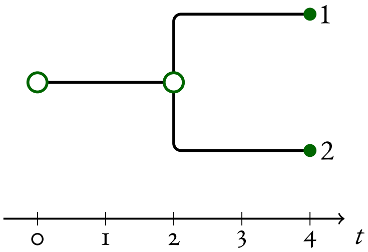


Figure 4.1: This is a graphical representation of the simple phylogeny, where the nodes represent sequences and the branches represent mutation events. In a duplication event, a branch divides at a node.

In fact the functional form of α can be found exactly and is given below

$$\alpha(m,p) = \exp(-\omega m/p) \ . \tag{4.1}$$

where ω is a constant. How this is derived is given below.

4.1 Covariance Between Related Sequences

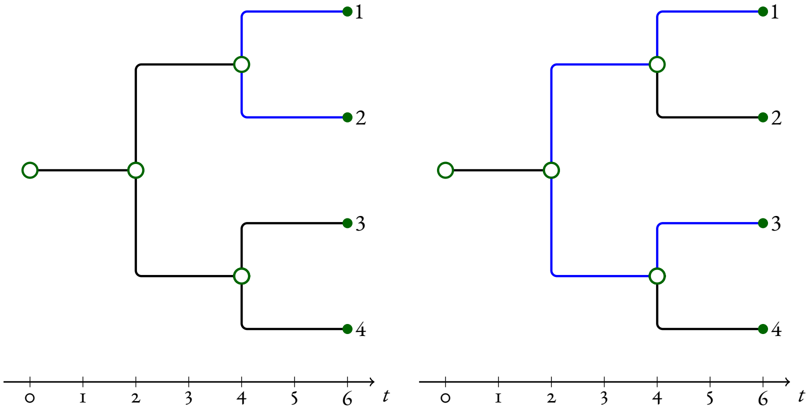


Figure 4.2: The blue lines highlight the mutation (coalescence) distance, D_{ij}, which separates the sequences $\mathbf{x}^i$ and $\mathbf{x}^j$ - this is the only factor governing the covariance between the sequences. The examples shown are the distance between sequences $\mathbf{x}^1$ and $\mathbf{x}^2$ on the left and $\mathbf{x}^1$ and $\mathbf{x}^3$ on the right.

To examine the expected covariance generated by a phylogenetic tree, we consider the expected covariance between two nodes separated by $2m$ mutation events. We use a substitution model with the assumption that each amino acid can mutate to any other with equal probability. This is reminiscent of the Jukes-Cantor model with discrete time[30], i.e. discrete mutations as a measure of time. Here, we will give a more detailed mathematical derivation of the expected covariance.

The derivation takes two steps. Firstly, we will show that this is an Ornstein-Uhlenbeck[22] process, thus the covariance/autocorrelation is proportional to an exponential where the exponent is dependent on a relaxation rate. The second step is to find the relaxation rate.

OU PROCESS : The Ornstein-Uhlenbeck process is otherwise known as stationary Gaussian Markov process, which means that the Markov chain is stationary and satisfies the Gaussian condition, see Eq. (4.2). To see that the phylogenetic process is stationary we note that when there are no preferences for mutation sites, the stationary state of the Markov chain is the uniform distribution. More concretely, consider an Ising sequence with two states -1, 1 evolving through the simple phylogeny shown in Fig 4.1. The probability distribution of the initial sequence, x^0, is uniform since it is randomly generated and after one mutation event the distribution can be found using the following

$$P(x^1 = 1) = P(x^1 = 1|x^0 = -1)P(x^0 = -1) + P(x^1 = 1|x^0 = 1)P(x^0 = 1) ,$$

which gives

$$P(x_{\mathrm{i}}^{1}=1)=\frac{1}{p}\times\frac{1}{2}+\frac{p-1}{p}\times\frac{1}{2}=\frac{1}{2}$$
$$P(x_{\mathrm{i}}^{1}=-1)=1-P(x_{\mathrm{i}}^{0}=1)=\frac{1}{2}\,.$$

Thus by induction, this shows the probability distribution stays uniform throughout the phylogeny process, thereby making the process stationary.

The Gaussian condition[26] is as follows

$$\mathbf{E}(x(t+1)|x(t),\cdots,x(0))=\mathbf{E}(x(t+1)|x(t))\,. \tag{4.2}$$

To see that the phylogenetic process is Gaussian, we note that the state at $t+1$ can be written as $x(t+1)=x(t)+v(t+1)$ where $v(t)$ is the change induced by the mutation; which can be viewed as a discrete velocity. Crucially, the expectation of $v(t+1)$ is only dependent on the state $x(t)$. This can be seen from the Ising model with states -1 and 1. The probability distribution of $v(t+1)$ is given by

$$P(v(t+1)=-2x(t))=\frac{1}{p} \tag{4.3}$$
$$P(v(t+1)=0)=\frac{p-1}{p}\,. \tag{4.4}$$

Therefore the instantaneous mean is

$$\mathbf{E}(v(t+1)|x(t)) = -\frac{2}{p}x(t)\ . \tag{4.5}$$

This implies that $x(t+1)$ is solely dependent on $x(t)$, consequently Eq. (4.2) is satisfied.

Relaxation Rate : The relaxation rate r is defined by this condition

$$\mathbf{E}(x(t+1)|x(t)) = (1-r)x(t).$$

Another property of an OU process is that the covariance between two states at time 0 and t, $x(t)$ and $x(0)$, is given by $\alpha(t) = \alpha(0)\exp(-rt)$. There are multiple ways of deriving r. We choose to map the q states onto the unit circle of a complex plane as this does not place any bias towards any states, i.e. the magnitude of the states are the same. The mapping is explicitly given by

$$F\colon \{1,\cdots,q\} \mapsto \{1, \exp(i\theta), \cdots, \exp(i(q-1)\theta)\}\ , \tag{4.6}$$

where $\theta = 2\pi/q$. The Gaussian condition, Eq. (4.2), can be rewritten as

$$\begin{aligned} \mathbf{E}(x_j(t+1)|x_j(t)) &= \mathbf{E}(d_j(t+1)x_j(t)|x_j(t)) \\ &= \underbrace{\mathbf{E}(d_j(t+1)|x_j(t))}_{(A)}\,x_j(t)\ , \end{aligned} \tag{4.7}$$

where $d_j(t) = x_j(t+1)/x_j(t)$ and $d_j(t)$ and $x_j(t)$ are the jth elements of $d(t)$ and $x(t)$ respectively. If we use mapping (4.6) then (A) in Eq. (4.7) becomes

$$\mathbf{E}(d_j(t+1)|x_j(t)) = \frac{p-1}{p} + \frac{1}{p(q-1)} \sum_{k=1}^{q-1} \exp(ik\theta)$$
$$= 1 - \frac{1}{p}\frac{q}{q-1} .$$

Substituting this back into Eq. (4.7) yields

$$\mathbf{E}(\mathbf{x}(t+1)|\mathbf{x}(t)) = \left(1 - \frac{1}{p}\frac{q}{q-1}\right) \mathbf{x}(t) .$$

Using the definition of relaxation rate, r, the relaxation rate is given by $r = q/(p(q-1))$. As a result, the covariance is

$$\alpha(t) = \alpha(0) \exp\left(-\frac{q}{q-1}\frac{t}{p}\right) . \tag{4.8}$$

The mapping to a complex circle means $\alpha(0) = \text{var}(x) = 1$. This can be seen if we consider $\mathbf{E}(x)$ and $\mathbf{E}(x^T x)$ separately. We note that for an uniform distribution, $\mathbf{E}(x)$ is proportional to the sum of the states on a complex circle, which is 0. Similarly, we note that since the magnitude of the nodes on a unit circle is one, we have $\mathbf{E}(x^T x) = \text{var}(x) = 1$. Consequently, for two nodes separated by $t = 2m$ mutations, the covariance is given by

$$\alpha = \exp(-2mq/p(q-1)) \tag{4.9}$$

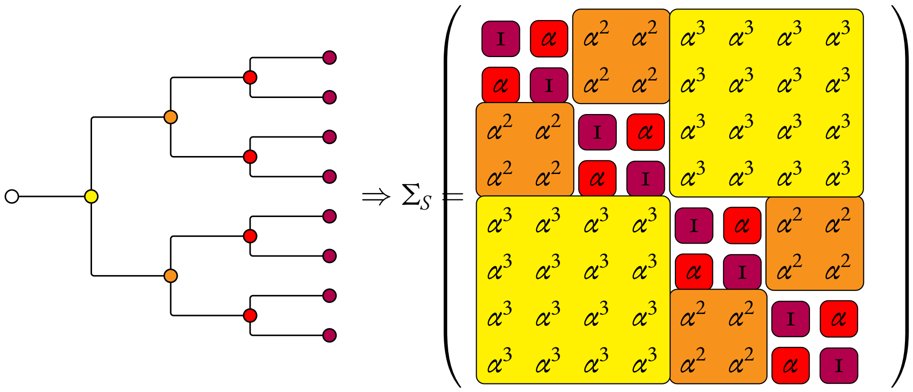

$\Rightarrow \Sigma_S =$
$1 \quad \alpha \quad \alpha^2 \quad \alpha^2 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3$
$\alpha \quad 1 \quad \alpha^2 \quad \alpha^2 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3$
$\alpha^2 \quad \alpha^2 \quad 1 \quad \alpha \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3$
$\alpha^2 \quad \alpha^2 \quad \alpha \quad 1 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3$
$\alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad 1 \quad \alpha \quad \alpha^2 \quad \alpha^2$
$\alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha \quad 1 \quad \alpha^2 \quad \alpha^2$
$\alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^2 \quad \alpha^2 \quad 1 \quad \alpha$
$\alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^3 \quad \alpha^2 \quad \alpha^2 \quad \alpha \quad 1$

Figure 4.3: We visualise the nested structure induced in the covariance matrix between the sequences from a binary tree. The color in $(\Sigma_S)_{ij}$ corresponds to the color of the node where the two sequences coalesce.

4.1.1 Nested Structure of the Covariance Matrix

To see the nested structure of the covariance matrix induced by a phylogenetic tree, as shown in Fig. 4.3, we consider the homogeneous tree with b branching events and m mutations per branch. Each pair of sequences are separated by $2\tilde{b}m$ mutations where $\tilde{b}$ is the number branching events since their most recent common ancestor. Hence, using the covariance to distance relation shown in Eq. (4.9), the covariance matrix Σ_S between a set of sequences generated by a homogeneous tree is given by

$$\Sigma_S = \exp\left(-\frac{q}{p(q-1)}D\right) \tag{4.10}$$

where

$$D = \begin{pmatrix} 0 & 2m & \cdots & 2bm & \cdots & 2bm \\ 2m & 0 & & \vdots & \ddots & \vdots \\ \vdots & & \ddots & 2bm & \cdots & 2bm \\ 2bm & \cdots & 2bm & \ddots & & \vdots \\ \vdots & \ddots & \vdots & & 0 & 2m \\ 2bm & \cdots & 2bm & \cdots & 2m & 0 \end{pmatrix}, \tag{4.11}$$

is the distance matrix, see Fig 4.2. The monotonicity of the exponential function means that the nested structure of D is reflected in Σ_S. This is explicitly demonstrated in Fig. 4.3.

4.2 Eigenvalues of the Covariance Matrix

The eigenvalues of Σ_S in Eq. (4.10) can be found analytically. This set of eigenvalues has a few distinct features, firstly, there are $b+1$ distinct eigenvalues for sequences generated with b branching events; secondly, the degeneracy of the eigenvalues increases as its magnitude decreases. The explicit mathematical formula is given by the following

$$\lambda_i = \begin{cases} 1 + \sum_{j=1}^{b} 2^{j-1}\alpha^{j} & i = 0 \\ (1-\alpha)\left(\sum_{j=0}^{i-1}(2\alpha)^{b-i}\right) & i > 0 \end{cases}, \tag{4.12}$$

where $\lambda_0 > \cdots > \lambda_b$ ($\alpha \neq 0$). We can view the degeneracy of the eigenvalues as proportional to the probability of drawing a particular eigenvalue, this probability distribution is given by

$$p_i = \begin{cases} 1/n & i = 0 \\ 2^{i-1}/n & i > 0 \end{cases}, \tag{4.13}$$

where $p_i = \mathbb{P}(\lambda = \lambda_i)$ and $n = 2^b$. The exponential decay in degeneracy as the eigenvalue gets larger indicates that the eigenvalues of its empirical covariance matrix will have a heavy tailed distribution.

A final proof of our ideas can only be obtained by detailed studies on the alterations produced in the amino acid sequence of a protein by mutations of the type discussed here.

Francis Crick

5

Protein Interactions and Covariance Analysis

So far, we have made the simplification that the amino-acid sequences generated are phylogenetically dependent only and there are no interactions within the sequence. Here, we extend this analysis to sequences with interactions, i.e. where elements in J are non-zero. In this thesis, we refer to these pairwise interactions as 'phenotypic' interactions, alluding to the importance of these interactions to the structure of the protein.

In the second half of this chapter, we will go through, in detail, the framework used to unravelling relevant signals induced by J in the presence of phylogeny.

5.1 Protein Interactions

To simulate the phenotypic interactions in a protein sequence with p residues chosen from q characters ($q = 21$ for amino acids including gaps), we adopted a generalised Potts model[64]. This model in its biological applications has been explored extensively[3,40,60].

For a generalized Potts model, we need to specify a positional interaction matrix $\tilde{J}$ of size $p \times p$ and an intrinsic interaction matrix between the q types of amino acids which is Θ of size $q \times q$. In protein sequences, Θ can capture the interaction between amino acids, for example two amino acids with similar charges will repel each other and the corresponding interaction within Θ will be positive, which means a positive increase in energy. The associated energy function of the generalized Potts model is given by

$$E(x) = -\sum_{a,b=1}^{q} \sum_{i<j} \tilde{J}_{ij} \Theta_{ab}\, \delta(x_i, a) \delta(x_j, b). \tag{5.1}$$

For the matrix Θ, we use the planar Potts model[64] which is a specific form of the generalized Potts model. This model extends the binary spin states in an Ising model to q spin states as follows

$$\Theta_{a,b} = \cos(2\pi(a-b)/q)$$

where $a, b \in \{1, \cdots q\}$. To derive the covariance induced by the pairwise interaction matrix $\tilde{J}$ we can express the amino acids with the same mapping as shown in Eq. (4.6). We find that

$$\begin{aligned} E(x_i x_j) - E(x_i)E(x_j) = E(x_i x_j) &= \sum_{a,b=1}^{q} e^{i2\pi(a-b)/q} \frac{e^{\tilde{J}_{ij}\Theta_{ab}}}{\sum_{a,b} e^{\tilde{J}_{ij}\Theta_{ab}}} \\ &= \frac{1}{\mathcal{Z}_{ab}} \frac{\partial \mathcal{Z}_{ab}}{\partial \tilde{J}_{ij}} \, , \end{aligned} \tag{5.2}$$

where $\mathcal{Z}_{ab} = \sum_{a,b=1}^{q} e^{\tilde{J}_{ij}\Theta_{ab}}$. To see that this is true we simply note that the left-hand side of the equation is real with the imaginary part set to zero. An important property of Eq. (5.2) is that as $\tilde{J}_{ij} \to \pm\infty, E(x_i x_j) \to \pm 1$, similarly as $\tilde{J}_{ij} \to 0, E(x_i x_j) \to 0$. In other words, the covariance function saturates and cannot exceed more than the magnitude one as the strength of the interactions increases. For example, if we consider the binary state case then Eq. (5.2) becomes

$$\begin{aligned} E(x_i x_j) &= \frac{1}{e^{\tilde{J}_{ij}} + e^{-\tilde{J}_{ij}}} \frac{\partial}{\partial \tilde{J}_{ij}} \left(e^{\tilde{J}_{ij}} + e^{-\tilde{J}_{ij}} \right) \\ &= \tanh(\tilde{J}_{ij}) \, , \end{aligned} \tag{5.3}$$

which is a function that saturates at large values of $\tilde{J}_{ij}$.

We can relate this derivation to commonly used covariance analysis from the literature for protein structure prediction [18,25,13,57,19,50] as follows. For protein structure prediction, it is common to map the sequences to a one-hot format. This is where we map the first state to the basis vector $(1, 0, \cdots, 0)$, and second state on to $(0, 1, 0, \cdots, 0)$ and so on. The mapping is given by:

$$\mathcal{X}\colon \{1, \cdots, q\} \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_q\} \, , \tag{5.4}$$

where $\mathbf{e}_i$ is the ith basis vector. Then, analogous to the Eq. (5.3) above, the expected covariance between the ith and jth position can be expressed as $E(x_i x_j^T) - E(x_i)E(x_j)^T$, where

$$E(x_i x_j^T) = \sum_{a,b=1}^{q} \mathbf{e}_a \mathbf{e}_b^T \frac{e^{\tilde{J}_{ij}\Theta_{ab}}}{\sum_{a,b} e^{\tilde{J}_{ij}\Theta_{ab}}} \;\Rightarrow\; E(x_i x_j^T)_{ab} = P(x_i = a, x_j = b) \tag{5.5}$$

$$E(x_i) = \sum_{a=1}^{q} \mathbf{e}_a \frac{e^{\tilde{J}_{ij}\Theta_{ab}}}{\sum_{a,b} e^{\tilde{J}_{ij}\Theta_{ab}}} \;\Rightarrow\; E(x_i)_a = P(x_i = a) \, , \tag{5.6}$$

putting the above together, yields

$$\mathrm{cov}(x_i, x_j)_{ab} = P(x_i = a, x_j = b) - P(x_i = a)P(x_j = b) \, . \tag{5.7}$$

Phenotype vs Phylogeny: Here, we note that the phenotypic interaction matrix J only affects the covariance of the residues which are involved, shown in Fig. 5.1. In contrast, phylogeny affects every single element in the covariance matrix. An example of this effect is shown in Fig. 5.1. As we will show later, this has significant impact on how the interaction signals should be extracted.

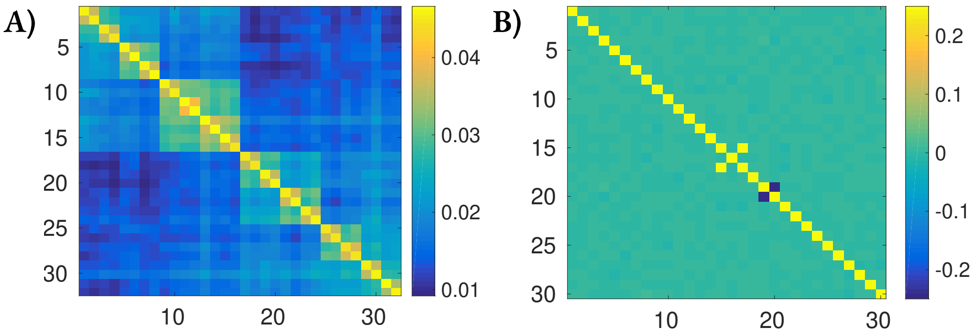


Figure 5.1: Empirical covariance matrices are shown. A) sequences of length 100 generated after 5 branching events and mutation rate 0.1. B) 4096 independent sequences of length 30 are generated with two disjoint interactions $J_{17,15} = -3.12$ and $J_{19,20} = 3.30$.

5.2 Covariance Analysis

Our goal is to find properties which distinguish phylogenetic and phenotypic interaction signals in a covariance matrix which contains a superposition of both. Towards this end, we develop a method that approaches this problem by examining characteristics in the covariance matrix from phylogenetic or phenotypic sources separately. To approximate the covariance matrix, in practice we use the empirical covariance matrix throughout our analysis; as defined below.

Definition 5.2.1. *For a matrix X of size $n \times pq$ where n is the number of samples and pq is the dimension of each sample then the phylogeny covariance matrix of X is $C_S = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})^T$, which is an $n \times n$ matrix, where $\overline{X}_{\mathrm{ij}} = \frac{1}{p}\sum_{\mathrm{j}=1}^{p} x_{\mathrm{j}}^{\mathrm{i}}$.*

Definition 5.2.2. *For a matrix X of size $n \times pq$ where n is the number of samples and pq is the dimension of each sample (p is the size of the amino acid sequence and q is the dimension of the one-hot vector) then the phenotypic covariance matrix of X is $C = \frac{1}{p-1}(X - \overline{X})^T(X - \overline{X})$, is an $pq \times pq$ matrix, where $\overline{X}_{\mathrm{ij}} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} x_{\mathrm{j}}^{\mathrm{i}}$.*

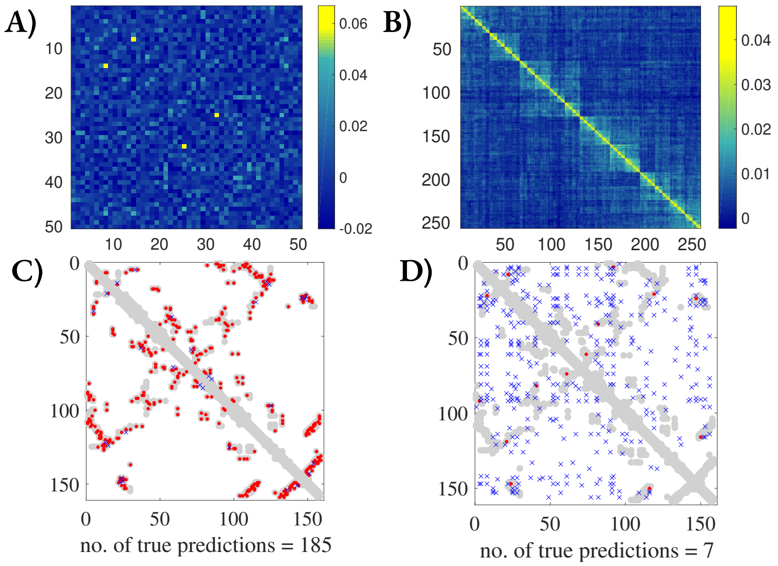


Figure 5.2: This plot demonstrates the detrimental effects of phylogeny when using the empirical covariance matrix C to infer contacts. A) vs B) are heatmaps of C without and with phylogenetic dependencies respectively. C) and D) shows when we use the top 200 correlations in C shown in A) and B) respectively to infer contacts. We can see that by adding the phylogenetic process the precision reduces from 185/200 to 7/200 which is signficant.

When sequences are generated with a phylogenetic tree, their corresponding covariance matrix exhibits a nested square structure (Fig. 5.1A). In contrast interactions between amino-acids appear to generate a spiky signal (Fig. 5.1B), at (i, j) position, if the positional residues i and j are interacting with each other. This entails that phylogeny imposes a global signature on the covariance matrix while the phenotype signature on the covariance matrix is local.

5.2.1 Significance of Phylogeny

Fig. 5.2 shows the ability to infer J from the empirical covariance matrix when the sequences are generated via a phylogenetic tree compared to independently generated. In Fig. 5.2C we see that, indeed, the empirical covariance matrix can infer 185/200 non-zero interactions for indepenent sequences, while when phylogeny is involved this drops down to 7/200 (Fig. 5.2D).

From this we ask the following key question: is it possible to disentangle phylogenetic correlation from the correlations induced by J in the empirical covariance matrix? Moreover, would it be possible to improve the inference of J if we removed the phylogenetic signals.

5.3 Eigenvalue Distribution

Most of the covariance analysis will be done with respect to its decomposition in eigenspace. The advantage to this method is that complicated features can be easily extracted through eigenvectors[28]. The eigen decomposition of the covariance is given by

$$C = \lambda_n U_n U_n^T + \cdots + \lambda_1 U_1 U_1^T\,, \tag{5.8}$$

where U_i denote the eigenvectors and λ_i the corresponding eigenvalues. Note if $\lambda_n \gg \lambda_i$ for all $i \neq n$ then the prominent features of Σ_S can be captured by a rank-one matrix given by $\lambda_n U_n U_n^T$ (also known as principal componenet analysis).

Definition 5.3.1. *Let C be a matrix of size $p \times p$ with eigenvalues $\lambda_1, \cdots, \lambda_p$. The empirical eigenvalue distribution function $P(\lambda)$ is given as*

$$P(\lambda) = \frac{1}{p}\sum_{i=1}^{p} \delta(\lambda - \lambda_i).$$

Here we show that the phenotypic covariance matrix, C shown in Definition 5.2.2, has almost the same eigenvalue distribution as phylogenetic covariance matrix (Definition 5.2.1).

To see this, first note that matrices $C_1 = X^T X$ and $C_2 = XX^T$ share the same eigenvalue distribution, with the exception of a degenerate eigenvalue at zero since both matrices have the same rank but are of different dimension.

We can thus use a singular value decomposition on X

$$X = U \Lambda V \,, \tag{5.9}$$

where U is an $n \times n$ unitary matrix, V is a $p \times p$ unitary matrix and Λ is a diagonal matrix of size $n \times p$. Next, we substitute this into C_1 and C_2 resulting in

$$\begin{aligned} C_1 &= U \Lambda \Lambda^T U^T \,, \\ C_2 &= V^T \Lambda^T \Lambda V \,. \end{aligned} \tag{5.10}$$

The eigenvalues of C_1 and C_2 are the diagonal elements of $\Lambda\Lambda^T$ and $\Lambda^T\Lambda$ respectively. The matrix of the larger dimension between C_1 and C_2 will have a degenerate eigenvalue at zero. This is simply a reflection on the non-zero dimensionality of the null space and is not essential for the detection of the correlation structure of either phylogeny or phenotype. The rest of the eigenvalues of C_1 and C_2 are the same. Similarly, both the phylogenetic and the phenotypic empirical covariance matrix, C_S and C, will have approximately the same eigenvalues distribution after scaling. From empirical observations, the removal of modes, $\overline{X}^T\overline{X}$ and $\overline{XX}^T$, only affect the largest eigenvalue but the rest of the eigenvalues remains the same.

In this chapter and above, we have shown that the empirical covariances matrices for both phylogeny and phenotype contains equivalent eigenmodes as well as non-zero eigenvalues. Crucially, we have found that the phylogenetic process and phenotypic process induces two very different kind of signals in its corresponding true covariance matrix. Phylogenetic dependence changes the covariance matrix globally (i.e. nested squares see Fig. 4.3), while phenotypic covariance induces only local spikes, i.e. we see that interactions between the *i*th and *j*th amino acid pair will only induce non-zero covariance in the *ij*th element of the matrix see Eqs 5.3 and 5.7. These two different ”frequency” of signals should be decomposable similar to Fourier decomposition where phylogenetic dependencies corresponds to signicantly lower frequency signal than that of point contacts induced by phenotype.

As a result, we hope to find a way to remove phylogenetic noise by looking at the eigendecomposition of the empirical covariance matrices. Towards this end, we turn towards random matrix theory to shed light into how to decompose these two different signatures.

We could not, for example, arrive at a principle like that of entropy without introducing some additional principle, such as randomness, to this topography.

Michael Polanyi

6

Eigenvalue Analysis using Random Matrix Theory

When the data is sampled independently, the empirical covariance matrix, $C \sim X^T X$, is a good way to deduce the significant correlations that come from phenotypic interactions. However, for protein families – the sequences are not independent. They are related via phylogeny and, as shown in the previous chapter, when the sequences are related via phylogeny, this correlation measure is no longer effective at all.

We have also seen from the previous chapter that the covariance induced by phylogeny and the one induced by the phenotype are strikingly different in nature. Namely, the phenotypic interactions induce sharp spiky signals in the covariance matrix, whereas phylogeny induces a global nested structure. Here, we examine the consequential effects these different signatures have on the empirical covariance measure. In particular, we would like to know the effect on its corresponding eigenvalue distribution.

In this chapter, we turn to random matrix theory (RMT) to show that, in the regime where we have a large number of sequences in the multiple sequence alignment, we can predict to high precision the eigenvalue distribution of the empirical covariance matrix induced by phylogeny. In order to do this, we develop a novel extension to a known theorem by Marčenko & Pastur [37].

With the ability to predict the eigenvalue distribution induced by phylogeny, we hope to shed light on why the covariance matrix is no longer a good indicator of phenotype in its presence.

RMT has been applied to a wide range of areas such as quantum, population genetics and finance to name a few[63,46,32,6,7,8,9]; we will limit the exposure in the following to our area of investigation but refer to books by Edelman & Rao [17], Tao [58], for in-depth guides.

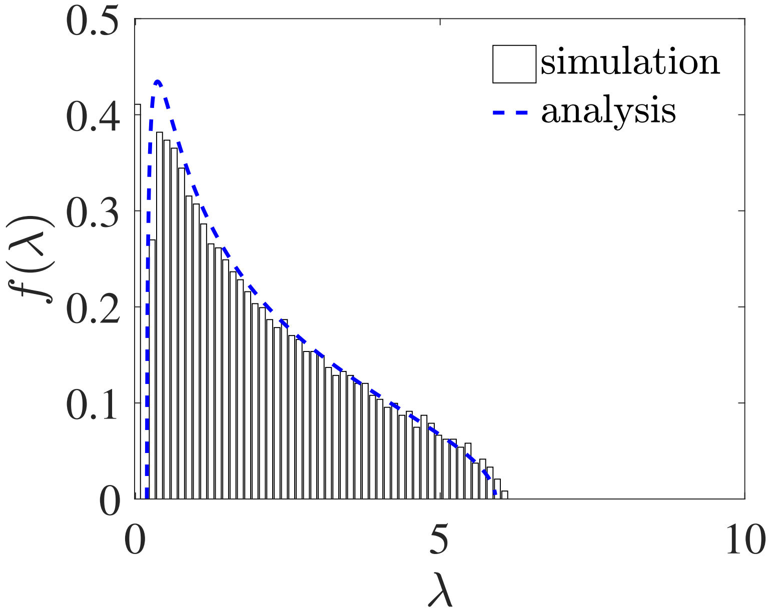


Figure 6.1: 4096 random sequences of length 100 are generated. Blue line is the MP distribution.

6.1 Marčenko-Pastur (MP) Distribution

For independent and identically distributed (iid) sequences with no phenotypic constraints, the empirical covariance matrix, $X^T X/n$, has some well known properties. One of which, is its eigenvalue distribution. This distribution was discovered by Marčenko & Pastur[37] in 1967 and has the following form:

$$f(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi\lambda c}, \tag{6.1}$$

where $a = \sqrt{1-c^2}$, $b = \sqrt{1+c^2}$ and $c = p/n$. Fig. 6.1 gives an example of the MP distribution. An important property about this distribution is that it is not dependent on the base distribution used to generate the elements in X and more importantly it is satisfied in the limit where $p, n \to \infty$.

We can view this as the theorem in RMT which is equivalent to Central Limit Theorem where we can find the exact distribution of the sample mean in the limit of infinite samples, i.e. $n \to \infty$. Note this distribution is compactly supported, which means it is interval bounded by $[a, b]$. Thus this distribution is an easy way to detect whether the data we have is completely random.

This compactness also allows us to distinguish the noisy modes (eigenvectors) from the important ones. For example, if we have the following interaction matrix

$$J = \begin{pmatrix} 0 & \varepsilon & 0 \\ \varepsilon & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} , \tag{6.2}$$

which is a rank one matrix. Its empirical covariance matrix $C = X^T X/n$ should tend towards a rank one matrix as we increase the sample size. A common way to make sure that the estimate we have is rank one is to use dimension reduction methods. A common dimension reduction method is principal component analysis[28] (PCA) which captures the matrix in lower dimensions by reducing the following loss function

$$L(C, \tilde{C}_d) = \sum_{i,j} |C_{ij} - \tilde{C}_{dij}|^2 , \tag{6.3}$$

where d is the dimension of the estimate. By expressing C in its eigendecomposition

$$C = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \cdots + \lambda_p \mathbf{v}_p \mathbf{v}_p^T ,$$

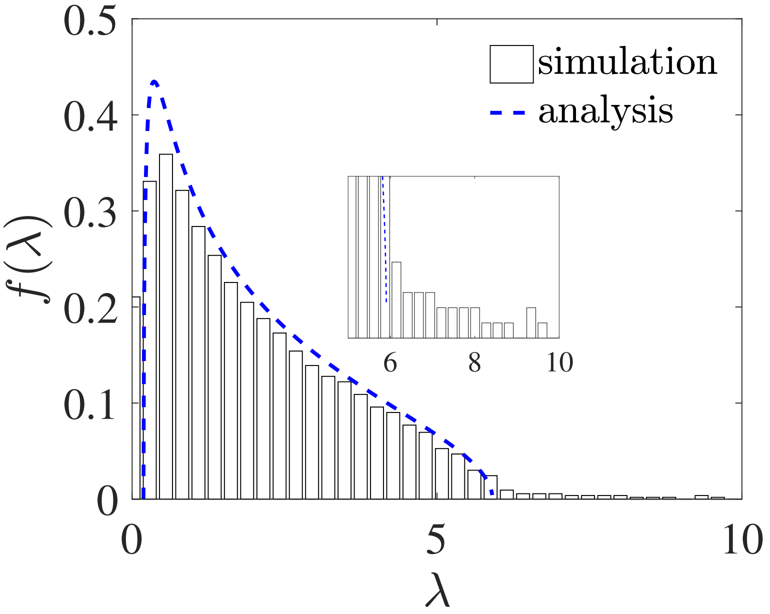


Figure 6.2: 4096 independent sequence of length 100 are generated with 100 phenotypic interactions, the non-zeros values of J are between -5 and 5. The blue line is the MP distribution. Note that many eigenvalues fall outside of the compat support of MP distribution, as we would expect from the interaction structure in J.

where $\lambda_1 \geq \cdots \geq \lambda_p$ and p is the dimension of the matrix, then the d-dimension PCA estimate is given by

$$\tilde{C}_d = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \cdots + \lambda_d \mathbf{v}_d \mathbf{v}_d^T \,. \tag{6.4}$$

For the problem where J has the form of Eq. (6.2), we know that $d = 1$ will give us a good estimate of C. However, if the rank of the interaction matrix, J, is not known beforehand then choosing d is not as straightforward. One method is to plot the loss function shown in Eq. (6.3) against d. The optimal d can then be determined as the lowerest dimension before the loss function saturates. However, this is not an elegant method as it gives us no indication of the dimension of the underlying physical system.

A more mathematically satisfying way of choosing d is by utilising the MP distribution shown in Eq. (6.1). In particular, the eigenvectors associated with the eigenvalues outside the compact support of the MP distribution must capture the relevant signals (i.e. the non-zero interactions in J). An example is shown in Fig. 6.2, here we see that the eigenvalues mostly satisfy the MP distribution, but a few lies outside of the compact support. This indicates that the underlying physical systems used to generate the sequences are not completely random.

Many fields have successfully adopted this idea to clean the covariance matrices for their physical system such as in finance[6,5,32,4], which made this an area of increasing interest.

6.2 Extension to Marčenko-Pastur

The compactness of the MP-distribution gave rise to an elegant way of distinguishing noise from relevant signals. However, when the true covariance matrix is not of low rank, as is the case for the covariance matrix induced by phylogeny (see Fig. 5.1), then the MP distribution no longer holds.

In this work, we extended the MP distribution to sequences generated via a phylogenetic tree.

For sequences with no phenotypic constraints simulated along a homogeneous tree, the covariance matrix, Σ_S, is given by Eq. (4.10). Correspondingly, its eigenvalue distribution is given by Eqs. (4.12) and (4.13). Marčenko and Pastur[37] formulated a connection between the expected eigenvalue distributions and the empirical eigenvalues of C_S. We extend Marčenko and Pastur's derivation for independent samples to the case of samples which are dependent via a tree structure. Surprisingly, the parameters of the phylogenetic tree, i.e. the number of mutations per branch and number of branching events, control the empirical eigenvalue distribution through a polynomial which we can analytically find. Edelman et al[49] coined the term ‘algebraic random matrices’ for such cases where the eigenvalue distribution is encoded in a polynomial. Additionally, we use the properties of these polynomials to find the eigenvalue distribution of C_S.

The connection which Marčenko and Pastur derived between the eigenvalue distribution of Σ_S, denoted as $T(\lambda)$, and of C_S, denoted by $f(\lambda)$, is via its Stieltjes Transform, $G(z)$. The Stieltjes transform of $f(\lambda)$ is given by

$$G(z) = \int_{-\infty}^{\infty} \frac{\mathrm{d}F(\lambda)}{\lambda - z} \,, \tag{6.5}$$

where $\mathrm{d}F(\lambda) = f(\lambda)\mathrm{d}\lambda$. The inversion formula is

$$f(\lambda) = \lim_{y \to 0} \frac{1}{\pi} \Im\{G(\lambda + iy)\} \,. \tag{6.6}$$

Marčenko and Pastur[37] found that $G(z)$ satisfies the following differential equation

$$\frac{-1}{G(z)} = z - c \int_{-\infty}^{\infty} \frac{\lambda \mathrm{d}T(\lambda)}{1 + \lambda G(z)} \,, \tag{6.7}$$

where $c = n/p$ when X is a matrix of size $n \times p$. This estabilishes a connection between $T(\lambda)$ to $f(\lambda)$ via $G(z)$. To apply Eq. (6.7) we simply use the expressions for the eigenvalues λ_i and their corresponding probabilities p_i from Eqs. (4.12) and (4.13), namely $\mathrm{d}T(\lambda) = \sum_{i=1}^{b+1} p_i \delta(\lambda - \lambda_i) d\lambda$. Thus

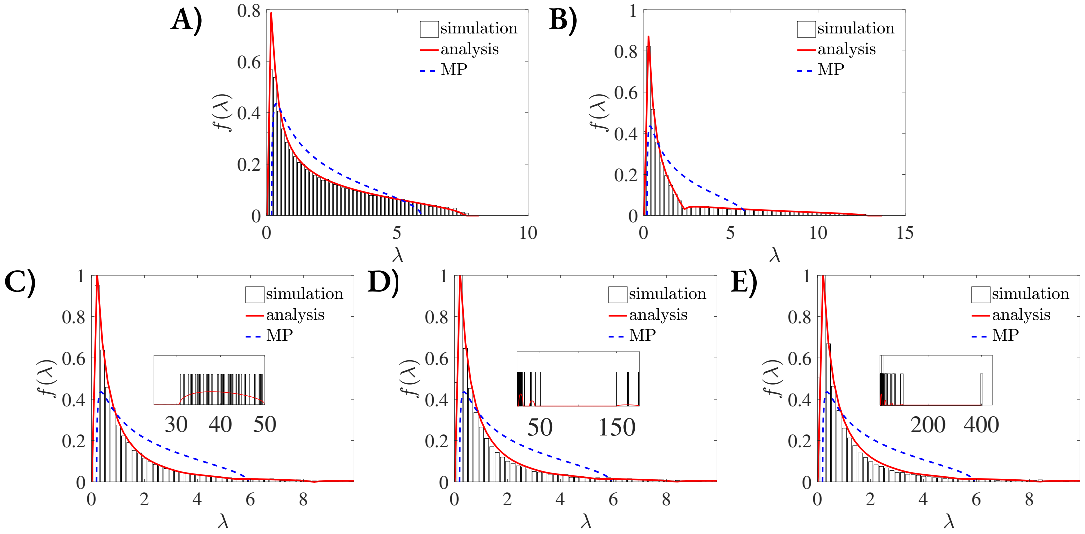


Figure 6.3: 4096 sequences of length 100 with no structural constraints are generated using the following branching events A)1; B)3; C)7 ; D) 10 ; E)12. Here analysis depicts our extension to MP, which yields an exact fit.

Eq. (6.7) becomes

$$\frac{-1}{G(z)} = z - c\sum_{i=1}^{b+1} \frac{p_i\lambda_i}{1+\lambda_i G(z)} \ . \tag{6.8}$$

By multiplying this with the factor $G(z) \times \prod_{i=1}^{b+1}(1+\lambda_i G(z))$, Eq. (6.8) is changed to the following polynomial

$$(zG+1)\prod_{i=1}^{b+1}(1+\lambda_i G) - c\sum_{i=1}^{b+1} p_i\lambda_i G\prod_{j\neq i}(1+\lambda_j G) = 0 \ . \tag{6.9}$$

$f(\lambda)$ can now be found using the inversion formula Eq. (6.6). As this requires us to find the roots to the polynomial Eq. (6.9), one limit to this method is in the accuracy of root finding algorithms for polynomials of high degree. However, we note that the eigenvalue distribution becomes station-

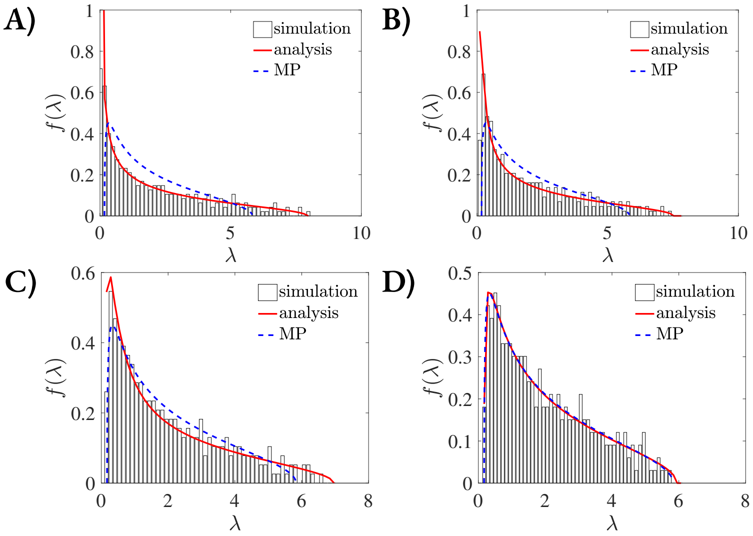


Figure 6.4: This shows $C_S = XX^T/p$, where X is a matrix consisting of 600 binary sequences of length 300 in one-hot format. Here $J = 0$ and the sequences are created with 300 random initial sequences going through one branching event. The mutations per branch are A) 3, B)20, C)50 and D)200.

ary as b increases, see Fig. 6.3. In this figure we plot sequences for different number of branching events as well as our analytically derived solution from above (red line in the figure) and the standard MP distribution (blue line). We can observe a clear difference between the exact solution we derived and the MP distribution. We further see that the change in the eigenvalue distribution between one, three and seven branching events is noticeable, while the bulk of eigenvalue distribution is almost exactly the same between 10 and 12 branching events. Thus we can approximate the eigenvalue distribution by finding the distribution of a tree with sufficiently large number of branching events.

Simple Phylogeny – Case Study As an example, consider a tree with just one branching event. The expected eigenvalue distribution is $p_1 = P(\lambda = 1+\alpha) = 1/2$, $p_2 = P(\lambda = 1-\alpha) = 1/2$

where $\alpha = \exp(-2mq/p(q-1))$. As a result, Eq. (6.9) becomes

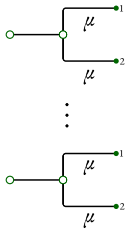
$$z(1-\alpha^2)G^3 + (2z + (1-c)(1-\alpha^2))G^2 + (z+2-c)G + 1 = 0 \ . \tag{6.10}$$

This is a polynomial of degree three, hence there are three roots, $G_1(z)$, $G_2(z)$ and $G_3(z)$. These roots are either all real, or we can have one real root and two complex conjugate roots. The eigenvalue distribution is given by $f(z) = \Im(G(z))$. Fig. 6.4 shows our analytical results (red line) compared to the empirical eigenvalue distribution, the two almost overlap exactly. Moreover, we show that as we increase the mutations along the branch, the eigenvalue distribution tends towards the Marčenko-Pastur distribution.

6.3 Spectra of Inhomogeneous Simple Phylogeny

So far, the analysis was done for trees with equal length branches. To push this towards a more realistic protein evolution model, we extend this analysis to inhomogeneous trees. Here, the mutations per branch are drawn from a probability distribution with mean $\mathbf{E}(m) \equiv \mu$. The distribution we consider is the Poisson distribution as it provides a realistic model of the frequency of events in a time interval.

We start again with the case of simple phylogeny, a tree with just one branching event. If we have n_0 initial sequences which all go through a homogeneous simple phylogenetic tree with μ mutations per branch, then the expected covariance matrix is given by :

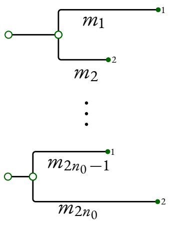


$$\Rightarrow \Sigma_S = \begin{pmatrix} 1 & \alpha_{2\mu} & & & \\ \alpha_{2\mu} & 1 & & & \\ & & \ddots & & \\ & & & 1 & \alpha_{2\mu} \\ & & & \alpha_{2\mu} & 1 \end{pmatrix} , \tag{6.11}$$

where $\alpha_{2\mu} = \exp(-2\mu q/p(q-1))$. Using Eq. (6.9), the Stieltjes transform for this homogeneous simple phylogenetic system satisfies the following

$$G\left(z - \frac{c}{2}\frac{1+\alpha_{2\mu}}{1+(1+\alpha_{2\ \mu})G} - \frac{c}{2}\frac{1-\alpha_{2\mu}}{1+(1-\alpha_{2\mu})G}\right) = -1\,, \tag{6.12}$$

where $c = n/p$ and $n = 2n_0$. For the equivalent inhomogeneous case, $m_1, m_2, \cdots, m_{2n_0-1}, m_{2n_0}$ are the branch lengths drawn from a Poisson distribution with mean μ, thus Σ_S is



$$\Rightarrow \Sigma_S = \begin{pmatrix} 1 & \alpha_{i_1} & & & \\ \alpha_{i_1} & 1 & & & \\ & & \ddots & & \\ & & & 1 & \alpha_{i_{n_0}} \\ & & & \alpha_{i_{n_0}} & 1 \end{pmatrix} , \tag{6.13}$$

where $i_1 = m_1 + m_2\ , \cdots, i_{n_0} = m_{2n_0-1} + m_{2n_0}$ and $\alpha_i = e^{-qi/p(q-1)}$. This notation satisfies three properties :$\alpha_{i+j} = \alpha_i \alpha_j; \alpha_{ij} = \alpha_i^j; \alpha_i^j = \alpha_j^i$. Using the additive property of a Poisson distribution, $i_1 , \cdots , i_{n_0}$ are independent and identically distributed variables drawn from a Poisson distribution with mean 2μ

$$\rho_i = \frac{(2\mu)^i e^{-2\mu}}{i!} \ , \tag{6.14}$$

where $\rho_i = P(i_1 = i)$. Using Eq. (6.8), G satisfies

$$G\bigg(z - \underbrace{\frac{c}{2} \sum_{i=0}^{\infty} \frac{\rho_i (1+\alpha_i)}{1 + (1+\alpha_i) G}}_{\text{(A)}} - \underbrace{\frac{c}{2} \sum_{i=0}^{\infty} \frac{\rho_i (1-\alpha_i)}{1 + (1-\alpha_i) G}}_{\text{(B)}} \bigg) = -1 \ , \tag{6.15}$$

where ρ_i is given by Eq. (6.14). We note that (A) can be rearranged into the following

$$\begin{aligned} \sum_{i=0}^{\infty} \frac{\rho_i (1+\alpha_i)}{1 + (1+\alpha_i) G} &= \frac{1}{G} - \frac{1}{G(1+G)} \sum_{i=0}^{\infty} \frac{\rho_i}{1 + \alpha_i \frac{G}{G+1}} \\ &= \frac{1}{G} - \frac{1}{G(1+G)} \sum_{i=0}^{\infty} \rho_i \sum_{j=0}^{\infty} \left(\frac{-G}{1+G} \alpha_i \right)^j \\ &= \frac{1}{G} - \frac{1}{G(1+G)} \sum_{j=0}^{\infty} \left(\frac{-G}{1+G} \right)^j \sum_{i=0}^{\infty} \rho_i \alpha_j^i \ , \end{aligned} \tag{6.16}$$

(B) can be rearranged in a similar fashion. Furthermore, we can make the following approximation

$$\sum_{i=0}^{\infty} \rho_i \alpha_j^i = \mathbf{E}_i(\alpha_j^i) \tag{6.17}$$

$$= \exp\left(2\mu (e^{-qj/p(q-1)} - 1) \right)$$

$$= \exp\left(2\mu (qj/p(q-1) + o(p^{-2})) \right) \sim \alpha_{2\mu}^j \ , \tag{6.18}$$

for large p. Substituting this back into Eq. (6.16) gives

$$\begin{aligned}\sum_{i=0}^{\infty}\frac{\rho_i(1+\alpha_i)}{1+(1+\alpha_i)G} &\sim \frac{1}{G}-\frac{1}{G(1+G)}\sum_{i=0}^{\infty}\left(\frac{-G}{1+G}\right)^i\alpha_{2\mu}^i\\ &=\frac{1+\alpha_{2\mu}}{1+(1+\alpha_{2\mu})G}\,. \end{aligned} \tag{6.19}$$

Eq. (6.19) shows that when p is sufficiently large, the summation simplifies as a function of the mean of the Poisson distribution which yields

$$G\left(z-\frac{c}{2}\frac{1+\alpha_{2\mu}}{1+(1+\alpha_{2\mu})G}-\frac{c}{2}\frac{1-\alpha_{2\mu}}{1+(1-\alpha_{2\mu})G}\right)\sim -1\,.$$

This is the same as Eq. (6.12). Intuitively, G can be approximated by the mean of the probability distribution, μ, when p is sufficiently large.

6.4 Spectra of Inhomogeneous Trees

We extend the analysis in the previous section for trees with just one branching event to inhomogeneous trees with an arbitrary number of branching events. This is done in two steps.

The first step is to show that Eq. (6.18) still holds when the following approximation can be made:

$$\mathbf{E}_i(\alpha_j^i)\sim\alpha_{\mathbf{E}(i)}^j.$$

For our purposes $\mathbf{E}(i) = 2\mu$. We show that this is a valid approximation for any probability distribution with convergent moment generating function (MGF). For the second step, we use this approximation for a tree with arbitrary number of branching events.

Definition 6.4.1. *A moment generating function (MGF), $M(t)$, of a random variable x is given by $M(t) = E(e^{tx})$.*

For the random variable i, where i is the number of mutations, the MGF is given by $\mathbf{E}_i(\alpha_i) = \mathbf{E}_i(e^{-qi/p(q-1)})$.

For the first step we want to show, for p sufficiently large, that

$$\mathbf{E}_i(\alpha_j^i) \sim \alpha_j^{\mathbf{E}(i)} = \alpha_{\mathbf{E}(i)}^j \; . \tag{6.20}$$

This condition can be equivalently expressed as $\mathbf{E}(e^{-\delta i}) \sim e^{-\delta \mathbf{E}(i)}$ where $\delta = O(1/p)$. Firstly, we note that the MGF satisfies the following:

$$\mathbf{E}(e^{-\delta x}) = \sum_{i=0}^{\infty} (-1)^i \frac{1}{i!} \delta^i \mathbf{E}(x^i) \; , \tag{6.21}$$

in particular, $\mathbf{E}(x^2) = \mathrm{var}(x) + \mathbf{E}(x)^2$. The functional form $f(x) = e^{-\delta x}$ is convex for positive δ and x. Therefore, we can apply Jensen's inequality which yields

$$e^{-\delta \mathbf{E}(x)} \leq \mathbf{E}(e^{-\delta x}) \; , \tag{6.22}$$

this gives a lower bound to Eq. (6.21). An upper bound can also be found by considering the following inequality:

$$e^{-\delta x} \leq 1 - \delta x + \frac{1}{2!} \left(\delta x\right)^2 \ , \tag{6.23}$$

we can apply the expectation operator on both sides to give:

$$\begin{aligned} \mathbf{E}(e^{-\delta x}) &\leq 1 - \delta\mathbf{E}(x) + \frac{1}{2!}\delta^2\mathbf{E}(x)^2 + \frac{1}{2!}\delta^2\mathrm{var}(x) \\ &= e^{-\delta\mathbf{E}(x)} + \frac{1}{2!}\delta^2\mathrm{var}(x) + O(\delta^3) \ . \end{aligned} \tag{6.24}$$

Consequently, Eq. (6.21) is bounded by the following :

$$e^{-\delta\mathbf{E}(x)} \leq \mathbf{E}(e^{-\delta x}) \leq e^{-\delta\mathbf{E}(x)} + O(\delta^2) \ . \tag{6.25}$$

As δ becomes sufficiently small, the upper and lower bounds both converges to $e^{-\delta\mu}$. Subsequently, we can approximate $\mathrm{E}(e^{-\delta x})$ by $e^{-\delta\mu}$ with an error term which is second order with respect to δ. Hence, as p becomes sufficiently large we can use the following approximation

$$\mathbf{E}(e^{-\delta x}) \sim e^{-\delta\mathbf{E}(x)} \ . \tag{6.26}$$

where the error of the approximation is $O(p^{-2})$.

Extending to Arbitrary Phylogeny: The Stieltjes transform, G, for a homogeneous tree with b

branching events and μ mutation events per branch is given by

$$G\left(z - c\sum_{i=1}^{b+1} p_i \frac{\lambda_i}{1+\lambda_i G(z;c)}\right) = -1\,, \tag{6.27}$$

where λ_i and p_i are given in Eq. (4.12), Eq. (4.13) respectively, further $\alpha = \exp(-2q\mu/(p(q-1))$.

For a inhomogeneous tree with l branches where the number of mutations is drawn from a distribution with mean μ, the equivalent expression to Eq. (6.27) is given by

$$G\left(z - c\sum_{i=1}^{b+1} p_i \sum_{\mathbf{M}} \frac{\rho_{m_1}\cdots\rho_{m_l}\lambda_i(\alpha_{m_1},\cdots,\alpha_{m_l})}{1+\lambda_i(\alpha_{m_1},\cdots,\alpha_{m_l})G}\right) = -1\,, \tag{6.28}$$

where $\{m_1,\cdots,m_l\}$ denotes the branch lengths, $\rho_i = P(m=i)$ and $\lambda_i(\alpha_{m_1},\cdots,\alpha_{m_l})$ satisfies $\lambda_i(\alpha_{\mathrm{\mu}},\cdots,\alpha_{\mathrm{\mu}}) = \lambda_i$.

We consider an inductive process, where we show that the following is satisfied

$$\begin{aligned}&\sum_{\mathbf{M}\in\mathbb{N}^l} \frac{\rho_{m_1}\cdots\rho_{m_l}\lambda(\alpha_{m_1},\cdots,\alpha_{m_l})}{1+\lambda(\alpha_{m_1},\cdots,\alpha_{m_l})G}\\ &\sim \sum_{\mathbf{M}\in\mathbb{N}^{l-1}} \frac{\rho_{m_1}\cdots\rho_{m_{l-1}}\lambda(a_{m_1},\cdots,\alpha_{m_{l-1}},\alpha_{\mathrm{\mu}})}{1+\rho_{m_{l-1}}\lambda(a_{m_1},\cdots,\alpha_{m_{l-1}},\alpha_{\mathrm{\mu}})G}\,.\end{aligned} \tag{6.29}$$

To do this, we first consider the Taylor expansion of the following function

$$h(x) = \frac{\lambda(\alpha_{m_1},\cdots,\alpha_{m_{l-1}},x)}{1+\lambda(\alpha_{m_1},\cdots,\alpha_{m_{l-1}},x)G} = \sum_{j=0}^{\infty} h_j x^j\,, \tag{6.30}$$

where the coefficients h_j are dependent on α_{m_i}. This Taylor expansion can then subsequently be

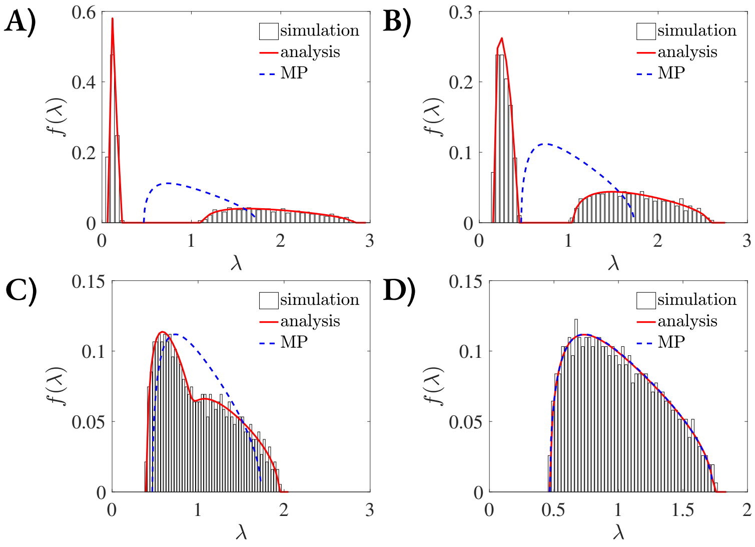


Figure 6.5: This shows $C_S = XX^T/p$ for sequences related via a simple phylogeny (one-branching event), where X is the matrix consisting of 600 21-state sequences of length 300 in one-hot format. Note that the eigenvalues wrt the true covariance matrix of simple phylogeny is give $1 \pm \alpha$ where $\alpha = \exp(-2mq/p(q-1))$. This induces a bimodal distribution when α is sufficiently large, this corresponds to when the mutation rate is low. While if the mutation rate is sufficiently high then $\alpha \to 0$ and the empirical eigenvalue distribution merges to the MP distribution. For these for plots we show how the bimodal distribution tends towards the MP distribution as we increase the mutation rate. The average mutations per branch used are: A) 20, B)50, C)200 and D)500.

used in the following way:

$$\begin{aligned}
&\sum_{i=0}^{\infty} \frac{\rho_i \lambda(\alpha_{m_1}, \cdots, \alpha_{m_{l-1}}, \alpha_i)}{1 + \lambda(\alpha_{m_1}, \cdots, \alpha_{m_{l-1}}, \alpha_i) G} \\
&= \sum_{i=0}^{\infty} \rho_i \sum_{j=0}^{\infty} h_j \alpha_i^j = \sum_{j=0}^{\infty} h_j \sum_{i=0}^{\infty} \rho_i \alpha_i^j = \sum_{j=0}^{\infty} h_j \sum_{i=0}^{\infty} \rho_i \alpha_j^i \\
&= \sum_{j=0}^{\infty} h_j \mathbf{E}_i(\alpha_j^i) \ . \qquad (6.31)
\end{aligned}$$

Using the approximation in Eq. (6.20), Eq. (6.31) can be approximated by the following

$$\sum_{j=0}^{\infty} h_j \alpha_\mu^j = h(\alpha_\mu) \ , \qquad (6.32)$$

thus Eq. (6.29) is satisfied. We can repeat this process l times yielding

$$\begin{aligned}
\sum_{\mathbf{M} \in \mathbb{N}^l} \frac{\rho_{m_1, \cdots, m_l} \lambda(\alpha_{m_1}, \cdots, \alpha_{m_l})}{1 + \lambda(\alpha_{m_1}, \cdots, \alpha_{m_l}) G} &\sim \frac{\lambda(\alpha_\mu, \cdots, \alpha_\mu)}{1 + \lambda(\alpha_\mu, \cdots, \alpha_\mu) G} \\
&= \frac{\lambda}{1 + \lambda G} \ . \qquad (6.33)
\end{aligned}$$

Substituting this back into Eq. (6.28), we find that this equation is approximated by Eq. (6.27). We can see from Fig. 6.5 that our analytical solution is indeed correctly overlaying the empirical eigenvalues observed.

Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge: the serenity to accept the things we cannot predict, the courage to predict the things we can, and the wisdom to know the difference.

Nate Silver

7

Cleaning Protein Spectra of Phylogeny

There have been many heuristic approaches in cleaning the phylogenetic bias, such as reweighting of sequences[2,41,18]; pruning the multiple sequence alignment to be rid of redundant sequences; or more advanced methods involving the inference of a tree such as the evolutionary trace (ET)[35,57], or phylogenetic significance testing[51,53,52]. However, these all rely on the inference of a phylogenetic

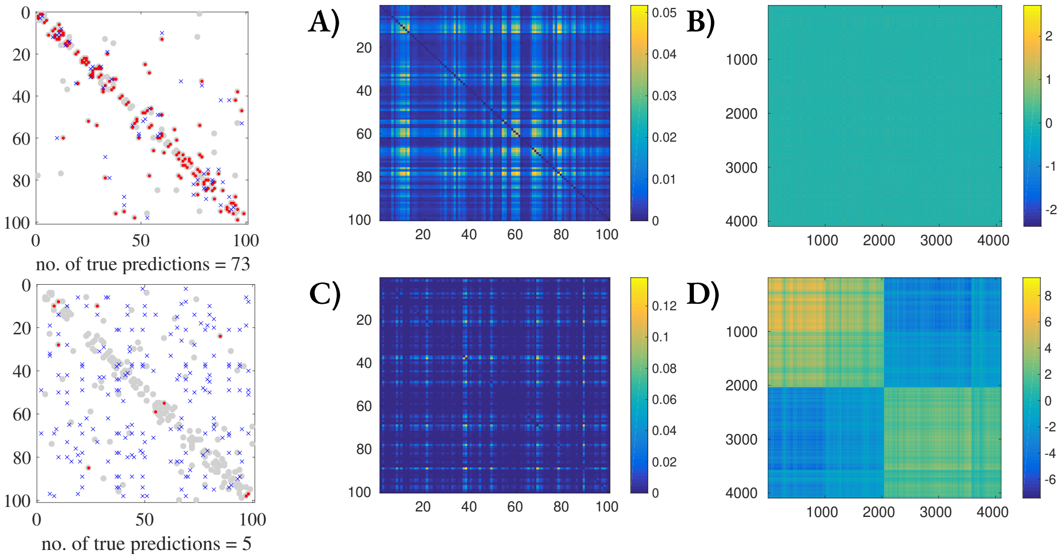


Figure 7.1: 4096 sequences of length 100 are generated using 100 pairwise interactions with strength uniformly distributed between -10 and 10. Top row are independent sequences and bottom row are sequences generated with 12 branching events and mutation rate 0.03. The pairwise interaction matrix is shown on the left where the grey dots are the true interactions, red dots represent correctly predicted contact and blue cross is a false predicted contact. A-D) shows the outer product of the top eigenvector, v_1 of C. A,C) $v_1^T v_1$, B,D) $(Xv_1)^T(Xv_1)$

tree using various methods such as parsimony or maximum likelihood. The inference of a phylogenetic tree is computationally expensive, further, it is unclear how the fine details of a phylogenetic tree can help contact prediction.

In the previous chapters we have shown that given a phylogenetic tree we can infer the eigenvalue distribution of the empirical covariance matrix $C \propto X^T X$. In this chapter, we now put this knowledge to good use, by demonstrating how we can disentangle phylogenetic signal from phenotypic by inspecting the eigenvalue distribution.

Finally, we demonstrate the effectiveness of our approach in removing phylogenetic noise for both our synthetic data and real protein data.

7.1 Relation between Eigenvectors

If the sequences are independent, then the maximum likelihood estimate (MLE) for the phenotypic pairwise interactions is given by $C = (X^TX)/n$. Similarly, if there are no phenotypic interactions, then the MLE for the dependence between the sequences is given by $C_S = (XX^T)/p$. However, when there are both phenotypic and phylogenetic dependencies then C and C_S are no longer optimal estimates since the matrices contain both phenotypic and phylogenetic information.

We have seen from Section 5.3 that the non-zero eigenvalues C and C_S are related via scaling. Here, we show that the eigenvectors are also related by a linear transformation. The scaling and the linear transformation is derived via the following

$$\begin{aligned} X^TX\mathbf{v} = \lambda\mathbf{v} \quad &\Rightarrow \quad Cv = \tfrac{1}{n}\lambda v \\ XX^T(X\mathbf{v}) = \lambda X\mathbf{v} \quad &\Rightarrow \quad C_S(Xv) = \tfrac{1}{p}\lambda(Xv)\ . \end{aligned} \tag{7.1}$$

In other words the following mapping transforms the eigenvectors in the residue space to its corresponding sequence space

$$X : \{v_1, \cdots, v_r\} \mapsto \{Xv_1, \cdots, Xv_r\}\ , \tag{7.2}$$

where $r = \max(n, p)$ and $\lambda_1 \geq \cdots \geq \lambda_r$ are the corresponding eigenvalues.

This correspondence between eigenvectors means that each eigenvector should contain information both about the phenotypic interactions and phylogenetic dependencies.

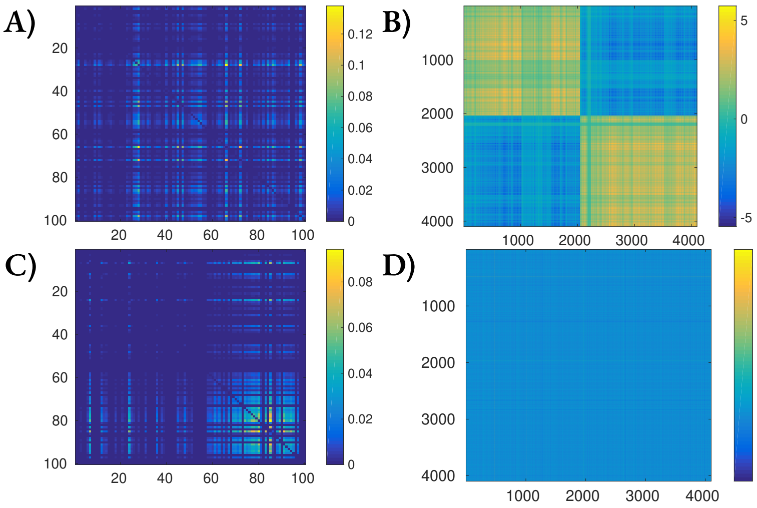


Figure 7.2: 4096 sequences of length 100 are generated with 12 branching events and 0.03 mutation rate using 100 pairwise interactions with strength uniformly distributed between -10 and 10. This shows the outer product of the eigenvectors, A-B) v_1, Xv_1 and C-D) v_{50}, Xv_{50}.

This sharing of information can be seen in the outer products as we switch between the residue and sequence space shown in Fig. 7.1. Crucially, in this figure we see the catastrophic effects which this sharing of information has on our ability to predict contacts if we compare the top and bottom row figures 7.1. While the correct contacts predicted using C is 73/100 for independent sequences, we only managed to predict 5/100 correct contacts when the sequences are related by phylogeny. This figure also exhibits the sharing of information in the principal eigenvectors. In particular, for Fig. 7.1C) we observe that the principal eigenvector of C, v_1, doesn't appear to pin down any particular phenotypic interaction pair, rather, it precisely captures the first duplication event when we transform it into sequence space (see Fig. 7.1D)), Xv_1. It would appear that this eigenvector was 'phylogenetic' rather than 'phenotypic'.

This observation motivated our approach in the rest of this chapter: maybe there are some eigenvectors which contain significantly more phylogenetic information and vice versa. This also raises the question, how much does phylogeny affect each of the eigenvectors of the phenotypic empirical covariance matrix? From Fig. 7.2 we see the outer products of v_1 and Xv_1, v_{50} and Xv_{50} where v_1 and v_{50} are the eigenvectors of C. The first duplication event is clearly captured in the principal eigenvector, Xv_1, while Xv_{50} appears to be white noise. This suggests that phylogeny is a more dominant factor in the eigenvectors with larger corresponding eigenvalues. This is intuitive, since phylogenetic factors are ubiquitous amongst the sequences while phenotypic factors affect only the local residues involved. If this indeed is the case, then this raises the exciting prospect of cleaning phylogenetic biases by removing the top eigenmodes.

7.2 Phenotypic Modes Exists in Conjugate Pairs

Although cleaning the phylogenetic bias by removing the top eigenmodes seems like a simple and attractive solution, it is paramount that we do not remove important phenotypic information by doing so. Thus, we want to understand how the information is encoded in C for a sequence alignment where there is no sequence dependence (i.e. phylogeny). We consider a system with only one pairwise interaction to gain some insight. Concretely, let us assume that the underlying covariance matrix for the system is

$$\Sigma = \begin{pmatrix} 1 & \beta & 0 & 0 \\ \beta & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

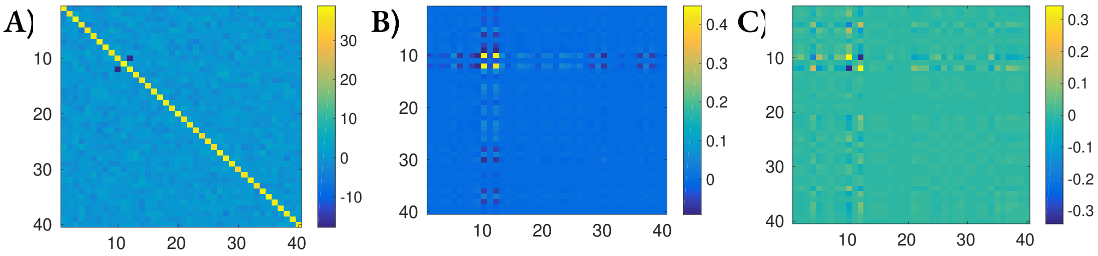


Figure 7.3: We visualise the outer product of the eigenvectors with largest and smallest eigenvalues, extracted from the covariance of 700 independent sequences of length 40 with one pairwise interaction. A) C (X in amino acid format), B) $\mathbf{v}_1$, C) $\mathbf{v}_r$.

this describes a system with an interaction between residue positions 1 and 2. Then the eigenmodes associated with the largest and lowest eigenvalue $(1 \pm \beta)$ are correspondingly $\mathbf{v}_+ = (1, 1, 0, 0)^T/\sqrt{2}$ and $\mathbf{v}_- = (1, -1, 0, 0)^T/\sqrt{2}$. Unlike PCA, this system suggests that information is encoded in the modes of the largest and smallest eigenvalues, precisely in those eigenvalues which are not one. Fig. 7.3 shows the outer product of the eigenvector with the largest and smallest eigenvalues. Surprisingly, the pairwise interaction is encoded in both of the eigenmodes. Furthermore, the modes at the top end of the spectrum and the bottom ends of the spectrum appear to hold the same phenotypic information albeit with different signs, in some sense they contain the same information but in a different orientation*. This suggests that if the top eigenmodes are corrupted by phylogeny, we can try to retain the right amount of lower eigenmodes and not lose any phenotypic information.

*The result of adding these outer products together is given by

$$\mathbf{v}_+\mathbf{v}_+^T + \mathbf{v}_-\mathbf{v}_-^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Apart from the diagonal imprint, there is no information in the off diagonal elements when the top and bottom modes are added together. This behaviour suggests that the information at the different ends of the spectrum are 'oriented' differently.

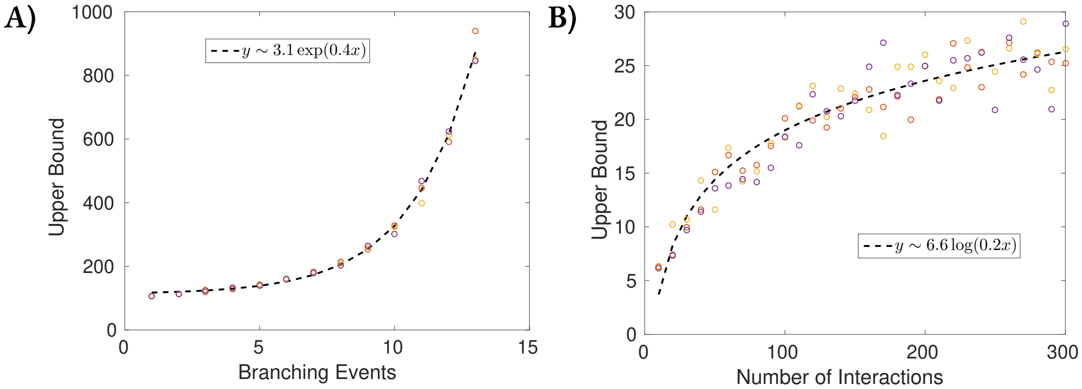


Figure 7.4: The change to the maximum eigenvalue of C_S with A) phylogenetic and B) phenotypic parameters. The line of best fit is found using MATLAB, the different runs are represented by different colors. A) 8192 ($n_0 \times 2^b = 8192$) amino acid sequences of length 100 are generated with no phenotypic signals; the maximum eigenvalue is measured against the number of branching events b. B) 4000 amino acid sequences of length 100 are generated independently of each other where the non-zero elements in the interaction matrix J is one; the maximum eigenvalue is measured against the number of interactions.

7.3 Principal Eigenvectors Dominated By Phylogeny

The sharing of information between matrices C and C_S curbs our ability to accurately infer phenotypic interactions when the sequences are related via a phylogenetic tree. We have speculated that the modes with higher eigenvalues are more influenced by phylogeny, giving rise to the possibility of separating phylogeny from phenotype information by removing the principal eigenvectors. Here, we will show that by analysing the spectra of C_S, the extremely large eigenvalues can only be conducive to phylogeny.

Fig. 7.4 shows the functional behavior of the maximum eigenvalue when we simulated sequences with either phylogeny (Fig. 7.4A) or phenotypic interactions (Fig. 7.4B). Here, we find that the largest eigenvalue increases exponentially as we increase the number of duplication events while the increase is only logarithmic when we increase the number of phenotypic interactions. This suggests

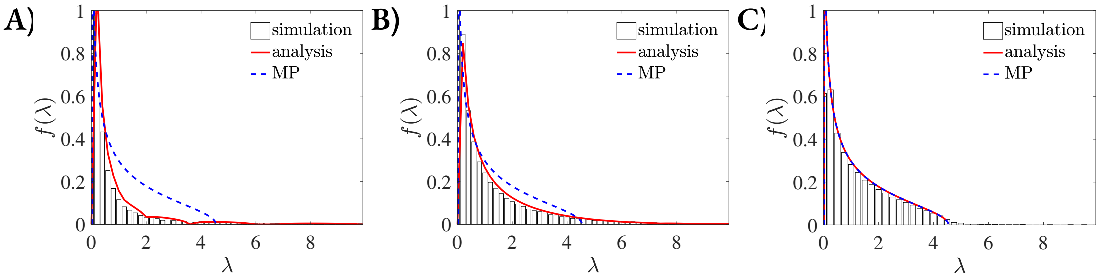


Figure 7.5: The sequences are generated with 12 branching events and phenotypic interactions that is akin to DYR-ECOLI (see Fig. 5.2). A, B, C) uses 20, 100, 500 mutations per branch respectively. Here we see the empirical eigenvalue distribution for $C_S = XX^T/p$. The red line is the analytical solution using Eq. (6.9) and the effective mutation rates, the blue line is the MP distribution.

that when the phylogenetic tree has a sufficiently large number of branching events (as is the case in protein sequences), then the largest eigenvalues in the spectrum of our empirical covariance matrix must be due to phylogeny.

Moreover, our empirical observations show that the shape of the eigenvalue distribution is mostly controlled by phylogeny and not by phenotype. This can be seen in Fig. 7.5. Here the red lines are the predicted analytical eigenvalue distribution from phylogeny. We see that it closely matches with the empirical eigenvalue distribution observed - the sequences generated have both phylogenetic dependencies and phenotypic interactions.

7.4 Power Law Tails

We have shown that the shape of the eigenvalue distribution of the empirical covariance matrix is dictated by phylogeny – crucially even when phenotypic constraints are present. As a result, we can extend many properties that apply to the eigenvalue distribution induced by only phylogeny to the

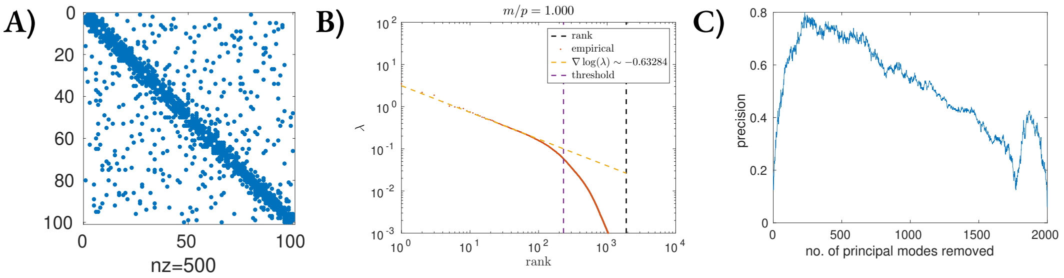


Figure 7.6: Sequences generated using 13 branching events with 500 interactions, the mutation rate is approximately 105. A) the interactions B) log log scaling of the spectrum, C) accuracy of predicted contacts given truncation.

eigenvalue distribution when both phylogeny and phenotype are present such as protein sequences. In particular, we have seen in Fig. 7.4 that the maximum eigenvalue increases exponentially as the number of branching events increase. Further, we have also shown that the probability decreases exponentially as the eigenvalues increase, see Eq. (4.13). Thus, the heavy tailed eigenvalue distribution is symptomatic of the phylogenetic process, while phenotypic information is present in the lower half of the spectrum which is much harder to detect. With the phylogenetic signals dominating the phenotypic signals in the principal modes, we hope to remove phylogenetic signals from the phenotypic by removing these - and only these - modes.

Since the eigenvalue distribution is mostly dictated by phylogeny, for this part of our analysis, we will, at first, make the simplification that the sequences have only phylogenetic dependence. Here, we will first show that the eigenvalue distribution induced by phylogeny follows a power law in the *low mutation regime*. To see this, we can use Eqs. (4.12) and (4.13) to deduce the power law in the tail of the empirical eigenvalue distribution when mutation rate is sufficiently low. We note that the expected eigenvalue distribution in Eq. (4.12) can be rewritten as

$$\begin{aligned} \lambda(r) &= (1-\alpha)(1+2\alpha+\cdots+(2\alpha)^{k}) \\ \lambda(2r) &= (1-\alpha)(1+2\alpha+\cdots+(2\alpha)^{k-1})\,, \end{aligned} \tag{7.3}$$

here $k = b - \lfloor \log_2 r \rfloor$, b is the number of branching events, $r > 1$ and $\lambda(r) = \lambda_{\lfloor \log_2 r \rfloor}$ (λ_i is defined in Eq. (4.12)). If we consider the bth generation of sequences where b is sufficiently large, we can evaluate the gradient of $\log(\lambda)$ as a function of $\log(r)$ by taking the approximation that $\lambda(r) \sim O((2\alpha)^{k+1})$ when $2\alpha > 1$. This results in the following

$$\lambda(r) \sim O((2\alpha)^{b-\log(r)/\log(2)+1}) \tag{7.4}$$

$$\lambda(r) \sim O\left(\exp(\log(2\alpha)(b-\log(r)/\log(2)+1))\right) \tag{7.5}$$

$$\to \nabla_{\log(r)} \log(\lambda) = -\frac{\log(2\alpha)}{\log(2)}\,. \tag{7.6}$$

For the case where $2\alpha \leq 1$ we note that $\nabla \log(\lambda) \sim 0$; using the approximation that k is sufficiently large in the tail of the eigenvalue distribution for large b.

We can further simplify the above by considering α as a function of the mutation rate m/p, $\alpha = e^{-2qm/(p(q-1))}$, this gives

$$\nabla_{\log(r)} \log(\lambda) \propto \begin{cases} \frac{2q}{\log(2)(q-1)}\frac{m}{p} - 1 & \frac{(q-1)\log(2)}{2q} > \frac{m}{p} \\ 0 & \text{otherwise.} \end{cases} \tag{7.7}$$

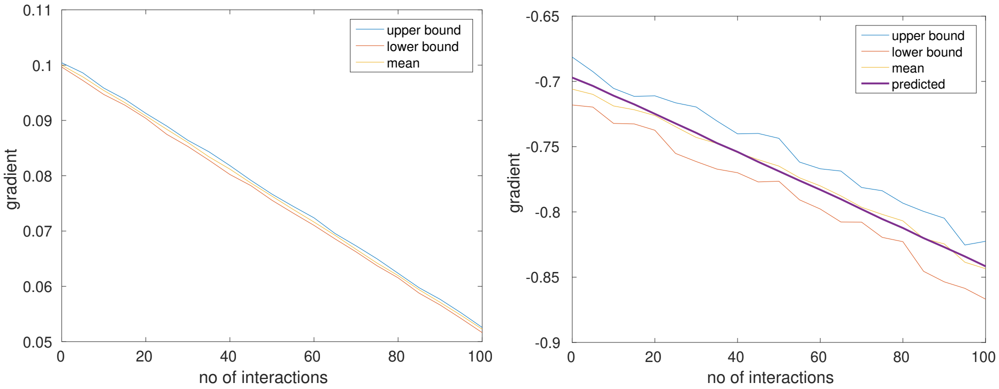


Figure 7.7: We used 12th generation sequences of length 100, the average mutations per branch is 10. The x-axis is the size of the block of interactions, the left panel shows the effect mutation rate when the true mutation rate is 0.1, the right panel is the corresponding gradient extracted from a prediction using MATLAB's fit function. The blue, yellow and red lines show the maximum, mean and minimum values, respectively, calculated from 20 runs. The purple is $y(m_{eff}) = -\log(2\alpha_{eff})/\log(2)$.

From this we see that $\nabla_{\log(r)} \log(\lambda)$ is piecewise constant, depending on the mutation rate m/p. Thus, $\lambda(r) \propto r^{\nabla \log(\lambda)} = r^{cst}$ when λ is sufficiently large; also note that the exponent cst is dependent on the mutation rate. Concretely, the exponent is always negative with a slower mutation rate generating a steeper gradient. Fig. 7.6 shows the prediction the slope of the power law using Eq. 7.7, we can see that the observed slope is exactly as predicted.

7.5 Effective Mutation Rate

One subtlety in the analysis we have done so far, is that the m used is the proposed number of mutations for the Metropolis Hastings algorithm; cf. Algorithm 1. However, the number of mutations accepted will be significantly different in the case of pathological interactions[27], where the phenotypic interactions are spatially correlated. Here, we use the term *'effective mutation rate'* to mean the rate of mutations which are accepted.

The spatial correlation mentioned above is necessary for a realistic protein model, as it is well known that a protein's 3D structure is composed of a variety of beta sheets, alpha helices and loops. This so called secondary structure projects a checkerboard-like pattern on the contact map between proteins. For example, Fig. 1.1 shows the pattern induced in the contact map by oxymyoglobin which consists of α-helix strands. An important question to answer thus is: What happens to the power law when the spatial correlation between the amino acid residues are high? To investigate this question, we impose a simple interaction matrix where the spatial correlation is high.

$$J = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} . \tag{7.8}$$

Although this model is unrealistic, it can be seen as a stress test that we can use to uncover to what extent imposing a strong spatial correlation affects the power law induced by phylogeny alone. Fig. 7.7 shows the relationship between the power tail and the block size of interactions. We make two noteworthy observations: firstly the effective mutation rate decreases linearly as we increase the block size, this behaviour is mirrored in the power tail, and secondly the average gradient of exponent of λ is well approximated by $-\log(2\alpha_{eff})/\log(2)$ where $\alpha_{eff} \propto \frac{m_{eff}}{p}$ and m_{eff}/p is the effective mutation rate.

For this simple synthetic model the linear behaviour can be analytically found. Consider the probability of accepting a mutation at position i which we denote as m_i, the ith residue has interacting neighbours if $i \leq x$, where x is the size of the block. We can apply the sum of probability to get

$$\begin{aligned} P(m_i) &= P(m_i|i \leq x)P(i \leq x) + P(m_i^C|i > x)P(i > x) \\ &= 1 - \frac{1 - e^{-\Delta E}}{p} x \,. \end{aligned} \tag{7.9}$$

The effective mutation rate would be given by $m_{eff}/p = P(m_i)m/p$ and this is a linear function of the block of interactions of size x.

7.6 Cleaning Protein Spectra

Finally, we arrive at the application of the above developed theory to cleaning protein spectra.

7.6.1 Pruning Sequence Alignments

To start, as a sanity check, we first want to see what a canonical phylogeny correction such as pruning the multiple sequence alignment would do to the eigenvalue distribution of the empirical covariance matrix. This is shown in Fig. 7.8. We see that the eigenvalue distribution remains almost exactly the same when we prune the alignment; indicating that phylogenetic bias is still very much present.

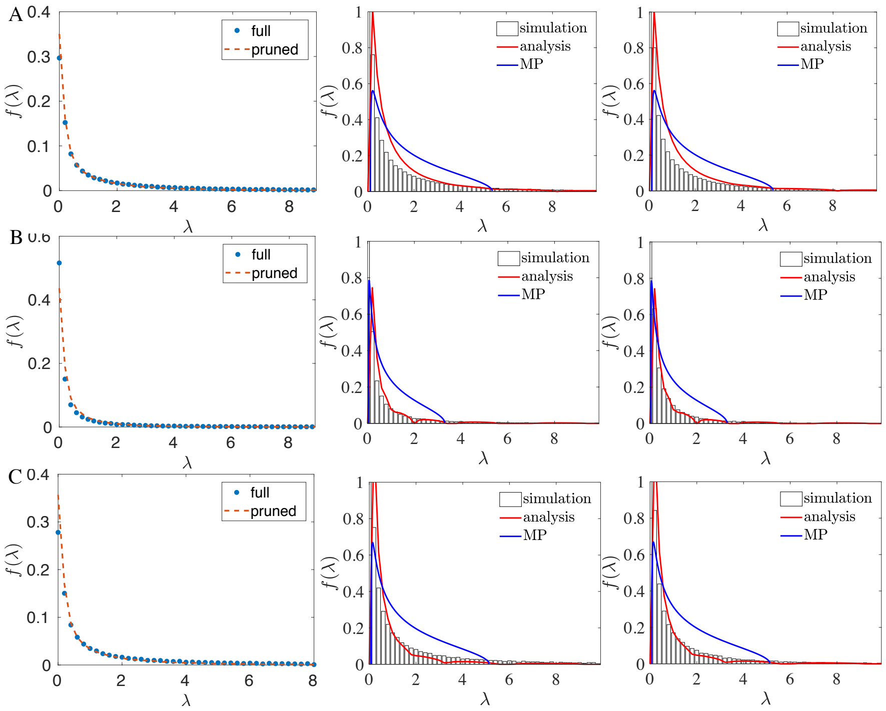


Figure 7.8: Comparison of the eigenvalue spectra of large protein sequence alignments with and without the canonical phylogeny correction used in the literature. Here we compare no treatment to pruning the alignments, whereby sequences are removed so that no two remaining sequences are more similar than a user-chosen threshold—often a Hamming distance of around 0.3 is chosen. These plots show analysis of the spectra resulting from alignments for (A) Trypsin, (B) DHFR, and (C) TRML HAEIN both before (A–C, Left and Center) and after (A–C, Left and Right) we prune the alignment.

7.6.2 Removing Phylogenetic Modes

Our approach then, which we present in this section, is derived from the signature behaviour of the eigenvalue distribution with respect to phenotypic and phylogenetic signals; as described in previous section. Furthermore, we note that this analysis carries through even when the sequences are generated with a superposition of both.

As we have argued above, the presence of phylogeny will introduce a heavy tailed eigenvalue distribution for the empirical covariance matrix. This remains to be true also in the presence of phenotype interactions.

This heavy tailed distribution induced by phylogeny follows a power law which is dependent on the mutation rate of evolution. Here, we hypothesise that in the presence of both phenotypical mutations and phylogeny the modes whose eigenvalues follow this power law are 'phylogenetic' modes. In other words, these are eigenvectors whose signal is mostly induced by phylogeny rather than the phenotypic interactions we seek.

We use our simulated data to test this hypothesis. Indeed, the analysis of simulated data suggests that the effects of phylogeny can be minimised by removing large modes from the covariance matrix, and enforcing the constraint that the remaining eigenvalues are all of the same size. Concretely, if our empirical covariance matrix $C \propto X^T X$ has eigenvalues $\lambda_1 \leq \cdots \leq \lambda_t \leq \cdots \leq \lambda_{p(q-1)}$ with corresponding eigenvectors v_i, then the most effective way to enhance phenotypic signals whilst removing phylogenetic signals is given by the truncated covariance

$$C(t) = \mathbf{v}_1\mathbf{v}_1^T + \cdots + \mathbf{v}_t\mathbf{v}_t^T, \quad \lambda_1 \leq \cdots \leq \lambda_t \leq \cdots \leq \lambda_{p(q-1)}, \tag{7.10}$$

here t is the parameter which controls how many principal modes we truncate.

7.6.3 Effectiveness on Simulated Sequences

Our results are highlighted in Fig. 7.6. In Fig. 7.6A we show the interaction matrix J used to simulate the sequences. In Fig. 7.6B we show the plot of $\log(r)$ vs $log(\lambda)$. Indeed we see a linear dependence for the highest eigenvalues, the yellow line in the figure is the predicted linearity shown in Eq. (7.6). The empirically observed eigenvalues match our prediction.

The most important part of this figure, is given in Fig. 7.6C. Here, we plot the accuracy of our predicted contacts (interactions) vs the number of principal modes we remove – namely the truncation t from Eq. (7.10). When no principal modes are removed the accuracy of the prediction of interactions is less than 10%, when we remove the first $O(200)$ principal modes out of 2000, the accuracy shoots up to 80%. This agrees with our analysis that the largest principal modes are highly corrupted by phylogeny. Moreover, the optimal number of modes to remove – to obtain the highest accuracy – is highlighted in the purple line shown in Fig. 7.6B) which corresponds to the cut-off where the eigenvalues are no longer following the power law. All of this is in exact agreement with our analysis.

7.6.4 Effectiveness on Protein Families

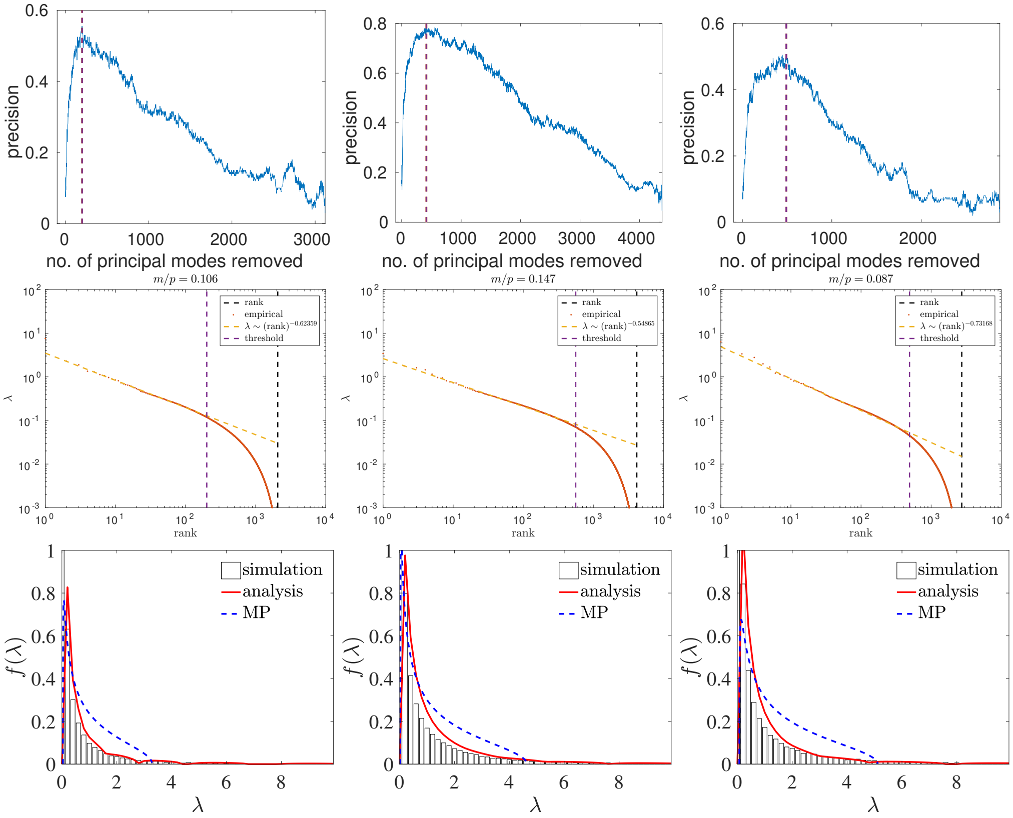


Figure 7.9: Protein sequences used DYR ECOLI,TRY2 RAT,TRML HAEIN. The top row shows the precision of the contact prediction when we remove the largest eigenmodes; the number of principal modes removed is shown on the x-axis. The middle row shows the power law induced by the design matrix of the proteins with the purple line showing the threshold at which the lower rank eigenvalues should be discarded to optimize the precision. No pseudocounts are used. The purple line shows that maximum contact prediction is reached when 200, 416, 491 eigenmodes are removed, respectively. The bottom row uses the power law to predict the spectrum of the proteins, the analytical result is shown by the red line.

We now examine the eigenvalue distributions of the covariance matrix formed from three, different, real protein families. Given the vastly different signatures in the eigenvalue distributions expected from phylogeny and phenotypic interactions, it is of great interest to see if such signatures arise in practice. We choose three representative protein families for our experiment. We selected these families adversarially, i.e. we chose them because covariance analysis has been shown to yield inaccurate contact predictions for them. The results when applying our method are given in Fig. 7.9.

The middle panels of Fig. 7.9 show that the eigenvalue distributions for each of the three families satisfy the power law tail, $\lambda \sim r^{-x}$, with x ranging from 0.4 to 0.9. Given the formula (7.6), this means that $m/p \sim O(0.1)$. The top panels of Fig. 7.9 show the results of $C(t)$ for predicting contacts. For each protein, the optimal threshold, T, is found, i.e when we remove $\mathbf{v}_t$ for $t > T$ the highest contact accuracy is obtained. The lower panels of Fig. 7.9 compare the optimal truncation (purple vertical line) with the eigenvalue distribution. There is an astonishing agreement between the modes best removed and the modes which follow the linearity of the power law. The result is congruent with our theory; the power law indicates the presence of strong phylogenetic effects. Further, the power law tails give us the estimated effective mutation rates allowing us to predict the spectra, which is the analytical solution (red lines) shown in the bottom panels. Overall, we observe a remarkable match between our analytical predictions and empirical data extracted from real protein sequences.

8

Conclusion

In this thesis we aimed to understand, at a fundamental leve, how we can use the protein data at our disposal to better infer protein structures. To this end, we focused on analysing properties of covariance matrices induced by a Multiple Sequence Alignment (MSA). In particular, we wanted to understand how phylogeny and residue-residue interactions differently impact the empirical covariance we observe in proteins. By building upon existing results from Random Matrix Theory (RMT), we were able to develop new approaches in RMT spectral analysis to handle random matrices that can be closely related to correlations extracted for different protein families. In particular, we can now analyse MSAs in which phylogeny is a major factor.

Both our theoretical and empirical analysis show that covariance matrices induced by MSAs are dominated by signals from phylogeny rather than phenotype (residue-residue interactions). Concretely, we find that phantom correlations induced by phylogeny causes eigenvalues in the empirical covariance matrix to follow a heavy-tailed distribution. In this distribution the largest eigenvalues/eigenvectors are largely due to phylogeny, following a power-law that depends on the mutation rate. This is a theoretical result that we built using simulated models of tree-related Boltzmann sequences. More importantly, we find that this theoretical result is consistent with what we observe empirically for real protein data. Namely, we find that for MSAs, the spectral distribution of an empirical covariance matrix is always heavy-tailed with the tail satisfying a power law.

This discovery allowed us to develop a simple and theoretically grounded method to disentangle the correlations induced by structure/phenotype and phylogeny by leveraging the shape of the eigenvalue distribution of the empirical covariance matrix. Indeed, the primary accomplishment of this manuscript is in identifying that phylogenetic relations between amino-acid sequences give rise to a power law tail in the eigenvalue distribution of covariance matrices. This distinct feature can be used to distinguish the covariation caused by phylogeny from that caused by phenotypic interactions. Crucially, we demonstrate that covariance matrices are most effective at predicting residue-residue contact when we remove eigenmodes for which the corresponding eigenvalues follow the power-law, further validating our theoretical result – and retain other eigenmodes.

The presence of power law tails in both synthetic Boltzmann sequences as well as the data from diverse protein families shows the dominance of covariant signals induced by phylogeny. At the same time, it means that removing the modes associated with the power law is an elegant way of deconvolving covariance induced by phenotypic interactions from the covariance that results from sequence phylogeny. Interestingly, this offers an alternative rationale for the covariance matrix inversion step (i.e. mean-field approximations) that enabled features of protein structure and function to be predicted from covariance analysis of large protein sequence alignments[10,56,62,43,29,19,45].

Understanding the extent to which the effects of phylogeny and structural/functional interactions can be disentangled is an important direction for further research; and this thesis only provides a first step in this direction. Many other open questions such as: In what circumstances can we accurately infer the strength of interactions? What happens in the low-data regime, e.g. when we have a multiple sequence alignment of only one or two sequences? still remain to be answered. Our hope is, that the approach outlined in this work can provide a mathematical framework that future work can build upon to make progress towards answering these questions.

References

[1] Altschuh, D., Lesk, A., Bloomer, A., & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology*, 193(4), 693–707.

[2] Altschul, S. F., Carroll, R. J., & Lipman, D. J. (1989). Weights for data related by a tree. *Journal of molecular biology*, 207(4), 647–653.

[3] Bialek, W. & Ranganathan, R. (2007). Rediscovering the power of pairwise interactions. *arXiv preprint arXiv:0712.4397*.

[4] Biroli, G., Bouchaud, J.-P., & Potters, M. (2007). On the top eigenvalue of heavy-tailed random matrices. *EPL (Europhysics Letters)*, 78(1), 10001.

[5] Bouchaud, J.-P., Laloux, L., Miceli, M. A., & Potters, M. (2007). Large dimension forecasting models and random singular value spectra. *The European Physical Journal B*, 55(2), 201–207.

[6] Bouchaud, J.-P. & Potters, M. (2003). *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge university press.

[7] Bray, A. & Moore, M. (1982). On the eigenvalue spectrum of the susceptibility matrix for random spin systems. *Journal of Physics C: Solid State Physics*, 15(23), L765.

[8] Brenner, M. P., Colwell, L. J., et al. (2016). Predicting protein–ligand affinity with a random matrix framework. *Proceedings of the National Academy of Sciences*, 113(48), 13564–13569.

[9] Burda, Z., Görlich, A., Jarosz, A., & Jurkiewicz, J. (2004). Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications*, 343, 295–310.

[10] Burger, L. & van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, 4, 165.

[11] Buslje, C. M., Santos, J., Delfino, J. M., & Nielsen, M. (2009). Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9), 1125–1131.

[12] Cocco, S., Monasson, R., & Weigt, M. (2013). From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS Comput Biol*, 9(8), e1003176.

[13] de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4), 249–261.

[14] Dunn, S. D., Wahl, L. M., & Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3), 333–340.

[15] Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

[16] Dutheil, J. Y. (2012). Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Briefings in bioinformatics*, 13(2), 228–243.

[17] Edelman, A. & Rao, N. R. (2005). Random matrix theory. *Acta Numerica*, 14, 233–297.

[18] Ekeberg, M., Hartonen, T., & Aurell, E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276, 341–356.

[19] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1), 012707.

[20] Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2), 211–221.

[21] Frenkel, D. & Smit, B. (2001). *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier.

[22] Gardiner, C. W. et al. (1985). *Handbook of stochastic methods*, volume 3. Springer Berlin.

[23] Georges, A. & Yedidia, J. S. (1991). How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9), 2173.

[24] Gerstein, M., Sonnhammer, E. L., & Chothia, C. (1994). Volume changes in protein evolution. *Journal of molecular biology*, 236(4), 1067–1078.

[25] Gouveia-Oliveira, R. & Pedersen, A. G. (2007). Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology*, 2(1), 1.

[26] Grimmett, G. & Stirzaker, D. (2001). *Probability and random processes*. Oxford university press.

[27] Huang, Y.-F. & Golding, G. B. (2014). Phylogenetic gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLOS Comput Biol*, 10(1), e1003429.

[28] Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

[29] Jones, D. T., Buchan, D. W., Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2), 184–190.

[30] Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3(21), 132.

[31] Khatri, B. S., McLeish, T. C., & Sear, R. P. (2009). Statistical mechanics of convergent evolution in spatial patterning. *Proceedings of the National Academy of Sciences*, 106(24), 9564–9569.

[32] Laloux, L., Cizeau, P., Potters, M., & Bouchaud, J.-P. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03), 391–397.

[33] Lapedes, A. S., Giraud, B. G., Liu, L., & Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Statistics in molecular biology and genetics - IMS Lecture Notes - Monograph Series*, volume 33 (pp. 236–256).

[34] Larson, S. M., Di Nardo, A. A., & Davidson, A. R. (2000). Analysis of covariation in an sh3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *Journal of molecular biology*, 303(3), 433–446.

[35] Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2), 342–358.

[36] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12), e28766.

[37] Marčenko, V. A. & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4), 507–536.

[38] McLeish, T. (2005). Protein folding in high-dimensional spaces: hypergutters and the role of nonnative interactions. *Biophysical journal*, 88(1), 172–183.

[39] Miller, C. S. & Eisenberg, D. (2008). Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14), 1575–1582.

[40] Mora, T. & Bialek, W. (2011). Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2), 268–302.

[41] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49), E1293–E1301.

[42] Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11), 2.

[43] N.Halabi, Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138, 774–786.

[44] Obermayer, B. & Levine, E. (2014). Inverse ising inference with correlated samples. *New Journal of Physics*, 16(12), 123017.

[45] Ovchinnikov, S., Kamisetty, H., & Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, 3, e02030.

[46] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genet*, 2(12), e190.

[47] Plefka, T. (1982). Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general*, 15(6), 1971.

[48] Qin, C. & Colwell, L. J. (2018). Power law tails in phylogenetic systems. *Proceedings of the National Academy of Sciences*, 115(4), 690–695.

[49] Rao, N. R. & Edelman, A. (2008). The polynomial method for random matrices. *Foundations of Computational Mathematics*, 8(6), 649–702.

[50] Rivas, E. (2005). Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC bioinformatics*, 6(1), 1.

[51] Rivas, E. (2016). R-scape user's guide.

[52] Rivas, E., Clements, J., & Eddy, S. R. (2016). A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. *Nature Methods*.

[53] Rivas, E. & Eddy, S. R. (2015). Parameterizing sequence alignment with an explicit evolutionary model. *BMC bioinformatics*, 16(1), 1.

[54] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.

[55] Shindyalov, I., Kolchanov, N., & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, 7(3), 349–358.

[56] Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6), 1043–1054.

[57] Sung, Y.-M., Wilkins, A. D., Rodriguez, G. J., Wensel, T. G., & Lichtarge, O. (2016). Intramolecular allosteric communication in dopamine d2 receptor revealed by evolutionary amino acid covariation. *Proceedings of the National Academy of Sciences*, 113(13), 3539–3544.

[58] Tao, T. (2012). *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI.

[59] Tao, T., Vu, V., Krishnapur, M., et al. (2010). Random matrices: universality of esds and the circular law. *The Annals of Probability*, 38(5), 2023–2065.

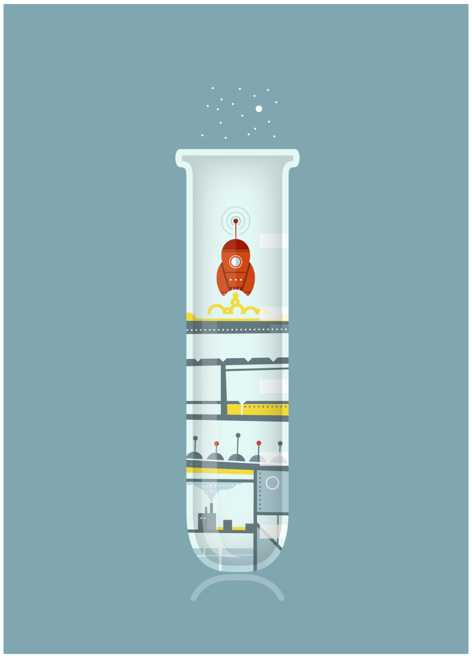
[60] Tkacik, G., Schneidman, E., Berry, I., Michael, J., & Bialek, W. (2009). Spin glass models for a network of real neurons. *arXiv preprint arXiv:0912.5409*.

[61] Townsend, P. D., Rodgers, T. L., Glover, L. C., Korhonen, H. J., Richards, S. A., Colwell, L. J., Pohl, E., Wilson, M. R., Hodgson, D. R., McLeish, T. C., et al. (2015). The role of protein-ligand contacts in allosteric regulation of the escherichia coli catabolite activator protein. *Journal of Biological Chemistry*, 290(36), 22225–22235.

[62] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., & Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1), 67–72.

[63] Wigner, E. P. (1993). Characteristic vectors of bordered matrices with infinite dimensions i. In *The Collected Works of Eugene Paul Wigner* (pp. 524–540). Springer.

[64] Wu, F.-Y. (1982). The potts model. *Reviews of modern physics*, 54(1), 235.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.