

Quantitative methods for reconstructing
protein-protein interaction histories

Ryan Topping

Imperial College London, Department of Life Sciences

PhD Thesis

July 27, 2013

© 2013 Ryan Topping

All rights reserved

Typeset by L^AT_EX

This report is the result of my own work.

No part of this dissertation has already been, or is currently being submitted by the author for any other degree or diploma or other qualification.

This dissertation does not exceed 100,000 words, excluding appendices, bibliography, footnotes, tables and equations. It does not contain more than 150 figures.

This work is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) grant and completed in the Division of Molecular Biosciences at Imperial College, London.

All trademarks used in this dissertation are acknowledged to be the property of their respective owners.

Acknowledgements

I would like to thank...

my supervisors Dr. John Pinney and Prof. Michael Stumpf,

all members of the Theoretical Systems Biology Group,

the funding and support of the BBSRC,

Philip Davidson for preparing the protein alignments and phylogenies in Chapters 4 and 5,

Dr. Dannie Durand for many helpful discussions regarding Chapters 4 and 5,

my family, for helping me through my education,

and finally,

Erica Webb for support of all kinds, throughout the production of this thesis.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Abstract

Protein-protein interactions (PPIs) are vital for the function of a cell and the evolution of these interactions produce much of the evolution of phenotype of an organism. However, as the evolutionary process cannot be observed, methods are required to infer evolution from existing data. An understanding of the resulting evolutionary relationships between species can then provide information for PPI prediction and function assignment.

This thesis further develops and applies the interaction tree method for modelling PPI evolution within and between protein families. In this approach, a phylogeny of the protein family/ies of interest is used to explicitly construct a history of duplication and speciation events. Given a model relating sequence change in this phylogeny to the probability of a rewiring event occurring, this method can then infer probabilities of interaction between the ancestral proteins described in the phylogeny.

It is shown that the method can be adapted to infer the evolution of PPIs within obligate protein complexes, using a large set of such complexes to validate this application. This approach is then applied to reconstruct the history of the proteasome complex, using x-ray crystallography structures of the complex as input, with validation to show its utility in predicting present day complexes for which we have no structural data.

The methodology is then adapted for application to transient PPIs. It is shown that the approach used in the previous chapter is inadequate here and a new scoring system is described based on a likelihood score of interaction. The predictive ability of this score is shown in predicting known two component systems in bacteria and its use in an interaction tree setting is demonstrated through inference of the

interaction history between the histidine kinase and response regulator proteins responsible for sporulation onset in a set of bacteria.

This thesis demonstrates that with suitable modifications the interaction tree approach is widely applicable to modelling PPI evolution and also, importantly, predicting existing PPIs. This demonstrates the need to incorporate phylogenetic data in to methods of predicting PPIs and gives some measure of the benefit in doing so.

Contents

1	Introduction	1
1.1	Protein interactions	2
1.1.1	What is a protein?	2
1.1.2	How do proteins interact?	4
1.1.3	Protein-protein interactions as networks	7
1.2	Protein evolution	9
1.2.1	How proteins evolve	9
1.2.2	Building protein alignments	12
1.2.3	Building phylogenies	15
1.3	Protein-protein interaction evolution	21
1.4	Why is Protein-protein interaction evolution important?	23
1.5	Studying Protein-protein Interaction Evolution Computationally	24
1.5.1	Network Level	25
1.5.2	Complex Level	32
1.5.3	Interaction Level	34
1.6	Phylogeny in Protein-protein Interaction Evolution	36
1.6.1	The Interaction Tree	37
2	Adapting the interaction tree for protein complexes	39

2.1	Introduction	39
2.2	Methods	41
2.2.1	Building an interaction tree	41
2.2.2	Scoring systems	46
2.2.3	Proteasome data	57
2.2.4	Clustering algorithm	59
2.2.5	Training set of paralog complexes	60
2.3	Results	62
2.3.1	Interaction Trees for Globular Proteins	62
2.3.2	Finding the Conditional Probability Distribution	63
2.3.3	Evaluating the models	68
2.3.4	Calibrating the Model	73
2.4	Discussion	74
2.5	Conclusion	75
3	Modelling PPI rewiring in protein complexes	78
3.1	Introduction	78
3.2	Methods	80
3.2.1	Structural and sequence data	80
3.2.2	Phylogeny building	83
3.2.3	Sequence reconstruction with PAML	88
3.3	Results	89
3.3.1	Proteasome phylogeny	89
3.3.2	Predicting present day structures	92
3.3.3	Reconstructing the history of protein complexes	100
3.4	Discussion	107
3.5	Conclusion	108

4	A method for predicting transient PPIs	110
4.1	Introduction	110
4.2	Methods	114
4.2.1	Training data	114
4.2.2	SCOTCH score	114
4.2.3	RPScore	116
4.2.4	MODELLER and FoldX energy	117
4.2.5	Likelihood Score	117
4.2.6	Test data	119
4.3	Results	120
4.3.1	Orphan Two Component Systems in <i>Bacillus subtilis</i>	125
4.3.2	Predicting Specificity Rewiring	126
4.3.3	Predicting Two-component Systems in 7 Bacterial Species	128
4.3.4	Effect of Alignment Quality on MILLscore	139
4.4	Discussion	140
4.5	Conclusion	143
4.6	Appendix : Full Predictions in 7 Bacterial Species	144
5	Reconstructing PPI history for transient interactions	152
5.1	Introduction	152
5.2	Methods	155
5.2.1	Two component system sequences	155
5.2.2	Phylogenetic trees	155
5.3	Results	157
5.3.1	Predicting present day PPIs	160
5.3.2	Predicting Clostridial PPIs from a Bacillus	167
5.3.3	Predicting ancestral PPIs	173

5.4	Discussion	178
5.5	Conclusion	179
6	Discussion and conclusions	181
6.1	Advantages of the approach	181
6.2	PPI evolution in obligate complexes	183
6.3	Transient PPI evolution	186

Chapter 1

Introduction

The aim of this thesis is to explore computational methods for modelling protein-protein interaction evolution and to show the applicability of these methods to a range of data. As will be set out in this chapter, proteins are the most common class of macromolecule in the cell and are involved in all major functional processes within the cell. However, proteins do not function alone but work in concert, forming larger permanent protein machines and fleeting transient complexes. Therefore, these interactions are vital to the functioning of an organism and represent a link from the genotype to phenotype.

This chapter begins with some basic background, covering what a protein is and how proteins evolve as a result of DNA mutation. Methods and approaches for modelling the evolution of individual proteins are then described as these methods become central to the later methods for modelling interactions between several proteins. Finally, the chapter ends with a description of how protein evolution leads to change in interaction partners over time and a motivation for why understanding of this process is important.

1.1 Protein interactions

1.1.1 What is a protein?

Proteins are macromolecules found within every living cell. These macromolecules are formed of linear chains of connected amino acids (polypeptide chains), the content and ordering of which is referred to as the proteins sequence or its primary structure. Varying protein sequences are produced by a cell in order to produce varied proteins of differing length and of differing utility to the cell. Proteins are often described as miniature machines that perform the vital functions required for the survival and propagation of the cell [1]. For instance, during DNA replication several proteins come together to form the replisome, an organic machine that unwinds and copies the DNA, the separate protein subunits moving and performing the task in concert [2]. The diversity of function of proteins in a cell is vast, being responsible for such disparate tasks as DNA replication, transcription, signal transduction, structure and protein degradation. To see how different protein sequences, built from the same amino acid building blocks, can be responsible for such a diverse set of functions, we have to consider what happens to the protein inside the cell. The proteins produced do not exist within the cell as formless chains of amino acids but curl up and pack in to a set 3D structure (Figure 1.1).

This process is known as protein folding and produces macromolecules of complex 3D structure determined by the sequence of amino acids constituting the protein. Within these varied structures we can identify local substructures which are found in proteins from all kingdoms of life, the most prevalent being alpha helices; right handed helices formed by the polypeptide chain and beta sheets; parallel arrangements of sections of the chain in to sheet-like structures. Both of these substructures are the result of hydrogen bonding between non-adjacent amino acids in the chain. These recurring substructures and their arrangement

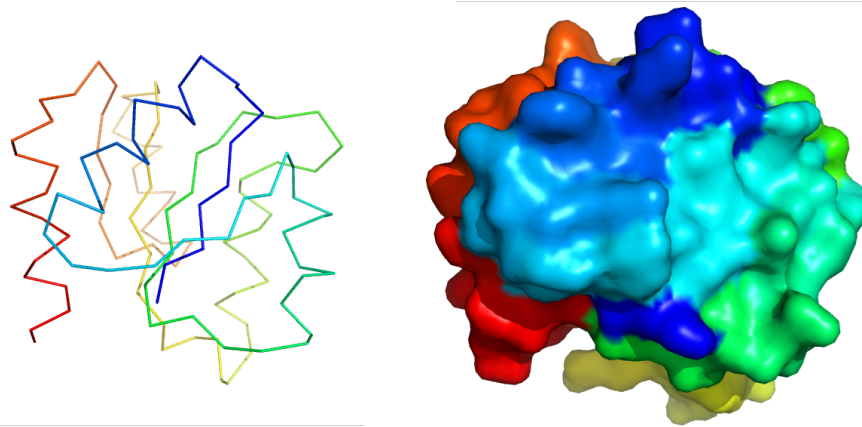


Figure 1.1: A folded protein. On the left the line traces the shape of the polypeptide chain in the folded protein coloured from red to blue along the length of the chain. On the right is shown the surface of the protein, defined by the surface area accessible by a 1.4Å probe

along the polypeptide chain are known as the protein's secondary structure and the overall 3D structure of the whole, folded protein is its tertiary structure.

It is a protein's detailed tertiary structure that allows it to perform a specific function for instance, the proteasome is a complex formed of many proteins in a barrel shape that is responsible for degrading proteins that are misfolded or no longer needed. To do this it brings the protein to be degraded into the barrel and disassembles it. It is the shape of the barrel structure and the positioning of the amino acids responsible for disassembly that allow the proteasome to perform this function. Or to give another example, actin is a protein that can assemble with many copies of itself to form a chain. These actin filaments are able to support the shape of cells like a macromolecular scaffold, because their long, thin shape allows them to form structural supports between different parts of the cell.

In these two examples the proteins are not working alone in order to perform a function but instead binding together with other proteins to perform complexes

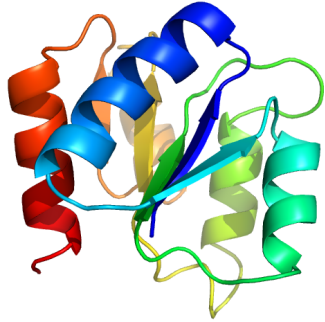


Figure 1.2: The same protein as shown in Figure 1.2, shown in a cartoon representation with alpha helices shown as wide helices and beta sheets as parallel arrows. The folded protein contains four helices surrounding a beta sheet.

of many proteins capable of performing some task. This is true in general and the majority of proteins function in concert with others. The event of two proteins binding together to form larger assemblies like this is the result of non-covalent binding between amino acids of either protein, at the protein-protein interface (Figure 1.3). Such events are called protein-protein interactions (PPI) and, being central to understanding how proteins perform cellular function, PPIs have been the subject of much study

1.1.2 How do proteins interact?

Not only are PPIs important for the examples in the last section but in fact, most of the biological processes within a cell are reliant on PPIs. Indeed, perturbation of PPIs has been implicated in several diseases for instance Huntington's disease [3], prion diseases [4], sickle cell anemia [5] and cancer [6]. The interactions between proteins can be categorised very broadly into two types. The first type are permanent interactions, in which the proteins form a complex and remain so for the

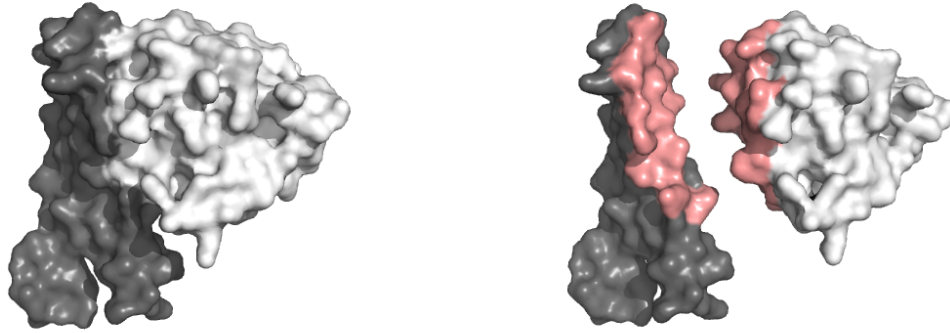


Figure 1.3: Two proteins can bind together to form a complex (left), this is the result of non-covalent binding between residues in the binding site of each protein (coloured pink, right).

duration of their existence. This type of interaction can be found when proteins assemble into some larger macromolecular machine such as the replisome or the proteasome. The function of these machines is impossible unless the proteins are complexed and so the proteins remain bound in order to perform this function.

The second class of PPI is the transient interactions, in which the proteins associate and disassociate repeatedly to perform some function. For example, in a signal transduction pathway proteins often bind temporarily in order to pass the signal, in the form of a phosphatase group, to the next protein in the cascade. In this case the two proteins in question can be found in bound and unbound states dependent on the signal in the system.

Several differences have been noted in the nature of these two classes of interaction: the contact area of the interaction tends to be larger and more hydrophobic in obligate interactions [7], with the residues mediating the interaction tending to evolve more slowly [8]. However, the fundamental physicochemical laws govern-

ing PPIs are the same for both types of interaction. For any physical process to spontaneously occur there must be a negative change in free energy as a result, the change in free energy being calculated according to

$$\Delta G = \Delta H - T\Delta S \quad (1.1)$$

with G equal to free energy, H enthalpy, S entropy and T the temperature in kelvin. The association of two objects to form one larger object, as when proteins bind to form a complex, represents a decrease in entropy and therefore contributes positively to the change in free energy. This positive contribution must be overcome for two proteins to bind and is counterbalanced by the increased entropy of the solvent after binding and the reduction in enthalpy as a result of salt bridges, hydrogen bonds and van der Waals interactions between residues across the protein-protein interface [9]. This second factor is made possible by the complementary nature of the amino acid composition and shape of the constituent proteins across the interface, for instance in matching hydrogen bond donor with hydrogen bond acceptor residues at the interface, allowing for energetically favourable hydrogen bonding. Thus when examining X-ray crystallography structures of PPIs it is often seen that complementary residues allowing such favourable interactions are found paired at the protein-protein interface [10] [11], [12]. It is these residue level interactions that cause two proteins to become bound in both permanent and transient interactions. Once bound the change in free energy determines the strength of the binding and thus how long the proteins remain complexed, resulting in interactions of differing permanence.

1.1.3 Protein-protein interactions as networks

Traditionally reductionist biology attempts to elucidate knowledge of cellular function by studying isolated parts of the cell in detail, for instance in the study of PPIs by examining in detail the structure of one interaction or by studying in detail one linear signalling pathway. More recently the rise of systems biology has encouraged study of the cell at the systems level, that is, by attempting to consider many parts of the cell at once and the relationships between these parts in order to get insights into how the system works as a whole to generate the observed phenotype (e.g. [13]). As applied to PPIs, this has led to the study of protein interaction networks (PINs).

A PIN is a graph in which nodes represent proteins and the edges between the nodes represent interactions between them, giving a formalised way of describing the interactions amongst a set of proteins (Figure 1.4). An attempt has been made in several organisms to experimentally find all PPIs within the proteome and to represent these datasets as PINs. This has allowed a study of the global properties of these networks, leading to insights into the organisation of cells that would not be possible in studying interactions in isolation. For instance, experiments to survey PINs in several organisms have looked at the distribution of the number of interaction partners across proteins, finding that the distribution has a long tail with some proteins having a very large number of interaction partners. These proteins have been termed hubs and several studies have attempted to study their particular properties, finding that these hubs are disproportionately essential for the integrity of the overall network [14] [15]. These particular insights would not have been possible without a systems level analysis. Of course such a large scale treatment has its inherent problems, it has been noted that PIN datasets can have poor accuracy, low coverage and present a sampling problem that has to

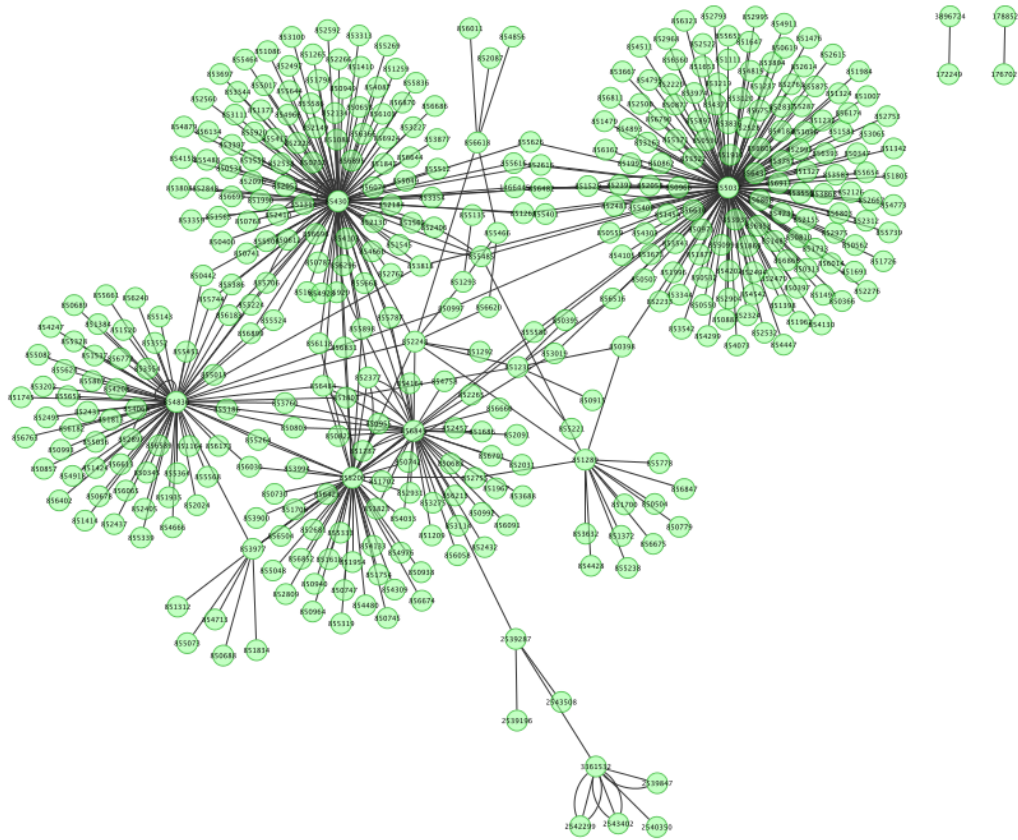


Figure 1.4: A PIN representation of the interactions between a set of proteins, nodes represent proteins and edge between nodes represent binding events.

be considered in any analysis [16]. These factors affect the conclusions that can be made from these datasets. There is also the problem that PIN generated in *in vitro* assays may not represent the true functional interactions in the cell as, *in vivo*, the proteins may be in separate cell compartments or have differential expression, preventing the biochemically possible interaction from ever occurring. This presents problems in interpreting PINs biologically, although recognition of this problem has led to approaches that attempt to produce biologically relevant PINs, for instance in producing specific PINs for each stage of the cell cycle in yeast [17].

1.2 Protein evolution

Of course proteins and their interactions are constantly evolving and are not static in time. Before considering the effect of evolution on PPIs, the evolution of individual proteins is described. An understanding of individual protein evolution and the methods used to study this process will become vital later in the thesis in relation to modelling PPI evolution as many of the ideas and algorithms are co-opted for use.

1.2.1 How proteins evolve

Proteins are forever evolving in organisms producing the variety of polypeptide sequences observed within and between species. This evolution begins with random changes in the DNA sequence including single nucleotide substitutions, insertions/deletions and gene duplications of the DNA. In the case of substitutions, a single character in the genome is changed, if this change occurs within a coding region that will be transcribed and translated to protein then the mutation can belong to one of two classes: synonymous mutations which do not change

the resulting polypeptide sequence, as defined by the genetic code (Table 1.1), or non-synonymous mutations which lead to a change in amino acid produced from the tri-nucleotide that the mutation occurs in. For instance, if a gene contains a CAT codon, this set of nucleotides will result in a histidine residue added to the polypeptide chain by the ribosome during translation. If the final nucleotide mutates to a C, a histidine is still produced (CAC). However, if the mutation produces CAA, a glutamine will result and so the final protein sequence of the gene is changed. It is these non-synonymous changes that lead to evolution of the proteins primary structure.

1st	2nd								3rd
	T		C		A		G		
T	TTT	Phenylalanine	TCT	Serine	TAT	Tyrosine	TGT	Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	Leucine	TCA		TAA	Stop	TGA	Stop	A
	TTG		TCG		TAG	Stop	TGG	Tryptophan	G
C	CTT	Leucine	CCT	Proline	CAT	Histidine	CGT	Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	CGA	A		
	CTG		CCG		CAG	CGG	G		
A	ATT	Isoleucine	ACT	Threonine	AAT	Asparagine	AGT	Serine	T
	ATC		ACC		ACC		AGC		C
	ATA		ACA		AAA	Lysine	AGA	Arginine	A
	ATG	Methionine	ACG		AAG		AGG		G
G	GTT	Valine	GCT	Alanine	GAT	Aspartic acid	GTT	Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Table 1.1: The genetic code determines how DNA triplets are translated in to amino acids by the cellular machinery.

Non-synonymous substitutions lead to changes in the resulting protein; as the amino acid sequence changes, the structure and physicochemical characteristics of

the protein change and so its function can be altered. The change of phenotype brought about can become the target of selection and overtime this evolutionary process leads to diversity of protein.

It is also possible for base pairs to be inserted or removed from DNA via mutation, these processes being called insertions and deletions respectively or indels collectively. If indels occur in protein coding sequence the effect on the resulting translated protein can be more dramatic. Specifically, if the length of the indel is not a multiple of three, then the translation of the protein will be put out of phase leading to potentially different codons and therefore amino acids at every position downstream of the indel. This will most likely lead to a non-functional protein. If however, the indel is of length multiple of three, then the mutation produces an insertion/deletion of amino acids at that position in the final protein, potentially altering its function.

A third important process driving protein evolution is that of gene duplication, wherein the length of DNA coding a protein is duplicated leading to two identical copies in the genome. Post-duplication, mutations can accrue in the duplicate genes independently, producing two proteins with different functions. Two broad scenarios have been proposed for how this differing function is produced. In the first scenario, termed neofunctionalisation, the presence of one duplicate gene relaxes the selective pressure on the other to perform its function allowing previously deleterious mutations to occur. This leads to a protein with some new function that proves useful to the organism and could not have evolved without the presence of a duplicate gene. The second scenario, subfunctionalisation, describes a situation in which the original function of the gene is split in to several parts or roles (role a and role b say) and the presence of a duplicate allows each gene to lose its ability to perform all roles and to instead perform those complementary to its duplicate i.e. duplicate 1 only performs role a, duplicate 2 only performs role

b.

Together, these related processes produce the variety of protein sequences that we can observe. Within these sequences we can identify protein families; groups of proteins related across species that have a detectable sequence similarity. These families are enlarged by gene duplication events, shrunk by gene loss events and their variation is produced through mutation. Members of a protein family often perform similar functions and have similar 3D structures once folded. Given such a family of related protein sequences the starting point in understanding how evolutionary processes created the family is to identify the substitutions, indels and gene duplications that led to its existence. Methods for doing so are briefly described in the next sections.

1.2.2 Building protein alignments

To discover the history of substitutions and indels between two related protein sequences it is necessary to identify equivalent positions in the sequences i.e. positions that are derived from the same position in the common ancestor of the two proteins, a task complicated by the fact that the residues at these positions may now be different as a result of substitution. Having defined the equivalent characters in the two sequences, indels then correspond to the characters with no equivalent in the other sequence (although with only two sequences it will not be clear if this is a result of an insertion in one sequence or a deletion in the other). The resulting representation of the relationship is called an alignment and can be represented by presenting aligned equivalent positions and using a gap character, typically "-", as shown in Figure 1.5.

Alignments can be produced for large sets of related sequences, this extension of pairwise alignment is called multiple sequence alignment, providing useful in-



Figure 1.5: Aligned protein sequences, equivalent characters are aligned vertically and a gap character shows the believed position of indels.

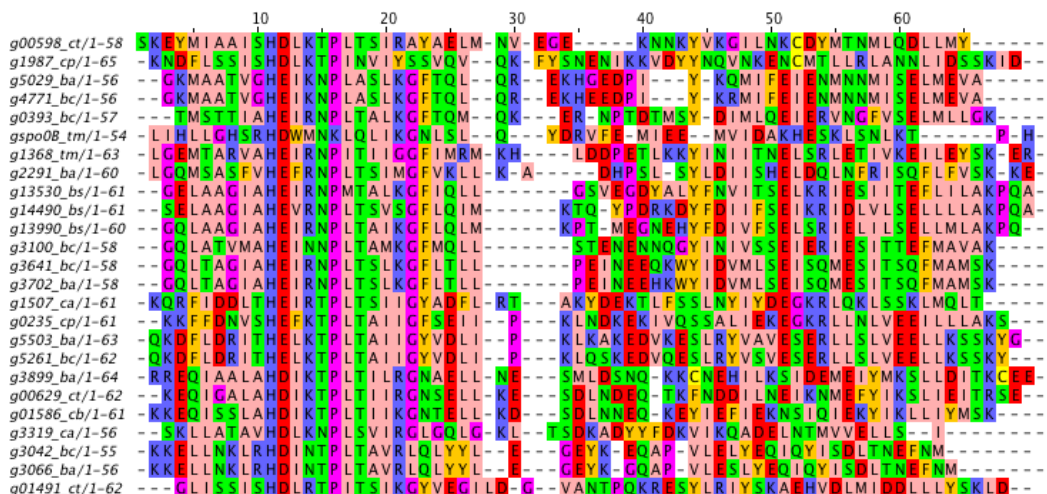


Figure 1.6: An example of a protein alignment. Residues are coloured according to their chemical characteristics and the gap character "-" is used to show the position of indels. Residues conserved across an alignment are often important for the functioning of the protein and in this example, the conserved histidine at position 11 (blue column to the left) is a phosphorylation site vital to the function of this group of related proteins.

formation about the proteins. For example, if the proteins are known to perform some conserved function then the parts of the sequences that are highly conserved could relate to this function, for instance constituting a binding site or phosphorylation site, as shown in Figure 1.6. Alignments can also be used to build profiles and models of sequences from the family represented by the alignment, as is done in the PFAM database for instance [18]. These can then be used to search protein sequence databases for new sequences that have a significant probability to belong to the family, proving important in genome annotation.

Several computational approaches exist for producing protein alignments. One such approach uses dynamic programming along with a substitution matrix to produce optimal alignments between proteins. The substitution matrix is a 20 by 20 matrix with the (i, j) entry encoding our knowledge of the likelihood that residue type i is substituted for residue type j during evolution, based on some model of amino acid substitution [19] [20] [21]. These models are usually empirical and based on alignments of closely related sequences in which we can be sure of the equivalency of positions. The resulting matrix assigns a cost for each possible transition between one amino acid to another, as for instance in BLOSUM matrices [22], and show that physicochemically similar amino acids are more likely to be substituted for each other during evolution. The Needleman-Wunsch dynamic programming algorithm [23] takes this matrix and a given penalty for introducing indels in to the alignment and is guaranteed to produce an optimally scoring alignment of the full length of the sequences. Alternatively, the related algorithm of Smith-Waterman [24] can be used to find optimally scoring alignments of subsequences within the proteins. Despite the attractive quality of guaranteed optimality, this class of algorithm proves prohibitively expensive computationally for aligning large groups of sequences.

The advent of large amounts of sequencing data has seen the development of

a range of suboptimal alignment algorithms that avoid the computational expense of dynamic programming. The most popular of these is the heuristic BLAST [25] algorithm which finds local alignments of two sequences by first finding matching short subsequences, once again using a substitution matrix, and building alignments from these. Whilst not guaranteed to give an optimally scoring result, the drastically reduced runtime of this algorithm makes it suitable for searching whole proteome sets, for example when searching for homologs of a given protein. There now exist a huge variety of sequence alignment algorithms, many are designed for use in specific situations, for example PSI-BLAST [25] uses the results of a BLAST search to build a profile and then iteratively looks for extra related sequences, this can be used to find distantly related sequences or 3DCoffee [26], which uses structural information from related proteins to improve the quality of the alignment. Suffice to say that whichever algorithm is used, no method is infallible; to produce the highest confidence alignments manual curation and expertise is often required. For instance, automated alignment may fail to align residues that are known to be functionally important, such as known phosphorylation sites, especially if pairwise alignment rather than multiple alignment is being used. In this situation, manual edit would be required to align the positions correctly.

1.2.3 Building phylogenies

Having described how protein sequence alignments can be used to identify substitutions and indels, methods will now be described for inferring the other process important to protein evolution mentioned above: gene duplication. Sequence alignments give us a snapshot of existing proteins but to infer duplication events, the history of the sequences must be inferred. This history consists of a sequence of speciation and gene duplication events that produce a branching tree of sequence

evolution, with the leaf nodes representing the sequences in the alignment, as shown in Figure 1.7. This tree representation is called a phylogeny, nodes in this tree represent proteins, whether ancestral or extant and edges represent evolution. Several algorithms exist to infer the topology of this tree based on the sequence alignment.

The most simple methods for reconstructing protein phylogenies are the distance based methods. To begin, pairwise distances between all protein sequences are computed. Often these distances are estimated based on a model of amino acid substitution and represent an estimate of the number of substitutions occurring between two proteins. Given the matrix of distances between proteins, some algorithm is then applied to produce the phylogenetic tree. Perhaps the most popular algorithm to do so is the Neighbour-Joining (NJ) algorithm [27] which looks for pairs of nodes that have a small distance between each other and a large distance between themselves and the rest of the nodes, this pairing is then joined to form a clade. This is repeated until a bifurcating tree is obtained. Whilst this method is very fast its accuracy is sometimes poor.

To this end, more sophisticated methods have been developed that take into account the differences at all sites in the protein alignment and attempt to find a tree that best explains them, as an alternative to distance based methods which only consider a whole sequence distance between proteins. The simplest of these type of methods are parsimony methods based on the idea of least evolution. These methods aim to reconstruct a phylogeny that requires a minimum of evolutionary change to explain the variation in the sequences. This approach began with [28], with an algorithm to calculate the least number of substitutions required to produce a set of sequences given a tree topology. This was later expanded to weighted parsimony in which a cost matrix, such as BLOSUM used for protein alignment, is used to penalise improbable substitutions [29]. Reconstruction proceeds by search-

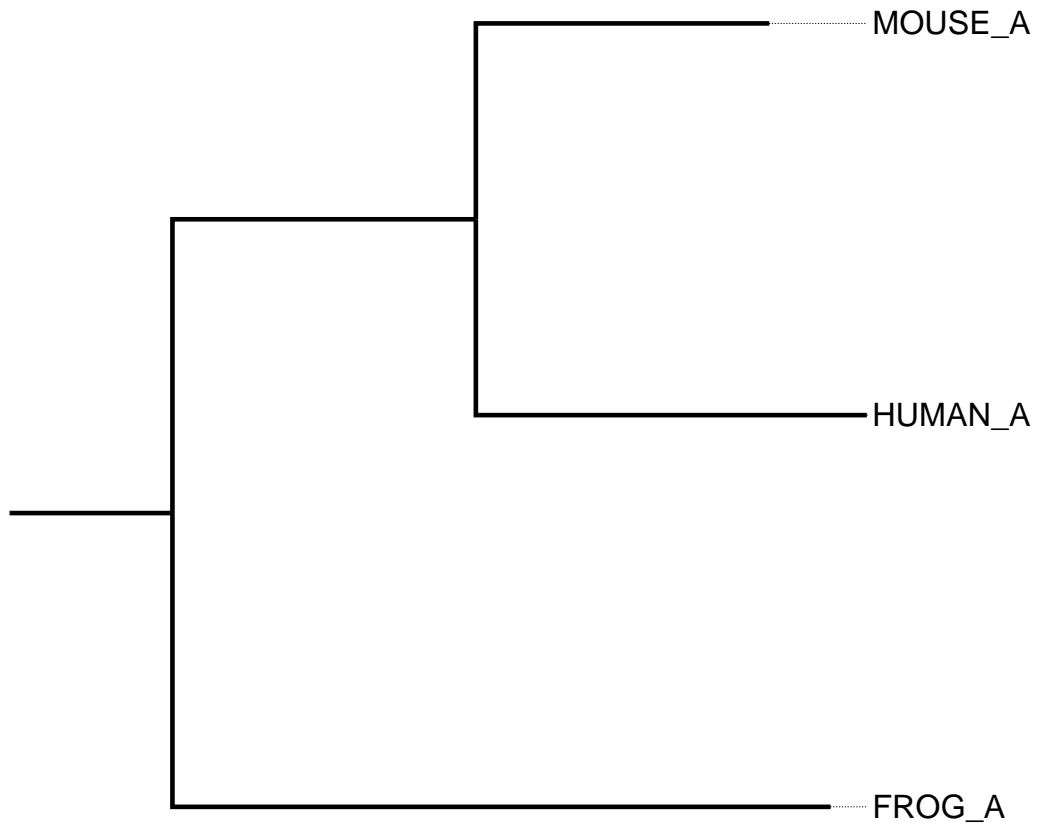


Figure 1.7: An example of a protein phylogeny. The leaves of the tree represent the existing proteins and the branching of the tree represents the evolutionary history of the sequences. The tree shows that the mouse and human proteins diverged recently whereas the frog protein is more distantly related.

ing through possible tree topologies, calculating the minimum required change and then finally selecting the tree requiring the least minimum change as the estimate for the phylogeny. Although parsimony has several benefits, such as the ability to reconstruct ancestral states, its simplicity can lead to documented inaccuracies in reconstruction [30].

Maximum Likelihood (ML) approaches are more sophisticated and can avoid some of these inaccuracies. Here a model of amino acid substitution is used to define, for each site, the probability of observing the amino acid composition at that site given a tree topology. This can be done by summing over all possible ancestral states in the tree at that site (although in practice methods are used that avoid calculating this entire sum) and if the sites are independent in the model then the probability for observing the entire alignment can be obtained by multiplying this probability across all sites. This quantity, known as a likelihood, is the probability of observing the alignment given the tree topology and the model of amino acid substitution. ML reconstruction searches through tree space and attempts to find a tree that maximises the likelihood.

The final approach to phylogenetic reconstruction to be discussed is the Bayesian approach, first introduced in [31]. Here Markov Chain Monte Carlo (MCMC) algorithms are used to compute posterior probabilities for a set of parameters given the sequence alignment. These parameters can include those used in the amino acid substitution model and those describing the topology and branch lengths of the phylogeny. Probably the most widely used of these approaches is the Mr. Bayes package [32] which implements several models of amino acid substitution as well as using methods for minimising tree search time to improve the convergence of the MCMC algorithm. One benefit of the Bayesian approach is that the output of a probability of a tree given the alignment avoids the need for bootstrapping to produce confidence estimates as required by other methods, however, the MCMC

algorithm needs to run for an extended time to converge for large datasets and so whilst convenient and easy to interpret, this benefit may not necessarily save time.

Given a reconstructed phylogeny for a protein sequence alignment we are almost in a position to identify the gene duplications that lead to the observed variation. Each interior node in the tree represents the diversion of one sequence into two. This occurs when a gene is duplicated but also when a speciation event occurs as defined by the species tree. If the protein phylogeny has the same topology as the species tree then no duplications have occurred and all interior nodes are speciation nodes. However more complicated situations will need an algorithmic approach to decide which nodes are duplications. An example of such a situation is given in Figure 1.8. This process of assigning interior nodes as duplication or speciation nodes is known as tree reconciliation and was first attempted in [33]. Several methods exist for performing reconciliation including parsimony like methods that seek to minimise the amount of duplications and/or losses [34], ML and Bayesian approaches [35]. Once we have a reconciled tree we have an estimate for the sequence of gene duplications that lead to a diverse set of proteins. This has the benefit of giving precise orthology relationships between the proteins i.e. orthologs are related by speciation events and represent equivalent proteins in different species whereas paralogs are related by duplication events. This distinction is known to be important in determining the function of homologous proteins, for instance see [36].

The ability to construct phylogenies is a powerful tool for understanding biology. On face value, it is a method for reconstructing the branching that led to present day, observable sequences, this has for instance allowed attempts to reconstruct species divergence [37] or reanalyse accepted taxonomy [38]. The information contained in phylogenies has also proved useful in other areas, for instance, the similarity of phylogenetic trees has been used fairly extensively to predict PPIs af-

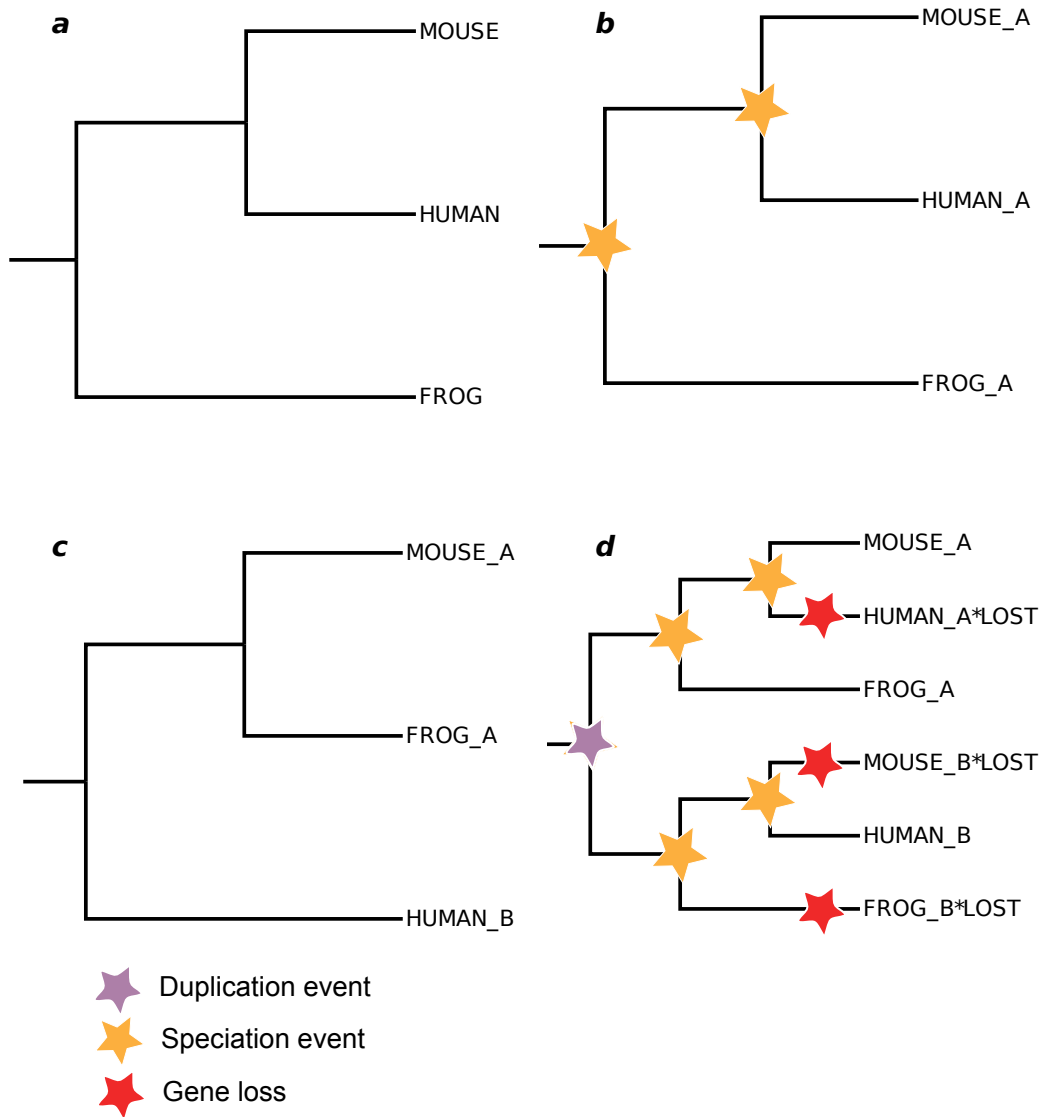


Figure 1.8: Gene tree reconciliation: Given a species tree (**a**), the aim of gene tree reconciliation is to decide if internal gene tree nodes represent speciation or duplication events. If the gene tree has the same topology as the species tree then all internal branching is a result of speciation (**b**). If the gene tree disagrees with the species tree (**c**), then a more complicated history is required to explain the disagreement (**d**). This includes prediction of genes that have been lost, these genes are suffixed *LOST in (**d**).

ter the observation that interacting proteins tend to belong to families with similar phylogenetic trees [39].

Having described the processes by which proteins evolve and the computational methods for analysing that evolution. Focus now moves to how PPIs can be said to evolve, to be followed by a discussion of the methods used for analysing such evolution.

1.3 Protein-protein interaction evolution

Having described the evolution of individual proteins and the tools for describing the process, the evolution of the interactions between proteins is considered. Proteins are able to interact due to the physicochemical and shape complementarity of the interacting proteins [10] [11] [12] which allows energetically favourable interactions between residues across the protein-protein interface e.g. acidic and basic residues forming salt bridges. As covered in section 1.2.1, protein evolution begins with mutations in the coding DNA that change the resulting amino acid sequence. These changes can then alter the binding affinity between the protein and its interaction partners. For instance, a protein contains a tyrosine residue, coded by a TAT codon, that forms a hydrogen bond with an asparagine residue in its interaction partner. A mutation occurs that changes the TAT for a TTT codon, resulting in a change from tyrosine to phenylalanine, which is not capable of forming hydrogen bonds. The loss of the energetically favourable pairing of residues at the interface of the two proteins could destabilise the interaction and reduce the probability of observing the proteins interacting, perhaps completely. If this leads to the loss of the PPI then the phenotype of the cell will be altered, for instance this could remove a link in a signalling pathway. This variation can become acted upon by natural selection. If the new phenotype has higher fitness,

a selective sweep can occur leading to a removal of the PPI completely from the population. Alternatively, for two non-interacting proteins, substitutions could change the amino acid sequences and resulting tertiary structures sufficiently to produce a *de novo* interaction between them. These two opposite processes of PPI gain and loss have been termed rewiring events and the relative occurrence of both PPI gains and losses has been studied and debated [40] [41] [42].

There is also another route by which PPIs can evolve. For two proteins to interact they must be expressed at the same time and the same place in the cell and so changes in expression patterns could lead to a given PPI being unobservable *in vivo*. For instance, mutations in the transcription factor binding site governing expression of one of the proteins could lead to recruitment of a different transcription factor, resulting in expression in some new tissue or cell compartment. Attempts have been made to quantify the effects of these types of changes (e.g. [43]) but this process is not the focus of this thesis.

An important consequence of the process of amino acid substitutions at the interface of a PPI is the emergence of correlated mutations. Often, mutations at the PPI interface that disrupt binding affinity will be removed from the population by selection [8], however, a compensating mutation can occur subsequently in the opposite protein, maintaining the affinity; the binding affinity may even increase. This will lead to correlated changes in the protein sequences at these sites, termed coevolution. These correlations are detectable [44] and can predict functionally important pairings of residues across the PPI interface. This has proved valuable for predicting which amino acids are physically close to each other for use in structure prediction [45] and for finding specificity determining residues [46].

1.4 Why is Protein-protein interaction evolution important?

Before methods for studying PPI evolution are discussed in the next section, it is firstly worth motivating interest in such methods with an argument of the importance and utility of studying PPI evolution. Proteins and their interactions within cells are responsible for the functioning and therefore the phenotype of the organism. Therefore, purely for the fact of understanding how organisms evolve, we need to understand how PPIs evolve, as this will explain the evolution of phenotype. To give a concrete example, it is thought that much of the diversity in eukaryote proteomes is a result of gene duplication followed by divergent evolution of the duplicates [47]. As mentioned in Section 1.2.1, after duplication it has been proposed that a protein's function can become subfunctionalised or a new function can appear. Understanding the relative importance of these processes will tell us if phenotypic roles in the cell tend to be shared among duplicates or if duplication drives innovation and attempts to deduce the relative importance of each scenario have been made, for example in [48]. Study of PPI evolution, particularly the behaviour after duplication events, can answer these basic biological questions.

In terms of applications, understanding PPI evolution is important in that it allows us to transfer knowledge between species by exploiting the evolutionary links between proteins. In the common ancestor of two species, proteins existed that interacted, forming complexes that performed the functions necessary for the survival of that ancestral organism. After the speciation event the two lineages leading to the present day species began evolving independently and their proteomes and interaction networks began to look different as new proteins and complexes appeared, evolving to perform new functions. However, there will exist conservation between the species as parts of the PIN are kept or co-opted to

perform some new function. Understanding and subsequently modelling the process that leads to this conservation will allow knowledge of PPIs to be transferred between species, for instance in using some knowledge of the PIN of a species to predict the PPIs in a related species.

To give a recent, concrete example of such reasoning, in [49] a group of interactions conserved between yeast and humans was identified (in this case genetic interactions not physical PPIs). The interactions were responsible for maintaining the cell wall in yeast but had evolved to regulate blood vessel growth in vertebrates. Blocking blood vessel growth is one tactic for fighting tumor growth and it was found that a drug previously used to perturb the interactions in yeast could be used to slow xenograft tumor growth by preventing blood vessel formation.

This is a good example of how an evolutionary approach can provide insights relevant to understanding of disease by integrating knowledge across species. In this particular application the modelling of evolution was limited to finding a conserved subnetwork within two interaction networks. More complex approaches are possible that incorporate more of the knowledge of PPI evolution. In the next chapter approaches and methods for studying PPI evolution are described and their various applications and limitations are discussed.

1.5 Studying Protein-protein Interaction Evolution Computationally

As set out above, understanding PPI evolution is essential for understanding the emergence of phenotype in biological systems [50] and can also be leveraged to make predictions of present day PPIs [51]. However, because it is impossible to measure the PPIs present in extinct species, designing experiments to test hy-

potheses of PPI evolution is often infeasible. As such, this is an ideal situation in which computational approaches can be used to infer the processes shaping PPI evolution. The following section describes some of the key research in this area at the level of whole protein interaction networks, protein complexes and, finally, individual interactions.

1.5.1 Network Level

As explained in Section 1.1.3, with the rise of systems biology has come a rise in popularity of describing large sets of PPIs as networks of interactions. These networks consist of nodes, representing proteins, and edges between the nodes, representing PPIs between the proteins. The set of proteins in a PIN may consist of, for instance, the complete proteome of a species and the edges some set of measured interactions, for instance [52] was a first attempt to measure PPIs genome-wide in yeast, producing a PIN with 3,278 nodes and 4,549 edges. There have also been attempts to combine several datasets to produce networks of high confidence interactions. The ultimate goal in producing such datasets is to produce a global picture of the set of possible PPIs within an organism, often regardless of the occurrence of the PPIs *in vivo*.

Given the large PIN datasets available for several organisms, for instance [52], [53], [54], [55], those interested in PPI evolution would like to know how these sets of interactions evolved from their most recent common ancestor (MRCA). Full knowledge of this process would include the history of gene duplication and loss relating the proteins in the PIN of the MRCA to those in the modern day PINS (i.e. phylogeny, as described in section 1.2.3) along with the interaction rewiring events that led to the differing sets of interactions in the existing species. Obtaining such a description is a hard problem and much of the focus in the research literature

has been on identifying the conserved regions amongst PIN datasets, i.e. finding the equivalent interactions that have remained in all the PINs considered since the MRCA. This problem is somewhat analogous to biological sequence alignment in that it is looking for evolutionary conservation but computationally more complex due to the non-linearity of PINs [56].

This approach is called network alignment (Figure 1.9) and as described above, has the aim of finding conserved interactions in a group of two or more PINs. To begin looking for conserved interactions, most algorithms first use sequence alignment, e.g. BLAST, to look for homologous proteins across the PINs. Having defined the protein orthologs algorithms then begin to look for interactions that are conserved among orthologous pairs in each network. To give a simple example, [57] looked for conserved PPIs across species by first finding the best BLAST hit of each protein and then looking for interactions between two proteins in one PIN for which their best hits interact in another PIN. The conserved interactions are termed interlogs and are assumed to represent PPIs that have been retained since the MRCA.

More complicated algorithms can detect conserved groups of interactions as opposed to single conserved interactions. In an early example, Kelly et al [58] used a network alignment approach to look for conserved PPI pathways between *Saccharomyces cerevisiae* and *Helicobacter pylori*, that is conserved linear paths in the PIN such as A interacts with B interacts with C interacts with D. The algorithm, named PATHBLAST, uses BLAST searches to find putative homologs between the two PINs and then uses a dynamic programming algorithm to find conserved linear paths between homologs in each PIN. Due to the large divergence time between the two species and the low coverage of the PINs, only 7 direct interlogs were found. The alignment algorithm found these as part of 5 larger conserved pathways by allowing for some missing conservation in a pathway. This approach

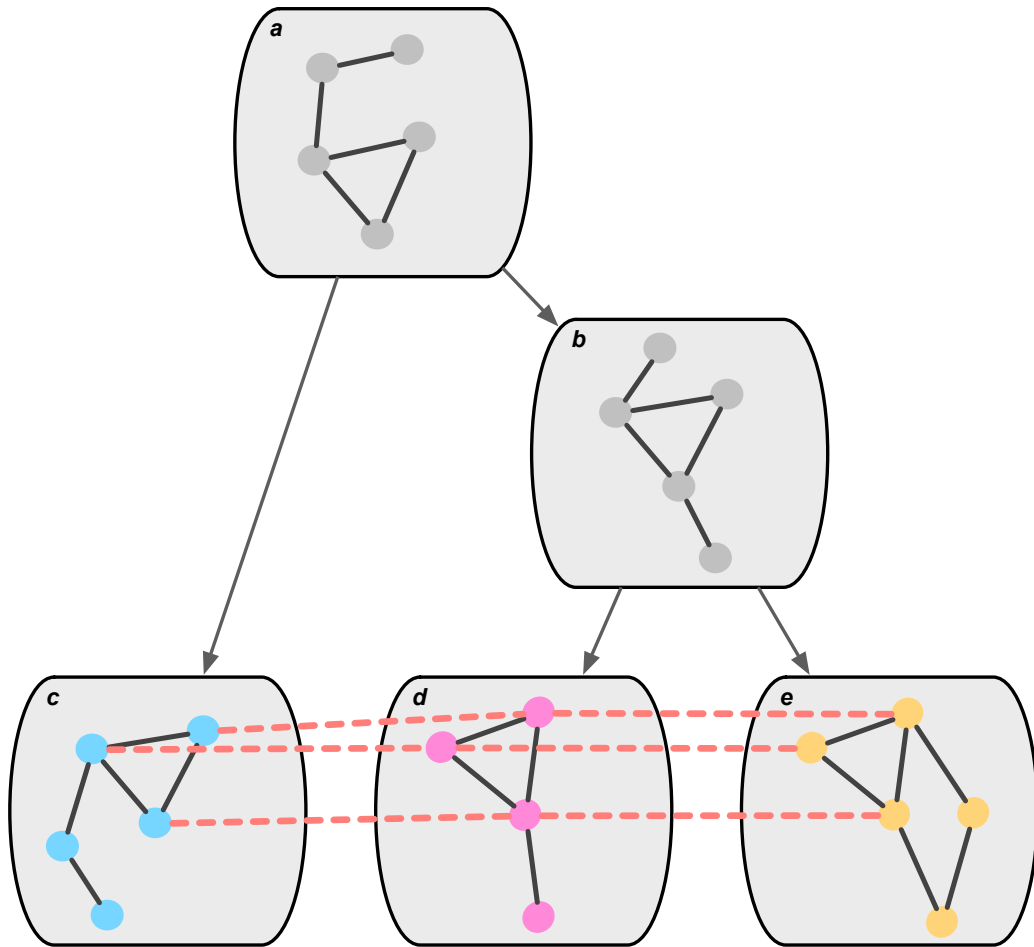


Figure 1.9: Schematic of network alignment applied to PINs. Starting with some ancestral network **a**, PPI evolution proceeds by gene duplication/loss and interaction gain/loss to produce the diverse PINs seen in existing species **c**, **d**, **e**. A complete model of this process would include the phylogeny of gene duplication and loss events leading to the proteins observed in each species (not shown), along with the history of interaction gains and losses. Network alignment does not give such a model but attempts to find conserved parts of the PINs that have remained intact throughout the evolutionary process (red lines).

was later extended [59] to look for conserved, densely interacting subgraphs of the same PINs studied above. It is known that groups of proteins in PINs that share many interactions amongst themselves often correspond to protein complexes and so the approach is applied to find 11 conserved complexes between the two PINs. This kind of approach can suggest functions for proteins predicted as part of a complex and can be used to predict complexes outright based on the results but, in terms of PPI evolution, it is limited in what it can tell about the events that led to the observed pattern of conservation as it does not attempt to model this process.

Many other attempts have been made at the network alignment problem, for instance [60], [61] [62], [63] [51] [56] [64]. These algorithms differ in the way they measure similarity between nodes in the separate PINs and how conserved edges are extracted. However, the focus of application has been overwhelmingly on identification of conserved complexes based on comparison of PINs. In terms of PPI evolution, this can give indication as to the timescales that PPIs can be conserved over, for instance given the results in [59] it is clear the PPIs can be conserved over large timescales, given the similarities found between a eukaryotic and bacterial species. However, network alignment is ill equipped to answer the more detailed questions concerning PPI evolution. For instance, with the low coverage and often poor quality of PIN datasets [65], [66], it is hard to infer rates of PPI loss given the amount of conservation between species as a lack of conservation may simply be because an interaction was not tested or gave a false negative in one species.

Besides the issue of data quality, the network alignment approach is essentially a comparison of snapshots during PIN evolution that can show which parts of a network have remained unchanged but cannot reconstruct the intermediate PINs that led from the ancestor to the present day PINs, nor the process of change

that led to this conservation. As described in section 1.3, a major part of this process is the sequence of gene duplications/losses and PPI rewiring events that occurred since the MRCA. Ideally, we would like to model these processes from some ancestral PIN up until the observed, present day PINs under comparison. This would allow estimation of gene duplication/loss rates and PPI gain/loss rates, showing how relatively important each of these events is.

One attempt to model these processes has been in the use of PIN growth models. This approach begins with some proposed ancestral PIN and then based on some model of PIN evolution (Figure 1.10), proceeds to add proteins to this network and rewire the interactions, artificially evolving the PIN according to the proposed model. Once the artificial PIN has grown to the desired size, it can then be compared, via various network statistics, to existing PINs based on experimental data. If the model of PIN evolution assumed is representative of the true evolutionary process then you might expect the artificial grown network to resemble a real PIN and so this approach is used to evaluate competing models of PIN evolution based on their ability to produce networks that resemble observed datasets.

The history of this approach began in [67], in which the authors proposed a preferential attachment model for PIN evolution. Under this model, new proteins are added to the PIN and gain interactions preferentially to proteins with many existing interaction partners. This model was used to generate artificial PINs and compare to experimental PIN datasets via their degree distribution (the distribution of number of interaction partners over all proteins in the PIN). PINs contain a few very highly connected "hub" proteins giving the degree distribution a long tail that would not be found if the interactions were distributed randomly through the PIN. It was shown that the preferential attachment model could reproduce this property and so the mechanism of new proteins tending to interact with existing

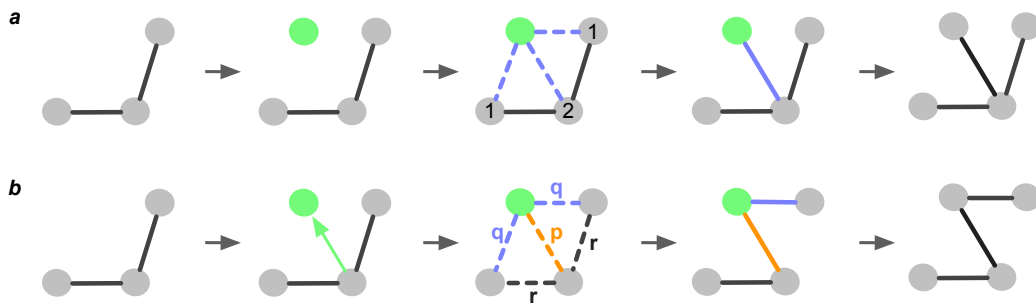


Figure 1.10: Schematic of PIN evolution models. **a** shows the preferential attachment model; starting with some seed PIN (leftmost), at each stage a new protein is added to the network (green) and edges are added to pre-existing proteins with probability proportional to their degree (their number of existing interaction partners). This "rich get richer" model in which proteins with many interactions tend to accumulate more interactions. **b** shows an alternative, the duplication-divergence model, in which at each step a randomly chosen protein is duplicated, an interactions is added between the duplicates with probability p , interactions are added between the new protein and the interaction partners of the original protein with probability q and the interactions of the original protein are removed with probability r .

highly connected proteins was proposed as a defining feature of PIN evolution.

Whilst the preferential attachment model can reproduce the degree distribution seen in PIN datasets, the model does not correspond to the known processes of PPI evolution, such as gene duplication and interaction rewiring. As a result, models that incorporate these known processes have been proposed, called Duplication-Divergence models (DD models). In these models of PIN evolution, at each step a protein is chosen in the network to be duplicated, with the duplicate inheriting the same interaction partners as the source protein. The interactions of both duplicates are then lost according to some probability. For instance, [68] proposed a model in which at each step a node i is chosen randomly and duplicated to give node i' , with an interaction between the duplicates added with probability p . Then for each node j connected to i and i' , one of the two interactions is chosen randomly and deleted with probability q . It was shown that this network generating model produces PINs with similar degree distributions to those found in an experimental PIN dataset [69] and that by tuning the parameters of the model, artificial networks with high modularity, similar to the experimental data, could be generated. Many other generative models have been described and their generated PINs compared to PIN data, for example [70], [71], [56].

The use of generative PIN models in this way has shown that, in principle, duplication of existing proteins and their interactions, followed by rewiring of the duplicate interactions can produce networks that look similar to measured PINs (in terms of degree distribution, clustering coefficient etc). The fact that the underlying mechanics of the DD models mimic the processes of gene duplication and interaction rewiring involved in PPI evolution has been used to argue for the importance of these processes in generating PIN structure [72].

There are however some general limitations to this approach. Firstly, validation of generative models has typically used comparison to the supposed power law

degree distribution of experimental PIN datasets. This is problematic as a method of model selection as PINs with the same degree distribution can look very different if measured by other metrics and so it is not clear how to best describe a PIN in terms of network statistics such as degree. Several statistical approaches to this problem of model selection have been proposed, for instance, Approximate Bayesian Computation was proposed as an approach to such model selection in [73]. The second, deeper limitation lies in the fact that, when comparing a simulated PIN to a 'real' PIN, the evolutionary process used to generate both networks may be entirely dissimilar, for instance, the gene duplications occurring in the history of each bear no relation to each other. This not only calls in to question the validity of evolutionary reasoning based on these models but also as a result of this there is no correspondence between the proteins in the simulated PIN and the 'real' PIN and so it is impossible to answer questions such as when a particular group of PPIs appeared during evolution. As seen in section 1.2.3, computational tools exist for reconstructing the evolutionary history (i.e. phylogeny) of a group of proteins and so one possible approach to prevent this problem would be to use this evolutionary history as a framework to model PPI evolution. That is, modelling the rewiring events that occurred during the evolution described by a phylogeny, leading to the present day PPIs. In contrast to the generative approach, this allows models of PIN evolution that are consistent with a phylogenetic history of the proteins in the PIN.

1.5.2 Complex Level

The methods described in the previous section focus on studying the evolution of PPIs at the network level. This is the broadest level of detail in studying PPIs, encompassing large groups of proteins (often the entire proteome of a species) and

their interactions. At a higher level of detail, PINs contain protein complexes; groups of proteins that associate to form larger assemblies capable of performing some task. These complexes can be seen as the indivisible functional units of the PIN; if some constituent protein is removed the complex will be unable to perform its function. For larger complexes of more than two proteins, a description of the evolutionary history of the complex will require an understanding of the evolution of the component proteins and of the order in which the PPIs forming the complex appeared. An attempt at deducing an evolutionary history has been made for several specific protein complexes, for instance, [74] used arguments from parsimony to propose a history of the proteasome complex. Similar analyses were undertaken in [75] and [76].

More general attempts to model the evolution of protein complexes have been made, particularly in describing the evolution of complexes formed of many copies of the same subunit (homomeric complexes). These types of protein complexes often display symmetry or quasi-symmetry. As described in [77], symmetry is extremely common within protein complexes and the symmetrical structures formed can often be vital for the function of the complex. For instance the 6-fold rotational symmetry of the sliding clamp of DNA polymerase allows formation of a ring that can encircle the DNA strand. Larger complexes can have more complicated symmetry, such as the GroEL complex having 7-fold dihedral symmetry, forming a barrel structure big enough to contain another protein. Newly translated and unfolded proteins are drawn in to this cavity and the hydrophobic conditions inside coerce the substrate protein to fold [78]. These functions are contingent on the symmetry of the complex [77].

Given the functional constraint of symmetry on complexes such as these, successful modelling of their evolution must attempt to model the evolution of symmetry. Such an attempt was made in [79] to describe the possible evolutionary

history of homomeric complexes in terms of their symmetry. Given a self interacting protein forming a complex, mutations occurring at the surface of the protein can form new binding sites and so change the quaternary structure of the complex. The authors argue that the possible symmetry groups of the new complex after emergence of a new binding site are dependent on the symmetry group of the original complex. Using this argument, the authors formulate the evolution of homomeric complexes as evolution between symmetry groups and use observed structures to estimate the rate of conversion between each type. This model is useful in describing the constraints of symmetry on complex evolution however there are several limitations to this approach. Firstly, it can not infer the history of a given complex, for instance given some protein complex with complicated symmetry, the method cannot infer the original ancestral complex nor the sequence of evolution that lead to the existing complex. Secondly, this model is restricted to homomeric complexes in which all subunits have the same evolving sequence. In fact, symmetry plays a role in other classes of protein complex, such as those formed of paralogs [80]. A general approach would be preferable in order to study these complexes.

1.5.3 Interaction Level

At the highest level of detail, PPI evolution can be studied computationally in terms of single interactions; that is, modelling and predicting how the sequence and structure of interacting proteins evolve and how this effects their dynamics. An important concept here is that of coevolution. As explained earlier in Section 1.1.2 PPIs are the result of specific and complementary interactions between residues across the protein-protein interface. The residues responsible for making these favourable interactions are crucial to the ability of the proteins to interact and thus

to perform the function resulting from the interaction. Therefore, mutations at these residues, disrupting the interaction are usually deleterious and removed from the population by selection [50]. However, it is possible that after such a mutation, a subsequent mutation in the interaction partner will restore complementarity and so make the double mutation non-deleterious. This process can thus lead to correlated mutations in interacting proteins.

Computational methods attempting to capture this process have mainly focussed on prediction of PPIs. One such approach is the MIRRORTREE method [81] that attempts to predict whether two protein families interact by looking for correlated evolutionary rates in the families. This is done by looking for correlated branch lengths across two phylogenetic trees describing each family. The method works as protein families that interact with each other tend to have more similar phylogenetic trees compared to families that do not interact, however, this is not necessarily due to the coevolutionary process described above and could be due to factors such as shared expression rate during evolution [82], [83]. There are several limitations of the MIRRORTREE approach; firstly, predictions can only be made for proteins for which a phylogenetic tree can be built and secondly, the predictions made are at the family level, that is, the method predicts whether there are any inter-family PPIs between two protein families but not the exact pattern of PPIs existing between the two families.

The existence of correlated mutations in interacting proteins has also been used to predict the structure of PPIs. The compensatory mutations described above often occur between residues that are close to each other at the PPI interface. It has also been observed that the bound structure of a PPI is often conserved at the level of the protein family, that is proteins from a given family have similar interaction structure when binding to proteins from another given family. These observations taken together have led to attempts to predict the binding structure

between two protein families (or at least the residues important for the interaction) by looking for correlated positions between multiple sequence alignments of those families. One recent, successful attempt at such a prediction [45] used a statistical method to find the correlated positions and then used these as constraints during a subsequent 3D structure prediction of the bound proteins.

1.6 Phylogeny in Protein-protein Interaction Evolution

As described in Section 1.3, PPI evolution is the result of changes in the proteins present in an organism (through gene duplication, gene loss, horizontal transfer) and changes in the interactions between those proteins, known as rewiring events. Any attempt at a full model of PPI evolution will therefore include these processes. The first process of gene gain and loss can be modelled by phylogenetics and the second process of interaction rewiring can be modelled using any method that can predict PPIs between individual proteins. Of the methods for modelling PPIs mentioned so far, none attempt to explicitly model these processes simultaneously (except perhaps for network growth models but here the phylogeny produced does not describe the relationship between any real group of proteins). Therefore, in order to produce a complete model of PPI history for a given group of proteins, a different approach is required, specifically using phylogeny as a basis.

One recent such approach is the PARANA method [84] for predicting the history of PPIs between two protein families, given the set of present day PPIs. This algorithm starts by producing phylogenetic trees for the two families and then given the known interactions at the leaves of the trees, attempts to find the least set of PPI rewiring events that could generate these interactions. The

resulting set of events form a proposed history of interaction and it was shown that the method could be used to recover the history of PINs generated using various network growth models [84]. This method may be preferable in some situations but a parsimony based method will not be preferable in cases where we know more about the PPIs than their existence. For instance, if the structure of the PPIs is known or can be modelled then changes in the sequence at important locations could be used to predict rewiring events. A more accurate and specific prediction could be made by a methodology that incorporates this kind of knowledge about the sequence and structure of the proteins.

1.6.1 The Interaction Tree

This research focuses on a novel model for describing the evolution of protein interactions called the interaction tree. This approach takes protein phylogenies, reconciled with a species tree, and produces a history of all possible interactions throughout the species tree. This history is represented as an interaction tree in which each node represents a pair of proteins (a possible interaction) and an edge represents evolution between pairs of proteins. Importantly, the interaction tree contains a node for every possible interaction i.e. for every pair of proteins that were jointly present in an organism not just for every pair believed to be interacting. This interaction tree can then be used as a framework in which interaction rewiring events are predicted between pairs of proteins, as their sequences evolve. This overcomes a major problem of previous, generative, models of PIN evolution in that the interaction tree describes the evolution of a real set of proteins rather than evolving an artificial network and then fitting network statistics to observed values. It also avoids the problem of many traditional network alignment methods in that the evolutionary processes leading to present day networks are explicitly

modelled, allowing inferences to be made about these processes and ancestral PPIs to be predicted.

The interaction tree framework has so far been applied to reconstruct ancestral networks [85], [86], identify complexes conserved across species [62] and perform data integration [87]. The flexibility of the approach allows models capable of incorporating phylogenetic information, noisy or error prone data and the specific structure and sequence of an interaction at the interface level. This type of model is expected to increase the ability of current research to understanding evolution of protein interactions by testing hypothesis and making predictions [88].

In this work, I firstly aim to apply the interaction tree framework to study the evolution of protein-protein interactions in protein complexes. Many protein complexes are built from paralagous subunits, related by gene duplication [80]. These subunits have a phylogenetic relationship that can be used as a basis for the interaction tree method which can then predict the rewiring events occurring between subunits during the phylogeny. For this to work, it is necessary to develop a model of PPI rewiring that can predict gains and loss of interaction between the paralogs of a complex, given some change in sequence in the phylogeny. The development of such a model is the subject of the next chapter.

Chapter 2

Adapting the interaction tree for protein complexes

2.1 Introduction

Large protein complexes are present in all organisms and are responsible for some of the core function of the cell e.g. DNA replication, protein synthesis, protein degradation. The structure and pattern of interactions within these complexes is divergent across species and so studying their evolution is important for understanding how and why such diversity arises.

One common type of multi subunit complex are those made of homologous subunits. These can be formed of many copies of the same subunit or formed of subunits from a family related by duplication. It is thought that such large complexes made from related subunits arise from an original self-interacting subunit which is duplicated to produce paralogs. These new paralogs are then incorporated into the complex allowing a inhomogenous complex to arise, formed of interacting, duplicate proteins [80]. This process is thought to be important in the evolution of

protein complexes based on the number of observed self-interacting proteins and the number of observed complexes containing paralogs [89], [90], [80], [91], [92].

Some ideas have been proposed to describe the evolution of these complexes [79] but there is a lack of detailed modelling of the emergence of these complexes. The interaction tree is a suitable methodology to use for this task as a phylogeny can be constructed relating all of the subunits involved in a complex. The issue to be decided in applying the interaction tree to this class of problem is how to model the PPI rewiring events on the phylogeny. Previously, [87] used the interaction tree approach with a constant probability of gain or loss of interaction after duplication as a model of rewiring. This approach was used to model the evolution of protein complexes, however using a constant probability across all subunits led to no distinction between the interaction patterns of the subunits; the complexes appeared as almost completely connected components. In fact, paralogous complexes are not completely connected in this way but each paralog has a distinct pattern of connections, vital to conferring function on the complex.

In order to detect interactions at this higher level of definition, a model of rewiring that is specific to each subunit needs to be defined. Previously, [85] defined such a model (based on branch lengths of the phylogeny) and applied it to reconstructing the history of interactions amongst the bZip family of transcription factors. In this chapter, this model is tested in its ability to predict rewiring events in protein complexes along with two other models, one based on the popular MIRRORTREE method and one based on a rough measure of the chemical complementarity at the protein-protein interface. As a test case for these models the proteasome complex is used as it is a well studied complex with many structural examples.

2.2 Methods

2.2.1 Building an interaction tree

The modelling of PPI evolution in this thesis uses the interaction tree framework, the concept of which will be explained here. As mentioned in the previous chapter, PPI evolution is the combination of changes to the set of proteins present in an organism and changes to the set of interactions between them. Phylogenetics can be used to model and predict the change in the set of proteins present in an organism over evolutionary time, with gene tree reconciliation able to predict the set of proteins present in ancestral organisms. Therefore, a sensible approach to modelling PPI evolution (in cases where you can build a phylogeny for the proteins) would be to use these phylogenetic predictions as a scaffold onto which the changes in the PPIs between them are modelled. The problem here is that the phylogenetic trees describe the evolution of single proteins but in order to model PPI rewiring, changes between pairs of proteins need to be described.

The interaction tree offers a route around this problem; in situations where the PPIs between two protein families are to be modelled, the method combines the two phylogenetic trees and produces a new tree structure (an interaction tree), in which each node represents a pair of proteins and edges represent evolution between these pairs. It is then possible to take this structure and model the PPI rewiring events occurring on its branches. An overview of this approach is shown in Figure 2.1 and a step by step description of the construction of an interaction tree for a toy example is shown in Figure 2.2

Interaction trees describe the totality of possible interactions amongst 1 or more protein families and track this set of possible interactions in evolutionary time using phylogenies constructed for each protein family (Figure 2.1). The history of possible interactions is represented as a tree in which each node rep-

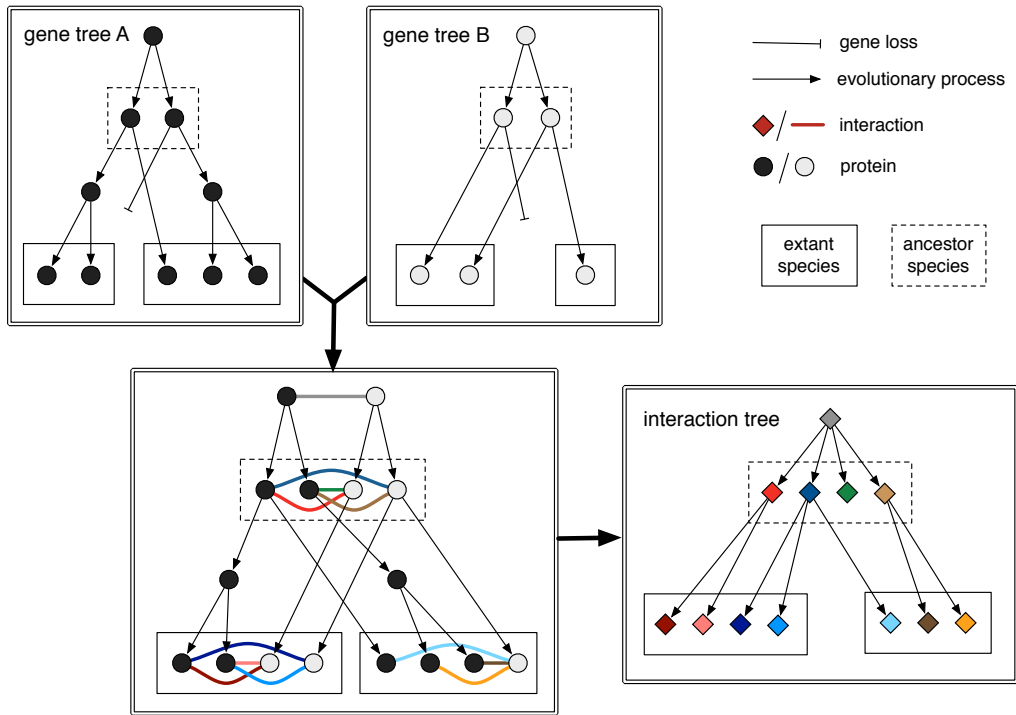


Figure 2.1: Schematic showing the construction of an interaction tree. Here an interaction tree is constructed describing the history of possible interactions between two protein families. To begin we construct a gene tree for each protein family (*top left*) using MUSCLE [93]. These trees are reconciled with a species tree, using NOTUNG [34], in order to classify each node in the gene trees as a duplication or speciation node (describing a gene duplication event or speciation event respectively) and to assign proteins to ancestral species. This allows description of all possible interactions between the two families in each species, including ancestor species (coloured lines, bottom left). From here we can construct a tree of interactions (bottom right), in which the coloured nodes represent the possible interactions and each interaction is the child of one ancestor interaction. Each node can be either 'on' or 'off', corresponding to presence of absence of interaction.

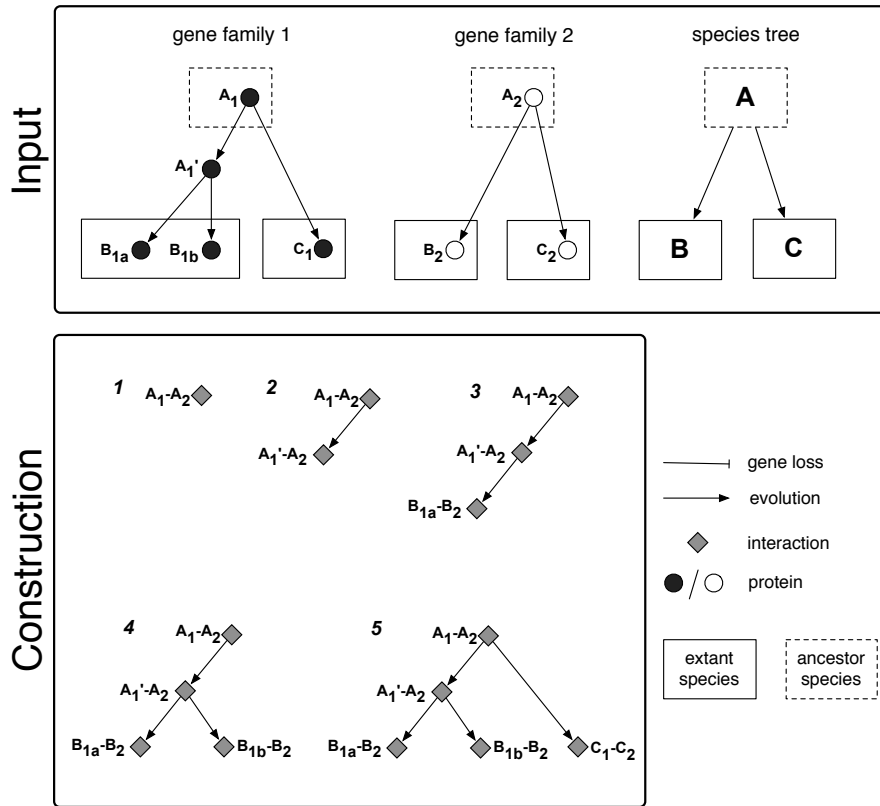


Figure 2.2: Schematic of construction of an interaction tree. Starting from two reconciled gene trees, a tree is constructed in which nodes represent possible interactions between nodes in the gene trees. Beginning at 1, the two root nodes are combined to form the root node of the interactions tree. This node represents the pair of ancestral proteins from which all other proteins in the two families are derived from (note, this is not assuming that the two ancestral proteins interacted, merely that they coexisted in some organism). Now, take the left branches from the root nodes to reach A_1' from gene family 1, this gene coexisted with a protein intermediate to A_2 and B_2 in gene family 2. To avoid introducing an intermediate node on this branch, A_1' is paired with A_2 to form the next interaction node in 2. The interaction nodes added in 3 and 4 complete the description of all possible pairings up to species B. Now the procedure is repeated on the right branches from the root to give a description of all interactions up to species C. The resulting tree structure describes all pairs of proteins thought to be coexisting in some organism at some time. The pairs can either be either interacting or non-interacting.

resents a possible interaction between 2 proteins and is the child of exactly one parent interaction. Each interaction node can be in one of two states, 'on' or 'off', corresponding to a present or absent interaction. The tree structure makes this framework well suited for building probabilistic graphical models describing the evolution of protein interactions. For instance, if we can find a probability function that describes the probability of a child node being 'on' or 'off', given the state of its parent node and the evolution between them, then message passing algorithms [94] can be used to compute probabilities for 'on' or 'off' at every node in the tree. Such a function would be written as

$$P(I(C) | I(A), D(A, C)) \tag{2.1}$$

Where (as shown in Figure 2.3), A is the ancestor interaction node, C is the child interaction node, $I(A)$ is the interaction state ('on' or 'off') of interaction node A and $D(A, C)$ is some measure of the evolutionary change between interaction nodes A and C . For instance, suppose that $D(A, C)$ is the total number of substitutions in both proteins from the ancestral pair A to the pair C . That is (from Figure 2.3), the number of substitutions on the branch from A_1 to C_1 plus the number of substitutions on the branch from A_2 to C_2 . Then, given some training set of PPI evolutionary histories, the probability function Equation 2.1 can be estimated and represented as a matrix (Table 2.1) for given values of D . If such a relationship can be found then it can be used to deduce probabilities of interaction at unobserved nodes given some observed interactions. In the simplest case, if A is known to be 'on' then the probability that C is also 'on' can be read from a matrix such as Table 2.1.

Previously such a function was defined based not on substitutions but similarly

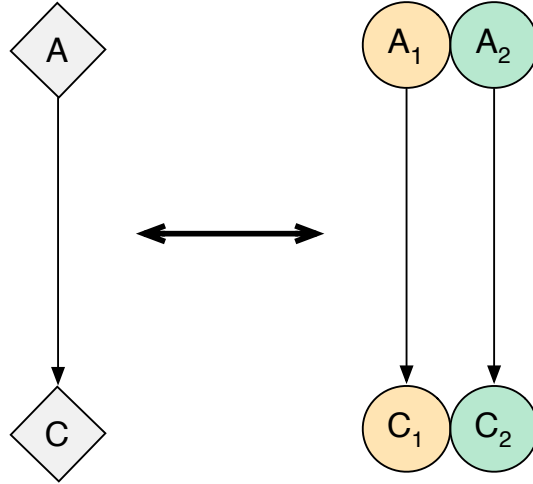


Figure 2.3: Decomposing an interaction tree branch. A branch of an interaction tree describes evolution from an ancestral interaction A to a child interaction C . Each interaction node can be decomposed into the two constituent proteins that make up the interaction (right). We label these A_1 , A_2 , C_1 and C_2 . A_1 is the ancestor of C_1 in the original gene phylogenies used to build the interaction tree and likewise for A_2 and C_2 .

$$D(A, C) = 10$$

	$P(C = on)$	$P(C = off)$
$A = on$	0.2	0.8
$A = off$	0.1	0.9

Table 2.1: A representation of Equation 2.1 in matrix form. The matrix shown is for some measure of PPI evolution D and describes the probability of C given A when $D=10$. For instance, if the ancestor node is non interacting, there is a 10% chance that the child node is interacting if $D=10$, that is, there is a 10% chance of a gain of interaction. Given a defined D , these probabilities can be estimated from some training set of observed PPI rewiring events/non-events.

on the distance between proteins in terms of the expected fraction of changed amino acids [85], as calculated under the Jones-Taylor-Thornton model of protein sequence evolution. The next section describes three possible D on which to base such a function, when applied to protein complexes formed of homologous subunits. Including that from [85] and two new candidates.

2.2.2 Scoring systems

Three possible D are described here with the the aid of some toy examples. Each of these D are candidates for defining a model of PPI evolution as described by Equation 2.1. The first model, D_{dis} is based on the the following quantity

$$D_{dis}(A, C) = E(A_1, C_1) + E(A_2, C_2) \quad (2.2)$$

Where A, C are interaction nodes as shown in Figure 2.3 and $E(i, j)$ is the distance between protein sequences i and j restricted to the binding site under a Jones-Taylor-Thornton (JTT) model [95] of amino acid substitution, as calculated using PROTDIST [96]. This is the same function used to formulate an interaction tree model of PPI evolution in [85]. The model assumes that larger values of D_{dis} are more likely to produce rewiring events.

To demonstrate the calculation of D_{dis} , a toy example is used. Given two protein families (Family 1 and Family 2) described by phylogenetic trees, we will focus on C_1 , an existing protein from family 1, and C_2 , an existing protein from family 2 (Figure 2.4). Now suppose that C_1 and C_2 interact; the distance D_{dis} between the C_1 - C_2 interaction and its ancestor interaction can be calculated and used to predict if the ancestor proteins interacted. For instance, assume that given a D_{dis} of between 2 and 3, it is determined from some training set that there is a 90% chance of PPI being maintained between two interacting proteins and a 50%

chance of PPI gain between proteins that do not interact.

Figure 2.4 shows the toy example, the sequence distances under the JTT model, between each ancestor-child protein are shown, on the respective branches of the phylogenetic trees (this can be calculated using a variety of software packages, e.g. PROTDIST). The D_{dis} between the ancestral A_1 - A_2 pair and the C_1 - C_2 pair is then calculated as

$$D_{dis} = 1.0465 + 1.6988 = 2.7453 \quad (2.3)$$

Now, given our previously described knowledge of the occurrence of rewiring events given $2 < D_{dis} < 3$, the probability of an ancestral interaction can be inferred. To do this, we can use Bayes' rule

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \quad (2.4)$$

where A is the event ' A_1 and A_2 interact' and C is the event ' C_1 and C_2 interact'. Here $P(A)$ represents some prior belief that the ancestral proteins interact. If there is thought to be a 50% prior probability of an ancestral interaction then the posterior probability, given the observed D_{dis} is

$$P(A|C) = P(C|A)P(A)/P(C) = \frac{0.9 * 0.5}{(0.9 * 0.5) + (0.5 * 0.5)} = 0.64 \quad (2.5)$$

The prior belief of a 50% chance of an ancestral interaction has been revised to a 64% chance, in light of the information contained in D_{dis} . Later, a training set of PPI evolution will be defined. This will be used to observe the frequency of interaction gain/loss, given changes in D_{dis} , which can then be examined to determine the suitability of this measure in defining a model of PPI evolution.

The second model, D_{dif} is defined similarly to D_{dis} ...

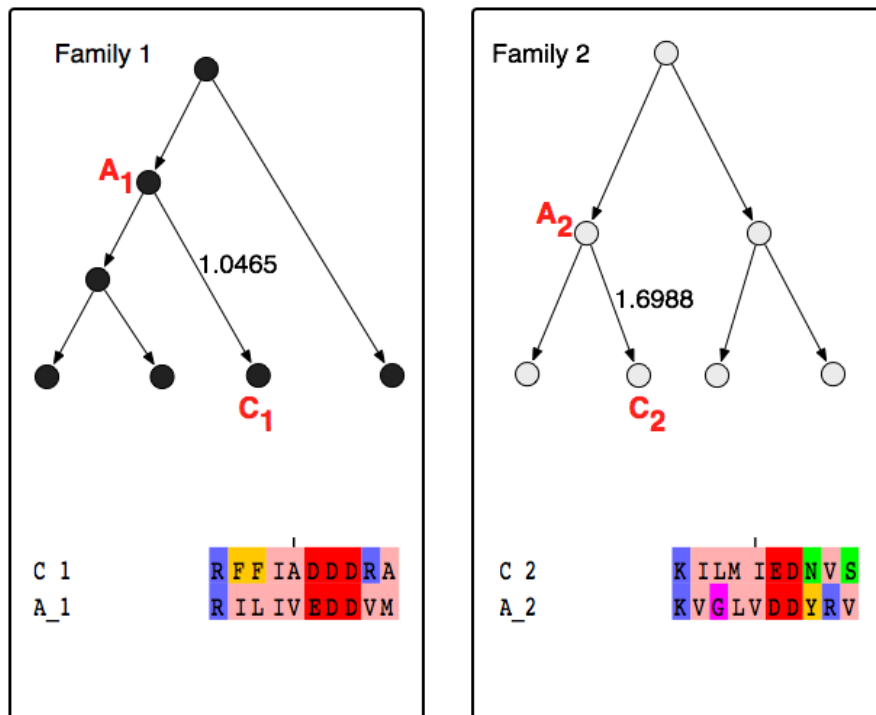


Figure 2.4: A toy model, described in the text, used to demonstrate the calculation of D_{dis} . Sequences for each described protein are shown, as are the relevant branch lengths used in calculating the metric. This example assumes that ancestral protein sequences are known with certainty. In reality, these sequences must be estimated using some sequence reconstruction method.

$$D_{dif}(A, C) = \left| \frac{E(A_1, C_1) - E(A_2, C_2)}{E(A_1, C_1) + E(A_2, C_2)} \right| \quad (2.6)$$

With $E(i, j)$ once again representing the distance between proteins i and j . Comparing this equation with the description of the previous D_{dis} model (Equation 2.2) we can see that rather than measure the total of the branch distances, this metric measures the asymmetry of the branches. Returning to the example shown in Figure 2.4, calculation of D_{dif} can be demonstrated in the same way as for the previous metric...

$$D_{dif} = \left| \frac{1.0465 - 1.6988}{1.0465 + 1.6988} \right| = 0.2376 \quad (2.7)$$

This metric is larger for pairs of proteins with asymmetric evolutionary rates and smaller for proteins evolving at a similar rate. It is hoped that this metric will take advantage of the observation that pairs of interacting proteins tend to evolve at similar rates, as described in [39], [97], [8]. To elaborate, if D_{dif} tends to be very low for pairs of proteins maintaining a PPI over time, then this information will be contained in the $P(C|A)$ of Equation 2.4. This will then allow updating of the probability of interaction for ancestral pairs, reflecting this information. In order to determine the relationship between D_{dif} and the occurrence of rewiring events, a training set of PPI evolution will again be required.

The final model, D_{com} , is slightly more complicated than the previous two. This model attempts to measure the change in chemical complementarity at the protein-protein interface as the sequence of the proteins evolves. The metric that

this model is based on is adapted from the SCOTCH [98] method for scoring docked protein models.

We start by defining a measure of complementarity at the protein-protein interface, called the complementary fraction. To calculate the complementary fraction we first divide the 20 amino acids into 4 groups; (GLY, ALA, VAL, LEU, ILE, MET, CYS, PHE, PRO, TRP, TYR), (SER, THR, ASN, GLN), (LYS, ARG, HIS), (ASP, GLU). These are the hydrophobic, polar, positively charged and negatively charged residues respectively. We define two amino acids to be complementary if they are both hydrophobic, both polar or one positively and one negatively charged (Figure 2.5).

We start with two protein sequences and a proposed interface between them represented as a list of interacting residue pairs (i, j) where i refers to residue position i of the first sequence and j refers to residue position j of the second protein sequence. Here a pair of residues are defined as interacting if they contain heavy atoms within 4.5\AA of one another. An example of such an (i, j) pair is shown in Figure 2.6.

To describe the complementarity of the interface, we will aggregate over all (i, j) pairs, allowing for complementarity to also be maintained by nearby residues. In detail, for each (i, j) pair in turn, we then find the two nearest structural neighbours of both i and j , residing in the same chain. This requires a proposed three dimensional structure for the interface, such as can be found in a solved crystal structure of the constituent chains in complex. An example is shown in Figure 2.6 where i_1, i_2 are the nearest two structural neighbours to i and j_1, j_2 are the nearest two structural neighbours to j . We then define positions i and j to be complementary if any of the pairs $(i, j), (i_1, j), (i_2, j), (i, j_1), (i, j_2)$ are complementary as described in Figure 2.5.

For the example of Figure 2.6, the residues at each of the positions are listed

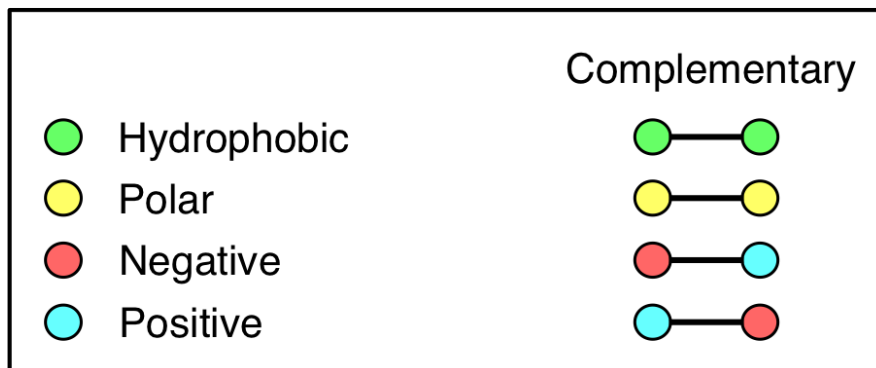


Figure 2.5: The categorisation of pairings of residues as complementary or non-complementary. All 20 amino acids are decomposed in to 4 groups; Hydrophobic, Polar, Negative and Positive (described in the text). Pairings of residues are defined as complementary if the pairing of their groups falls in to one of four cases (right of figure). All other pairings are non-complementary.

Position	Residue
i	ALA
j	SER
i_1	THR
i_2	ARG
j_1	ILE
j_2	PRO

Table 2.2: A list of residues at relevant positions in the toy example of Figure 2.6. This example is used to demonstrate the process of calculating the complementary fraction of two proteins.

in Table 2.2. So, we are examining the pairs (ALA, SER), (THR, SER), (ARG, SER), (ALA, ILE) and (ALA,PRO). As at least one of these pairs is complementary (the 2nd, 4th and 5th are) we say that the positions (i, j) are complementary. Notice that the *positions* (i, j) are declared complementary even though the actual complementary interaction is maintained by a neighbouring residue(s). This process is repeated for all of the interacting (i, j) pairs at the protein-protein interface. The extent of complementarity across the interface is then described by the complementary fraction; the fraction of all interacting (i, j) pairs that are complementary. For instance, suppose for a given PPI there are 40 pairs of residues (one from each chain) with heavy atoms within 4.5Å (note that a residue can appear in more than one pairing here if there is more than one residue within the distance threshold). Then suppose that of these pairs, 30 are found to be complementary, taking in to account structural neighbours as described here. The complementary fraction would then be calculated as $\frac{30}{40} = 0.75$.

This score works under the assumption that during coevolution of maintained interactions, a residue mutation going to fixation at an interface will most likely be physiochemically complementary (e.g. hydrophobic to hydrophobic) to the residues it interacts with. However, this complementarity can be maintained within clusters of residues at an interface and so does not have to be maintained through specific pairs but can be conserved by nearby residues in the interface (i.e. the structural neighbours). The complementary fraction aims to measure this coevolution occurring between interacting proteins and so could be used here to detect changes in interaction state as changes in complementary fraction.

In many situations, particularly when considering ancestral proteins, no structure of the PPI will be available. In this situation a strategy is proposed using an homologous structure as a template. For instance, if we wish to compute the complementary fraction for two proteins that are homologous to those shown in

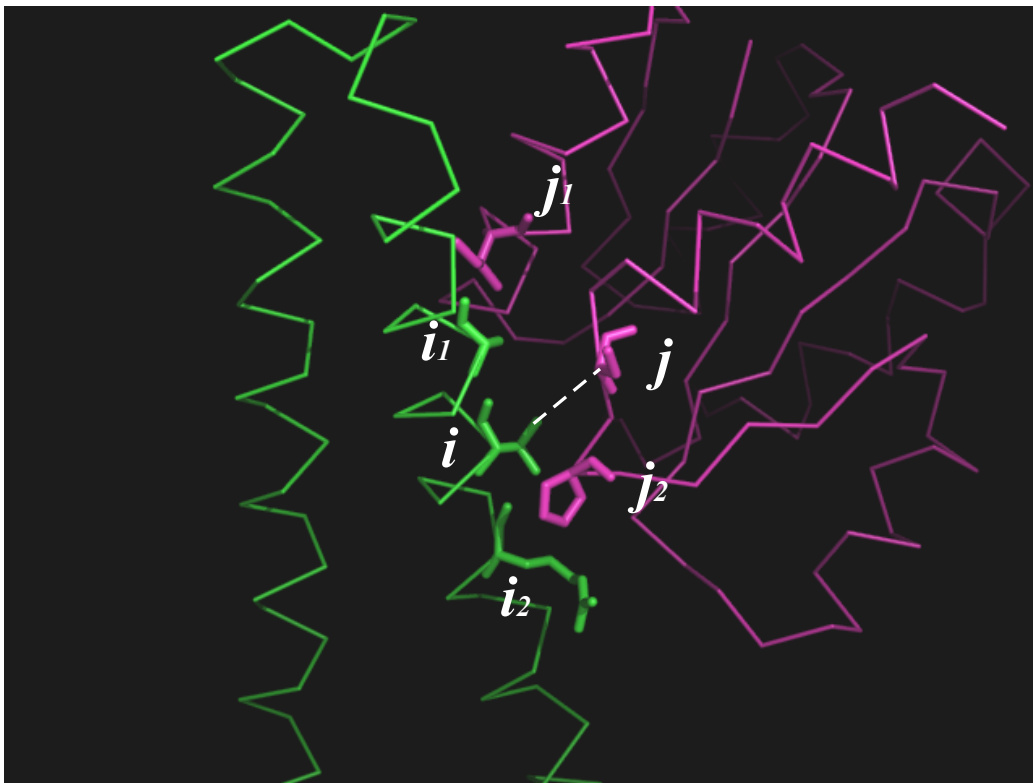


Figure 2.6: An example of residues pertinent for classifying two nearby residues as complementary or not. The residues i (green protein) and j (purple protein) are within 4.5\AA of each other. To classify this pair of positions as complementary or not, their nearest two structural neighbours in the same chain must be identified. These are i_1 and i_2 for i (labelled on figure) and j_1, j_2 for j (also labelled). If any of the pairings of residues (i, j) , (i_1, j) , (i_2, j) , (i, j_1) , (i, j_2) are complementary, as defined in Figure 2.5 then the positions i and j are said to be complementary. The classification for this example is carried out and described in the text.

Figure 2.6. Suppose the sequences of these proteins are known but we do not have a structure for their proposed PPI. The complementary fraction can be estimated for our new proteins, using the original structure as a template. To do this, the two new proteins are firstly aligned to their respective template proteins (Figure 2.7). For each (i, j) interacting pair in the template structure we can then identify an analogous, aligned (i, j) pair in the new proteins. The same approach can be used to estimate the nearest structural neighbours i_1, i_2, j_1 and j_2 (see Figure 2.7). The categorisation (as complementary or not) of each (i, j) pair of positions in the new sequences can then be estimated using these analogous positions.

In the example shown in Figure 2.7, to decide if (i, j) are complementary we must check if any of the $(i, j), (i_1, j), (i_2, j), (i, j_1), (i, j_2)$ are complementary in our new sequences. The new residues to be compared are listed in Table 2.3. In this case that means checking (ALA, SER), (ALA, SER), (GLY, SER), (ALA, THR) and (ALA, ASN) for at least one complementary pair. As each of these pairs a hydrophobic and a polar residue, none are complementary. As such, the positions (i, j) are not complementary in the new sequences. This process can be repeated for all (i, j) pairs within 4.5\AA in the template structure. The fraction of these positions that are complementary in the new sequences is then an estimate for the complementary fraction of the new sequences.

Now, having defined a measure of physicochemical complementarity between proteins, there follows a description of how this measure can be used to model PPI gains and losses during evolution. Given that interacting proteins are expected to have a higher complementary fraction, it is proposed that changes in complementarity are tracked during evolution with the hope that increases correspond to gains of interaction and opposite for PPI losses. Thus a model (the D_{com} model) is defined based on the following quantity

Using a template structure to estimate the complementary fraction.

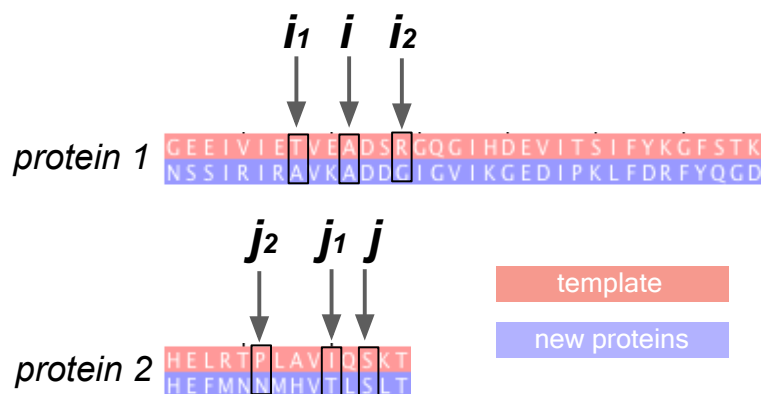


Figure 2.7: If a PPI structure is not available for two proteins (because the structure has not been solved but also when the proteins do not interact) the complementary fraction can be estimated using some homologous structure. Here the proteins from our original example are shown in red and some new pair of proteins for which we have no PPI structure is shown in blue. For a given interacting pair of positions (i, j) in the template sequence there is an analogous aligned pair in the new sequences (ALA, SER in this example). Equally, there are analogous residues to each of the structural neighbours. The comparison of these analogous residues to classify (i, j) as complementary or not in the new sequences is then as before. This is repeated for all interacting positions in the template to produce an estimate of the complementary fraction in the new pair of sequences. An example of such a calculation is given in the text.

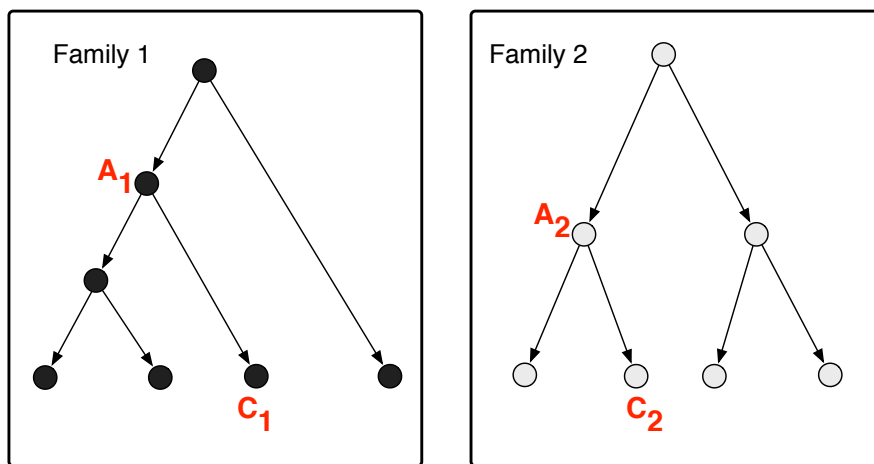
Position	Residue
i	ALA
j	SER
i_1	ALA
i_2	GLY
j_1	THR
j_2	ASN

Table 2.3: A list of relevant residues to be compared to classify the positions (i, j) as complementary in the example given in the text. These residues are defined using a template structure.

$$D_{com}(A, C) = F(C_1, C_2) - F(A_1, A_2) \quad (2.8)$$

Where $F()$ is the complementary fraction as described above. To demonstrate the use of this model in practice, a toy example is given once again (Figure 2.8). This example has the same format as the example for the previous two models; Two proteins C_1 and C_2 are known to interact and we wish to infer if their ancestral precursors A_1 and A_2 also interacted. Assuming some known template structure to which these proteins can be aligned, the complementary fraction of C_1 and C_2 is 0.7. After inferring the sequences of the ancestral proteins, their complementary fraction is found to be 0.5. So the D_{com} metric is calculated as

$$D_{com}(A, C) = 0.7 - 0.5 = 0.2 \quad (2.9)$$



$$F(A_1, A_2) = 0.5$$

$$F(C_1, C_2) = 0.7$$

Figure 2.8: A toy example of the calculation of D_{com} in the context of PPI evolution. The complementary fraction of C_1 and C_2 is calculated as 0.7 using either a structure for the interaction or a template structure. The complementary fraction for the pair of ancestral proteins is 0.5. Thus, D_{com} here is $0.7 - 0.5 = 0.2$.

As before, if, using some training set of PPI evolution, we can estimate the probability of a PPI rewiring event given a D_{com} of 0.2, then Bayes theorem can be used to infer the probability of interaction between the ancestral proteins.

2.2.3 Proteasome data

In order to test the ability of these 3 models to infer the history of a protein complex, a test case is required. For a test case, the 20S proteasome complex from *Saccharomyces cerevisiae* is used (Figure 2.9). This complex is a multi-subunit protease responsible for degrading proteins that have been tagged with

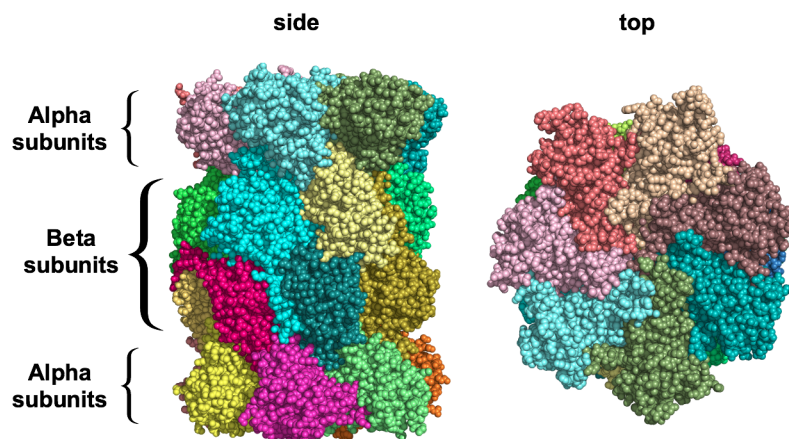


Figure 2.9: The yeast 20S proteasome. The barrel structure is shown from the side (left) and from above with the subunits coloured differently to distinguish. The four layers/rings forming the barrel can be seen in the side-on view. The outer most ring is shown face on in the right hand view. This is the alpha ring that controls access to the cavity in the centre of the complex. Here the ring is in closed conformation and access to the cavity is restricted.

ubiquitin. In doing so, the complex plays a major role in regulation of protein levels in the cell, including degradation of misfolded or denatured proteins. All 28 subunits show homology and can be split in to two subtypes; the alpha and the beta subunits. The complex itself is shaped like a barrel, formed of four stacked heptameric rings with the outer rings of alpha subunits regulating access to a cavity formed by the inner two rings of beta subunits. It is in this cavity that the active sites responsible for the protease activity are found. The 1RYP structure of this yeast protein complex was downloaded from the Protein Data Bank. From this structure, PPIs are defined as existing between any two subunits having heavy atoms within 4.5\AA of each other. These PPIs are used as a starting test set for comparison of the three models of PPI evolution previously defined.

2.2.4 Clustering algorithm

Given a set of interfaces we use a clustering algorithm to identify interface types that can then be treated separately. This clustering algorithm has been designed to cluster interfaces between two protein families i.e. interfaces between protein A and protein B where A belongs to the first family and B belongs to the second family for all of the interfaces. We start by aligning all proteins involved, for each family, to produce an alignment containing m columns for the first protein family and n columns for the second. Then for each interface in our set, we produce a list of residues within 4.5\AA across the interface, each of which can be represented by an (i, j) pair with i representing the relevant column number in the A alignment and similarly for j in the B alignment. For each interface this gives a representation

$$P = \{(i, j) : 0 < i \leq m, 0 < j \leq n\} \quad (2.10)$$

For two interfaces, with representations P_1 and P_2 , the *Jaccard distance* between the interfaces can be defined as

$$J(P_1, P_2) = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|} \quad (2.11)$$

i.e. the fraction of (i, j) pairs shared by the interfaces out of all observed pairs from either interface. This distance is then used to produce average linking clustering with a threshold of 0.1 for groups of interfaces. This produces clusters of similar PPIs having different parts of their proteins responsible for mediating the interaction.

The clustering as applied to the yeast proteasome is shown in Figure 2.10. Six clusters are identified within the stacked ring topology of the proteasome. These clusters correspond to topologically consistent groupings within the structure of the complex. For instance, one cluster contains the PPIs forming the alpha rings and one cluster contains the PPIs forming the beta rings.

2.2.5 Training set of paralog complexes

After comparing the performance of each model using the yeast proteasome structure described above, the analysis is expanded to cover a larger set of PPIs. A training set of PPIs taken from large, obligate complexes is used to define a training set of interaction tree branches which are then used to fit a model of PPI evolution (i.e. by finding a conditional probability function of the form Equation 2.1). To build this training set of interaction tree branches we firstly take from 3D complex [99], the list of protein complexes containing 14 or more subunits non-redundant in topology and sequence to the QS30 level. The QS30 level comprises protein complexes that are non-redundant in terms of their quarternary topology and a 30% sequence identity threshold. The biological assemblies of these complexes were downloaded from PDB [100]. From this list of complexes we now want to identify those that are composed entirely of subunits from one paralogous gene family.

To identify the complexes formed from paralogs, we employ pairwise BLAST [25] searches. It is not sufficient to just identify those complexes in which every subunit aligns to at least one other subunit as, for instance, this would mistakenly identify complexes formed from two gene families as a complex of paralogs. Instead we define a *homology graph* in which subunits are nodes and edges are placed between nodes whose subunits produce an alignment, with E-value < 0.05 in a pairwise

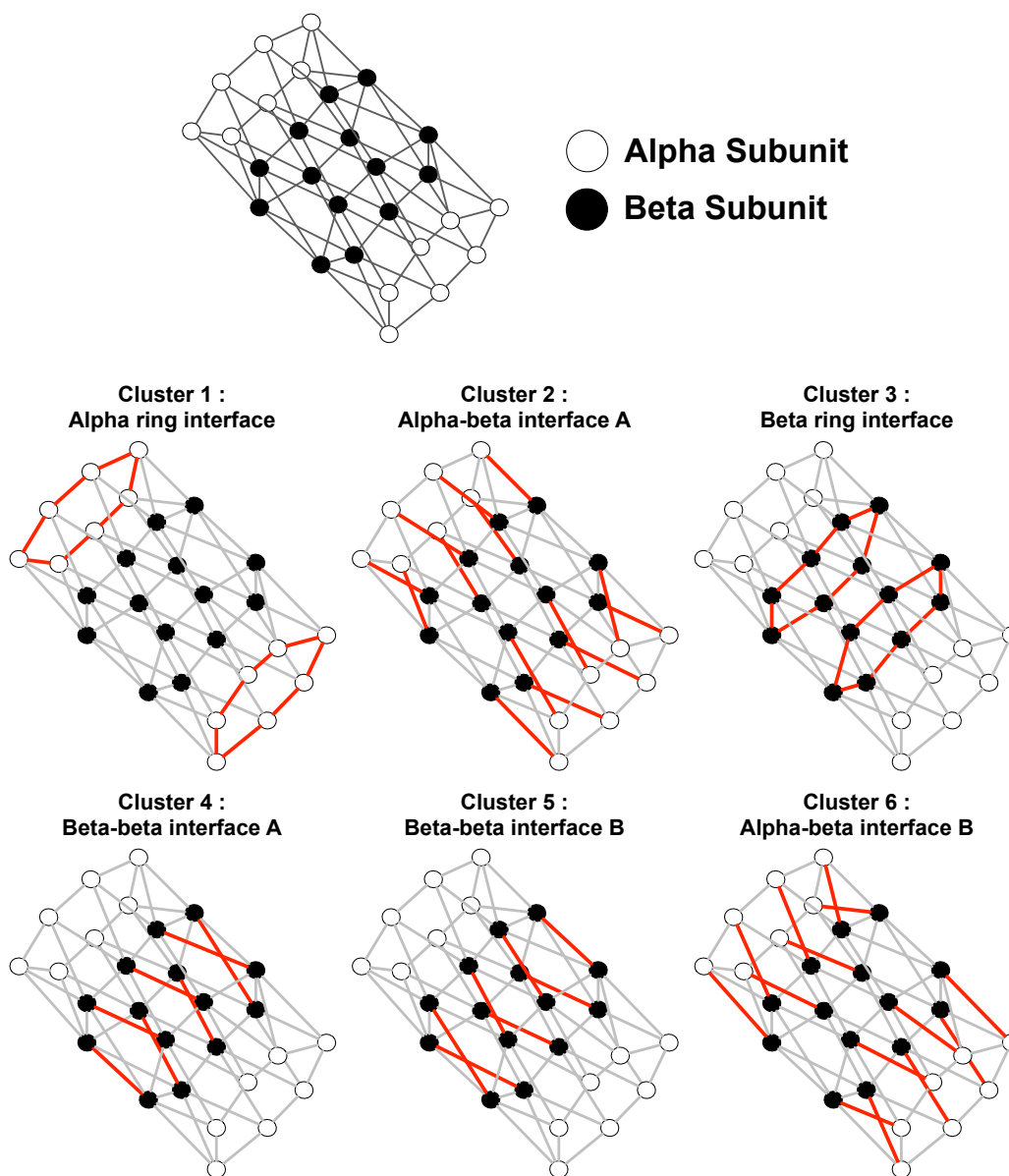


Figure 2.10: Visualisation of interaction clustering in the yeast proteasome. The nodes represent proteins, black nodes are beta subunits and white nodes represent alpha subunits. Edges represent interactions between proteins with two proteins being classified as interacting if any residue of one protein is within 4.5\AA of any residue in the second protein. For each cluster the interactions belonging to that cluster are highlighted red.

BLAST search. Complexes formed from one family of paralogs correspond to complexes whose homology graph is formed from one connected component. Homology graphs were constructed for each complex, allowing identification of 47 complexes of paralogs.

Given this set of complexes, interfaces are defined and then clustered as described in Section 2.2.4. For each interface type defined by the clustering and given one of these complexes, we can define interaction branches for our training set as follows: take two subunits from the complex to form the ancestor interaction node, take two subunits to form the child node and allow hypothetical sequence evolution between these two nodes. We can then define the interaction state of both nodes by defining pairs of subunits that interact using the interface type under consideration in the crystal structure to be interacting and to be non-interacting otherwise. We now have a hypothetical or simulated interaction tree branch, for which we know the ancestor and child interaction states. Using this method we produce a large set of interaction tree branches from the set of complexes defined above.

2.3 Results

2.3.1 Interaction Trees for Globular Proteins

To begin, three models of PPI evolution, as modelled by an interaction tree (see Methods) are compared, a complete description of each model can be found in Methods. Each model attempts to predict PPI gains or losses on a branch of the interaction tree based on the sequence changes along that branch (Figure 2.3). An interaction tree branch describes the evolution from a pair of ancestral proteins A_1 , A_2 to a pair of proteins C_1 , C_2 and can be decomposed into the two phylogenetic branches from A_1 to C_1 and from A_2 to C_2 . The first model, D_{dis} ,

predicts PPI rewiring on an interaction tree branch based on the total sequence change across its two constituent phylogeny branches, as calculated under the Jones-Taylor-Thornton model [20]. This is the same model used successfully to model the history of interactions within the bZip transcription family in [85]. The second model, D_{dif} is based on the observation that interacting protein families tend to have similar phylogenetic trees and quantifying that similarity has been used as the basis of PPI prediction [39]. As such, this model predicts PPI rewiring based on the similarity in branch length of the two constituent phylogeny branches. The final model considered, D_{com} , is different in that it uses structural information to predict PPI rewiring. This model requires a structural template to define the residue-residue contacts believed to mediate a PPI, once these are defined, we take a simple measure of chemical complementarity adapted from [98] and calculate its change along an interaction tree branch. PPIs are made possible due to the chemical complementarity between the contacting residues of the interacting proteins and so we hypothesise that changes in this measure of complementarity can predict PPI rewiring events.

2.3.2 Finding the Conditional Probability Distribution

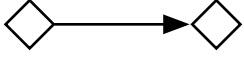

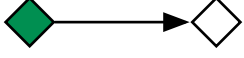
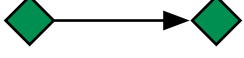
Each of the three models considered here aims to predict rewiring events based on some measure of the evolution occurring on a branch of the interaction tree. Table 2.4 shows the expected behaviour of each model along an interaction tree branch containing either type of rewiring event. These assumptions could be used to predict the occurrence of events. For instance, Table 2.4 shows we expect the largest values of D_{dis} when there is no gain of interaction along a branch (branch type A). If trying to distinguish these from branches containing a gain of PPI, a threshold α could be set and all branches having $D_{com} < \alpha$ predicted as containing

a gain of PPI.

However, in this case this description is not sufficient; a conditional probability distribution of the type shown in Table 2.1 is required to describe the probability of a rewiring event on a branch, given the change in D on that branch. This probability distribution is required by the interaction tree (Equation 2.1) to infer rewiring events during PPI evolution. These probabilities can be estimated directly given a training set of interaction tree branches for which we know the ancestral state, the child state and the value of D on each branch (Figure 2.11).

The problem here is that it is impossible to observe the ancestral state (or the value of D) of a true interaction tree branch as the ancestral proteins no longer exist. One solution to this problem is to produce hypothetical branches between present day, existing nodes (Figure 2.12). For instance, in generating a training set for the interaction tree shown in Figure 2.2 a hypothetical branch could be constructed between $B_{1a}-B_2$ and $B_{1b}-B_2$, with $B_{1a}-B_2$ being the hypothetical ancestor. Evolution is then imagined between these two nodes, for which we can observe the state of each node and calculate the value of D on the branch. We would like to know whether the basic assumptions outlined in Table 2.4 remain true for these branches. To begin, we consider how a hypothetical branch relates to the true interaction tree (Figure 2.12).

Given this relationship we can then calculate the expected value of each measure on each type of hypothetical interaction branch (Figure 2.13). We can see that for D_{com} the expected value for each hypothetical branch type matches that of the true branch types in Table 2.4 in every case. For D_{dis} and D_{dif} this is not the case, however, the relationship between the values on each branch type is maintained. This shows that despite the recourse to hypothetical evolution to calculate the conditional probability distribution in Equation 2.1, the resulting

	Branch type	D_{dis}	D_{dif}	D_{com}	
A		+	+	0	+
B		δ	δ	+	δ
C		δ	δ	-	ϵ
D		ϵ	0	0	0

+

δ

ϵ

0

-

positive

small positive

smaller positive

zero

negative

Table 2.4: Expected behaviour of the three evolutionary measures on the four interaction tree branch types. Interacting nodes are shown in green and non-interacting in white, for instance branch type B shows a gain of interaction. For D_{dis} we assume that interacting proteins evolve slower at the interface [101], leading to a prediction of the greatest distance on branch type A, during which the proteins never interact, an intermediate distance on branches B and C on which the proteins interact for some part of the branch and the smallest D_{dis} for branch D throughout which the proteins interact. For D_{dif} , we assume that interacting proteins evolve at similar rates and so have similar branch lengths in the phylogeny [97]. This leads to prediction of positive D_{dif} in branches A,B and C and zero D_{dif} on branch D (corresponding to symmetric branch lengths). Finally for D_{com} we assume that interacting proteins have a higher complementary fraction [98]. This leads to a prediction of zero D_{com} on branches A and D and positive/negative D_{com} on branches B/C. It is worth noting that D_{com} is the only non-symmetric measure, in that branches B and C have different predictions, meaning that the direction of the arrow in the branch is important.

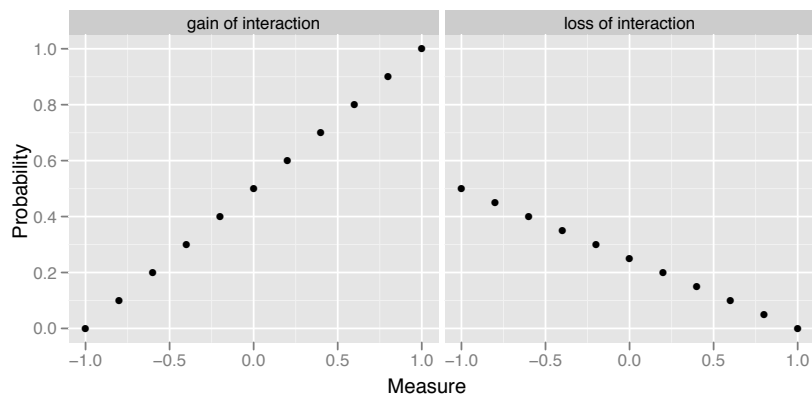
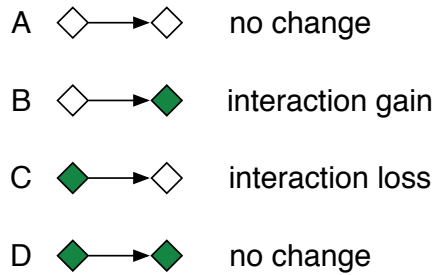


Figure 2.11: Probabilities for gain and loss of interaction given a change in some measure along a branch of an interaction tree. To compute these probabilities we first consider individual branches of an interaction tree (top left, grey diamonds indicate no interaction, green diamonds indicate interactions) and specifically what events can occur on a branch e.g. no change in state, gain or loss of interaction. Given a set of such branches to use as a training set, we first compute the change in our evolutionary parameter across each branch and then bin the branches according to the amount of change (bottom, here we are using ten bins). We then calculate a probability of gain of interaction for each bin by counting the proportion of branches in each bin starting with non-interaction ancestor, that finish with an interacting child (i.e. using the labels in the figure $\text{total}(B)/\text{total}(B)+\text{total}(A)$). Probability for interaction loss is computed in the same way as $\text{total}(C)/\text{total}(C)+\text{total}(D)$. In the plotted example we have some measure associated with interaction change e.g. change in binding affinity, and after binning the branches in to ten bins we see that a positive increase in this measure gives higher probabilities for interaction gain and vice versa for interaction loss.

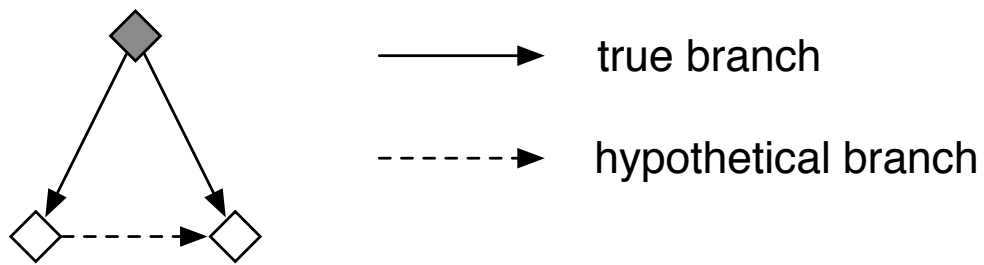


Figure 2.12: A hypothetical interaction tree branch and its relation to the true interaction tree. The hypothetical branch is represented as a dashed arrow between two existing interaction nodes (white diamonds). These existing nodes have a most recent common ancestor (MRCA) interaction (grey diamond) in the interaction tree, much in the same way that sequences have a MRCA in sequence phylogenies. This allows us to describe the evolution along our hypothetical branch as the sum of evolution across the two true branches (solid arrow) linking the end nodes to their MRCA. This allows us to determine if inferences made from hypothetical branches are valid given our assumptions in Table 2.4.

probability distribution is still expected to reflect the true evolutionary process.

2.3.3 Evaluating the models

In order to evaluate the 3 models, a test case is first required. As a test system we will be using the proteasome, a large, compartmentalised protease present in eukaryotes and archaea with a homologous HslV protease in bacteria [74]. This complex is responsible for degrading misfolded or damaged proteins that have been targeted with ubiquitin and the protease activity of the complex is also responsible for regulating cellular processes such as cell division. The proteasome consists of a central 20S core particle along with associated regulatory complexes such as AAA+ ATPases [102]. We consider here only the 20S core particle, consisting of alpha and beta-type subunits arranged in four stacked heptameric rings, two outer alpha subunit rings and two inner beta subunit rings. The homologous bacterial HslV complex consists of two stacked hexameric rings of identical subunits. The alpha, beta and HslV subunits are all homologous. This is a well studied complex, with well understood structure and structural examples across a wide range of organisms, making a useful system to evaluate the method.

Having defined these three models of PPI evolution, they are now evaluated in their ability to predict PPI rewiring in the proteasome complex from *Saccharomyces cerevisiae*. To begin, structure 1RYP [103] was downloaded from the Protein Data Bank [104] and the sequences of all subunits aligned with MUSCLE [93], with some manual realignment. As explained above, the D_{com} model of PPI evolution requires a template structure for an interaction in order to be calculated. The problem with defining such a structure for the proteasome is that proteasome subunits interact in a variety of orientations and so in evaluating the D_{com} score it is required to specify the particular orientation. To classify all possible orientations

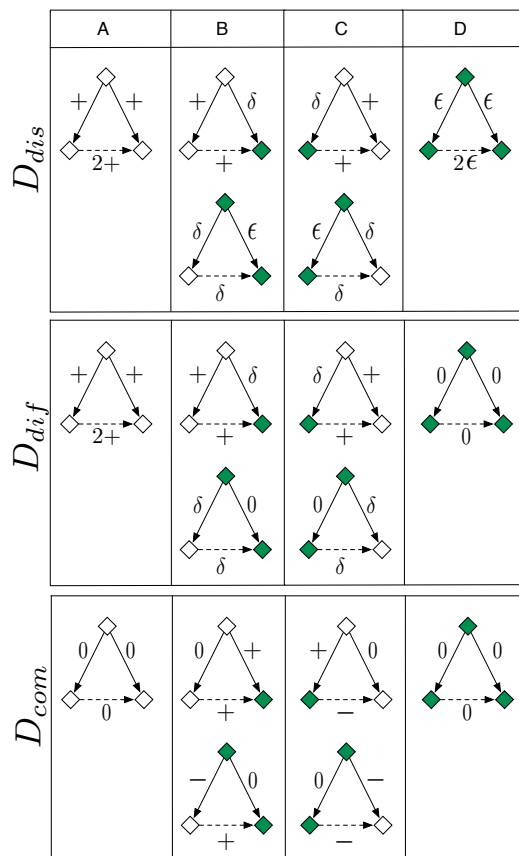


Figure 2.13: Expected value of the three measures of interaction evolution on each of the four types of hypothetical interaction tree branch. We consider every possible combination of 'on' and 'off' nodes for which only one change has occurred on either true branch since the MRCA. Here we assume that $+ \gg \delta \gg \epsilon$. So that, for instance, $\delta + \epsilon \simeq \delta$. There are differences to be noted when calculating the expected value on the hypothetical branch for each measure. The value of D_{dis} and D_{dif} along the hypothetical branch is just the sum of the distances across the true branches. This makes sense as evolutionary distance is the same going up or down a branch. However, D_{com} is not symmetric in this way (see Table 2.4) and so the value of D_{com} changes depending on which direction you travel along a branch. When calculating D_{com} on the hypothetical branches, this has the effect of reversing the sign on the leftmost true branch, as we have to travel against this arrow and then down the rightmost arrow to describe the evolution occurring on the hypothetical branch.

a clustering algorithm was applied based on overlapping amino acid contacts (see Methods), identifying 6 PPI types in the proteasome, with symmetric arrangement within the complex (Figure 2.10). For instance, cluster 3 contains all PPIs forming the ring of beta proteins. All analysis can now proceed independently for each PPI orientation.

To evaluate the three models of PPI evolution within the proteasome we require a set of interaction tree branches for which we know the interaction state of the ancestral proteins and the child proteins. As explained in Section 2.3.2 it is impossible to use true interaction tree branches for this task and so we construct hypothetical branches between existing nodes. Using the yeast proteasome, a large set of test branches is built in this way; two existing proteins are chosen to form the ancestral interaction node and two to form the child node. This is repeated until all combinations of existing proteins have been chosen to form an interaction tree branch. For each of the branches in the training set, the nodes are classified as interacting if their constituent proteins are within a distance of 4.5\AA in the 1RYP structure. The value of D can also be calculated on each of these branches, producing the required training set. Before the models are used to generate a CPD, each model is evaluated on its ability to detect both types of rewiring, that is gains of interaction (i.e. distinguishing branch type B from type A) and loss of interaction (C vs D) and a ROC curve is produced for each (Figures 2.14, 2.15).

In predicting losses of interaction, the D_{com} model performs best but all three models perform better than random suggesting that a loss of PPI between two proteins increases the rate and asymmetry of substitutions whilst decreasing the chemical complementarity at the protein-protein interface. In predicting gains however, D_{com} clearly outperforms the other methods, the two models based on sequence distances alone do not produce accurate predictions of interaction gains. It is worth noting here that using hypothetical interaction branches in place of real

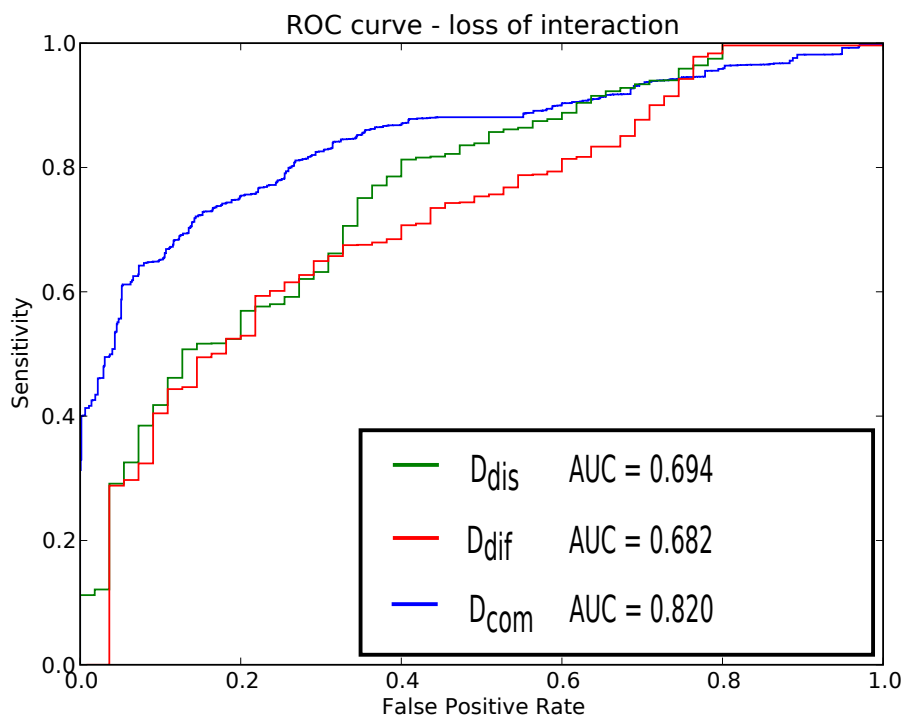


Figure 2.14: ROC curves comparing each of the 3 models of PPI evolution in their ability to predict both losses (left) and gains (right) of interaction.

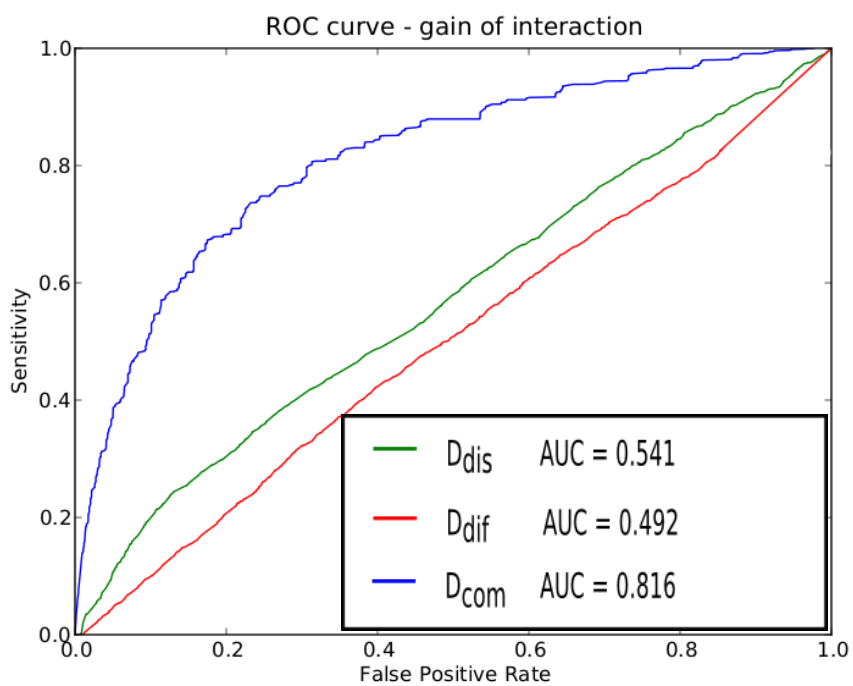


Figure 2.15: ROC curves comparing each of the 3 models of PPI evolution in their ability to predict both losses (left) and gains (right) of interaction.

branches could distort the behaviour of D_{dis} and D_{dif} as set out in Supplementary Material.

2.3.4 Calibrating the Model

From the previous ROC analysis it is concluded that D_{com} is the most promising candidate for developing an interaction tree model to apply to complexes of paralagous subunits. The model will now be used to produce a CPD of the type shown in Table 2.1. In order to do this in general and not just in the proteasome, a large set of large protein complexes, composed of paralagous subunits was formed, (Section 2.2.5) cluster the PPIs and construct hypothetical interaction branches as before for each complex.

This allows plotting of changes in the D_{com} score against probability of both types of rewiring events, as described in Figure 2.11 across this large, non redundant set of complexes (Figure 2.16). This shows that there is a generalisable relationship between the D_{com} score and rewiring events and suggests that a interaction tree model with branch-specific rewiring probabilities based on D_{com} could be applied to complexes of paralogs. Specifically, after a gene duplication event, changes in the D_{com} score between the duplicates and all other possible interaction partners could be used to predict subfunctionalization (loss of interactions) and neofunctionalization (gain of interactions). The probability of such events can be defined according to fitted exponential functions as plotted in Figure 2.16.

The fitting of exponential curves to the results was done using the *nls* command in R. The form of the exponential functions used are shown below in Equations 2.12, where D is the D_{com} score and A_g , B_g , A_l , B_l are parameters to be found. The fit parameters were $A_g = 0.670$, $B_g = -2.195$, $A_l = 0.995$, $B_l = 23.314$.

$$D_{com}(A, C) = x$$

	$P(C = on)$	$P(C = off)$
$A = on$	$1 - (A_g - A_g e^{B_g x})$	$A_g - A_g e^{B_g x}$
$A = off$	$A_g e^{B_g(1-x)} - A_g e^{B_g}$	$1 - (A_g e^{B_g(1-x)} - A_g e^{B_g})$

Table 2.5: A representation of Equation 2.1 in matrix form. The matrix shown is for $D=10$ and describes the probability of C given A . For instance, if the ancestor node is non interacting, there is a 10% chance that the child node is interacting if $D=10$, that is, there is a 10% chance of a gain of interaction.

$$P(gain) = A_g e^{B_g(1-D)} - A_g e^{B_g} \quad (2.12)$$

$$P(loss) = A_g - A_g e^{B_g D} \quad (2.13)$$

This now allows a description of the conditional probability distribution (as described in Table 2.1) for any value of D_{com} on an interaction tree branch. The probabilities forming the probability table are taken from Equations 2.5 and 2.6, substituting the D_{com} value for that branch (Table 2.5) with the fact that the rows of the table sum to 1 used to find the other two probabilities.

2.4 Discussion

This chapter presents a method for predicting PPIs and evolutionary history within protein complexes, specifically obligate complexes formed of homologous subunits. The method is based upon the previously described interaction tree approach which attempts to explicitly model the process of PPI evolution, composed of rewiring

events and protein gains and losses. Most previous interaction tree approaches have modelled rewiring events uniformly throughout the evolution described, that is the probability of a rewiring event is the same throughout the tree. The exception is [85] in which the probability of a rewiring event was related to the substitution rates of the proteins in question, in particular that increased amino acid substitution is predictive of rewiring events. This led to accurate evolutionary inference in the bZip family of transcription factors.

The aim in this chapter is to apply a similar model of rewiring to protein complexes. It was shown initially that a model based on substitution/ phylogeny branch length as in [85] does not predict rewiring events in an example complex. This could be due to the increased structural complexity of the interaction as opposed to the simple coiled coil interactions studied in [85]. Prediction is also attempted unsuccessfully based on similarity of branch length, as this has been used to predict PPIs at the family level elsewhere. A successful model was found based on a simple measure of physico-chemical complementarity that requires a template structure for the interaction and predicts change in the complementarity of two proteins as their sequences evolve. This allows prediction of rewiring events between two proteins as their sequences change, the downside being the requirement of a structure for the interface.

2.5 Conclusion

The D_{com} model of PPI evolution can predict gains and losses of interaction (rewiring events) in complexes formed of homologous subunits. This is generally true for a large set of such complexes and this relationship offers a way of reconstructing the evolutionary history of such complexes using the interaction tree

methodology. This can be done in a way that is branch specific unlike previous attempts to model complex evolution using the interaction tree.

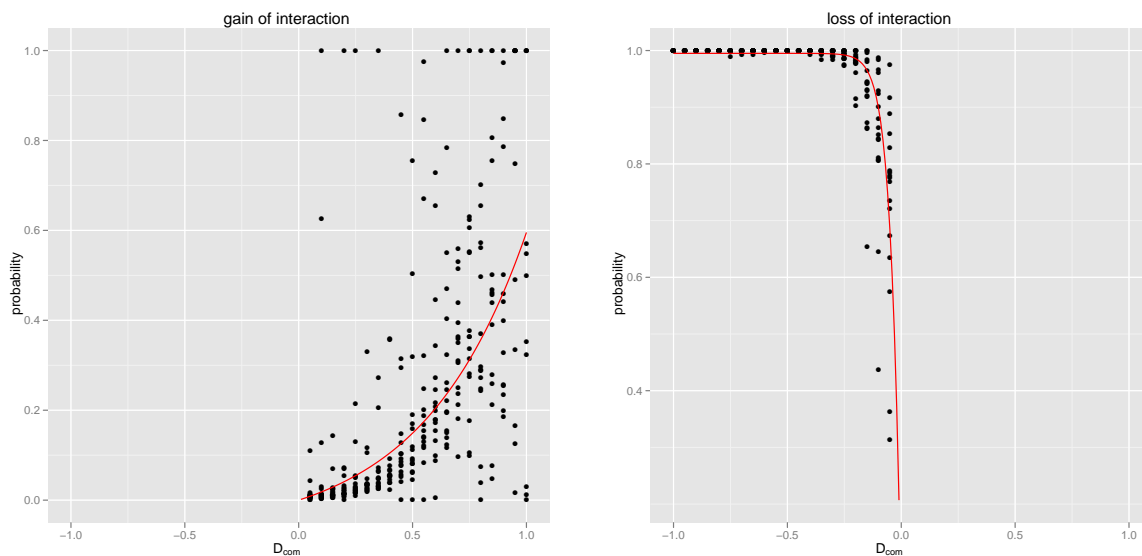


Figure 2.16: The relationship between the change in interaction score and probability of interaction gain and loss across the entire training set of paralog complexes. All data is binned in the same way and over-plotted for all interaction clusters for all complexes to produce the scatter plots shown above. We assume that for negative changes in the interaction score the probability of interaction gain is 0 and for positive changes the probability of interaction loss is 0. Then, to each scatterplot we fit a logistic curve (red lines) to describe the general relationship between change in D_{com} score and interaction gain and loss.

Chapter 3

Modelling PPI rewiring in protein complexes

3.1 Introduction

In the last chapter, a new approach to using the interaction tree method was defined and validated. This methodology addressed several problems with previous interaction tree applications, primarily the fact that previous attempts have assigned a constant probability of PPI rewiring after duplication across all proteins. The approach described here bases the probability of rewiring on the change in amino acid composition at the protein-protein binding site and so allows probability of rewiring, specific to each protein, dependent on the evolutionary events happening in the sequence (amino acid substitution). This method was specifically designed with a problem in mind: the reconstruction of the history of protein complexes of paralogs.

In this chapter the methodology is applied to that problem. As a test case, the proteasome and its homologues are chosen. The 20S proteasome is a multi-

meric protease, responsible for degrading misfolded proteins and controlling gene expression through targeting proteins tagged with ubiquitin for destruction. Some eukaryotes have been observed to express a variety of proteasomes, with subunits being substituted to produce a complex with subtly different function. Most bacterial species do not have a 20S proteasome but the related HslV complex. This protease has a similar structure but is less complicated in terms of the number of subunits and their variety. This variation throughout the tree of life, paired with the large amount of solved crystal structures of this complex, make it a suitable first test case for the methodology.

Previously, scenarios for the evolution of this complex have been proposed, based on reasoning from the principle of parsimony. The detailed modelling here will allow testing of these arguments and also, a deeper understanding of the emergence of the topology of the complex. For instance, [105] proposed that a complex with a barrel-like structure would have evolved either from a single ring structure or a simple dimer structure. Each of these ancestral structures relates to a specific binding site between the subunits of the complex, i.e. the site responsible for forming a ring structure of the symmetric binding site responsible for forming a dimer. The ancestral complex can then be predicted by predicting what is the oldest binding site within the complex.

By clustering the interactions in a complex (as described in the last chapter), these distinct binding sites can be defined and the interactions occurring at these sites tracked. This allows predictions for PPI evolution for each interface type, including predicting the probability of each binding site being ancestral. In this chapter, the methodology is first validated through prediction of known interactions in the cattle proteasome using evidence of interactions from other species. Then, the method is used to infer the evolutionary history of the proteasome complex. Insights are gained in to the evolution of the complex, including evidence as

to what the ancestral complex's quaternary structure may have been.

3.2 Methods

3.2.1 Structural and sequence data

In the previous chapter, the yeast 20S proteasome was used as a test set for proving the suitability of the D_{com} metric as a basis for a model of PPI evolution. A larger training set of complexes was then used to fit the model. Now we return to the proteasome, to apply this model to infer the evolutionary history of this complex.

Before showing how this inference can be performed using the interaction tree methodology, a description of the distribution and variation of proteasomes across species is given. As described in the previous chapter and shown here in Figure 3.1, the yeast proteasome consists of 4 stacked rings of 7 subunits. These subunits can be categorised into alpha or beta subunits based on sequence similarity. The beta subunits being catalytically active and the alpha subunits regulating access to the active sites of the beta subunits. In this structure there are 7 unique beta subunits in each beta ring and similarly 7 unique alpha subunits in each alpha ring. This gives a total of 14 unique proteasome proteins in this 28-mer.

This is the typical eukaryote 20S proteasome. Whilst all eukaryote proteasomes are 28-mers built of 14 unique proteins, some variation is seen in the number of proteins available in the genome for formation of proteasomes. For instance, in mammals an alternative proteasome, the immunoproteasome, has been discovered. This complex substitutes 3 of the beta subunits of the standard proteasome for 3 new subunits, altering the activity of the complex.

The archaeal proteasome is also formed of 28 subunits. However, in this case all alpha subunits are identical, as are all beta subunits. Figure 3.1 shows a typical

archaeal proteasome from *T. acidophilum*. The complex is also formed of 4 stacked rings of 7 subunits, however, here the alpha subunits forming the end of the barrel structure are in an "open" conformation. As such, the channel leading to the cavity containing the active sites can be seen in the top down view.

With a few exceptions, bacteria do not possess a 20S proteasome. However, most do have the homologous HslV complex (bottom of Figure 3.1). This complex is formed of just one repeated subunit that is homologous to the proteasome subunits and more similar to the beta subunits than the alpha subunits. This single protein assembles into two stacked hexameric rings forming the HslV protease.

Given this variation amongst proteasomes and homologous structures there are several questions concerning its evolution, such as; what was the ancestral complex? was it some simpler HslV like structure? How did the eukaryotic structures come to have extra subunits? Were these all gained in quick succession?

In order to answer some of these questions, in this chapter the D_{com} model will be used in conjunction with the interaction tree to infer the history of PPIs within this complex. As before, the D_{com} model requires some structural example(s) of PPIs to be calculable. A set of proteasome structures listed in Table 3.1 are included in this analysis for this purpose.

The interaction tree framework allows inference of the history of the interactions in these structures, using the D_{com} model. We already know the PPIs present in these species (from the solved structures) however, the interaction tree can also infer the PPIs present in species for which we do not have this information. To this end, a set of extra proteasome sequences, from species for which we have no structural information, are included in the analysis (Table 3.2) The expanded set also contains some extra *Bos taurus* proteins that are absent from the solved 1IRU structure [106]. This set was generated by firstly choosing a set of species with a wide coverage of the tree of life and then taking all sequences classified as

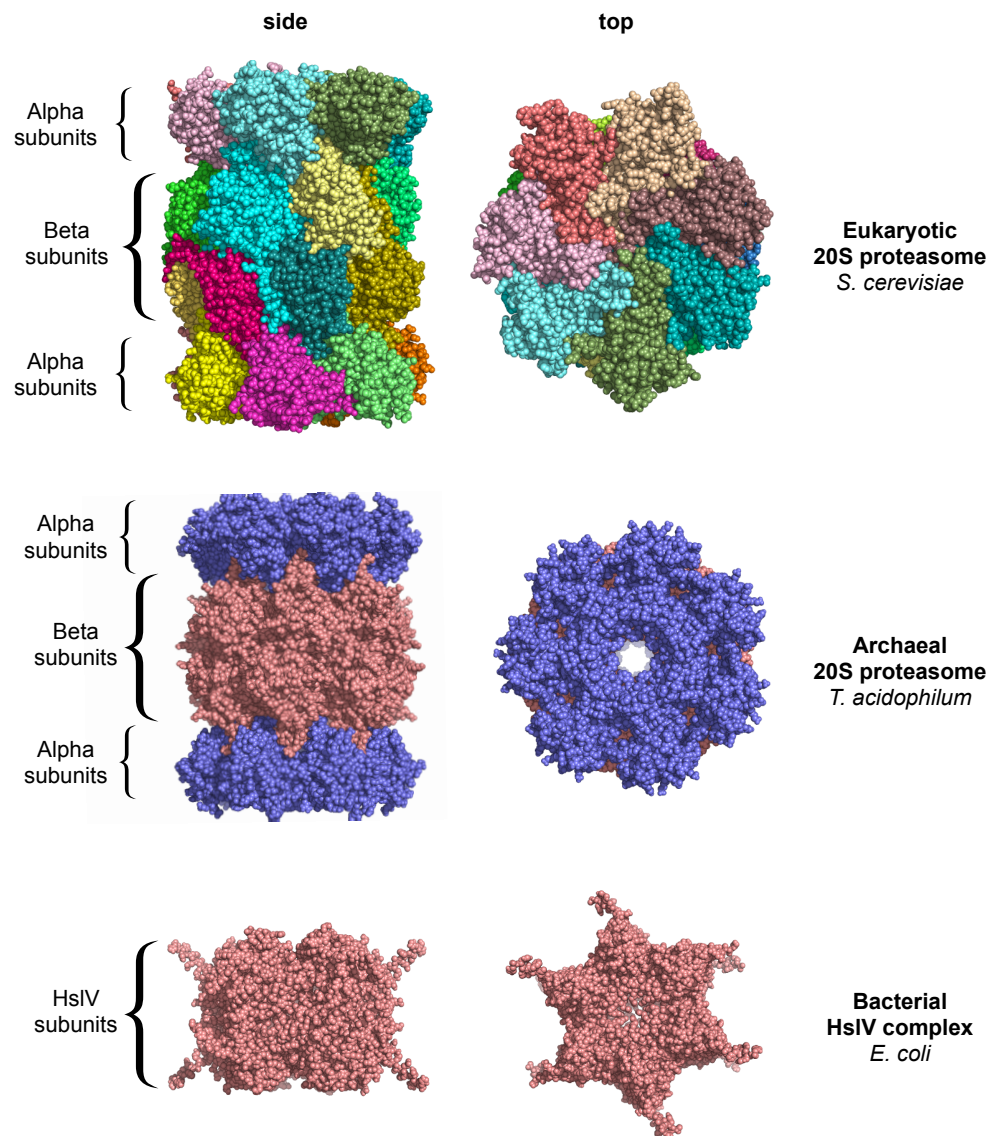


Figure 3.1: A selection of proteasomes and proteasome-like complexes described in the text.

PDB code	Species	Number of subunits	Unique subunits
1NED	<i>Escherichia coli</i>	12	1
1G3K	<i>Haemophilus influenzae</i>	12	1
1M4Y	<i>Thermotoga maritima</i>	12	1
1PMA	<i>Thermoplasma acidophilum</i>	28	2
3H4P	<i>Methanocaldococcus jannaschii</i>	28	2
1RYP	<i>Saccharomyces cerevisiae</i>	28	14
1IRU	<i>Bos taurus</i>	28	14

Table 3.1: List of proteasome structures used in the analysis, in increasing order of complexity. The three bacterial complexes are HslV proteases formed of two stacked hexameric rings of identical subunits. The two archaeal proteasomes are composed of four heptameric rings, two stacked rings of beta subunits capped on either end by a ring of alpha subunits. The alpha subunits are all identical in these complexes, as are the beta subunits. The eukaryote complexes are the most complicated, having the same subunit topology as the archaeal structures but 7 unique alpha subunits and 7 unique beta subunits.

belonging to the "proteasome subunits" family in the Superfamily [107] database.

3.2.2 Phylogeny building

In order to build a phylogeny for the expanded set of proteasome sequences, the set was first aligned using MUSCLE [93] with default parameters. As mentioned in Table 3.2, *P. falciparum* has a protein that appears most similar to the bacterial HslV proteins (PfHslV) [108] and clustered with them in initial phylogeny building attempts. It is postulated that this is the result of horizontal transfer and due to the fact that the interaction tree algorithm can not currently incorporate horizontal events, this protein was removed from the analysis.

Species	Alpha subunits	Beta subunits
<i>Bordetella bronchiseptica</i>	0	1
<i>Rickettsia prowazekii</i>	0	1
<i>Helicobacter pylori</i>	0	1
<i>Aquifex aeolicus</i>	0	1
<i>Bacillus subtilis</i>	0	1
<i>Haloferax volcanii</i>	2	1
<i>Natronomonas pharaonis</i>	1	1
<i>Pyrococcus furiosus</i>	1	2
<i>Plasmodium falciparum</i> *	7	7
<i>Arabidopsis thaliana</i>	14	14
<i>Dictyostelium discoideum</i>	7	7
<i>Bos taurus</i> **	1	5

Table 3.2: List of extra proteasome sequences included in this analysis. Here the bacterial HslV proteins are listed as Beta subunits as they are more similar to this subfamily. We include *H volcanii* as this archaea has two alpha subunits compared to most archaea which have one. We similarly include *P furiosus*, as this archaeon has 2 beta subunits. * *P falciparum* also has a HslV-like protein (PfHslV) but this was removed from the analysis due to problems including it in the phylogeny. ** These are extra proteasome proteins found in *B taurus* but not present in the 1IRU structure described in Table 3.1

This produces an alignment of 110 proteasome or HslV sequences, with an average pairwise identity of 25.94%. To begin phylogeny construction, 100 bootstrap samples are taken using the SeqBoot program from the Phylip [96] suite for phylogenetics. A bootstrap sample consists 565 columns sampled with replacement from the 565 columns of the alignment, to produce a bootstrap alignment. These 100 alignments were then used to generate 100 neighbour-joining trees using the Neighbour program from Phylip. From these 100 trees a consensus tree is generated using the Consense program from Phylip, with the Majority Rule (extended) algorithm. To summarise this algorithm, each internal node in a tree corresponds to a set of proteins (the proteins present in the clade descendant from that node). For each internal node, Consense counts how many of the 100 trees the clade is present, with any clade present in more than 50% of the trees included in the final consensus tree. From this initial consensus tree, internal nodes with less than 50% support are then considered in decreasing order of support and added to the tree if possible (some may be inconsistent with the internal nodes already added to the tree, e.g. if (a,b,c) is a clade then (b,c,d) can not be). This is continued until the tree is fully binary, that is, each internal node has exactly 2 descendants. This produces a consensus tree in which each internal node has an associated bootstrap value indicating the support for that node in the alignment. The resulting tree for the proteasome data is described in the Results section.

Before a phylogeny can be used to construct an interaction tree, it must be reconciled with a species tree, as described in Chapter 1. Each internal node in the phylogeny corresponds to either a speciation event or a duplication event; proteins can diverge either after duplication (within a species) or after speciation (in separate species). Reconciliation classifies internal nodes as either the result of speciation or duplication events and this in turn defines which proteins were present in a species at the moment of speciation (the speciation nodes correspond

to these). This knowledge of the proteins present at speciation is required to construct the interaction tree as described in the previous chapter. To begin reconciliation, a species tree is first required. Here, a species tree describing the relationship between all species included in the analysis is downloaded from ITOL [109] (Figure 3.2).

The reconciliation is then performed using the NOTUNG [34] program with default parameters. A given reconciliation implies gene losses in certain species, for instance, after a speciation event a gene/protein may be lost in one of the species (see Chapter 1 for an example). NOTUNG performs reconciliation by attempting to minimise both the number of duplication events and the number of these gene losses in the resulting tree. Thus the reconciled tree follows the principle of parsimony by containing the minimum number of events to explain the observed gene phylogeny in terms of the species tree. NOTUNG weights the duplication and loss events separately; by default a weight of 1.5 is given to duplication events and a weight of 1 to loss events.

In parts of the phylogeny, inferring the correct branching order may be difficult and low bootstrap values will result. For instance, if there are many speciation events inferred in a short space on the tree, getting these events in the right order and with a high bootstrap support may be difficult. Subsequently, during reconciliation, extra duplication and loss events will be inferred by NOTUNG to account for this discrepancy. To help in this situation, NOTUNG offers a rearrange function that can reorder internal nodes with bootstrap values less than a given cutoff, in order to agree with the branching order in the species tree. This follows the principle of parsimony in that in cases where the bootstrap values do not give strong support for a particular tree topology, the topology that requires less duplication/loss events (i.e. the topology that follows that of the species tree) should be preferred.

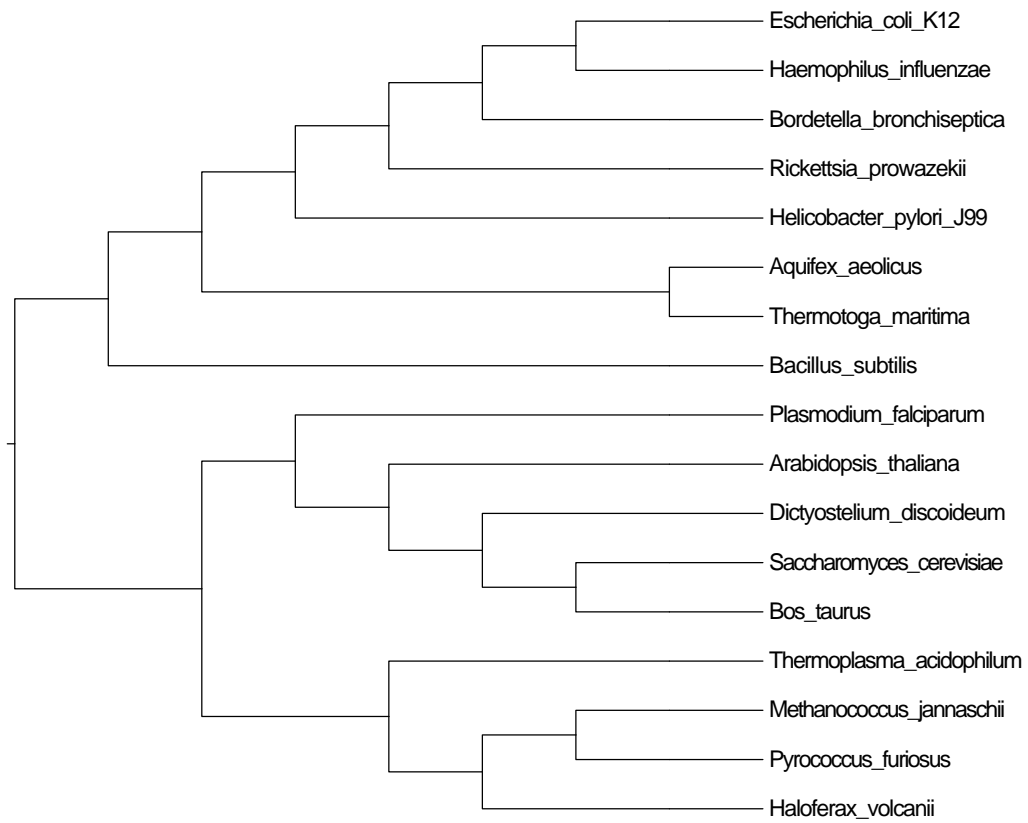


Figure 3.2: The species tree used for reconciliation with NOTUNG, taken from the ITOL website.

3.2.3 Sequence reconstruction with PAML

Given a phylogeny that has been reconciled as above, an interaction tree can be constructed as described in the previous chapter. Then, to use the D_{com} model, the D_{com} score has to be calculated for every branch in the interaction tree. To do this sequences are required for the interior nodes of the phylogeny. The PAML [110] package can be used to reconstruct the ancestral sequences using a likelihood based approach. This has an advantage over parsimony based methods (Fitch 1971) in that rather than one likely sequence being inferred, a distribution of sequences is produced at each node with a posterior probability assigned to each one. It has been previously observed [111] that using such a distribution as opposed to point estimates may be preferable.

In order to calculate the D_{com} value on each branch, the complementary fraction ($F()$ in the previous chapter) has to be calculated at the ancestor node and the child node of that branch. To estimate the complementary fraction at an interaction node, the following procedure is undertaken; a sequence is sampled, for each of the constituent proteins of the interaction node, from the sequence distribution calculated by PAML. The complementary fraction is calculated between these sequences as usual, this process is repeated 1000 times and then the average is taken over these 1000 samples. This gives an estimate of the complementary fraction at an interaction node, allowing the change along an interaction branch to be calculated, giving the D_{com} value.

3.3 Results

3.3.1 Proteasome phylogeny

To begin with the extended set of protein sequences, included those listed in Table 3.2, is aligned using muscle and a bootstrapped phylogeny built as described in Section 3.2.2. As described in the Methods section, this tree contains all identifiable proteasome subunits for each included species in Figure 3.2 except for *P. falciparum*. This species has an extra proteasome subunit (PfHslV) which is more similar to the bacterial HslV proteins than the Eukaryotic proteasome subunits. It is postulated that this is the result of some horizontal transfer and so it has been excluded as the interaction tree construction assumes vertical descent in the phylogeny. The consensus phylogeny, before any reconciliation, is shown in Figure 3.3.

The tree as shown was then reconciled with the species tree, allowing for rearrangement of branches with low bootstrap support as described in Methods. The reconciled and rearranged tree is shown in Figure 3.4. Clades corresponding to the different types of proteasome or proteasome-like proteins are identified in the tree, showing the phylogeny building process is working as expected. After rearrangement the tree is more congruent with the species tree, for instance the *B. subtilis* HslV branches first (red clade, right of Figure 3.4) in the reconciled tree as *B. subtilis* branches first of the bacteria in the species tree. The sub clades within the eukaryote Alpha and Beta clades, correspond to the seven distinct alpha and beta subunits in eukaryotes, Figure 3.5. The tree shows that there were 6 consecutive duplications of both the alpha and beta proteins, prior to the divergence of the eukaryotic species but after the divergence from archaea, to produce the diverse set of subunits in their proteasomes. The 7 alpha proteins come together to form a heptameric ring, as do the beta proteins. No such duplications happened in

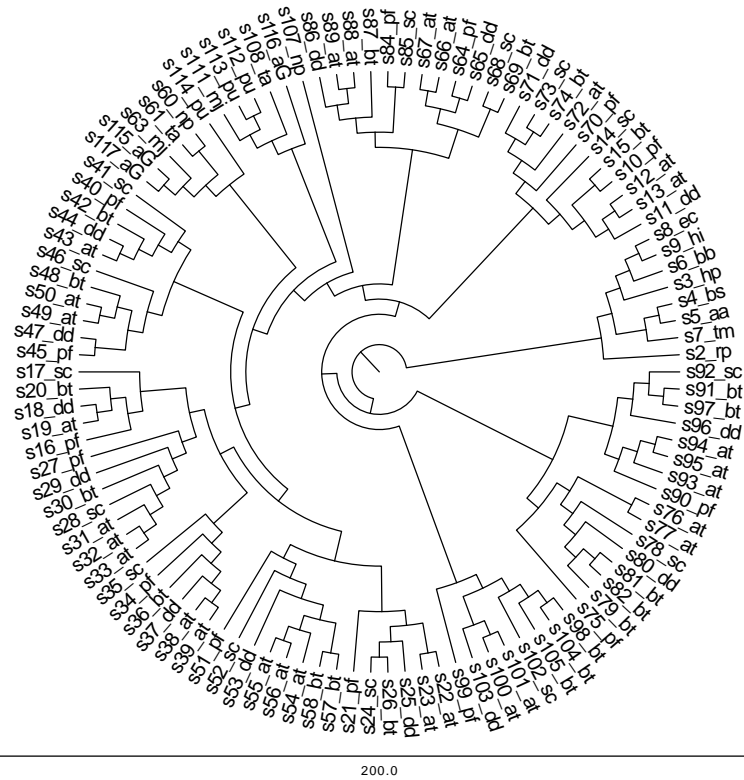


Figure 3.3: The bootstrap tree constructed for the set of proteasome subunits. Each protein has been given a unique identifier and is suffixed by a short species id (listed in Table 3.3). It can be seen that the proteins cluster as expected, for instance the bacterial HslV proteins form a clade (right of diagram). However, the branching order differs in most cases from that of the species tree, e.g. the HslV proteins branch in a different order.

Species	Short ID
<i>Escherichia coli</i>	ec
<i>Haemophilus influenzae</i>	hi
<i>Bordetella bronchiseptica</i>	bb
<i>Rickettsia prowazekii</i>	rp
<i>Helicobacter pylori</i>	hp
<i>Aquifex aeolicus</i>	aa
<i>Thermatoga maritima</i>	tm
<i>Bacillus subtilis</i>	bs
<i>Plasmodium falciparum</i>	pf
<i>Arabidopsis thaliana</i>	at
<i>Dictyostelium discoideum</i>	dd
<i>Saccharomyces cerevisiae</i>	sc
<i>Bos taurus</i>	bt
<i>Thermoplasma acidophilum</i>	ta
<i>Methanococcus jannaschi</i>	mj
<i>Haloferax volcanii</i>	aG
<i>Natronomonas pharaonis</i>	np
<i>Pyrococcus furiosus</i>	pu

Table 3.3: Long species names and two letter Superfamily species ids used in this analysis.

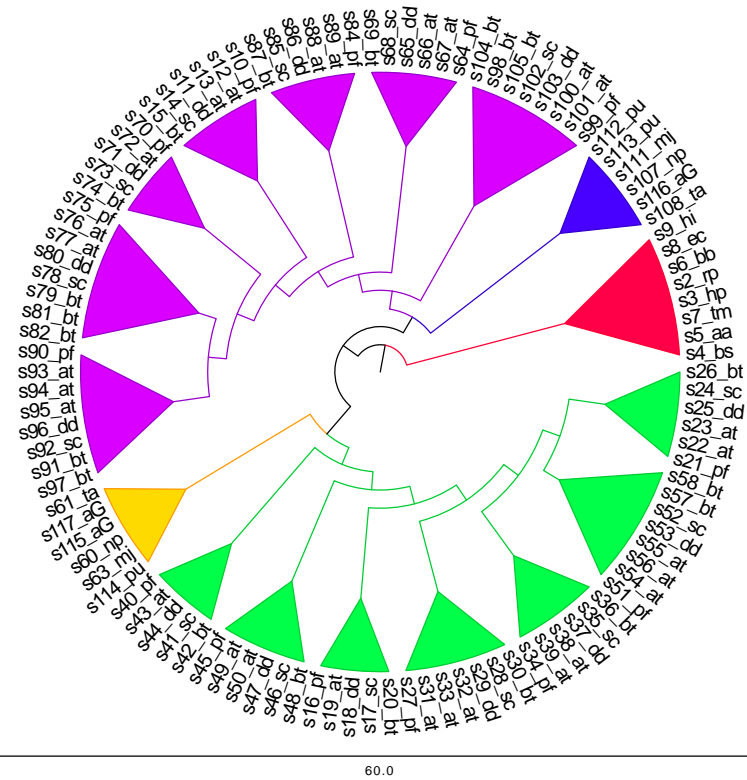
the archaea and so it is that the archaeal species simply have one alpha protein repeated 7 times in the alpha ring and one beta protein repeated 7 times in the beta ring. There have also been lineage specific duplications, particularly in *A. thalania*, producing extra subunits within some species.

3.3.2 Predicting present day structures

To begin, the ability of the method to predict present day structures will be tested. The interaction tree approach can be used to infer the history of protein interactions but can also be used to infer probability of interaction between existing proteins; observed PPIs can be used as input and the evidence propagated up the interaction tree and then, propagated back down to the leaves, in order to produce probability of interaction between proteins in species for which we have no PPI data.

To demonstrate the use of the interaction tree and the D_{com} model in this way, a simple toy example is used (Figure 3.6). This example can be thought of as a simplified version of the proteasome data set. In this simplification only yeast and cow are included in the analysis and each has one alpha subunit (CowA, YeastA in diagram) and one beta subunit (CowB, YeastB in diagram). To begin, a phylogeny is constructed, as described previously, for these proteins (top of figure). In the previous chapter, the construction of an interaction tree to describe interactions between two distinct protein families was described. In this example we wish to look at interactions *within* a family and so the phylogeny is combined with a duplicate of itself in order to produce the interaction tree structure (middle of figure).

Now, assume that we are using crystal structures to decide which present day proteins interact and a structure for PPIs between these proteins is only available



60.0

Figure 3.4: The reconciled and rearranged tree of proteasome subunits. Clades are identified corresponding to HslV subunits (red), beta subunits from Archaea (blue), beta subunits of eukaryotes (purple), alpha subunits from archaea (yellow) and alpha subunits from eukaryotes (green). The tree has 35 inferred duplication events and 0 inferred gene losses.

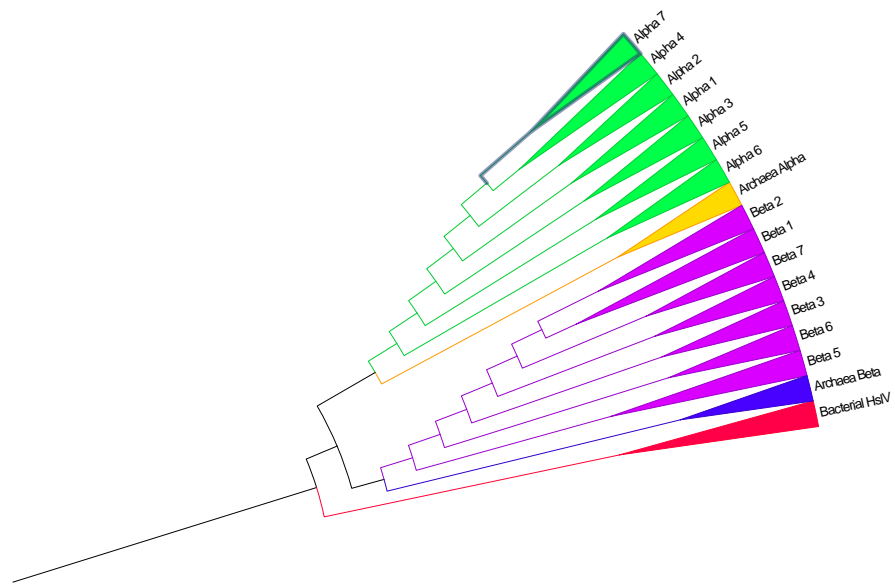


Figure 3.5: A simplified representation of the reconciled tree, clades are collapsed to show the relationship between the classes of subunits, using the same colour scheme as before.

in yeast. Firstly, we cluster the PPIs in the structure as before, giving the same 6 proteasome PPI clusters. We then consider each PPI cluster separately, let's say we start by looking at cluster 3: the beta ring interface. Using a distance threshold as before, we know from the yeast structure that the YeastB protein interacts with itself using a cluster 3 interaction. Equally, we observe that YeastA does not interact with itself or YeastB using this cluster. This information is shown in step 2 of the figure; known interactions (i.e. probability of interaction = 1) are coloured green and known non-interactions are coloured red.

Now, after reconstructing the ancestral sequences for all ancestral proteins in the phylogeny using PAML, the complementary fraction can be computed for every interaction node in the interaction tree (each interaction node corresponds to a pair of proteins). To do this we use the observed PPI structure from yeast as a template, as was described in the previous chapter. Now, we can calculate the change in complementary fraction on each branch of the interaction tree to give the D_{com} metric on each branch (blue numbers in figure). With these numbers calculated, we can use the previously fitted model, as described in Figure 2.16, to produce probabilities of PPI gain or loss on each of these branches, given D_{com} . For instance, given a D_{com} of 0.4, looking at the previously fitted exponential curves in Figure 2.16 gives probability of gain of interaction of around 0.1. Similarly, the probability of loss of interactions is 0.

In the last chapter it was described how, given a simpler example, Bayes' theorem can be used for inference given these probabilities. In this example, we have observed probabilities for three nodes and wish to produce probabilities of interaction for all other nodes in the tree. For this, the message passing algorithm of [94] is used. This produces probabilities of interaction for each node given the already observed probabilities and the conditional probability function on each branch (as defined by D_{com}). In this example, the final probabilities are shown

in step 3 of the figure. Pairs of proteins with probability of interaction greater than 0.5 are coloured green and other pairs coloured red. We see that the CowB protein is also predicted to interact with itself using this interaction type and also with CowA. It could be that the CowA-CowB interaction is a false positive or that there is indeed a different pattern of interaction in this species.

Having elaborated on the use of the D_{com} model to infer interaction histories, we can now go ahead and test the method on some real data. Before proceeding we first repeat the clustering of PPIs across our larger data set. The same clustering algorithm is applied as earlier but now to cluster the PPIs across all 7 structures simultaneously. Remarkably, the PPIs still fall into 6 main clusters showing the high level of PPI structure conservation between even bacteria and eukaryotes in the proteasome structures. Each PPI orientation is now treated independently with the results combined in the final analysis unless stated otherwise.

For this part of the analysis, only the set of sequences in Table 3.1 are considered, thus reducing the phylogeny to include just proteins from 7 species. The phylogeny is pruned to contain only these sequences and then ancestral sequences are reconstructed for the pruned tree, using PAML, as described in the Methods section. In order to test the predictive ability of the model, we propose a strategy in which the one multicellular eukaryote structure (from cow) is removed from the data and predicted using the other structures and the interaction tree model. This is carried out as described in the example of Figure 3.6. The rationale behind this is that the cow structure is the most complicated in terms of the number of available subunits and so a useful method would be able to predict this more complex structure from the relatively simpler structures (for example in bacteria). Several predictions are made; in order to test how the quality of the prediction depends on the relatedness of the proteasome structures used as evidence PPIs. The first prediction is made including only the distantly related bacterial species as evidence

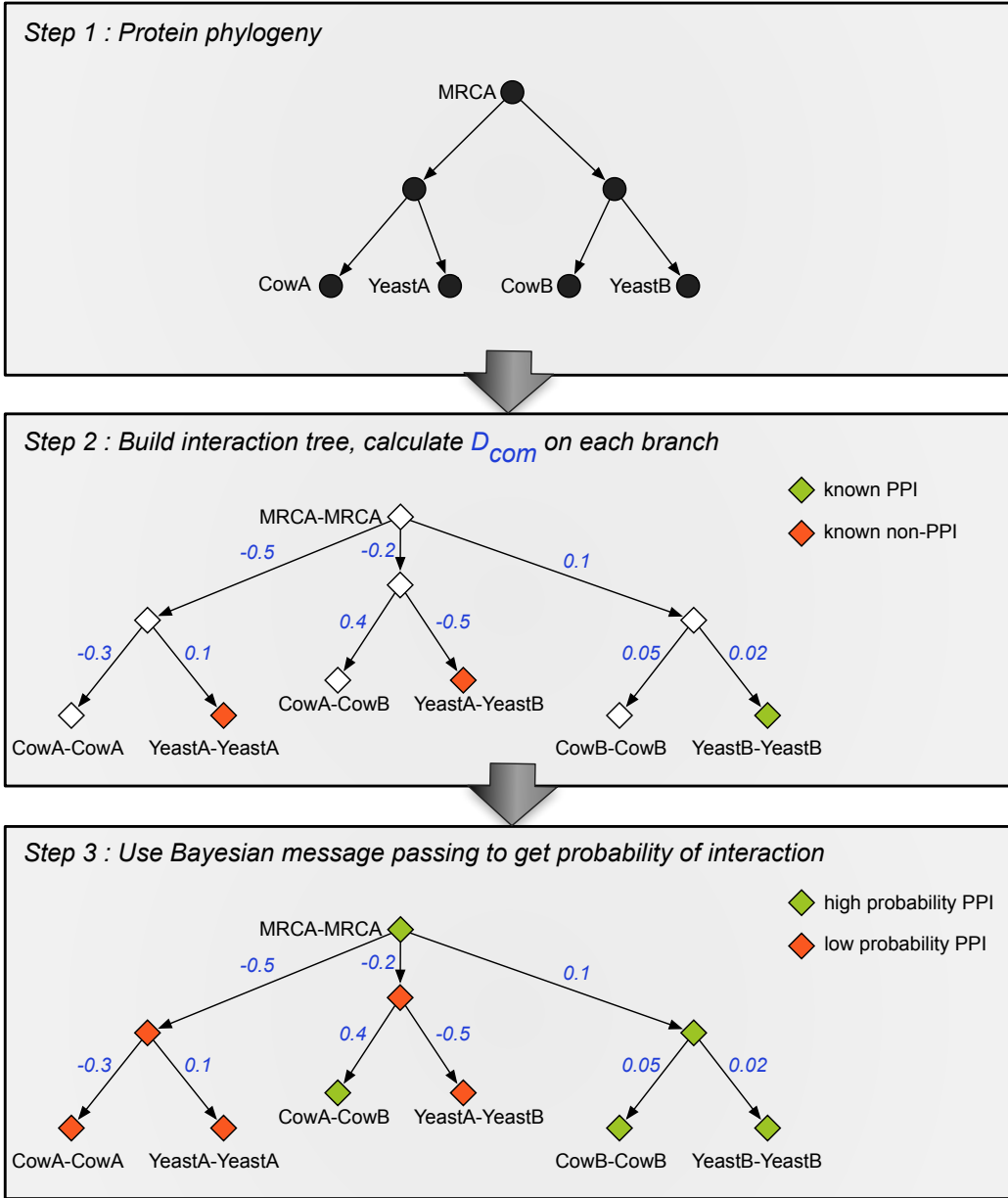


Figure 3.6: A toy example showing use of the D_{com} model, in conjunction with an interaction tree, to infer PPI histories.

PPIs, the second prediction uses bacterial and archaeal species as evidence and the final prediction uses all data as input evidence.

ROC curves were produced for each prediction (Figure 3.7) and comparisons were made with two other methods for predicting PPIs using evolutionary arguments: a simple method based on interologs and a prediction based on the PARANA algorithm for inferring PPI evolution. Predictions were also made using just the complementary fraction from the D_{com} model, with no interaction tree inference. To do this, the complementary fraction was calculated for all pairs of proteasome subunits and a threshold applied to these scores to predict which pairs interact. One popular method for predicting PPIs using evolutionary information is the MIRRORTREE method [39] [112], however, this algorithm predicts interaction between protein families. In this case we are predicting the individual PPIs at the protein level and so the MIRRORTREE algorithm is unsuitable.

The prediction based just on scores with no interaction tree performs better than random but is the worst performing prediction in terms of the area under the curve statistic (AUC). Using the interaction tree method we can see that including just the very distant bacterial structures improves prediction over using the raw scores. This is improved further when including the archaeal data (the archaeal structures have a similar topology to the eukaryotic structures) and becomes a perfect prediction when including the closest related structure from yeast and predicting using all available structural data. These results clearly show the benefit of using the phylogeny to explicitly model the PPI evolutionary process as opposed to using just the scores.

A prediction is also made based on the PARANA algorithm [84]; this algorithm takes present day PPIs as input and tries to infer the most parsimonious history of rewiring events that can produce these PPIs, given the phylogeny of the proteins. As packaged, this algorithm only supports predictions in ancestral species and

does not support making predictions in other existing species as here. So, to make a prediction in *B. taurus* using PARANA the following approach is taken: the history of interactions are first inferred for the input evidence PPIs, then, the nearest ancestor for which predictions were made is taken and using the principle of parsimony (i.e. no rewiring events), all descendant pairs of interacting pairs in the ancestor species are predicted to interact. This prediction adheres to producing the most parsimonious history and can produce predictions in unobserved species.

The PARANA prediction shown in Figure 3.7 was made using default parameters, using all structural PPIs as input. In this prediction all PPIs are gained at the leaves of the trees as this leads to the least number of rewiring events in the history of the complex and this produces a prediction on no PPIs in the *B. taurus* complex. This is certainly not the application that this algorithm was designed for; the algorithm was designed to reconstruct large protein interaction network histories, but this result highlights the unsuitability of parsimony arguments in this example. This is a result of the large number of PPIs lost during the history of the proteasome: the complex began as homomeric, in which all subunits interact, and in present day eukaryotes has 14 subunits which do not all interact with each other, in fact of the 14^2 possible interactions only 70 exist in eukaryotes. Parsimony is ill equipped to predict this large number of PPI losses.

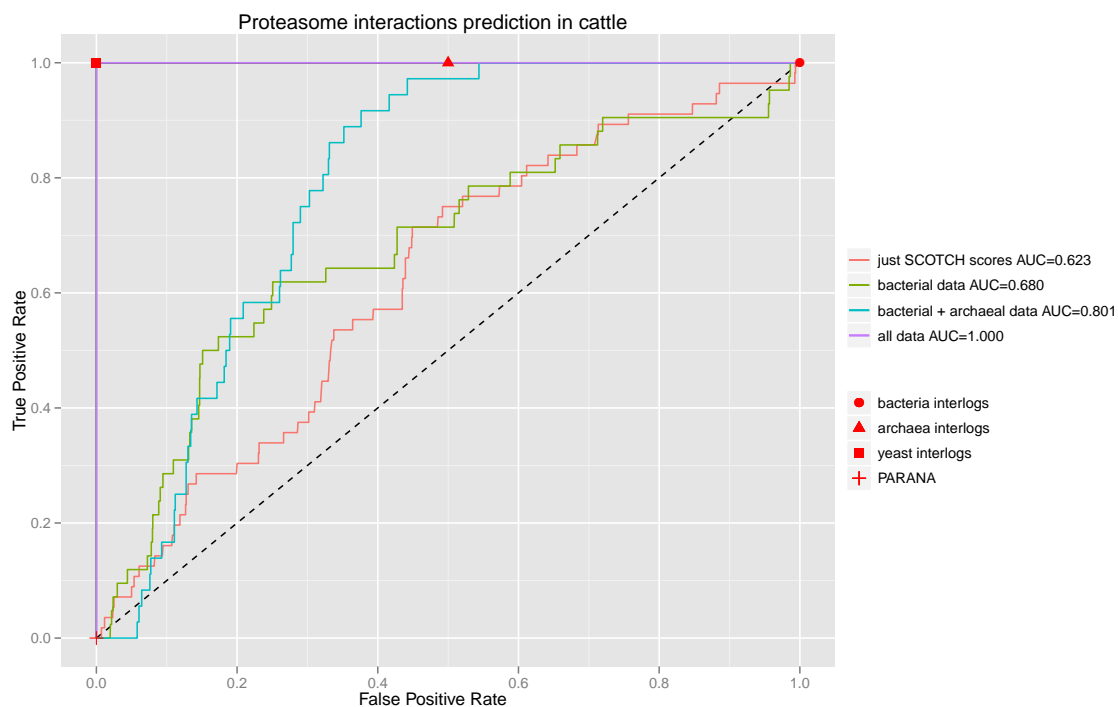
Predictions are also made using an interolog approach [57], in which each pair of *B. taurus* proteins are predicted as interacting if their nearest bacterial proteins (from one of the bacterial structures) interact. Similar predictions are made using the archaeal and eukaryotic structures. The eukaryotic interologs produce a perfect prediction similar to the interaction tree, this suggests that there has been no rewiring between the *B. taurus* and the nearby eukaryotic input structures and explains why the interaction tree produces a perfect prediction in this case. The archaeal and bacterial interolog predictions tend to overpredict

interactions, producing a fairly high false positive rate.

3.3.3 Reconstructing the history of protein complexes

Aside from the predictions reported above, the interaction tree also allows predictions of ancestral PPIs. To demonstrate this, an expanded analysis was performed including the extra proteasome sequences described in Table 2.4. A phylogenetic tree was constructed for all proteins and then the interaction tree was used, along with the model of rewiring events, to predict all rewiring events between the proteins described by the phylogeny. This history can then be used to produce predictions for the PPIs present in the proteasome of each species, including the ancestral species, Figure 3.8. This prediction broadly agrees with what we expect to see; the method predicts high rates of loss of interaction at the base of the eukaryote clade, this coincides with the gene duplication events that produce the variation in the eukaryote complexes. If we assume that these duplications did not change the topology of the complex then interactions between duplicates will have been lost to maintain the correct number of subunits in the complex and the method predicts this process. It is worth noting that this is in contrast to the parsimonious predictions from PARANA wherein all interactions are gained near the tips of the tree. This highlights the unsuitability of parsimony methods in reconstructing the history of complexes formed of duplicates in this way.

The predictions shown in Figure 3.8 are for all PPIs across the 6 interaction types classified by the clustering algorithm. It is also possible to look at the history of PPIs for each interaction type separately, in particular this can predict what binding site was the first to exist within the proteasome by looking at the interactions predicted at the root of the phylogeny. Based on this prediction it can be inferred if the original complex was a dimer or a ring structure. The history



]

Figure 3.7: Validation of method in predicting *B. taurus* structure. On the right is shown the 6 scenarios for which we test the method. Each scenario is illustrated by a species tree showing the relationship of the 7 species for which we have structural data, the *B. taurus* structure is show on the leftmost branch on each tree. In scenario 1 we leave out all structural data except the most distantly related species (*E. coli*) and try to predict the interactions in the *B. taurus* structure. In scenario 2 we include the next furthest structure and repeat. At each stage we plot the effectiveness of our predictions using a ROC curve as shown on the left.

inferred here predicts 5 of the 6 possible interaction types to be present at the root of the tree (Figure 3.8) and so is inconclusive as to what binding interface was the original state of the proteasome (Table 3.4). The results do however clearly predict that the interface forming the ring of alpha subunits in the proteasome was the last to emerge.

These predictions can be compared to existing hypotheses of proteasome evolution. The first attempt at a comprehensive description of proteasome evolution was due to [74], in which, after a survey of proteasome-like proteins across a diverse set of genomes, it was concluded that HslV was the ancestral complex with the 20S proteasome (from archaea and eukaryotes) derived from this simpler complex. The exception to this hypothesis is the actinobacteria which contain 20S proteasomes instead of the HslV (e.g. *Mycobacterium tuberculosis*). It was proposed in [74] that this bacterial proteasome was the result of horizontal transfer, however the source of the transfer could not be found. The results of the inference here disagree somewhat with this hypothesis; the alpha-beta interactions are predicted to be present in the Last Universal Common Ancestor (LUCA). However under the hypothesis of [74], the LUCA is predicted to form an HslV-type complex, in which these interaction types are not present (e.g. as in *E coli*).

A second hypothesis for proteasome evolution was recently proposed in [113]. Here, the authors surveyed a larger number of genomes than previously and looked at the distribution of proteasome-like subunits across these genomes. In this analysis, the authors demonstrate the existence of another sub-family of proteasome-like proteins, distinct to the alpha, beta and HslV proteins. This sub-family is named *Anbu* and after analysing its distribution across the tree of life, the authors conclude that this protein is the LUCA of the proteasome subunits superfamily. Under this hypothesis, the predictions of the interaction tree analysis at the root refer to the Anbu ancestor protein. In [113], homology modelling of Anbu proteins is used

in an attempt to deduce the quaternary structure of the complexes they form. It is argued that Anbu forms a barrel structure formed of heptameric rings, as in the 20S proteasome, as opposed to hexameric rings such as those found in HslV. The authors then propose a strategy for determining whether the Anbu protein forms a barrel of 4 stacked rings (as in the 20S proteasome) or 2 stacked rings (as in HslV). However, it is not possible to determine which is the complex structure of Anbu.

The predictions presented here are consistent with an ancestral Anbu complex formed of 4 stacked rings. Such a complex would be formed of 28 subunits, arranged in to four rings of seven, with binding sites between the subunits corresponding to those found in the 20S proteasome. The only discrepancy here is the prediction of no Alpha ring interface. This could be explained if the Alpha ring interface was not present in the ancestral complex; alpha subunits could maintain a ring like structure through their association to the beta ring via interaction clusters 4 + 5. The alpha ring interface could then have evolved later in the 20S proteasome in order to strengthen/refine the structure of the complex (this is exactly what is predicted using the interaction tree). This hypothesis is supported in structures of bacterial 20S proteasomes, in which the contact area of the alpha ring interface is much smaller (e.g. *Rhodococcus* proteasome [114]). It is possible that that these 20S proteasomes have an alpha subunit interaction pattern similar to the ancestral Anbu complex.

Taken together this analysis supports the hypothesis of [113] and provides further evidence that the LUCA proteasome-like complex was formed of the Anbu protein, arranged in 4 stacked rings of 7. This kind of detailed, structural prediction is possible because of two advantages of the method presented here: firstly, clustering of the interactions in the complex allows independent analysis of their evolution. Each of these clusters plays a distinct topological role within the com-

Cluster ID	Description	Probability of interaction at root
1	Alpha ring interface	0.065
2	Alpha-beta interface 1	0.704
3	Beta ring interface	0.998
4	Beta-beta interface 1	0.998
5	Beta-beta interface 2	0.999
6	Alpha-beta interface 2	0.948

Table 3.4: Probabilities of interaction at the root of the protein phylogeny.

plex and so analysis of their evolution individually allows insight in to the evolution of the quaternary structure of the complex. Secondly, the method takes in to account the distinct sequences of the proteins whilst making predictions. To give a good example, the evolution of both the alpha and beta subunits in eukaryotes is explained by 6 consecutive duplications to produce 7 paralagous subunits. A method such as [87], postulates a uniform probability of PPI rewiring after each of these duplications and so the inferred pattern of PPI rewiring would be identical for both the alpha and beta rings in this case. The method presented here bases the probability of rewiring on changes in the sequence, this allows detection of the differing rewiring behaviours of the alpha vs. beta subunits.

Two archaeal species with an unusual number of proteasome subunits were included in this analysis: *H volcanii* has two alpha subunits compared to most archaea which have one, *P furiosus* has 2 beta subunits. These species were included to test the interaction tree’s ability to deduce the role of the extra subunits in the assembly of the complex. In *H volcanii*, the predictions of interaction are very similar (using a probability cutoff of 0.5) for the 2 possible alpha subunits: both are predicted to bind to the beta subunit and both are predicted to interact with

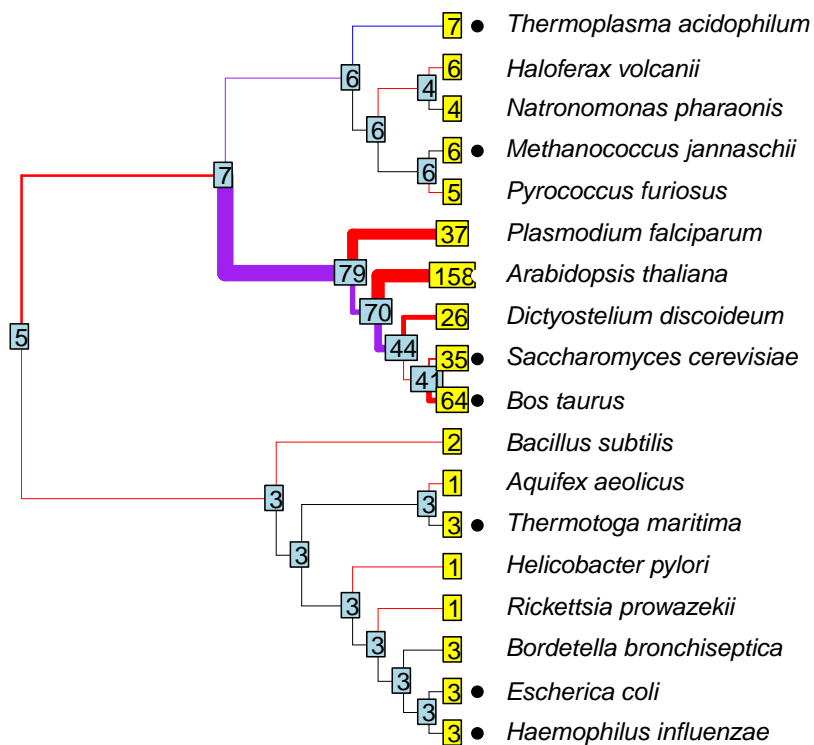


Figure 3.8: Reconstructed history of interactions in the proteasome. This figure shows the species tree relating all genes in the analysis with a number at each node representing the number of predicted interactions across all 6 interface clusters. The predictions are obtained by placing a 0.5 probability threshold on the results and marking every interaction with a probability higher than this as a positive prediction. Species for which we have interaction data in the form of the structures listed in Table 3.1 are marked with a black circle. Edges on the tree are marked red if interactions are lost along this branch, blue if they are gained and purple if interactions are both lost and gained along the branch with the edge width representing the number of such events.

themselves and each other in forming the ring of alpha subunits. This prediction is consistent with experimental results showing that both alpha subunits are capable of forming rings within the 20s proteasome [115].

In *P. furiosus* both beta subunits are predicted to bind to the alpha subunits however the predictions for binding between the beta subunits, in forming the catalytic ring, suggests differing roles for the two beta subunits, PF1404 ($\beta 1$) and PF0159 ($\beta 2$). The probabilities for each pairing of beta subunits interacting using the beta ring interface are shown in Table 3.5. Using a cutoff of 0.5 the method predicts that $\beta 2$ can interact with itself to form rings but $\beta 1$ cannot, $\beta 1$ instead interacting with $\beta 2$ to form mixed rings. This is consistent with experimental results reporting that only $\beta 2$ forms homomeric rings, with $\beta 1$ being incorporated into the structure at higher temperatures [116].

The ability of the interaction tree methodology to distinguish between the interaction patterns of these recent duplications sets it apart from competing methods. Firstly, the interolog method described in the previous section would return identical predictions for the pair of beta subunits in *P. furiosus*. Secondly, competing interaction tree-type methods, e.g. [87] place a constant value on the probability of rewiring in either duplicate after duplication. This means that the inferred interaction probabilities for the duplicates are derived from the interaction probabilities of their ancestor identically, the result being identical predictions for the interactions of the duplicates. However, the method proposed here takes into account the individual sequences of the duplicates, specifically the changes in the sequences since the ancestral protein. This allows the subtle differences between the duplicates to be recognised and used to predict the difference in interaction patterns.

Subunit 1	Subunit 2	Probability of interaction
$\beta 1$	$\beta 1$	0.278
$\beta 1$	$\beta 2$	0.530
$\beta 2$	$\beta 1$	0.301
$\beta 2$	$\beta 2$	0.530

Table 3.5: Probabilities of interaction using the beta ring interface (cluster 3) in *P. furiosus*. Note that the structure of the interaction is asymmetric and so both possible orientations are reported.

3.4 Discussion

The first important finding in this chapter was presented in Figure 3.7. Here it was shown that the D_{com} score alone provides fairly weak ability of predicting PPIs in the cattle 20S proteasome. However, inclusion of PPI evidence from even distantly related bacterial species, via the interaction tree, improves these predictions. This effect is magnified as evidence from closer species is incorporated. This result highlights the relevant information present in protein phylogenies; PPI patterns are often conserved by orthology relationships (and this is certainly true in the proteasome). The modelling of the rewiring events on top of these phylogenetic relationships allows boosting of a predictor that would not be very useful to a predictor that is very useful, with comparable accuracy to other methods. The success seen here echoes previous arguments that attempting to model the true evolutionary process improves predictions of protein structure and function [117].

There is one factor central to the assembly (and evolution) of the proteasome that is ignored by this model; that is the working of chaperones. Chaperones are proteins that help with the assembly of multi subunit complexes such as the proteasome but are not found in the final complex. In many cases, a chaperone's

main function is to prevent proteins from forming a non-functional aggregate (e.g. [118]) but in other cases (e.g. [119]) the chaperone can be vital for the correct assembly of the complex. That is, the complex will not assemble correctly in the absence of the chaperone. This is true in complexes with a ring structure, for which the chaperone can be important in determining the order of the proteins in the ring. As such, it is not their sequence alone that determines the binding patterns in these complexes but also the effect of the chaperones. This may make prediction in these systems difficult.

In the final part of this chapter, some specific predictions are made about ancestral topology of the proteasome. These predictions are made at the level of the binding sites in the proteins and allow predictions such as what ring structures were present in the ancestor. These predictions are easy to make from the results in the ancestor as there is only one subunit at the root of the phylogeny. Predicting the topology/quaternary structure in species containing several subunits is less trivial. Therefore, an extension to this work would be to explore the feasibility of an algorithm to infer quaternary structure from a list of pairwise scores between proteins, for each binding site. A starter test for such an algorithm would be to predict the topology of the cattle proteasome from the predictions presented in this chapter.

3.5 Conclusion

This chapter has demonstrated the utility of the interaction tree approach in inferring the evolution of a specific class of protein complexes. These complexes are formed of paralogous subunits with obligate interactions between the subunits. An obvious question now is whether the methodology can be extended to other classes of PPI. As opposed to obligate, permanent PPIs, the other main class of interac-

tion is the transient interaction. These interactions are temporary and prevalent in functions that require dynamic, temporal PPIs e.g. in signalling networks. In the next chapter, the method is extended to transient interactions. These interactions are fundamentally different in terms of the energetics of their binding and so the first question to answer is whether the D_{com} measure described here is predictive for this class of PPI.

Chapter 4

A method for predicting transient PPIs

4.1 Introduction

Transient PPIs are an important class of interactions found within living organisms [8], distinct from the obligate PPIs that were the subject of previous chapters. The term transient is applied to interactions that are temporary, the component proteins often being found separately in the cell and only coming together to interact under some condition. This kind of PPI is utilised by the cell in situations requiring response to some condition. For instance, sensory proteins embedded in the cell membrane pass information of the external environment to the interior of the cell via interactions with proteins in the cytoplasm. These interactions are dependent on the presence of some external signalling molecule and so the interaction will only happen under certain conditions. After the interaction has occurred, the internal protein then disassociates from the sensory protein and then goes on to effect some outcome. This class of interactions has been observed to

be fundamentally different to obligate interactions in terms of the residues used to mediate the interaction [120], [121]. Therefore, it is important to test the robustness of the PPI rewiring model to these changes.

This chapter focuses on extending the interaction tree methodology described previously to transient PPIs. The previous D_{com} model of PPI rewiring, based on the SCOTCH score [98], is taken as a starting point. This model worked in the previous chapters because the SCOTCH score could successfully distinguish interacting pairs from non-interacting pairs of proteins. As a result, tracking the change of this score along a branch of the interaction tree allowed prediction of rewiring events on that branch. For the model to work successfully for transient interactions, the SCOTCH score used as its starting point needs to successfully distinguish interacting from non-interacting pairs in transient systems. To test the suitability of the D_{com} model, the SCOTCH score is compared against several competing metrics in order to find which is best for a model of PPI rewiring in this case.

Several possible metrics exist for predicting PPIs. The only restriction for use in the interaction tree is that the method be automatable and fast: when computing the metric for ancestral proteins the sequence of the proteins is uncertain, to deal with this uncertainty, a large number of sequences are sampled from a distribution of sequences and the metric is computed in each case. This allows computation of the mean of the metric across the samples. If the scoring system used is not fast and automatable then this approach will not be feasible. Given this restriction, 3 competing approaches are considered. The first uses residue level statistical pair potentials, the second uses basic homology modelling and the last uses a purely knowledge based approach (see Methods for details).

As a test case for these approaches, bacterial two component systems [122] are used. Two component systems are formed of a histidine kinase (HK) embedded

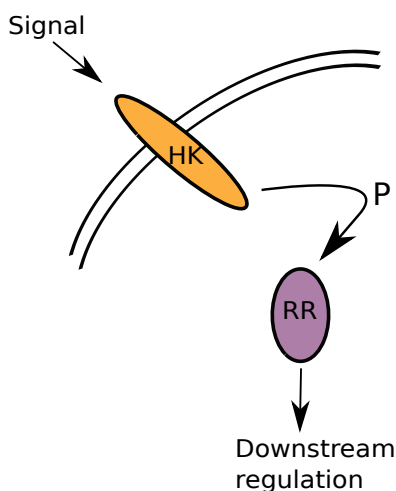


Figure 4.1: A cartoon representation of a two component system PPI. The Histidine Kinase (HK) is embedded in the cell membrane. On receiving some external signal, this protein auto-phosphorylates and subsequently phosphorylates a Response Regulator (RR) via a PPI. This RR then becomes activated and goes on to produce some downstream effect.

in the membrane and a response regulator (RR) in the cytoplasm (Figure 4.1). In these systems, the HK receives some external signal and autophosphorylates as a result. The transient interaction between the HK/RR allows passing of the phosphoryl group from the HK to the RR, activating the RR. The RR is a transcription factor and on receiving the phosphoryl group is activated and able to interact with the DNA, effecting some downstream response. This simple system is extremely common in bacteria and represents the typical method by which bacteria sense and respond to their environment. Understanding the evolution of these systems is therefore key to understanding the evolution of bacterial sensing in response to a changing environment.

There is an important property of these systems that makes them a good test case for PPI prediction. In a large proportion of cases a HK is found to be

neighbouring an RR. These pairs of co-located genes are known as *cognate* pairs (illustrated in Figure 4.2) and the majority of them produce proteins that interact. This means that a large number of examples of two component system PPIs can be inferred from genome sequences alone, simply by finding neighbouring HK/RR genes. Indeed, resources exist that have already performed this search [123]. In addition, it has been observed [124] that the majority of cognate pairs only interact with each other. This lack of crosstalk allows an equally simple strategy for finding HK/RR pairs that are unlikely to interact; choose a HK and an RR from different cognate pairs.

The fact that positive and negative PPI examples can be generated easily for this system means that large training sets of PPIs can be defined, with some confidence, and used as a basis for fitting a model of PPI evolution. Despite the fact that many PPIs can be easily predicted using genome location, there is a subset of two component systems that cannot be predicted in this way. These HK/RR proteins are not paired with another RR/HK protein on the genome and are called *orphans* (illustrated in Figure 4.2). For instance, an orphan HK is a HK protein for which neither the downstream or upstream neighbour on the genome is an RR protein. Similarly an RR protein can be an orphan. Given the lack of syntenic information in these cases, alternative approaches to PPI prediction are required. It is proposed that, using the cognates as a training set, a predictor can be defined and used to predict the PPIs participated in by the orphan proteins of a species.

Cognate pairs and orphans

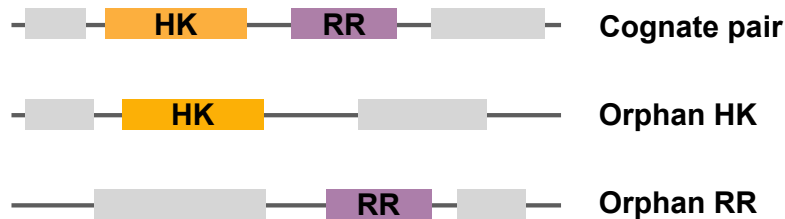


Figure 4.2: The difference between cognate and orphan two component system proteins. Cognate proteins (top) are colocated on the genome as HK/RR pairs. These pairings usually encode proteins that interact. Orphan HKs or RRs (bottom) are not found located on the genome and their interacting partners can not be inferred by genome location.

4.2 Methods

4.2.1 Training data

Two training sets of two-component systems are used for benchmarking. The first is taken from [125] and contains 1,299 pairs of two-component system proteins (a histidine kinase and a response regulator) that are co-located in a given bacterial genome (i.e. cognate pairs). In this training set the sequences of the histidine kinases are aligned as described in [125] and the response regulators are aligned similarly. The second training set is taken from [126] and contains 8,999 two-component system cognate pairs (i.e. 8,999 histidine kinases and 8,999 response regulators, arranged in cognate pairings) aligned as described in the above paper.

4.2.2 SCOTCH score

The first prediction method is adapted from the SCOTCH [98] method for scoring docked protein models. This was explained in detail in Section 2.2.2 but is ex-

plained again here briefly. To calculate the SCOTCH score we first divide the 20 amino acids into 4 groups; (GLY, ALA, VAL, LEU, ILE, MET, CYS, PHE, PRO, TRP, TYR), (SER, THR, ASN, GLN), (LYS, ARG, HIS), (ASP, GLU). These are the hydrophobic, polar, positively charged and negatively charged residues respectively. We define two amino acids to be complementary if they are both hydrophobic, both polar or one positively and one negatively charged. We start with two protein sequences and a proposed interface between them represented as a list of interacting residue pairs (i, j) where i refers to residue position i of the first sequence and j refers to residue position j of the second protein sequence. For each residue we then define the two nearest structural neighbours as the nearest two residues in three dimensional space within the same chain in the interface. This requires a proposed three dimensional structure for the interface such as can be found in a solved crystal structure of the constituent chains in complex. We then define positions i and j to be complementary if any of the pairs (i, j) , (i_1, j) , (i_2, j) , (i, j_1) , (i, j_2) describe complementary amino acids as described above, where i_1 , i_2 are the nearest two structural neighbours to i and j_1 , j_2 are the nearest two structural neighbours to j . The SCOTCH score is then given by the fraction of complementary pairs in the list of interacting residue pairs.

As mentioned above, to produce a SCOTCH score for a pair of proteins a structure is needed for the interface (or at least a set of interacting residue pairs responsible for the interaction). To do this a related interaction structure (called the evidence structure) is used and aligned to the proteins to be scored. Then, a 4.5Å distance threshold is applied to the evidence structure to produce a set of interacting residue pairs. These can be used to propose the interacting residues in the new pair of proteins by taking the residues aligned to the structure pairs in each case.

This score works under the assumption that during coevolution of maintained interactions, a residue mutation going to fixation at an interface will most likely be physiochemically complementary (e.g. hydrophobic to hydrophobic) to the residues it interacts with. However, PPI interfaces can exhibit plasticity and this complementarity can be maintained by nearby residues and does not have to be maintained through specific pairs. The SCOTCH score aims to measure this coevolution occurring between interacting proteins and could be used here to detect changes in interaction state as changes in SCOTCH score.

4.2.3 RPScore

Predictions are also made using the RPScore residue pair potential from [127]. The RPScore assigns to each pairing of amino acids a propensity for being within 8\AA across the protein-protein interface. To score an interaction the scores for all residue-residue pairs within this distance cutoff are added together to produce the RPScore. To score an unlabelled pair of proteins it is necessary to have a proposed set of residue-residue pairings; these are generated here using an evidence structure as described above.

An adapted version of the RPScore, RPSmax, was also benchmarked. RPSmax uses the same propensity matrix as the RPScore but each pair of interacting positions i and j are scored as follows: firstly define i_1, i_2, j_1, j_2 as above, then take the maximum propensity from the pairs of positions $(i, j), (i_1, j), (i_2, j), (i, j_1), (i, j_2)$. The maximum propensities are then added across the set of interacting positions before. This scoring method aims to allow for some plasticity in the structure of the PPI interface in the same way as the SCOTCH score above.

4.2.4 MODELLER and FoldX energy

The PPI prediction methods described so far do not explicitly model the 3D structure of the interaction; instead protein alignments are used to define residue-residue pairs in pairs of proteins that are equivalent to the residue-residue pairs responsible for maintaining an interaction in some related evidence structure. One obvious limitation here is that the residue-residue pairs important for the PPI may be different in the unlabelled protein pairs when compared with the evidence structure. One approach that could model these differences is homology modelling, in which a template PPI structure is used to model the 3D structure of the potential interaction between two proteins.

This approach is tested using the MODELLER homology modelling suite [128] to produce a predicted 3D structure for the interaction between two proteins. Subsequently, the FoldX [129] energy function is used to estimate the change in free energy on binding of the two proteins. This quantity is then used to discriminate between PPIs and non-PPIs: true PPIs should have a reduction of free energy on binding.

4.2.5 Likelihood Score

The final PPI prediction method uses the training sets of interactions described in Section 4.2.1 to characterise the two-component system interaction. Firstly, a 4.5\AA distance threshold is applied to a known structure of the interaction (in this case a known two component system interaction), as described above to give a list of interacting residue pairs i, j responsible for mediating the interaction. The sequences from this evidence structure are aligned to the training alignments, allowing comparison of the interacting pairs for all interactions in the training set. To begin the frequency of observing amino acids A, B at a pair of interacting

positions i, j can be calculated from the training set of m interactions as

$$F_{i,j}(A, B) = \frac{1}{20^2 + m} \left(\sum_{k=1}^m \delta_{i,j,k}(A, B) + \lambda \right) \quad (4.1)$$

where $\delta_{i,j,k}(A, B) = 1$ iff there is amino acid A at position i , B at position j in interaction k from the training set and λ is a pseudocount (equal to 1 throughout the applications in this thesis). This can be compared to the frequency of observing amino acid A in column i of the alignment

$$F_i(A) = \frac{1}{20 + m} \left(\sum_{k=1}^m \delta_{i,k}(A) + \lambda \right) \quad (4.2)$$

allowing definition of the log likelihood ratio

$$LL_{i,j}(A, B) = \log \frac{F_{i,j}(A, B)}{F_i(A)F_j(B)} \quad (4.3)$$

which describes the likelihood of observing residues A and B at positions i, j due to an interaction between the proteins as opposed to by chance in two non-interacting proteins. This score can then be added across all i, j pairs to give a likelihood of interaction for a given pair of sequences as

$$LLscore = \sum_{(i,j)} LL_{i,j}(A_i, B_j) \quad (4.4)$$

where A_i is the amino acid at position i and B_j is the amino acid at position j . It is known that amongst the interacting residue pairs i, j there are some that are more important for maintaining the PPI. Previously the importance of pairs has been described using Mutual Information (MI) calculations to identify positions having correlated positions. In order to include this information, another likelihood score is calculated using the MI to weight the contribution from each i, j residue pair, as

Species	HK proteins	RR proteins
<i>Bacillus subtilis</i>	35	33
<i>Bacillus anthracis</i>	46	46
<i>Bacillus cereus</i>	50	45
<i>Clostridium acetobutylicum</i>	35	40
<i>Clostridium difficile</i>	47	53
<i>Clostridium botulinum</i>	36	42
<i>Clostridium perfringens</i>	29	24

Table 4.1: List of species included in the analysis of this chapter and the number of histidine kinase and response regulator proteins included from each. These represent all known two component system proteins from these 7 species and the set for each species contains both cognate and orphan proteins (the difference between cognates and orphans is explained in the introduction to this chapter and Figure 4.2)

$$MILLscore = \frac{\sum_{(i,j)} MI_{i,j} * LL_{i,j}(A_i, B_j)}{\sum_{(i,j)} MI_{i,j}} \quad (4.5)$$

where $MI_{i,j}$ is the mutual information between columns i, j in the training data, calculated as described in [130].

4.2.6 Test data

Several further datasets are then used to test the scoring methods. All HK/RR proteins, as classified in MISTdb [123], were downloaded for all species listed in Table 4.1. These proteins are used as a further test set and also as a system in which to generate testable predictions. In addition, mutated sequences of a set of *E coli* two component proteins are used, as described in [131].

4.3 Results

For this analysis we use and compare two training sets of two component system interactions taken from [125] (referred to as the Laub training set) and [126] (referred to as the Weigt training set). As explained in detail in the introduction to this chapter, it has been observed that HK and RR proteins have a high specificity with a given HK often only observed to interact with one RR (and vice versa) and this pairing of genes being colocated on the genome. These pairs are termed *cognate* pairs and it has been noted that crosstalk between these cognates is rare [124]. The training sets are both made entirely of examples of these cognate pairs of HK/RR that are assumed to be interacting. Each cognate HK/RR pair is concatenated to form one long sequence and these sequences are aligned to form the training sets (the sets were each provided as aligned by their authors). These sets will act as training and testing data in benchmarking various methods for predicting interacting pairs from the sequences of the individual components. To begin these training sets are compared in terms of their size and overlap (Figure 4.3) and in terms of their redundancy (Figure 4.4).

It is clear that the two training sets differ in their size, composition and alignment. To test the effect of these differences on prediction of interaction, all benchmarked prediction methods were applied to both training sets and also to both training sets restricted to their intersection (as described in 4.4). 6 prediction methods are tested in their ability to predict two component PPIs, they are referred to as SCOTCH, RPScore, RPSmax, FoldX, LLscore and MILLscore (see Methods). The SCOTCH, LLscore and MILLscore methods require a proposed PPI structure in order to calculate. For this, the 3DGE pdb structure (a two component system interaction from *Thermatoga maritima*) is used. These methods require a list of interacting residue pairs across the protein-protein interface. The

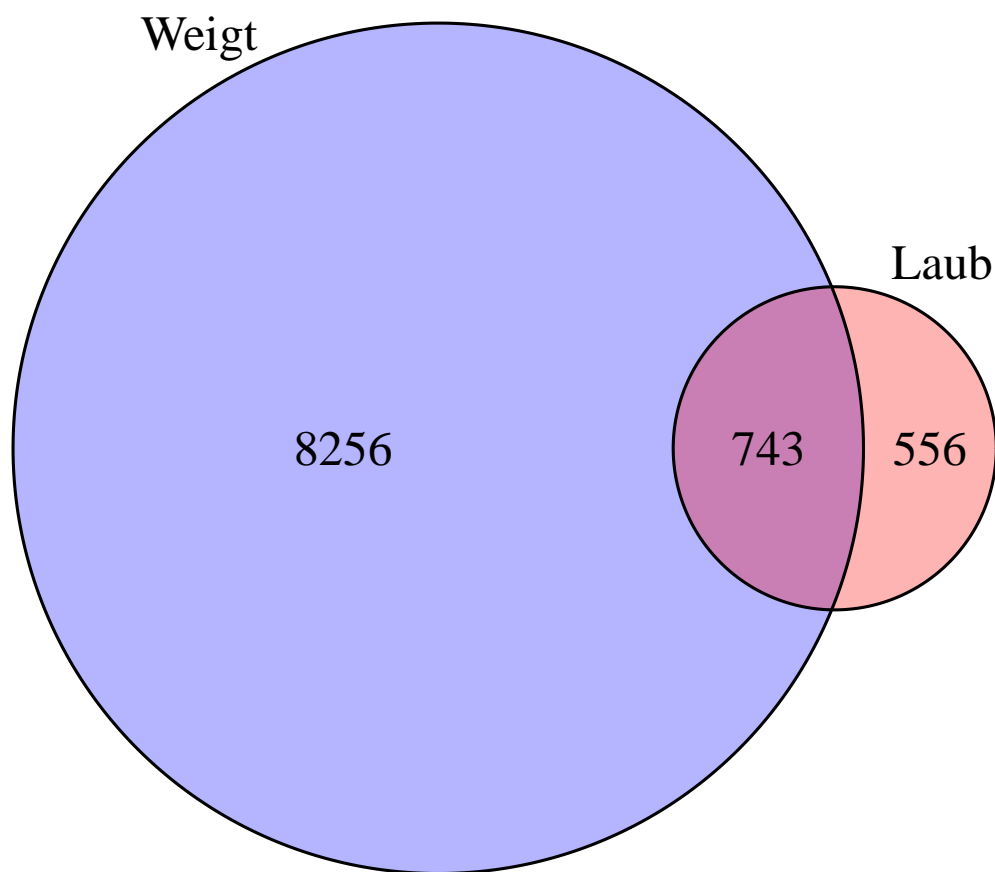


Figure 4.3: Venn diagram of the two training sets of aligned two component system interactions used in this paper. Each interaction is a pairing of a Histidine Kinase(HK) and a Response Regulator(RR) and two interactions are considered identical if the GI numbers of both the HK and RR match between the two pairs. The training set from Weigt et al is considerably larger than that from Laub et al, however there are still pairings present in the smaller set that are not in the larger set.

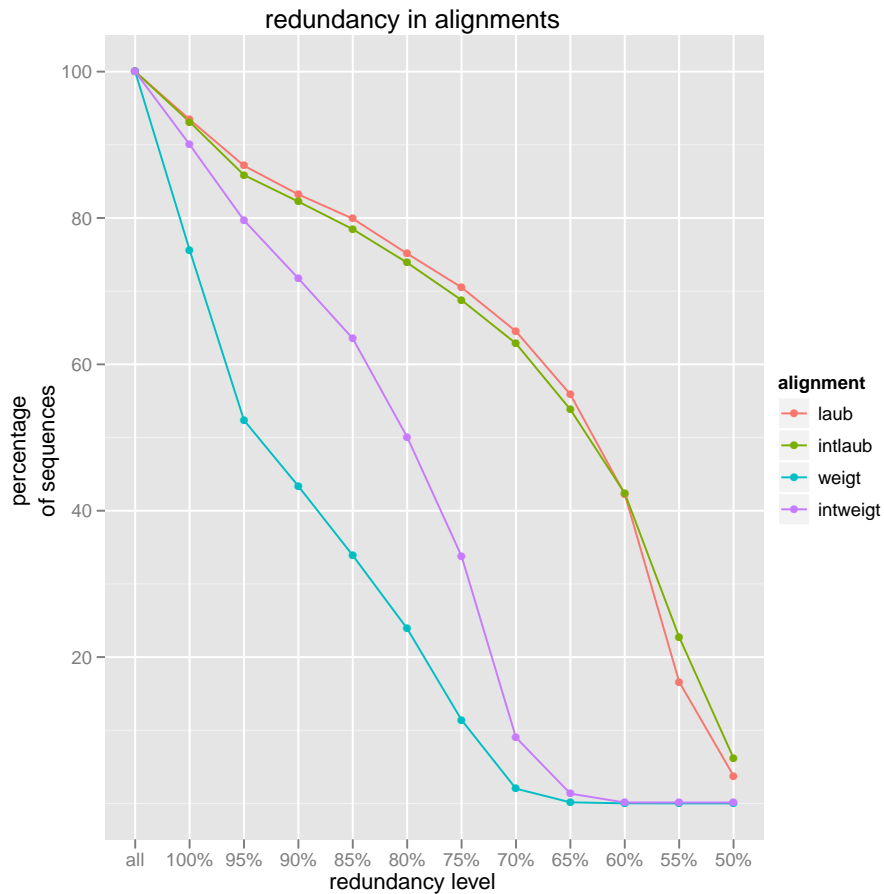


Figure 4.4: Redundancy in the training sets; the percentage of sequences left in each training set is shown after filtering at various sequence identity thresholds. Results are shown for the Laub set, the Weigt set and for each set restricted to the intersection (the 743 cases shown in Figure 4.3). Whilst the intersection contains the same set of sequence pairs, the intersection sets can still differ in how the sequences are aligned. It is clear that the Weigt training set contains a higher level of redundancy, with around 50% of the sequences having over 95% similarity to another. However, even when restricting each set to the 743 intersection cases differences are seen in the redundancy; the Weigt set still appears more redundant. As the intersection contains the same set of sequences in each case, this difference is due to the differing alignment in each case and shows that the Weigt alignment tends to align identical residues more often.

required pairs are defined, as in previous chapters, using a 4.5Å distance threshold, producing 38 interacting residue pairs from the 3DGE structure. These pairings are between 19 distinct residues in the histidine kinase and 15 residues in the response regulator. The same templating procedure as previously (Figure 2.7) is used to map sequences to the template 3DGE structure to calculate the SCOTCH score, the LLscore and the MILLscore. The LLscore and MILLscore methods also require a training set of interacting sequences (whereas the other methods require no training). Thus in order to benchmark the 6 methods on the same test data, the following approach is taken for each training set: the training set is divided in to 10 equal parts, a part is chosen in turn to be the test set, the other 9 parts are used as a training set for the LLscore and MILLscore. The test set contains only positive examples, random selection of pairings from the test set is used to generate an equal number of negative examples which are added to the test set. All 6 scores are evaluated on the test set, the results recorded, then a new test set is chosen and the process repeated until all 10 parts have been used as the test set. This allows construction of a ROC curve for each of the 6 prediction methods across each of the 4 training sets, Figure 4.5.

The SCOTCH scoring method was used in the preceding chapters as a basis for predicting rewiring in obligate protein complexes. This was successful as the score can discriminate between interacting and non-interacting proteins in these systems. However, the SCOTCH score does not perform well in predicting these transient PPIs. One hypothesis for this difference is that the SCOTCH score can detect the hydrophobic pairings that are common at obligate interfaces. In the case of transient interactions there tends to be a smaller set of specific contacts that the coarse grained SCOTCH score may not be detecting.

The FoldX score also performs badly. This score is based on the change in free energy on binding calculated from a homology model of the interaction. The

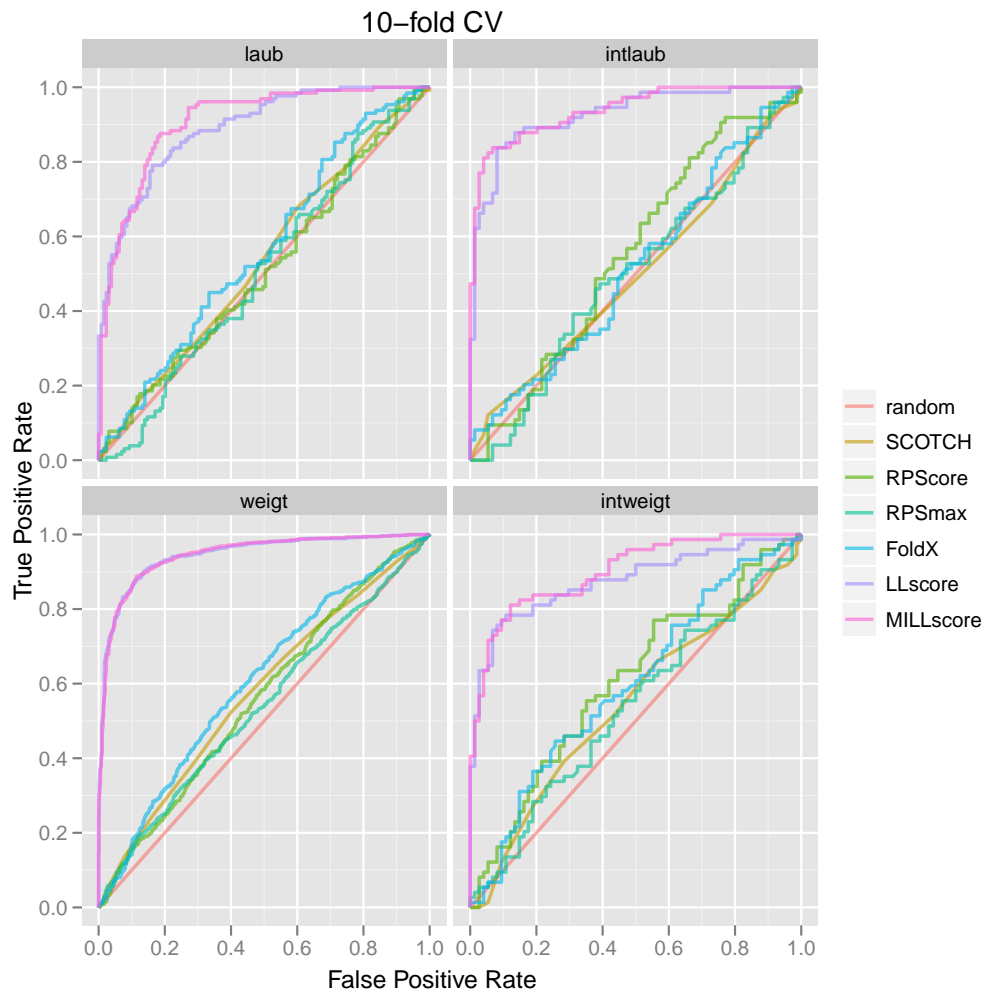


Figure 4.5: ROC curve for each of the 6 prediction measures (plus a straight line showing random prediction) across the 4 training sets. The likelihood based methods produce consistently better results with the MILLscore marginally better in some cases.

homology model was generated in an automated fashion with no human curation or verification. This could be leading to poor quality homology models that are uninformative or misleading in predicting interaction.

From these results it is clear that the two likelihood based predictions, LLscore and MILLscore give by far the most accurate predictions with the other 4 prediction methods performing slightly better than random guessing. MILLscore performs marginally better than LLscore and although this is not the case for the largest training set (weight, bottom left Figure 4.5), these two scores perform equally well here.

A similar predictive method (to MILLscore) was recently employed in [126] to successfully predict interactions between HK and RR domains, given the fact that MILLscore is consistently the best performing predictor, it remains to ask how it compares to this approach. The full test data from this paper was unavailable, and so to compare the method presented here, predictions were made, using MILLscore, for two specific examples given in [126]. Firstly, predictions were made for the orphan two component systems of *Bacillus subtilis* and secondly for the orphan two component systems of *Caulobacter crescentus*.

4.3.1 Orphan Two Component Systems in *Bacillus subtilis*

As mentioned earlier, many HK/RR proteins will be located adjacently on the genome and interact solely with each other, these pairings are referred to as cognate pairs. However, there are orphan HK and RR proteins that are not located with a potential partner. For these orphan proteins, interactions can not be inferred based on genome co-location and so it would be valuable to have some method to predict the partners of a given HK/RR. The vast majority of these orphan HK/RRs are

uncharacterised in terms of their interactions however the five orphan histidine kinases responsible for the onset of sporulation in *B. subtilis* have all been shown to interact with the Spo0F orphan response regulator. So one test of a method for predicting the interactions of orphan two component systems would be to evaluate the methods ability to predict these known PPIs. Previously in [126], predictions identified Spo0F as the highest scoring orphan partner for 4 out of 5 of these kinases.

Here, the same set of 5 orphan kinases are scored against the 4 orphan RRs from *B. subtilis* using the MILLScore method, trained on the Weigt training set. This score successfully predicts Spo0F as the partner (by assigning the interaction with this RR the highest score) of all five of the kinases (Figure 4.6). Not only does the MILLScore give improved predictions but it also obtains these in vastly reduced computational time. The original predictions from [126] took 2 days of computer time to train the scoring system whereas the MILLScore calculations take 15 minutes. This huge reduction in computation time comes as the MILLScore method uses a template structure to estimate the residue-residue contacts between two proteins whereas [126] estimates the residue-residue contacts using computationally expensive direct-coupling analysis.

4.3.2 Predicting Specificity Rewiring

It has been shown experimentally that a small number of residues at the protein-protein interface are responsible for the high specificity of two component system interactions. Indeed it has been possible for the specificity of HK-RR pairings to be changed by mutating a few important residues; [131] changed the specificity of the EnvZ histidine kinase from the EnvZ/OmpR cognate pair to that of RstA from the RstA/RstB pair, via a sequence of mutations in a small set of important residues.

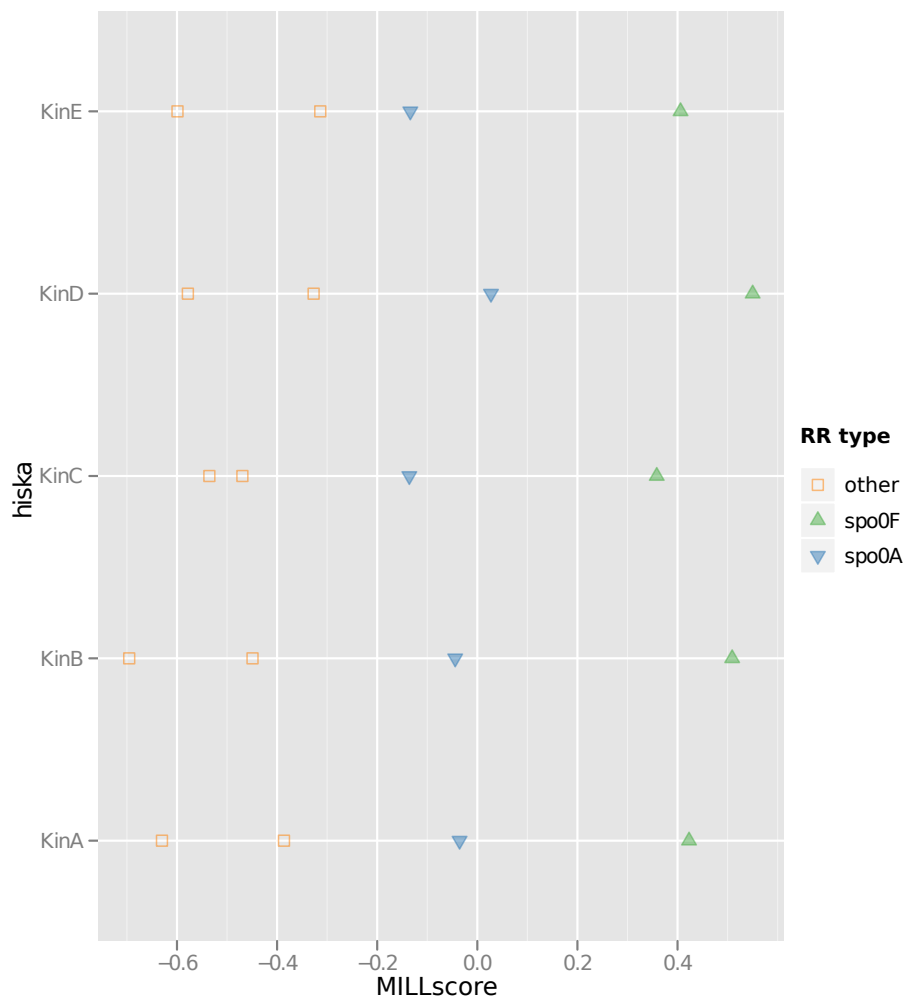


Figure 4.6: Interaction predictions for five orphan kinases in *B. subtilis*. Spo0F (green triangle) is correctly predicted as the interacting partner of each of the five kinases. In *B. subtilis* these interactions form part of a phosphorelay pathway in which Spo0F is phosphorylated by one of the kinases, then passing on the phosphate group to an intermediate Spo0B which finally passes on to Spo0A, a downstream response regulator that initiates sporulation. In many species (e.g. Clostridia) such a relay doesn't exist; the HK phosphorylates Spo0A directly and so a good scoring method will be able to distinguish between these two alternatives. The scores between each of the orphan kinases and Spo0A is shown here (blue triangles); the score successfully discriminates between the two possible modes of sporulation onset.

That is, after the sequence of mutations EnvZ was shown to phosphorylate RstB but no longer OmpR. The set of residues mutated in this study are contained within the set of residue pairings identified and included in the MILLscore procedure. To test the ability of the MILLscore to predict changes in specificity at this resolution, the MILLscore was evaluated after each mutation in the sequences, as presented in [131].

The predictions of the MILLscore are shown in Figure 4.7. Each bar chart corresponds to one kinase, named according to the original paper, with wild type kinases shown at the top and successive mutants shown below. Three residues of the kinases were mutated from their wild type T,L,A in EnvZ and V,Y,R in RstB. The mutants are named with the three amino acids in brackets if they differ from the wild type i.e. EnvZ(TLR) has an arginine at the third position in place of an alanine. Two trajectories are shown mutating the three residues of EnvZ to those of RstB and vice versa and the CpxA/CpxR cognate pair are included as a control.

The predictions made by MILLscore (Figure 4.7) agree with the core experimental findings of [131]; firstly, the cognate RR scores highest for each of the three wild type HKs (although marginally so for the CpxA/CpxR pair). Secondly, the control case CpxR maintains the lowest score throughout the mutations. Finally, the sequence of mutations switches the highest scoring RR from OmpR to RstB and vice versa, indeed the final mutant resembles the other wild type in both cases (EnvZ(VYR) vs RstB and RstB(TLA) vs EnvZ).

4.3.3 Predicting Two-component Systems in 7 Bacterial Species

Having demonstrated the applicability of the MILLscore in predicting two component system interactions both in the training set and in some specific known cases,

predictions are now made in some unknown cases. Sequences for all HK/RR proteins (as defined in MIST db [123]) were downloaded for the 7 species listed in Table 4.1. The HKs and RRs were each then aligned using 3Dcoffee [26] with the 3DGE pdb structure as a template. This crystal structure is a two component system from *Thermatoga maritima*. Each kinase can then be scored against all RRs in a species in order to produce predictions of interaction in each species.

Before beginning, the coverage of the 7 species by each of the training alignments is explored as shown in Figure 4.8. The proportion of each training set being from each of the species varies across the two training sets and in their intersection. It can be seen, however, that the Weigt set is the only set containing representative sequences across the range of 7 species and so this training set remains the set of choice in making predictions. It is important to mention that some of the sequences downloaded from MIST may be present in the training set. However, the training set only contains cognate pairs (known HK/RR partners) and so this will only bias predictions for known pairs. Predictions of the interactions of orphan HK/RR will not be biased in this way.

In particular, of interest here are the predictions of HK/RR pairs involved in sporulation. The 7 species can be divided into Bacilli and Clostridia and in each case sporulation is initiated by activation of the orphan RR Spo0A, a conserved RR that is easy to identify in each species due to its conserved nature. However, due to their unconserved nature, the HKs responsible for this activation are largely unknown. It is known that Bacillus species activate Spo0A by a phosphorelay [132] (Figure 4.9) with the sensor kinase phosphorylating the Spo0F RR which then transfers the phosphoryl group to Spo0A via the cytoplasmic phosphotransferase Spo0B. In the model species *B subtilis*, the five orphan kinases responsible for phosphorylation of Spo0F are known [133] (Figure 4.6), although their activating signal is not. In Clostridia, there is a simpler model of sporulation onset; Spo0F

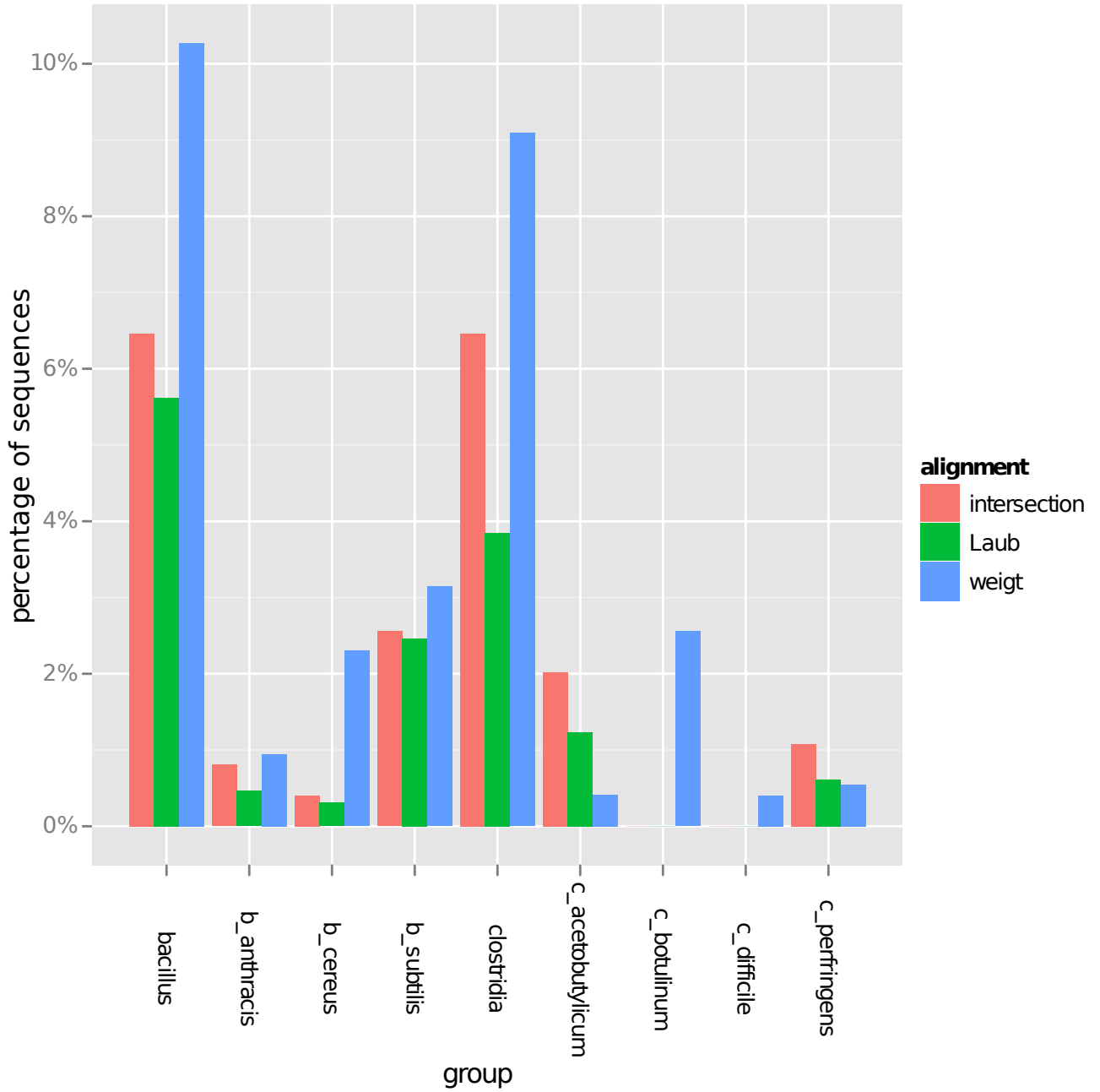


Figure 4.8: Coverage of the seven species by the training alignments. For each species (and for the *Bacilli* and *Clostridia* as a whole), the percentage of sequences from that species is shown in each training alignment and in their intersection.

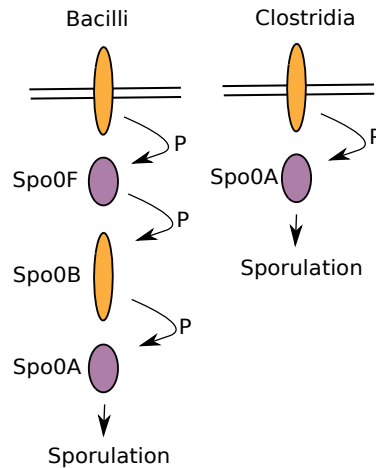


Figure 4.9: A cartoon representation of the two component systems responsible for sporulation onset in Bacteria. In the *Bacilli* a phosphorelay is used (right) and in *Clostridia* a simpler, canonical pathway of two proteins is found.

proteins are easily identifiable due to their being conserved, *Clostridia* are lacking Spo0F [134] and so the sensor kinases activate Spo0A directly (Figure 4.9). Some work has been done to identify the kinases responsible [135], although the system here is less well understood.

The identification of sporulation kinases is then a good problem to which to apply the MILLscore. This can be done by scoring kinases against Spo0F in bacilli and Spo0A in clostridia to look for high scoring pairs. As both Spo0F/Spo0A are orphans, no cognate pairs exist and so computational predictions like this are extremely useful in guiding experimentation to determine the kinases responsible.

To begin, predictions were produced for all kinases in *B subtilis* (Figure 4.12, Appendix). A MILLscore was computed for every pairing of a HK with an RR in *B subtilis*. These pairings are then split in to six types as described in Table 4.2. The MILLscore performs well here in predicting the cognate pairs with 11 kinases (of 28 non-orphans) having their highest score with their cognate RR. More than 50% (17

out of 28) have their cognate RR in their top 3 scores. To quantify the performance of the prediction of PPIs between the cognate pairs in this species, standardised z-scores are computed. A z-score is computed for the MILLscore of each HK with its cognate RR, compared with the distribution of scores of the HK with all RRs in the species. This measures the degree of separation between the MILLscore of cognate pairs (which are known to interact) and an average MILLscore of pairs of proteins thought not to interact. Such a z-score is not computed for the orphan proteins in the set, as these, by definition, do not have cognate partners.

The average z-score for the cognate MILLscores across the *B subtilis* HKs is 1.56. Assuming that the MILLscore is normally distributed (Figure 4.10 shows that this assumption is reasonable), this means that the MILLscore between an HK and its cognate RR is, on average, higher than the score with 94% of the other RRs in the species. As it is known that the vast majority of cognate pairs interact, this shows that the MILLscore is strongly predictive of PPIs in *B subtilis*.

The orphan predictions also perform well here, with Spo0F being the top scoring RR for 3 out of 5 of the known sporulation kinases and the second best scoring for the 2 other known sporulation kinases. These are the same scores as seen in Figure 4.6 except here scores against a larger set of RRs (i.e. including non-orphan RRs) are shown. It is encouraging to see that the sporulation kinase predictions are still accurate despite the much larger set of scores being compared; the MILLscore is performing well for this understood model species.

Next, the same analysis is presented for *Bacillus anthracis* (Figure 4.13, Appendix). Here the cognate predictions are accurate with 15 of 34 kinases having their cognate RR in their top 3 scores. These predictions are worse than those in *B subtilis*, as measured by this metric. However, *B anthracis* has many more RRs to score against than *B subtilis* and so achieving a top 3 score is harder. If the threshold is lowered to top 5 predictions, 24 of the 34 non-orphan kinases have

Type	Description
Cognate	A pairing of an HK and its cognate RR
Spo0A	A pairing of any HK and Spo0A
Spo0F	A pairing of any HK and Spo0F
RR orphan	A pairing of a non-orphan HK and an orphan RR (not Spo0A/Spo0F)
Double orphan	A pairing of an orphan HK and an orphan RR (not Spo0A/Spo0F)
Other	A pairing of a non-orphan HK and a non-orphan RR (not cognates)

Table 4.2: List of six types of HK/RR pairings as shown in figure legends

their cognate RR as a top 5 scoring partner. By computing standardised z-scores as before, the strength of the cognate PPI predictions can be compared directly to those in *B subtilis*. The average z-score of the cognate predictions, calculated as before, is found to be 1.57. This is comparable to the score of 1.56 in *B subtilis* and shows that the MILLscore has similar performance in both species.

Less is known about which kinases are responsible for sporulation in *B anthracis* and so predictions here are useful to explore the possibilities and corroborate existing experimental results. Previously, [136] proposed a set of 9 candidate sporulation kinases, based on their homology to the known *B subtilis* kinases (Table 4.3). Of these 9 candidates, 8 have Spo0F in their top 3 scoring RRs, based on the MILLscore. This shows striking agreement since these 8 kinases were the only orphans with Spo0F as a top 3 prediction; the two sets of predictions show almost perfect overlap. [136] went on to test the ability of some of these kinases to initiate sporulation by measuring sporulation when inserted in to sporulation kinases deficient *B subtilis*. Of the 7 candidates tested, 4 were able to produce sporulation in the mutant strain. All 4 of these kinases have Spo0F in their top 3 MILLscores. 3 of the tested candidates were unable to recapitulate sporulation, however this

<i>B anthracis</i> HK	Proposed as candidate	Spo0F in top 3 MILLscore	Experimental evidence
BA_5029	Y	Y	Y
BA_4223	Y	Y	Y
BA_3702	Y	Y	N
BA_2644	Y	Y	-
BA_2636	Y	Y	N
BA_2291	Y	Y	Y
BA_1478	Y	N	-
BA_1356	Y	Y	Y
BA_1351	Y	Y	N

Table 4.3: Comparing existing evidence for PPIs between orphan kinases in *B anthracis* and Spo0F. Each row describes a candidate orphan histidine kinase, proposed as likely to interact with the Spo0F RR in [136]. It is shown which of these kinases has Spo0F in its top 3 scoring RRs and which of these showed experimental evidence of Spo0F interaction in [136]. 4 out of 7 of the kinases with Spo0F in their top 3 MILLscores have been experimentally verified to have some interaction with Spo0F.

is not conclusive evidence that they are not sporulation kinases. The candidates were tested only in their ability to produce sporulation in *B subtilis* i.e. to interact with the *B subtilis* Spo0F. Based on the limited experimental data, the MILLscore has produced accurate predictions of sporulation kinases in *B anthracis*.

Similar results were obtained for a third *Bacillus* species, *Bacillus cereus*. The full set of predictions are shown in the Appendix to this chapter, Figure 4.14. The average z-score of the cognate predictions was again calculated and found to be 1.46 in this species. Whilst the cognate PPI predictions are slightly worse in this

species, this score means that, on average, the MILLscore between an HK and its cognate RR is still higher than the score between the HK and 92% of the other RRs in the species.

Having shown the ability of the MILLscore to predict the interactions and thereby function of orphan kinases in Bacilli species, similar analysis is now undertaken for the four Clostridia species considered in this study. In this case, the sporulation pathway is simplified (Figure 4.9) with the sensory kinases activating Spo0A directly. Thus, to find the sporulation kinases for these species, those interacting with Spo0A need to be identified (as opposed to Spo0F as earlier in the Bacilli).

To begin, results are shown for *Clostridium acetobutylicum* (Figure 4.15). Here the cognate predictions display similar accuracy as in the earlier Bacilli predictions; 16 out of 27 non-orphan kinases have their cognate RR in their top 3 MILLscores and the average z-score for the cognate predictions is 1.45. For sporulation predictions, only one orphan kinase has Spo0A in its top 3 scores, CA_C0903. Less is known about the exact set of kinases responsible for sporulation onset in this species. In [135], the authors attempted to determine this experimentally, by assessing the effect of knockdowns on sporulation and then by measuring the phosphotransfer of various kinases to Spo0A. It was found that knockdown of CA_C0323, CA_C0903 and CA_C3319 all reduced sporulation activity. Follow up experiments demonstrated the phosphotransfer from CA_C0903 and CA_C3319 to Spo0A, with results unavailable for CA_C0323. This is partially encouraging as the CA_C0903 prediction has been experimentally verified, however, the other two probable sporulation kinases have not been predicted by the MILLscore.

The other 3 Clostridia species show slightly decreased accuracy in predicting cognate pairs; the average z-score for cognate predictions is 1.40 in *C botulinum*, 1.36 in *C difficile* and 1.36 in *C perfringens*. Considering the predictions for

the orphan proteins, these other 3 Clostridia do not generally show any strong predictions for interaction with Spo0A. The exception is perhaps *C difficile*, in which Spo0A is in the top 3 MILLscores for 4 orphan kinases (CD0576, CD1492, CD1579, CD2492). So it appears that the MILLscore is less effective in uncovering the sporulation pathway in Clostridia as compared to Bacilli. This is illustrated in Figure 4.10. This figure shows the distribution of scores across the six types of HK/RR pairs described in Table 4.2, across Bacilli and Clostridia. The distributions have the same colour coding as the scatter plots in this chapter (e.g. Figure 4.12). Here it can be seen that in the Bacilli, there are some high scores (> 0.5) for cognate pairs and high scores for some kinases against Spo0F but no high scores against Spo0A. This is as expected given the structure of the sporulation pathway in Bacilli.

In contrast, in the Clostridia species, high scores are only seen within the cognate pair group. The distribution of Spo0A scores does not extend past the 0.5 mark, i.e. there are no strong predictions of Spo0A interacting kinases in Clostridia, as we would hope to find. There are two immediate possible explanations for this; firstly, there could be something different about the Spo0A proteins that make them hard to predict. This could be the result several factors including a different structure of the interaction with Spo0A, a bias of the training set against detecting Spo0A interactions or poorer alignment of the Spo0A proteins. Secondly, there could be something hard about predicting interactions of the orphan kinases in Clostridia. For instance, it might be the case that these proteins tend to be poorly aligned and receive inaccurate scores as a result.

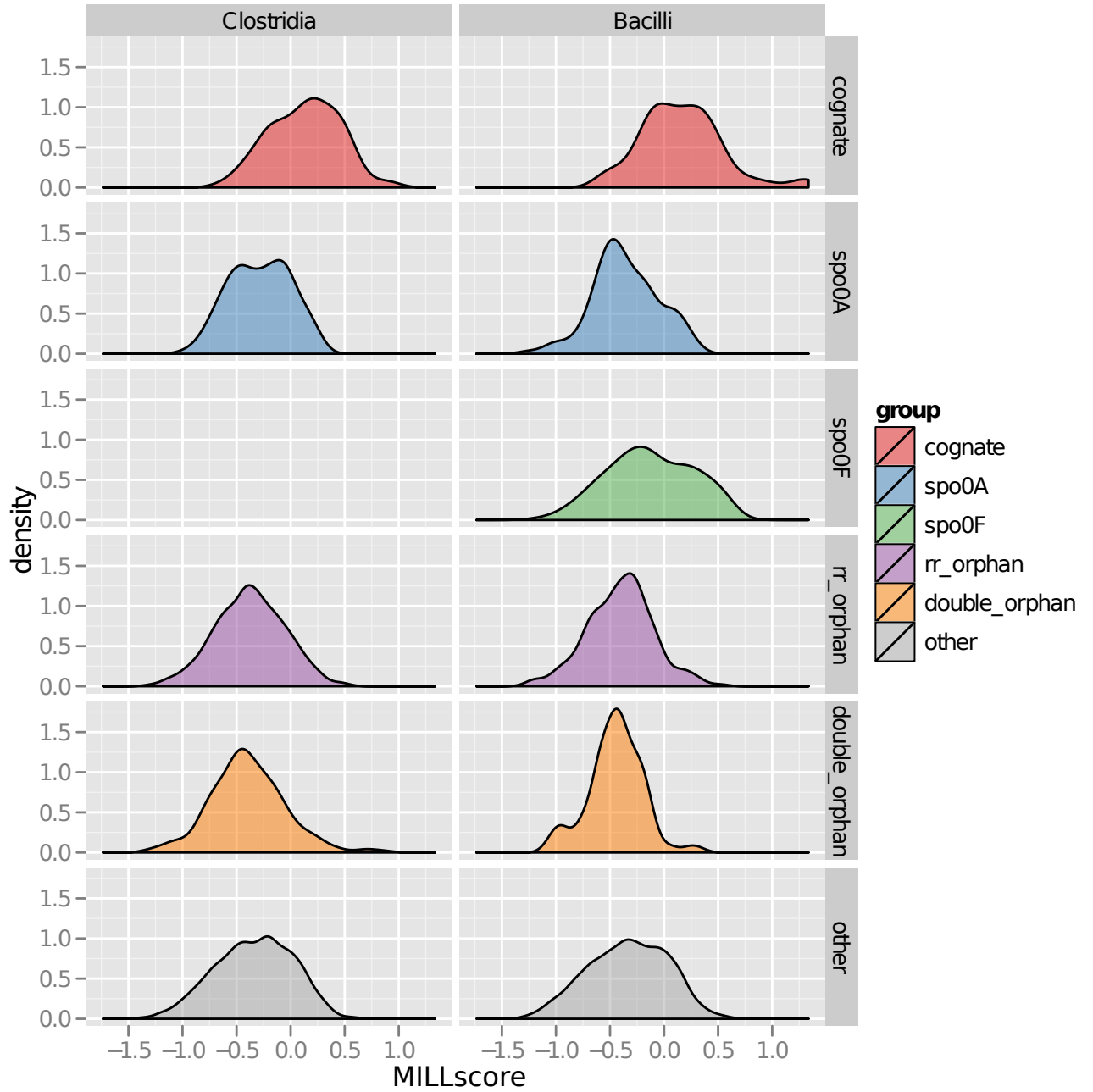


Figure 4.10: The distribution of MILLscore for the different classes of HK/RR pairings listed in Table 4.2, in *Bacilli* and *Clostridia*

4.3.4 Effect of Alignment Quality on MILLscore

In order to explore one possible source of inaccuracy in the Clostridia sporulation kinase predictions it is first noted that a kinase with inaccurate predictions tends to have lower scores across all RRs and a lower score with its known partner RR (if there is one). For instance, CA_C0323 has consistently low scores across all RRs and a low score against Spo0A (this was identified as a sporulation kinase in [135]). The same is true for some kinases in *B subtilis* for which the MILLscore failed, e.g. BSU07580 has consistently low scores and a low score with its cognate RR.

It is hypothesised that this behaviour could be due to these kinases being poorly aligned. The MILLscore procedure uses a template structure to define important residue pairings for predicting PPIs. As such, it is important that the alignment of query sequences to this template is accurate to ensure the equivalent residues are correctly identified. If this is not the case then the score against all RRs will be affected, leading to the behaviour outlined above. In practice, query sequences will be placed in a multiple alignment with the sequence from the template structure. As such, the alignment of each query sequence to the template can be assessed in terms of the robustness of alignment of the sequence, within the multiple alignment.

To evaluate the effect of alignment quality on these scores, the GUIDANCE score method [137] is used to measure how robustly each sequence is aligned within the multiple alignment. The GUIDANCE score is calculated by perturbing the order in which the sequences are aligned to produce the multiple alignment. This produces an ensemble of alignments from which a measure of how consistently each sequence is aligned is calculated. This is the GUIDANCE score and serves as a measure of how robustly aligned each sequence is.

The GUIDANCE score of each aligned protein can then be compared to its MILLscore with its cognate and also to its mean MILLscore (Figure 4.11). This can be done for both HK and RR subunits. Assuming that each two-component subunit interacts only with its cognate, a useful predictor of specificity would assign the interaction with the cognate a higher score than with non-cognate. As expected, the MILLscore does so, however its ability to do so appears to be dependent on alignment quality; the robustly aligned sequences have higher scores with their cognate and a higher separation against the average score. Although the HKs are universally less robustly aligned than the RRs, as shown by the difference in GUIDANCE scores between the two groups, this relationship is true for both HKs and RRs.

4.4 Discussion

Recently, a similar approach to the MILLscore was described in [126]. The test set used in this paper was unavailable for comparison, however, several specific examples were described in [126], to which the MILLscore can be applied. The MILLscore produces more accurate predictions of the partners of 5 orphan kinases in *B. subtilis*. The accuracy of the two approaches being similar, a big advantage of the MILLscore is the time of computation. The previously described approach uses an expensive computational approach (~ 2 days) to estimate from a training alignment, the residue-residue pairings important for mediating the PPI. The MILLscore estimates these pairings directly from a template structure taking around 15 minutes for a similar computation. The obvious downside is the requirement of a template structure, however, for many well studied systems for which a large training set exists, interaction structures are available for at least 1

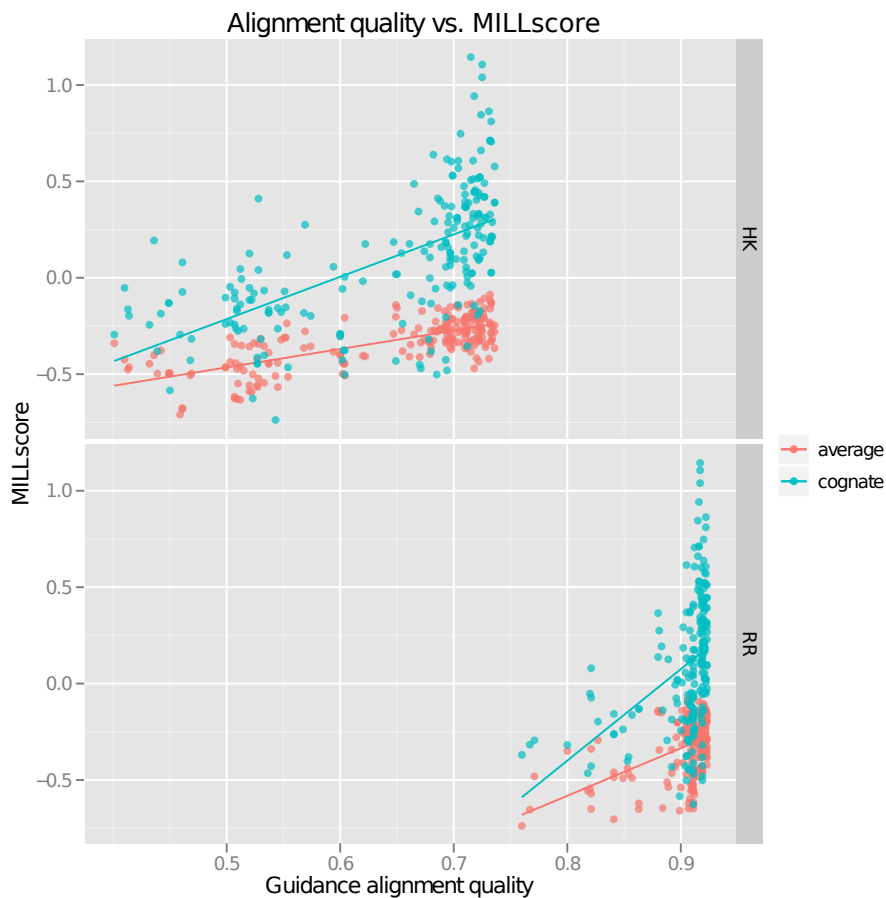


Figure 4.11: Comparison of Guidance alignment scores with MILLscores against cognate and average partners for both HK and RR subunits. The top frame shows, for each histidine kinase in the data set, the average MILLscore across all RRs in the same species (red points). The MILLscore for that kinase with its cognate is also shown in blue for HKs having cognate pairs (i.e. there are no blue dots for orphans). We can see that as expected, cognates are scored higher than average in the majority of cases. However, this is less true of proteins that are hard to align. The x-axis shows the GUIDANCE score and it can be seen that the gap between average and cognate scores is lower for proteins with a low GUIDANCE score (i.e. they are hard to align). A similar pattern is seen when considering the alignment of RR proteins (bottom frame)

PPI.

Two component systems display high specificity, with most pairings of HK/RR observed to be exclusive. It has been observed that a small number of residues are responsible for maintaining this specificity and pairings can be rewired by a few mutations at these positions. The ability of the MILLscore to predict the effect of individual substitutions on PPI specificity was tested by comparison of MILLscore predictions with the results from [131]. In [131], a sequence of mutations was shown to convert the specificity of the RstA HK in *E. coli* to that of the OmpR HK and vice versa. The MILLscore can accurately recapitulate the effect of this sequence of mutations on specificity, however the specificity at intermediate steps along the sequence appears to be less accurate.

One obvious application of the MILLscore, outside of its use for the interaction tree, is prediction of interactions between orphan kinases and response regulators in two component systems. The ability of the MILLscore to recapitulate the interactions of known sporulation kinases in *B subtilis* was shown. Predictions of interactions were then made for the orphan components in several Bacilli and Clostridia species. In many cases, such as the prediction of sporulation kinases in *B anthracis*, the MILLscore agrees with the available experimental evidence. In other cases, such as the prediction of interaction for the orphan kinases of *C acetobutylicum*, the agreement is not as strong. It is hypothesised that one possible contribution to the failure of the MILLscore in some cases could be the unreliable alignment of some kinases. It is then shown that the MILLscore is indeed sensitive to alignment quality, as would be expected given the reliance on aligning query sequences to a template structure. This could present a problem with hard to align sequences. However, this also means that higher confidence can be assigned to predictions of well aligned sequences, as measured by GUIDANCE.

4.5 Conclusion

In conclusion, the MILLscore method is a fast and accurate approach to predicting transient PPIs. The accuracy of this approach in predicting bacterial signalling interactions is shown to equal that of [126] while the MILLscore is two orders of magnitude faster. This speed up comes as the result of using available structural information to inform predictions and so is only possible when such a structure is available. The sensitivity of the predictions to alignment quality is demonstrated.

Species	Average z-score of cognate PPI prediction
<i>Bacillus subtilis</i>	1.56
<i>Bacillus anthracis</i>	1.57
<i>Bacillus cereus</i>	1.46
<i>Clostridium acetobutylicum</i>	1.45
<i>Clostridium difficile</i>	1.36
<i>Clostridium botulinum</i>	1.40
<i>Clostridium perfringens</i>	1.36

Table 4.4: The average z-score for the MILLscore of cognate HK/RR pairs in each of the 7 species included in the analysis. Although predictions are stronger in some species, in all cases these scores mean that (on average) the MILLscore of an HK with it's cognate RR is higher than 90-94% of other RRs from that species.

4.6 Appendix : Full Predictions in 7 Bacterial Species

This appendix contains the plots of two-component system predictions in the 7 species, presented in the same order as listed in Table 4.1. The standardised z-scores of the prediction of PPIs between cognate HK/RR pairs are shown in Table 4.4. The calculation of these scores is explained in the main text.

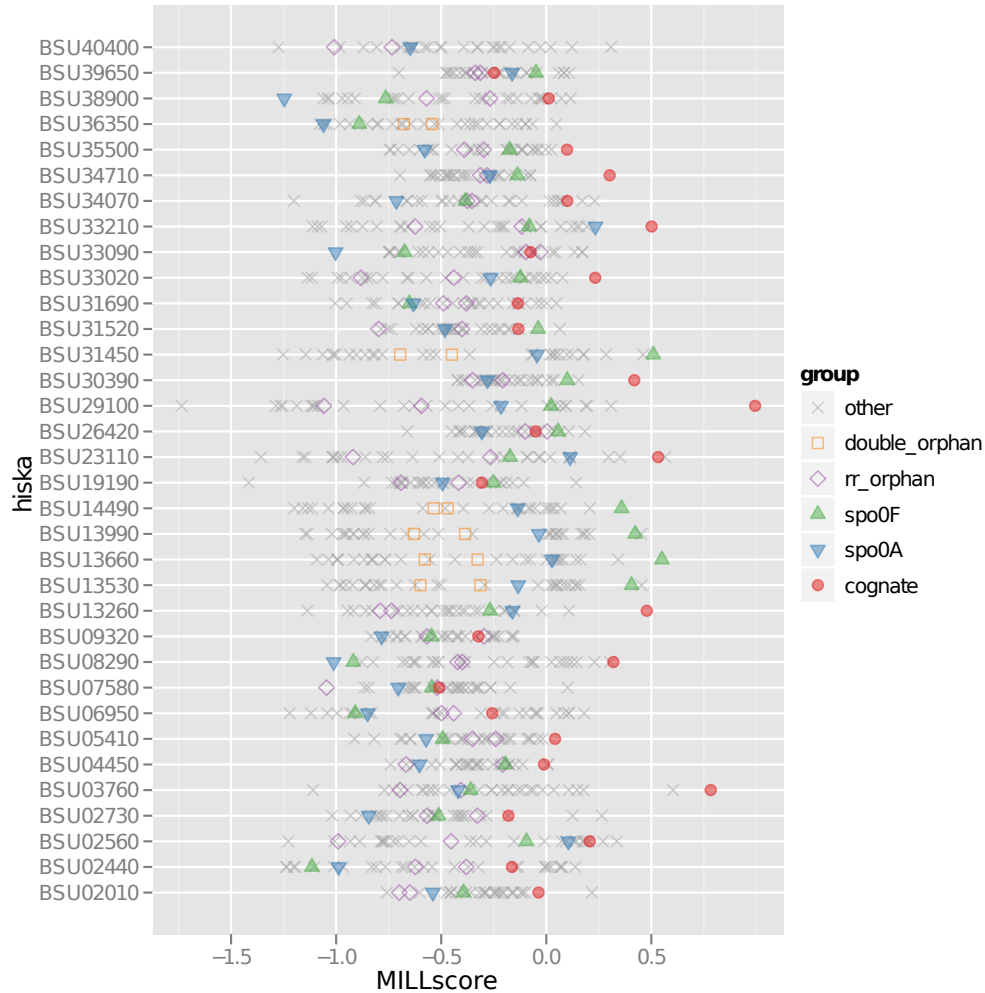


Figure 4.12: MILLscores for the full set of histidine kinases from *Bacillus subtilis* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case.

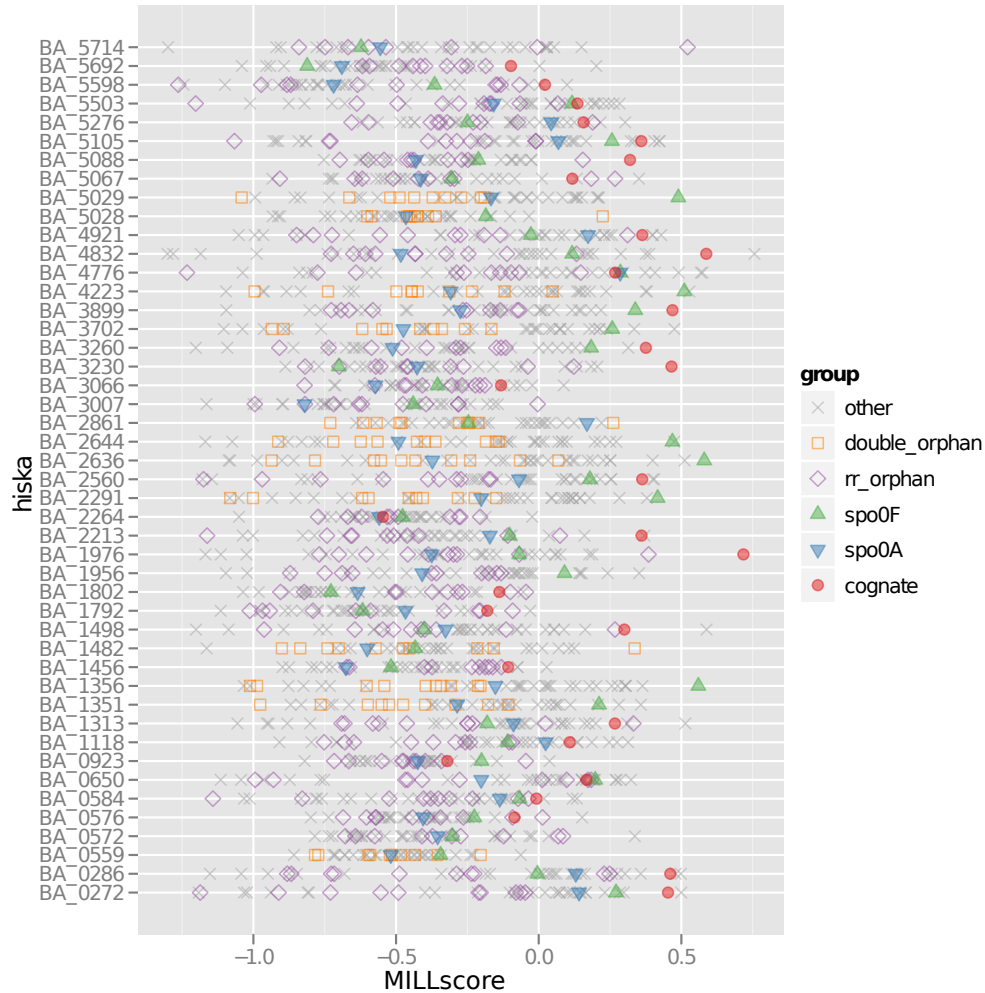


Figure 4.13: MILLscores for the full set of histidine kinases from *Bacillus anthracis* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case.

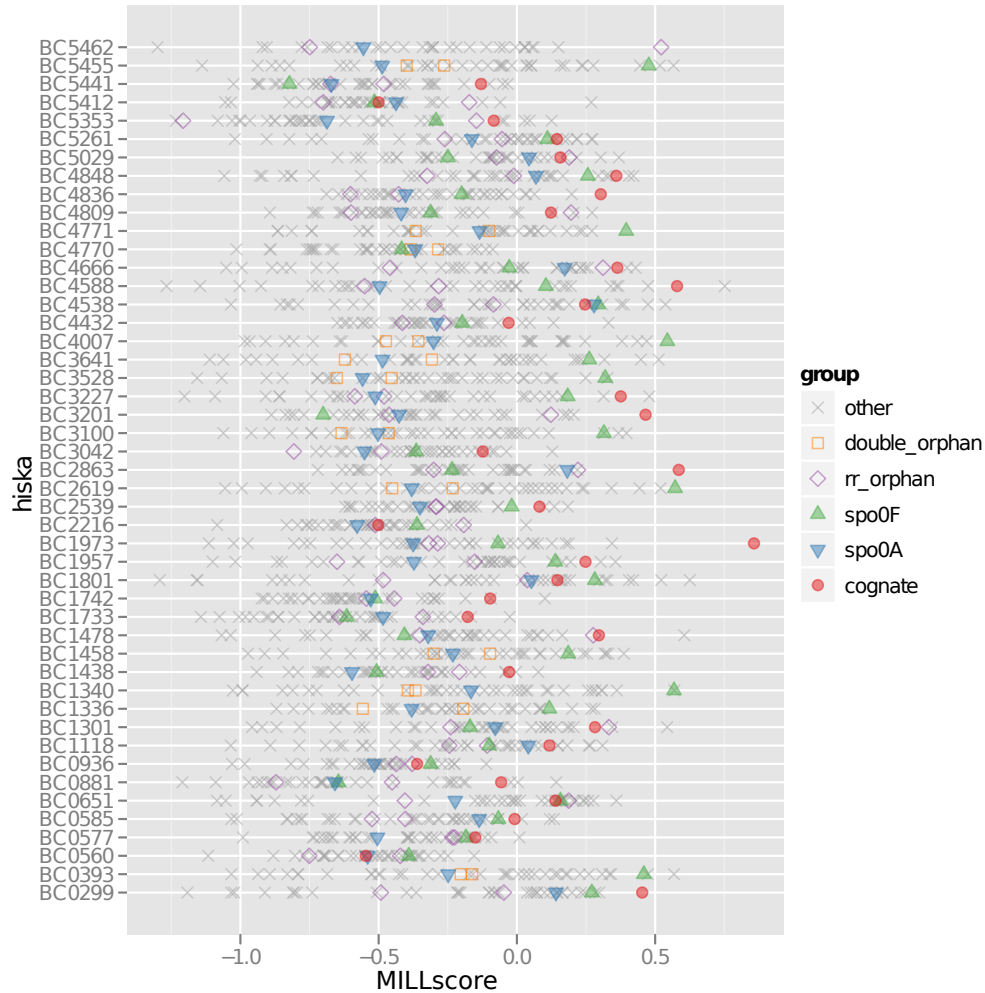


Figure 4.14: MILLscores for the full set of histidine kinases from *Bacillus cereus* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case

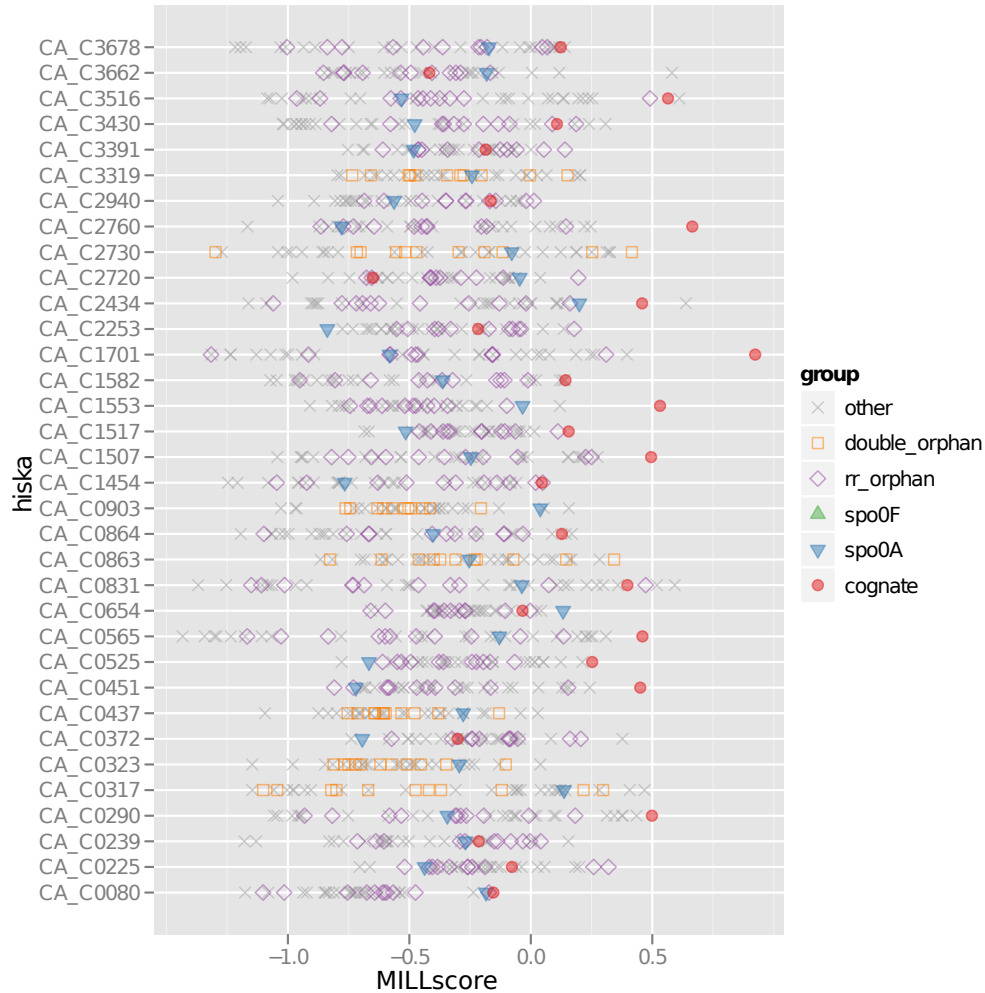


Figure 4.15: MILLscores for the full set of histidine kinases from *Clostridia acetobutylicum* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case

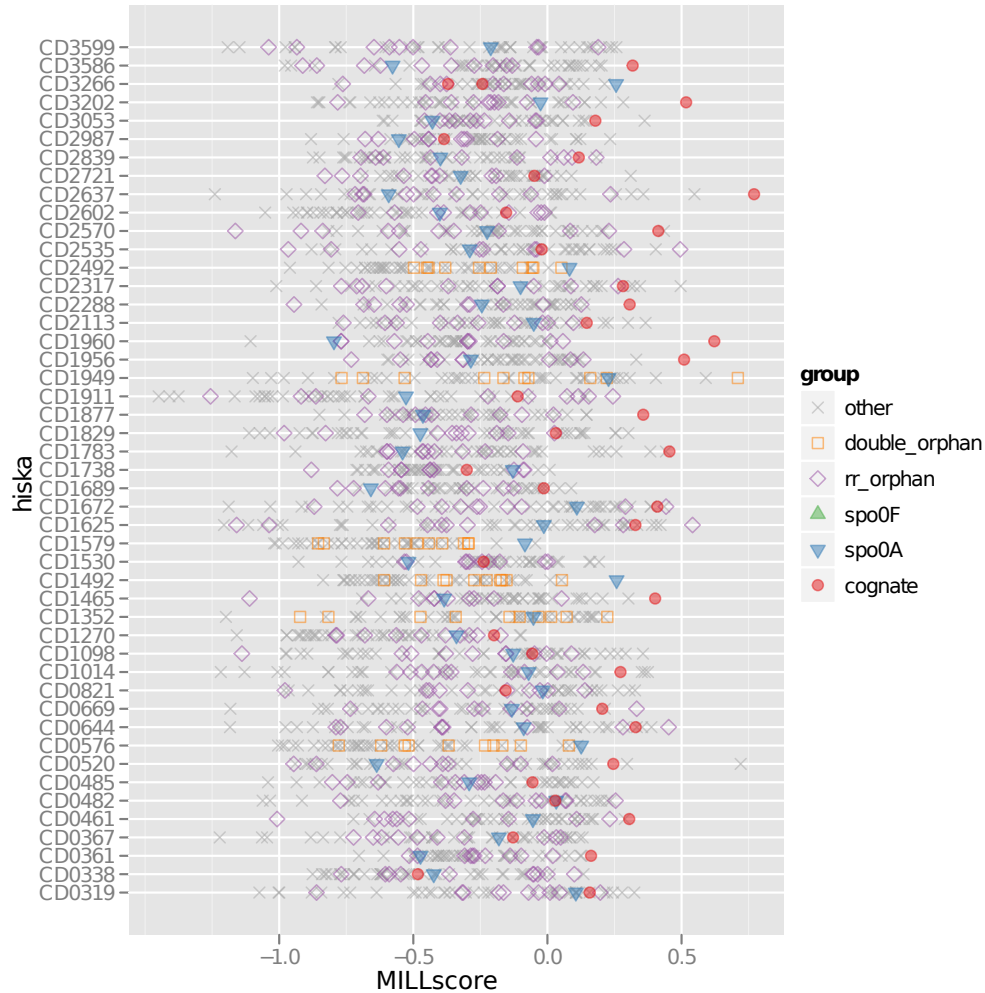


Figure 4.16: MILLscores for the full set of histidine kinases from *Clostridia difficile* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case

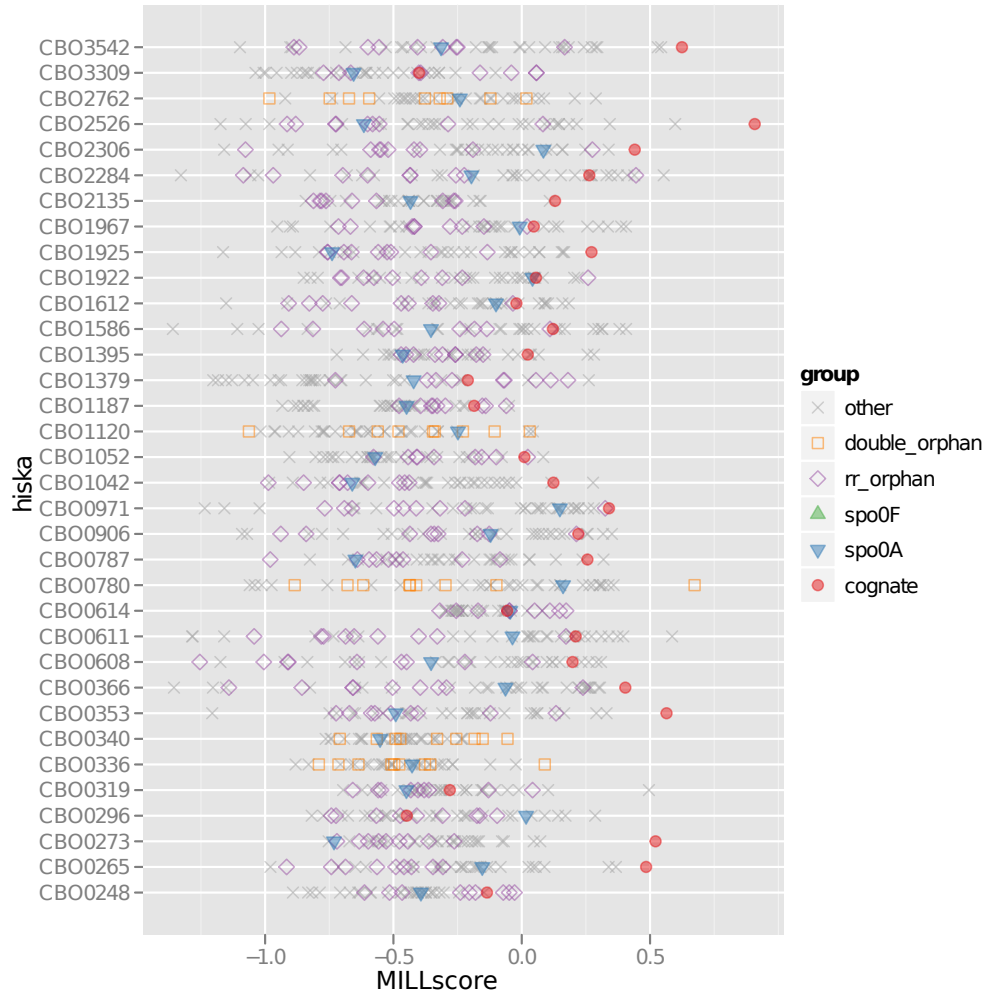


Figure 4.17: MILLscores for the full set of histidine kinases from *Clostridia botulinum* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case

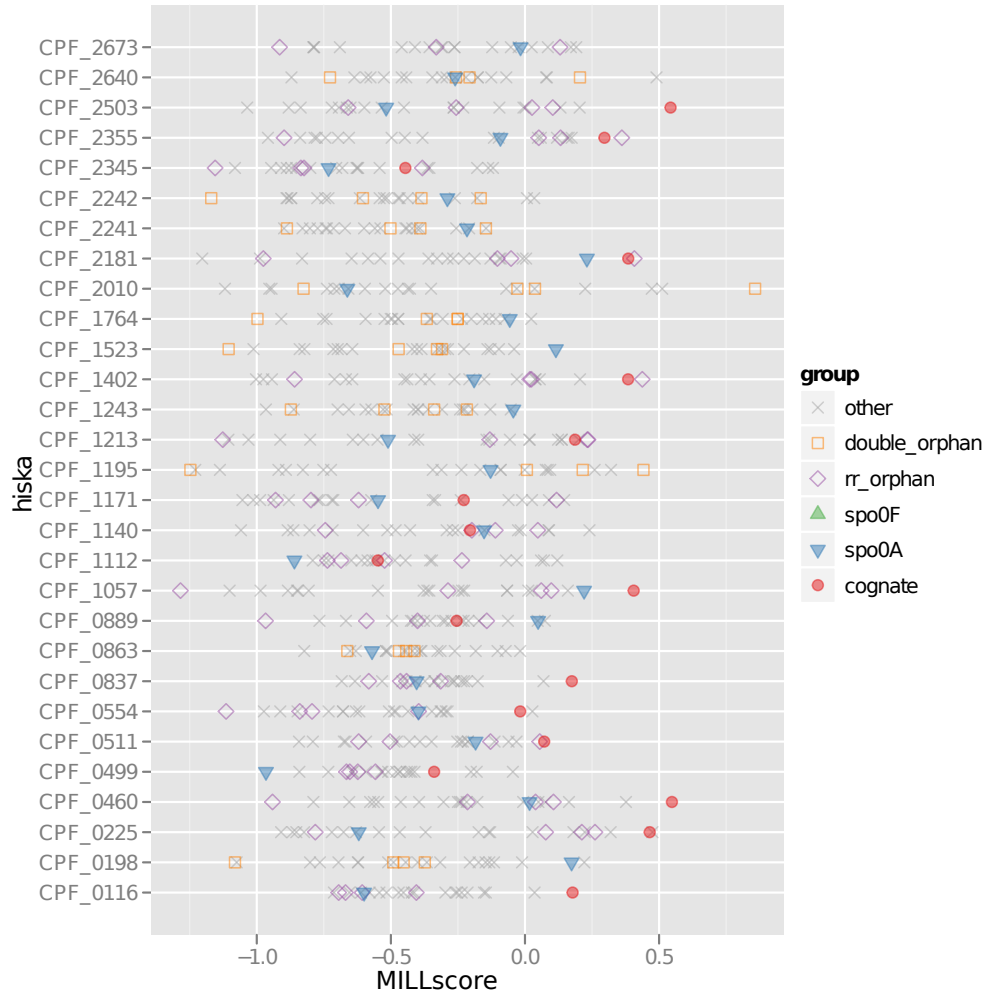


Figure 4.18: MILLscores for the full set of histidine kinases from *Clostridia perfringens* scored against the full set of response regulators. The scored pairs fall in to 6 categories (described in Table 4.2). Cognate pairs are shown as red points. Each row of points shows the MILLscore of a given histidine kinase against every RR from the species. Given that cognate pairs interact in the vast majority of cases, a high scoring cognate (i.e. red point to the right of the whole set of points on that row) means that the MILLscore can predict the PPI in this case

Chapter 5

Reconstructing PPI history for transient interactions

5.1 Introduction

Previous interaction tree applications have considered transient interactions (in much the same way as permanent interactions) but have assigned constant probability of rewiring events at each step of evolution. This assumption is clearly wrong and leads to very coarse grained models that may miss the finer detail. For instance, the previous model of [87] assigned a constant probability of rewiring after a duplication/speciation event and used this in an interaction tree framework. This was used to model PPIs in the proteasome complex, producing a dense set of predicted PPIs amongst the proteasome subunits. The authors argued that the detection of such a clustered set of PPIs shows the models ability to detect protein complexes (as densely connected areas of a PIN). While this is certainly true, there are many situations in which the fairly coarse grained question “What are the protein complexes?” is replaced with the finer grained question “What are

the interactions within the protein complexes?”. Quite clearly, the PPIs within the proteasome do not form the densely connected set described in [87], on average a yeast proteasome subunit interacts only with 5 others, not with every other subunit in the complex. In order to model the rewiring within a complex, a model is required that can assign probabilities of rewiring independently for pairs of proteins, based on their sequence change during evolution.

In chapters 2 and 3 such an approach that avoids assigning constant probabilities to rewiring events was described and applied to model successfully the evolution of obligate/permanent PPIs within a protein complex. This method assigns probabilities of rewiring based on the sequence change seen at the protein-protein interface allowing a more detailed model of PPI evolution. The method of assigning probabilities here is based upon tracking the evolutionary change in a simple measure of physicochemical complementarity that is predictive of interaction between proteins, in the obligate examples. In the last chapter, it was shown that this approach does not generalise to transient interactions (in particular, two-component systems), as the previously used scoring measure is not predictive of rewiring events in transient systems.

In the previous chapter, a major step in generalising the interaction tree approach was made. The MILLscore was shown to be able to predict transient PPIs quickly and accurately, making it a good candidate to replace the use of the SCOTCH score used in chapters 2/3. The MILLscore can be used to predict PPIs between proteins of one family with proteins from another family. For instance, given an alignment of proteins from family A and an alignment of proteins from family B, which A-B pairs interact? The ability to produce predictions is dependent on the existence of a known structure of a PPI and a set of pairs of A-B sequences known to interact. These stipulations restrict the class of problems that the MILLscore can be applied to, however there are systems of interest fulfilling

these requirements. For this chapter, the two component systems will remain the PPI family of interest.

In order to complete the formulation of an interaction tree model, using the MILLscore, the MILLscore needs to be used to describe a model of PPI rewiring. This was achieved for the previous SCOTCH-based model by considering “hypothetical” interactions between existing pairs of proteins. A similar approach is taken here to define the probability of a rewiring event occurring given the change in MILLscore between two proteins. This model can then be applied to transient PPIs, allowing a protein specific (i.e. not a constant probability of rewiring) model of PPI evolution for this class of interactions, where the previous model failed.

The model has two immediate uses which are explored here. Firstly, the model is used to enhance, where possible, the prediction of interaction amongst existing proteins. As shown in Chapter 3, the interaction tree can be used to produce a probability of interaction for existing protein pairs. This approach is used here in an attempt to improve on the predictions of sporulation kinases at the end of the last chapter and progress is made in describing when this approach is suitable. The second use is in reconstructing the history of two component system PPIs. Of particular interest here is the evolution of the two component systems responsible for sporulation in bacteria. As described in the previous Chapter, the PPIs responsible for this process differ between the *Bacilli* and the *Clostridia*: *Bacilli* employ a phosphorelay, *Clostridia* a single canonical two component interaction. One open question is this: What was the ancestral state of the system? *Bacilli*-like or *Clostridia*-like? The interaction tree offers answers to these questions and this approach in the second part of this Chapter.

5.2 Methods

5.2.1 Two component system sequences

The two component system remains the focus of application in this chapter. The set of bacterial species included in the analysis of the previous chapter are once again used as a test set when applying the interaction tree methodology. To recap, the sequences of all HK/RR proteins were downloaded for the species listed in Table 5.1. All HKs in this set are aligned using 3Dcoffee [26] and the RRs are aligned similarly, using the 3DGE [138] structure from the Protein Data Bank [104] (this is the structure of a two component system interaction from *Thermatoga maritima*). This produces an alignment of 278 HKs and a separate alignment of 283 RRs to use as a test set for an interaction tree model in this system.

5.2.2 Phylogenetic trees

To build an interaction tree describing the evolution of PPIs in this set, phylogenetic trees must first be produced for both the HK and RR alignments. Each of these tree describes a branching process that gives rise to the observed set of HK/RR proteins and their sequence divergence, as shown in the multiple alignment. A tree was generated for both alignments using PHYML [139], with default parameters, using 1000 bootstrap resamples to produce bootstrap values at each internal branching node of the tree. Each internal node of a tree can be either the result of a gene duplication event (a gene is duplicated to produce two genes in a species) or a speciation event (a species becomes two species, with an independent copy of the gene in each). In order to assign each node to one of these two categories, NOTUNG [34] is used to reconcile the tree with a species tree taken from ITOL [109]. Rearrangement of poorly supported branches to produce agree-

Species	HK proteins	RR proteins
<i>Bacillus subtilis</i>	35	33
<i>Bacillus anthracis</i>	46	46
<i>Bacillus cereus</i>	50	45
<i>Clostridium acetobutylicum</i>	35	40
<i>Clostridium difficile</i>	47	53
<i>Clostridium botulinum</i>	36	42
<i>Clostridium perfringens</i>	29	24
Total	278	283

Table 5.1: List of species included in the analysis of this chapter and the number of histidine kinase and response regulator proteins included from each. These represent all known two component system proteins from these 7 species and the set for each species contains both cognate and orphan proteins (the difference between cognates and orphans is explained in the introduction to this chapter and Figure 4.2)

ment with the branching order of the species tree is performed as before (reference previous chapter), using a bootstrap threshold of 50%. This process produces a phylogeny for both the HK and RR proteins in set of Table 5.1. The procedure outlined in Chapter 3 can then be used to combine these trees to produce an interaction tree.

5.3 Results

The utility of applying the interaction tree methodology to transient interactions is twofold: firstly, the interaction tree can produce predictions of the evolutionary history of PPIs in a system. This allows the ancestral PPIs to be inferred. For instance, as previously described in Chapter 4 the two component system PPIs responsible for sporulation onset are fundamentally different in Bacilli versus Clostridia; the Bacilli have a phosphorelay containing four proteins, the Clostridia have just one PPI involving two proteins. An obvious question is this: what was the ancestral state of this system? The interaction tree approach can answer this question by predicting the PPIs present in an ancestral species.

The second use of the interaction tree is in predicting PPIs between existing proteins. As shown earlier (Chapter 3) in the proteasome, the interaction tree framework can use a model of PPI rewiring to predict existing PPIs. This is done by modelling the rewiring events between some known set of PPIs and the PPIs of interest. Previously, this approach was shown to improve predictions using a PPI rewiring model based on the SCOTCH score compared to using the SCOTCH score alone.

In order to apply the interaction tree in these ways to the transient interactions of two component systems, a model of PPI rewiring based on the MILLscore needs to be defined. Previously, such a model was produced based on the SCOTCH

score by considering hypothetical evolution between existing pairs of proteins. The same approach is taken here in order to relate changes in the MILLscore to changes in interaction state between two proteins. The Weigt alignment used to train the MILLscore is used to generate the training set of hypothetical interaction tree branches required. This alignment consists of cognate HK/RR pairs which are taken to be known PPIs and an equally sized set of non-interacting pairs is generated by randomly choosing a HK and RR from the alignment, that are not cognates. These observed interaction nodes are then combined as before to produce a set of observed (hypothetical) interaction tree branches to be used as a training set.

The probability of either a gain or loss of interaction between proteins, for a given change in MILLscore, can then be calculated empirically from this training set (Figure 5.1) in exactly the same way as done previously in Chapter 3. As is expected, increases in MILLscore are associated with increasing probability of gains of PPI. Conversely, decreasing MILLscore produces increasing probability of loss of interaction. Unlike the previous SCOTCH based model, the two empirically derived probability functions are almost perfectly symmetric about 0. This implies that under this MILLscore model, negating the change in score switches $P(gain)$ and $P(loss)$. In order to describe these functions formally, curves are fitted to the points with the following form

$$P(gain|M) = \frac{1}{1 + e^{((A_g - M)/B_g)}} \quad (5.1)$$

$$P(loss|M) = \frac{1}{1 + e^{((A_l + M)/B_l)}} \quad (5.2)$$

with fit parameters $A_g = 2.032$, $B_g = 1.709$ and $A_l = 2.313$, $B_l = 1.678$

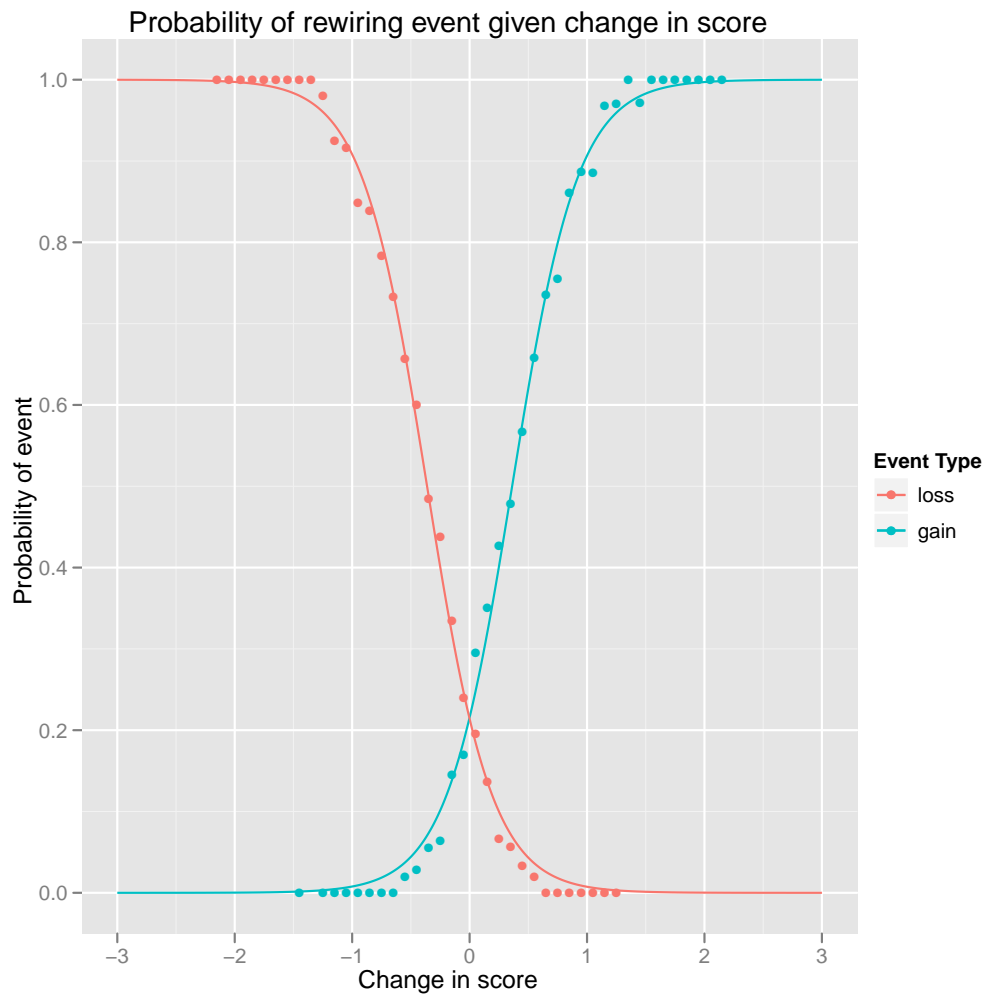


Figure 5.1: The relationship between change in MILLscore and probability of rewiring event. The points are the empirical probabilities derived as described in the text. The curves are fitted functions of the form described in Equation 5.1

using the *nls* function in R with default settings. The symmetry of the gain and loss functions is captured in the similarity of the fit parameters in the two cases ($A_g \approx A_l$ and $B_g \approx B_l$).

These functions define a model of the same type as the previously defined SCOTCH model. Given an interaction tree structure, with reconstructed sequences at the ancestral nodes, this can model the gain and loss of PPIs during evolution.

5.3.1 Predicting present day PPIs

The first application, of this new model of transient PPI evolution, is predicting PPIs between existing proteins. Previously, in Chapter 3, the D_{com} model was used in this way to predict PPIs in the cattle proteasome. This provides an obvious way to validate the model by generating predictions for known interacting pairs and evaluating the agreement of predictions and reality. This approach can also be used to generate new predictions of PPIs.

In the case of the MILLscore model, a similar approach is taken. Using the two component systems of the species of Table 4.1 as a test set, the MILLscore interaction tree model is applied to predict existing PPIs given some known input PPIs. Producing results that can be directly compared to the predictions of the last chapter, based on the MILLscore alone.

Unfortunately, in this scenario, the full set of 278 HKs and 283 RRs produces an interaction tree that is too large for analysis in Matlab using the Bayes Net Toolbox (producing an interaction tree of 23,239 interaction nodes). Instead of considering the full phylogeny, a set of smaller, manageable analyses are undertaken. Firstly, a smaller analysis is undertaken, in which the known two component system interactions of *B subtilis* are used as input to predict the two component

systems in *B anthracis*. To do this, the full phylogeny of the HKs and the full phylogeny of the RRs are firstly pruned to contain only the proteins from these two species (i.e. a tree of 81 HKs and a tree of 79 RRs). These are combined to produce an interaction tree of 6,541 nodes. As before, these interaction nodes can have state 1 corresponding to a PPI or state 0, corresponding to no PPI. As input, all cognate pairs in *B subtilis* are taken as observed, as are the five known sporulation kinases paired with Spo0A. This corresponds to assigning the corresponding interaction nodes to state 1. All other pairings of a *B subtilis* HK and RR are then taken to be observed with interaction state 0.

The interaction tree framework can then be used to model the PPI gains and losses between these observed inputs and the two component proteins of *B anthracis*. This is achieved using the same message passing algorithm as before, with the conditional probability function replaced by that described in Figure 5.1. The output of this algorithm is a probability for each HK/RR pairing in *B anthracis* to be in state 1, i.e. to interact. This output is shown in Figure 5.2.

These predictions can be validated by assessing their ability to detect the cognate pairs (Table 5.2). This ability can be measured, in the same way as in the previous chapter, by examining the rank of the cognate RR amongst all scores for non-orphan HKs. 15 out of 32 non-orphan HKs have their cognate RR as their most probable interaction. This compares favourably with the predictions of the previous chapter (using just the MILLscores with no interaction tree based inference) in which 6 out of 32 of these cognates were predicted correctly. However, when the threshold is extended to count any HK with its cognate RR in its top 5 scores, this situation is reversed: the interaction tree approach places the cognate RR in the top 5 for 20 of the non orphan kinases, the previous approach in 24 cases (Table 5.2). The average z-score of these predictions is computed (as described in the previous chapter) and is found to be 1.37 for the interaction tree predictions.

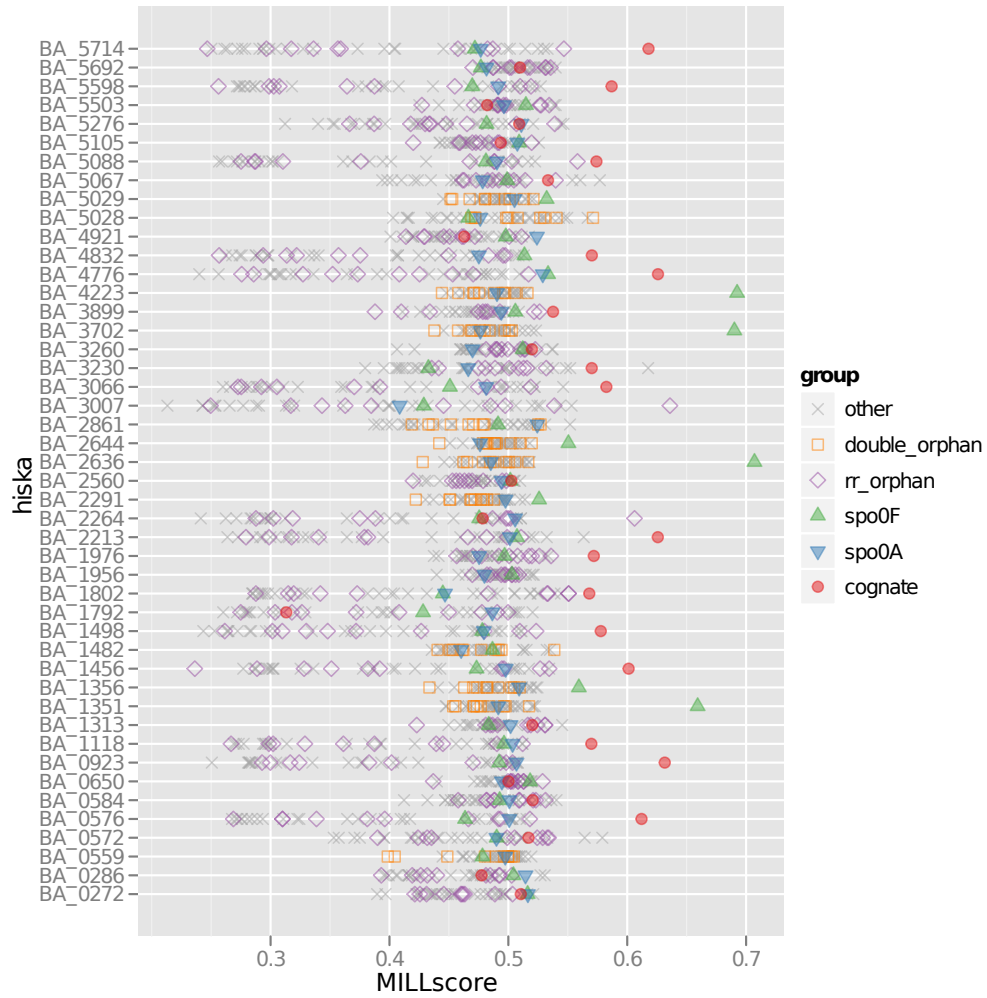


Figure 5.2: Interaction tree predictions of two component system interactions in *B anthracis*, using a model based on the MILLScore and the interactions of *B subtilis* as input.

This shows that despite the increased number of top 1 predictions, on average the cognate pairs are not predicted as strongly using the interaction tree method. A score of 1.37 means that, on average, for a given HK, its cognate RR scores higher than 91% of the other RRs.

	Top 1 predictions	Top 3 predictions	Top 5 predictions	Avg Z-score
MILLScores	6 of 32	15 of 32	24 of 32	1.57
MILLScores + interaction tree	15 of 32	16 of 32	20 of 32	1.37

Table 5.2: Comparison of two component system predictions in *B anthracis* using the MILLScore alone and using the MILLScore interaction tree model. There are a total of 32 cognate histidine kinases in *B anthracis* so, for instance, 15 of 32 (47%) of kinases had their highest probability of interaction with their cognate RR using the MILLScores + interaction tree method.

It seems then that using the interaction tree approach produces more very high confidence interactions than using the MILLScore alone. This shows the benefit of including the phylogenetic information from the interaction tree in to predictions. However, the predictions are also worse for a proportion of the kinases, as can be seen by the deteriorating “Top 5 predictions” metric. It is clear then, that in these cases the phylogenetic information is not helping the prediction, indeed it may be misleading. To understand why the interaction tree drastically improves predictions for some kinases and worsens for others, a successful prediction is first examined.

BA_4776 has its cognate RR *BA_4777* as its highest scoring RR. This is an improvement from being the 8th highest scoring in the previous chapter, based

on the MILLscore alone. The HK/RR phylogenies provide a clue as to why this might be (Figure 5.3). The *B anthracis* HK *BA_4776* neighbours the *BSU02560* HK from *B subtilis* in the HK tree (left side Figure 5.3). Its cognate RR *BA_4777* is neighbouring *BSU02550* in the RR tree (right side Figure 5.3). This means that the *BA_4776* – *BA_4777* interaction node neighbours the *BSU02550* – *BSU02560* interaction node in the interaction tree (Figure 5.4).

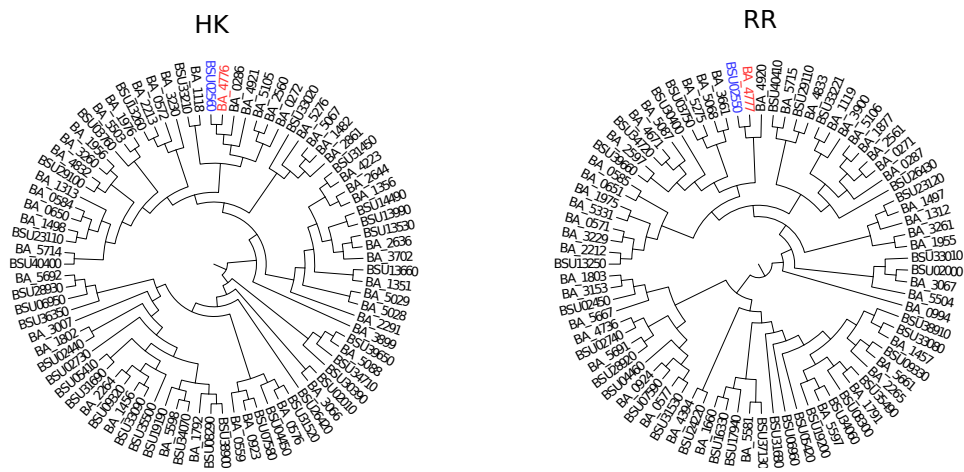
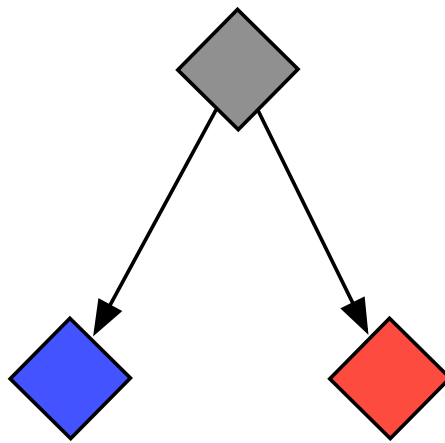


Figure 5.3: Preservation of cognate relationships. The HK tree of *B subtilis* and *B anthracis* proteins is shown on the left, the tree of RRs is on the right. *BA_4776* is highlighted red in the HK tree and its cognate RR *BA_4777* is highlighted red in the RR tree. In this case, the orthologues of these proteins in *B subtilis* (highlighted in blue) are themselves cognates.

The *BSU02550* – *BSU02560* node is itself a cognate pair and is included in the input evidence. Given the short distance in the tree, this evidence is easily transferred to the *BA_4776* – *BA_4777* node, resulting in a high probability of a PPI at this node.

In contrast, the *BA.1792* HK has a very low score against its cognate RR,



BSU02550-BSU02560 **BA_4776-BA_4777**

Figure 5.4: As a result of the preserved cognate relationship in this example, there is a local structure within the interaction tree as shown above. In this case the blue interaction node is observed with state 1. The closeness in the tree of the red node means that this information is easily transferred to this node, resulting in a useful prediction.

BA_1791. Again, the phylogeny provides clues as to the cause of this: the cognate relationship is not preserved in the tree structure in the same way for this pair (Figure 5.5). It is not clear if this is because the reconstructed phylogeny is wrong or the true phylogeny simply misleading in this case. It is clear however that the interaction tree predictions are unreliable in such situations.

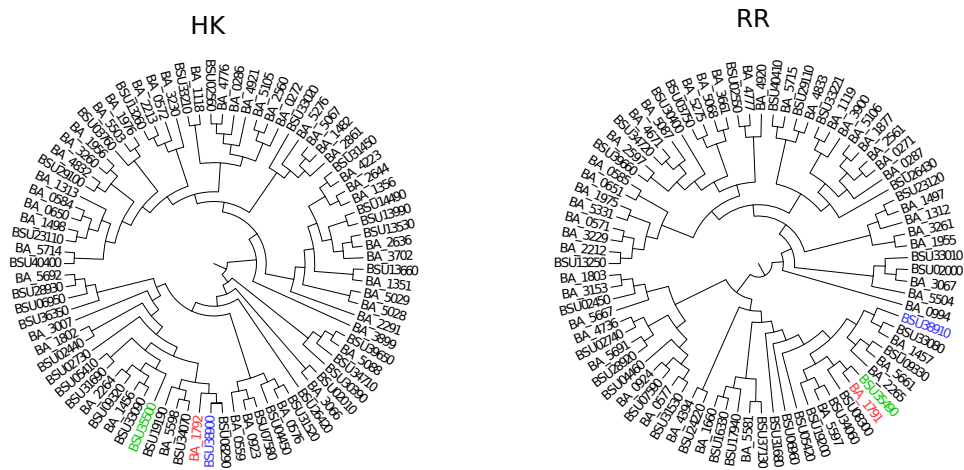


Figure 5.5: In this example, the BA_1792 and BA_1791 cognate pair are highlighted in red, as before. However, the closest ortholog of the BA_1792 HK (highlighted blue) has its cognate located elsewhere in the RR tree (coloured blue). Conversely, the closest ortholog of the BA_1791 RR (green) has its cognate HK located distantly from BA_1792. In this case the cognate relationships are not preserved in the tree. That is, the orthologs of cognate pairs are not themselves cognates.

This analysis also produces predictions for the interactions of the orphan kinases. In the previous chapter, predictions were made based on the MILLscore, for the orphan kinases of *B anthracis*. These predictions were then compared to the experimental evidence available [136]. Now the predictions from this interaction tree analysis can be added to the comparison (Figure 5.3). These predictions are very consistent with the previous predictions, showing the same agreement to the

<i>B anthracis</i> protein	Proposed as candidate	Top 3 MILLScore	Top 3 interaction tree	Experimental evidence
BA_5029	Y	Y	Y	Y
BA_4223	Y	Y	Y	Y
BA_3702	Y	Y	Y	N
BA_2644	Y	Y	Y	-
BA_2636	Y	Y	Y	N
BA_2291	Y	Y	Y	Y
BA_1482	Y	N	N	-
BA_1356	Y	Y	Y	Y
BA_1351	Y	Y	Y	N

Table 5.3: Comparison of the predictions of orphan kinases that interact with Spo0F based on the interaction tree analysis (from Figure 5.2) with experimental evidence available from [136]. The predictions agree precisely with the previous predictions based on MILLScore alone and agree also with the experimental evidence.

experimental evidence. However, it is worth noting that this ability to predict the orphan kinases disappears if the known orphan interactions from *B subtilis* are removed from the input evidence, Figure 5.6 (i.e. the prediction is made based on the cognate pairs of *B subtilis* alone).

5.3.2 Predicting Clostridial PPIs from a Bacillus

For the next application of the MILLScore interaction tree model, the two component system interactions of *C acetobutylicum* are predicted using the same set of input PPIs from *B subtilis*. To begin, the full phylogenies are pruned to contain

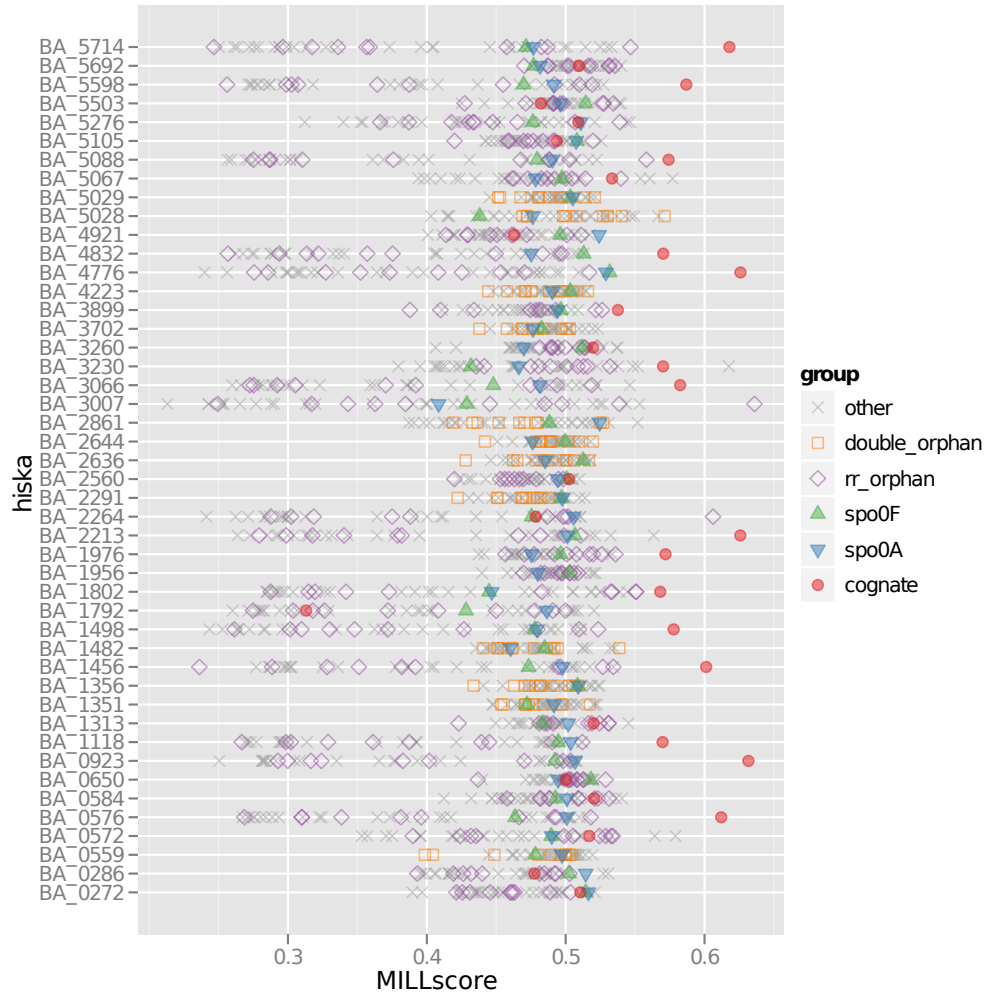


Figure 5.6: Prediction of two component system interactions in *B anthracis*, using cognate two component pairs from *B subtilis* as observed input. No information regarding the sporulation kinases of *B subtilis* is included. The previous predictions of kinases that are likely to interact with Spo0F disappear (compare to Figure 5.2), i.e. there are no green triangles to the right of the figure.

only two component proteins from *B subtilis* and *C acetobutylicum*. This produces a tree of 70 HK proteins and a tree of 73 RR proteins that are combined to produce an interaction tree as before. This tree structure is then used to generate probabilities of two component system PPIs as before but now in *C acetobutylicum* Figure 5.7.

In the same way as before, these predictions can be compared to those made in the last chapter based on the success in predicting cognate pairs (Table 5.4). It is clear that in this case the interaction tree does not bring any benefit to the predictions. Indeed, the phylogenetic information appears to be misleading. This is also the case for the predictions of sporulation kinases amongst the orphan kinases: CA_0317 has the highest probability of interaction of Spo0A but was found not to interact with this RR in [135]. Of the three kinases linked to sporulation in [135] (CA_C0323, CA_C0903 and CA_C3319), only CA_C0903 has Spo0A as its most probable interactor.

So, the interaction tree can potentially produce more high confidence predictions of PPIs in a *Bacilli* species, using known PPIs from another *Bacilli* as input (whilst decreasing the average prediction accuracy). However, predictions in a *Clostridia* species were not improved using *Bacilli* interactions as input. Consistent with the findings of Chapter 3, it seems that the interaction tree works best when PPI evidence is included from some closely related species. In this case, it appears that evidence from the same genus is required to produce improvements on predictions based on the MILLscore alone.

To test this hypothesis, the two component system interactions of a *Clostridia* species are predicted using the known interactions in another *Clostridia* species. To allow direct comparison with the above analysis, the two component system interactions of *C acetobutylicum* are predicted using the cognate pairs of *C botulinum* as the observed evidence PPIs (Figure 5.8).



Figure 5.7: Predictions of two component PPIs in *C. acetobutylicum* using the MILLscore interaction tree model and the cognates in *B. subtilis* as input.

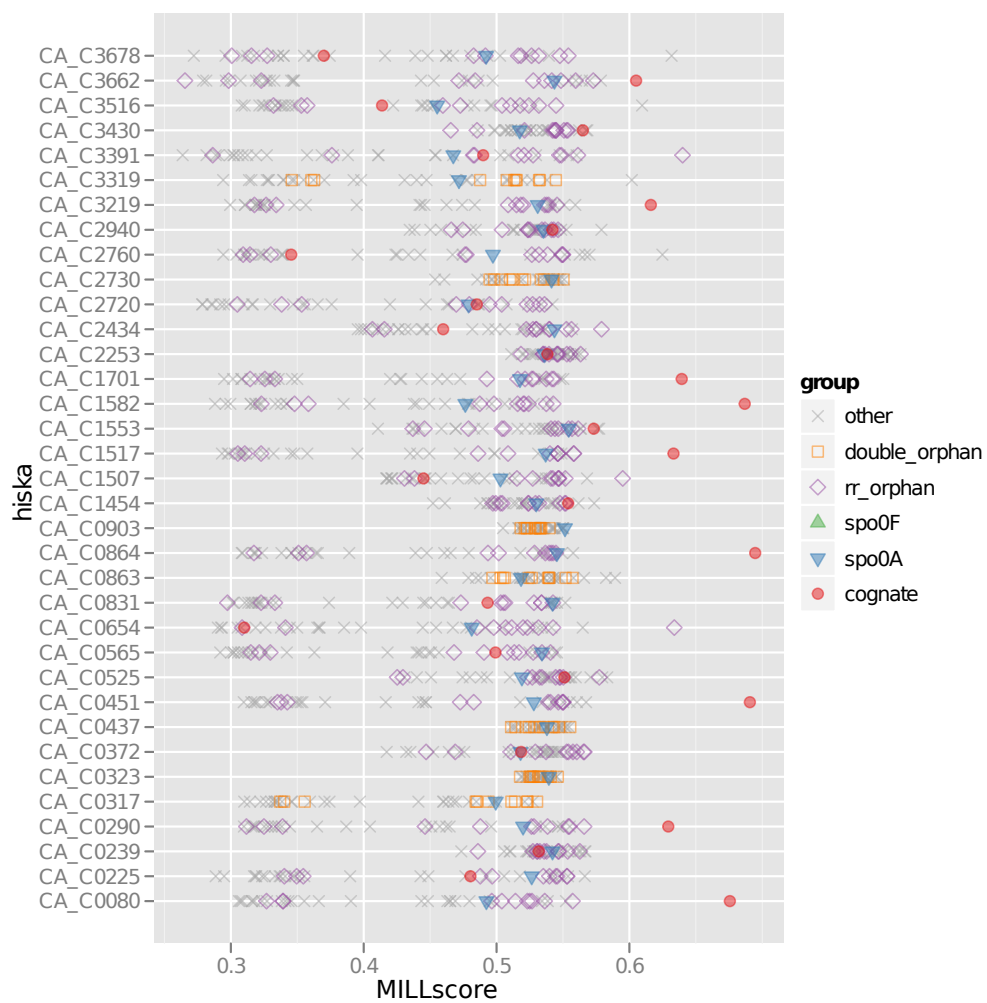


Figure 5.8: Predictions of two component PPIs in *C. acetobutylicum* using the MILLscore interaction tree model and the cognates in *C. botulinum* as input.

	Top 1 predictions	Top 3 predictions	Top 5 predictions	Avg Z-score
MILLScores	11 of 27	16 of 27	18 of 27	1.45
MILLScores + interaction tree (<i>B subtilis</i>)	3 of 27	9 of 27	10 of 27	0.82
MILLScores + interaction tree (<i>C botulinum</i>)	9 of 27	12 of 27	12 of 27	1.15

Table 5.4: A comparison of three methods of predicting the two component system interactions of *C acetobutylicum*: Using the MILLScore alone, using the MILLScore interaction tree model with the PPIs from *B subtilis* as input and using the MILLScore interaction with the PPIs from *C botulinum* as input.

These predictions can be compared to the previous predictions in *C acetobutylicum* (Table 5.4). It is clear that including evidence from the more closely related *C botulinum*, improves the interaction tree predictions over using evidence from *B subtilis*. This is consistent with the findings in Chapter 4; including PPI evidence from closely related species improves predictions from the interaction tree. However, in this situation the interaction tree has not improved predictions beyond using the raw MILLScores, even when including the closer evidence. This is in contrast to *B subtilis*, in which the interaction tree increased the number of “Top 1” predictions. One possible explanation is the different origin of two component proteins in *Bacilli* and *Clostridia*. A higher proportion of the HK/RR proteins in *Clostridia* are the result of horizontal transfer [140]. The process of horizontal gene transfer is not modelled in the current implementation of the interaction tree model. This could mean that this approach is less suitable in the

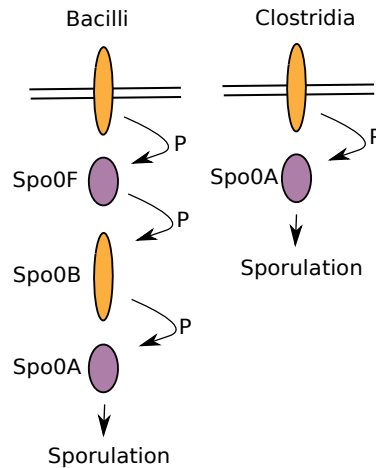


Figure 5.9: A cartoon representation of the two component systems responsible for sporulation onset in Bacteria. In the *Bacilli* a phosphorelay is used (right) and in *Clostridia* a simpler, canonical pathway of two proteins is found.

Clostridia, producing inferior predictions compared to the *Bacilli*.

5.3.3 Predicting ancestral PPIs

The interaction tree can also be used to predict ancestral PPIs. As described in the last chapter, sporulation onset is controlled by two component systems in both *Bacillus* and *Clostridia*. In both types of Bacteria the Spo0A RR is responsible for initiating sporulation however in *Clostridia* Spo0A is phosphorylated directly by the sensory HKs, in *Bacillus* the sensor HKs phosphorylate Spo0F, with Spo0A being activated downstream (Figure 5.9). One question here then is this: what was the ancestral mode of sporulation activation? Did the sporulation HKs interact with Spo0A directly in the ancestral species? or did the sporulation HKs interact with Spo0F?

To test these competing hypotheses, firstly the RR phylogeny is used to iden-

tify putative ancestral Spo0A and ancestral Spo0F (Figure 5.10). The known sporulation interactions from *B subtilis* and the experimentally verified sporulation interactions from *C acetobutylicum* [135] are used as evidence input (Table 5.5).

Species	HK	RR
<i>B subtilis</i>	BSU31450	BSU37130 (Spo0F)
<i>B subtilis</i>	BSU14490	BSU37130 (Spo0F)
<i>B subtilis</i>	BSU13990	BSU37130 (Spo0F)
<i>B subtilis</i>	BSU13660	BSU37130 (Spo0F)
<i>B subtilis</i>	BSU13530	BSU37130 (Spo0F)
<i>C acetobutylicum</i>	CA_C0323	CA_C2071 (Spo0A)
<i>C acetobutylicum</i>	CA_C0903	CA_C2071 (Spo0A)
<i>C acetobutylicum</i>	CA_C3319	CA_C2071 (Spo0A)

Table 5.5: The known sporulation PPIs used as input to the interaction tree in inferring the ancestral sporulation interactions. This set includes the 5 known sporulation kinase interactions with Spo0F (BSU37130) from *B subtilis* and the three experimentally verified kinase-Spo0A (CA_C2071) interactions from [135]

The HK and RR phylogenies are once again pruned to only include *B subtilis* and *C acetobutylicum*. The interaction tree is then used to infer probability of interaction with the ancestral Spo0A (aSpo0A) and ancestral Spo0F (aSpo0F) for each kinase present in the ancestral species, at the point of speciation. The number of ancestral kinases with aSpo0A/aSpo0F as their highest scoring partner can then be counted (Table 5.6).

Taking just the kinases with aSpo0A or aSpo0F as their top predicted interactor, only aSpo0A interacting kinases are predicted (Figure 5.11). The predicted

RR

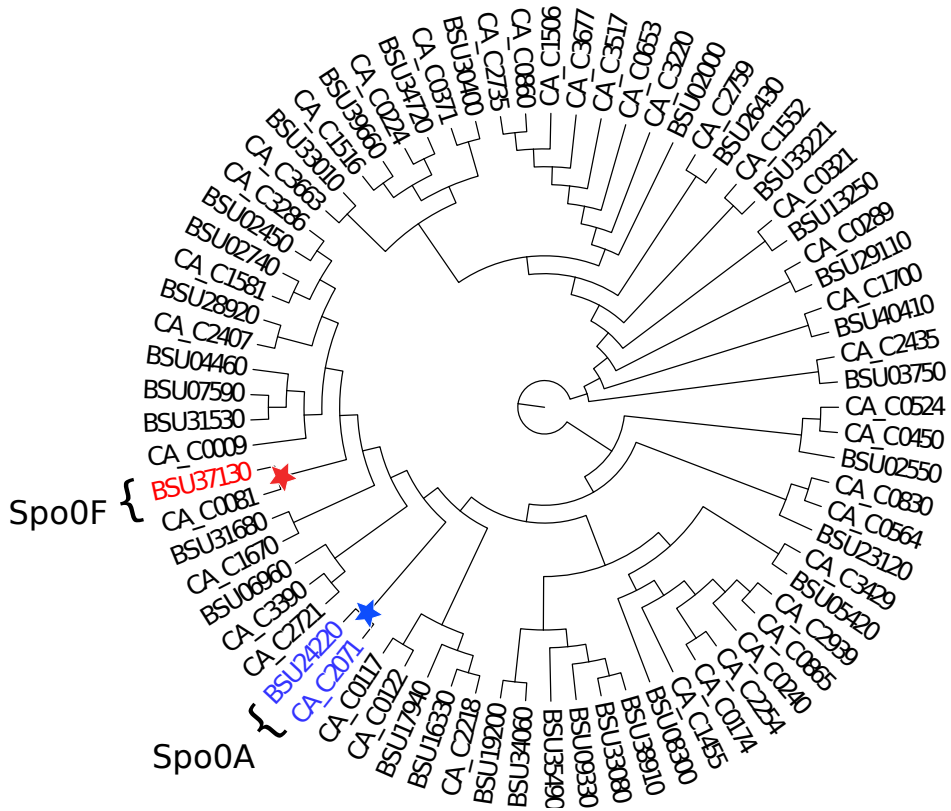


Figure 5.10: Identifying ancestral Spo0A and Spo0F in the *Clostridia/Bacilli* ancestor. The figure shows the RR phylogeny, restricted to *B subtilis* and *C acetobutylicum*. Spo0A in both species is highlighted in blue, the ancestral Spo0A (aSpo0A) is identified as the MRCA of these proteins, present in the ancestor (blue star). Spo0F from *B subtilis* is highlighted in red, aSpo0F is defined as the ancestor of this protein, present at the point of speciation in the ancestral species (red star).

	Top 1 predictions	Top 3 predictions	Top 5 predictions
aSpo0A	6	8	13
aSpo0F	0	1	1

Table 5.6: The number of kinases in the *B subtilis*/*C acetobutylicum* ancestor with aSpo0A or aSpo0F in their top 1/3/5 most likely interaction partners, according to the interaction tree model

interactors of aSpo0A include the ancestor of CA_C0323/CA_C0903. These two *C acetobutylicum* sporulation proteins are identified as recent duplicates in the phylogeny and it appears that their ancestor also interacted directly with Spo0A. The CA_C3319 kinase forms a clade with the five sporulation kinases from *B subtilis*. According to the phylogeny, all 6 of these proteins descend from one ancestral kinase in the ancestor species. Interestingly, this ancestral kinase has aSpo0A as its most probable RR partner. The probability that this kinase interacts with aSpo0A is higher than that with aSpo0F (0.79 vs 0.59). If the threshold is relaxed to include top 3 and top 5 most probable RR interactors, then aSpo0F is predicted as an interaction partner of this ancestral kinase (however the interaction with aSpo0A obviously remains more probable).

This analysis suggests several things concerning the evolution of sporulation activation. Firstly, the ancestor of *Bacillus* and *Clostridia* used the simpler method of sporulation activation: direct interaction with Spo0A. Secondly, the *B subtilis* sporulation kinases are all descended from one ancestral kinase that interacted directly with Spo0A. At some point in evolution, these descendant kinases in *B subtilis* developed the ability to interact with Spo0F.

HK

Key

interacts with Spo0F
interacts with Spo0A

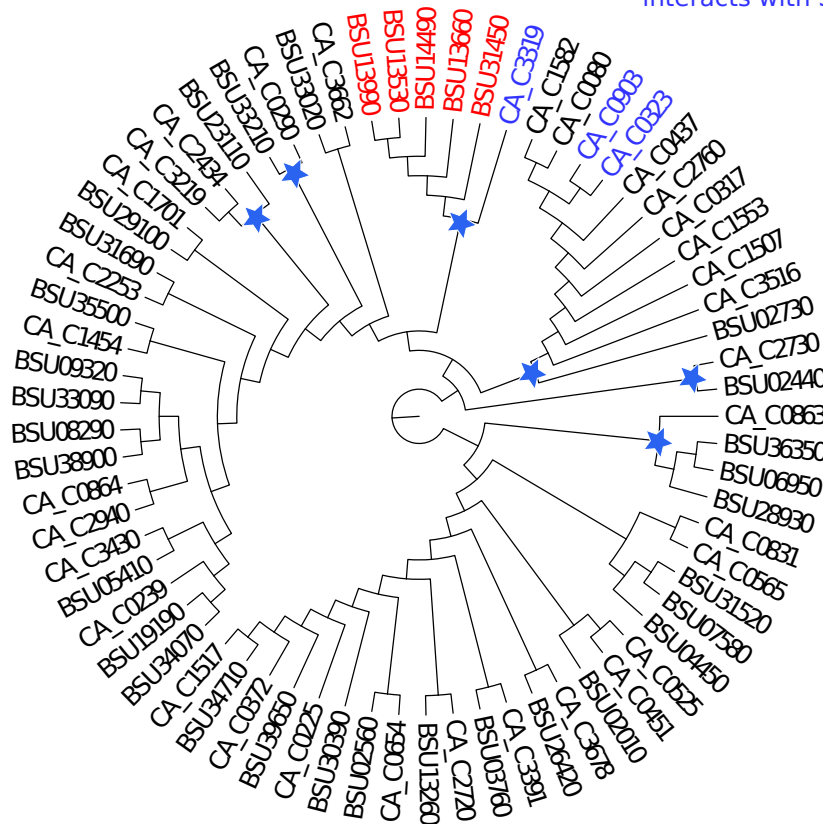


Figure 5.11: The figure shows the HK tree used in the ancestral reconstruction. The *B subtilis* Spo0F interacting kinases are highlighted in red, the *C acetobutylicum* Spo0A interacting kinases in blue. Ancestral kinases with aSpo0A as their most likely interaction partner, according to the interaction tree model, are highlighted with blue stars. No ancestral kinases had aSpo0F as their most likely interaction partner.

5.4 Discussion

In this Chapter an interaction tree model of transient PPI evolution based on the MILLscore was defined. Unlike many previous such models, this approach relates evolutionary sequence change directly to probability of rewiring events. This allows probabilities of such events to be assigned independently to pairs of proteins. As mentioned in the introduction to this chapter, such an approach is likely to allow a finer level of detail in modelling PPI evolution.

The first application of this model was in updating predictions of two component system interactions between existing proteins. In the previous chapter, good predictions were obtained for two component system interactions between existing proteins although there were cases in which prediction was failing. The interaction tree offers one solution to this prediction problem: by allowing known PPIs to be included as observed in the interaction tree, this knowledge can be incorporated in to predictions, via the phylogeny. Such an approach was shown to clearly improve predictions of interaction in some subset of two component systems in *B subtilis*. In some cases the interaction tree approach worsens the predictions and this seems to be the result of the phylogeny being misleading in some way. It is hard to say conclusively if this is a failing of the model or a failure of the phylogeny reconstruction. One possible hypothesis is that the presence of horizontal transfer (not taken in to account by the method of phylogeny reconstruction) is leading to incorrect phylogeny and thus incorrect predictions.

In support of this hypothesis, the interaction tree approach performs poorly in predicting existing PPIs in *Clostridia*. Analysis by [140] shows that *Clostridia* have a higher proportion of two component proteins resulting from horizontal transfer. This would result in a more incorrect phylogeny as these events are not included and so the tree structure misleads the interaction tree algorithm to make incorrect

predictions.

The second application of the MILLscore interaction tree model was in predicting ancestral two component PPIs. Of interest here are the two component system PPIs responsible for sporulation initiation. The model was used to infer the PPIs responsible for this process in the ancestor of *Bacilli* and *Clostridia*. The model predicts that the ancestor used the simpler method: the sporulation kinases interacted with Spo0A directly with the more complicated activation via Spo0F being evolved subsequently in *Bacilli* alone. It is worth noting that such a switch to the complicated pathway need not happen in one step. According to the predictions, several ancestral kinases existed that interacted with Spo0A directly. It is possible that the majority of these kinases remained responsible for sporulation whilst one of them evolved the ability to interact with Spo0F, with the subsequent evolution of the phosphorelay pathway. Once the phosphorelay pathway was established, the Spo0F interacting kinases could duplicate and diversify to respond to a mix of different external stimuli (producing the five sporulation kinases of *B subtilis*). The other canonical Spo0A interacting kinases could then be lost from the organism, at no point interrupting the ability to sporulate and allowing the Bacteria to adapt to a changing environment at each step.

5.5 Conclusion

This chapter has demonstrated that the interaction tree approach of Chapters 2-3 can be extended to apply to transient PPIs. Using a model of PPI rewiring based on the MILLscore, the interaction tree can improve predictions of PPI between existing proteins. However, there are caveats to this application. Firstly, some nearby PPI evidence is required to make good predictions using this model. Secondly, systems in which horizontal transfers are known to exist should perhaps

be avoided until the model can incorporate such events. The interaction tree has also been applied to infer the history of PPIs responsible for sporulation onset in bacteria. These results seem to suggest that the more complicated phosphorelay approaches to sporulation onset used by some species, evolved from an earlier, simpler system, similar to that found in the *Clostridia*.

Chapter 6

Discussion and conclusions

This thesis aimed to develop and apply a method for modelling the evolution of protein-protein interactions. PPIs are a major link in the conversion of genotype to phenotype (Figure 6.1) and so understanding the evolution of PPIs is to understand the evolution of phenotype. This understanding provides insight in to how living organisms adapt to their environment. The interaction tree models developed in this thesis can aid in understanding the PPI evolutionary process; after reconstructing a probable history of PPI evolution, various characteristics of this history can be measured. The key is in developing a model that can accurately reproduce the evolutionary history.

6.1 Advantages of the approach

As set out in the introductory Chapter, there are several limitations to existing approaches to modelling PPI evolution, specifically there is one major drawback to previously described interaction tree models. These models assume that after gene duplication events, there is a constant probability of a PPI rewiring event between

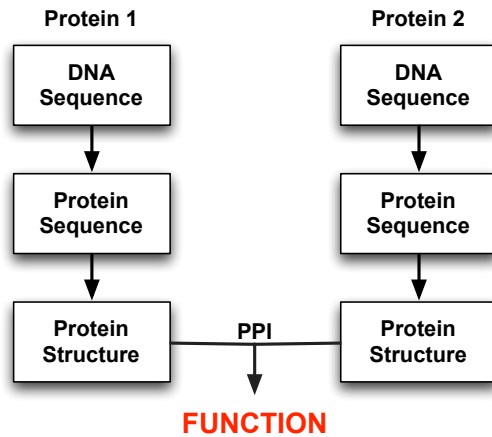


Figure 6.1: The role of protein-protein interactions in producing phenotype. DNA sequence (genotype) is translated to protein. The protein sequence determines the folded 3D structure of the protein. The structure of a protein determines its interaction partner and protein interactions produce functions (phenotype) within the cell.

any pair of proteins. This probability can be estimated in various ways and the resulting models have been used to successfully model PIN evolution and predict protein complexes. In [87] such an interaction tree model was used to predict protein complexes within PINs, the resulting complexes were almost completely connected subgraphs (this is far from the true, sparse pattern of interactions). Whilst this is suitable for the task of predicting complex membership, it highlights the drawback of the simple model; it is hard to model the rewiring events within a complex (or some similar closed system of PPIs). As all proteins are treated identically by the model, there is no distinguishing the pattern of gain and loss at this level of detail.

Obviously the proteins modelled are different and what makes them different is their differing sequences. A model that defines a probability of rewiring distinctly to a pair of proteins must do so based on the sequences of the proteins. Such a model would represent a mapping between changes in sequences and changes in

interaction specificity. The first attempt at using such an interaction tree model was in [85], defining a probability of interaction rewiring between proteins given the total number of substitutions in both proteins. This thesis aimed to expand this approach, extending the model to apply to a wider range of PPIs than in [85].

6.2 PPI evolution in obligate complexes

The interaction tree approach was first expanded to model PPIs within large obligate protein complexes. Using the proteasome complex as a test system it was first shown that the model of [85] could not predict rewiring within protein complexes. It was hypothesised that this model was too simplistic to capture the behaviour of this more complex system (the model was previously applied to simple coiled-coil interactions of bZip transcription factors). A new model was formulated, linking changes in interaction partner to changes in the SCOTCH score, a simple measure of physicochemical complementarity. The model assumes that the structure of a PPI remains constant during evolution allowing calculation of the change in complementarity between two proteins as their sequences change (the assumption of conserved PPI structure is reasonable given e.g. [141]). Probabilities for rewiring events (i.e. gain or loss of PPI) can then be defined given this change in complementarity, in this case using a large training set of protein complexes. This model has obvious advantages over previous interaction tree approaches, primarily the model is more detailed in that it treats PPIs uniquely, as described above. The disadvantages here are the need of a known structure of the PPI family being modelled and the increased computational time needed, compared to simpler approaches.

This model was then applied to reconstruct the history of rewiring within the proteasome complex, given a set of known proteasome structures. The first

outcome of this is predictions of PPIs amongst the proteasome subunits of species for which no structure has been measured. Based on the validation in Chapter 3, these predictions appear to be reliable, especially when some phylogenetically “nearby” structure is available. The case of prediction in the proteasome may be complicated by the role of chaperone proteins in the formation of the complex. These proteins influence the order in which subunits bind to form the complex and influence the specificity of interaction between the subunits [119] [142]. As such, it is not the sequence of the proteins alone that determine the PPIs and the model may be ignoring an important determinant of interaction. A larger analysis comparing predictions in complexes with/without chaperones would gauge the extent of their detrimental effect on prediction.

The predictions themselves come in the form of probabilities of interaction between each pair of proteins in a species. These can be converted to predictions of interacting pairs by placing a threshold on these probabilities (say all pairs with interaction probability > 0.5 are classified as interacting, for instance). However, the resulting set of PPIs may not form the correct topology of the complex e.g. in forming the rings of the proteasome. One obvious question (and opportunity for further development) is this: can the quaternary structure of a complex be inferred from these probabilities, given some “known” topology for the complex. This is clearer with an example; if it is assumed that a eukaryotic proteasome is formed of four stacked heptameric rings, can a probable structure be inferred from a set of pairwise interaction probabilities? An example of this problem is given in Figure 6.2

The second outcome of the analysis is a predicted history of PPIs in the proteasome. This kind of reconstruction can potentially answer questions of how protein complexes have evolved. Specifically, [79] predict that a complex such as the proteasome is either descended from an ancestral dimer or an ancestral ring

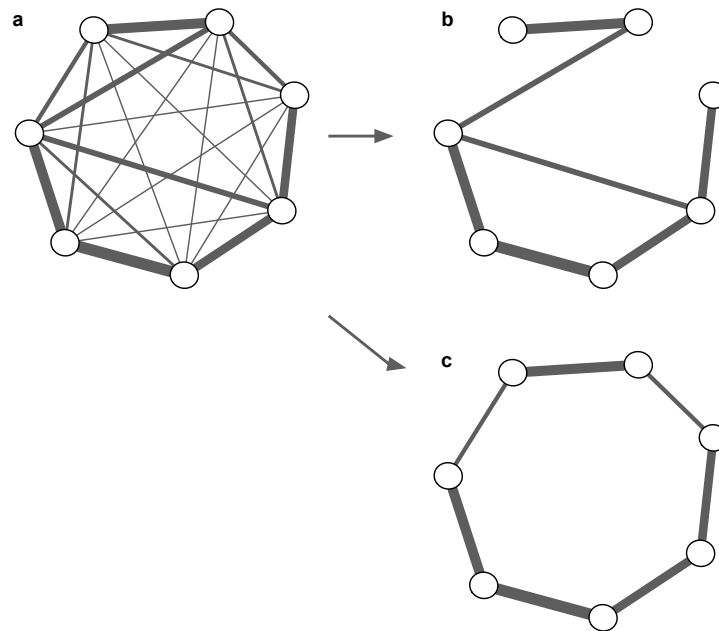


Figure 6.2: Producing structural predictions from the interaction tree predictions. Here the example of predicting an alpha ring of the proteasome is given. The output of the interaction tree algorithm are probabilities of PPI between every pair of proteins (*a*, probabilities shown by thickness of connection). The problem is to infer the structure of the alpha ring structure from these probabilities. Assuming that the alpha ring consists of seven PPIs between seven proteins, an algorithm could predict the seven most probable PPIs as true PPIs (*b*). The problem here is that these do not necessarily form a ring. A better algorithm would constrain the predictions further to produce a ring structure that is supported by the probabilities (*c*).

complex (Figure 6.3). The power of the model presented here is in clustering the PPIs in the complex before undertaking the reconstruction. This allows the model to distinguish between the “dimer” interfaces (clusters 2, 4, 5 & 6 from Chapter 3) and the “ring” interfaces (clusters 1 & 3 from Chapter 3) in the complex. The model can then predict which of these interfaces was the most probable to exist in the ancestral proteasome subunit, at the root of the phylogeny.

In the case of the proteasome, the reconstructed history fails to produce a clear prediction of the very first proteasome interaction and so can not distinguish between the dimer/ring hypotheses. However, the prediction does indicate that most of the interaction interfaces found in existing proteasomes were present in the Last Universal Common Ancestor, challenging previous assumptions that existing 20S proteasomes evolved from a simpler HslV-like complex [74]. This finding is in agreement with a recent study asserting the existence of a related proteasome-like protein, named Anbu. In [113], the authors assert the existence of an ancestral Anbu protein. The results here agree with this assertion and furthermore predict a topology for the ancestral Anbu complex: 4 stacked rings, as in the 20S proteasome. More work needs to be done to confirm this prediction, a starting point would be solving a crystal structure of a present day Anbu complex. The structure of this complex could then be compared to the predicted ancestral complex, specifically the prediction that the complex consists of 4 stacked rings with a smaller (or at least structurally different) alpha ring interface.

6.3 Transient PPI evolution

The methods developed in Chapters 2 and 3 were specifically for application to obligate protein complexes. Of course only a subset of proteins participate in these permanent PPIs; many proteins form transient, impermanent complexes in

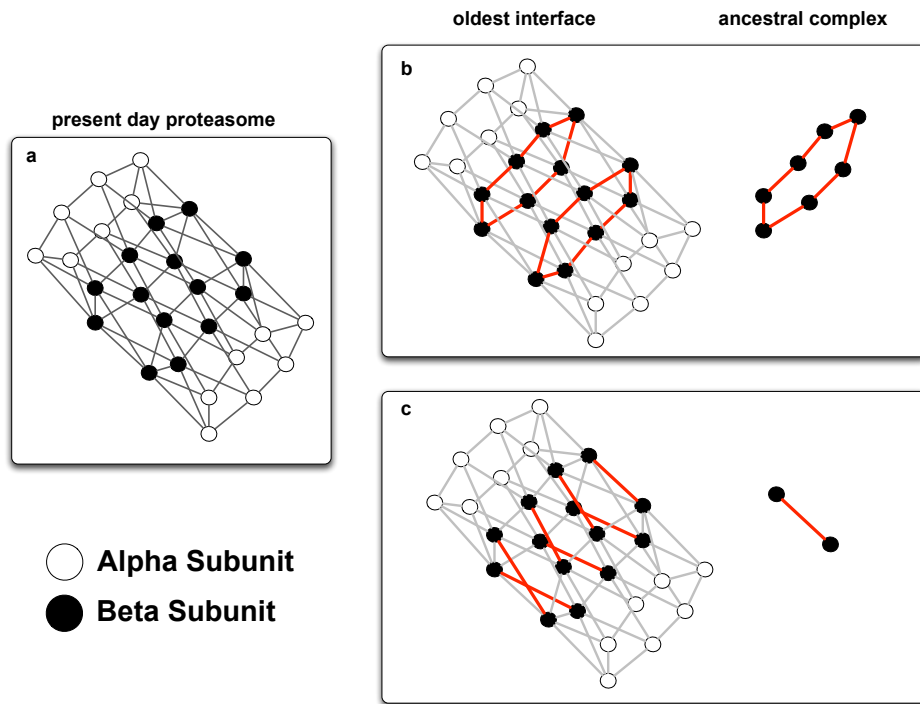


Figure 6.3: The PPIs of the present day yeast proteasome are shown in *a*. The structure of the very first proteasome complex is determined by which PPI in the complex is the oldest. If a ring forming PPI is the oldest, then the ancestral proteasome was a ring *b*. If the oldest PPI was a dimer forming interface (*c*), then the ancestral proteasome complex was a dimer.

order to perform their function. These transient PPIs are known to have different physical and permanent properties to permanent PPIs [143] [8]. Therefore, before these transient PPIs can be modelled, it is first necessary to check that the model is applicable given these differences.

In Chapter 4 it was shown that the D_{com} does not accurately model transient PPIs. This was shown by demonstrating that the SCOTCH measure, that the model is based on, cannot predict transient PPIs. This is likely due to the chemical differences of the protein-protein interface of permanent vs. transient PPIs. For example, it has been observed that permanent PPIs tend to utilise large binding sites composed of hydrophobic residues. This means that these PPIs are often mediated by contacts between hydrophobic residues, which are detected as 'complementary' by the SCOTCH scoring measure. In contrast, transient interactions tend to be mediated by specific hydrogen bonding across the interface [143]. For instance, it might be the case that a hydrogen bond between a serine and a histidine becomes impossible when the serine is replaced by a tyrosine, due to the different shape/orientation of the side chains. However, these two situations would be judged as equally beneficial by the SCOTCH scoring (by pairing a polar with a polar residue). It appears then that the previous model is not detailed enough to capture the specificity of transient PPIs.

Due to this failure of the original model, the MILLscore was developed as a basis for a model of transient PPI evolution. Taken on its own, the MILLscore is a method for predicting if two proteins interact, based on their sequence (although the training of the score uses extra sources of information in the form of large alignments and interaction structures). This method of prediction is similar in approach and accuracy to the recently described method of [126]. The MILLscore has a key advantage in that it is much quicker to train the scoring method (minutes vs days). The large reduction in computational time is due to one difference: the

method of [126] uses a complex algorithm to estimate the important pairings of residues mediating the interaction, the MILLscore observes these pairings directly in a known structure of a homologous PPI. This approach is successful as these pairings are conserved during sequence evolution; the same pairings of residue positions are often important in the PPI to be predicted as in the homologous structure.

An obvious drawback of the MILLscore is the need of a solved, homologous crystal structure. In [144] it was estimated that a quarter of single domain protein families have a solved crystal structure. Obviously, the coverage of PPIs in terms of structures will be less than this, simply given the number of pairings of protein families producing PPIs. Nevertheless, structures do exist for enough interactions to make the MILLscore applicable in a number of cases. For instance, [145] identified a network of 873 yeast proteins, having 1,269 interactions, each of which had a homologous interaction structure in the PDB. It is also worth noting that in cases of interest where no structure is available, the scoring method of [126] could be applied as a replacement of the MILLscore, in order to produce a similar interaction tree model.

It is worth mentioning here that there are a range of other PPI prediction methods that were not considered as a basis for the interaction tree model. In particular there have been several attempts to use machine learning classification to predict PPIs. This has the advantage of not needing a structural example of a homologous PPI. However, it appears that this type of approach is good for determining co-complex membership or functional association but not suitable for determining direct, physical interaction [146]. In this thesis, the aim was to model direct, physical interactions and their change within protein families. It would seem that these general, machine learning approaches are not a good choice of tool in this case.

The MILLscore certainly does prove useful in predicting transient PPI specificities in two component systems. Many HK/RR proteins come in pairs, colocalised on the genome, called cognate pairs. These cognate pairs interact with each other to perform some function (for instance the EnvZ/OmpR cognate pair is responsible for osmoregulation in *E coli* [147]) and the majority of cognate pairs interact only with each other [124]. Given this set of easily determinable interactions, the knowledge to be gained using the MILLscore is in prediction of interactions amongst orphan HKs and RRs. These are HK/RR proteins that are not colocalised with an interaction partner, many having unknown interactions.

In these orphan cases the MILLscore has made several testable predictions. Particular attention was paid to prediction of orphan HKs responsible for sporulation onset via their interaction with the RRs Spo0A or Spo0F. The predictions in bacterial species for which experiments are yet to be performed can be used to guide the search for sporulation kinases in these species. For instance, in *B cereus*, 4 orphan kinases have Spo0F as their highest scoring RR partner (Table 6.1). These 4 orphan kinases would be a good place to start in searching for the HKs responsible for sporulation in this species. This could be investigated further by gene knockdown to ascertain the effect on sporulation (as in [135]) or through direct phosphotransfer assays (as in [131]). Similarly, there are 3 clear predictions of sporulation kinases in *C difficile* and 1 in *C perfringens*. These predictions highlight the role of computational PPI predictions in guiding efficient experimental design.

In Chapter 5, the MILLscore was used as a basis for an interaction tree model of PPI evolution. It was first investigated whether the interaction tree framework could improve the prediction of existing PPIs. It was not clear if this is the case, with some predictions appearing to be improved but others not. One possible reason for the failure of the interaction tree is the failure to model horizontal

Species	Sporulation Kinase	Target RR
<i>B cereus</i>	BC1340	Spo0F
<i>B cereus</i>	BC2619	Spo0F
<i>B cereus</i>	BC4007	Spo0F
<i>B cereus</i>	BC4771	Spo0F
<i>C difficile</i>	CD0576	Spo0A
<i>C difficile</i>	CD1492	Spo0A
<i>C difficile</i>	CD2492	Spo0A
<i>C perfringens</i>	CPF_1523	Spo0A

Table 6.1: The specific predictions of sporulation kinases made by the MILLscore. These predictions consider orphan kinases in each species only. Reported are all orphan kinases in a species for which Spo0A or Spo0F were the highest scoring RR. Refer to the appendix of Chapter 4 for visualisation of these predictions.

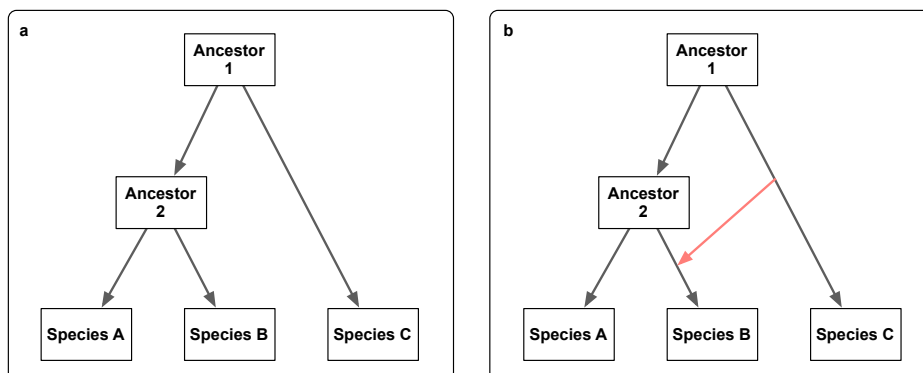


Figure 6.4: Horizontal gene transfer, shown on a species tree. In canonical, vertical evolution (a), genetic material is passed from organisms to their offspring and the relationship of the genes follow the relationship of the species tree. However, it is possible that genetic material is passed between species, in a process called horizontal transfer (b). This is especially common in bacteria and leads to genetic relationships that do not follow a species tree.

transfer (Figure 6.4). Horizontal transfer is especially prevalent in Bacterial species as are analysed here, although further investigation would be needed to show the detrimental effect on prediction of this process. Although not implemented now, horizontal transfer could be included in the model. The latest version of the NOTUNG [34] package allows prediction of horizontal transfer, as do other described reconciliation algorithms [148]. Given a phylogeny including predicted horizontal transfers, the interaction tree construction algorithm could then be modified slightly to incorporate this new type of event.

The interaction tree was also applied to predict the ancestral interactions responsible for sporulation onset. Here, a clear prediction was made: the ancestral species had HKs that phosphorylated Spo0A directly (as in present day *Clostridia*) as opposed to the more complicated phosphorelay system. In order to test this further, a larger phylogenetic analysis could be undertaken to ascertain if components

of the phosphorelay were present in the ancestral species. The phosphorelay consists of a HK, the RR Spo0F, the phosphotransferase Spo0B and the RR Spo0A, with the phosphate group passed in that order. If the ancestral species can be shown to lack Spo0F and/or Spo0B, through a phylogenetic analysis, then this would support the argument that the phosphorelay was not present in the ancestor. The ancestral predictions could also be tested further through recreating the predicted ancestral Spo0A (aSpo0A) and Spo0F (aSpo0F), based on their reconstructed sequences from the analysis. The same could then be done for the predicted ancestral sporulation kinases and a phosphotransfer assay used to confirm the predicted ancestral interactions.

In retrospect, given the versatility of the MILLscore (it is agnostic to the chemical differences of permanent vs transient PPIs for instance) it would be interesting to return to permanent complexes and apply the MILLscore here. Having a generally applicable method that can be applied across several classes of PPI is appealing and the MILLscore is a good candidate for such a model. Given such a generally applicable model of PPI evolution, it could then be applied to study PPI evolution at the network level. This would give an interaction tree model of PIN evolution, similar to [86] for instance, but with a much more detailed predictive model of the PPI rewiring events. Given the need of a structure to compute the MILLscore a starting point would be the dataset of [145]; a PIN with an interaction structure associated with every edge. The MILLscore also requires a training alignment of interacting protein pairs for each edge of the network. A starting point here could be identifying extended sets of paralogs for each node of the PIN (using [149] for instance) and then using publicly available PPI datasets such as BIOGRID [150] or DIP [151] to identify interacting pairs of sequences from these sets, for each edge. These paired sequences could form training alignments, giving a trained MILLscore for each edge of the PIN.

In summary, the interaction tree is a versatile methodological framework, useful for modelling both transient and permanent PPIs. It has two main uses: prediction of existing PPIs and making specific predictions about ancestral PPIs. It is most useful when phylogenetically close PPI evidence exists to guide prediction, although some benefit is seen even when including distant evidence. The MILLscore developed in this thesis provides accurate predictions of specificity for transient interactions which are promising avenues for further investigation. The applicability of the MILLscore in interaction tree modelling can not be conclusively shown here. This may be due to horizontal transfer not modelled during the analysis; adaptation of the method could address this.

References

- [1] B. Alberts, “The cell as a collection of protein machines: preparing the next generation of molecular biologists,” *Cell*, vol. 92, no. 3, pp. 291–4, 1998.
- [2] K. J. Mariani, “Understanding how the replisome works,” *Nature Structural and Molecular Biology*, vol. 15, pp. 125–127, Feb. 2008.
- [3] F. Giorgini and P. J. Muchowski, “Connecting the dots in huntington’s disease with protein interaction networks,” *Genome Biology*, vol. 6, no. 3, p. 210, 2005.
- [4] E. D. Ross, A. Minton, and R. B. Wickner, “Prion domains: sequences, structures and interactions,” *Nature Cell Biology*, vol. 7, no. 11, pp. 1039–1044, 2005.
- [5] S. J. Watowich, L. J. Gross, and R. Josefs, “Analysis of the intermolecular contacts within sickle hemoglobin fibers: effect of site-specific substitutions, fiber pitch, and double-strand disorder,” *Journal of structural biology*, vol. 111, no. 3, pp. 161–179, 1993.
- [6] A. Wittinghofer and H. Waldmann, “Rasa molecular switch involved in tumor formation,” *Angewandte Chemie International Edition*, vol. 39, no. 23, pp. 4192–4214, 2000.

- [7] S. Jones and J. M. Thornton, “Protein-protein interactions: a review of protein dimer structures,” *Progress in biophysics and molecular biology*, vol. 63, no. 1, pp. 31–65, 1995.
- [8] J. Mintseris and Z. Weng, “Structure, function, and evolution of transient and obligate protein-protein interactions,” *Proceedings of the National Academy of Sciences*, 2005.
- [9] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, “Conformational entropy in molecular recognition by proteins,” *Nature*, vol. 448, pp. 325–329, July 2007.
- [10] M. C. Lawrence and P. M. Colman, “Shape complementarity at protein/protein interfaces,” *Journal of molecular biology*, vol. 234, no. 4, pp. 946–50, 1993.
- [11] A. McCoy, V. Epa, and P. Colman, “Electrostatic complementarity at protein/protein interfaces,” *Journal of molecular biology*, vol. 268, no. 2, pp. 570–584, 1997.
- [12] D. Xu, C. J. Tsai, and R. Nussinov, “Hydrogen bonds and salt bridges across protein-protein interfaces,” *Protein engineering*, vol. 10, no. 9, pp. 999–1012, 1997.
- [13] E. M. Schmid and H. T. McMahon, “Integrating molecular and network biology to decode endocytosis,” *Nature*, vol. 448, no. 7156, pp. 883–888, 2007.
- [14] J. D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, and F. P. Roth, “Evidence for

dynamically organized modularity in the yeast protein–protein interaction network,” *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

- [15] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
- [16] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, “Protein-protein interaction networks and biologywhat’s the connection?,” *Nature biotechnology*, vol. 26, no. 1, pp. 69–72, 2008.
- [17] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork, “Dynamic complex formation during the yeast cell cycle,” *Science’s STKE*, vol. 307, no. 5710, p. 724, 2005.
- [18] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, and E. L. L. Sonnhammer, “The pfam protein families database,” *Nucleic acids research*, vol. 32, no. suppl 1, pp. D138–D141, 2004.
- [19] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, “A model of evolutionary change in proteins,” *In Atlas of Protein Sequences and Structure*, vol. 5, pp. 345–352, 1978.
- [20] D. T. Jones, W. R. Taylor, and J. M. Thornton, “The rapid generation of mutation data matrices from protein sequences,” *Computer applications in the biosciences: CABIOS*, vol. 8, no. 3, pp. 275–282, 1992.
- [21] S. Whelan and N. Goldman, “A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach,” *Molecular biology and evolution*, vol. 18, no. 5, pp. 691–699, 2001.

- [22] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, p. 10915, 1992.
- [23] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443–453, Mar. 1970.
- [24] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [25] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–402, 1997.
- [26] O. O'Sullivan, K. Suhre, C. Abergel, D. G. Higgins, and C. Notredame, "3dcoffee: combining protein sequences and structures within multiple sequence alignments," *Journal of molecular biology*, vol. 340, no. 2, pp. 385–395, 2004.
- [27] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.
- [28] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971.

- [29] D. Sankoff, “Minimal mutation trees of sequences,” *SIAM Journal on Applied Mathematics*, pp. 35–42, 1975.
- [30] J. Felsenstein, “Cases in which parsimony or compatibility methods will be positively misleading,” *Systematic Biology*, vol. 27, no. 4, pp. 401–410, 1978.
- [31] B. Rannala and Z. Yang, “Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference,” *Journal of molecular evolution*, vol. 43, no. 3, pp. 304–311, 1996.
- [32] J. P. Huelsenbeck and F. Ronquist, “Mrbayes: Bayesian inference of phylogenetic trees,” *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.
- [33] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda, “Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences,” *Systematic Biology*, vol. 28, no. 2, pp. 132–163, 1979.
- [34] D. Durand, B. V. Halldorsson, and B. Vernot, “A hybrid Micro-Macroevolutionary approach to gene tree reconstruction,” *Journal of Computational Biology*, vol. 13, no. 2, pp. 320–335, 2006.
- [35] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren, “Simultaneous bayesian gene tree reconstruction and reconciliation analysis,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 5714–9, 2009.
- [36] E. V. Koonin, “An apology for orthologs-or brave new memes,” *Genome Biology*, vol. 2, no. 4, p. 1005, 2001.
- [37] W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O’Brien, “Molecular phylogenetics and the origins of placental mammals,” *Nature*, vol. 409, no. 6820, pp. 614–618, 2001.

- [38] I. Seibold and A. J. Helbig, “Evolutionary history of new and old world vultures inferred from nucleotide sequences of the mitochondrial cytochrome b gene,” *Philosophical Transactions of the Royal Society of London.*, vol. 350, no. 1332, pp. 163–178, 1995.
- [39] F. Pazos and A. Valencia, “Similarity of phylogenetic trees as indicator of proteinprotein interaction,” *Protein Engineering*, vol. 14, pp. 609–614, Sept. 2001.
- [40] A. Wagner, “How the global structure of protein interaction networks evolves,” *Proc Biol Sci*, vol. 270, no. 1514, pp. 457–66, 2003.
- [41] C. Shou, N. Bhardwaj, H. Y. K. Lam, K. Yan, P. M. Kim, M. Snyder, and M. B. Gerstein, “Measuring the evolutionary rewiring of biological networks,” *PLoS Comput Biol*, vol. 7, pp. e1001050+, Jan. 2011.
- [42] A. Presser, M. B. Elowitz, M. Kellis, and R. Kishony, “The evolutionary dynamics of the *saccharomyces cerevisiae* protein interaction network after duplication.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 950–954, Jan. 2008.
- [43] I. Tirosh and N. Barkai, “Comparative analysis indicates regulatory neofunctionalization of yeast duplicates.,” *Genome biology*, vol. 8, pp. R50+, Apr. 2007.
- [44] I. Halperin, H. Wolfson, and R. Nussinov, “Correlated mutations: advances and limitations. a study on fusion proteins and on the cohesin-dockerin families,” *Proteins*, vol. 63, no. 4, pp. 832–45, 2006.

- [45] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, “Protein 3D structure computed from evolutionary sequence variation,” *PLoS ONE*, vol. 6, pp. e28766+, Dec. 2011.
- [46] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub, “Rewiring the specificity of Two-Component signal transduction systems,” *Cell*, vol. 133, pp. 1043–1054, June 2008.
- [47] J. Zhang, “Evolution by gene duplication: an update,” *Trends in Ecology and Evolution*, vol. 18, no. 6, pp. 292 – 298, 2003.
- [48] T. A. Gibson and D. S. Goldberg, “Questioning the ubiquity of neofunctionalization,” *PLoS Comput Biol*, vol. 5, pp. e1000252+, Jan. 2009.
- [49] H. J. Cha, M. Byrom, P. E. Mead, A. D. Ellington, J. B. Wallingford, and E. M. Marcotte, “Evolutionarily repurposed networks reveal the well-known antifungal drug thiabendazole to be a novel vascular disrupting agent,” *PLoS Biol*, vol. 10, no. 8, p. e1001379, 2012.
- [50] S. C. Lovell and D. L. Robertson, “An integrated view of molecular co-evolution in protein-protein interactions,” *Molecular Biology and Evolution*, vol. 27, pp. 2567–2575, 2010.
- [51] W. Ali and C. Deane, “Functionally guided alignment of protein interaction networks for module detection,” *Bioinformatics*, 2009.
- [52] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 4569–4574, Apr. 2001.

- [53] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dämpfung, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, “Proteome survey reveals modularity of the yeast cell machinery,” *Nature*, vol. 440, pp. 631–636, Jan. 2006.
- [54] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aa-nensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg, “A protein interaction map of drosophila melanogaster,” *Science*, vol. 302, pp. 1727–1736, Dec. 2003.
- [55] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal, “A map of the interactome network of the metazoan *c. elegans*,” *Science*

(*New York, N.Y.*), vol. 303, pp. 540–543, Jan. 2004.

- [56] N. Przulj, O. Kuchaiev, A. Stevanović, and W. Hayes, “Geometric evolutionary dynamics of protein interaction networks,” *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 178–89, 2010.
- [57] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal, “Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or ”interologs” .,” *Genome research*, vol. 11, pp. 2120–2126, Dec. 2001.
- [58] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, “Conserved pathways within bacteria and yeast as revealed by global protein network alignment,” *Proceedings of the National Academy of Sciences*, vol. 100, pp. 11394–11399, Sept. 2003.
- [59] R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R. M. Karp, “Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data.,” *Journal of computational biology : a journal of computational molecular cell biology*, vol. 12, pp. 835–846, July 2005.
- [60] M. Kalaev, M. Smoot, T. Ideker, and R. Sharan, “Networkblast: comparative analysis of protein networks,” *Bioinformatics (Oxford, England)*, vol. 24, no. 4, pp. 594–6, 2008.
- [61] J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou, “Graemlin: general and robust alignment of multiple large interaction networks,” *Genome Res*, vol. 16, no. 9, pp. 1169–81, 2006.

- [62] J. Dutkowski and J. Tiuryn, “Identification of functional modules from conserved ancestral protein protein interactions,” *Bioinformatics*, vol. 23, pp. i149–i158, 2007.
- [63] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, “Pairwise alignment of protein interaction networks,” *Journal of Computational Biology*, vol. 13, no. 2, pp. 182–99, 2006.
- [64] H. T. T. Phan and M. J. E. Sternberg, “PINALOG: a novel approach to align protein interaction networksimplications for complex detection and function prediction,” *Bioinformatics*, vol. 28, pp. 1239–1245, May 2012.
- [65] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, “Comparative assessment of large-scale data sets of protein-protein interactions,” *Nature*, vol. 417, pp. 399–403, May 2002.
- [66] G. D. Bader and C. W. Hogue, “Analyzing yeast protein-protein interaction data obtained from different sources.,” *Nature biotechnology*, vol. 20, pp. 991–997, Oct. 2002.
- [67] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509–512, Oct. 1999.
- [68] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, “Modeling of protein interaction networks,” *ComPlexUs*, vol. 1, Aug. 2003.
- [69] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, “A comprehensive analysis of

- protein-protein interactions in *saccharomyces cerevisiae*,” *Nature*, vol. 403, pp. 623–627, Feb. 2000.
- [70] R. Sole, R. Pastor-Satorras, E. Smith, and T. Kepler, “A model of large-scale proteome evolution,” *Advances in Complex Systems*, 2002.
- [71] J. Berg, M. Lässig, and A. Wagner, “Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications,” *BMC Evolutionary Biology*, vol. 4, no. 1, p. 51, 2004.
- [72] T. A. Gibson and D. S. Goldberg, “Improving evolutionary models of protein interaction networks,” *Bioinformatics (Oxford, England)*, vol. 27, pp. 376–382, Feb. 2011.
- [73] O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson, “Model criticism based on likelihood-free inference, with an application to protein network evolution,” *Proceedings of the National Academy of Sciences*, vol. 106, pp. 10576–10581, June 2009.
- [74] C. Gille, A. Goede, and C. Schlöetelburg, “A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome,” *Journal of molecular biology*, vol. 326, pp. 1437–1448, 2003.
- [75] N. Chia, I. Cann, and G. J. Olsen, “Evolution of dna replication protein complexes in eukaryotes and archaea,” *PLoS ONE*, vol. 5, no. 6, p. e10866, 2010.
- [76] T. Gabaldón, D. Rainey, and M. Huynen, “Tracing the evolution of a large protein complex in the eukaryotes, nadh: ubiquinone . . .,” *Journal of molecular biology*, 2005.

- [77] D. Goodsell and A. Olson, “Structural symmetry and protein function,” *Annual review of biophysics and biomolecular structure*, 2000.
- [78] Z. Lin and H. S. Rye, “GroEL-mediated protein folding: Making the impossible, possible,” *Critical Reviews in Biochemistry and Molecular Biology*, vol. 41, pp. 211–239, Jan. 2006.
- [79] E. D. Levy, E. B. E. C. V. Robinson, and S. A. Teichmann, “Assembly reflects evolution of protein complexes,” *Nature*, vol. 453, no. 7199, pp. 1262–1265, 2008.
- [80] J. Pereira-Leal, E. Levy, C. Kamp, and S. Teichmann, “Evolution of protein complexes by duplication of homomeric interactions,” *Genome Biology*, vol. 8, no. 4, p. R51, 2007.
- [81] F. Pazos and A. Valencia, “Protein co-evolution, co-adaptation and interactions,” *The EMBO journal*, 2008.
- [82] M. G. Kann, B. A. Shoemaker, A. R. Panchenko, and T. M. Przytycka, “Correlated evolution of interacting proteins: looking behind the mirrortree,” *Journal of molecular biology*, vol. 385, no. 1, pp. 91–8, 2009.
- [83] L. Hakes, S. Lovell, S. G. Oliver, and D. L. Robertson, “Specificity in protein interactions and its relationship with sequence diversity and coevolution,” *Proc Natl Acad Sci USA*, vol. 104, no. 19, pp. 7999–8004, 2007.
- [84] R. Patro, E. Sefer, J. Malin, G. Marcais, S. Navlakha, and C. Kingsford, “Parsimonious reconstruction of network evolution,” *Algorithms for Molecular Biology*, vol. 7, no. 1, pp. 25+, 2012.
- [85] J. Pinney, G. Amoutzias, M. Rattray, and D. Robertson, “Reconstruction of ancestral protein interaction networks for the bzip transcription factors,”

Proceedings of the National Academy of Sciences, vol. 104, no. 51, p. 20449, 2007.

- [86] T. A. Gibson and D. S. Goldberg, “Reverse engineering the evolution of protein interaction networks,” *Pacific Symposium on Biocomputing*, vol. 14, pp. 190–202, 2009.
- [87] J. Dutkowski and J. Tiuryn, “Phylogeny-guided interaction mapping in seven eukaryotes,” *BMC bioinformatics*, vol. 10, p. 393, 2009.
- [88] O. Ratmann, C. Wiuf, and J. W. Pinney, “From evidence to inference: probing the evolution of protein interaction networks,” *Adv. Online Pub. HFSP J.*, vol. 1, no. 1, p. 114, 2009.
- [89] N. Marianayagam, M. Sunde, and J. Matthews, “The power of two: protein dimerization in biology,” *Trends in Biochemical Sciences*, 2004.
- [90] I. Ispolatov, A. Yuryev, I. Mazo, and S. Maslov, “Binding properties and evolution of homodimers in protein-protein interaction networks,” *Nucleic acids research*, 2005.
- [91] J. B. Pereira-Leal, E. D. Levy, and S. A. Teichmann, “The origins and evolution of functional modules: lessons from protein complexes,” *Philos Trans R Soc Lond, B, Biol Sci*, vol. 361, no. 1467, pp. 507–17, 2006.
- [92] Z. Itzhaki, E. Akiva, Y. Altuvia, and H. Margalit, “Evolutionary conservation of domain-domain interactions,” *Genome Biology*, 2006.
- [93] E. Robert, “MUSCLE: a multiple sequence alignment method with reduced time and space complexity,” *BMC Bioinformatics*, vol. 5, pp. 113+, Aug. 2004.

- [94] J. Pearl, “Probabilistic reasoning in intelligent systems: networks of plausible inference,” *Morgan Kaufmann*, 1988.
- [95] D. T. Jones, W. R. Taylor, and J. M. Thornton, “The rapid generation of mutation data matrices from protein sequences,” *Comput Appl Biosci*, vol. 8, no. 3, pp. 275–82, 1992.
- [96] J. Felsenstein, “Phylip-phylogeny inference package (version 3.2),” *Cladistics*, 1989.
- [97] D. Juan, F. Pazos, and A. Valencia, “High-confidence prediction of global interactomes based on genome-wide coevolutionary networks,” *Proc Natl Acad Sci USA*, vol. 105, no. 3, pp. 934–9, 2008.
- [98] H. Madaoui and R. Guerois, “Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking,” *Proc Natl Acad Sci USA*, vol. 105, no. 22, pp. 7708–13, 2008.
- [99] E. Levy, J. Pereira-Leal, C. Chothia, and S. Teichmann, “3d complex: a structural classification of protein complexes,” *PLoS Comput Biol*, vol. 2, no. 11, p. e155, 2006.
- [100] E. Krissinel and K. Henrick, “Inference of macromolecular assemblies from crystalline state,” *Journal of molecular biology*, vol. 372, no. 3, pp. 774–97, 2007.
- [101] S. J. de Vries, A. D. J. van Dijk, and A. M. J. J. Bonvin, “Whisky: what information does surface conservation yield? application to data-driven docking,” *Proteins*, vol. 63, no. 3, pp. 479–89, 2006.

- [102] K. Tanaka, “The proteasome: overview of structure and functions,” *Proc Jpn Acad, Ser B, Phys Biol Sci*, vol. 85, no. 1, pp. 12–36, 2009.
- [103] M. Groll, L. Ditzel, J. Löwe, D. Stock, M. Bochtler, H. D. Bartunik, and R. Huber, “Structure of 20s proteasome from yeast at 2.4 a resolution,” *Nature*, vol. 386, no. 6624, pp. 463–71, 1997.
- [104] J. Sussman, D. Lin, J. Jiang, N. Manning, J. Prilusky, O. Ritter, and E. Abola, “Protein data bank (pdb): Database of three-dimensional structural information of biological macromolecules,” *Acta Crystallogr D*, vol. 54, pp. 1078–1084, 1998.
- [105] E. Levy, E. Erba, and C. Robinson, “Assembly reflects evolution of protein complexes,” *Nature*, 2008.
- [106] M. Unno, T. Mizushima, Y. Morimoto, Y. Tomisugi, K. Tanaka, N. Yasuoka, and T. Tsukihara, “The structure of the mammalian 20S proteasome at 2.75 a resolution,” *Structure (Camb)*, vol. 10, pp. 609–618, May 2002.
- [107] J. Gough, K. Karplus, R. Hughey, and C. Chothia, “Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure,” *Journal of Molecular Biology*, vol. 313, no. 4, pp. 903 – 919, 2001.
- [108] B. Mordm?ller, R. Fendel, A. Kreidenweiss, C. Gille, R. Hurwitz, W. G. Metzger, J. F. Kun, T. Lamkemeyer, A. Nordheim, and P. G. Kremsner, “Plasmodia express two threonine-peptidase complexes during asexual development,” *Molecular and Biochemical Parasitology*, vol. 148, no. 1, pp. 79 – 85, 2006.

- [109] Letunic and P. Bork, “Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation.,” *Bioinformatics (Oxford, England)*, vol. 23, pp. 127–128, Jan. 2007.
- [110] Z. Yang, “PAML 4: Phylogenetic analysis by maximum likelihood,” *Molecular Biology and Evolution*, vol. 24, pp. 1586–1591, Aug. 2007.
- [111] P. D. Williams, D. D. Pollock, B. P. Blackburne, and R. A. Goldstein, “Assessing the accuracy of ancestral protein reconstruction methods,” *PLoS Comput Biol*, vol. 2, no. 6, p. e69, 2006.
- [112] F. Pazos, J. A. G. Ranea, D. Juan, and M. J. E. Sternberg, “Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome,” *Journal of molecular biology*, vol. 352, no. 4, pp. 1002–15, 2005.
- [113] Valas, Ruben, Bourne, and Philip, “Rethinking proteasome evolution: Two novel bacterial proteasomes,” *Journal of Molecular Evolution*, vol. 66, pp. 494–504, May 2008.
- [114] Y. D. Kwon, I. Nagy, P. D. Adams, W. Baumeister, and B. K. Jap, “Crystal structures of the rhodococcus proteasome with and without its pro-peptides: Implications for the role of the pro-peptide in proteasome assembly,” *Journal of Molecular Biology*, vol. 335, no. 1, pp. 233 – 245, 2004.
- [115] S. J. Kaczowka and J. A. Maupin-Furlow, “Subunit topology of two 20s proteasomes from *haloferax volcanii*,” *J. Bacteriol.*, vol. 185, no. 1, pp. 165–74, 2003.
- [116] L. S. Madding, J. K. Michel, K. R. Shockley, S. B. Connors, K. L. Epting, M. R. Johnson, and R. M. Kelly, “Role of the alpha-1 subunit in the func-

- tion and stability of the 20s proteasome in the hyperthermophilic archaeon *pyrococcus furiosus*,” *J. Bacteriol.*, vol. 189, no. 2, pp. 583–590, 2007.
- [117] M. J. Harms and J. W. Thornton, “Analyzing protein structure and function using ancestral gene reconstruction,” *Current Opinion in Structural Biology*, vol. 20, no. 3, pp. 360 – 366, 2010.
- [118] A. De Maio, “Heat shock proteins: facts, thoughts, and dreams.,” *Shock*, vol. 11, no. 1, pp. 1–12, 1999.
- [119] H. C. Besche, A. Peth, and A. L. Goldberg, “Getting to first base in proteasome assembly.,” *Cell*, vol. 138, pp. 25–28, July 2009.
- [120] S. Jones and J. M. Thornton, “Principles of protein-protein interactions,” *Proc Natl Acad Sci USA*, vol. 93, no. 1, pp. 13–20, 1996.
- [121] I. Nooren and J. M. Thornton, “Structural characterisation and functional significance of transient protein-protein interactions,” *Journal of molecular biology*, vol. 325, no. 5, pp. 991–1018, 2003.
- [122] A. H. West and A. M. Stock, “Histidine kinases and response regulator proteins in two-component signaling systems,” *Trends in Biochemical Sciences*, vol. 26, no. 6, pp. 369 – 376, 2001.
- [123] L. E. Ulrich and I. B. Zhulin, “The MiST2 database: a comprehensive genomics resource on microbial signal transduction,” *Nucl. Acids Res.*, pp. gkp940+, Nov. 2009.
- [124] E. J. Capra, B. S. Perchuk, J. M. Skerker, and M. T. Laub, “Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families,” *Cell*, vol. 150, pp. 222–232, July 2012.

- [125] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub, “Rewiring the specificity of Two-Component signal transduction systems,” *Cell*, vol. 133, pp. 1043–1054, June 2008.
- [126] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt, “Dissecting the specificity of Protein-Protein interaction in bacterial Two-Component signaling: Orphans and crosstalks,” *PLoS ONE*, vol. 6, pp. e19729+, May 2011.
- [127] G. Moont, H. A. Gabb, and M. J. E. Sternberg, “Use of pair potentials across protein interfaces in screening predicted docked complexes,” *Proteins: Structure, Function, and Genetics*, vol. 35, no. 3, pp. 364–373, 1999.
- [128] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Y. Shen, U. Pieper, and A. Sali, “Comparative protein structure modeling using MODELLER.,” *Current protocols in protein science*, vol. Chapter 2, Nov. 2007.
- [129] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serano, “The FoldX web server: an online force field,” *Nucleic Acids Research*, vol. 33, pp. W382–W388, July 2005.
- [130] S. D. Dunn, L. M. Wahl, and G. B. Gloor, “Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction,” *Bioinformatics*, vol. 24, pp. 333–340, Feb. 2008.
- [131] E. J. Capra, B. S. Perchuk, E. A. Lubin, O. Ashenberg, J. M. Skerker, and M. T. Laub, “Systematic dissection and Trajectory-Scanning mutagenesis of the molecular interface that ensures specificity of Two-Component signaling pathways,” *PLoS Genet*, vol. 6, pp. e1001220+, Nov. 2010.

- [132] D. Burbulys, K. A. Trach, and J. A. Hoch, "Initiation of sporulation in *b. subtilis* is controlled by a multicomponent phosphorelay.," *Cell*, vol. 64, pp. 545–552, Feb. 1991.
- [133] M. Jiang, W. Shao, M. Perego, and J. A. Hoch, "Multiple histidine kinases regulate entry into stationary phase and sporulation in *bacillus subtilis*," *Molecular microbiology*, vol. 38, pp. 535–542, Nov. 2000.
- [134] K. Stephenson and J. Hoch, "Evolution of signalling in the sporulation phosphorelay.," *Mol Microbiol*, vol. 46, no. 2, pp. 297–304, 2002.
- [135] E. Steiner, A. Dago, D. Young, J. Heap, N. Minton, J. Hoch, and M. Young, "Multiple orphan histidine kinases interact directly with *spo0a* to control the initiation of endospore formation in *clostridium acetobutylicum*," *Mol Microbiol*, vol. 80, no. 3, pp. 641–54, 2011.
- [136] R. Brunsing, C. La Clair, S. Tang, C. Chiang, L. Hancock, M. Perego, and J. Hoch, "Characterization of sporulation histidine kinases of *bacillus anthracis*," *J Bacteriol*, vol. 187, no. 20, pp. 6972–81, 2005.
- [137] O. Penn, E. Privman, G. Landan, D. Graur, and T. Pupko, "An alignment confidence score capturing robustness to guide tree uncertainty," *Molecular Biology and Evolution*, vol. 27, pp. 1759–1767, Aug. 2010.
- [138] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, and E. L. L. Sonnhammer, "The pfam protein families database," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D138–D141, 2004.

- [139] S. Guindon and O. Gascuel, “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.,” *Systematic biology*, vol. 52, pp. 696–704, Oct. 2003.
- [140] E. Alm, K. Huang, and A. Arkin, “The evolution of Two-Component systems in bacteria reveals different strategies for niche adaptation,” *PLoS Comput Biol*, vol. 2, pp. e143+, Nov. 2006.
- [141] Q. C. Zhang, D. Petrey, R. Norel, and B. H. Honig, “Protein interface conservation across structure space,” *Proceedings of the National Academy of Sciences*, vol. 107, pp. 10896–10901, June 2010.
- [142] S. Murata, H. Yashiroda, and K. Tanaka, “Molecular mechanisms of proteasome assembly.,” *Nature reviews. Molecular cell biology*, vol. 10, pp. 104–115, Feb. 2009.
- [143] J. Thornton, “Diversity of protein–protein interactions,” *The EMBO journal*, 2003.
- [144] M. Levitt, “Nature of the protein universe,” *Proceedings of the National Academy of Sciences*, vol. 106, pp. 11079–11084, July 2009.
- [145] P. Kim, L. Lu, Y. Xia, and M. Gerstein, “Relating three-dimensional structures to protein networks provides evolutionary insights.,” *Science (New York)*, vol. 314, no. 5, pp. 1938–41, 2006-12-22 00:00:00.0.
- [146] Y. Qi, Z. Bar-joseph, and J. Klein-seetharaman, “Evaluation of different biological data and computational classification methods for use in protein interaction prediction,” *Proteins*, vol. 63, pp. 490–500, 2006.
- [147] S. J. Cai and M. Inouye, “EnvZ-OmpR interaction and osmoregulation in escherichia coli,” *J. Biol. Chem.*, vol. 277, pp. 24155–24161, June 2002.

- [148] M. S. Bansal, E. J. Alm, and M. Kellis, “Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss,” *Bioinformatics*, vol. 28, pp. i283–i291, June 2012.
- [149] S. Heinicke, M. S. Livstone, C. Lu, R. Oughtred, F. Kang, S. V. Angiuoli, O. White, D. Botstein, and K. Dolinski, “The princeton protein orthology database (P-POD): A comparative genomics analysis tool for biologists,” *PLoS ONE*, vol. 2, pp. e766+, Aug. 2007.
- [150] C. Stark, B.-J. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets.,” *Nucleic acids research*, vol. 34, pp. D535–D539, Jan. 2006.
- [151] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, “DIP: the database of interacting proteins,” *Nucleic Acids Research*, vol. 28, pp. 289–291, Jan. 2000.