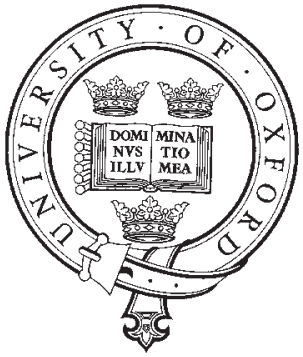
University of Oxford



A Bayesian approach to predicting protein-protein interactions

by

Pao-Yang Chen

St Anne's College

A research proposal submitted in partial fulfilment of the requirements for the transfer to D.Phil status.

Department of Statistics, 1 South Parks Road,
Oxford OX1 3TG

December 2005

This is my own work (except where otherwise indicated)

Candidate: Pao-Yang Chen

Signed:.....

Date:.....

August 7, 2006

Abstract

To predict novel protein-protein interactions, the protein interaction networks are upcast by assigning SCOP structural classifications to as many of the interacting proteins as possible. Two probabilistic models, the frequency-based approach and the odds ratio-based approach are developed, in which both a maximum likelihood method and a Bayesian method with H.pylori as the prior and Yeast as the test data are implemented. Our models are also applied to the predicting of protein structure. In addition, we calculate the network statistics of the protein interaction networks of Yeast and H.pylori, demonstrating that these follow the expected pattern for small world graphs.

Acknowledgements

I would like to thank Professor Gesine Reinert and Dr. Charlotte Deane for the very helpful discussion and guidance on my work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many biological characteristics arise from the complex interactions between the numerous constituents in the cell, such as proteins, DNA, RNA, and small molecules [12]. The interactions of proteins are part of this complicated puzzle. Proteins interact to maintain various biological functions in cells. These interactions can be experimentally detected by many techniques including high-throughput Yeast two hybrid screens (abbreviated as Y2H) and protein complex purification techniques using mass spectrometry, correlated messenger RNA expression profiles and genetic interaction data [54]. Data about protein-protein interactions are publicly available from several databases such as DIP [75], MIPS [41] and IntAct [26]. The goal of my project is to predict Protein-protein interactions (abbreviated as PPI) as the correct inference of missing or unknown interactions will improve the understanding of the biological mechanism and hence benefit drug design.

PPI are affected by many factors including subcellular location, function and structure. Subcellular location affects PPI because proteins are less likely to interact with those which are far away or even unreachable. Function is also important because proteins frequently bind together in pairs or larger complexes to take part in a biological process. Proteins which interact therefore may share the same or similar functions [61]. The specificity of the protein interaction implies that the structure of the protein domain will affect the interaction [2]. Therefore, it is possible to predict, to some extent, the subcellular location, protein function , protein structure or even protein interactions based on a combination of the other factors [14, 48]. A number of computational approaches have been proposed to predict PPI from various aspects. These *in silico* methods are described in Chapter 2.

In this project we study PPI using a statistical approach. Probabilistic models are constructed to analyze biological information. We develop both a frequentist and a Bayesian approach taking advantage of prior information from H.pylori to predict PPI in S. cerevisiae. The model provides estimates of the probability of interaction, given that the structures of two interacting proteins are known.

To validate our models, earlier datasets are used to predict later datasets. The result shows that the Bayesian method performs well when not many interactions are known. The use of prior information from H.pylori can only improve the prediction in the early stage. After more PPI are available for training, the maximum likelihood method gives better estimates. The prediction does not give promising results, indicating that better models are needed. Our models can in principle predict not only the unknown PPI but also the structures of proteins, which

are both important. The method uses cross-species inference and can use not only structural information but also other factors such as functional classes.

Furthermore, if proteins are nodes and their interactions are edges, protein-protein interaction networks (abbreviated as PIN) can be viewed as networks. It is thought that PIN display the small world behaviour, which means that networks are highly clustered with short path lengths [59]. There are studies suggesting a relationship between the network structure on the one hand and the functional role and the subcellular location of proteins on the other hand. Functional classes appear as segregated subnetworks [78]. Meanwhile, whether the PIN are scale-free is still not clear [64, 73]. To investigate the network structure of PPI, we calculate some networks statistics [4] and we evaluate the fit to the theoretical Watt-Strogatz model of small worlds [5–7]. A better understanding of network structure should improve our prediction methods [1] and provide insights on less studied organisms.

Future work is aimed at extending the model considering network structure and also at integrating more biological information. More details of future work are included in Chapter 7.

1.1 Overview

Chapter 2 Here we survey the background of my project, including the experimental techniques to detect PPI and also some recently proposed computational methods for predicting unknown interactions. The high false-positive rate and low coverage from the experiments are reported. In addition, we describe some network characteristics for better understanding of PIN.

Chapter 3 A probabilistic model, the frequency-based approach, is constructed for PIN and two statistical methods, the maximum likelihood method and the Bayesian method are used in the estimation. The data matrix and the frequency tables are built and the statistical model are described in detail.

Chapter 4 Another probabilistic model based on odds ratios is developed that the maximum likelihood method and the Bayesian method are applied as well. We estimate the probabilities of odds ratios between present and absent interactions for each class-class categories.

Chapter 5 Our method is applied to the Yeast data (H.pylori as prior in the Bayesian method). The evaluation of the performance is carried out by comparing χ^2 statistics of two methods. The smaller the χ^2 statistic is, the better the model fitting. Also, our approaches are applied to predict protein structures.

Chapter 6 We explain how to incorporate the structure of PIN in our model, how to calculate some network statistics namely vertex degree, clustering coefficients and the shortest path length, for both Yeast and H.pylori. The small world behaviour in PIN indicates the potential to improve our model using network structure.

Chapter 7 My future work includes five aspects, the handling of experimental errors, integrating multiple sources of data, incorporating the clustering effect in the model, studying network modelling in more detail and applying our model to various datasets. Each aspect is given a brief description.

Chapter 2

Background

2.1 Protein-Protein Interactions

A protein is a complex, high molecular weight organic compound that consists of amino acids joined by peptide bonds. Proteins are essential to the structure and function of all living cells and viruses. The definition of a protein-protein interaction refers to the physical binding, docking, between proteins. Proteins interact with other proteins to achieve biological functions that maintain life. In particular, proteins transmit regulatory signals throughout the cell, catalyze a tremendous number of chemical reactions, and are critical for the stability of numerous cellular structures.

Many methods exist for the detection of protein-protein interactions. These various methods can be roughly defined to be experimentally or computationally derived. Some of them are mixed approaches to cover as many as possible interactions.

2.1.1 Experimental approaches

There are two major approaches, Yeast two-hybrid screening methods, and affinity purification coupled with mass spectroscopy. Other techniques include the use of gene interactions before inferring protein interactions and the study of protein 3-D structures for docked complexes.

- Yeast two-hybrid screening [31, 69]

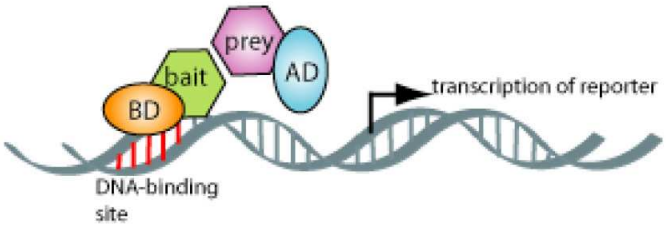


Figure 2.1: Yeast two-hybrid screening
(Picture from http://www.biology.duke.edu/model-system/ymsg/index.html)

The Yeast two-hybrid technique uses two protein domains that have specific functions: a DNA-binding domain (BD), that is capable of binding to DNA, and an activation domain (AD), that is capable of activating transcription of the DNA (see Figure 2.1). Two fusion proteins of interests, bait and prey, are designed one with BD and one with AD. If two proteins interact, AD and BD will be in close distance to activate transcription, so that the report gene is transcribed and its product or the activity can be detected. Thus, it is possible to detect whether or not two proteins interact.

- Affinity purification coupled with mass spectrometry [53]
  PPI can be analyzed directly by precipitation of a tagged bait, a testing protein, followed by mass spectrometric identification of its binding partners.

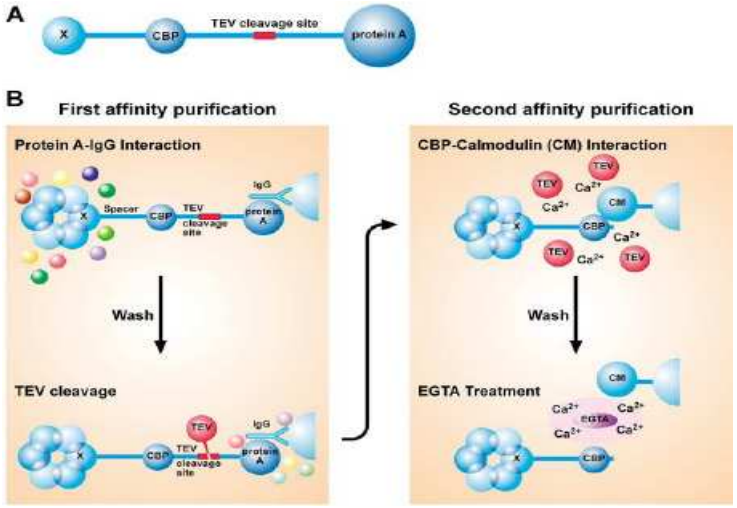


Figure 2.2: Schematic representation of the affinity purification method
A. Structure of the TAP-tag. B. To remove unspecifically bound proteins, two affinity purifications are performed to reduce the affinities of igG and calmodulin binding peptide (CBP). (Picture from A. Bauer & B. Kuster, 2003 [9]).

First, as in Figure 2.2, the "TAP-tagged" protein is expressed in cells, maintaining the expression of the fusion protein at its natural level, to form a complex with the endogenous components. Along with associated partners, it retrieved via interaction of the Protein A tag with the igG that are fixed on agarose beads.

Secondly, in order to remove proteins that are unspecifically bound to the column, the retrieved protein complex is released by proteinase cleavage using the TEV (Tobacco etch virus). After washing, the TEV protease is added to release the bound material. In the second affinity step, the complex is immobilized to calmodulin coated beads via the calmodulin binding peptide (CBP) tag. This step removes the TEV protease and further contaminants that may present. Finally, the resulting complexes can be identified by mass spectrometry [22].

- Correlated mRNA expression
  Messenger RNA (mRNA) expression levels are measured and clustering analysis is used to group genes according to the similarity of their expression across different experimental

conditions and genetic backgrounds. Genes within the same group are assumed to mediate related biological functions and to encode physically interacting proteins. By studying the mRNA level, it is possible to detect protein complexes. The disadvantages are that the method is a relatively inaccurate predictor of direct protein interactions, and it depends on the clustering method used. [38, 39]

- Synthetic lethality
  Two mutations are synthetically lethal if cells with either of the single mutation are viable but cells with both mutations are inviable. Such mutant genes are often functionally related and their encoded proteins may also interact. Through the detection of the synthetic lethality, it is possible to identify putative PPI. [66]

- Docking (3-D)
  In 3-D docking we model their structures, the properties of contact surfaces, forces involved in the interaction, as well as kinetic and thermodynamic parameters. Through the computational simulations, the most likely structure of PPI can be predicted. [24, 58]

Enormous amounts of data have been generated, but unfortunately all these techniques suffer from experimental error [40]. The popular Yeast two-hybrid screening method, one of the high throughput approaches, is most affected. There have been many attempts to assess the reliability of different experimental methods [15, 43].

As most of the PPI-detecting methods are labour-intensive, the development of *in silico* methods is much in demand.

2.1.2 Computational approaches

A number of *in silico* approaches for predicting either physical interactions or functional relationships between proteins have been developed. Genomic context analysis in the prediction of protein-protein interactions includes several approaches, such as *gene fusion* [18, 39], *gene neighborhood* [28, 47], *phylogenetic profiles* [29, 50] and *similarity of phylogenetic trees* [49, 70]. These methods use the evolutionary relationship of genes as well as proteins to detect potential protein-protein interactions. Alternatively, based on the protein-domain information and physical interaction, probabilistic models are built to predict unknown interactions [62]. Recently, more statistical methods were applied to predict unknown interaction [16, 46, 67, 68] and to integrate multiple sources of data [30, 32, 36].

Review of the current approaches

Given a known protein-protein interaction, the ability to predict structured aspects of this interaction would be of great biological value. Here we review some computational approaches used for predicting protein interactions.

- Studies which focus on the protein domain level
  Protein-domain information can be obtained from protein-domain family databases such as PFAM [8]. A protein may contain one or more domains. The existing methods calculate how often two domains are found on two interacting proteins to predict PPI.

1. Domain counting [62]
   For each domain pair a log-odds value compares its observed frequency with the expected frequency from randomly distributed domains.

$$\text{log-odds ratio} = \log \frac{p_{ij}}{p_i p_j}$$

   where p_{ij} is the observed domain pair (i, j) among data, p_i and p_j are the frequency of domain i and domain j in the data.

   The log-odds ratio is a measure of over-represented domain pairs. The domain pair with the highest log-odds ratio is inferred to be the binding pair in the two proteins. Therefore, this protein pair is given a probability score that is equal to the highest log-odds ratio of this domain pair. By setting up a threshold, the probability score is dichotomized to predict the protein-protein interactions. The domain counting method provides a probability score for each protein pair, and a way to detect the probable binding domains.

   The disadvantages of the domain counting method are, firstly, when it calculates the log-odds value, other possible interacting domains are ignored. Secondly, it assumes the independence of domain-domain interactions, though it has been observed that some domains frequently coexist in a protein. Thirdly, the experimental error is totally ignored. In addition, the incompleteness of data will seriously affect the result.

2. Maximum likelihood methods
   An enhanced method based on the domain counting has been proposed by Deng el al. [16]. It uses a maximum likelihood methods with the EM algorithm to estimate the probability of interaction for every domain pair.

   Here is a brief overview.

   Data The protein-domain relationships, which can be obtained from PFAM and the pairwise protein-protein interaction data

   Method Let $D_1, \cdots, D_M$ be M domains and $P_1, \cdots, P_N$ be N proteins. S_{ij} is the set of domain pairs between P_i and P_j. We put $P_{ij} = 1$ if protein i interacts with protein j, and $P_{ij} = 0$ if protein i and protein j have no interaction. Then let $\lambda_{mn} = 1$ indicate that domain m and domain n interact, and let $\lambda_{mn} = 0$ indicate that domain m and domain n have no interaction.

   Two assumptions are made, namely independence between domain-domain interactions and that the protein-protein interaction is based on the presence of the interaction on at least one domain pair. Then we express the probability of protein i interacting with protein j, $Pr(P_{ij} = 1)$, as

$$Pr(P_{ij} = 1) = 1 - \prod_{(m,\, n) \in S_{ij}} (1 - \lambda_{mn}) \quad .$$

   To consider the experimental error from high-throughput experiments, both false positives and false negatives are included in the model as the parameters.

$$f_P = Pr(O_{ij} = 1 | P_{ij} = 0)$$
$$f_N = Pr(O_{ij} = 0 | P_{ij} = 1),$$

where O_{ij} is the binary value for the observed interaction between protein i and protein j.

Thus, the probability of for the observed protein-protein interaction is

$$Pr(O_{ij} = 1) = Pr(P_{ij} = 1)(1 - f_N) + (1 - Pr(P_{ij} = 1))f_P \quad .$$

The likelihood function, the probability of the whole observed data, is the product of the probability of observing the presence/absence of all possible protein pairs,

$$L = \prod (Pr(O_{ij} = 1))^{O_{ij}} (1 - Pr(O_{ij} = 1))^{1-O_{ij}} \quad .$$

The parameters to be estimated are (λ_{mn}, f_P, f_N). Applying the EM algorithm, the likelihood function is iterated to obtain $\hat{\lambda}_{mn}$, the estimate of the probability that two domains (m, n) interact. The probability of the interaction happening on a given protein pair is assumed to be the highest $\hat{\lambda}_{mn}$ among all domain pairs between two proteins. Hence, each protein pair is assigned a probability score. With an appropriate threshold, the prediction of the protein-protein interaction based on the observed interaction is established.

Results This method is then applied on 5719 PPI in Yeast. For a given specificity, the maximum likelihood method outperforms the domain counting method, with better sensitivity. As to the performance, it achieves specificity 42.5% and sensitivity 77.6% by setting the threshold at 0.80. The probability model includes the experiment error and hence makes the prediction more accurate as currently available data have high false-positive rate and low coverage. Additionally, the likelihood function considers the whole interactome at the same time to avoid a local bias on $\hat{\lambda}_{mn}$.

Unfortunately, it is computationally expensive to carry out this operation on a large number of domain pairs. Due to computational difficulties, only the local maximum of the likelihood is obtained with prefixed (f_P, f_N). In the later verification, different initial values of λ_{mn} caused the estimates to vary greatly, which shows the lack of robustness of the approach.

3. Random Shuffling [46]

This is a Monte Carlo approach, as follows.

Aim To predict the most likely pair of domains mediating a given protein interaction; p-values are given to all potential domain superfamily pairs [1].

Data Pairwise PPI (such as the data from Yeast two-hybrid experiments), domain superfamily - protein information (proteins are decomposed into one or several domain superfamilies based on structural classification [44]).

[1] Domains are grouped by SUPERFAMILY [23] and proteins in the same superfamily usually have the same 3D formation. Here a domain superfamily pair refers to two domains with their predicted superfamily.

Method Given the experimental data, the expected number of contacts between a given domain pair j across the entire proteome is E_j. The quantity E_j is calculated based on the number of observed contacts, the total number of possible domain-pair contacts within each protein pair, and also the experimental errors. The total number of different possible contacts of the given domain pair j within the proteome, N_j, is sum of the type j contacts for all protein pairs.

A measure, the odds ratio, of interaction for each domain superfamily pair j is constructed below.

$$s^{(j)} = \log \frac{E_j \sum_{k \neq j}(N_k - E_k)}{(N_j - E_j)(\sum_{k \neq j} E_k)} \quad .$$

It is assumed that the network of interactions between proteins remains fixed, and that the number and the type of domain superfamilies in the observed data remain unchanged. Now all domain superfamilies in the proteome are shuffled. For n random shuffles, n odds ratios for the domain pair j are calculated as $s_1^{(j)}, s_2^{(j)}, \cdots, s_n^{(j)}$. By counting the number of times the simulated statistics exceed the observed statistics $s_0^{(j)}$ the p-value is obtained from the observed data :

$$p_j = \frac{1}{n} \times (\text{number of times that the statistics } s_1^{(j)}, s_2^{(j)}, \cdots, s_n^{(j)} \text{ exceed } s_0^{(j)}) \quad .$$

The performance of the random shuffling approach is compared with the domain counting method and the maximum likelihood method [16, 62]. Generally, their accuracies are similar. When the number of potential contacts increases, the random shuffling method outperforms the others.

Several factors may affect the prediction: firstly, the incompleteness of the data. In addition, repeated domain superfamily pairs in a protein pair with different contacts will receive same p-values that are undistinguishable. As there exists some gaps in the sequence used for superfamily prediction, they are ignored in the analysis and could lead to bias. The removal of proteins without superfamily assignment could also cause a bias.

- Kernel methods [34]

Kernel-based methods handle the relationships among many "items" by describing their similarities as kernels. The nature of kernel matrices allows them to consider several characteristics of protein pairs simultaneously. The integrated information should be of great help when predicting protein-protein interactions.

Aim To classify unknown protein-protein interactions into a positive class (interacting protein pairs) and a negative class (non-interacting protein pairs)

Data Any description helpful in inferring the interaction between proteins, such as amino acid sequences, protein complex data, gene expression, known protein-protein interactions, clustering coefficients of proteins

Kernels The measure of similarity between two genes or proteins based on knowledge of them. The data are included as a matrix of kernel similarity values.

Method All data representing the known relationship between every pair of proteins is summarized as a matrix K produced by the kernel function. A kernel function is a function that describes the similarity between two proteins in the corresponding dataset.

Different kernel functions are designed for different types of data, including vectors, strings, trees and graphs. Selected sources of data are expressed as kernel matrices in which each element is the kernel between two proteins. Multiple matrices can be integrated. The Support Vector Machine (SVM) is a binary classification algorithm that classifies protein pairs as recorded in the kernel matrix, by a linear boundary. The process of the classification is to maximize the distance between the positive and the negative class by optimizing the coefficient of the boundary. Once the optimized boundary is found, unknown PPI can be predicted.

Errors

Experimental errors are observed in the experimental approaches and thus also affect the prediction based on the experimental data.

Recent work indicates that interactions found by experimental screens are far from complete, with thousands to tens of thousands of interactions as yet unknown within Yeast [43, 65]. Among the interactions in Yeast identified by different high throughput methods, only a small number of them overlap and no method covers more than 60% of the proteins. The two main reasons are the coverage and the false positive rate described as follows.

1. Each method is only able to detect a certain distribution of interactions. Different methods cover protein interactions in certain functional categories [40]. Some methods only catch those proteins from certain subcellular locations.

2. The false positive rate is high. In a comparison between the interactions identified by different methods with a reference dataset (gold standard), the accuracy of the data from Yeast two-hybrid and other experimental methods are less than 40% [40]. Deane et al. [15] examined $\sim$ 8000 PPI of Yeast from a large and diverse collection by two computational methods. They estimated that only 50% of them are reliable and only 3000 of them are confident. Other estimates of false positive rates varied from 10 to 95% according to different sources of PPI data [46]. Though the reference datasets might not be the true ones, these diverse rates reveals the low reliability of the data.

An more accurate dataset can be obtained by selecting interactions at the intersections of multiple methods. However, this greatly reduces the size due to low coverage and the high diversity of the detected data by different methods [40].

2.1.3 Network characteristics

There is a series of studies which detect specific patterns of subgraphs in biological networks including the transcription-regulation networks and the PIN, in which the nodes are genes,

transcription factors or proteins and the edges are the interactions [10, 13, 57, 80]. In the study of network motifs (frequently observed patterns) approaches exist to find out the relationship between the network motifs and functional modules [37]. The detection of network motifs allows better inference of missing links [55]. Yet, the underlying reasons for the quantity of different subgraph types, their propensity to form clusters, and their relationship with the networks' global organization remain poorly understood [71].

Basic network statistics, such as vertex degree, clustering coefficient and the shortest path length, are common statistics used for describing networks. Here are their definitions.

Let $G = (V, E)$ be a graph with vertex set V and edge set E, we write $i \sim j$ if $\{i, j\} \in E$ (here the edges are considered undirected.).

- The degree of vertex i is $v(i) = \sum_{j \neq i} \mathbf{1}(i \sim j)$.
- The clustering coefficient of i is

$$C_i = \begin{cases} 0, & \text{if } \sum_{j \neq k} \mathbf{1}(i \sim j,\ i \sim k) = 0 \\ \frac{\sum_{j \neq k} \mathbf{1}(i \sim j \sim k \sim i)}{\sum_{j \neq k} \mathbf{1}(i \sim j,\ i \sim k)}, & \text{otherwise.} \end{cases}$$

  The average clustering coefficient is then $C = \frac{1}{|V|} \sum_{i \in V} C_i$.

- A path of length L from i to j is a collection of edges $(k_l, k_{l+1}) \in E$, where $l = 1, \cdots L-1$ and $k_1 = 1$, $k_L = j$. The length of the shortest path between i and j is the smallest L such that there is a path from i to j of length L.

Further insight into network structure would greatly benefit our understanding of the mechanism of biological interaction. The study of highly connected nodes (hub nodes) shows their conserved property and implies the essential role in interacting with others [20]. Furthermore, the modularity of networks reflects a potential hierarchical structure, in which the modules reflect the existence of protein complexes in protein networks [21, 61]. Several network models have been discussed such as the random network model, the scale-free model and the small-world model [6, 59, 72, 73, 78]. The approach to understand real networks is still vague [76]. The underlying model for PIN is not clear due to the incompleteness of the networks and the potential sampling bias [25, 60, 64, 79]. More details about the network statistics in PIN are described in Chapter 6.

2.1.4 Data Sources

Several public databases (DIP, MIPS, IntAct, etc.) store experimental data of PPI [26, 41, 75]. The DIP, Database of Interacting Proteins [2], stores experimental determined interactions from various sources. DIP also contains a high confidence core subset. The IntAct[3] database from EMBL-EBI provides numbers of PPI from small-scale experiments thought in general to be more accurate than their high throughput counterparts. It also includes datasets from various species.

The datasets from different species are provided at different time points. Here are the two datasets we analyzed in this project.

[2] http://dip.doe-mbi.ucla.edu/

[3] http://www.ebi.ac.uk/intact/index.jsp/

DIP Yeast is the Yeast subset of DIP containing all the pairs of interacting proteins identified in the budding Yeast, *Saccharomyces cerevisiae*. There are 49 datasets provided chronologically from 2001 to 2005 with 7800-17500 interactions each.

DIP H.pylori includes 1420 PPI from 710 proteins. The size of DIP H.pylori stays unchanged from 2003 to 2005.

2.2 Classification of Proteins

The characteristics of proteins, including their 3D structures, functions, or subcellular locations, are useful information for understanding protein interactions [14]. To utilize the protein structures for inferring PPI, a classification of protein structures was employed.

SCOP[4], is a protein structure classification database with four hierarchical levels, Family, Superfamily, Fold and Class [44]. Class, as the top level in SCOP, is based only on the presence of different secondary structure elements (Table 2.1 & Figure 2.3). A SCOP assignment of a protein is a protein structure successfully predicted by Superfamily [23]. Therefore, not all proteins have SCOP assignments. In the DIP Yeast data that we downloaded, there are about 46% to 56% of proteins being successfully predicted by Superfamily and thus have structure assignments. Multi-domain proteins that include domains of different structures are assigned to more than one class. In this case, one protein may be described by multiple SCOP classes.

Table 2.1: The Class level in SCOP †

Class	Description
a	All α proteins
b	All β proteins
c	Alpha and beta proteins (α/β), mainly parallel beta sheets ($\beta-\alpha-\beta$ units)
d	Alpha and beta proteins ($\alpha+\beta$), mainly antiparallel beta sheets (segregated α and β regions)
e	α and β, folds consisting of two or more domains belonging to different classes
f	Membrane and cell surface proteins and peptides, not including proteins in the immune system
g	Small proteins, usually dominated by metal ligand, heme, and/or disulfide bridges

† http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html

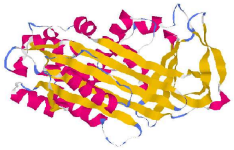
[4]Structural Classification of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop/)

(a)

(b)

(c)

(d)



(e)

(f)

(g)

Figure 2.3: Pictures (a) $\sim$ (g) represent 7 SCOP classes a $\sim$ g in Table 2.1

Chapter 3

Frequency-based Approach

To predict PPI based on the experimental data of Yeast, we first propose a frequency-based approach. This approach estimates the probability of a protein-protein interaction from the relative frequency of the binding between two specific SCOP classes. Assuming that the frequency distribution of class-class interactions is known, the probabilistic model is constructed to obtain the estimates of the probabilities. A maximum likelihood method and a Bayesian approach are both used to obtain the estimates.

3.1 Data Matrix

In this project, the data of PPI are **DIP Yeast** and **DIP H.pylori**. For the n proteins observed in the Yeast data, each protein i corresponds to a protein vector T_i that records all interacting events t_{ij} as binary outcomes with other proteins j. The notation $i \backsim j$ stands for the undirected PPI between protein i and protein j, i.e. $i \backsim j = j \backsim i$. Self-interactions, $i \backsim i$, are also included. To avoid repetition, the interaction $i \backsim j$ or $j \backsim i$ is treated as $i \backsim j$ where $i \leq j$ (in alphabetic order). The order of the binary outcome, t_{ij}, in the protein vector is sorted by the superfamily classes that protein j belongs to.

$$T_i = (\overbrace{t_{i,1}, t_{i,2}, ..}^{a}, \overbrace{.., t_{i,j}}^{b}, .. \overbrace{., t_{i,n}}^{abcdefg}) \qquad t_{ij} = \begin{cases} 1 & \text{if} \quad i \backsim j \\ 0 & \text{if} \quad i \nsim j \quad \text{or} \quad i > j \end{cases}$$

The data matrix, $\mathbb{D}$, is a square matrix of n protein vectors storing all the information of the PPI. As the interactions are undirected, $\mathbb{D}$ is an upper triangular matrix.

$$\mathbb{D} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix}$$

3.2 Frequency Table of Class Interactions

To focus on class interactions between 7 basic SCOP classes predicted by Superfamily, a frequency table of class interactions is calculated from the data of PPI, with data matrix $\mathbb{D}$. The class-class interaction between SCOP class u and SCOP class v is expressed as $u \leftrightarrow v$ and is

undirected.

First, all possible class interactions from every protein interaction in $\mathbb{D}$ are identified. For example, consider a protein x that interacts with a protein y, $x \backsim y$. If protein x is assigned two SCOP classes, a and b, and protein y is assigned two classes, c and d, then four possible class interactions, $a \leftrightarrow c$, $a \leftrightarrow d$, $b \leftrightarrow c$, $b \leftrightarrow d$, are observed.

Second, the number of each type of class interactions counted from all observed PPI are calculated. Let $F_{(i)} = \{\text{superfamily class(es) of protein } i\}$,

$$F_{(i)} \subset \{a, b, c, d, e, f, g\}$$

For example: if protein x belongs to three superfamily classes, a, b and c, then $F_{(x)} = \{a, b, c\}$.

The combination of two classes from two interacting proteins is indexed by k, where their relationships are expressed in Table 3.1. For example, if the SCOP class of one protein is b and its interacting protein is assigned the SCOP class d, then this interaction is indexed as $k = 10$. As the PPI are treated as undirected interactions, the inverse case, one protein is d and its interacting protein is assigned the SCOP class b, is also indexed as $k = 10$.

Table 3.1: The relationships between index k and SCOP classes

k	a	b	c	d	e	f	g
a	1	2	3	4	5	6	7
b	2	8	9	10	11	12	13
c	3	9	14	15	16	17	18
d	4	10	15	19	20	21	22
e	5	11	16	20	23	24	25
f	6	12	17	21	24	26	27
g	7	13	18	22	25	27	28

The frequency of observed class interactions between two classes is n_k, in which the lower index k refers to which two classes are considered. (see Table 3.1 for the index k and Table 3.2 for the arrangement of n_k). For the category k, its frequency is calculated by summing all PPI that one interacting protein is in the corresponding row/column and another protein is in the corresponding column/row, and is as follows.

$$n_k = \sum_i \sum_j t_{ij} \mathbf{1}(\exists\, u, v : u \in F_{(i)}, v \in F_{(j)}, u, v \text{ are in category } k)$$

where $\mathbf{1}$ is the indicator function.

Due to the nature of the available data, only present PPI are observed. The information about proteins which do not interact is very limited, which makes statistical inference more challenging.

Table 3.2: The frequency table of SCOP class interactions

	a	b	c	d	e	f	g	total
a	n_1	n_2	n_3	n_4	n_5	n_6	n_7	
b		n_8	n_9	n_{10}	n_{11}	n_{12}	n_{13}	
c			n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	
d				n_{19}	n_{20}	n_{21}	n_{22}	
e					n_{23}	n_{24}	n_{25}	
f						n_{26}	n_{27}	
g							n_{28}	
total								$n_.$

3.3 Parameters

The parameter to be estimated is the vector of probabilities with the element being P_k, the probability of a protein-protein interaction that its class-class interaction is index as k, one of the 28 types of class-class interaction.

$$\theta = \{P_k |\, k = 1, 2, \cdots, 28\}$$

In total, 7 SCOP classes create 28 different class-class interactions. Thus, 28 probabilities need to be estimated.

3.4 Maximum Likelihood Estimator

The maximum likelihood method is a common method used in estimating parameters in probabilistic models. The maximum likelihood estimates of the probabilities can be directly obtained from the data. The DIP Yeast datasets from different years are used in this project. The first step is to construct the likelihood function based on a probabilistic model. Then the estimates of the probabilities are the values that maximise the likelihood.

Likelihood function The likelihood is comprised of the probabilities from all class interactions decomposed into PPI. The aim is to find out the estimates of probabilities of the 28 class interactions. The estimates will allow further inference on protein interactions. Here, the probability model of class interactions is assumed to be a *multinomial model* in which the observations of PPI are assumed to be independent. In this model, the 28 probabilities of class interactions, $P_k \in [0, 1]$ are parameters. The sample size is the number of all class interactions from the observed data. The sum of probabilities of all 28 class interactions equals to 1,

$$\sum_{k=1}^{28} P_k = 1 \quad .$$

Recall that θ are probabilities of class-class interactions, t are the observed PPI, k indicates the type of two SCOP classes in the class-class interactions.

$$\theta = (P_1, P_2, \cdots, P_{28})$$

$$P_k \geq 0, \quad k = 1, 2, \cdots, 28$$

The likelihood can be expressed as products of probabilities of 28 class interactions with their "observed" class interactions as powers,

$$\begin{aligned} \textit{Likelihood} &= \mathcal{L}(\theta, t) \\ &= \frac{n_.!}{\prod_{k=1}^{28} n_k!} P_1^{n_1} P_2^{n_2} \cdots P_{28}^{n_{28}} \end{aligned}$$

Under this probability model, it is assumed that each protein interaction is independent of all others and each class-class interaction is also independent of all others.

Maximum Likelihood Estimate We compute the log-likelihood and the first partial derivative to obtain the maximum likelihood estimator (MLE). (See for example [52])

$$\hat{\theta}_{mle} = \arg\max_{\theta} \mathcal{L}(\theta, t)$$

Therefore,

$$\begin{aligned} \hat{\theta}_{mle} &= \hat{P}_k = \frac{n_k}{n_.} \\ &= \frac{\text{total of observed class interactions indexed as } k}{\text{total of observed class interactions}} \quad . \end{aligned} \tag{3.1}$$

The MLE of the probability of class interactions is the relative frequency of observed class interactions among all class interactions.

3.5 Bayesian Approach

It is of interest to find out whether it is possible to employ the PPI data from other organisms, to improve the prediction of a less studied organism. Such an approach would be helpful when we explore the PPI of new organisms, especially for those complicated organism where only a small portion of PPI is available. The Bayesian approach provides a way to incorporate prior knowledge of PPI in the estimation. The Bayesian estimate is based on both the prior knowledge gathered from the training dataset and the testing dataset. Therefore, it is possible to predict an organism by using another organism as the prior.

In this project, the DIP H.pylori is employed as the training dataset that gives the prior information of class interactions. The DIP Yeast dataset is the test dataset. The Bayes theorem is recalled before the implementation of the Bayesian method is described.

3.5.1 Bayes Theorem

Let $B_1, B_2, \cdots, B_k$ be a set of mutually exclusive and exhaustive events, For any event A with $P(A) > 0$,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)} \quad .$$

Equivalently,

$$P(B_i|A) \propto P(A|B_i)P(B_i),$$

where we use the notations:

- $P(B_i)$ is the prior probability of B_i
- $P(A|B_i)$ is the likelihood of A given B_i
- $P(B_i|A)$ is the posterior probability of B_i
- $P(A)$ is the predictive probability of A implied by the likelihoods and the prior probabilities.

In general, a Bayesian statistical model consists of

1. A parametric statistical model $f(t|\theta)$ for the data t, where θ is the parameter.
2. A prior distribution $\pi(\theta)$ on the parameter.

The distribution of θ given t is given by

$$\pi(\theta|t) = \frac{f(t|\theta)\pi(\theta)}{\int f(t|\theta)\pi(\theta)d\theta} \text{ or } \frac{f(t|\theta)\pi(\theta)}{\sum_\theta f(t|\theta)\pi(\theta)} \quad ,$$

the posterior distribution of θ given t. The probability of an event is calculated from the data and from the prior knowledge of that event.

3.5.2 Bayesian Estimates

All PPI are classified into one of the 28 categories of class-class interactions. We assume again that the probability model of class interactions follows a multinomial distribution, in which all 28 probabilities of class interactions are parameters in the multinomial distribution and the sample size is the number of all "observed" class interactions, $n_\cdot$.

It is common to further assume an informative prior from the training data, H.pylori subset in this project. The conjugate prior to the multinomial distribution is the Dirichlet distribution, so that the posterior distribution is also a Dirichlet distribution.

Here is the model in more detail.

1. Parametric model – multinomial
   In the model, t is the set of all observed PPI in the data matrix, and n_k is the frequency of PPI falling into a class-class interaction indexed as k (see Table 3.1). Class interactions are computed from all observed PPI, t. The parametric model is set to a multinomial model,

   $$f(t|\theta) = \frac{n_\cdot!}{\prod_{k=1}^{28} n_k!} P_1^{n_1} P_2^{n_2} \cdots P_{28}^{n_{28}} \quad .$$

   Because every class interaction falls into one and only one of 28 categories, the sum of all probabilities is 1, i.e., $\sum_{k=1}^{28} P_k = 1$.

2. Prior $\pi(\theta) \sim$ Dirichlet Distribution
   The prior distribution is set to be a Dirichlet distribution, that is

$$\pi(\theta) = P(\theta|\alpha) = P(P_1, \cdots, P_{28}|\alpha_1, \cdots, \alpha_{28}) = \frac{\Gamma(\sum_{k=1}^{28} \alpha_k)}{\prod_{k=1}^{28} \Gamma(\alpha_k)} \prod_{k=1}^{28} (P_k^{\alpha_k - 1}) \quad ,$$

   where α is the set of hyperparameters in the Dirichlet prior and

$$\alpha_k > 0, \quad \sum_{k=1}^{28} P_k = 1 \quad .$$

3. Choices of hyperparameter, α
   In the prior distribution (from the training dataset, DIP H.pylori subset), parameters are called hyperparameters. In the case of Dirichlet prior, all α are hyperparameters and can be assigned to different values, informative or non-informative. Here, informative hyperparameters are chosen, so that α is set to be the number of the corresponding class interactions,

$$\alpha_k = h_k \quad ,$$

   where h_k is calculated as the same way we did on n_k,

$$h_k = \sum_i \sum_j t_{ij} \mathbf{1}(\exists\ u, v : u \in F_{(i)}, v \in F_{(j)}, u, v \text{ are in category } k) \quad ,$$

   and expressed in the frequency table from the training dataset(s) (see Table 3.3).

Table 3.3: Frequency table of class interaction for training data

	a	b	c	d	e	f	g	total
a	h_1	h_2	h_3	h_4	h_5	h_6	h_7	
b		h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	
c			h_{14}	h_{15}	h_{16}	h_{17}	h_{18}	
d				h_{19}	h_{20}	h_{21}	h_{22}	
e					h_{23}	h_{24}	h_{25}	
f						h_{26}	h_{27}	
g							h_{28}	
total								$h_.$

4. The joint distribution of θ and t given the prior distribution is now calculated as

$$\begin{aligned} P(\theta, t|\alpha) &= f(t|\theta)\pi(\theta) \\ &= \prod_i P_i^{n_i} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i (P_i^{\alpha_i - 1}) \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i (P_i^{n_i + \alpha_i - 1}) \quad . \end{aligned}$$

5. The prior predictive distribution is calculated as

$$\begin{aligned} P(t) &= \int_\theta f(t|\theta)\pi(\theta)d\theta \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int_\theta \prod_i (P_i^{n_i+\alpha_i-1}) d\theta \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \times \frac{\prod_i \Gamma(n_i+\alpha_i)}{\Gamma(\sum_i (n_i+\alpha_i))} \\ &= \frac{\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{28})} \times \frac{\Gamma(\alpha_1+n_1)\cdots\Gamma(\alpha_{28}+n_{28})}{\Gamma(\sum_i n_i + \sum_i \alpha_i)} \\ &= \frac{\prod_i \alpha_i^{[n_i]}}{(\sum_i \alpha_i)^{[n_\cdot]}} \quad , \end{aligned}$$

where $x^{[n]} = x(x+1)\cdots(x+n-1)$, here we used that

$$\frac{\Gamma(\alpha_i+n_i)}{\Gamma(\alpha_i)} = (\alpha_i)(\alpha_i+1)\cdots(\alpha_i+n_1+1) = \alpha_i^{[n_i]}$$

and

$$\frac{\Gamma(\sum_i \alpha_i + \sum_i n_i)}{\Gamma(\sum_i \alpha_i)} = (\sum_i \alpha_i)^{[\sum_i n_i]} = (\sum_i \alpha_i)^{[n_\cdot]} \quad .$$

6. The posterior $\pi(\theta|t)$ is calculated as follows.
Conditioned on observed interactions, the posterior density of 28 categories of SCOP class interactions also follows a Dirichlet distribution and is given by

$$\begin{aligned} \pi(\theta|t) &= \frac{f(t|\theta)\pi(\theta)}{\int_\theta f(t|\theta)\pi(\theta)d\theta} \\ &= \left[\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i (P_i^{n_i+\alpha_i-1})\right] \times \left[\frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \frac{\Gamma(\sum_i n_i+\alpha_i)}{\prod_i \Gamma(n_i+\alpha_i)}\right] \\ &= \frac{\Gamma(\sum_i n_i+\alpha_i)}{\prod_i \Gamma(n_i+\alpha_i)} \prod_i (P_i^{n_i+\alpha_i-1}) \\ &\backsim Dirichlet(n_1+\alpha_1, n_2+\alpha_2, \cdots, n_{28}+\alpha_{28}) \quad . \end{aligned}$$

7. The posterior mean and the posterior variance can be calculated as follows.
The posterior mean $E(\theta|y)$ is the unique Bayes estimate of θ under the squared loss function, $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$, that minimizes the posterior expected loss, $E(L(\hat{\theta}|y))$ [52]. Here, the posterior means are calculated from the posterior distribution, the *Dirichlet* distribution, and are the estimates of probabilities of class interactions.

In general, for variables Θ, $\Theta = \{\theta_j \geq 0, j = 1, \cdots, r\}$ satisfying $\sum_1^r \theta_j = 1$ and following a Dirichlet Distribution, $Dirichlet(\alpha_1, \cdots, \alpha_r)$, the means and variances of Θ are

$$E(\theta_i) = \frac{\alpha_i}{\sum_{j=1}^r \alpha_j}, \quad Var(\theta_i) = \frac{\alpha_i(\sum_{j=1}^r \alpha_j - \alpha_i)}{(\sum_{j=1}^r \alpha_j)^2(\alpha_i+1)} \quad .$$

The explicit formulae for the posterior mean and the posterior variance in $Dirichlet(n_1 + \alpha_1, n_2 + \alpha_2, \cdots, n_{28} + \alpha_{28})$ are given by

$$E(P_k) = \frac{\alpha_k + n_k}{\sum_i (\alpha_i + n_i)} = \frac{\alpha_k + n_k}{(\sum_i \alpha_i) + n_.} \tag{3.2}$$

and

$$Var(P_k) = \frac{(\alpha_k + n_k)(\sum_i (\alpha_i + n_i) - \alpha_k + n_k)}{(\alpha_k + n_k)^2(\alpha_k + n_k + 1)} \quad , k = 1, 2, \cdots 28. \tag{3.3}$$

3.5.3 Prediction

Based on the data matrix $\mathbb{D}$ and the predictive distribution, we can predict the probabilities of having a new protein-protein interaction, if the SCOP assignments to both interacting proteins are available. The prediction of the PPI is the conditional probability of a protein-protein interaction given the SCOP classes of two proteins.

To predict a new interaction between Protein x and Protein y, $x \backsim y$, for Protein x with superfamily assignment $F_{(x)}$ and Protein y with superfamily assignment $F_{(y)}$, we proceed as follows.

Let $P(x \backsim y | t, \alpha, u \leftrightarrow v)$ be the probability that two proteins x and y interact, conditioning on the specific class interaction, $u \leftrightarrow v$. Here, $P(u \leftrightarrow v | t, \alpha)$ is the probability that the specific class interaction, $u \leftrightarrow v$, happened among all potential class interactions between x and y. The probability of having a new interaction conditioning on the known data is to sum up each conditional probability that two proteins interact, given the class interaction between them being observed, and is given by

$$P(x \backsim y | t, \alpha) = \sum_{\substack{u \in F_{(x)} \\ v \in F_{(y)}}} [P(x \backsim y | t, \alpha, u \leftrightarrow v) \cdot P(u \leftrightarrow v | t, \alpha)] \quad .$$

At present, the parametric model only deals with single-class proteins, so the chance that two specific SCOP classes being observed from a given protein pair is one, i.e., there exists no other class-class interaction between them. In this case, the above equation can be simplified as,

$$P(x \backsim y | t, \alpha) = \sum_{\substack{u \in F_{(x)} \\ v \in F_{(y)}}} P(x \backsim y | t, \alpha, u \leftrightarrow v)] \quad .$$

Given the class-class interaction, $u \leftrightarrow v$, between two proteins x and y is in the category k (Table 3.1), the probability that they interact can be estimated by the maximum likelihood estimate and the Bayesian estimate as follows.

- The prediction can be carried out by using maximum likelihood estimate,

$$P(x \backsim y | t, \alpha) = P(x \backsim y | t, \alpha, u \leftrightarrow v) = \hat{P}_k = \frac{n_k}{n_.} \quad .$$

- The prediction can be carried out by using Bayesian estimate,

$$\begin{aligned} P(x \backsim y|t,\alpha) &= P(x \backsim y|t,\alpha,u \leftrightarrow v) \\ &= \frac{P(x \backsim y,t,\alpha|u \leftrightarrow v)}{P(t,\alpha)} \\ &= \frac{\alpha_1^{[n_1]} \cdots \alpha_k^{[n_k+1]} \cdots \alpha_{28}^{[n_{28}]}}{(\sum_i \alpha_i)^{[n_\cdot+1]}} \cdot \frac{(\sum_i \alpha_i)^{[n_\cdot]}}{\alpha_1^{[n_1]} \cdots \alpha_{28}^{[n_{28}]}} \\ &= \frac{\alpha_k + n_k}{(\sum_i \alpha_i) + n_\cdot} \quad , \end{aligned}$$

  using that

$$\frac{a^{[x+1]}}{a^{[x]}} = \frac{a(a+1)\ldots(a+x-1)(a+x)}{a(a+1)\ldots(a+x-1)} = a+x \quad .$$

Before we show the results, we explain a second approach based on odds ratios.

Chapter 4

Odds Ratio-based Approach

Instead of looking at the occurrences of class interactions, the odds ratio of the class interaction gives a measure of the relative count of the class interaction which is found between present links and absent links. The odds ratio of the class interaction calculates the proportion of a certain type of class interaction in the group of present divided by the proportion of the same type of interaction in the group of absent links. In the case that some classes are rare, the odds ratio can reflect the fact that rare observations may still reveal critical information, while they are easily neglected in frequency-based approach.

To be more specific about the links, the present links are the observed class-class interactions decomposed from PPI. The absent links are the links in the complete network of proteins except those observed interactions (present links). As the current data show that PIN are quite sparse, i.e., there are many proteins with relatively small number of links, unknown links are treated as absent links.

4.1 Frequency Table

The complete PIN is constructed using all proteins with SCOP class assignments. The PPI are decomposed into 28 categories of class-class interactions. As in the Table 3.1, each type of class-class interaction is a category with an index k. For each k, three frequencies are calculated for the present link, the absent link and the full link.

The present links are the observed PPI data. The absent links are the difference between full links and present links. The frequency of full links can be estimated using the numbers of proteins in the two SCOP classes. The details are described below.

4.1.1 Inference of full links in the full graph

Consider a PIN with S proteins from 7 SCOP classes,

$$S = m_1 + m_2 + \cdots + m_7$$

$$\mathbf{m} = \{m_1, m_2, \cdots, m_7\}$$

where $M_1, \cdots, M_7$ are the numbers of proteins in SCOP class $a, \cdots, g$, respectively.

The frequency of potential class interactions between class u and class v (full links), which is the row total f_k of category k, is calculated from M_u and M_v as follows,

$$f_k = \begin{cases} \binom{M_u}{2} & \text{if} \quad u = v \qquad \text{proteins from the same class } u \\ M_u \cdot M_v & \text{if} \quad u \neq v \quad \text{proteins from different classes } u \text{ and } v \ . \end{cases} \tag{4.1}$$

where $k \in 1, 2, \cdots, 28$.

4.1.2 Frequency table for odds calculation

The frequency table is designed for the calculation of the odds ratios. It is comprised of the frequencies of present links, absent links and the complete links.

- The frequency of the complete links, f_k, is the frequency of all interactions, including present links and absent links, from proteins in the category k.
- The frequency of the present links, $f_{k,1}$, is the frequency of the observed interactions. The vector of the present links is as $\mathbf{f_{k,1}} = (f_{1,1}, f_{2,1}, \cdots, f_{28,1})$.
- The frequency of the absent links, $f_{k,2}$, can be calculated from $f_{k,2} = f_k - f_{k,1}$.

The frequency table is constructed so that the row represents the category of class-class interactions.

Table 4.1: The frequency table of positive links and false links

class-class interaction	k	present links $f_{k,1}$	absent links $f_{k,2}$	complete links f_k
a $\leftrightarrow$ a	1	$f_{1,1}$	$f_{1,2}$	f_1
a $\leftrightarrow$ b	2	$f_{2,1}$	$f_{2,2}$	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
g $\leftrightarrow$ g	28	$f_{28,1}$	$f_{28,2}$	f_{28}
		$f_{.,1}$	$f_{.,2}$	$f_.$

4.2 Probability Model

To estimate the odds of class interactions, the parameters of interest are π_p and $\pi_{1|k}$ and $\pi_{2|k}$, where π_p is the probability of a randomly picked protein falling into SCOP class p. The parameters $\pi_{k,1}$ and $\pi_{k,2}$ are the probabilities of observing present links and absent links in the class-class interaction indexed k. Then, the conditional probability of the present links and the absent links are $\pi_{1|k}$ and $\pi_{2|k}$, given the class-class interactions is in the category k. Therefore, the set of parameters is

$$\Theta = \{\pi_p, \pi_{1|k}, \pi_{2|k} \quad p = 1, 2, \cdots, 7, \quad k = 1, 2, \cdots, 28\} \quad .$$

The probability model includes two layers of probability models, *binomial* and *multinomial*.

1. Given the frequencies of full links, $(f_1, f_2, \cdots, f_k)$, that are derived from the frequencies of proteins in 7 SCOP classes, $(M_1, M_2, \cdots, M_7)$, each type of class-class interaction, row k, is binomially distributed with probability $\pi_{1|k}$.

$$\mathbf{f_{k,1}} \sim Bin(f_k, \pi_{1|k}) \quad .$$

2. Given the frequencies of proteins in 7 SCOP classes, $(M_1, M_2, \cdots, M_7)$, the rows are conditionally independent.

3. The number of proteins in each SCOP class, $\mathbf{M}$, is multinomial with parameters $\{s, \pi_1, \cdots, \pi_7\}$, where

$$\pi_p = P(\text{a protein } x \text{ drawn at random is in class } p) = P(p \in F_{(x)}) \quad p = 1, 2, \cdots, 7$$

and

$$\sum_{i=1}^{7} \pi_i = 1 \quad .$$

The probability model for the k-th class-class interaction is binomial given $\mathbf{M}$,

$$P(F_{k,1} = f_{k,1} | \mathbf{M} = m) = \binom{f_k}{f_{k,1}} \pi_{1|k}^{f_{k,1}} (1 - \pi_{1|k})^{f_k - f_{k,1}}$$

The joint probability is therefore

$$\begin{aligned} P(\mathbf{F_{k,1}} = f_{k,1}, \mathbf{M} = m) &= P(\mathbf{M} = m) \prod_{i=1}^{28} P(F_{i,1} = f_{i,1} | \mathbf{M} = m) \\ &= \binom{s}{m_1 \cdots m_7} \pi_1^{m_1} \cdots \pi_7^{m_7} \prod_{i=1}^{28} \binom{f_i}{f_{i,1}} (\pi_{1|i})^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}} \quad . \end{aligned}$$

The probability model for all protein interactions is the product of the binomial and the multinomial,

$$\begin{aligned} P(\mathbf{F_{k,1}} = f_{k,1}) &= P(f_{1,1}, \cdots, f_{28,1}) \\ &= \sum_m P(m_1, \cdots, m_7) \prod_{i=1}^{28} P(f_{i,1} | \mathbf{M} = m) \\ &= \sum_m \binom{s}{m_1 \cdots m_7} \pi_1^{m_1} \cdots \pi_7^{m_7} \prod_{i=1}^{28} \binom{f_i}{f_{i,1}} (\pi_{1|i})^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}} \end{aligned}$$

where $m = (m_1, m_2, \cdots, m_7)$

4.3 Maximum Likelihood Estimator

The likelihood is the sum of the joint probabilities over all possible m.

$$\begin{aligned} \textit{Likelihood} &= \mathcal{L}(\Theta, M) \\ &= \sum_m \binom{s}{m_1 \cdots m_7} \pi_1^{m_1} \cdots \pi_7^{m_7} \prod_{i=1}^{28} \binom{f_i}{f_{i,1}} (\pi_{1|i})^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}} \quad . \end{aligned}$$

We compute the first partial derivative to obtain maximum likelihood estimates,

$$\hat{\theta}_{mle} = \arg\max_{\theta} \mathcal{L}(\Theta, M)$$

under the constraint

$$\pi_1 + \pi_2 + \cdots + \pi_7 = 1 \quad .$$

MLE of $\pi_{1|k}$ Differentiating with respect to $\pi_{1|k}$ ($\pi_{1|k} > 0$), $k = 1, \cdots, 28$ and equating the resulting expressions to zero, we get

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \pi_{1|k}} &= \sum_m A \times \left[\prod_{\substack{i=1 \\ i \neq k}}^{28} \binom{f_i}{f_{i,1}} (\pi_{1|i})^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}} \right] \\
&\times \binom{f_k}{f_{k,1}} (\pi_{1|k})^{f_{k,1}} (1 - \pi_{1|k})^{f_k - f_{k,1}} \left[\frac{f_{k,1}}{\pi_{1|k}} - \frac{(f_k - f_{k,1})}{1 - \pi_{1|k}} \right] \\
&= 0
\end{aligned}$$

where $A = \binom{s}{m_1 \cdots m_7} \pi_1^{m_1} \cdots \pi_7^{m_7}$, yielding

$$\frac{f_{k,1}}{\pi_{1|k}} - \frac{(f_k - f_{k,1})}{1 - \pi_{1|k}} = 0$$

so that

$$\hat{\pi}_{1|k} = \frac{f_{k,1}}{f_k} \tag{4.2}$$

The second derivative is calculated to identify the estimator is to maximize the likelihood.

$$\begin{aligned}
\frac{\partial \mathcal{L}^2(\hat{\pi}_{1|k})}{\partial {\pi_{1|k}}^2} &= \sum_m A \times \left[\prod_{\substack{i=1 \\ i \neq k}}^{28} \binom{f_k}{f_{k,1}} \pi_{1|k}^{f_{k,1}} (1 - \pi_{1|k})^{f_k - f_{k,1}} \right] \\
&\times \binom{f_k}{f_{k,1}} \pi_{1|k}^{f_{k,1}} (1 - \pi_{1|k})^{f_k - f_{k,1}} \left[-\frac{f_{k,1}}{(\hat{\pi}_{1|k})^2} - \frac{(f_k - f_{k,1})}{(1 - \pi_{1|k})^2} \right]
\end{aligned}$$

if $f_{k,1} \geq 0$, the we have $\hat{\pi}_{1|k} = \frac{f_{k,1}}{f_k} \in [0, 1]$.
It is straightforward to verify that

$$\frac{\partial \mathcal{L}^2(\hat{\pi}_{1|k})}{\partial {\pi_{1|k}}^2} \leq 0 \quad .$$

When, the present links exist, the second derivative is negative, so that the estimator, $\hat{\pi}_{1|k}$, is a MLE.

Hence, the MLE of the probability of observing positive links in the class-class interaction ($u \leftrightarrow v$), row k, is the ratio of observed count of class-class interactions between class u and class v and the total of all potential interactions between class u and class v.

MLE of π_p To find the MLE, the method of Lagrange multiplier is applied [52]. The extreme values for the likelihood of parameter π_p, under the constraint $\sum_{i=1}^{7} \pi_i - 1 = 0$, are to be found on the surface $g = g(\underline{\pi}) = \sum_{i=1}^{7} \pi_i - 1 = 0$ at the points where

$$\triangle \mathcal{L} = \lambda \nabla g$$

for some scalar λ.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_p} &= \sum_m B \times \binom{s}{m_1 \cdots m_7} (\prod_{\substack{i=1 \\ i \neq p}}^{7} \pi_i^{m_i})(m_p \cdot \pi_p^{m_p - 1}) \\ &= \sum_m B \times \binom{s}{m_1 \cdots m_7} (\prod_{i=1}^{7} \pi_i^{m_i})(\frac{m_p}{\pi_p}) \\ &= \lambda \frac{\partial g}{\partial \pi_p}\end{aligned}$$

where $B = \prod_{i=1}^{28} \binom{f_i}{f_{i,1}} \pi_{1|i}^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}}$ and $\frac{\partial g}{\partial \pi_p} = 1$.

Because λ is a constant for every equation with respect to different π_p, $p = \{1, 2, \cdots, 7\}$, by equating λ in each equation we get

$$\begin{aligned}\frac{m_1}{\pi_1} = \frac{m_2}{\pi_2} = \cdots = \frac{m_7}{\pi_7} &\underset{\text{let}}{=} \rho \\ \Rightarrow \pi_i = \frac{m_i}{\rho} \quad &.\end{aligned}$$

Under the constraint, we have $\sum_{i=1}^{7} \frac{m_i}{\rho} = 1$ implying $\rho = s$.
Therefore, the MLE for π_p is

$$\hat{\pi}_p = \frac{m_p}{s}, \quad \text{where } p = 1, \cdots, 7 \quad . \tag{4.3}$$

4.4 Bayes Estimator

The Bayesian approach is applied in parallel to estimate the probabilities of class-class interactions in PPI. In addition to the target organism, Yeast in this project, another organism, H.pylori, is used to provide prior information.

1. Parametric model: Multinomial × Binomial
   Following the probability model constructed in the previous section, the parametric model is as follows.

$$f(\mathbf{F_{k,1}} = f_{k,1}, \mathbf{M} = m | \underline{\pi}) = \binom{s}{m_1 \cdots m_7} \pi_1^{m_1} \cdots \pi_7^{m_7} \prod_{i=1}^{28} \binom{f_i}{f_{i,1}} (\pi_{1|i})^{f_{i,1}} (1 - \pi_{1|i})^{f_i - f_{i,1}}$$

   where $\underline{\pi} = \{\pi_1, \pi_2, \cdots, \pi_7; \pi_{1|1}, \cdots, \pi_{1|28}\}$.

2. Distribution of prior: Dirichlet × Beta
   The informative prior from H.pylori on $\underline{\pi}$ are chosen to be the Dirichlet distribution as the conjugate prior to the multinomial distribution in the parametric model. Another prior on $\pi_{1|k}$ is the conjugate prior to the binomial distribution, the Beta distribution, $Beta(\alpha, \beta)$. Here, we choose a non-informative prior, the uniform distribution, which is a

special case of the Beta distribution when $\alpha = 1$ and $\beta = 1$. So, the prior distribution in our Bayesian approach is a Dirichlet $\times$ Beta distribution,

$$\begin{aligned}
\pi(\underline{\pi}|\underline{\alpha}) &= P(\underline{\pi}|\underline{\alpha}) \\
&= P(\pi_1, \cdots, \pi_7|\alpha_1, \cdots, \alpha_7) \cdot P(\pi_{1|1}, \pi_{1|2}, \cdots, \pi_{1|28}) \\
&= \frac{\Gamma(\sum_{i=1}^7 \alpha_i)}{\prod_{i=1}^7 \Gamma(\alpha_i)} \prod_{i=1}^7 \pi_i^{\alpha_i - 1} \cdot 1
\end{aligned}$$

where $\underline{\alpha}$ is the set of hyperparameters in the Dirichlet prior, $\alpha_p > 0$, and $\sum_{p=1}^7 \pi_p = 1$.

3. Joint distribution
The joint distribution of $\underline{\pi}$ and $(m, f_{k,1})$ is given by

$$\begin{aligned}
P(m, f_{k,1}, \underline{\pi}|\underline{\alpha}) &= f(m, f_{k,1}|\underline{\pi}) \cdot \pi(\underline{\pi}|\underline{\alpha}) \\
&= \binom{s}{m_1 \cdots m_7} \times \prod_{i=1}^7 \pi_i^{m_i} \left[\prod_{j=1}^{28} \binom{f_j}{f_{j,1}} (\pi_{1|j})^{f_{j,1}} (1 - \pi_{1|j})^{f_j - f_{j,1}} \right] \\
&\times \left[\frac{\Gamma(\sum_{i=1}^7 \alpha_i)}{\prod_{i=1}^7 \Gamma(\alpha_i)} \prod_{i=1}^7 \pi_i^{\alpha_i - 1} \right] \times 1 \\
&= C \times B \times \frac{\Gamma(\sum_{i=1}^7 \alpha_i)}{\prod_{i=1}^7 \Gamma(\alpha_i)} \prod_{i=1}^7 \pi_i^{\alpha_i + m_i - 1} \quad ,
\end{aligned}$$

where $C = \binom{s}{m_1 \cdots m_7}$ and $B = \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} (\pi_{1|j})^{f_{j,1}} (1 - \pi_{1|j})^{f_j - f_{j,1}}$.

4. Prior predictive distribution
To sum up all $\underline{\pi}$ we will have the prior predictive distribution on $(m, f_{k,1})$.

$$\begin{aligned}
P(m, f_{k,1}) &= \int_{\underline{\pi}} f(m, f_{k,1}|\underline{\pi}) \cdot \pi(\underline{\pi}|\underline{\alpha}) d\underline{\pi} \\
&= \int_{\underline{\pi}} \binom{s}{m_1 \cdots m_7} \prod_{i=1}^7 \pi_i^{m_i} \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} (\pi_{1|j})^{f_{j,1}} (1 - \pi_{1|j})^{f_j - f_{j,1}} \\
&\times \frac{\Gamma(\sum_{i=1}^7 \alpha_i)}{\prod_{i=1}^7 \Gamma(\alpha_i)} \prod_{i=1}^7 \pi_i^{\alpha_i - 1} d\underline{\pi} \\
&= C \times D \times \int_{\pi_1 \cdots \pi_7} \prod_{i=1}^7 \pi_i^{\alpha_i - 1} d\pi_1 \cdots \pi_7 \\
&\times \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} \int_0^1 (\pi_{1|j})^{f_{j,1}} (1 - \pi_{1|j})^{f_j - f_{j,1}} d\pi_{1|j} \\
&= C \times D \times \int_{\pi_1 \cdots \pi_7} \prod_{i=1}^7 \pi_i^{\alpha_i - 1} d\pi_1 \cdots \pi_7 \\
&\times \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} \frac{\Gamma(f_{j,1} + 1) \times \Gamma(f_j - f_{j,1} + 1)}{\Gamma(f_j + 2)} \quad ,
\end{aligned}$$

where $C = \binom{s}{m_1 \cdots m_7}$ and $D = \frac{\Gamma(\sum_{i=1}^7 \alpha_i)}{\prod_{i=1}^7 \Gamma(\alpha_i)}$.
Note that the Beta function[1] is used in the integration.

5. Posterior Distribution
With the Dirichlet×Beta prior, the posterior distribution is shown to remain a Dirichlet×Beta distribution,

$$\begin{aligned}
\pi(\underline{\pi}|m, f_{k,1}) &= \frac{f(m, f_{k,1}|\underline{\pi})\pi(\underline{\pi}|\underline{\alpha})}{\int_{\underline{\pi}} f(m, f_{k,1}|\underline{\pi})\pi(\underline{\pi}|\underline{\alpha})d\underline{\pi}} \\
&= \frac{C \times \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} \pi_{1|j}^{f_{j,1}} (1-\pi_{1|j})^{f_j - f_{j,1}} \times D \times \prod_{i=1}^7 \pi_i^{\alpha_i + m_i - 1}}{C \times D \times \frac{\prod_{i=1}^7 \Gamma(\alpha_i + m_i)}{\Gamma(\sum_{i=1}^i \alpha_i + m_i)} \times \prod_{j=1}^{28} \binom{f_j}{f_{j,1}} \frac{\Gamma(f_{j,1}+1)\Gamma(f_j - f_{j,1}+1)}{\Gamma(f_j+2)}} \\
&= \left[\frac{\prod_{i=1}^7 \Gamma(\alpha_i + m_i)}{\Gamma(\sum_{i=1}^7 \alpha_i + m_i)}\right] \cdot \left[\prod_{j=1}^{28} [\frac{\Gamma(f_j+2)}{\Gamma(f_{j,1}+1)\Gamma(f_j - f_{j,1}+1)} \pi_{1|j}^{f_{j,1}} (1-\pi_{1|j})^{f_j - f_{j,1}}]\right]
\end{aligned}$$

$\backsim Dirichlet(\alpha_1 + M_1, \alpha_2 + M_2, \cdots, \alpha_7 + M_7) \times \prod_{j=1}^{28} Beta(f_{j,1} + 1, f_j - f_{j,1} + 1)$.

6. Posterior mean
Since the posterior distribution is a product of a Dirichlet and a Beta distribution, the estimates the posterior probabilities of π_p and $\pi_{1|k}$ can be calculated through their posterior means as the section 3.5.2.

The posterior mean of π_p is the mean in the Dirichlet distribution,

$$\hat{\pi}_p = E(\pi_p|m) = \frac{\alpha_p + m_p}{\sum_{i=1}^7 (\alpha_i + m_i)} \quad p \in \{1, 2, \cdots, 7\} \quad . \tag{4.4}$$

The posterior mean of $\pi_{1|k}$ is the mean in the Beta distribution. As the mean for a random variable following a Beta distribution, $Beta(\alpha, \beta)$, is $\frac{\alpha}{(\alpha+\beta)}$, therefore the mean in $Beta(f_{j,1} + 1, f_j - f_{j,1} + 1)$ is

$$\hat{\pi}_{1|k} = E(\pi_{1|k}|f_{k,1}) = \frac{f_{k,1} + 1}{(f_{k,1} + 1) + (f_k - f_{k,1} + 1)} = \frac{f_{k,1} + 1}{f_k + 2} \quad k \in \{1, 2, \cdots, 28\}. \tag{4.5}$$

7. Posterior variance
The posterior variance for π_p is calculated from $Dirichlet(\alpha_1 + M_1, \alpha_2 + M_2, \cdots, \alpha_7 + M_7)$ and is given by,

$$Var(\pi_p|m) = \frac{(\alpha_p + m_p)[\sum_{i=1}^7 (\alpha_i + m_i) - (\alpha_p + m_p)]}{(\sum_{i=1}^7 \alpha_i + m_i)^2 (\sum_{i=1}^7 (\alpha_i + m_i) + 1)} \quad .$$

The posterior variances for $\pi_{1|k}$ is calculated from $Beta$ distribution, $Beta(\alpha, \beta)$. The variance in Beta distribution is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. So the variance of $\pi_{1|k}$ in $Beta(f_{j,1} + 1, f_j - f_{j,1} + 1)$ is given by

$$Var(\pi_{1|k}|f_{k,1}) = \frac{(f_{k,1} + 1)(f_k - f_{k,1} + 1)}{(f_k + 2)^2 (f_k + 3)} \quad .$$

[1]Beta function: $\beta(n, m) = \int_0^1 t^{n-1}(1-t)^{m-1} dt = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)}$

Chapter 5

Results

The data analysis is performed using both the frequency-based approach and the odds ratio-based approach. These two approaches view the relationship between the protein classes and the protein interaction differently. The first one estimates the probability an interaction falls into a specific class-class interaction while the second one estimates how likely an interaction is observed given that the protein classes are known. Though they have different properties, these two approaches should both help us to understand the relationship between the protein structure and protein interactions.

A Bayesian method and the maximum likelihood method are used for estimation. The Bayesian method employs a prior derived from H.pylori data. Different priors may affect the Bayesian method greatly and therefore the choice of prior needs to be considered carefully. In general, the Bayesian method performs better when few interactions are known and the MLE when more data is available.

Different subsets of the data are used to explore the possible factors affecting the results. Considering interacting proteins from a smaller subset of SCOP classes may reduce the noise from minor classes. Here we select PPI from the first four SCOP classes, a, b, c and d, that comprised 84% of all PPI.

5.1 Size of DIP Yeast Subsets

There are 49 Yeast datasets of PPI in DIP from the year 2001 to the year 2005. The datasets are uploaded in a monthly fashion. We downloaded all subsets of Yeast in the DIP database. The older interactions are included in newer datasets which are presented in the Figure 5.1. The blue line on the top is the total number of PPI and the two lines below are restricted on PPI from single class proteins in 7 and 4 SCOP classes. Two major jumps indicate that many interactions are identified in 2002-02 and in 2003-01. Reduced datasets only include proteins from the major SCOP classes, a, b, c, and d.

5.2 Results of the Frequency-based Approach

Following the previous chapter, the frequency-based approach is applied to the analysis of PPI from DIP Yeast. The probabilities of the PPI from a specific class-class interaction are estimated by both the maximum likelihood method and the Bayesian method. Two methods are

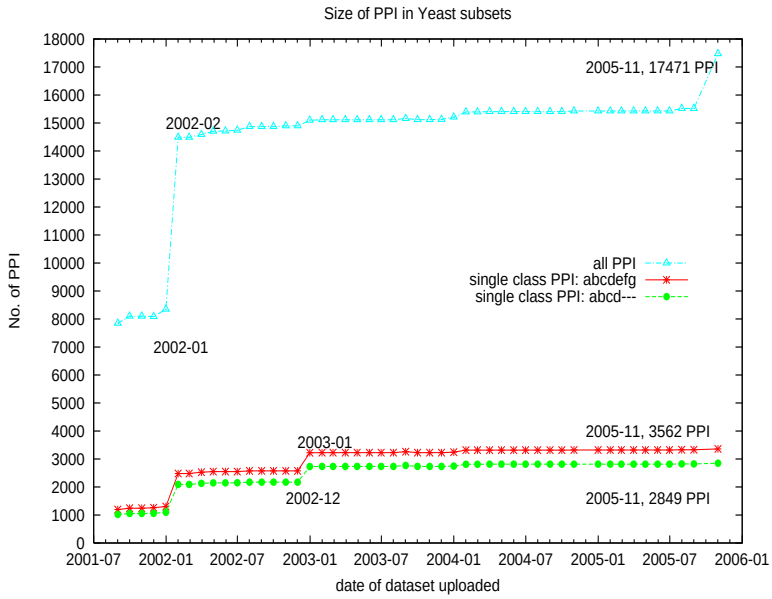


Figure 5.1: Number of PPI in DIP Yeast subsets

compared via their Pearson's χ^2 statistics, as it measures the deviance between observed data and predicted data. Earlier datasets are used to predict new interactions in the later datasets.

Here, our selection criteria is that we select only the interactions formed by two single class proteins. In the case when the multiple class protein(s) involved in interactions, it is not clear which structure class binds which.

1. Maximum likelihood estimates
   The maximum likelihood estimates are calculated from the frequency table of class interactions. It is the relative frequency of the class interaction, $\frac{n_{kl}}{n_{..}}$ (equation 3.1).

   Initially, we analysed the Yeast dataset uploaded on 2001-12 (DIP Yeast 2001). In the DIP Yeast 2001, there are 8087 PPI observed from 4145 proteins. Among those proteins, 1981 (47.8%) are identified to be single-class. We select 887 single class proteins with 1260 PPI among them, in order to meet our selection criteria. Table 5.1 shows the distribution of these PPI in the 28 class-class categories

Table 5.1: Frequency table of class-class interactions (DIP Yeast 2001)

SCOP class	a	b	c	d	e	f	g	total
a	89(0.071)	126(0.100)	127(0.101)	114(0.090))	15(0.012)	2(0.002)	28(0.022)	501
b		96(0.076)	92(0.073)	82(0.065)	5(0.004)	2(0.002)	22(0.017)	299
c			132(0.105)	100(0.079)	11(0.009)	7(0.006)	26(0.021)	276
d				99(0.079)	12(0.010)	3(0.002)	40(0.032)	154
e					3(0.002)	1(0.001)	3(0.002)	7
f						0(0.000)	0(0.000)	0
g							23(0.018)	23
total								1260(1.0)

The MLE is shown in parenthesis.

2. Bayesian estimates
The prior data are the PPI from DIP H.pylori. It includes 1420 PPI from 710 proteins. Among them, 434 (61.13%) proteins have the SCOP class assignment. Totally, 185 PPI formed by single class proteins are selected as in Table 5.2. In our model (section 3.5.2), the account of each cell h_{kl} is the hyperparameters, α_{kl} in the Bayesian estimator.

Table 5.2: Frequency table of class-class interactions (DIP H.pylori)

SCOP class	a	b	c	d	e	f	g	total
a	11(0.059)	1(0.005)	26(0.141)	2(0.011)	1(0.005)	4(0.022)	0(0.000)	45
b		3(0.016)	9(0.049)	5(0.027)	0(0.000)	0(0.000)	0(0.000)	17
c			51(0.276)	40(0.216)	2(0.011)	11(0.059)	0(0.000)	104
d				17(0.092)	0(0.000)	1(0.005)	0(0.000)	18
e					0(0.000)	0(0.000)	0(0.000)	0
f						1(0.005)	0(0.000)	1
g							0(0.000)	0
total								185(1.0)

The relative frequencies are shown in parenthesis.

Since an empty cell, i.e, zero count, gives the estimate of the posterior probability zero, it is risky to give zero probability to any type of class-class interaction, when it could be due to insufficient observations of PPI rather than biological reality. To counter this, we add 1 to all cells in the frequency table of H.pylori.

The estimate of the posterior probability is given by the posterior mean. The Dirichlet posterior mean is $\frac{\alpha_{kl}+n_{kl}}{(\sum_{ij}\alpha_{ij})+n_{..}}$ with $\alpha_{kl} = h_{kl}$ from the frequency table of H.pylori. (Table 5.2 and 5.3)

Table 5.3: Dirichlet posterior means ± standard deviation

SCOP class	a	b	c	d	e	f	g
a	0.068±0.006	0.086±0.007	0.103±0.008	0.079±0.007	0.012±0.003	0.005±0.002	0.020±0.004
b		0.067±0.006	0.069±0.007	0.059±0.006	0.005±0.002	0.003±0.001	0.016±0.003
c			0.123±0.008	0.095±0.008	0.010±0.003	0.013±0.003	0.019±0.003
d				0.079±0.007	0.009±0.002	0.004±0.002	0.028±0.004
e					0.03±0.001	0.002±0.001	0.003±0.002
f						0.002±0.001	0.001±0.001
g							0.017±0.003
total							1.0

The later dataset uploaded on 2002-12 (DIP Yeast 2002) is predicted. Table 5.4 shows the observed PPI in DIP Yeast 2002 and their expected frequencies in each of the 28 class-class categories. The expected frequencies are calculated from the Dirichlet posterior means in Table 5.3 times 2575, the total number of PPI in DIP Yeast 2002.

Table 5.4: Frequency table of observed data vs. expected data (DIP Yeast 2002)

SCOP class	a	b	c	d	e	f	g	total
a	167 (175.0)	237 (221.3)	255 (265.9)	247 (202.4)	40 (30.9)	3 (13.7)	46 (51.5)	
b		161 (173.3)	179 (176.7)	180 (152.7)	23 (12.0)	3 (6.9)	42 (41.2)	
c			218 (317.4)	301 (243.6)	41 (25.7)	11 (34.3)	40 (48.0)	
d				226 (202.4)	53 (24.0)	6 (10.3)	52 (72.1)	
e					6 (8.6)	1 (5.1)	9 (8.6)	
f						0 (5.1)	0 (3.4)	
g							28 (42.9)	
total								2575 (2575.0)

The expected frequency by the Bayesian method is shown in parenthesis.

3. Evaluation of the performance
   Our models are tested using earlier datasets to predict later datasets. The evaluation of the performance is carried out using the Pearson's χ^2 statistics, where the degree of freedom is $28-1$. A smaller χ^2 means a smaller difference between the observed frequency and the expected frequency, thus a better prediction.

   In Figure 5.2, the results show that the Bayesian estimate performs better, which means smaller Pearson's χ^2 statistics, when the data are not intensively explored (i.e., less PPI data available in the early stage) so that the prior from H.pylori provides useful information in the Bayesian model. Meanwhile, the maximum likelihood method yields good results after accumulating many known interactions. Hence, the maximum likelihood method would be more useful in those species where large amounts of PPI data are available.

   In addition, the reduced datasets are analysed because the reduction of the noise from rarely observed class interactions enlarges the differences between the two methods, and the χ^2 statistic clearly decreases as the analysing domains are cleaner. As 84% of interactions are observed in the first four SCOP classes, the result therefore can still be considered representative.

   In the righthand of Figure 5.2, the Bayesian method has constantly smaller χ^2 values than the maximum likelihood when it is trained with the earlier dataset. Though similar trends are found in the two figures in Figures 5.2, the reduced dataset shows a clearer difference between the MLE and the Bayesian method.

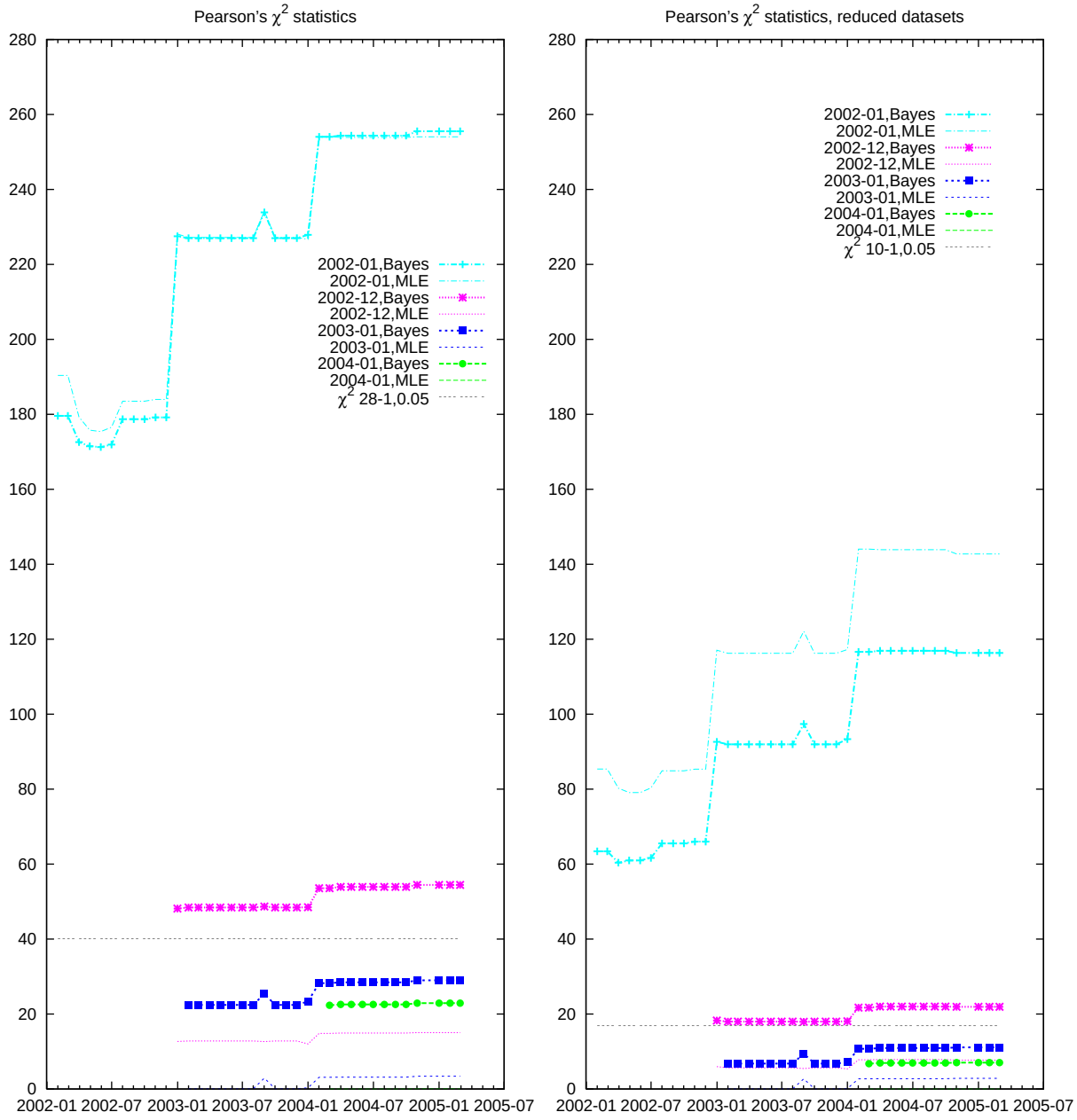


Figure 5.2: χ^2 statistics from Bayesian estimates and MLE

The χ^2 statistics from the two methods are showed in Figures 5.2. In the lefthand of Figure 5.2 is the analysis using proteins from all 7 SCOP classes (i.e., Class a to Class g) and the righthand figure uses proteins from the 4 major SCOP classes, a, b, c and d. In each figure, four test datasets are selected to establish the probability model.

5.3 Results of the Odds Ratio-based Approach

The odds ratio-based approach is also applied to the analysis of the DIP Yeast datasets. Similar to the previous analysis (section 5.2), in order to demonstrate the computation of the Bayesian estimates and the maximum likelihood estimates, the subset of PPI in Yeast dated 2002-01 is used to as the training dataset and a later subset dated 2002-02 is set as the target dataset to be predicted. The prior in the Bayesian method is still counted using the PPI in H.Pylori. Again, only single-class proteins are considered here.

Aim The aim of the odds ratio-based approach is

1. To estimate $\hat{\pi}_p$, the probabilities of proteins in the SCOP class, where $p = 1, \cdots, 7$, (see equations 4.3 & 4.4).
2. To estimate $\hat{\pi}_{1|k}$, the probability of the odds of two proteins interacting given their SCOP classes, where $k = 1, \cdots, 28$, (see equations 4.2 & 4.5).

First, the distribution of the number of proteins in every SCOP class is summarized in Table 5.5.

Table 5.5: Number of protein in the SCOP classes

SCOP class	a	b	c	d	e	f	g	total
H.pylori	19	12	90	39	3	8	0	171
Yeast 2002-01	194	138	282	221	25	6	53	919
Yeast 2002-02	245	189	367	287	37	9	62	1196

Secondly, from the observed frequencies of class distribution, the numbers of all interactions (links in the full graph) are calculated according to the formula (equation 4.1).

Table 5.6: Frequency table of the observed interactions and the full links (in parentheses) from DIP Yeast 2002-01

SCOP class	a	b	c	d	e	f	g	total
a	90 (18721)	134 (26772)	129 (54708)	116 (42874)	15 (4850)	2 (1164)	28 (10282)	
b		109 (9453)	97 (38916)	88 (30498)	5 (3450)	2 (828)	25 (7314)	
c			133 (39621)	100 (62322)	11 (7050)	7 (1692)	26 (14946)	
d				100 (24310)	12 (5525)	3 (1326)	40 (11713)	
e					3 (300)	1 (150)	3 (1325)	
f						0 (15)	0 (318)	
g							23 (1378)	
total								1302 (421821)

The number of full links is shown in parenthesis.

Estimators Both the MLE and Bayesian estimator are employed.

The estimators for π_p are given in (4.3) and (4.4), namely

$$\hat{\pi}_{p\ [mle]} = \frac{m_p}{s} \quad p = 1, \cdots, 7$$

and

$$\hat{\pi}_{p\ [bayes]} = \frac{\alpha_p + m_p}{\sum_{p=1}^{7}(\alpha_p + m_p)} \quad p = 1, 2, \cdots, 7 \quad .$$

The estimators for $\pi_{1|k}$ are given in (4.2) and (4.5), namely

$$\hat{\pi}_{1|k\ [mle]} = \frac{f_{k,1}}{f_k} \quad k = 1, \cdots, 28$$

and

$$\hat{\pi}_{1|k\ [bayes]} = \frac{f_{k,1} + 1}{f_k + 2} \quad k = 1, 2, \cdots, 28 \quad .$$

Data analysis The subset of DIP Yeast in 2002-01 are selected as the test data to predict another subset of DIP Yeast in 2002-2. The estimates are computed by both methods and compared against the real data. In addition, the 95% credible intervals are provided for the Bayesian estimates of probabilities.

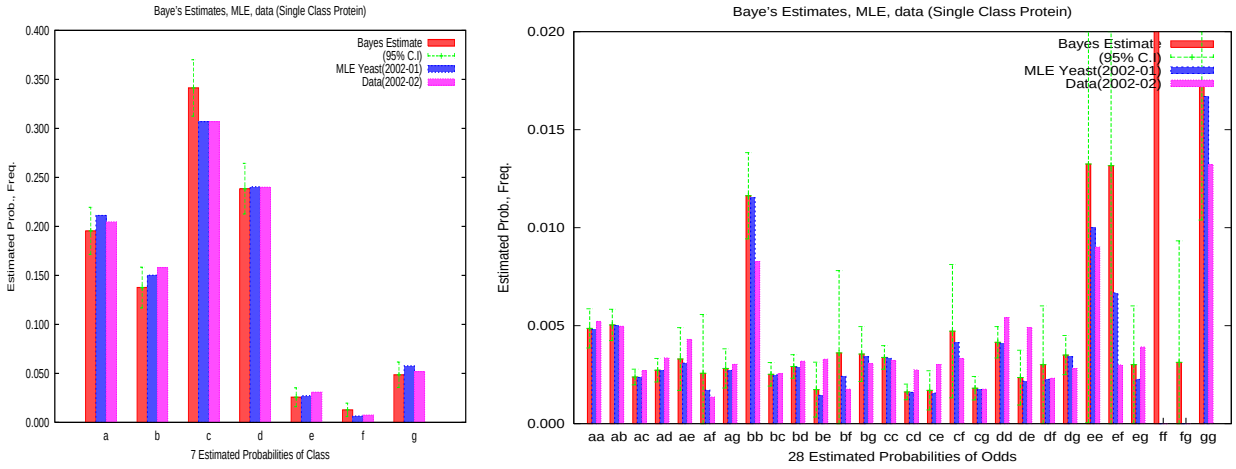


Figure 5.3: Comparison among Bayesian estimate, MLE, and real data with proteins from 7 SCOP classes

Figures 5.3 and 5.4 show the comparison of the two estimates against the relative frequency of the target dataset. The left figure is the estimate for $\hat{\pi}_p$ and the right figure is for $\hat{\pi}_{1|k}$. It is not surprising that both methods perform well in catching the pattern of the real distribution. It does not seem that one method is better than another from the histogram. The two methods obtain close estimates in most predictions.

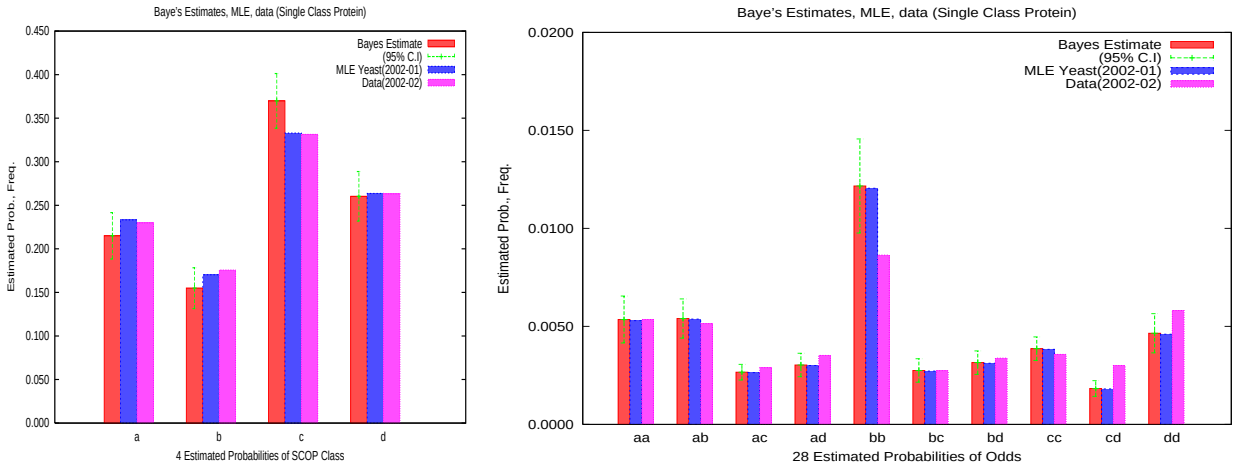


Figure 5.4: Comparison between Bayesian estimate, MLE and real data with proteins from 4 SCOP classes (a,b,c,d)

In Figure 5.3, the Bayesian method has several bad estimates on $\hat{\pi}_{1|k}$ due to several zero/small frequencies from the prior. This disadvantage could be avoid by the removal of these counts. Consequently, the analysis carried out with proteins from only the first 4 SCOP classes is presented in Figure 5.4.

Comparison of the methods We evaluate the two methods by comparing their Pearson's χ^2 statistics. Exactly as when evaluating the frequency-based approach, four subsets of DIP Yeast dated from 2001-09 to 2004 -02 are picked as the test datasets to predict later identified PPI.

In Figure 5.5, the figure on the righthand is the analysis χ^2 statistics using proteins from 7 SCOP classes whereas the left one uses only major SCOP classes. In general, the χ^2 statistics in the estimation of $\hat{\pi}_p$ shows that the maximum likelihood method performs better (smaller χ^2 statistics).

In Figure 5.6, the Bayesian method has better performance in the estimation of $\hat{\pi}_{1|k}$ (smaller χ^2 statistics), when only a few PPI data are available for the target organism. The prior helps in gaining information from another organism. On the righthand figure, the analysis using proteins from only major SCOP classes reduces the χ^2 statistic. Both methods have similar performance. Thus, the use of prior is not much helpful here. It might suggest that better models and more analyses are needed.

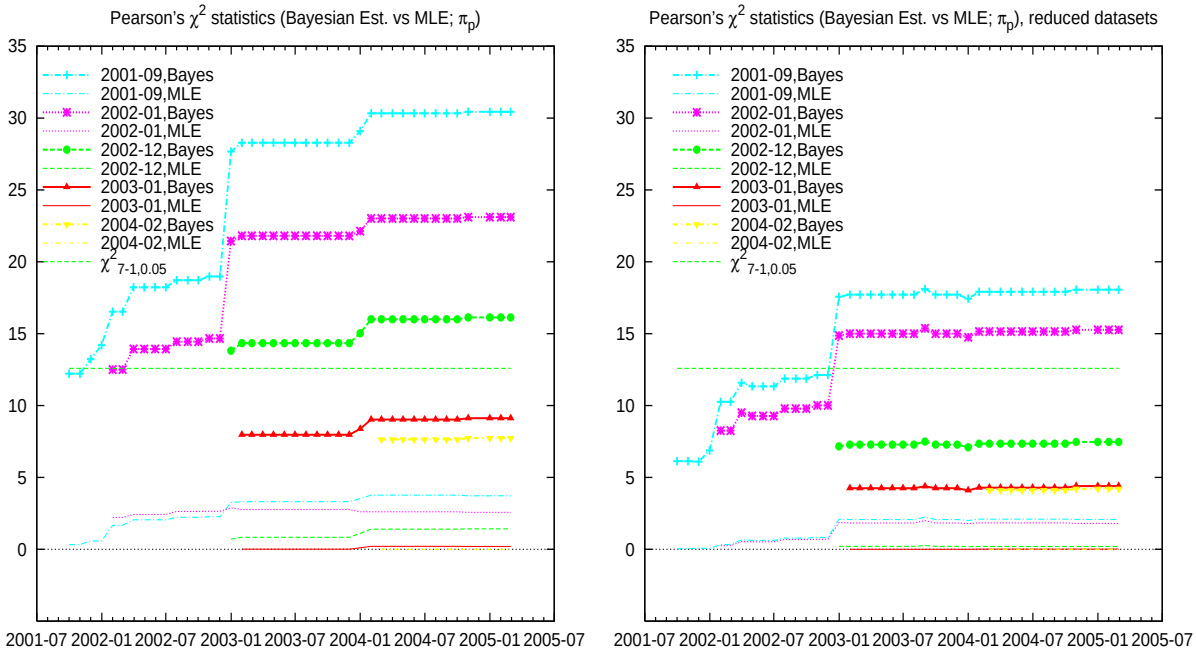


Figure 5.5: Comparison of the two methods via χ^2 statistics ($\hat{\pi}_p$)

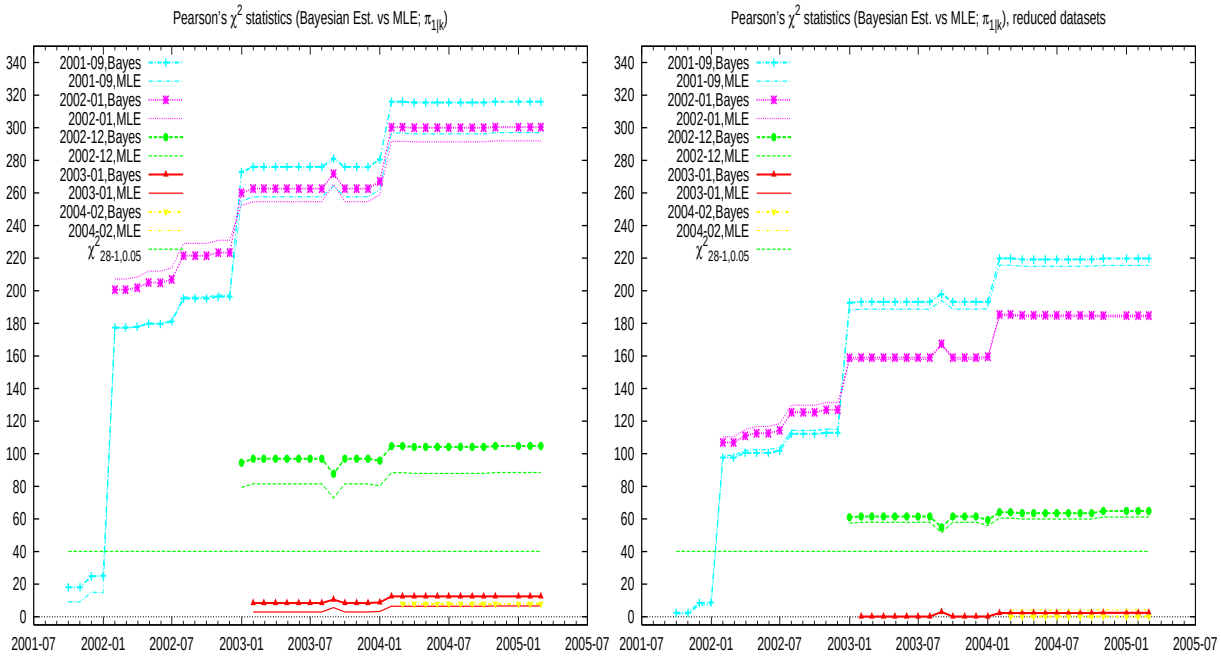


Figure 5.6: Comparison of the two methods via χ^2 statistics ($\hat{\pi}_{1|k}$)

5.4 Structure Prediction

An important step in understanding the complicated biological processes within a cell is the structural description of the protein interaction. The prediction of protein structure can also be approached using our models. Both the frequency-based approach and the odds ratio-based approach are able to provide an estimate of the probability of the class-class interaction.

Here we describe step by step the implementation of the prediction and present our prediction results.

Target proteins Given a newly identified interaction constructed by an annotated protein (existing protein) and an unannotated protein (new protein), we want to predict the structure of this unannotated protein. (see Figure 5.7)

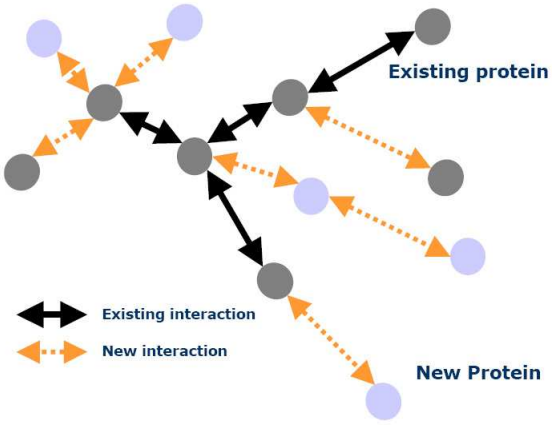


Figure 5.7: New interactions from an existing network

Within the protein networks, newly identified interactions (new interactions) from one existing protein and one new protein are selected for analysis. For those existing proteins we know their structures, while for those new ones we do not. Here, we choose proteins where the 3D structure of two interacting proteins are both known in order to evaluate the goodness of the prediction (see Figure 5.8).

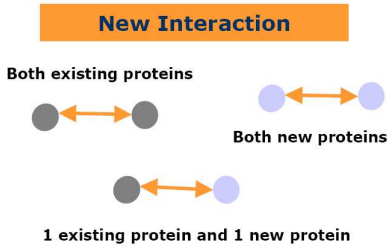


Figure 5.8: Three types of new interaction

Assigning probabilities from 1-step interacting partners Initially, proteins from any of the 7 SCOP classes are included. The probability, that a new protein is in each structural

class, is calculated based on the classes of its interacting partner(s), for which structures are already known.

In our prediction, two types of proteins are ignored.

1. Proteins isolated from the major PIN are ignored, because there are not enough proteins in the neighborhood to infer its structure.
2. The interacting proteins from unknown structure classes are excluded, because for these proteins we can not validate our prediction.

Assume that a new protein x has observed interactions with m existing proteins, $I_1, \cdots, I_m$, with structure classes, $F_{(I_1)}, \cdots, F_{(I_m)}$, where the function $F_{(.)}$ returns the structure class of a protein. Let W be the set of 7 SCOP classes. The probability that the new protein x is in class k can be estimated by

$$\hat{P}(F_{(x)} = k| \text{ 1-step interaction}) = \frac{\prod_{i=1}^{m} P(k \leftrightarrow F_{(I_i)})}{\sum_{c \in W} \prod_{i=1}^{m} P(c \leftrightarrow F_{(I_i)})} \quad .$$

The estimate of $P(F_{(I_i)} \leftrightarrow k)$ can be obtained from the $\hat{P}_k$ in the frequency-based approach or the $\hat{\pi}_{1|k}$ in the odds ratio-based approach.

Prediction of structures The prediction of protein structure is carried out by the frequency-based approach and the odds ratio-based approach. In each approach, the maximum likelihood method and the Bayesian method are implemented. The subset of DIP Yeast in 2002-01 is used to predict later datasets of yeast PPI. The DIP H.pylori is used as the prior in the Bayesian method. The aim is to predict the structure of new proteins in Yeast subsets.

Among all new interactions between existing proteins and new proteins, we select only those interactions for which we know the structure classes in both interacting proteins, so that the verification of the prediction is possible (see Figure 5.9). Finally, according to the datasets, there are 116 to 300 proteins selected for the prediction. The prediction is based on its 1-step interacting partner as described above. In both approaches, the accuracies (i.e.,the numbers of correctly predicted proteins over all proteins) are calculated.

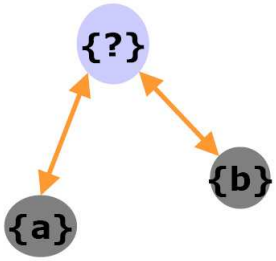


Figure 5.9: A new protein with two annotated interacting proteins

In the odds ratio-based approach, in addition to $\hat{\pi}_{1|k}$, the probability of protein from a SCOP class, $\hat{\pi}_p$, is also used to predict protein structures. A protein is randomly

assigned its SCOP class according to $\hat{\pi}_p$ estimated by two methods. After all proteins are predicted their classes, the number of correct predictions is recorded. The procedure is then replicated for 100 time for each dataset to obtain an averaged number of correct predictions.

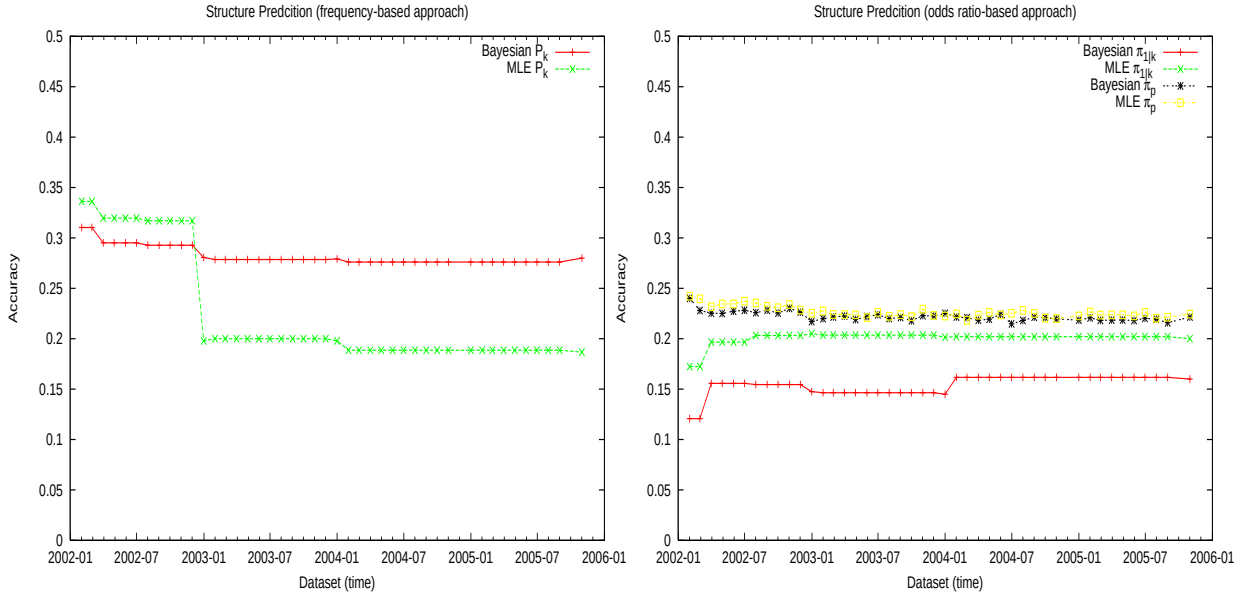


Figure 5.10: Accuracies of the structure prediction (7 SCOP classes)
The figures show the accuracies (numbers of correctly predicted proteins over all predicted proteins) given by the frequency-based approach (lefthand figure) and the odds ratio-based approach (righthand figure).

The total number of proteins to be predicted increases with time. Among two estimates used in two approaches as in Figure 5.10, the Bayesian estimate $\hat{P}_k$ in the frequency-based approach correctly predicted $\sim 30\%$ of proteins, which are about 84 correct predictions. If second chances are given, that is the wrongly predicted proteins are re-assigned their structures with the second high probabilities, it can reach 133 (44%) correctly predicted proteins totally. The Bayesian estimates $\hat{\pi}_{1|k}$ in the odds-ratio based approach does not predict well. The reason could be the use of the non-informative prior. It might indicate that the selection of prior affects the performance.

In the odds ratio-based approach, the predictions from the Bayesian estimates $\hat{\pi}_p$ and the MLE $\hat{\pi}_p$ shows, however, more correctly predicted proteins than those predicted by $\hat{\pi}_{1|k}$. The reason could be that the number of available PPI is not sufficient to provide good estimation when they are split into 28 categories, though their structure classes do help in understanding the protein interaction.

In Figure 5.11, the predictions are based on the reduced datasets that proteins from four major classes are selected. Though the total number of proteins are reduced, the accuracies are slightly increased. A reason might be that more proteins are from major classes,

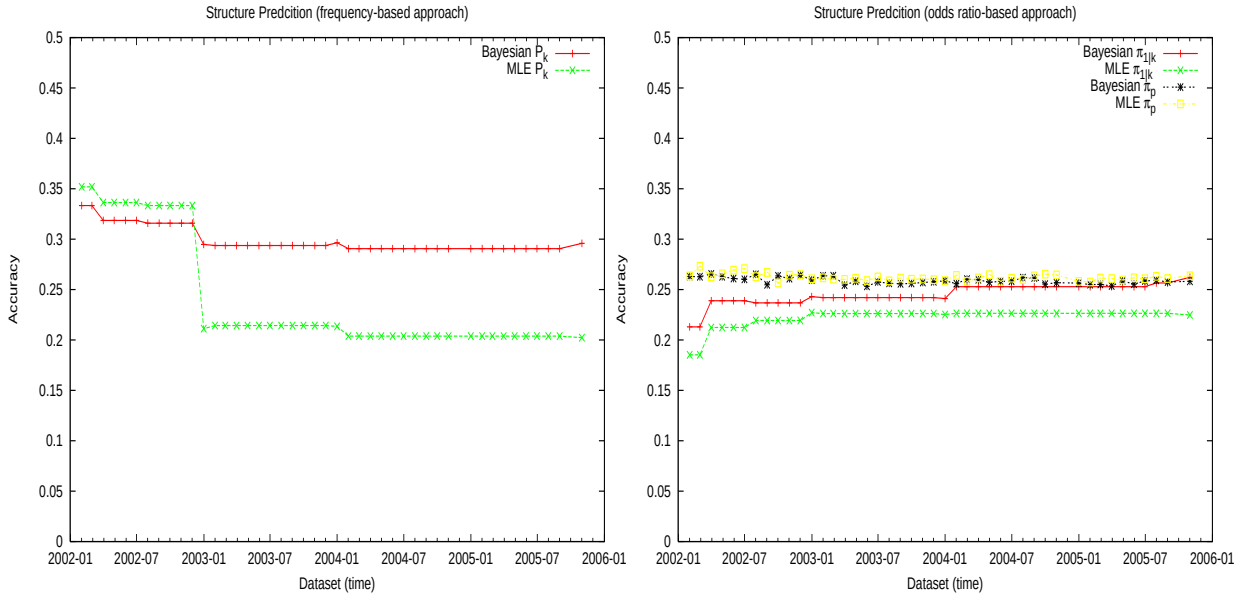


Figure 5.11: Accuracies of the structure prediction (4 SCOP classes)
The figures are from the reduced datasets with only 4 major classes, (a, b, c, d), proteins.

so that more information are provided. Whereas proteins from other classes (e, f, g) are less available, that make the prediction less accurate. Again, the Bayesian estimate $\hat{P}_k$ has better performance that its accuracy stays at 30% to 35% throughout the predictions.

In particular, as in the righthand figures the accuracy of Bayesian estimate $\hat{\pi}_{1|k}$ has been increased gradually and is slightly higher than the MLE $\hat{\pi}_{1|k}$, when the datasets are reduced (Figure 5.10 right and Figure 5.11 right). The reasons might be that the proteins in classes e, f, g in H.pylori do not help in the prediction of the Yeast PPI and are removed in the reduced datasets, and the major classes dominate most PPI. Nevertheless, the low accuracy suggests that a better model is needed.

Evaluation of the performance We can evaluate the performance using the ROC curve, which is based on the *Specificity* (SP) and the *Sensitivity* (SN).

Table 5.8: Evaluation of the performance

	Prediction	
Annotation	positive	negative
positive	TP	FN
negative	FP	TN

Specificity is the ability to reject "false positive" matches. The most specific search will return only true matches, but might have lots of false negatives. Sensitivity is the ability to detect "true positive" matches. The most sensitive search finds all true matches, but

might have lots of false positives. Therefore, the ideal ROC curve shall close to the top left corner.

The specificity and the sensitivity are calculated using the following formulas,

$$Specificity = \frac{TN}{TN + FP}$$
$$Sensitivity = \frac{TP}{TP + FN} .$$

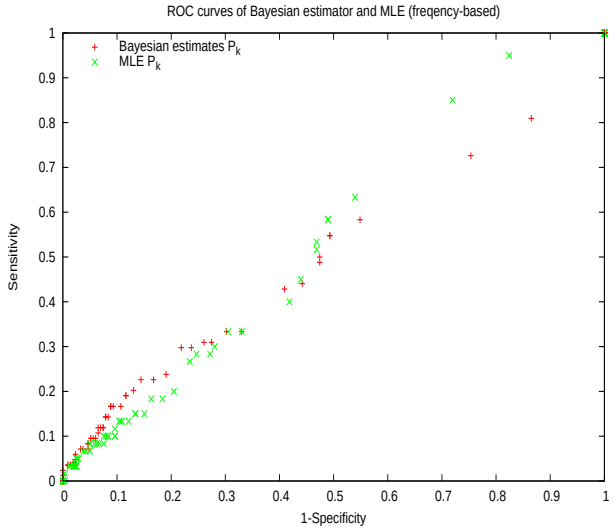


Figure 5.12: ROC curves - frequency-based

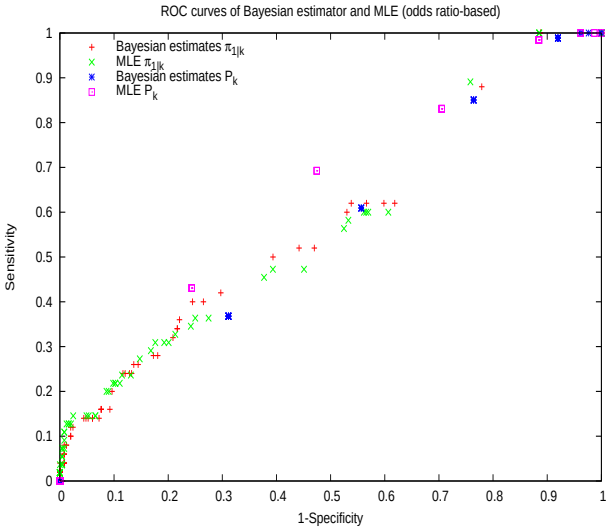


Figure 5.13: ROC curves - odds ratio-based

Unfortunately, the ROC curves from both approaches are prone to the right indicting the bad performance in the structure prediction (see Figures 5.12 and 5.13). There are several possible factors which could affect the result. Firstly, the relationship between PPI and protein structure are not yet clearly understood. Protein structures are not the only factor affects PPI. It is difficult to address their relationship well when other potential factors are not considered. Secondly, the incompleteness of PPI data might give biased predictions. The earlier PPI data will not necessarily provide useful information for the understanding of later PPI data. Insufficient of PPI data that makes the statistical analyses more difficult. Several parameters in our models are estimated only from very few observations. Lastly, multiple domain proteins are excluded from our analyses.

Then we move to the individual approach, with the frequency-based approach (Figure 5.12), the Bayesian method has a higher specificity than MLE in most of the time. The use of prior improves the prediction of the protein structures. However, different priors have to be tested.

With the odds ratio-based approach (Figure 5.13), both the Bayesian method and the maximum likelihood method perform similarly when $\pi_{1|k}$ are used for the prediction. This may be explained by the fact that a non-formative prior is used in the Bayesian model so that two estimates have similar forms (equations 4.2 & 4.5). Hence, the use of informative prior may needed to improve our model. Besides, the $\hat{\pi}_p$ seems to reach higher specificities then $\pi_{1|k}$ suggesting our models are needed to be improved.

Figures 5.14 and 5.15 show the ROC curves from the reduced datasets. In general, it is similar to the analysis using all proteins, except the specificities are improved slightly.

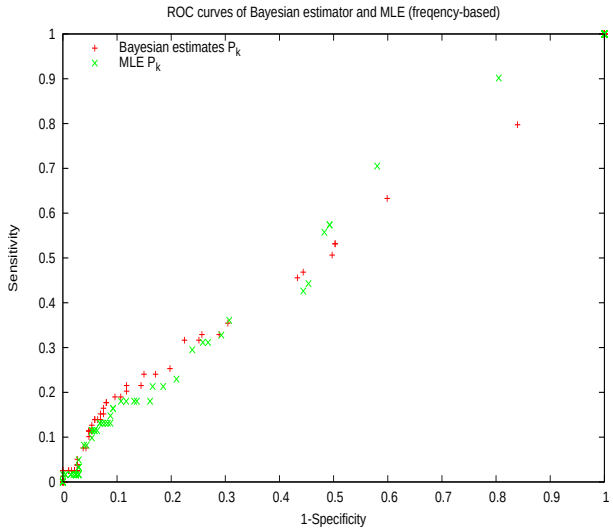


Figure 5.14: ROC curves - frequency-based

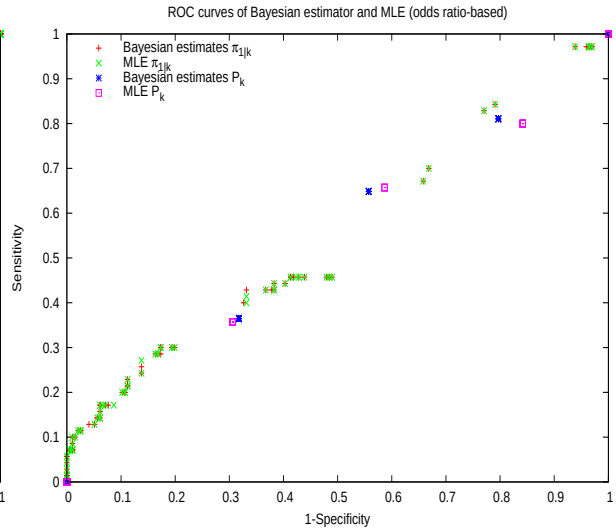


Figure 5.15: ROC curves - odds ratio-based

Chapter 6

Network Statistics

Various biological networks such as metabolic networks, gene transcriptional regulatory networks, protein folding networks, and PIN provide plenty of data for analysis. The availability of these genome-scale biological networks has enabled the analysis of their topological structures, which may allow us to answer biological questions by taking the whole picture into account [31, 41, 69].

If the proteins are nodes and the interactions are edges, PPI can be viewed as networks. Some topological properties in PIN have been observed and biological explanations have been given [27]. The *modularity* of the networks indicates proteins form clusters to achieve biological functions [21, 61]. The protein clusters appear in frequently observed patterns, *network motifs*, that work as function modules and are conserved in different organisms [56]. On the other hand, the highly connected nodes, called hub proteins, are found to show evolutionary conservation [20]. They are thought to play critical roles in communicating among clusters in networks [74].

From the macroscopic point of view, knowledge about the network will be helpful in understanding the mechanism of protein interactions. A concrete description of a network model would provide a clearer picture for protein networks. Several network models, such as the small world model and a random geometric model, have been proposed [51, 59, 72]. However, neither statistical model has been well studied for the PIN, nor does any of them explain the data properly. The difficulty comes from (1) The fact that the experimental data suffer from high experimental error [40]. (2) Different reliabilities of data make it complicated to integrate multiple data sources [11]. (3) Whether subnets of the networks maintain the same properties as the whole networks is not clear [60, 64]. The sampling bias from different sampling schemes also need to be studied.

Statistical models for networks are an important area to be studied and the application on PIN is one of the most interesting areas to be explored. At present, the following network statistics are calculated to describe the behavior of PIN [79]. The *vertex degree*, a measure of the connectivity of one node. The *clustering coefficient* of a node which characterizes the connectivity of any two interaction partners in its neighborhood. These both describe the clustering in the networks. In addition, the *shortest path length* is the smallest number of links between two selected nodes. In this report, the probabilistic models of the distribution based on the network statistics are studied as the beginning for network modelling [6]. The results provide an insight of the structure of PIN and show the potential of network structure in improving the

prediction of PPI, as described in Chapter 7.

6.1 Network Models

The topology of networks is an interesting area of mathematics. It was first developed by Leonard Euler in 1735 in the problem of "The Seven Bridges of Königsberg". In 1960, Erdös and Rényi [19] proposed a probabilistic model for random graphs, in which each of the possible edges in the network exists with a probability p. Hence, the degree distribution is a binomial distribution.

However, the random model does not appear to fit social networks and biological networks well. In the case of social networks, the common situation is that two friends of one individual tend to know each other as well, which makes the social interaction less random. An experiment on social networks claims to exhibit the famous six-degrees of separation phenomenon [42]. It says the average number of social links needed to connect two people in the United States was less than six. This special characteristic, short path links, has been observed in other networks including biological networks.

The small-world phenomenon includes two conditions [3]:

1. A small number of links between any two nodes
2. The existence of clustering

These two conditions seem contradictory in that the first one is based on the concepts from the random graph while the second one is from the ordered lattice. A small world graph is like a graph somewhere between a random graph and a fixed lattice.

In 1998, Watts and Strogatz [72] applied a random rewiring scheme to introduce randomness into a fixed lattice, called the Watts-Strogatz model. In this model, small-world networks emerge as the result of randomly rewiring a probability p of the links, i.e., the shortcuts, in the ordered lattice. The Watts-Strogatz small world model was later modified that the shortcuts are added with a probability p rather than are rewired from the fixed edges. The characteristics of the Watts-Strogatz model might explain the modularity found in the PIN and the nature of biological communication. A rigorous mathematical proof of the distribution of the shortest path lengths in the Watts-Strogatz model is given by Barbour and Reinert [6, 7]. Recently, Lin [35] provided a proof for the compound Poisson distribution of clustering coefficients [1] in the Watts-Strogatz model . Both results enable the verification of the model on real networks and statistical testing.

In some real networks such as the biological networks, there exist some highly connected nodes. The degree distribution of the nodes show a heavy tail toward the right end, resembling a power law on the log-log plot. The heavy-tailed distribution would imply an infinite variance and is named "scale-free" [63]. Scale-free networks are commonly claimed to be observed on biological interaction data including PIN [78]. However, it should be noted that the observed power law of degree distribution is based on an approximate result, rather a mathematical

[1] A different but similar definition is used in his work. Instead of the total number of edges among the neighbours, the expected number is used in the calculation of the clustering coefficient.

proof. The low coverage of experimental data implies the possibility of bias inference. Without statistical testing, we can not conclude that the observed distribution follows a power law.

6.2 Connectivity

The connectivity of a protein, its *degree*, is the total number of its interacting partners. The relative frequency versus degree, the degree distribution, are shown in Figure 6.1.

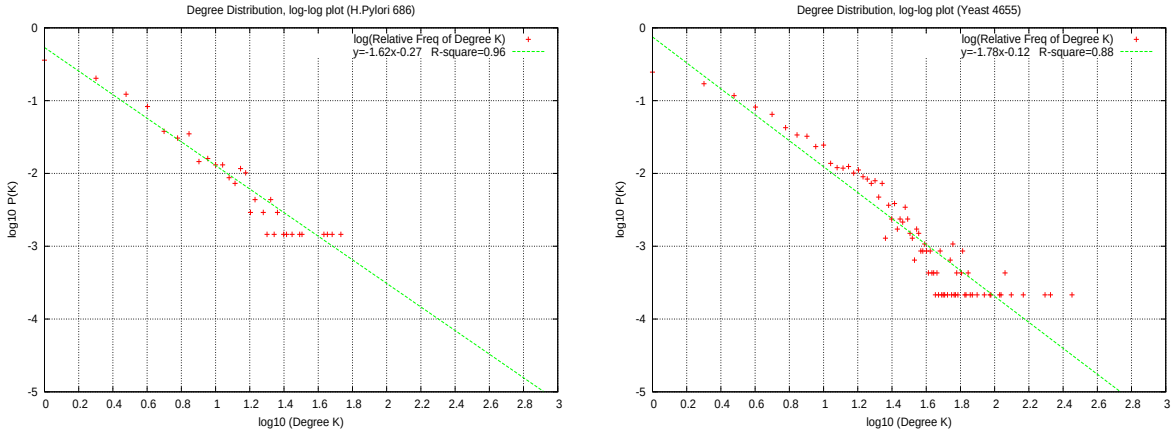


Figure 6.1: Degree Distribution in the H.pylori (left) and in the Yeast (right)
The horizontal axis is the degree (k) and the vertical axis is the relative frequency $P(k)$ on a log-log scale.

In Figure 6.1, there exist many proteins with low connectivity and also a few highly connected proteins. The existence of the proteins with many interacting partners, hub proteins, is thought to be related to the evolution of proteins, that many hub proteins are found to be old proteins [74]. They are more conserved and play important roles in biological processes.

The log-log plots of degree distributions from Yeast PIN and the H.pylori PIN fit linear regression reasonably well ($R^2 = 0.88, 0.96$, respectively). They seems to follow power-law distributions, which are characteristic for a scale-free model. However, it is not clear whether the PIN are really scale-free. The sampling bias and the fitness of power-law might mislead the statistical inference.

6.3 Clustering

Clustering is an important characteristic in the PIN in that it demonstrates the density of connection. It can be measured through the calculation of the clustering coefficient. The clustering coefficient of a vertex i, C_i is defined as the ratio between y, the number of the edges connecting all 1-step neighbours, z, and all possible edges connecting 1-step neighbours $\binom{z}{2}$ [6].

$$C_i = \frac{y}{\binom{z}{2}} = \frac{2y}{z(z-1)}$$

The statistics $C(k)$ is defined as the average of clustering coefficient for all vertices with degree k. The high average clustering coefficient C indicates the existence of loops in the networks, whereas the random graph only has a few loops.

The protein interaction network is composed by many protein complexes or functional modules. These complexes and modules are based on certain PPI to achieve particular functions, that proteins work as a team during the biological process [21, 74]. Even the size and the pattern of the module, often called motif, will affect its role in the process. It is expected that the probability of having a within-module interaction will be higher than that of a random pair [80]. In order to further understanding the networks, the distribution of the clustering coefficient against the degrees shall be helpful in describing the structure of the networks. In Figure 6.2, the low-degree proteins with high clustering coefficients suggest the local clustering in the protein interaction network.

Regarding the average clustering coefficient C of the network, we performed 100 simulations of random networks with similar sizes, i.e, same number of vertices and same number of edges, of Yeast PIN (4655 nodes and 15382 edges) and H.pylori PIN (686 nodes and 1404 edges). The average values of all average clustering coefficients from simulated random networks are 0.0014 and 0.0056, respectively, while they are 0.093 and 0.016 in Yeast PIN and H.pylori PIN. These two average clustering coefficients from PIN are also the highest comparing with those simulated random networks. This may suggests the existence of clustering in PIN, as one condition required for the Watts-Strogatz small world model.

6.4 The Shortest Path Length

For a pair of nodes in a network, there might be many paths from one to another. The shortest path length is the smallest number of links between two selected nodes. Therefore, each pair of nodes in a network has a shortest path length. The distribution of the shortest path length reflects the reaction time it requires for passing the message between two nodes. In biological networks, it is important to maintain a rapid response so that the efficient communication and the fast reaction against danger are possible. Therefore, we would expect short path lengths.

In this project, the shortest path lengths between each pair of proteins in Yeast and in H.pylori are calculated and stored. The network diameter, the longest shortest path length, is 12 in Yeast and 9 in H.Pylori.

PIN have quite small average shortest path lengths, only 4.17 in Yeast PIN (4655 proteins) and 4.13 in H.pylori PIN (686 proteins), the small average shortest path lengths in PIN meet the condition required for Watts-Strogatz small world model, too.

Hence, we are interested in comparing the PIN with the Watts-Strogatz model. The theoretical model of the approximate shortest path length distribution in the small-world network is provided by Barbour and Reinert [7] with a rigorous proof . This model is compared with the empirical distribution in PIN and the results are presented in the following graphs with statistical tests for goodness of fit.

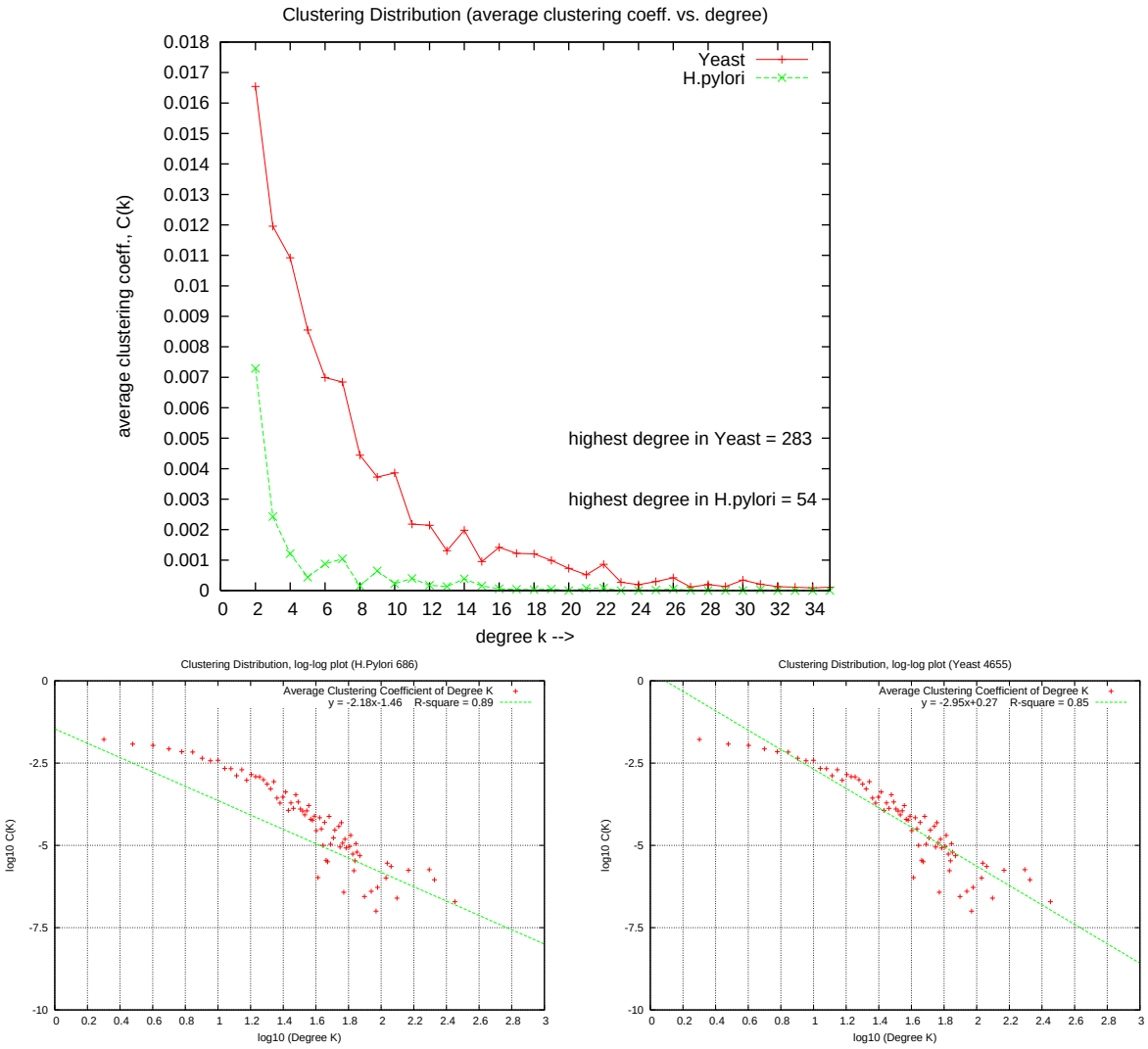


Figure 6.2: Clustering distribution of Yeast and H.pylori (top) and the log-log plot for H.pylori (bottom-left) and for Yeast (bottom-right)
The clustering distribution is presented by the degree versus the clustering coefficient. The trends on log-log plots imply the possibility of having a hierarchical structure in PIN.

6.4.1 Comparison between real data and the Watts-Strogatz small world model

Theoretical Distribution of the shortest path length *Watts-Strogatz small world model*: A circle of circumference L, vertices, includes $\binom{L}{2}$ directly or indirectly links. Pick any pair at random, uniformly calculate the shortest path length, D, between these two vertices. The expected number of the shortcuts is $(\frac{L\rho}{2})$.

Barbour and Reinert [7] provided a mathematical proof, which is a continuous approximation for the distribution of the shortest path length in the Watts-Strogatz small world model. Moreover, the bound on the approximation is given in the proof.

$$E(\hat{D}) \approx \frac{1}{\rho}\,(\frac{1}{2}\log(L\rho) + 0.2886) \tag{6.1}$$

$$\mathcal{P}(D > \frac{1}{\rho}\,(\frac{1}{2}\log(L\rho) + x)) \approx \int_0^\infty \frac{e^{-y}}{1 + e^{2x}y}\,dy \tag{6.2}$$

Parameter Estimates Before fitting the model, the parameters are needed from both the Yeast (DIP Yeast)and the H.pylori (DIP H.pylori).

1. Two datasets of H.pylori and Yeast provide their L, number of proteins, in the major interaction networks.
2. The computation on the adjacency matrices of PIN gives the shortest path length, D.
3. The parameter ρ is approximated from the equation (6.1) above.
4. Finally, the parameters in the Watts-Strogatz small world model are estimated as in Table 6.1. With these parameters, the theoretical distribution can then be obtained from the equation (6.2).

Table 6.1: Estimates of the parameters

	number of proteins L	average shortest path length D	ρ
H.pylori	686	4.137637	0.8375
Yeast	4655	4.176134	1.0907

Figures 6.3 and 6.4 show the distribution of the shortest path length from PIN in red bars and the distribution from the Watts-Strogatz small world model in the green line. In both the Yeast figure in the left and the H.pylori figure on the right, they look to fit well. Statistical tests are carried out as follows.

Tests for goodness-of-fit Statistical tests are carried out to test between the empirical and the approximate theoretical distributions in Yeast and H.pylori. The shortest path lengths between any pair of proteins in the PIN are calculated. They are summarized as 9 categories by their distances. The relative frequency in each category is then the observed frequency (O_i). In parallel, the Watts-Strogatz model in continuous distribution is also grouped into 9 categories. The relative frequency in each category is the expected frequency (E_i) under the model.

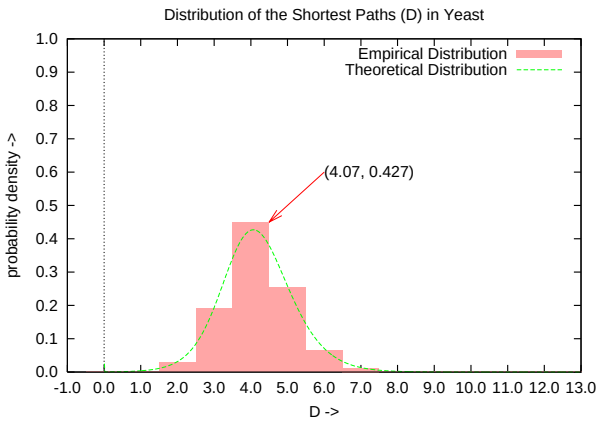


Figure 6.3: Comparison between Yeast and the small-world model

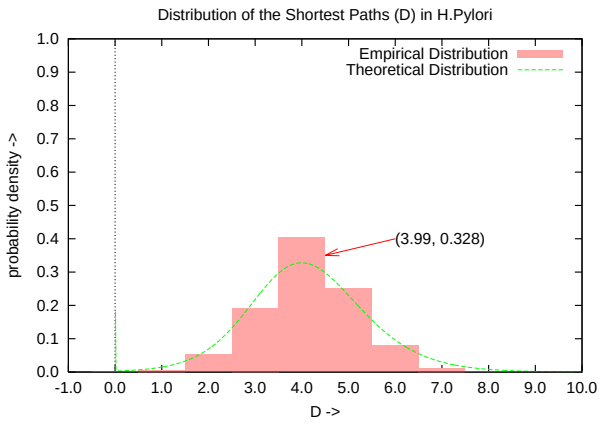


Figure 6.4: Comparison between H.pylori and the small-world model

Pearson's χ^2 test is performed to compare between O_i and E_i. There are 9 categories which give the degree of freedom $9-1$. In the test for Yeast PIN, the χ^2 value is calculated as $\sum_{i=1}^{9} \frac{(O_i-E_i)^2}{E_i} = 125505$, which is significant ($\chi^2_{8,0.05} = 15.5$). In another test for H.pylori PIN, its χ^2 value is 10257, which is also larger than the 95% significance level. On the other hand, the Kolmogorov-Smirnov tests are also carried out. The results fail to conclude that either the Yeast PIN or the H.pylori PIN has the same distribution as the Watt-Strogatz model.

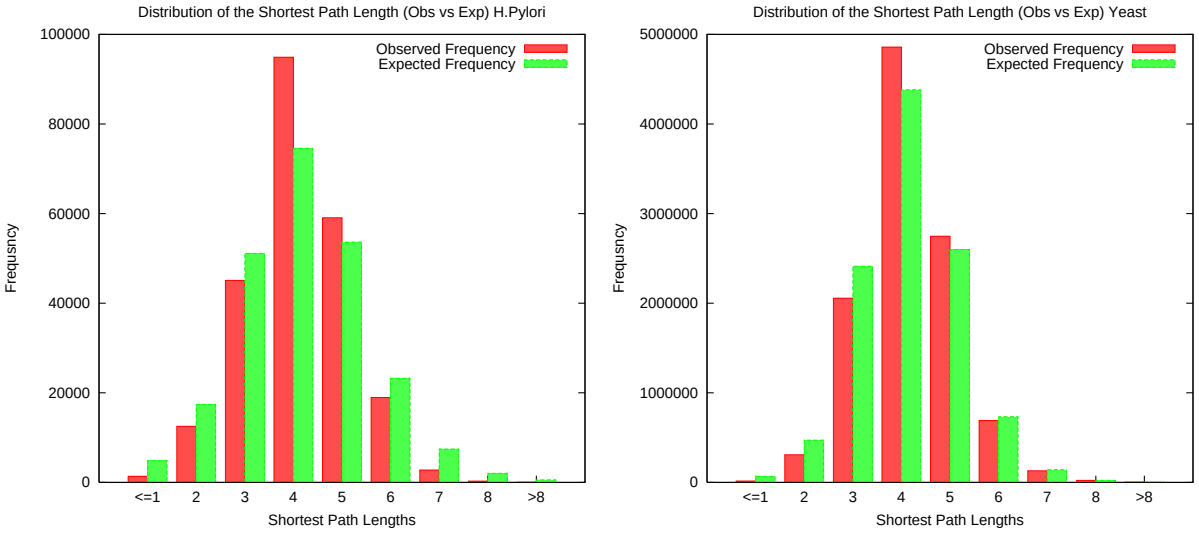


Figure 6.5: Comparison of the theoretical distribution and the empirical distribution The theoretical distributions (green bar) and the empirical distributions (red bar) are compared for Yeast (lefthand figure) and for H.pylori separately.

It appears that the results of both the χ^2 test and the Kolmogorov-Smirnov test do not support that the Watts-Strogatz small world model fits the PIN well. This might due to

the fact that the theoretical model we used is for continuous approximating distribution while the observed distribution is actually discrete. In Figure 6.5 the maximal absolute difference does not seem to be large at all and the graphs seem to indicate that the model fits quite well. It is worth further exploring the discrete model using the PIN.

In addition to the test of the shortest path length distribution, the distribution of the clustering coefficient can be used to explore the Watts-Strogatz small world model. Recently, Lin [35] has proposed a compound Poisson distribution as a theoretical model for the clustering coefficients in small world networks. It is of interest to see the goodness of fit on PIN in future work.

Chapter 7

Future Work

Based on my first-year work, there is much to be improved and even more to be explored. I believe the keys to success in prediction of protein-protein interactions are, firstly, consider the experimental error in the prediction. Secondly, multiple sources of information shall be integrated into the analysis. Lastly, a biologically realistic model shall be used as the aim is to improve our understanding in a biological issue. However, neither the proposed approaches nor other currently available methods are able to satisfy these conditions.

In order to construct a statistical model that fits the characteristics of PPI, it is apparently important that the network structure should be incorporated in the model. Meanwhile, the small-worlds properties shall be further studied to see the goodness of fits with the PIN.

From a biological aspect, it is of interest to identify the factors that influences PIN, we need to understand, more specifically, to what extent they affect PIN and how these factors interact with each other. In our approaches, through the SCOP classification applied in the model it is possible to see, roughly, the connection between the protein structure and the PPI. However, with current methods it is not easy to clarify the relationships among these factors. Enhanced statistical models should be studied and developed.

Finally, as our method has already provided a way to make use of prior knowledge for the prediction of PIN and protein structure prediction, we will try more cross-species predictions to test the method and to apply other classifications in the model.

7.1 Improvement of the model

Here I describe several aspects that shall be considered to improve our models.

Inclusion of explicit network structure The basic concept of the clustering effect is that the interacting partners of a protein are more likely to interact. The local clustering in biological networks has been identified in this report and also in many other studies. The network statistics of PPI point towards the existence of clustering, the degree distribution indicates there are a few highly connected proteins. These hub proteins form local clusters in networks and are related to the evolution. It is of interest to study the interacting pattern within a cluster as well as the inter-connection between clusters.

To date, only a few studies include the clustering effect in the statistical model for the prediction of PPI. The nature of clustering may increase the complexity in modelling. Deng et al. [17] applied the concept that proteins in the same complex have similar functions in their Markov random field model, not mentioning the explicit relationship among physical interactions within the complex. Yu et al. [79] provide a tool for investigating network statistics in the networks without further discussion of the prediction. However, to date no clustering coefficient has been included. We propose to study this in my future work.

Considering the experimental error As discussed in Chapter 2, the experimentally obtained data are reported to be seriously affected by experimental error [15, 16, 40, 43]. It is important to know how much the errors affect the model. Essentially, the false positives and the false negatives have to be estimated, which has been carried out by either empirical estimation, that is an inference from the estimate that one protein is likely to interact with 5 to 50 proteins [16], or by comparing with a positive reference dataset and a negative reference dataset [32].

The positive reference dataset can be obtained from the datasets built by highly confident methods or from the validation with multiple datasets. The subset in DIP CORE and the MIPS datasets[1] have been used as the gold standards in some reports [15, 41]. The construction of a negative reference dataset is more difficult. One possibility is to assume that proteins from different cellular locations will not interact. Other approaches include identifying the directed interactions in gene regulatory networks which identify specific protein pairs that do not interact [45].

Increasing amount of data Firstly, as more experimental PPI, such as the C. elegans, D. melanogaster, H. sapiens, become publicly available, we shall apply our method on more PPI data. These sets can be used as the prior, training data or the target data. The large amount of PPI will allow us to evaluate our method correctly and provide a direction for model improvement.

Secondly, to construct a whole picture of the PIN, it is essential to integrate more of the potentially related sources of data. The information that could help in deciphering the PPI includes the various sources of experimentally observed PPI data (with different reliabilities), correlated mRNA expression data that helps to infer PPI, the synthetic lethality of gene mutations, information about the constituent protein domains, subcellular locations, protein functions, domain structures and more. It has been shown that integrated data can improve prediction [17, 33, 56]. In previous studies, a Markov random field method and the kernel method are used to handle multiple datasets. In modelling, more efforts have to be made for considering different types of data and their reliabilities in one model. Our Bayesian approach has the advantage that it includes prior information from other sources of data in a straightforward manner, which might also allow for data integration.

Network modelling The study of network models is just in a preliminary stage. The explanation of networks statistics, the study of network motifs, the inference of unknown

[1] Munich Information Center for Protein Sequence, http://mips.gsf.de

links and the design of network models are all very exciting topics [4, 6, 77]. The study of network models may not only stimulate the new ideas for biological networks but also improve the statistical modelling from pairwise interactions to a whole network.

Network modelling may help not only in untangling the PIN, but also in analysing many other networks. It involves knowledge from probability, graph theory, statistics and also background knowledge about the domain being studied. This relatively new subfield has great potential for better understanding complicated networks in the modern world.

Application For future applications, our Bayesian model shall be improved to predict PPI; improvements include the following aspects,

1. To investigate potential factors which have effects on the physical interactions; in addition to the SCOP classification, other classifications such as the functional category can be applied in the model to see if they improve the prediction. It is also possible to combine other factors in the model.
2. To handle multi-class proteins; the current method is limited to analysing single-class proteins. Multi-domain proteins have been shown to play an important role in PPI, the Bayesian model shall be modified to analyse multi-domain proteins and to deal with proteins from multiple categories. On the other hand, it may also be possible to apply the method in detecting the docking domain pair within the protein interaction [46]. Then the multiple domain proteins can be handled as the single domain proteins.
3. To carry out cross-species prediction; the current model uses H.pylori as prior along with the Yeast data to predict the new interactions in Yeast. The Bayesian method performs well when the organism has not yet been intensively studied as it takes advantages of using the prior information from other species. Therefore, it is of interest to carry out more prediction on other species, especially on those lacking experimental data. It would also be interesting to apply different priors for prediction and to study the phylogenetic connection between the training species and the test species.

Bibliography

[1] Albert, I. and Albert, R. (2004) Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, **20** (18), 3346–52.

[2] Aloy, P. and Russell, R. B. (2003) Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, **19** (1), 161–2.

[3] Amaral, L. A. N. and Ottino, J. M. (2004) Complex networks - augmenting the framework for the study of complex systems. *European Physical Journal B*, **38** (2), 147–162.

[4] Barabasi, A. L. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5** (2), 101–13.

[5] Barbour, A. and Reinert, G. (2003a) Discrete small world networks. *Preprint*.

[6] Barbour, A. D. and Reinert, G. (2001) Small worlds. *Random Structures & Algorithms*, **19** (1), 54–74.

[7] Barbour, A. D. and Reinert, G. (2003b) Small world networks. *Extended abstract for MaPhySto and Dynstoch Workshop on Dynamical Stochastic Modelling in Biology, Copenhagen*.

[8] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004) The pfam protein families database. *Nucleic Acids Research*, **32** (Database issue), D138–41.

[9] Bauer, A. and Kuster, B. (2003) Affinity purification-mass spectrometry - powerful tools for the characterization of protein complexes. *European Journal of Biochemistry*, **270** (4), 570–578.

[10] Berg, J. and Lassig, M. (2004) Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. USA*, **101** (41), 14689–14694.

[11] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I. and Marcotte, E. M. (2004) Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, **14** (3), 292–9.

[12] Bray, D. (2003) Molecular networks: the top-down view. *Science*, **301** (5641), 1864–5.

[13] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G. and Chen, R. (2003) Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, **31** (9), 2443–50.

[14] Chou, K. C. (2000) Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci.*, **1** (2), 171–208.

[15] Deane, C. M., Salwinski, L., Xenarios, I. and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular Cell Proteomics*, **1** (5), 349–56.

[16] Deng, M., Mehta, S., Sun, F. and Chen, T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, **12** (10), 1540–8.

[17] Deng, M., Chen, T. and Sun, F. (2004) An integrated probabilistic model for functional prediction of proteins. *Journal of Computational Biology*, **11** (2-3), 463–75.

[18] Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402** (6757), 86–90.

[19] Erdos, P. and Renyi, A. (1959) On random graphs. *I. Publ. Math. Debrecen*, **6**, 290–297.

[20] Fraser, H. B. (2005) Modularity and evolutionary constraint on proteins. *Nat. Genet.*, **37** (4), 351–2.

[21] Gagneur, J., Krause, R., Bouwmeester, T. and Casari, G. (2004) Modular decomposition of protein-protein interaction networks. *Genome Biology*, **5** (8), R57.

[22] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415** (6868), 141–7.

[23] Gough, J. and Chothia, C. (2002) Superfamily: Hmms representing all proteins of known structure. scop sequence searches, alignments and genome assignments. *Nucleic Acids Research*, **30** (1), 268–72.

[24] Gray, J. J., Moughon, S. E., Kortemme, T., Schueler-Furman, O., Misura, K. M. S., Morozov, A. V. and Baker, D. (2003) Protein-protein docking predictions for the capri experiment. *Proteins-Structure Function and Genetics*, **52** (1), 118–122.

[25] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430** (6995), 88–93.

[26] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004) Intact: an open source molecular interaction database. *Nucleic Acids Research*, **32** (Database issue), D452–5.

[27] Herrgard, M. J. and Palsson, B. O. (2005) Untangling the web of functional and physical interactions in yeast. *Journal of Biology*, **4** (2), 5.

[28] Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Research*, **10** (8), 1204–1210.

[29] Huynen, M. A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. USA*, **95** (11), 5849–56.

[30] Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J. S., White, K. P. and Rzhetsky, A. (2004) Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, **20** (8), 1205–13.

[31] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98** (8), 4569–74.

[32] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003) A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302** (5644), 449–53.

[33] Kato, T., Tsuda, K. and Asai, K. (2005) Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, **21** (10), 2488–2495.

[34] Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I. and Noble, W. S. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Proceedings of the Pacific Symposium on Biocomputing*, 300–11.

[35] Lin, K. (2005) Compound poisson approximations for clustering coeffcient and the joint counts of triangles and 4-cycles in small-world networks. *Preprint*.

[36] Lin, N., Wu, B., Jansen, R., Gerstein, M. and Zhao, H. (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5** (1), 154.

[37] Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA*, **100** (21), 11980–5.

[38] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285** (5428), 751–3.

[39] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402** (6757), 83–6.

[40] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417** (6887), 399–403.

[41] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) Mips: a database for genomes and protein sequences. *Nucleic Acids Research*, **30** (1), 31–4.

[42] Milgram, S. (1967) The small world problem. *Psychology Today*, 60–67.

[43] Mrowka, R., Patzak, A. and Herzel, H. (2001) Is there a bias in proteome research? *Genome Research*, **11** (12), 1971–3.

[44] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247** (4), 536–40.

[45] Nariai, N., Tamada, Y., Imoto, S. and Miyano, S. (2005) Estimating gene regulatory networks and protein-protein interactions of saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics*, **21 Suppl 2**, ii206–ii212.

[46] Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M. and Teichmann, S. A. (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21** (7), 993–1001.

[47] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, **96** (6), 2896–901.

[48] Park, K. J. and Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19** (13), 1656–63.

[49] Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14** (9), 609–614.

[50] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, **96** (8), 4285–8.

[51] Przulj, N., Corneil, D. G. and Jurisica, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20** (18), 3508–15.

[52] Rice, J. A. (1995) *Mathematical statistics and data analysis*. Belmont, Calif: Duxbury Press, second edition.

[53] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17** (10), 1030–2.

[54] Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (2004) A structural perspective on protein-protein interactions. *Current Opinion in Structural Biology*, **14** (3), 313–24.

[55] Samanta, M. P. and Liang, S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA*, **100** (22), 12579–83.

[56] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. and Ideker, T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA*, **102** (6), 1974–9.

[57] Shen-Orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of escherichia coli. *Nat. Genet.*, **31** (1), 64–8.

[58] Smith, G. R. and Sternberg, M. J. E. (2002) Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, **12** (1), 28–35.

[59] del Sol, A., Fujihashi, H. and O'Meara, P. (2005) Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, **21** (8), 1311–5.

[60] Song, C., Havlin, S. and Makse, H. A. (2005) Self-similarity of complex networks. *Nature*, **433** (7024), 392–5.

[61] Spirin, V. and Mirny, L. A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA*, **100** (21), 12123–8.

[62] Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, **311** (4), 681–692.

[63] Strogatz, S. H. (2001) Exploring complex networks. *Nature*, **410** (6825), 268–276.

[64] Stumpf, M. P., Wiuf, C. and May, R. M. (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. USA*, **102** (12), 4221–4.

[65] Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C. and Cesareni, G. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295** (5553), 321–4.

[66] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V., Bussey, H., Andrews, B., Tyers, M. and Boone, C. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294** (5550), 2364–2368.

[67] Tsuda, K. and Noble, W. S. (2004) Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, **20 Suppl 1**, I326–I333.

[68] Tsuda, K., Shin, H. and Scholkopf, B. (2005) Fast protein classification with multiple networks. *Bioinformatics, ECCB05*.

[69] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403** (6770), 623–7.

[70] Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, **12** (3), 368–373.

[71] Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J. P., Oltvai, Z. N. and Barabasi, A. L. (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl. Acad. Sci. USA*, **101** (52), 17940–17945.

[72] Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393** (6684), 440–2.

[73] Wuchty, S. (2001) Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, **18** (9), 1694–702.

[74] Wuchty, S. (2004) Evolution and topology in the yeast protein interaction network. *Genome Research*, **14** (7), 1310–4.

[75] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. and Eisenberg, D. (2000) Dip: the database of interacting proteins. *Nucleic Acids Research*, **28** (1), 289–91.

[76] Yang, S. J. (2005) Exploring complex networks by walking on them. *Physical Review E*, **71** (1).

[77] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U. and Margalit, H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc. Natl. Acad. Sci. USA*, **101** (16), 5934–9.

[78] Yook, S. H., Oltvai, Z. N. and Barabasi, A. L. (2004) Functional and topological characterization of protein interaction networks. *Proteomics*, **4** (4), 928–42.

[79] Yu, H., Zhu, X., Greenbaum, D., Karro, J. and Gerstein, M. (2004) Topnet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Research*, **32** (1), 328–37.

[80] Zhang, L. V., King, O. D., Wong, S. L., Goldberg, D. S., Tong, A. H., Lesage, G., Andrews, B., Bussey, H., Boone, C. and Roth, F. P. (2005) Motifs, themes and thematic maps of an integrated saccharomyces cerevisiae interaction network. *Journal of Biology*, **4** (2), 6.