EFFICIENT DISCOVERY OF BINDING MOTIF PAIRS FROM PROTEIN–PROTEIN INTERACTIONS

HAIQUAN LI

(M.Engineering, Huazhong University of Science and Technology, P.R.China)

(B.Engineering, Huazhong University of Science and Technology, P.R.China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE FOR INFOCOMM RESEARCH

NATIONAL UNIVERSITY OF SINGAPORE

2006

To my parents and yuehong

Acknowledgements

I am very grateful to Dr. Jinyan Li and Associate Professor Wee Sun Lee, the supervisors in my Ph.D. candidature.

It is Jinyan who showed me the way of research. When I was upset for my work, he encouraged me and drove my worries away. Once I made some progress or discovery, he enlightened me for deeper insight. In case my work was close to publication, he reminded me the importance of presentation. His seriousness in examining results and writing skills during these moments impressed me all along. More importantly, his careful plan for my Ph.D. candidature facilitates my thesis writing in a great way.

As the principal supervisor, Professor Wee Sun Lee has given me perfect supervision on planning, progress controlling, in addition to good environments for my research and life.

Meanwhile, I would like to extend special thanks to Professor Limsoon Wong, the research director of the institute. Although he was quite busy, he didn't mind giving careful guidance and response to every research question. His guidance benefited my research both from theoretical and practical aspects, which was highly appreciated.

I would like to thank Dr. See-Kiong Ng, the department manager, for his support and valuable hints during my candidature. Beyond, I would express special appreciations to my colleagues Mr. Soon Heng Tan and Mr. Han Hao, for all the biological suggestions

and helps. Without their helps, the thesis could never be completed or probably could not be started.

There were lots of discussions and help from my colleagues in the knowledge discovery department, including those from Dr. Huiqing Liu, Donny Soh, Dr. Guimei Liu, Kelvin Sim, Judice Koh, Sundar and Guanglan Zhang. In particular, Mr. Kelvin Sim helped the polish of a chapter of the dissertation. All those are fully acknowledged.

On the personal side, I wish to thank my parents for their strong support during my Ph.D. life. They shared my happiness and pains during the long duration. I also wish to thank my wife, Yuehong, for choosing me in the difficult time and supporting me all the way. Besides, thanks to my two sisters, for their compromise for my study.

Finally, I would like to acknowledge Institute for Infocomm Research for providing me the scholarship and facilities for my research, and National University for offering me some extra fellowships and for supporting my thesis work and coursework.

Preface

The dissertation contains seven chapters together with the table of contents and the Bibliography. The first two chapters provide an introductory outline and a literature review for the dissertation. The last chapter concludes this dissertation with an overall discussion on current and future research issues. The remaining chapters cover the main research topics. The Bibliography lists all the references used in this dissertation. No part of the dissertation has ever been submitted for any degree, or ever been conducted under employment.

An expanded version of Chapter 3 and some results from Chapter 4 were published by *IEEE transactions on Knowledge and Data Engineering (TKDE)* in August, 2005. The basic ideas and results of Chapter 4 were published in the *Proceedings of the Ninth Pacific Symposium on Biocomputing (PSB)*, Hawaii, 2004. Most results of Chapter 4 were published by *Bioinformatics* in February, 2005. The whole Chapter 5 was published in the *Proceedings of the Ninth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Portugal, 2004, and an expanded version of this chapter has been submitted to TKDE. Chapter 6 was published by *Bioinformatics* in April, 2006.

Contents

Summary

Protein interaction sites mediate protein interactions in all living organisms, and they play crucial roles in drug design. Current methods to identify the interaction sites are limited by the low-throughput in the experimental approaches and insufficient structure information in protein-protein docking approaches. To break the bottleneck, this dissertation aims to define and capture signature patterns at protein interaction sites using abundant protein interaction data together with their associated sequence data. The discovered patterns at protein interaction sites are originally termed by us as *binding motif pairs*, each of which consists of two traditional protein motifs. Two methods are proposed in this dissertation to discover binding motif pairs.

The first method is based on *a fixed-point theorem*. This idea reflects biochemical stabilities exhibited in protein-protein interactions, where the stability is the resistance to some transformation under some special points, i.e., the points keep unchanged after transformation by a function. We define a point of the function as a protein motif pair. This transformation function is closely associated with a large protein interaction sequence dataset. Discovery of the fixed points (or the *stable motif pairs*) of the function is an iterative process, undergoing a chain of changing but converging patterns. Selection of the starting points for this function is difficult. We use an experimentally determined protein complex dataset (a subset of the PDB) to help identifying meaningful starting points, so that the biological evidence is enhanced and the computational complexity is relaxed. The consequent stable motif pairs are evaluated for statistical significance, using the unexpected frequency of occurrence of the motif pairs in the interaction sequence

dataset. The final stable and significant motif pairs are the binding motif pairs that we are interested in.

The second method is based on our observation that there exist frequently occurred substructures in protein interaction networks, called *interacting protein group pairs*. The properties of such substructures reveal common binding mechanism between the two protein sets attributed to the all-versus-all interaction between the two sets. We find that the problem of mining interacting protein groups can be transformed into the classic problem of mining closed patterns, a problem extensively studied in data mining. As a motif can be derived from the sequences of a protein group by standard motif discovery algorithms, thus a motif pair can be easily formed from an interacting protein group pair.

For both of the two methods, we demonstrate their effectiveness from various aspects, including random experiments, systematic validations with some reference databases, literature validations and detailed case studies. The evaluation results confirm the high efficiency and reliable effectiveness of our methods, which indicate a promising future for the usefulness of the concept of binding motif pairs.

List of Tables

List of Figures

List of Symbols

The following symbols are frequently used throughout this dissertation.

Σ	the alphabet of the 20 amino acids
	a, c, d, e, f, g, h, i, k, l, m, n, p, q, r, s, t, v, w, y or their capital letters.
$\mathcal{A}, \mathcal{B}$	a set of amino acids from Σ
$\mathcal{P}, \mathcal{Q}$	a protein: a sequence of amino acids
$\mathcal{M}$	a motif: a sequence of amino acid sets
PPr	$= \{\mathcal{P}_1, \mathcal{P}_2\}$, a protein pair
MPr	$= \{\mathcal{M}_L, \mathcal{M}_R\}$, a motif pair
$\mathbb{P}$	a protein database
$\mathbb{D}$	a sequence dataset of interacting protein pairs
f	a transformation function
$\mathbf{G}$	a protein interaction network
DB	a transaction database
C	a close pattern in DB
X, Y	main variables
β	neighborhood relation of $\mathbf{G}$
occ	the occurrence set of a pattern in DB
τ	the size threshold for some sets
x, y, z	three-dimensional coordinates

Chapter 1

Introduction

Recent development of biotechnologies has changed our view to biological science significantly. The biological data are traditionally obtained through laborious lab work and the data size is often small. However, this situation has been changed dramatically since last decades. More and more high-throughput biotechnologies have emerged which can easily produce voluminous and high-dimensional data, for example, by polymerase chain reaction (PCR, a technology for sequencing) (Mullis, 1990) or by yeast two-hybrid (a technique to assay protein-protein interactions) (Uetz et al., 2000; Ito et al., 2001). The huge amount of data is far beyond the capability of biologists to analyze them efficiently. For example, the genome project produced giga-byte data, a mount to be dizzy even for computer scientists.

This tremendous amount of data has brought up at least two challenges. The first is the extrapolation of current unbalanced information. For example, protein sequences are widely available nowadays, but their corresponding structures are often limited, as they are constrained by current protein structure determining techniques which are far behind the pace of sequencing techniques. Therefore, theoretical models or simulations of biological processes can provide a preview of future experiments and may even cut down some unnecessary experiments. Under this background, a discipline named *computational biology* has been developed:

> **Computational Biology** is the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems (Huerta and et al., 2000).

The second challenge comes from the management and analysis of the huge amount of data, especially revealing the underlying knowledge or biological mechanisms in the historic data. This leads to a new inter-discipline called *bioinformatics* mainly between molecular biology and computer science (the term first occurred in 1977).

> **Bioinformatics** is the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Since bioinformatics emphasizes the revealing of underlying mechanism from the huge amount of data, it is necessarily related to another field called *data mining* (or knowledge discovery in databases). A definition of data mining is as follows:

> **Data mining** is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Han and Kamber, 2000).

Due to the complexity and enormity of biological data, bioinformatics brings new challenges and opportunities to traditional data mining techniques, such as pattern mining, classification, clustering, Hidden Markov Model (HMM) and expectation maximization (EM).

With the rapid growth of biological data, on one hand, many geneticists, physicists and bio-chemists have been trying to study simulating and modeling problems, on another hand, many mathematicians and statisticians have been using biological data as their testbed. This makes both computational biology and bioinformatics multidisciplinary. Although the two fields are highly overlapped and can be alternatively referred in most cases, computational biology emphasizes more on simulation and modeling while bioinformatics emphasizes more on data mining and data integration. The scope of this thesis is in the field of bioinformatics.

1.1 Biology Background

1.1.1 From DNAs to Proteins

As computational biology/bioinformatics deal with the data from the field of molecular biology, this section presents a short introduction to it.

The central dogma of molecular biology is the biological mechanism that transcribes and translates from DNA to proteins. A DNA (Deoxyribonucleic acid) is a kind of macromolecules in the cells of organisms, carrying the genetic codes. It is a polymer assembled by four kinds of nucleotides (A, T, G, C as abbreviations). The four nucleotides are assembled in form of base pairs (A is with T and G is with C) in the long strand of a DNA with a double-helix structure. Therefore, DNA can be represented as a sequence consisting of four characters from one particular direction. A DNA has specified sub-structures in the strand, where the basic function unit of heredity is called a *gene*. Each gene can be transcribed independently into one or more message ribonucleic acid (mRNA). The transcribed mRNA has similar nucleotides as its original DNA, except T in the original DNA is replaced by U. Besides, mRNA becomes a single strand after transcription. Each mRNA will be translated into a protein, a basic functional unit in cells. The complete set of genes in an organism is called a *genome*. Although the genome in different cells of the same organism is identical, the expressed (transcribed and translated) set of proteins varies from one to another, which leads to the diversity of cells. The set of proteins in a cell is called *proteome* (Wilkins et al., 1996).

Proteins are the main targets in this dissertation. They are another kind of macromolecules in the cells of organisms. A protein is a polymer of twenty kinds of amino acids (or residues after polymerization). A protein has at least three levels of structures. The primary structure of a protein is the sequence of its amino acids, namely, its primary sequence. The secondary structure is the sequence of its local folding units, such as α helixes, β strands and turns. The tertiary structure includes the three-dimensional

coordinates for all atoms of every amino acid after the protein folds from the primary sequence to three-dimensional space. After folding, some parts of a protein are exposed to the outside environment, which is thus called the *surface* (Connolly, 1983) of the protein.

The surface atoms of a protein are directly related to its metabolic function. Since the location of the surface atoms is determined by the primary sequence of the protein, it is not surprising that similar protein sequences exhibit similar structures, and similar structures lead to similar functions generally. But this is not always true. Similar sequences may have quite divergent structures and similar structures may have totally different functions, owing to the crucial changes caused by the mutated amino acids or structural patches. On the other hand, totally different sequences may shape similar structures, or completely different structures may perform the same functions.

1.1.2 Protein Interactions

The functions of a protein are achieved by interacting with its partners, perhaps another protein, a peptide, a DNA, or a small compound molecule (called a ligand usually). For example, protein-DNA interactions implement the central dogma of biological systems. As another example, protein-protein interactions regulate the signal transduction, the intercellular communication and catalytic reaction. Protein-protein interactions may also be related to some diseases owing to deleterious aggregation during protein association.

In principle, protein-protein interactions accompany the formation of protein complexes (Dziembowski and Seraphin, 2004), either in the form of a permanent structure such as homo-dimers, or a transient structure such as antigen-antibody complexes, enzyme-substrate or enzyme-inhibitor complexes. The formation of protein complexes is achieved through a process named confirmation change, which is the structure change in some regions of one protein to favor the counterpart. A protein is in a free state before conformation change, and is in a bound state after confirmation change. The bound state of a protein often exists in a protein complex, either permanently, or transiently, as mentioned above.

Protein-protein interactions can be influenced by the environment in cells. Therefore, protein interaction networks of cells, consisting of all interactions between the proteins in the cells, may vary from one to another in the same organism. This contributes to the diversity of cells, besides the diversity of proteome (expressed proteins) in cells. If we ignore the chronological order and location of the protein interactions, the set of protein interactions in a species is termed as the *interactome* (Ito et al., 2001) of the species.

1.1.3 Protein Interaction Sites

Protein-protein interactions are mediated by short sequences (usually 10-20 in length) of residues (amino acids), not by the whole sequence (Sheu et al., 2005). These short sequences dominate the conformation change during protein association. The atoms in the short sequence form the contact surfaces between interacting proteins, often referred to as *interfaces* (Miller, 1990). The residues in the interfaces are termed as *interaction sites* (Evans and Levine, 1979). Generally, the residues in the interfaces contacting directly with some residues in the counterpart protein are referred as *binding sites* (Rossmann and Argos, 1978). Sometimes, interaction sites and binding sites are alternatively used if their differences are not important.

Protein interaction sites have some distinct properties compared with other residues in protein surfaces. The residues at interaction sites are often highly favorable to the counterpart residues so that they can bind together (Keskin and Nussinov, 2005). The preferences include geometric complementarity, electrostatic compatibility and hydrophobic complementarity (Gabb et al., 1997). Some interaction sites even exhibit obvious cavities (or pockets) (Edelsbrunner et al., 1996) such as hinge-like scaffolds in three-dimensional space.

There are only limited types of protein interaction sites in nature. Many interaction sites are found to be similar to others in three-dimensional structures. It can be postulated that some favorite combinations of hinges have been repeatedly applied during

evolution (Keskin and Nussinov, 2005). The set of similar interaction sites (interfaces) is called *an interaction type* (Aloy and Russell, 2004). By estimation, there are around ten thousand interaction types in biological systems (Aloy and Russell, 2004).

1.1.4 A Challenge in the Post-genome Era

Biotechnologies take a crucial role to reveal the above biological units and processes. Here we give a brief review to the current status of biotechnologies regarding above issues. With PCR techniques, many genomes have been sequenced including human genome (Roberts et al., 2001). With microarray techniques, gene expression can be assayed in vitro nowadays (Schena et al., 1995). By the protein structure initiative, a complete set of representative protein structures is being determined. The project is expected to finish in five years, with US$5K as the single unit cost and 1000 structures as annual output (Terwilliger, 2004). Although there are many other details besides sequences and structures, it is a problem of resource and time rather than the bottleneck of biotechnologies. With emerging high-throughput technologies for protein interactions such as yeast two-hybrid (Uetz et al., 2000; Ito et al., 2001), abundant interaction data are being produced. Currently, it is a problem of data quality rather than data quantity in protein interactions.

Comparing with all adventures in biotechnology, the methods to determine protein interaction sites (or protein interfaces) are still in the low-throughput stage (a more detailed review will be given in Chapter 2). As a result, only a small number of interaction sites have been determined. It is expected to take at least 20 years to accomplish all interaction types using present techniques (Aloy and Russell, 2004). Since interaction sites are crucial to many metabolic processes and protein functions, they should be challenging targets for biotechnologists in the post-genome era.

1.2 Binding Motif Pairs: Patterns at Protein Interaction Sites

Before the emergence of high-throughput experimental techniques, protein-protein docking methods, which predict complex structures based on the structures of individual proteins (Mendez et al., 2005), have dominated the prediction of interaction sites. More details are given in Chapter 2. Since only a small proportion of proteins have a solved tertiary structure, more work should be carried out to make full use of existing information such as determined protein complexes or binary protein interactions. Our work is motivated from this point.

Our idea originated from the observation that interaction sites are conserved within the same protein interaction type (Keskin et al., 2005). We propose a novel pattern to represent such conservation, with the term *binding motif pairs*. A binding motif pair consists of two traditional motifs, where a motif (most likely corresponding to some biological functions) represents a pattern on one side of interaction sites. It may have multiple formats, such as regular expression, position weighted matrix (PWM), profile, Hidden Markov Model (HMM) or even structure profile. A pair of motifs usually holds the same kind of format.

We highlight some features of the concept of binding motif pairs as follows:

- **Novel**: Although the term of *motif pairs* occurred in a few publications (Spalholz et al., 1988), it had never been presented formally and applied specifically to describe protein interaction sites or interfaces prior to our first publication in 2004 (Li et al., 2004).
- **General**: A motif pair is a general concept about the pattern at a cluster of similar interaction sites. The format of representations is not fixed, as mentioned above. They can be sequential motif pairs, or, they can also be structural motif pairs, although this dissertation does not touch too much on the structural ones.

- **Correlated between two binding motifs**: Binding motif pairs are patterns to describe interaction sites, by specifying the residue composition on the whole interaction sites. Our patterns emphasize more on the correlation between the two motifs, while the individual composition on each side is not stressed in our assumption, i.e. every motif can be a part of interaction sites, as long as it can match a partner motif.

- **Summarized a set of interaction sites**: Unlike traditional experimental and computational methods targeting on individual protein interaction sites or interfaces, motif pairs are essentially designed for representing a cluster of interaction sites. Therefore, our discovered motif pairs are able to to predict novel interaction sites or to predict protein interactions.

1.3 Organization and Main Contribution

This dissertation elaborates two distinct methods to discover binding motif pairs from different types of protein interaction data:

- Discovery of binding motif pairs in the form of regular expressions from protein interaction sequence data and protein complex structural data using a fixed point model, in Chapter 3 and Chapter 4.

- Discovery of binding motif pairs in the form of blocks (matrixes) from only protein interaction sequence data using a maximal complete bipartite (named interacting protein group pairs) model, in Chapter 5 and Chapter 6.

The organization of the dissertation and our main contribution are outlined below.

1.3.1 Organization

In Chapter 2, we give a review about the techniques to assay protein interactions and protein interaction sites. We review the experimental methods as well as computational methods that determine protein interactions, to clarify where the data come from. We also discuss the quality of the current protein interaction data, since our work focus on the data. In the remainder of the chapter, a detailed review is conducted for methods to determine protein interaction sites, to locate our research in a whole picture. Experimental methods are reviewed in the first part, including X-ray crystallography, NMR spectroscopy, phage display, mutagenesis and biochemical methods, as they are related to our validation methods. Then, computational methods are reviewed, including protein-protein docking, conservation methods such as homologous motif discovery and classification methods which learn from existing complexes or protein interacting sequences. Through the comparison with these mostly related works, the significance and necessity of our work are unveiled.

In Chapter 3, a fixed point model is introduced to discover binding motif pairs from protein interaction sequence data. This model is motivated by the stability of many biological phenomena. A point in this model is defined as a motif pair consisting of two traditional protein motifs with regular expression format. A transformation function upon any point (a motif pair) is proposed which is closely related to a protein interaction sequence dataset. Motif pairs resisted to this transformation function are defined as stable motif pairs, which are originated from other points and keep unchanged after some steps of transformation. Many interesting properties of this transformation function and the algorithmic issues related to the properties are discussed in this chapter.

The approach by the fixed point model is interesting and effective. However, it has some drawbacks such as: the difficulty to find a complete solution to identify all fixed points under this transformation from a large interaction dataset; and the statistical significance of the stable motif pairs. We address these two issues and results of our proposed solutions in Chapter 4. To tackle the first issue, we describe a heuristic algorithm

to find a special subset of such fixed points (stable motif pairs). The starting motif pairs are generalized efficiently from continuous interaction sites in a protein complex dataset, to obtain biological support. To tackle the second issue, we introduce some statistical measurements to evaluate the significance of stable motif pairs and single motifs. In the reminder of the chapter, some experiments are conducted on a yeast protein interaction dataset and a subset of protein data bank (PDB), to demonstrate the effectiveness of the heuristic approach and the statistical measurements, especially some random experiments to demonstrate the impacts of choosing different interaction sites from complexes and the impacts of choosing different starting points to derive stable motif pairs. A few of literature validations are also carried out to indicate the effectiveness of the model from another direction.

In Chapter 5, we introduce another new model for the discovery of binding motif pairs, using only protein interaction sequence data. This model is motivated from the observation that many protein interaction networks exist a kind of substructures with an *all-versus-all* or *most-versus-most* interaction between two protein-sets, termed as *interacting protein group pairs* by us. In this Chapter, we focus only on the *all-versus-all* relationship, which corresponds to *maximal complete bipartite subgraphs* in graph theory. We try to transform the mining of interacting protein group pairs from a protein interaction network into the mining of closed patterns, a problem studied extensively in data mining. More specifically, we aim to reveal the correspondence between every interacting protein group pair and a closed pattern pair in the adjacency matrix of the protein network (regarded as a graph).

Then in Chapter 6, we apply the interacting protein group pairs (maximal complete bipartite subgraphs) to discover binding motif pairs. We believe that the all-versus-all interaction between a protein group pair indicates a common binding mechanism between proteins in the pair, which belong to the same interaction type as we mentioned earlier. We extract a motif from each protein group in the pair and then form a motif pair, to represent the interaction sites shared by this interaction type.

In Chapter 7, we summarize the research results presented in the dissertation, and point out how the two approaches could be improved and what the future work are about.

1.3.2 A Brief History

At the end of 2002 when I was deciding on what research topics for my Ph.D. thesis, I was attracted by one of the projects initialized by my colleague Chris Soon Heng Tan. He intended to search for motif pairs with significant emerging values (Dong and Li, 1999). As motif pairs are usually very short, he was trying a brute-force approach [the work was published later in 2004 (Tan et al., 2004)]. The approach is vulnerable to longer motif pairs and is difficult to identify the natural length of motifs. So, we turned to examine some natural interaction sites with flexible lengths in protein complexes as they may provide clues for longer motif pairs. Soon, I formalized the interaction sites in protein complexes as *maximal contact segment pairs* and worked out the mining algorithm in Feb. 2003. Chris gave very positive feedbacks to the segment pairs I identified after conducting some literature validation. The feedback encouraged me greatly as it was my first research work. After a few months, we obtained the first set of binding motif pairs by generalizing the segment pairs with structure-similar mutants and refined them on a protein interaction dataset. The paper with some preliminary results was submitted to PSB in July, 2003 and published in January, 2004 (Li et al., 2004).

Just before the submission of the PSB paper, Dr. Ng, our lab head and a co-author of the paper, suggested us to conduct some random experiments to demonstrate the statistical significance of the discovered patterns. I felt it was quite constructive and I followed his suggestion. Some statistical measurements were studied from Sep. 2003 to Dec. 2003 and significant differences in the measurements were found between our discovered patterns and random patterns. The paper was first submitted to ISMB in January, 2004 and later to bioinformatics in March 2004 and published in February 2005 (Li and Li, 2005a).

While conducting the random experiments, I found that all random motif pairs are converged to some stable motif pairs after a few rounds of refinement (less than 7),

which made me quite puzzled. Dr. Li supposed that this might be related to fixed point phenomena in mathematics, which is also to say: under some transformation by a contract mapping function, every point will go to a fixed point in the space. Then, we studied fixed point theorems and proved that our transformation function during refinement did satisfy the property of contract mapping. Although the idea was first mentioned in the bioinformatics paper (2005), the formal description and discussion of the fixed point model was not publish until the TKDE paper [submitted in July 2004 and published in August, 2005 (Li and Li, 2005b)].

Although the fixed point model was interesting and useful, it highly depended on the limited complex data. Hereby I had a strong motivation to find a pure sequence-based approach since March 2004. By chance, I observed an interesting relationship in a protein interaction network in April 2004. It was an all-versus-all interaction between two protein sets, which was named by me as a *interacting protein group pair*. Then I worked on the mining of these group pairs, starting from studies of their properties. Many properties indicated that the problem is highly similar to the problem of mining frequent patterns. An important transformation was made in August, 2004 which led to a solution of this problem. In the following days, I spent more time for the validations of these motif pairs, by comparing them with other interaction sites, such as segment pairs and domains/domain pairs. Significant results were achieved around December, 2004.

There is an episode during which I was blocked by the validations. Dr. Li asked me to join his project about mining generators and closed patterns in October 2004. Two papers were co-authored with Dr. Li and Prof. Limsoon Wong: a PODS paper (Li et al., 2005a) and an AAAI paper (Li et al., 2006b). The research works inspired Dr. Li and me that the problem of mining interacting protein group pairs can be transformed to the mining of closed patterns. This problem transformation greatly improved the efficiency of the mining algorithm. We summarized the theoretical and practical results (mainly the validations) of the approach into a paper and submitted it to ISMB 2005, ECCB 2005 and finally to bioinformatics in August, 2005. Thanks to the critical comments from the reviewers during the long course, the paper became more and more comprehensive and

professional from the biology perspectives including motivations, data sources and results. The paper finally appeared in April, 2006 (Li et al., 2006a).

While studying on the concept of interacting protein groups, Donny Soh also gave us an idea that the relationship is very similar to *maximal complete bipartite* in graph theory. Then we studied the relationship between interacting protein group pairs, maximal complete bipartite subgraphs and closed patterns. We found a correspondence between maximal complete bipartite subgraphs and closed patterns (our interacting protein group pairs have no substantial difference with maximal complete bipartite subgraphs) in April 2005. The theoretical issues was submitted to PKDD 2005 and got published (Li et al., 2005b).

Finally, the whole picture about the discovery of binding motif pairs using both approaches is clear, which makes the dissertation coming into being, although there are still lots of future work in both approaches especially in interacting protein group pairs.

1.3.3 Main Contribution

We have made the following contribution to bioinformatics and data mining:

1. The conceptualization of binding motif pairs, to represent the conserved patterns in interaction sites at an interaction type.

2. The proposal of a fixed point model and a definition of a simple transformation function.

3. The proposal of a combined learning strategy to integrate the advantage of two types of protein interaction data.

4. The introduction of the concept of interacting protein group pairs, with promising applications in other areas such as protein function predictions.

5. The theoretical association between two distinct problems, the maximal complete bipartite subgraph listing problem and the closed pattern mining problem.

6. The proposal of a method to discover binding motif pairs using only protein interaction sequence data, which tackled the barrier for proteins with no credible structures.

7. Preliminary results about the relation between binding motifs and domains.

1.4 Significance of the Study

The significance of the study covers but is not limited to:

- Potential to predict or validate protein-protein interactions using our discovered binding motif pairs.
- Enhancement of our understanding to the mechanisms of protein-protein interactions.
- Potential to reveal more details about domain-domain interactions.
- Potential to narrow down search space in protein-protein docking.
- Promising future for drug design, with the discovered motif pairs as potential drug targets.
- Potential to extend both models to the protein-DNA interaction problem.
- Potential to work as libraries for some experiments such as phage display (Smith, 1985a), to improve the hit ratio.
- Other applications in biological processes involving binding behaviors.

Chapter 2

Literature Review

We provide a literature review in this chapter which is closely related to the topics in the dissertation. We first review the methods to produce protein-protein interaction data, including the prior assumptions behind the methods and the appropriate usage of the data, since our research mainly works on protein-protein interaction data and their associated information.

Then, we describe the framework of the methods to determine protein interaction sites, as our research work is located in the domain. Our review aims to reveal the significance of our work and to suggest approaches to validate our results.

2.1 Approaches to Determine Protein-Protein Interactions

Protein-protein interactions can be assayed in-vivo/in-vitro or predicted in-silico. We review the two aspects and summarize the current status of protein-protein interaction data.

Theoretically, any method to pinpoint protein interaction sites can be regarded as a method to determine protein-protein interactions. In this section, we focus only on those determining protein interactions but not specifying the detailed interaction sites.

2.1.1 Experimental Approaches

Traditionally, protein-protein interactions are determined in a low-throughput manner with biological, biomedical and biophysical methods. Recently, they are advanced by high-throughput methods with the capability to determine protein interactions in large scales or even in proteome-wide scales. Consequently, formidable protein-protein interaction data are produced.

Low-throughput methods

Protein-protein interactions are essentially assayed by biological methods. Many of them are based on the affinity among two or more interacting proteins, which means: if we fix one protein (or pick one up), the interacting partners will also be attached to the protein owing to the affinity (either direct or indirect) with the protein. The method is called co-immunoprecipitation if some antibodies are chosen to pick up proteins. The method is called co-purification or affinity purification if some proteins are used as baits. It is called column chromatography if a bait protein is immobilized to a column (Phizicky and Fields, 1995), as all interacting (directly or indirectly) partners will also been attached in the column. There are other affinity methods such as affinity precipitation, with similar principles. Another large category of biological methods are based on genetical linkage between interacting proteins, for instance, synthetic lethal screening (Bender and Pringle, 1991). In such methods, mutations are exerted on individual proteins or pairs of proteins, and then, the changes of phenotype (functions) are monitored to determine the interactions of the protein pairs.

Protein-protein interaction can also be revealed by biochemical methods. For example, cross-linking method (Muller et al., 2001; Vasilescu et al., 2004) links some special atomic groups of proteins through cross-linking agents. If a protein interacts with other proteins, the cross-linked sites of the protein may be changed and the interactions can be inferred through the changed cross-linked sites. Obviously, the cross-linking method is very specific and complicated.

Moreover, protein-protein interactions can be detected by biophysical methods, based on mass or electronic properties of protein segments. Examples include mass spectrometry (Figeys et al., 2001), native gel, gel overlay and gradient centrifugation.

In summary, low-throughput methods are accurate and suitable to validate high-throughput protein-protein interactions. However, they are usually highly specific (only applicable to a small set of proteins), expensive, laborious, and inefficient.

High-throughput methods

To tackle the inefficiency in low-throughput methods, numerous high-throughput methods have emerged in recent years. Among the methods, some detect physical interactions such as yeast two-hybrid and phage display method, while others only infer functional linkages (indirect interactions) among proteins such as affinity purification, protein microarray and gene expressions.

1. **Yeast Two-hybrid**

   Yeast two-hybrid is a widely applied high-throughput method first described by Fields and Song (1989). In the method, two target proteins are fused separately with two domains of a transcription factor (a DNA-binding domain and a transactivation domain). If the two proteins interact with each other, the two domains will reunify as a whole transcription factor to transcribe a report gene. The expression of the report is then revealed visually. The method was applied to yeast genome separately

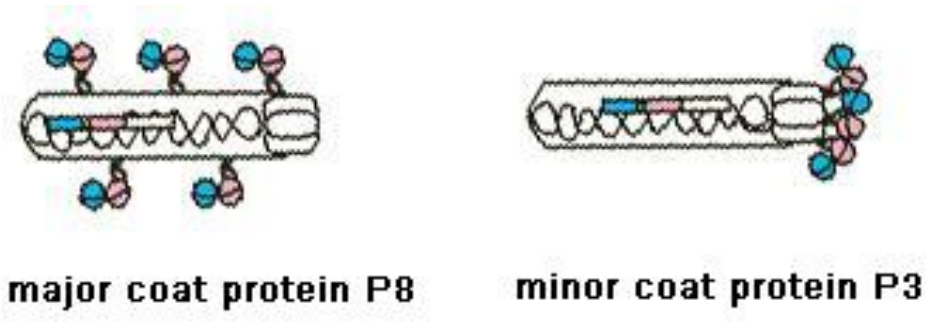


Figure 2.1: The basic principle of phage display, figure revised from (Hoogenboom and Chames, 2000).

by Uetz et al. (2000) and Ito et al. (2001). Besides those, it was applied in other genomes such as *C.elegans* (Walhout et al., 2000) and *H. pylori* (Rain et al., 2001).

Although yeast two-hybrid methods are very efficient, they are inaccurate as can be seen from the low-overlap between results from multiple experiments on the same genome (von Mering et al., 2002). The low-overlap may be caused by the transcription factor, which produced lots of false positive.

2. **Phage Display**

   Phage display was first proposed by Smith (1985a) with a phage named M13, where the phage or bacteriophage is a virus being able to infect and lyse certain bacteria. The basic characteristic of a phage is the non-exclusion of external DNAs. When an external DNA sequence is fused into the single strand DNA of the phage, the phage will not exclude the DNA. Instead, it will express the external DNA into a corresponding peptide or protein at the surface of the phage. The phenomenon is often called phage display, as demonstrated in Figure 2.1. Through the phage display, a physical linkage is constructed between genetic codes and their corresponding functions. The linkage is able to reproduce very fast via the host (bacteria) cellular machinery which is infected by the phage.

   Unlike yeast two-hybrid, which assays interactions between two proteins directly, phage display method determines protein interactions in a more indirect way. The

method immobilizes a full-length protein into a solid surface as a bait protein and mixes it with a library consisting of phages which are fused with short sequences of other proteins. If some short sequences bind to the bait protein, the corresponding phages will be attached in the solid surface and the sequences will be identified by sequencing techniques. The proteins containing the bound short sequences are thought to bind the bait protein (Sidhu et al., 2003). The protein interactions are thus determined.

For example, Tong et al. (2002) utilized the phage display method to predict interactions between proteins containing SH3-domains and proteins containing SH3-binding motifs. Most predicted interactions are confirmed by the yeast two-hybrid method.

As an indirect method, the phage display method can never guarantee the accuracy of the determination. The reason is the structure of an individual short sequence (peptide) may be quite different with that of the corresponding part in the whole protein.

3. **Mass Spectrometry**

   Mass spectrometry is a technique to separate particles with respect to distinct masses. This technique can be used to identify protein sequences. On sequence identification, a protein is first cleaved into peptides with limited types of masses, then, the whole sequence is recovered according to the composition of the mass spectrum. MALDI (Stults, 1995) is a popular method for protein cleaving and sequence recovering.

   Mass spectrometry can be applied to identify protein interactions, through the sequence identification in protein complexes (Figeys et al., 2001). Although mass spectrometry can be applied independently to determine protein interactions, it is often combined with other methods such as cross-linking (Vasilescu et al., 2004) and affinity purification, to improve the confidence of determination.

4. **Affinity Purification**

   The basic principle of affinity purification has been depicted previously in low-

throughput methods. Recently, it was combined with mass spectrometry techniques to reach high-throughput scales. Two methods, namely TAP[1] (Puig et al., 2001; Gavin et al., 2002) and HMS-PCI [2] (Ho et al., 2002) can purify protein complexes and identify the component proteins efficiently. Both methods work in proteome-wide manners and they greatly improve the purification techniques.

Although TAP and HMS-PCI produce high-throughput interaction data, they often assume full interactions among components in protein complexes. Therefore, the interactions they determined are functional interactions rather than physical ones.

5. **Protein Microarray**

   Protein microarray works similarly to DNA microarray (for in-vitro gene expressions), but the underlying principle is similar to affinity purification. Protein microarray is a matrix with each cell spotted with a particular protein. Proteins that interact (directly or indirectly) with the spotted protein will be congregated into a complex and attached in the cell. The complex are then detected automatically (MacBeath and Schreiber, 2000). The interactions produced by the method are apparently indirect ones (functional interactions).

6. **Gene Expressions**

   Gene expression methods infer protein interactions from the tremendous amount of gene expression data produced by DNA microarrays . The underlying principle is: since the interacting proteins are often involved in common functions or pathway, they should be expressed together generally. Hereby, we can predict protein interactions based on the correlated mRNA expressions. If two proteins are co-expressed in most cases, they are likely to be interacting partners (Ge et al., 2001).

[1] tandem-affinity purification

[2] high-throughput mass spectrometric protein complex identification

2.1.2 Computational Approaches

Although it is getting more sophisticated to determine protein interactions with experimental methods, computational methods take important roles all along for their inherent advantages. Historically, computational methods predicted many high-confidence interactions before the emergence of high-throughput experimental methods. Currently, computational methods not only work as necessary supplements for experimental methods, but also work as validation methods to remove false positive interactions generated by high-throughput experiments.

Computational methods can be categorized using different angles. One particular angle based on methodologies is: genome-based methods, homology methods, machine learning methods and simulation methods.

Genome-based methods

Genome-based methods compare a series of genomes and exploit underlying patterns among them. The properties examined in the genomes include occurrences, locations and similarities of genes/proteins. We elaborate four typical genome-based methods in the following.

1. **Gene order**

   The method assumes that the genes of interacting proteins should hold the same order among a set of genomes, i.e., the gene neighborhood of interacting proteins should be conserved (Dandekar et al., 1998). The assumption may be generally true for prokaryotic species such as bacteria, due to the simple rules at the start of evolution, but may not be true for eukaryotic species. Moreover, the predicted interactions should be functional linkages in pathways/processes, rather than physical interactions.

2. **Gene fusion**

   The method assumes two proteins in one genome are likely to interact if they have been fused into two domains of a protein in another genome (Marcotte et al., 1999). The reason may be the protein interaction is a favorite association and thus is reused as a basic unit by evolution. That is, interactions among domains in the same proteins may be originated from the interactions among different proteins containing these domains in the ancient era. The assumption is reasonable for both prokaryotic and eukaryotic species, but the coverage of the method is a concern.

3. **Phylogenetic profiles**

   The method assumes two proteins are likely to interact if their genes are either co-presence or co-absence in all genomes. The reason is: since interacting proteins often function together, they should appear or disappear simultaneously. The method can be regarded as an extension to the gene order method where the location of genes is extended to the co-occurrence of genes. After extension, the detection of such phylogenetic profiles can be transformed to a clustering problem (Pellegrini et al., 1999). Obviously, the interactions predicted by the method are functional linkage rather than physical interactions since indirect interactions in pathways are also present simultaneously.

4. **Phylogenetic trees**

   The method assumes that two proteins are likely to interact if they are co-mutated in all genomes, where the co-mutation means the mutation of one protein will trigger the mutation of the other protein in the same genome to maintain the interactions. This method can be regarded as a non-trivial extension to the phylogentic profile method, where the co-presence requirement in the phylogentic profile method is replaced by the co-mutation between protein sequences or protein regions (Pazos et al., 1997). After extension, the co-mutation of two proteins can be obtained through the correlation analysis between the similarity matrices of the two proteins among all genomes (Pazos and Valencia, 2001).

By analyzing the entire genome of different species, protein interactions (usually functional linkages) can be predicted with high accuracy. However, protein interactions vary from species to species, therefore, the genome-based methods are essentially weak to handle the variations. Also, the low coverage of these methods is another concern.

Homologous methods

Unlike genome-based methods, which are essentially slow due to the expensive search of multiple genomes, homologous methods infer protein interactions from homologous proteins with known interaction behaviors. Hence, they will be more efficient. The principle is: if a homologous protein of a target protein interacts with other proteins, the target protein may also interact with these other proteins. The homology in the methods may be estimated by sequences or structure similarities. With sequence similarities, they may be evaluated globally between proteins as done interolog methods (Matthews et al., 2001), or evaluated between some local regions as these regions dominate the protein interactions, for example, domains, which are widely believed to conserve interaction behaviors (Sprinzak and Margalit, 2001; Wojcik and Schachter, 2001; Deng et al., 2002; Ng et al., 2003) (we will review more details later in the chapter). In particular, we believe binding motifs are the most specific regions to dominate protein interactions compared with domains. Therefore, we discuss them and their interactions carefully in the dissertation.

Besides sequence similarities to infer protein interactions, structure similarities are more suitable to infer protein interactions, as done in multimeric threading (Lu et al., 2002). The rationale is that many similar interactions only hold among proteins with similar structures but with quite divergent sequences.

Machine learning methods

Unlike genome-based methods and homology which either work on whole genomes or homologous proteins, machine learning methods predict protein interactions by the pat-

terns learned from known interaction data. Various machining learning techniques have been applied so far. An association rule approach was presented to learn the frequent co-occurred feature sets in interacting protein pairs (Oyama et al., 2002). A number of support vector machine (SVM) methods (Joel and David, 2001) have been applied to learn important segment pairs, to distinguish positive interaction pairs from negative ones, provided that negative segments (peptides) or location of interaction sites are available. Neural networks have also been applied to the problem if time constraint is not a concern but accuracy is highly required (Fariselli et al., 2002). Note that both SVM and neural networks belong to classification methods, therefore, the reliability of negative samples is essential. Besides these methods, there are other approaches such as Bayesian networks (Jansen et al., 2003).

For all machine learning methods, the selection of features is crucial. The selected features may be the whole sequences, sequence segments or structure patches. Other biological information such as locations and functions may be utilized as features. On the other hand, multiple interaction data sources can be utilized simultaneously, such as binary interactions, protein complexes, individual structures and sequences.

Simulation methods

Without checking historical data, protein interactions can be predicted from structures of individual proteins by modeling and simulation methods, provided the structural data is available. The approach is referred as protein-protein docking (Smith and Sternberg, 2002), with more details depicted in the next section.

2.1.3 Characteristics of Protein-protein Interaction Data

From the review of the experimental and computational methods to determine protein interactions, we can see that the essential difference between the protein interaction data

and other traditional data, especially typical machine learning data. These characteristics of the protein interaction data are summarized as follows, which deserve careful consideration.

- **Negative data unavailable**

  Unlike other classification data, protein interaction experiments generally do not specify negative data, because protein interactions are determined by circumstances, which means non-interaction in one circumstance could not be extrapolated to the nonexistence of interactions in another circumstance. This characteristic makes some machine learning methods ineffective or even infeasible.

- **Inaccurate**

  Even for experimental methods, especially for high-throughput methods, the produced interactions data are not guaranteed to be accurate. An example was published in the two-hybrid method (Uetz et al., 2000; Ito et al., 2001), where only a few overlaps exist between experiments with the same method on the same proteome. This increases the difficulties for data mining methods such as classifications where the quality of the data is crucial.

- **Functional and physical interactions**

  The interaction data produced may be physical, functional or genetic. Consequently, interaction properties should be well considered before analysis.

- **Large scale**

  Protein interaction data are often enormous, from thousands to millions. The large scale data bring new challenges to data mining techniques which usually work on small scale dataset especially on a small number of dimensions, compared with protein interaction data.

- **Information in multiple levels**

  Protein interaction data has multiple levels, including binary protein interaction pairs, protein complexes without interaction details, or protein complexes with coordinates and physicochemical properties. Besides, the associated information also

has multiple levels, such as primary sequences, structures, contained domains, locations and functions. All information may be related to interactions and may be useful in the analysis.

2.2 Approaches to Determine Protein Interaction Sites

Protein interaction sites are short segments of resides which dominate protein-protein interactions in the long stretches of protein sequences (Sheu et al., 2005). Interaction sites are believed to consist of favorable structural scaffolds frequently reused by evolution (Keskin et al., 2004), therefore, they are well conserved (Pazos et al., 1997; Keskin et al., 2005) and reveal distinguishing characteristics such as accessibility distributions (Lo Conte et al., 1999; Chakrabarti and Janin, 2002).

Protein interaction sites are believed to be regulated by complicated, weak and non-covalent interactions such as Van Der Waals electrostatic forces and hydrophobic interactions. The formation of interaction sites led to the bury of large protein surfaces (1000-5000 $\dot{A}^2$) and the change of residue properties such as accessibility and hydrophobicity (Jones and Thornton, 1996).

To determine the complicated interaction sites, a handful of modern experimental and computational methods from various fields have emerged. In this section, we review the major ones, focusing only on the composition of residues but ignoring the detailed tertiary structures of the interaction sites.

2.2.1 Experimental Approaches

Recently, various techniques have been invented to determine interaction sites from multiple-disciplines, such as biochemistry, biophysics and molecular biology. We describe each one of them in the following.

Biochemical methods

Traditionally, biochemical methods were utilized to locate protein interaction sites. Multiple solvent crystal structure method (MSCS), a probe-based method, is an example. In MSCS, the X-ray crystal structures of a target protein are solved in a variety of organic solvents. Each type of solvent molecules simulates the side chain of a specific residue and works as the probe for the corresponding interaction sites in the protein. The probe distribution on the protein surface and the structures in different solvents provide the clues to locate the interaction sites of the protein and to characterize the potential ligand types (Ringe, 1995; Mattos and Ringe, 1996). For instance, the method was applied to analyze the interaction sites of elastase/inhibibitor complexes. Although the method is effective in some cases, it needs to determine a series of crystallography structures, which is essentially expensive. Furthermore, it can not guarantee the exact locations of interaction sites in the absence of the interacting partners.

It is easy to see that MSCS depends on the differences of the structures upon different solvents, but the differences may not be large enough to be observed. To tackle this problem, another method called chemical cross-linking relies directly on the complex structures. It uses cross-linking reagents which build covalent links between some specific reactive functional groups of proteins or protein complexes. The cross-linked complexes can be hydrolyzed by enzymes without breaking the interaction sites. Then, the interaction sites can be revealed comparatively from the mass spectrometric map (the mass spectrum of hydrolyzed peptides) of the cross-linked complexes and that of the non-linked complexes. For example, the method was applied to characterize the interaction sites of Op18 and tubulin (Muller et al., 2001). It is easy to see that the method requires the solubility and proteolysis of the target proteins and their complexes, which is not always applicable.

X-ray crystallography

X-ray crystallography, based on x-ray diffraction, are suitable to measure atomic structures of particles due to the intense, monochromatic and short-waved ($< 1nm$) X-rays. Recently, the technique has been widely used to solve the structures of protein complexes. Protein interaction sites are consequently determined from the structures of protein complex by examining the distance between inter-molecular residues (atoms).

Generally, x-ray crystallography contains four steps to solve the structures of protein complexes.

Step 1: Purification of protein complexes Firstly, enough number of proteins are formed and purified through DNA cloning and expression. Secondly, the purified proteins are mixed and incubated to form target protein complexes. Finally, the formed complexes are purified by removing individual proteins and other impurities. Tandem affinity purification (TAP), as described before, is such a technique to purify protein complexes efficiently (Rigaut et al., 1999; Puig et al., 2001).

Step 2: Crystallization of protein complexes In this step, crystals are formed from high concentration of purified protein complexes after the solution is volatilized. This step is very slow, varying from a couple of days to several months.

Step 3: Measurements of protein complexes and collection of data The crystals of protein complexes are cooled and mounted into the center of a diffractometer. A transfer robot will locate the crystals at liquid nitrogen temperatures. A sensitive CCD (charged coupled device) system is used to detect the scattered beams by the crystals. The detected raw data is collected finally, which is called a X-ray spectrum.

Step 4: Analysis of data The X-ray spectrum is analyzed to get the model structure of a protein complex. The constraints among atoms in the protein complex are generated and solved with the help of corresponding sequence information. The structure of the complex will be finalized with energy minimization refinement.

From above, we can see that the basic steps of protein complex determination are similar to those of individual proteins except the purification step. However, the small number of protein complexes having solved is much less than that of individual proteins due to failures in the purification and crystallization. The failures are caused by the dynamics of protein complexes especially in transient complexes and membrane-protein complexes. By May 30, 2006, less than 1350 protein complexes have been deposited in PDB, with a proportion $< 5\%$ out of 31223 solved structures by the X-ray crystallography. However, comparing to other techniques to determine protein complexes, X-ray crystallography is more productive due to the capability to tackle large molecules and the possibility to automatize.

For example, a domain interaction was solved by X-ray crystallography, which exists between two $C\epsilon 3$ domains in the Fc fragments of antibody lgE and the D1/D2 domain in the α chain of its receptor $Fc\epsilon RI$ (Garman et al., 2000).

NMR

Nuclear magnetic resonance (NMR) is also able to solve the structures of particles through the nuclear dynamics. This technique has been used to solve the structures of protein complexes by capturing the changes of parameters in NMR spectra cased by conformation changes during the protein complex association. Moreover, NMR can determine protein interaction sites before determining the whole structures of protein complexes. We depict three typical NMR methods here.

1. **Hydrogen-deuterium exchange**

   The method is based on the possible slowdown of hydrogen-deuterium exchange rate in protein interaction sites when proteins are associated with other proteins in isotope solvent D_2O. If the rate changes are large enough to be detected by 2D NMR spectra (Hoofnagle et al., 2003; Lanman and Prevelige, 2004), the corresponding interaction sites can be postulated. The method was initially applied in 1990 to

Residue	Hydrogen-bond acceptor	k_{free} (hour^{-1})	k_{bound} (hour^{-1})	k_{free}/k_{bound}
Gln^{12}	Lys^{8}	0.12	0.006	20
Phe^{36}	H_2O	0.074	<0.001	>75
Gly^{37}	Trp^{59}	0.33	0.0028	120
Arg^{38}	Leu^{35}	1.44	0.0042	340
Trp^{59}	Arg^{38}	0.065	<0.0005	>130
Lys^{60}	Thr^{63}	0.018	<0.0003	>60
Leu^{64}	Lys^{60}	0.014	<0.0003	>50
Met^{65}	Lys^{61}	0.004	<0.0005	>10
Glu^{66}	Glu^{62}	0.62	0.073	8
Tyr^{67}	Thr^{63}	0.04	0.0058	7
Lys^{100}	Ala^{96}	0.27	0.0011	250
Ala^{101}	Tyr^{97}	0.41	0.0018	230

Figure 2.2: Hydrogen-deuterium exchange rate of antigen CYT C in free state and in bound state with antibody E8 MAB, figure from (Paterson et al., 1990).

study interaction sites between protein cytochrome c (CYT c, a horse antigen) and protein E8 MAB (the antibody of CYT c) (Paterson et al., 1990). Figure 2.2 shows the exchange rate of CYT C before and after bound to E8 MAB in D_2O.

The method is very slow, taking more than 11 days in above example. It is only applicable to kinetically stable complexes with observable difference on hydrogen-deuterium exchange rates. Furthermore, the interaction sites may not be accurate and may be different with those revealed by X-ray crystallography.

2. **Chemical shift perturbation**

   The method is based on the conformation changes of side chains or backbones in protein association. The confirm changes cause parameter changes in the corresponding NMR spectrum. The residues which have obvious parameter changes (called chemical shifts) in the spectrum before and after associated with interacting partners are determined as interaction sites (see Figure 2.3 for example). The method provides a rapid approach to identify interaction residues without determining the three-dimensional structure of the protein complex, provided sequences and the residue

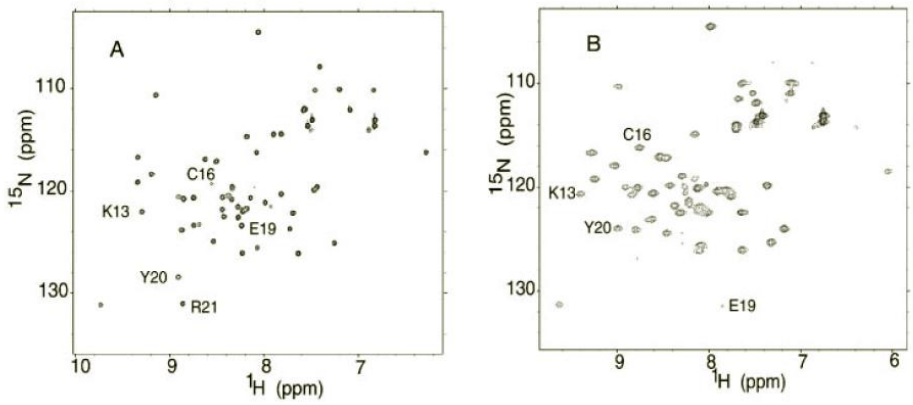


Figure 2.3: Chemical shifts in (^{15}N,^{1}H)-HSQC spectra of proteinase avian ovomucoid third domain when bound with bovine chymotrypsin Aα, figure from (Song and Markley, 2001).

assignments in the 2D spectra of the target proteins are known before hand. The method has been applied in a handful of works (Swanson et al., 1995; McKay et al., 1998; Song and Markley, 2001). However, its accuracy is still a problem owing to the chemical shifts caused by uncertain factors (Takahashi et al., 2000).

3. **Cross-saturation**

   The principle of the method is illustrated in Figure 2.4 (Nakanishi et al., 2002). The target protein with interaction sites to be examined is uniformly labeled with ^{2}H and ^{15}N and then is mixed with another unlabeled protein. The labeled (target) protein should have lower proton density than the unlabeled protein. When a radio frequency (RF) field irradiates non-selectively into the unlabeled protein, its protons will be saturated (from a lower energy level to a high energy level). Since the proton density in the unlabeled protein is higher than that in the labeled protein, the saturation will transfer to the labeled protein by cross-relaxation. If the proton density in the labeled protein is low enough, the cross-saturation will only be limited at the interaction sites between the two proteins. Thus, the cross-saturated interaction sites can be revealed by 2D HSQC spectra.

   This method is recently applied to analyze the interaction sites between B domain

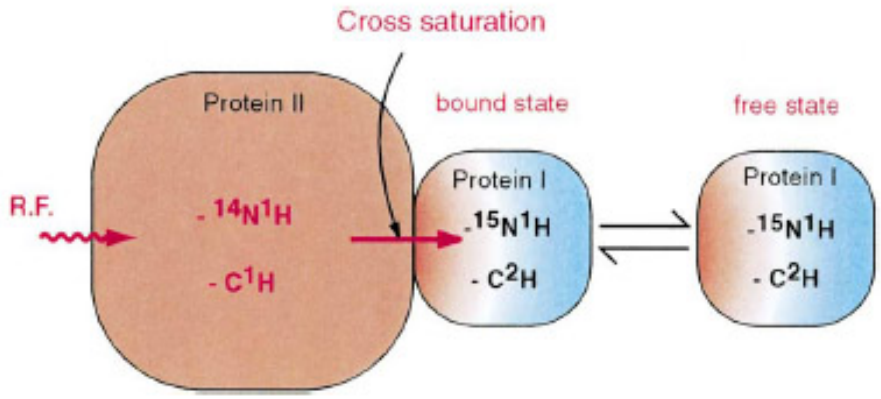


Figure 2.4: The principle of cross-saturation, figure from (Nakanishi et al., 2002).

of protein A (FB) and Fc fragment of immunoglobubin(lg) G. The determined interaction sites match well with those revealed by X-ray crystallography. This method is suitable to tackle protein complexes with large interface ($M_r > 50,000$) and large molecular weight ($> 50kDa$) (Takahashi et al., 2000; Shimada, 2005), but replies on the unbalanced proton densities between the two interacting proteins. The interaction sites of proteins with higher proton density can not be measured by this method.

Despite of various NMR methods, the solved interaction sites are still very limited so far (Wand and Englander, 1996; Nietlispach et al., 2004). The reason may be that NMR techniques can only handle soluble protein complexes with small molecular weight, which is a considerable limitation.

Phage display

Unlike X-ray crystallography and NMR, which locate interaction sites directly from interacting protein pairs (or complexes), bacteriophage (phage) display identifies interaction sites in an indirect way. The basic principle of phage display to identify interaction sites is the same as that to determine protein interactions, which was addressed previously in Section 2.1.1.

With phage display, a protein is first immobilized into a column and the candidate peptides (either from another protein or from a library) are screened. The binding peptides, which passed the screening, correspond to the interaction sites or the variants. The method has been applied to study affinity antibodies of antigens (Kretzschmar and von Ruden, 2002), inhibitors or substrates of enzymes (Fernandez-Gacio et al., 2003), epitopes (the energy contributors) or active sites (the functionally crucial sites) of interaction sites.

As mentioned, phage display can not locate interaction sites accurately from the interacting proteins, because the mimic peptides may exhibit quite different structures with the corresponding parts in the proteins. But phage display can tackle large interaction sites in both extracellular and intracellular interactions, which can not be achieved easily by X-ray crystallography and NMR. Moreover, phage display can pinpoint the most crucial residues and discover more affinity mutants for existing interaction sites.

Mutagenesis

Natural mutagenesis phenomena are the mutation of residues in proteins. Although they were discovered some decades ago, they are rare and are difficult to control. Alanine scanning mutagenesis was developed later (Cunningham and Wells, 1989), which replaces the residues with alanine one by one for the generality of alanine. The replaced residues which cause significant reduction of binding affinity correspond to the interaction sites for the irreplaceability (Cunningham and Wells, 1989). Another technique is called site-directed mutagenesis, invented in 1985 (Smith, 1985b), which replaces a residue with any other residues rather than only alanine. Site-directed mutagenesis not only identifies interaction sites, but also points out the substitutions.

Application examples of mutagenesis include studying the active sites of enzymes (Wagner and Benkovic, 1990) and analyzing the dynamics of interaction sites (Clemmons, 2001). Mutagenesis can also be combined with phage display or X-ray crystallography to get a comprehensive view about interaction sites (Sidhu et al., 2003).

Summary of experimental approaches

Experiment methods have been developed separately but they function complementarily, rather than competitively (Brunger, 1997). X-ray crystallography needs crystallizable protein complexes with adequate molecular sizes, while NMR prefers soluble complexes with small molecular weights [up to 50 kDa (Shimada, 2005)]. X-ray crystallography provides detailed tertiary structures but NMR provides more information about molecular dynamics. Similarly, phage display and mutagenesis often work together to identify the affinity variants of interaction sites which are determined by X-ray crystallography or NMR.

Although experimental methods are quite different from each other from technical perspectives, they share some common issues. Most methods require highly concentrated and purified proteins or protein complexes (Kellogg DR, 2002), and only a few of them work with impurities. Most methods need identification techniques such as labeling atoms (Kainosho, 1997). Finally, most methods involve post-analysis to the limited data obtained from the experiments.

In summary, the current experimental methods are still preliminary. Less than 2000 types of interaction sites are determined so far, with a proportion less than $\sim 20\%$ out of the total estimated ten thousand interaction types (Aloy and Russell, 2004). It would take more than 20 years to acquire a full representative set according to the present rate of experimental determination techniques (Dziembowski and Seraphin, 2004).

2.2.2 Computational Approaches

Since current experimental techniques to determine interaction sites are in low-throughput manner, expensive and inefficient (thousand US dollars and several months per site), it is valuable to predict interaction sites or protein complex structures using computational

methods. The predicted interaction sites can narrow down the search space for the expensive experimental methods, or work as putative interaction sites in applications such as the docking-based drug design.

Taxonomy of computational methods

Computational methods can be categorized using different angles. By the data source they depended on, computational methods can be categorized into complex-based, structure-based, sequence-based and combined methods. The dependent data may be experimental ones, or predicted ones, especially for structure data. More specifically, complex-based methods mainly work on existing experimental protein complex data. Structure-based methods work on individual protein structure data, both for experimental ones and predicted ones, as done in protein-protein docking (Halperin et al., 2002). Sequence-based methods work only on high accumulated non-structure data such as protein binary interaction data (Uetz et al., 2000; Ito et al., 2001) and their associated sequence data, generated by high-throughput techniques. Combined methods work on multiple data sources simultaneously, including our method based on the fixed point model (introduced in Chapter 3 and Chapter 4).

By the methodologies they used, computational methods can be roughly categorized into simulation methods and knowledge-based methods. Simulation methods simulate or model the protein-protein interactions and their interaction sites from biological, biochemical or biophysical perspectives, usually only taking individual proteins into consideration, as done by protein-protein docking. Knowledge-based methods learn from large set of historic interaction data and induce rules for their interaction sites. They can be further divided into classification methods and conservation methods, based on the requirements of negative data. Classification methods search the discriminative features of interaction sites such as linear or spatial characteristics from both positive and negative data. Since negative data are not always credible and available, conservation methods search for conserved patterns only from a set of related proteins or interactions, as done by homologous

methods and the methods in this dissertation. In the following description, we focus on this categorization and if necessary, we also specify the data sources they mainly used.

Simulation methods: protein-protein docking

Docking is a typical simulation method, which predicts complex structures based on individual structures. Traditional docking was developed some decades ago and contributed to the drug industries greatly, through searching for small molecule-like compounds as drugs (Halperin et al., 2002). Protein-protein docking was proposed in 1978 (Wodak and Janin, 1978), to face the challenge of large and flat interfaces (from 800 to 5000 $\dot{A}^2$) among proteins (Peters et al., 1996). The techniques were ignored for some time until the launch of protein-protein docking benchmark databases (Chen et al., 2003) and critical assessment of predicted interactions (CAPRI) (Janin, 2001). Benchmark databases consist of representatives of published complexes (Chen et al., 2003). Current version (2.0) includes 72 unbound-unbound complexes and their component proteins (Mintseris et al., 2005). CAPRI is a blind test, in which the target complexes are unpublished and used to test the quality of predictions (Janin et al., 2003; Mendez et al., 2003). At least 5 rounds have been undertaken and significant progress has been made according to the report by Mendez et al. (2005). The report shows enzyme-inhibitor complexes and antigen-antibody complexes are predictable but transient complexes involved in signal transduction are extremely difficult to handle owing to the substantial conformation change during association (Vajda and Camacho, 2004; Vajda, 2005).

Protein-protein docking simulates the confirmation change such as side-chain and backbone movement (Ehrlich et al., 2005) in the contact surfaces (Connolly, 1983) when proteins are associated into complexes. Most docking methods assume the final conformation change stops at the point of minimal free energy. Thus, the definition of free energy is crucial for the docking methods. Shape complementarity is the major factor or even the only factor (Peters et al., 1996) in the energy function due to the ubiquitous existence in protein interfaces. Electrostatic complementarity is another contributor to free energy,

with less importance compared with shape complementarity (Heifetz et al., 2002). Biochemical compatibility such as hydrophobic complementarity is one more factor in free energy, as it is energetically favorable in binding (Berchanski et al., 2004). Note that these factors are often combined in free energy functions, such as the geometric-electrostatic combined factor in SCHem (Fernandez-Recio et al., 2002), the geometric-hydrophobic factor or the geometric-electrostatic-hydrophobic factor (Gabb et al., 1997; Chen and Weng, 2002) in ClusPro (Comeau et al., 2004).

The search of global minimal free energy for protein-protein docking is extremely challenging owing to the huge search space caused by various flexibility. The search is roughly categorized into four steps. In the first step, one protein is fixed and the other protein is superimposed into the fixed protein to find the best docking position, evaluated in each candidate position through translation and rotation. Most algorithms apply a grid-body strategy at the step, without scaling and distortion of any part of the protein. There are a couple of strategies to reduce the huge search space in this step, including Monte Carlo Minimization sampling (MCM) as in RosettaDock (Schueler-Furman et al., 2005), Fast Fourier transformation (FFT) (Katchalski-Katzir et al., 1992) as in 3DDock (Carter et al., 2005) and ClusPro (Comeau et al., 2004), Pseudo-Brownian dynamics as in ICM-DISCO (Fernandez-Recio et al., 2003), molecular dynamics as in HADDOCK (Dominguez et al., 2003) and reduced model as in ATTACK (Zacharias, 2005). In the second step, the flexibility of side chains (such as torsion angles) is considered, as done in 3DDOCK (Carter et al., 2005), ATTRACK (Zacharias, 2005), ClusPro (Comeau et al., 2004), ICM-DISCO (Fernandez-Recio et al., 2003) and RosettaDock (Schueler-Furman et al., 2005). A few of algorithms also consider backbone flexibility using techniques such as principal component analysis as done in HADDOCK (Dominguez et al., 2003), ZDOCK+RDOCK (Wiehe et al., 2005), PatchDock (Schneidman-Duhovny et al., 2005) and ED-Hex (Mustard and Ritchie, 2005). Consequently, a series of solutions with different local minimum are produced after the first two steps. These solutions are clustered in the third step and representatives are selected. In the fourth step, re-evaluation is conducted, to improve the ranks for nearly native solutions, for the mismatch between

highly ranked solutions and native solutions owing to the limitation of score functions and search algorithms, as done in Berchanski et al. (2004), ICM-DISCO (Fernandez-Recio et al., 2003) and Schem (Fernandez-Recio et al., 2002). Note that the search of minimal free energies is influenced by the initial states, so unbound proteins are more difficult to handle compared with bound proteins. Also note that in all steps, biological information can be merged to aid the searching, such as binding sites (Carter et al., 2005), mutagenesis data and chemical shift perturbation data (Dominguez et al., 2003).

Return back to the interaction site problem, without the guidance of binding sites in the docking, the top-ranked interfaces in the final step are the predicted interaction sites; with the guidance of binding sites, the following steps may be dominated by the binding sites, thus the docking algorithms may not contribute remarkably to the prediction of interaction sites.

Although protein-protein docking is the dominant approach to predict protein interaction sites, the number of current experimentally determined structures are much less than that of sequences. Even using predicted structures, 40% proteins could not be modeled for putative structures (Aloy et al., 2005). This leaves a critical gap in the docking approach.

Classification methods

Classification methods assume that the features (sequence or spatial patches) in proteins distinguish positive protein interactions from negative protein interactions. Therefore, they correspond to protein interaction sites. The assumption holds generally but not always. The data sources used in classifications span protein sequences, structures and complexes.

A crucial issue in classification is the encoding of features. There are mainly two encoding methods. One encodes continuous residues with their associated physicochemical properties in the primary sequences (Joel and David, 2001; Ofran and Rost, 2003; Yan

et al., 2004). The other encodes a central residue and several spatially nearest neighbors, often named patches (Jones and Thornton, 1997; Zhou and Shan, 2001; Fariselli et al., 2002). The latter encoding is more accurate because structures are more important for interaction sites. Besides, the encoding of class labels in classification is another issue. It is trivial to generate class labels for binary interactions, but it is non-trivial for protein complexes. Usually, one property of the central residue in a patch is encoded as the class label in protein complexes.

Support vector machine (SVM) (Joel and David, 2001) and neural network are two traditional classification methods to predict interaction sites (Zhou and Shan, 2001; Fariselli et al., 2002; Ofran and Rost, 2003). Recently, a two-stage method was proposed. During the learning phase, both SVM and Bayesian network produce a model for the encoded continuous residues. While prediction, SVM model is first applied to predict a value, then Bayesian model is applied to predict the final value based on the predicted values in SVM model, exploiting the fact that interfacial residues tend to form clusters (Yan et al., 2004).

Although classification methods have many advantages, such as the suitability to handle transient complexes which is tough in docking (Ofran and Rost, 2003), they have several disadvantages. The first disadvantage is the unavailability and low quality of negative data. The second disadvantage is that many algorithms apply fixed-length windows to fit the basic requirement in classification, which conflicts the fact that many interaction sites have variable lengths. The third disadvantage is the incomprehensibility owing to the complicated encoding (Joel and David, 2001). The final disadvantage is the errors caused by different bound states in the training and the prediction.

Conservation methods

Conservation methods assume interaction site are highly conserved (frequently occurred) in proteomic data, including interface data, homologous protein data and protein in-

teraction data. Thus, the patterns of interaction sites are quite different from random expectation and can be revealed even in the absence of negative proteomic data.

1. **Analyzing the conserved patterns at protein interfaces**

   Interaction sites can be predicted by analyzing the characteristics such as residue conservation levels and hydrophobicity distribution in existing protein interfaces. Keskin et al. (2004) clustered the existing experimentally determined complexes and found many conserved patterns at protein interfaces, such as distribution rules of different conservation levels in the hot spot residues (the energy contributed residues in interaction sites) Keskin et al. (2005). Ma et al. (2003) also found similar rules and claimed that the hot spots can distinguish interfacial residues from other surface residues by their structural conservation. As the characteristics of hydrophobicity distribution in primary sequences, Gallet et al. (2000) analyzed the existing protein complexes and found them could predict interaction sites.

   Although these methods are reliable, the current protein complexes are very limited. Therefore, we should use other types of proteomic data which are abundantly available, such as sequence data and binary interaction data.

2. **Searching for motifs in homologous proteins**

   Homologous proteins share a common ancestor in the evolution process. They are likely to inherit some important properties from the common ancestor such as folding and binding mechanisms. On the contrary, given a group of homologous proteins [often obtained from biology evidences or estimated by sequence similarity ($> 30\%$)], the inherited properties can be recovered by searching the local conserved patterns (called motifs) from the homologous proteins, which is often referred as motif discovery.

   Before searching for motifs, the prior knowledge within the homologous proteins should be aware. There are three cases generally. In the first case, the locations of motifs are known, but the detailed patterns are unknown, such as the binding peptides of SH3 domains produced by phage display (Tong et al., 2002). EMO-

TIF (Nevill-Manning et al., 1998) is efficient to handle this case. In the second case, the rough locations of motifs are known, but the detailed positions and patterns are unknown, such as proteins containing a common domain. In the last case, no prior knowledge except homology is known in the proteins.

Besides, the representations of motifs also influence the motif discovery. They may vary from deterministic patterns to statistical patterns, including consensus sequences in regular expressions, gapped alignments, blocks/weight matrices, templates/profiles, Bayesian networks and even HMM models (Brazma et al., 1998).

Motif discovery is a NP-hard problem, so the search methods are heuristic ones. Most methods work on primary sequences and they can be roughly categorized into pattern-driven, sequence-driven and combined methods. Pattern-driven methods enumerate and test all possible motifs with a special format (often in pre-set ranges) and output the ones with enough occurrences. For instance, MOTIF (Smith et al., 1990) searches all frequent motifs with the format $a_1 - x(d_1) - a_2 - x(d_2) - a_3$, where a_1, a_2, a_3 are residues on three fixed positions and d_1 and d_2 are constant spacings (under pre-set ranges) between them; Pratt (Jonassen et al., 1995) searches regular patterns with ambiguous positions and flexible spacings. Sequence-driven methods restrict the candidate patterns to have at least some occurrences in the group of sequences. For example, CULSTAL (Higgins and Sharp, 1988) searches multiple sequence alignments from a phylogenetic tree built from pairwise sequence comparisons and hierarchical clustering. Combined methods integrate the strengths of pattern-driven and sequence-driven methods. For example, PROTOMAT extends and merges motifs found by MOTIF to form longer motifs, evaluating the extended ones by their corresponding occurrences, namely, blocks (Henikoff and Heinikoff, 1991; Jonassen, 1997); TEIRESIAS merges motifs with common sub-motifs and connects them with a graph based approach (Rigoutsos and Floratos, 1998). Lastly, other approaches are also possible, for example, statistical models such as Hidden Markov models (HMM) and expectation maximization (EM) models (Bailey and Elkan, 1995).

Motifs can also be represented by the structure similarities (called structure motifs) searched from the homologous proteins/peptides (Leibowitz et al., 2001). The problem of structural motif discovery is studied from various perspectives, such as multiple structure alignment (Shatsky et al., 2004; Lupyan et al., 2005), structure classification (Hadley and Jones, 1999), frequent common substructure mining (Leibowitz et al., 2001; Huan et al., 2004; Yan et al., 2005), or spatial sequence search (Jonassen et al., 2001), stimulated by the rapid growth of determined structures.

Motif discovery generally identifies only single motifs without specifying their interacting patterns. To study the impacts in interaction sites, we can randomly pair them and evaluate their correlation in a protein interaction dataset, which is done by Wang et al. (2005) with an expectation maximization (EM) method. However, neither sequence similarity search nor structure similarity search can guarantee the discovered motifs are binding motifs at interaction sites, because binding and folding are often interrelated and they could not be distinguished only from homologous proteins (Kumar et al., 2000). The discovered motifs are more likely to be folding motifs rather than binding motifs because homologous proteins share more folding regions than bindings regions. To identify the binding motifs, protein interaction information should be taken into consideration in the early stage of learning.

3. **Inferring domain-domain interacting pairs**

   Protein interaction sites are very related to a biological concept known as domains, which are well conserved regions in homologous proteins and believed to involve many biological processes and functions. Many interaction sites completely reside in a single domain. Therefore, inferring domain-domain interactions from protein interactions have been widely studied in recent years, especially those between Pfam domains (Sonnhammer et al., 1997).

   We glimpse at some well-known works in this area. Experimentally, iPfam (Finn et al., 2005) and 3DID (Stein et al., 2005) collect credible domain-domain interacting pairs from structural database PDB. In particular, 3DID also covers domain pairs within proteins, not limited to those between proteins. Computationally, Sprinzak

and Margalit (2001) extracted all domain pairs over-represented (having much larger occurrence than expected) in protein interaction data and termed them as correlated sequence-signatures initially; Wojcik and Schachter (2001) generated interacting domain pairs from protein cluster pairs with enough interactions, where the protein clusters are formed by proteins with enough sequence similarities and common interacting partners; Deng et al. (2002) used maximum-likelihood to infer interacting domain pairs from a protein interaction dataset, by modeling domain pairs as random variables and protein interactions as events; Ng et al. (2003) inferred domain pairs with enough integrated scores, integrating evidences from multiple interaction data.

Note that domains are usually very lengthy, in which most regions are related to folding rather than binding. Therefore, interaction sites are only a crucial part of domains. On the contrary, some interaction sites may not occur in any domain. Therefore, many interaction sites can not be revealed through the study of domain-domain interactions.

2.3 Summary

In this chapter, we first reviewed the major methods to produce protein-protein interaction data, including experimental and computational ones. This part of review manifested the characteristics of the protein-protein interaction data especially the low quality in high-throughput data. The characteristics increase the difficulty to predict protein interaction sites from protein-protein interaction data. We then reviewed major experimental methods to determine interaction sites from various disciplines, including biochemical methods such as solvent probing method and cross-linking, biophysical methods such as X-ray crystallography and multiple magnetic resonance (NMR), biological methods such as phage display and mutagenesis. Although experimental methods cooperate complementarily, they have only solved $< 20\%$ types of interaction sites due to their inefficiency and expensiveness. Thus, computational methods have a very crucial role. Protein-protein docking

is the dominant computational method but it is constrained by the limited amounts of protein structures. Classification methods are traditional machine learning methods but they lack of credible negative data. Conservation methods are unable to distinguish binding patterns from folding patterns. Overall, current computational methods are far from perfect. Therefore, it is valuable to develop novel computational methods in the near future to improve the coverage, specificity and accuracy. Since protein interactions are essentially related to interaction sites and the data are more and more widely available, we aim to predict protein interaction sites from protein-protein interaction data. The proposed methods will be explained in the remainder of the dissertation.

Chapter 3

Using Fixed Points to Model Binding Motif Pairs

3.1 Introduction

As discussed in previous chapters, it is appropriate to represent protein interaction sites as binding motif pairs which consist of two traditional protein motifs rather than only individual binding motifs. To reveal binding motif pairs at the large quantity of protein interaction sites, we should make full use of large amounts of protein interaction data. For this purpose, we propose a novel approach in this chapter. The motivation of our approach comes from the study of correlated mutation (or co-evolution) in the evolution of protein interactions (Pazos and Valencia, 2001). The co-evolution means that the mutations at interaction sites in a interacting protein pair are interactively happening: if a residue change incurred in one protein disrupts the interaction with its partner, some compensatory residue changes will also occur in its interacting partner to sustain the interaction, otherwise, the interaction will be eliminated. For instance, co-evolution has long been observed in well-known interacting protein pairs like dockerins and cohesins (Pages et al., 1997), as well as insulin and its receptors (Fryxell, 1996).

The correlated mutations in evolution imply a chain of motif pairs. We can assume that the binding motif pairs (recently survived motif pairs) should occur more frequently than those ancient motif pairs, and should be more frequent than those non-binding motif pairs. Also, the binding motif pairs should be more stable than others. Otherwise, they would be mutated further. These assumptions are very close to the mathematical notion of fixed point theorems which describes stability generally.

The fixed point theorems can be briefly described as: Let f be a function and x be a point in its domain, if $f(x) = x$, then x is called a *fixed point* for f. A famous fixed point theorem in modern mathematics, proposed by L. Brouwer in 1911, says that any continuous function $f : B \to B$, where B is a closed ball in R^n, has at least one fixed point (Mohamed and William, 2001). An easy example of fixed points is $x = 1$ for $f(x) = 2x - 1$. Hence, the idea of fixed points is to find conditions under which a function possesses a point that maps into itself.

An interesting instantiation of this mathematical notion is in life science: The DNA of a cell can be split into two parts, then, they grow, in two separate cells, to become the same DNA as the original one after self-replicating. In this example, the x is the DNA, and the $f(x)$ is the laws of physics and chemistry applied to the DNA. Recently, an important discovery for fixed points was made by Meng et al. (2004) at protein type level. The discovery is on genomic sequences of a gene family. This family of genes is called C2H2 Zinc-Finger genes, consisting of 226 members. A characteristic of this gene family is the frequent presence of tandem repeats. An interesting problem about these genes is whether they can be translated into the same type of protein before and after a *frameshift*. Twelve of them were found to be translated into the same type of protein after frameshifts. This is a fixed point phenomenon, where the x is the protein type and the function $f(x)$ is the frameshift.

From the description of fixed point theorems, we can see that they may be suitable to model the chain of motif pairs in the evolution of an interaction site. More specifically, a point in this model is a protein motif pair consisting of two traditional protein motifs

and a fixed point is a binding motif pair (termed as *a stable motif pair* from now on). To transform every motif pair to become a stable motif pair, we propose a transformation function to model the evolution of motif pairs.

The remainder of the chapter is organized as follows: In Section 3.2, we give a formal description of the problem, including the basic notations used in this chapter. In Section 3.3, we introduce a function $f_{\mathbb{D}}$ that is closely related to a sequence dataset $\mathbb{D}$ of protein interactions. The function will be used to transform protein motif pairs such that they can become stable ones. In Section 3.4, we prove and discuss the properties of $f_{\mathbb{D}}(\mathbf{X})$, including the convergence property and the forest-like decomposition of its domain. We conclude this chapter in Section 3.5.

3.2 Problem Statement under the Fixed Point Model

3.2.1 Basic Notations

Since a protein is a chain of amino acids, it can be mathematically represented by a *string* of the abbreviations of the 20 standard amino acids, allowing repetitions. We use Σ to denote the alphabet set of the 20 standard amino acids. All the amino acids are denoted by lower-case letters in fixed point model; but proteins and amino acid patterns are denoted by capital letters with mathematical calligraphic fonts. A protein $\mathcal{P}$ is defined as *a sequence* (a string) of amino acids. For example, $\mathcal{P}$ can be $a_1a_2\cdots a_v$, where $a_i \in \Sigma$ for $i = 1, \cdots, v$. This $\mathcal{P}$ is also called a v-length protein. A *segment* of a protein $\mathcal{P}$ is a substring of $\mathcal{P}$ where amino acids are connected continuously.

An *amino acid pattern*, also called a protein *motif*, is defined as a sequence (a string) of subsets of Σ. Hence, a motif $\mathcal{M}$ can be written in the form $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where $\mathcal{A}_i \subseteq \Sigma$ for $i = 1, \cdots, k$.

The following is an example protein motif that was found to be biologically important in signal transduction (Sparks et al., 1996; Kay et al., 2000). This protein motif is $\{p\}\Sigma\{l\}\{p\}\Sigma\{kr\}$ that binds to the SH3 domain of the protein *CrkA*. The length of this motif is 6; the second position of this motif is the whole alphabet set, meaning "don't care what is matched". It can also be written as $\{p\} * \{l\}\{p\} * \{kr\}$ in a traditional way by replacing Σ with the sign "*".

Definition 3.1. *Let a motif $\mathcal{M}$ be $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where at least one $\mathcal{A}$ is not $\emptyset$. $\mathcal{M}$ is defined to be* contained *in a protein $P = a_1a_2\cdots a_v$ if there exists a k-length segment of $\mathcal{P}$, denoted $a_{i+1}a_{i+2}\cdots a_{i+k}$ for some i, such that $a_{i+j} \in \mathcal{A}_j$ for all $\mathcal{A}_j$, $1 \leq j \leq k$, that are not $\emptyset$. If a motif is a sequence of only empty sets, we define that there is no protein containing such a motif.*

A motif $\mathcal{M}$ contained in a protein $\mathcal{P}$ is denoted by $\mathcal{M} \subseteq \mathcal{P}$, and the segment $a_{i+1}a_{i+2}\cdots a_{i+k}$ is said to *match* the motif $\mathcal{M}$.

Next, we give definitions related to interactions. A pair of interacting proteins $\mathcal{P}_1$ and $\mathcal{P}_2$ is called a *protein pair PPr*. This pair is denoted by the set of the two proteins, that is, $PPr = \{\mathcal{P}_1, \mathcal{P}_2\}$. A *motif pair*, denoted *MPr*, is a set of two motifs. One of the most important definitions used in this chapter is about the inclusion relationship between a motif pair and a protein pair.

Definition 3.2. *Let $MPr = \{\mathcal{M}_1, \mathcal{M}_2\}$ be a motif pair and $PPr = \{\mathcal{P}_1, \mathcal{P}_2\}$ be a protein pair. MPr is* contained *in PPr, denoted $MPr \subseteq PPr$, if (1) $\mathcal{M}_1 \subseteq \mathcal{P}_1$ and $\mathcal{M}_2 \subseteq \mathcal{P}_2$, or (2) $\mathcal{M}_1 \subseteq \mathcal{P}_2$ and $\mathcal{M}_2 \subseteq \mathcal{P}_1$.*

Let two proteins: $\mathcal{P}_l = eanftw$, $\mathcal{P}_r = wefc$, and three motifs: $\mathcal{M}_1 = \{ard\}\{nc\}$, $\mathcal{M}_2 = \{e\}\{f\}$, and $\mathcal{M}_3 = \{ard\}\emptyset\{nc\}$. Then the protein P_l contains the motif $\mathcal{M}_1$, i.e. $\mathcal{M}_1 \subseteq P_l$. This is because there exists a 2-length segment an in P_l such that $a \in \{ard\}$ and $n \in \{nc\}$. Similarly, $\mathcal{M}_2 \subseteq \mathcal{P}_r$. Hence, the motif pair $\{\mathcal{M}_1, \mathcal{M}_2\}$ is contained in the protein pair $\{\mathcal{P}_l, \mathcal{P}_r\}$.

However, the motif $\mathcal{M}_3 = \{ard\}\emptyset\{nc\}$ is not contained in any of the two proteins because there does not exist any 3-length segment in $\mathcal{P}_l$ or $\mathcal{P}_r$ that can match M_3. Therefore, motif pairs $\{\mathcal{M}_1, \mathcal{M}_3\}$ or $\{\mathcal{M}_2, \mathcal{M}_3\}$ cannot be contained in the protein pair $\{\mathcal{P}_l, \mathcal{P}_r\}$. But, if $\mathcal{M}_3$ is changed to $\mathcal{M}'_3 = \{erd\}\emptyset\{nc\}$, then both $\mathcal{P}_l$ and $\mathcal{P}_r$ contain $\mathcal{M}'_3$. Note that the empty set $\emptyset$ in $\mathcal{M}_3$ or $\mathcal{M}'_3$ has the same semantic meaning as that of Σ in this case (See Definition 1).

We denote a *sequence dataset* $\mathbb{D}$ of n protein pairs by $\{PPr^i = \{\mathcal{P}_1^i, \mathcal{P}_2^i\}, i = 1, \ldots, n\}$, where $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$ have interactions.

Definition 3.3. *The support of a motif pair* $MPr = \{\mathcal{M}_1, \mathcal{M}_2\}$ *in a protein sequence dataset* $\mathbb{D}$ *is defined as the number of protein pairs in* $\mathbb{D}$ *that contain* MPr*, denoted by* $|\{PPr^i |\ PPr^i \in \mathbb{D}\ , MPr \subseteq PPr^i\}|$.

3.2.2 Problem Statement

Let $\mathbb{D}$ be a sequence dataset of interacting protein pairs, the problem studied in this chapter is to design a function $f_{\mathbb{D}}$ that is closely related to $\mathbb{D}$, and then to discover stable motif pairs that are fixed points with regard to $f_{\mathbb{D}}$.

The domain of the function $f_{\mathbb{D}}$ is the set of all possible motif pairs. Let us first discuss the possibilities of single motifs. Recall that a motif is a sequence of subsets of Σ, denoted by $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$, where $\mathcal{A}_i \subseteq \Sigma$ for $i = 1, \cdots, k$. Hence, if $k = 1$, then the set of all possible motifs is the power set of Σ, denoted $\mathcal{POW}(\Sigma)$. Then, possibilities of k-length motifs $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$ can be represented by the following set union:

$$\bigcup\{\mathcal{A}_1\cdots\mathcal{A}_k \mid \mathcal{A}_i \in \mathcal{POW}(\Sigma) \text{ for } i = 1, \cdots, k\}$$

Since motif pairs are pairs of motifs, the set of all possible motif pairs has a much larger size than the domain of single motifs. We use $\mathbb{M}$ to denote all possibilities of motif pairs.

Therefore, in a formal way, the problem can be described as follows. Let $\mathbb{D}$ be a sequence dataset of protein pairs, our objective is to design a function

$$f_{\mathbb{D}} : \mathbb{M} \rightarrow \mathbb{M},$$

and to find those stable motif pairs $\mathbf{X} \in \mathbb{M}$ such that

$$f_{\mathbb{D}}(\mathbf{X}) = \mathbf{X}$$

by using an efficient algorithm.

3.3 Transformation Function of the Fixed Point Model

Given a motif pair $MPr = \{\mathcal{M}_1, \mathcal{M}_2\}$, our proposed $f_{\mathbb{D}}$ involves three steps to transform MPr. In the first step, it discovers a *subset* of $\mathbb{D}$ such that for every protein pair PPr in this subset, PPr contains the given motif pair MPr. We denote this subset by

$$s_{\mathbb{D}}^{MPr} = \{PPr \mid PPr \in \mathbb{D}, MPr \subseteq PPr\}. \tag{3.1}$$

In the second step, $f_{\mathbb{D}}$ moves to extract a *segment pair* from every protein pair in $s_{\mathbb{D}}^{MPr}$. Let $\mathsf{Y} = \{\mathcal{P}_l, \mathcal{P}_r\} \in s_{\mathbb{D}}^{MPr}$, then $MPr \subseteq \mathsf{Y}$. Therefore, there must exist: (1) A segment in $\mathcal{P}_l$ that matches $\mathcal{M}_1$ and a segment in $\mathcal{P}_r$ that matches $\mathcal{M}_2$, or (2) A segment in $\mathcal{P}_r$ that matches $\mathcal{M}_1$ and a segment in $\mathcal{P}_l$ that matches $\mathcal{M}_2$. If the both cases are true, we choose either of them. In each case, we denote the segment that matches $\mathcal{M}_1$ by $segment_1$, and the segment that matches $\mathcal{M}_2$ by $segment_2$. Observe that $\mathcal{M}_1$ and $segment_1$ have the same length, and so, for $\mathcal{M}_2$ and $segment_2$. Suppose there are u protein pairs in $s_{\mathbb{D}}^{MPr}$, then we can get u number of $segment_1$ and u number of $segment_2$. Let the length of $segment_1$ be w. Then, the u $segment_1$ can be represented as the following matrix $[a_{ij}]$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1w} \\ a_{21} & a_{22} & \cdots & a_{2w} \\ & & \cdots & \\ a_{u1} & a_{u2} & \cdots & a_{uw} \end{bmatrix}$$

This matrix is denoted by $aln_s^{\mathcal{M}_1}$. It is called the *alignment* of $\mathcal{M}_1$ with regard to $s_{\mathbb{D}}^{MPr}$ in the bioinformatics literature. Similarly, we can represent those u $segment_2$ as another matrix, denoted by $aln_s^{\mathcal{M}_2}$.

In the third step, our $f_{\mathbb{D}}$ moves to find a *consensus pattern* from the matrix $aln_s^{\mathcal{M}_1}$ and a consensus pattern from the matrix $aln_s^{\mathcal{M}_2}$. In the matrix $aln_s^{\mathcal{M}_1}$, for every column j, denoted by $[a_{ij}], i = 1, \cdots, u$, we choose those a_{ij}, whose population in this column is larger than a *threshold*, to form a set denoted by $\mathcal{A}_j$. If none of these a_{ij} satisfies the condition, we set this position as $\emptyset$. Then, the sequence $\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_w$, a motif, is called the consensus pattern of $\mathcal{M}_1$. This consensus pattern is denoted by $\mathcal{M}_1'$. Similarly, we can find the consensus pattern $\mathcal{M}_2'$ for $\mathcal{M}_2$. Then $\{\mathcal{M}_1', \mathcal{M}_2'\}$ is a transformed motif pair for $MPr = \{\mathcal{M}_1, \mathcal{M}_2\}$ by $f_{\mathbb{D}}$. Therefore, we can write $f_{\mathbb{D}}(\{\mathcal{M}_1, \mathcal{M}_2\}) = \{\mathcal{M}_1', \mathcal{M}_2'\}$.

The threshold for the amino acids' population in a column is important for the consensus pattern discovery. In this chapter, we use 20 percent, a percentage value, as the threshold. That is, if the occurrence rate of an amino acid at a column is less than 20 percent, then we drop it, not allowing it to get into the consensus pattern. Absolute support numbers are also possible for the threshold, but we explain later why percentage thresholds are better than absolute ones.

The discussion above assumes that $s_{\mathbb{D}}^{MPr}$ is non-empty. To let $f_{\mathbb{D}}$ be well-defined, we define the following extreme case for $f_{\mathbb{D}}$: Given a motif pair $\mathbf{X} = \{\mathcal{M}_1, \mathcal{M}_2\}$, if $s_{\mathbb{D}}^{\mathbf{X}} = \emptyset$, we define $f_{\mathbb{D}}(\mathbf{X}) = \{\emptyset\cdots\emptyset, \emptyset\cdots\emptyset\}$, where the number of empty sets in the first sequence is the length of $\mathcal{M}_1$, and the number of empty sets in the second sequence is the length of $\mathcal{M}_2$. Note that if a motif pair $\mathbf{X} = \{\emptyset\cdots\emptyset, \emptyset\cdots\emptyset\}$, then $f_{\mathbb{D}}(\mathbf{X}) = \mathbf{X}$. Such a motif pair is a trivial fixed point for $f_{\mathbb{D}}$.

Next, we use an example to show how $f_{\mathbb{D}}$ proceeds. Let a motif pair $\mathbf{X}$ be $\{\mathcal{M}_1, \mathcal{M}_2\}$, where $\mathcal{M}_1 = \{a\}\{g\}\{g\}\{g\}\{iy\}$ and $\mathcal{M}_2 = \{fv\}\{g\}\{ek\}\{ae\}\{ens\}\{il\}\{a\}$. Let $\mathbb{D}$ be a sequence dataset of interacting protein pairs. Suppose $s_{\mathbb{D}}^{\mathbf{X}}$ contains the following seven protein pairs

$$\{qqq\mathbf{agggi}yy, \quad eei\mathsf{fgkasia}ss\}$$
$$\{aa\mathsf{fgkasia}yy, \quad sss\mathbf{agggy}qy\}$$
$$\{yy\mathbf{agggi}qqq, \quad vx\mathsf{fgkasia}kk\}$$
$$\{kks\mathbf{agggy}ssa, \quad ggq\mathsf{vgeaeia}ii\}$$
$$\{vv\mathbf{agggi}yy, \quad iii\mathsf{vgeaeia}sss\}$$
$$\{qqq\mathsf{vgeaeia}kk, \quad yyy\mathbf{agggi}qqq\}$$
$$\{qqq\mathbf{agggy}qqq, \quad qqq\mathsf{vgeenla}yy\}.$$

Then, $aln_s^{\mathcal{M}_1}$—the segments from the seven protein pairs that match $\mathcal{M}_1$—is the following matrix:

$$\left[\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \hline a & g & g & g & i \\ a & g & g & g & y \\ a & g & g & g & i \\ a & g & g & g & y \\ a & g & g & g & i \\ a & g & g & g & i \\ a & g & g & g & y \end{array}\right]$$

The consensus pattern $\mathcal{M}_1'$ for this matrix is

$$\{a\}\{g\}\{g\}\{g\}\{iy\}.$$

Observe that $\mathcal{M}_1'$ is equal to $\mathcal{M}_1$. This is because that at the fifth column of this matrix, both i and y occur more than 20 percent. Hence, they are kept in the consensus pattern.

Similarly, $aln_s^{\mathcal{M}_2}$—the segments that match $\mathcal{M}_2$—is the following matrix:

$$\left[\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline f & g & k & a & s & i & a \\ f & g & k & a & s & i & a \\ f & g & k & a & s & i & a \\ v & g & e & a & e & i & a \\ v & g & e & a & e & i & a \\ v & g & e & a & e & i & a \\ v & g & e & \mathbf{e} & \mathbf{n} & \mathbf{l} & a \end{array}\right]$$

The consensus pattern $\mathcal{M}_2'$ for this matrix is

$$\{fv\}\{g\}\{ke\}\{a\}\{se\}\{i\}\{a\}.$$

Note that $\mathcal{M}_2'$ is not equal to $\mathcal{M}_2$. Also observe that the amino acids e, n, l at columns 4, 5, and 6 (in bold font), respectively, are dropped. Therefore, they do not appear in the fourth, fifth, and sixth set of $\mathcal{M}_2'$.

Since $f_{\mathbb{D}}(\{\mathcal{M}_1, \mathcal{M}_2\}) = \{\mathcal{M}_1, \mathcal{M}_2'\}$, $\mathbf{X} = \{\mathcal{M}_1, \mathcal{M}_2\}$ is not a fixed point of $f_{\mathbb{D}}$.

This example has illustrated that $f_{\mathbb{D}}$ uses three steps—discovery of a subset of $\mathbb{D}$, extraction of segments from this subset, and discovery of consensus patterns—to transform a given motif pair.

Table 3.1 gives an example showing the transformation from a starting motif pair to a fixed point, where three rounds of transformations by $f_{\mathbb{D}}$ are experienced before stable status is reached.

3.4 Properties of the Transformation Function

This section presents some important properties of $f_{\mathbb{D}}$. In the first part, we prove the convergence property of $f_{\mathbb{D}}$ for every starting motif pair and also discuss the forest structure

Table 3.1: A starting motif pair becomes a fixed point of our function $f_{\mathbb{D}}$ after three rounds of transformation

convergence	motif pairs $\mathbf{X}$											$\lvert s_{\mathbb{D}}^{\mathbf{X}} \rvert$
starting	{ek}	{g}	{l}	{l}	,	{k}	{ek}	{ek}	Σ	{g}	{iv}	31
$\mathbf{X}^{(1)}$	{ek}	{g}	{l}	{l}	,	{k}	{ek}	{ek}	{a}	{g}	{iv}	11
$\mathbf{X}^{(2)}$	{ek}	{g}	{l}	{l}	,	{k}	{e }	{ k}	{a}	{g}	{ v}	10
M_{fixed}	{ek}	{g}	{l}	{l}	,	{k}	{e }	{ k}	{a}	{g}	{ v}	10

of the domain of $f_{\mathbb{D}}$. In the second part, we discuss some specific properties of $f_{\mathbb{D}}$ when the consensus pattern threshold is set as percentage values or set as absolute numbers. In the third part, we explain why using percentage thresholds is a better choice than using absolute numbers for our fixed point theorems to model the binding in protein–protein interactions.

3.4.1 Convergence Properties

Proposition 3.1. *Given a motif pair* Y *and a sequence dataset* $\mathbb{D}$ *of interacting protein pairs, let* $\mathbf{X} = f_{\mathbb{D}}(\mathsf{Y})$ *and* $\mathbf{X}' = f_{\mathbb{D}}(\mathbf{X})$*, then* $s_{\mathbb{D}}^{\mathbf{X}'} \subseteq s_{\mathbb{D}}^{\mathbf{X}}$.

Proof. If $s_{\mathbb{D}}^{\mathbf{X}'} = \emptyset$, of course, $s_{\mathbb{D}}^{\mathbf{X}'} \subseteq s_{\mathbb{D}}^{\mathbf{X}}$. Next we prove this proposition for $s_{\mathbb{D}}^{\mathbf{X}'} \neq \emptyset$. Denote $\mathbf{X} = \{\mathcal{M}_1, \mathcal{M}_2\}$, $\mathcal{M}_1 = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_v$, $\mathcal{M}_2 = \mathcal{B}_1\mathcal{B}_2\cdots\mathcal{B}_w$; $\mathbf{X}' = \{\mathcal{M}_1', \mathcal{M}_2'\}$, $\mathcal{M}_1' = \mathcal{A}_1'\mathcal{A}_2'\cdots\mathcal{A}_v'$, $\mathcal{M}_2' = \mathcal{B}_1'\mathcal{B}_2'\cdots\mathcal{B}_w'$. Because $\mathbf{X}$ is a motif pair resulting from Y after a transformation by $f_{\mathbb{D}}$, then $\mathcal{A}_i' \neq \emptyset$ and also $\mathcal{A}_i' \subseteq \mathcal{A}_i$ for those i satisfying $\mathcal{A}_i \neq \emptyset$. Similarly, $\mathcal{B}_i' \neq \emptyset$ and also $\mathcal{B}_i' \subseteq \mathcal{B}_i$ for those i satisfying $\mathcal{B}_i \neq \emptyset$. That is, if $\mathcal{A}_i \neq \emptyset$ (respectively $\mathcal{B}_i \neq \emptyset$), $\mathcal{A}_i'$ (respectively $\mathcal{B}_i'$) would never become an empty set under the percentage thresholds such as 20 percent used in this chapter. (Note that this is not true when $\mathbf{X}$ is an arbitrary motif pair. That is why we need to set $\mathbf{X} = f_{\mathbb{D}}(\mathsf{Y})$ for some Y.)

Let $PPr \in s_{\mathbb{D}}^{\mathbf{X}'}$, we prove $PPr \notin \mathbb{D} - s_{\mathbb{D}}^{\mathbf{X}}$. Assume $PPr \in \mathbb{D} - s_{\mathbb{D}}^{\mathbf{X}}$, then $PPr \not\supseteq \mathbf{X}$. Therefore, for each two segments from PPr, they cannot match $\mathcal{M}_1$ and $\mathcal{M}_2$ at the same time. Therefore, they furthermore cannot match $\mathcal{M}_1'$ and $\mathcal{M}_2'$ at the same time. This is because $\mathcal{A}_i' \subseteq \mathcal{A}_i$ for those i satisfying $\mathcal{A}_i \neq \emptyset$, and $\mathcal{B}_i' \subseteq \mathcal{B}_i$ for those i satisfying $\mathcal{B}_i \neq \emptyset$. Here is a contradiction. Thus our assumption, that $PPr \in \mathbb{D} - s_{\mathbb{D}}^{\mathbf{X}}$, must be false. Therefore, we can conclude that $PPr \in s_{\mathbb{D}}^{\mathbf{X}}$. □

This proposition is useful for efficiently computing $s_{\mathbb{D}}^{\mathbf{X}'}$. By definition, $s_{\mathbb{D}}^{\mathbf{X}'}$ is a subset of $\mathbb{D}$ in which every protein pair contains the motif pair $\mathbf{X}'$. Therefore, a naive way to compute $s_{\mathbb{D}}^{\mathbf{X}'}$ is to check whether every protein pair in $\mathbb{D}$ contains $\mathbf{X}'$. Having the proposition, this naive method becomes unnecessary because the check within $s_{\mathbb{D}}^{\mathbf{X}}$ is sufficient. Since $s_{\mathbb{D}}^{\mathbf{X}}$ is much smaller than $\mathbb{D}$, we can gain much efficiency.

Theorem 3.1. *Let $\mathbb{D}$ be a sequence dataset of interacting protein pairs. Then for every starting motif pair $\mathbf{X}$, $f_{\mathbb{D}}(\mathbf{X})$ converges to a fixed point $\mathbf{X}_F$. That is, there exists an integer $t_0(\geq 1)$ such that $f_{\mathbb{D}}^{(t_0)}(\mathbf{X}) = \mathbf{X}_F$, and $f_{\mathbb{D}}(\mathbf{X}_F) = \mathbf{X}_F$, where $f_{\mathbb{D}}^{(1)}(\mathbf{X})$ represents $f_{\mathbb{D}}(\mathbf{X})$, $f_{\mathbb{D}}^{(2)}(\mathbf{X})$ represents $f_{\mathbb{D}}(f_{\mathbb{D}}(\mathbf{X}))$, and $f_{\mathbb{D}}^{(t+1)}(\mathbf{X})$ represents $f_{\mathbb{D}}(f_{\mathbb{D}}^{(t)}(\mathbf{X}))$.*

Proof. Denote $\mathbf{X}^{(0)} = \mathbf{X}$, $\mathbf{X}^{(1)} = f_{\mathbb{D}}^{(1)}(\mathbf{X})$, $\cdots$, $\mathbf{X}^{(t)} = f_{\mathbb{D}}^{(t)}(\mathbf{X})$.

By Proposition 3.1, we know that $s_{\mathbb{D}}^{\mathbf{X}^{(t+1)}} \subseteq s_{\mathbb{D}}^{\mathbf{X}^{(t)}}$ for every $t \geq 1$. Since $s_{\mathbb{D}}^{\mathbf{X}^{(1)}}$ is a limited set, there must exist a $t \geq 1$ such that $s_{\mathbb{D}}^{\mathbf{X}^{(t)}} = s_{\mathbb{D}}^{\mathbf{X}^{(t+1)}}$. Therefore, the consensus pattern from $s_{\mathbb{D}}^{\mathbf{X}^{(t)}}$ is equal to the consensus pattern from $s_{\mathbb{D}}^{\mathbf{X}^{(t+1)}}$. Because the consensus pattern from $s_{\mathbb{D}}^{\mathbf{X}^{(t)}}$ is represented as $\mathbf{X}^{(t+1)}$, and the consensus pattern from $s_{\mathbb{D}}^{\mathbf{X}^{(t+1)}}$ is represented as $\mathbf{X}^{(t+2)}$, we have $\mathbf{X}^{(t+1)} = \mathbf{X}^{(t+2)}$. That is, $f_{\mathbb{D}}(\mathbf{X}_F) = \mathbf{X}_F$, where $\mathbf{X}_F = \mathbf{X}^{(t+1)}$, as desired. □

From this theorem, we can understand: (1) that every starting motif pair will converge to a fixed point (likely an empty pattern) and (2) that different starting motif pairs may converge to the same fixed point. Therefore, the domain of $f_{\mathbb{D}}$ can be partitioned into non-overlapping clusters with each cluster corresponding to one fixed point. More specifically,

each cluster is a tree, as proved by the following proposition. Which trees are interesting and biologically meaningful? In the next chapter, we provide a heuristics.

Proposition 3.2. *The domain (search space) of $f_{\mathbb{D}}$ is a forest, with each root node as a fixed point (a stable motif pair).*

Proof. We denote a motif pair $\mathbf{X}$ as a node. If an edge is set from all possible $\mathbf{X}$ to $f_{\mathbb{D}}(\mathbf{X})$, the search space can be viewed as a graph. Since $f_{\mathbb{D}}(\mathbf{X})$ is a unique motif pair, the out-degree of each node should be no more than one. Meanwhile, it is impossible to have a circle in the graph. Assume $\mathbf{X}_0, \mathbf{X}_1 \ldots \mathbf{X}_k, \mathbf{X}_0$ is a circle. According to Proposition 3.1, $s_{\mathbb{D}}^{\mathbf{X}_0} \supseteq s_{\mathbb{D}}^{\mathbf{X}_1} \ldots \supseteq s_{\mathbb{D}}^{\mathbf{X}_t} \supseteq s_{\mathbb{D}}^{\mathbf{X}_0}$. Then $s_{\mathbb{D}}^{\mathbf{X}_0} = s_{\mathbb{D}}^{\mathbf{X}_1} \ldots = s_{\mathbb{D}}^{\mathbf{X}_t} = s_{\mathbb{D}}^{\mathbf{X}_0}$. Therefore, $\mathbf{X}_0 = \mathbf{X}_1 = \ldots = \mathbf{X}_t$. Hence, $\mathbf{X}_0$ is a fixed point. Thus it is impossible to have an out edge to $\mathbf{X}_1$. Also, by Theorem 3.1, every motif pair can lead to a fixed point, with the out degree as zero, which is the corresponding root of that tree. □

3.4.2 Specific Properties

Recall that the definition of $f_{\mathbb{D}}$ involves a step for consensus pattern discovery. To find consensus patterns, we need a threshold to filter out those minor amino acids from the alignments. As mentioned, we have two options to select the threshold: one is to use percentage values as the threshold; the other is to use absolute numbers. We denote the former approach as $f_{(\%,\mathbb{D})}$, and the latter as $f_{(\pi,\mathbb{D})}$.

The following proposition shows that the stability of a fixed point of $f_{(\pi,\mathbb{D})}$ can be transferred to its submotifs. Here, a motif M' is a submotif of motif $\mathcal{M}$ if M' is a segment of $\mathcal{M}$.

Proposition 3.3. *Let a motif pair $\mathbf{X} = \{\mathcal{M}_1, \mathcal{M}_2\}$ be a fixed point of $f_{(\pi,\mathbb{D})}$, then every of its submotif pairs $\mathbf{X}' = \{\mathcal{M}_1', \mathcal{M}_2'\}$ is a fixed point of $f_{(\pi,\mathbb{D})}$ as well, where $\mathcal{M}_1'$ is a submotif of $\mathcal{M}_1$, and $\mathcal{M}_2'$ is a submotif of $\mathcal{M}_2$.*

Proof. Because $\mathbf{X}'$ is a submotif pair of $\mathbf{X}$, for $\forall PPr \in s_{\mathbb{D}}^{\mathbf{X}}$, we have $PPr \in s_{\mathbb{D}}^{\mathbf{X}'}$, i.e. $s_{\mathbb{D}}^{\mathbf{X}} \subseteq s_{\mathbb{D}}^{\mathbf{X}'}$. Since $\mathbf{X}$ is a fixed point of $f_{(\pi,\mathbb{D})}$, $\forall a_{ij} \in \mathcal{A}_i$ either from $\mathcal{M}_1$ or from $\mathcal{M}_2$, its population in $s_{\mathbb{D}}^{\mathbf{X}}$ must be above the threshold. Since every occurrence of a_{ij} in $s_{\mathbb{D}}^{\mathbf{X}}$ is also an occurrence of a_{ij} in $s_{\mathbb{D}}^{\mathbf{X}'}$, the occurrence of $\forall a_{ij}$ in $\mathbf{X}'$ is also above the threshold. Therefore, $\mathbf{X}'$ is also a fixed point of $f_{(\pi,\mathbb{D})}$. □

Proposition 3.3 says that the fixed points of $f_{(\pi,\mathbb{D})}$ satisfies the famous Apriori-property (Agrawal and Srikant, 1994) known in data mining field. That is, if a submotif pair of a motif pair is not a fixed point, the motif pair is impossible to be a fixed point. Therefore, the mining of fixed points of $f_{(\pi,\mathbb{D})}$ should be similar to those algorithms for mining frequent itemsets.

Note that Proposition 3.3 does not hold if we replace $f_{(\pi,\mathbb{D})}$ with $f_{(\%,\mathbb{D})}$.

Proposition 3.4. *Let* $\mathbf{X}$ *and* Y *be two equal-length stable motif pairs of* $f_{(\pi,\mathbb{D})}$*, where* $\mathbf{X} = \{\mathcal{M}_{\mathbf{X}1}, \mathcal{M}_{\mathbf{X}2}\}$*,* $\mathsf{Y} = \{\mathcal{M}_{\mathsf{Y}1}, \mathcal{M}_{\mathsf{Y}2}\}$*,* $|\mathcal{M}_{\mathbf{X}1}| = |\mathcal{M}_{\mathsf{Y}1}|$ *and* $|\mathcal{M}_{\mathbf{X}2}| = |\mathcal{M}_{\mathsf{Y}2}|$*. Then the union motif pair* $\mathbf{X} + \mathsf{Y} = \{\mathcal{M}_{\mathbf{X}1} + \mathcal{M}_{\mathsf{Y}1}, \mathcal{M}_{\mathbf{X}2} + \mathcal{M}_{\mathsf{Y}2}\}$ *is also a fixed point of* $f_{(\pi,\mathbb{D})}$*. The union operation '+' of two motifs is defined as follows: suppose* $M = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$*, and* $M' = \mathcal{A}'_1\mathcal{A}'_2\cdots\mathcal{A}'_k$*, then* $M + M' = \mathcal{A}''_1\mathcal{A}''_2\ldots\mathcal{A}''_k$*, where* $\mathcal{A}''_i = \mathcal{A}_i \cup \mathcal{A}'_i$*,* $1 \leq i \leq k$*.*

Proof. Observe that $\forall PPr \in s_{\mathbb{D}}^{\mathbf{X}}$, then $PPr \in s_{\mathbb{D}}^{\mathbf{X}+\mathsf{Y}}$. Hence, we have $s_{\mathbb{D}}^{\mathbf{X}} \subseteq s_{\mathbb{D}}^{\mathbf{X}+\mathsf{Y}}$. Similarly, we can get $s_{\mathbb{D}}^{\mathsf{Y}} \subseteq s_{\mathbb{D}}^{\mathbf{X}+\mathsf{Y}}$. Since $\mathbf{X}$ and Y are fixed points of $f_{(\pi,\mathbb{D})}$, for $\forall a_{ij} \in \mathcal{A}_i$ either from $M_{\mathbf{X}1}$ or from $M_{\mathbf{X}2}$, its support in $s_{\mathbb{D}}^{\mathbf{X}}$ is above the threshold. Since every occurrence of a_{ij} in $s_{\mathbb{D}}^{\mathbf{X}}$ is also an occurrence of a_{ij} in $s_{\mathbb{D}}^{\mathbf{X}+\mathsf{Y}}$, the occurrence of $\forall a_{ij}$ in $\mathbf{X} + \mathsf{Y}$ is also above the support threshold. Therefore, $\mathbf{X} + \mathsf{Y}$ is also a fixed point. □

Note that this proposition may not hold if $f_{(\pi,\mathbb{D})}$ replaced with $f_{(\%,\mathbb{D})}$. This is because the occurrence of the union motif pairs not only covers the occurrences of the two original fixed points, but also covers some occurrences from new combinations. Therefore, it is difficult to determine whether the occurrence rate is still above the percentage threshold. Another interesting thing is that if $\mathbf{X}$ is not a fixed point, $\mathbf{X} + \mathsf{Y}$ is not impossible to be a fix point of $f_{(\pi,\mathbb{D})}$.

Proposition 3.5. *Let $f_{(\%,\mathbb{D})}$ be the $f_{\mathbb{D}}$ under the percentage threshold in the consensus pattern discovery. Let a motif pair $\mathbf{X} = \{\mathcal{M}_1, \mathcal{M}_2\}$, where $\mathcal{M}_1 = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_v$, $\mathcal{A}_i \subseteq \Sigma$, for $i = 1, \cdots, v$; $\mathcal{M}_2 = \mathcal{B}_1\mathcal{B}_2\cdots\mathcal{B}_w$, $\mathcal{B}_j \subseteq \Sigma$, for $j = 1, \cdots, w$. If all $\mathcal{A}_i$ and $\mathcal{B}_j$ are singleton sets, and $s_{\mathbb{D}}^{\mathbf{X}} \neq \emptyset$, then $\mathbf{X}$ is a fixed point of $f_{(\%,\mathbb{D})}$.*

Proof. Denote $\mathcal{A}_i = \{a_i\}$ for $i = 1, \cdots, v$, and $\mathcal{B}_j = \{b_j\}$ for $j = 1, \cdots, w$. Suppose $s_{\mathbb{D}}^{\mathbf{X}}$ contains m protein pairs $PPr^i, i = 1, \cdots, m$. Then the segment from the protein pair PPr^i for every i that matches $\mathcal{M}_1$ must be $a_1a_2\cdots a_v$; similarly, the segment from the protein pair PPr^i for every i that matches $\mathcal{M}_2$ must be $b_1b_2\cdots b_w$. Therefore, the two alignments $aln_s^{\mathcal{M}_1}$ and $aln_s^{\mathcal{M}_2}$ are the following two special matrixes:

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_v \\ a_1 & a_2 & \cdots & a_v \\ & & \cdots & \\ a_1 & a_2 & \cdots & a_v \end{bmatrix}$$

and

$$\begin{bmatrix} b_1 & b_2 & \cdots & b_w \\ b_1 & b_2 & \cdots & b_w \\ & & \cdots & \\ b_1 & b_2 & \cdots & b_w \end{bmatrix}$$

Then, the consensus pattern for $aln_s^{\mathcal{M}_1}$ and $aln_s^{\mathcal{M}_2}$ are $\{a_1\}\{a_2\}\cdots\{a_v\}$ and $\{b_1\}\{b_2\}\cdots\{b_w\}$ respectively, under percentage threshold, as the occurrence rate is 100 percent in this case. Hence, we can see that $\mathbf{X}$ is a fixed point of $f_{(\%,\mathbb{D})}$. □

3.4.3 Discussions of Properties

In this subsection, we give a comparison between $f_{(\%,\mathbb{D})}$ and $f_{(\pi,\mathbb{D})}$, and explain the reasons that $f_{(\%,\mathbb{D})}$ is better than $f_{(\pi,\mathbb{D})}$ for modeling the binding in protein–protein interactions.

First, let us examine the most likely lengths of fixed points derived by $f_{(\%,\mathbb{D})}$ and $f_{(\pi,\mathbb{D})}$. According to Proposition 3.3, for a *long* stable motif pair $\mathbf{X}$ of $f_{(\pi,\mathbb{D})}$, all submotif pairs of

X are also fixed points of $f_{(\pi,\mathbb{D})}$. In extreme cases, those many 1-1 pairs are stable motif pairs. In biology, they are called residue–reside interaction pairs (Glaser et al., 2001). Though they may be fundamental components of some interaction sites, they may have very high false positive rate. One way to solve this problem is to discover only those *maximal fixed points* of $f_{(\pi,\mathbb{D})}$ which are similar to a well studied data mining concept called maximal frequent patterns (Burdick et al., 2001; Grahne and Zhu, 2003). On the other hand, both very short and very long motif pairs are unlikely to be fixed points of $f_{(\%,\mathbb{D})}$ due to the equal possibility for short motif pairs and rare possibility for long motif pairs. This property of $f_{(\%,\mathbb{D})}$ is very consistent with the observations in biology (Sheu et al., 2005) that most interaction sites generally include more than 10 but less than 20 residues. In fact, the lengths of our discovered stable motif pairs of $f_{(\%,\mathbb{D})}$ match very well to those of real motif pairs.

Second, let us discuss the union ($'+'$) operation for $f_{(\%,\mathbb{D})}$ and $f_{(\pi,\mathbb{D})}$. According to Proposition 3.4, the union of *any* two equal-length fixed points of $f_{(\pi,\mathbb{D})}$ is also a fixed point of $f_{(\pi,\mathbb{D})}$, but this flexibility does not hold for fixed points of $f_{(\%,\mathbb{D})}$. In the real biology circumstances, this union property does not usually hold for interaction sites either. For example, a study on active sites (Doray and Kornfeld, 2001) shows only specially selected amino acids (not arbitrarily united) are possible to compose an interaction site or an active site. The union property of fixed points of $f_{(\pi,\mathbb{D})}$ also leads to another bad consequence: the motif pairs with large set in all positions are more likely to be fixed points. In the extreme case, the motif pairs which contain only full alphabet sets in each position are most likely to be fixed points. It is obviously meaningless from a biology perspective. However, $f_{(\%,\mathbb{D})}$ does not produce such fixed points.

Hence, $f_{(\%,\mathbb{D})}$ is better than $f_{(\pi,\mathbb{D})}$ for modeling the binding motif pairs in protein–protein interactions, as it reflects more properties of the real interaction sites. However, $f_{(\%,\mathbb{D})}$ has the singleton problem as discussed in Proposition 3.5. By this proposition, every segment pair from any protein pair of $\mathbb{D}$ is a fixed point of $f_{(\%,\mathbb{D})}$. Hence, it seems that there are many easy fixed points for $f_{(\%,\mathbb{D})}$. Therefore, we need other statistical measurements to remedy this, for example, using the support level or P-score of these fixed points in $\mathbb{D}$

or biological evidence as discussed in the Chapter 4 to filter out some easy ones. In the remainder of the chapter, every $f_{\mathbb{D}}$ refers to $f_{(\%,\mathbb{D})}$.

3.5 Summary

Motivated from correlated mutations and stability for many biological phenomena, we have proposed a fixed point model in this chapter to emulate the evolution of motif pairs at protein interaction sites, where a point is defined as a protein motif pair consisting of two traditional protein motifs and a fixed point (a stable motif pair) of this model is defined as a binding motif pair. To transform every motif pair to a stable motif pair, we proposed a mathematical function $f_{\mathbb{D}}$ which is closely related to a sequence data of interacting protein pairs. The transformation of a motif pair by $f_{\mathbb{D}}$ involves three steps: the discovery of a subset of $\mathbb{D}$, the extraction of alignments from this subset, and the discovery of two consensus patterns. We have proved that $f_{\mathbb{D}}$ is a convergent function for every starting motif pair, that is, mathematically, it is a chain of changing but converging patterns from every unstable starting motif pair to a stable motif pair. In this chapter, we have also discussed that $f_{(\%,\mathbb{D})}$ is better than $f_{(\pi,\mathbb{D})}$ to model the evolution of motif pairs, as it reflects more properties of the real interaction sites.

The discovery of all stable motif pairs from large amounts of protein-protein interaction data is an interesting and challenging problem. But unfortunately, it is extremely tough to find a complete solution. We will discuss it the next chapter and present a heuristic approach.

Chapter 4

Selection of Starting Motif Pairs and Significance of Stable Motif Pairs

4.1 Motivation

In the last chapter, we propose a fixed point model to discover binding motif pairs at protein interaction sites from a sequence dataset of interacting protein pairs. Although the model is promising, it has several weaknesses: (1) Computational difficulties. From Theorem 3.1, we know that every starting motif pair will converge to a stable motif pair after a couple of transformations by the function $f_{\mathbb{D}}$ we propose. Since the domain of the function is enormous (shown in Section 3.2.2), it is a computational challenging problem to work out stable motif pairs from large amounts of interacting protein sequence pairs; and (2) The singleton problem. As described in Proposition 3.5, every segment pair in an interacting protein pair is a stable motif pair of our proposed transformation function $f_{\mathbb{D}}$. This is obviously against the biological law; and (3) Significance issues. We can not guarantee all stable motif pairs of the fixed point model are statistically significant, some of which may still be chance ones.

To tackle the first weakness, we turn to work out only a subset of meaningful stable motif pairs since the complete solution is not available. Recall that there are two types of protein interaction data: protein interaction sequence data and protein complex structural data (refer to Chapter 1 and 2). Protein interaction sequence data are abundantly produced by existing high-throughput interaction detection techniques, but they are not accurate enough. On the other hand, protein complex structural data contains the most reliable three-dimensional coordinate information about interacting proteins, through which the exact locations of interaction sites can be figured out by calculating the distances of amino acids between interacting proteins in the complexes. However, the complex data is expensive and time-consuming to produce so that only a limited amount of data is available. Hence, we propose a heuristic approach which makes use of both types of data. We select good candidates for starting motif pairs which are guided from the protein complex structural data, so that the resulting stable motif pairs can have good biological significance.

A key idea in the heuristic approach is the detection of so-called *maximal contact segment pairs* between two proteins residing in a complex, to present interaction sites between the two proteins. First, all possible pairs of spatially contacting residues are determined from the three-dimensional structure data of a protein complex. These contact residues are extended to capture as many continuous contact residues as possible along the two proteins, thus derived the maximal contact segment pairs. Computationally, the derivation of maximal contact segment pairs is a challenging problem. We describe an algorithm to discover them efficiently. Then, we generalize these maximal contact segment pairs into starting motif pairs from a protein interaction sequence dataset, to search for stable motif pairs through our transformation function from the interaction sequence dataset. By this way, we can obtain high confidence to the discovered stable motif pairs since they are stemmed from the biologically reliable protein complex structural data. The heuristic approach reduces the formidable search space of interacting protein sequences while providing some biological support for the motif pairs discovered. Indeed, many of our motif pairs discovered this way can be confirmed by biological patterns reported in

the literature, as shown in this chapter.

Although the heuristic approach is effective, the resulting stable motif pairs may still be insignificant. To eliminate the problem and other two weaknesses, we introduce the concept of *significant motif pairs* in this chapter to capture more information for binding motif pairs, so that the chance motif pairs and insignificant singleton segment pairs can be filtered out. We require the resulting stable motif pairs to be significant not only for single motifs but more importantly for their co-occurrence as pairs. By significance, we mean that their observations or supports should be much higher than their random expectations. Thus, the final *stable* and *significant* motif pairs are modeled as *binding motif pairs* in the fixed point model.

The remainder of this chapter is organized as follows: In Section 4.2, we describe a heuristic approach which generates starting motif pairs from maximal contact segment pairs. In Section 4.3, we depict the significance measurements to evaluate stable motif pairs derived from the starting motif pairs. We review the overall algorithm and results in Section 4.4. Then, we examine the results of the heuristic approach in details. We conduct comprehensive random experiments and report them in Section 4.5. We perform a series of literature validations to our discovered binding motif pairs and report them in Section 4.6. We discuss the heuristic approach in Section 4.7 and summarize the chapter in the final section.

4.2 Generating Starting Motif Pairs from Maximal Contact Segment Pairs

4.2.1 Concept of Maximal Contact Segment Pairs

To define maximal contact segment pairs, we first clarify a basic concept, which is called a contact site. Given a pair of proteins in a complex, a *contact site* is an elemental pair

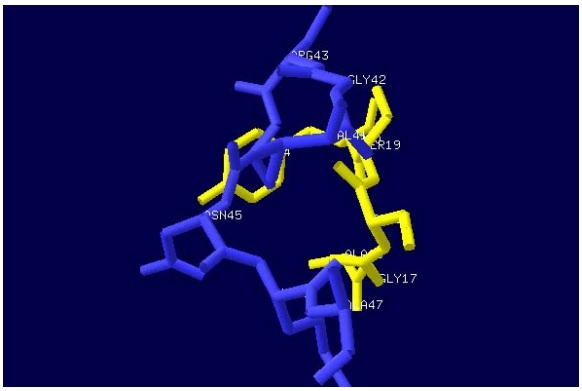
such as two residues or two atoms, each coming from one of the two proteins, which are close enough in space. As a protein complex usually consists of multiple proteins, in this study we consider all pairs of proteins in a protein complex to obtain all contact sites in this step.

We define a *contact site* mathematically as follows: *Suppose two proteins with 3-D structural coordinates in* (x, y, z), $\mathcal{P}_a = \{(\mathbf{a}_i, x_{\mathbf{a}_i}, y_{\mathbf{a}_i}, z_{\mathbf{a}_i}), i = 1...m\}$ *and* $\mathcal{P}_b = \{(\mathbf{b}_j, x_{\mathbf{b}_j}, y_{\mathbf{b}_j}, z_{\mathbf{b}_j}), j = 1...n\}$. *The pair* $(\mathbf{a}_i, \mathbf{b}_j)$ *is a contact site if* $dist(\mathbf{a}_i, \mathbf{b}_j) \leq \varepsilon$, *where* $\mathbf{a}_i$ *and* $\mathbf{b}_j$ *are the atom id in the protein* $\mathcal{P}_a$ *and* $\mathcal{P}_b$ *respectively, and* ε *is an empirical threshold for the Euclidean distance function* $dist(.,.)$. *Such a pair is denoted* $Contact(\mathbf{a}_i, \mathbf{b}_j)$, *or equivalently* $Contact(\mathbf{b}_j, \mathbf{a}_i)$.

Note that a contact site in the atom level directly implies a contact site in residue level because each atom is a part of a unique residue. Hereafter, we will discuss contact sites only at the residue level. Since two residues are said to be in contact if one of the atoms in a residue is in contact with one atom in the other residue, it is possible for a residue to be in contact with multiple residues.

Next, we extend the concept of contact sites to the concept of *contact segment pairs*, aiming to search for large areas of contact sites in a pair of interacting proteins. Figure 4.1 shows our idea, depicting a typical scenario where segments of residues in one protein are continuously in contact with segments of residues in the other protein. As an illustration, the segment $[\mathbf{a}_{16}, \mathbf{a}_{17}]$ in protein A of Figure 4.1 is in contact with the segment $[\mathbf{d}_{44}, \mathbf{d}_{46}]$ in protein D. That is, they are a contact segment pair. But the segment $[\mathbf{a}_{16}, \mathbf{a}_{17}]$ in protein A and the segment $[\mathbf{d}_{44}, \mathbf{d}_{47}]$ in protein D are collectively not a contact segment pair.

Formally, the definition is: *A contact segment pair is a segment pair* $([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}], [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}])$ *satisfying: (1) for* $\forall \mathbf{a}_i \in [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$, $\exists \mathbf{b}_j \in [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$ *such that* $(\mathbf{a}_i, \mathbf{b}_j)$ *is a contact site; (2) for* $\forall \mathbf{b}_i \in [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$, $\exists \mathbf{a}_i \in [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$ *such that* $(\mathbf{b}_j, \mathbf{a}_i)$ *is a contact site, where* $\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \mathbf{b}_{j_1}, \mathbf{b}_{j_2}$ *are residue ids in two proteins* $\mathcal{P}_a$ *and* $\mathcal{P}_b$. *Such a pair of segments is sometimes denoted* $Contact([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}], [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}])$.



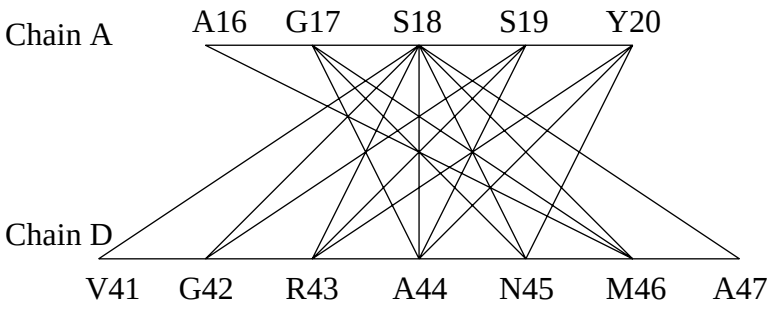


Figure 4.1: An example of maximal contact segment pair taken from the *pdb1mbm* complex. The maximal contact segment pair is $([\mathbf{a}_{16}, \mathbf{a}_{20}], [d_{41}, d_{47}])$ between chain A and chain D with sequence $(agssy, vgranma)$.

A **maximal contact segment pair** is then defined as a contact segment pair such that no other contact segment pair can contain the both segments of this contact pair. In above example, $([\mathbf{a}_{16}, \mathbf{a}_{20}], [\mathbf{d}_{41}, \mathbf{d}_{47}])$ is a maximal contact segment pair.

4.2.2 Extracting Maximal Contact Segment Pairs from Protein Complexes

The problem of extracting maximal contact segment pairs can be formally defined as follows:

Definition 4.1. ***Maximal Contact Segment Pair Problem:*** *Given a pair of interacting proteins $\mathcal{P}_a$ and $\mathcal{P}_b$, suppose $\mathcal{T} = \{(\mathbf{a}_i, \mathbf{b}_j)|\ Contact(\mathbf{a}_i, \mathbf{b}_j)$ with respect to the two proteins $\mathcal{P}_a$and $\mathcal{P}_b\}$, the problem is how to find all possible maximal contact segment pairs from $\mathcal{T}$ with their segment lengths all longer than a threshold.*

A naive approach to solving this problem would require testing all possible segment pairs. Suppose two proteins $\mathcal{P}_a$ and $\mathcal{P}_b$ have m and n residues respectively, then, the proteins $\mathcal{P}_a$ and $\mathcal{P}_b$ will have m^2 and n^2 possible segments respectively. For each combination, $O(mn)$ time complexity would be required for the computation. So, the total time complexity for such a naive approach will be $O(m^3 * n^3)$ per pair of proteins in each complex. This is very expensive particularly when the protein complexes are large and there are hundreds or thousands of protein complexes need to be examined. We present a more efficient method to compute maximal contact segment pairs here.

Observe that for each residue, it may be in contact with multiple residues in the opposite protein (see Figure 4.1). We introduce a concept named *coverage* to capture this phenomenon; it will be shown later that this is a useful concept for improving the efficiency of our discovery algorithm. The coverage of a residue $\mathbf{a}_i$, denoted $Cov(\mathbf{a}_i)$, is the set of all residues in the opposite protein that are in contact with this residue, namely $Cov(\mathbf{a}_i) = \{\mathbf{b}_j|(\mathbf{a}_i, \mathbf{b}_j) \in \mathcal{T}\}$. The coverage of a segment $[\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$, denoted $Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}])$, is the union of the coverage of all its residues, namely,
$Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]) = \cup_{\mathbf{a}_i \in [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]} Cov(\mathbf{a}_i)$.

The following proposition is useful in our algorithm to compute maximal contact segment pairs efficiently.

Proposition 4.1. *A segment pair $([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}], [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}])$ is a contact segment pair iff the coverage of each of the two segments contains the other segment, i.e. $Contact([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}], [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}])$ $\Longleftrightarrow (Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]) \supseteq [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]) \wedge (Cov([\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]) \supseteq [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}])$.*

Proof. $\Rightarrow$: We use contradiction to prove. Suppose $Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]) \supseteq [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$ is not true, then there exists a $\mathbf{b}_j \in [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$ but this $\mathbf{b}_j \notin Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}])$. This means there

is no $\mathbf{a}_i \in [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$ in contact with $\mathbf{b}_j$. This contradicts the assumption. Therefore, $Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]) \supseteq [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$. We can prove $Cov([\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]) \supseteq [\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$ in a symmetrical manner.

$\Leftarrow$: If $Cov([\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]) \supseteq [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$, this means that for each $\mathbf{b}_j \in [\mathbf{b}_{j_1}, \mathbf{b}_{j_2}]$, there exist at least one contact site in $[\mathbf{a}_{i_1}, \mathbf{a}_{i_2}]$. Similarly, the residues in the other segment have the same property. □

Our algorithm is a top-down recursive algorithm. At the initial step, each entire protein in a pair is treated as a segment. A series of recursive breaking-down are then performed to output maximal contact segment pairs, using the above proposition to determine when to break-down a segment into several smaller segments and when to terminate producing a new candidate segment pair. The details of our algorithm are as follows:

Input: An initial segment pair $[\mathbf{a}_1, \mathbf{a}_m]$, and $[\mathbf{b}_1, \mathbf{b}_n]$, and $\mathcal{T} = \{(\mathbf{a}_i, \mathbf{b}_j) |\ Contact(\mathbf{a}_i, \mathbf{b}_j), 1 \leq i \leq m, 1 \leq j \leq n\}$.

Output: A set of maximal contact segment pairs.

Preparation Step: Compute $Cov(\mathbf{a}_i)$ and $Cov(\mathbf{b}_j)$ for all $1 \leq i \leq m$, $1 \leq j \leq n$.

Initialization Step: Put the initial segment pair $([\mathbf{a}_1, \mathbf{a}_m], [\mathbf{b}_1, \mathbf{b}_n])$ into the candidate list.

repeat

Segment Coverage Step: Remove the first segment pair from the candidate list, denoted $([x_{i_1}, x_{i_2}], [y_{j_1}, y_{j_2}])$; Compute the coverage for $Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]$.

Splitting Step:

if $(Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]) == [y_{j_1}, y_{j_2}]$ **then**

if $(Cov([y_{j_1}, y_{j_2}]) \cap [x_{i_1}, x_{i_2}]) == [x_{i_1}, x_{i_2}]$ **then**

Output the segment pair.

else

Add $([y_{j_1}, y_{j_2}], [x_{i_1}, x_{i_2}])$ into the candidate list.

end if

else

Split $Cov([x_{i_1}, x_{i_2}]) \cap [y_{j_1}, y_{j_2}]$ into continuous sub-segments through a linear scan, denoted $[y_{k_{2t-1}}, y_{k_{2t}}], t = 1....w$, where w is the resulting number of sub-segments. Put each segment pair $([y_{k_{2t-1}}, y_{k_{2t}}], [x_{i_1}, x_{i_2}]), t = 1...w$, into the candidate list.

end if

until The candidate list is empty.

A detailed example is shown in Figure 4.2.

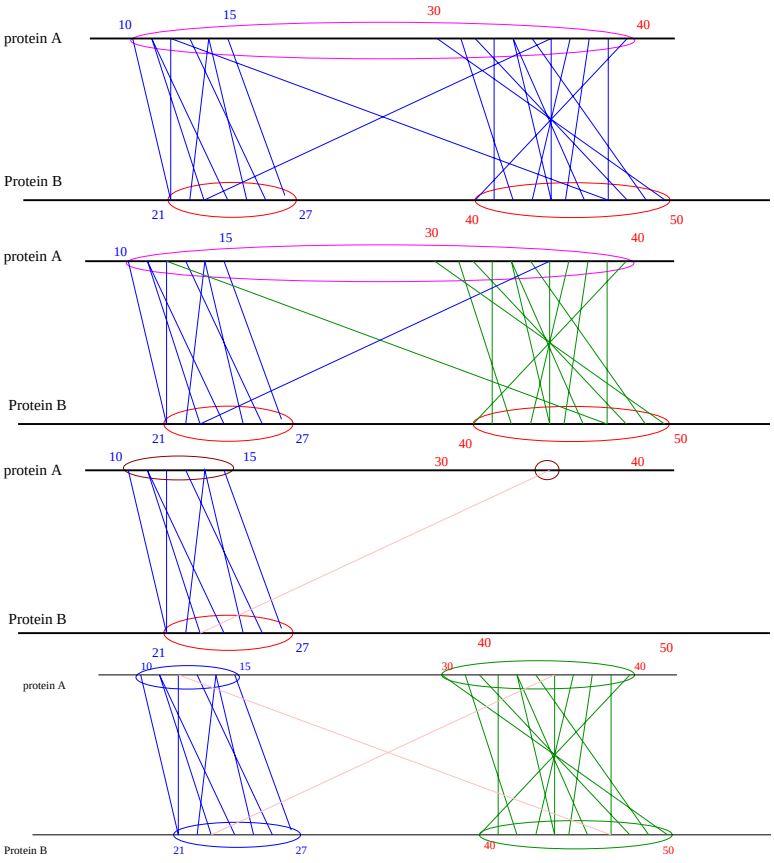


Figure 4.2: An example of computing a contact segment pair which includes four steps.

4.2.3 Generating Starting Motif Pairs

Directly using maximal contact segment pairs as starting motif pairs is not a smart choice. Because these segment pairs are highly specific in corresponding species, they may not occur in other interacting protein dataset $\mathbb{D}$. So, we need to *generalize* these contact segment pairs. We achieve this goal by using the principle proposed in Azarya-Sprinzak et al. (1997). The principle says that even some residues in some positions are changed to other residues, their structures are still unchanged. Since the structures maintain the same, the interacting behavior is highly likely to maintain as well. Basically, we use local alignment and consensus discovery to implement this generalization and to get satisfactory starting motif pairs.

Given a maximal contact segment pair *SPr* and a protein interaction dataset $\mathbb{D}$, the generalization of *SPr* is as follows:

1. Find a subset of $\mathbb{D}$, denoted $s_{\mathbb{D}}^{SPr} = \{PPr \in \mathbb{D} \mid Local\_Alignment(SPr, PPr) \geq \lambda\}$, where λ is an empirical threshold,

2. Discover the consensus pattern *MPr* from $s_{\mathbb{D}}^{SPr}$ as in Section 3.3.

Thus, *MPr* is a generalized pattern for *SPr*. Then we use *MPr* as a starting point to discover a stable motif pair. For instance, from the example maximal segment pair $(agssy, vgranma)$ mentioned in Section 4.2.1, we found 34 interactions for its $s_{\mathbb{D}}^{SPr}$ from a yeast interacting protein dataset $\mathbb{D}$. From this cluster, we induced a consensus motif pair, $\{\{a\}\{g\}\{dgs\}\{gs\}\{ivy\}, \{fv\}\{g\}\{ek\}\{ae\}\{dens\}\{il\}\{a\}\}$, which was then used as the starting point to derive a stable motif pair $\{\{a\}\{g\}\{g\}\{g\}\{iy\}, \{f\}\{v\}\{g\}\{ek\}\{a\}\{es\}\{i\}\{a\}\}$ from the same $\mathbb{D}$.

4.3 Significance Measurements of Motif Pairs

We begin with definitions for the absolute support and statistical score of single motifs and their efficient computation. Then we explain significant motif pairs and give efficient methods to compute their significance indices.

4.3.1 Significance Measurements for Single Motifs

Before discussing the significance of motif pairs, we consider the component motifs, as their significance is the prerequisite to make the whole pair significant.

Definition 4.2. [Support for a motif] *The absolute support of a motif* $\mathcal{M}$ *in* $\mathbb{P} = \{\mathcal{P}_i | i = 1 \dots m\}$ *is the number of proteins in* $\mathbb{P}$ *that contain* $\mathcal{M}$*, denoted by* $\pi(\mathcal{M}, \mathbb{P}) = |\{\mathcal{P}_i \in \mathbb{P} | \mathcal{M} \subseteq \mathcal{P}_i\}|$*, or simply denoted by* $\pi(\mathcal{M})$*.*

The Z-score measurement is widely used to evaluate the significance of single motifs (Atteson, 1998). The Z-score of a motif $\mathcal{M}$ is defined as

$$z_s(\mathcal{M}, \mathbb{P}) = \frac{\pi(\mathcal{M}, \mathbb{P}) - exp(\mathcal{M}, \mathbb{P})}{\sigma(\mathcal{M}, \mathbb{P})} \tag{4.1}$$

where $exp(\mathcal{M}, \mathbb{P})$ is the expectation support for $\mathcal{M}$ in $\mathbb{P}$, $\sigma(\mathcal{M}, \mathbb{P})$ is the standard deviation for the random occurrence (support) of $\mathcal{M}$ in $\mathbb{P}$. With Z-scores, we can distinguish significant motifs from random ones. If the occurrence of a motif is far away from its random expectation, this motif is considered to be statistically significant.

The exact computation of Z-scores is nontrivial. With the help of the software package provided by Nicodeme et al. (2002), the expectation and deviation for a motif $\mathcal{M} = \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k$ with respect to $\mathbb{P}$ can be calculated approximately as follows, where m is the

number of proteins in $\mathbb{P}$:

$$\begin{aligned} p(\mathcal{M}) &= \prod_{i=1}^{k} \frac{|\mathcal{A}_i|}{|\Sigma|} \\ &= \frac{\prod_{i=1}^{k} |\mathcal{A}_i|}{|\Sigma^k|} \\ exp(\mathcal{M}, \mathbb{P}) &= p(\mathcal{M}) * \sum_{i=1}^{m} (|\mathcal{P}_i| - k + 1) \\ &= p(\mathcal{M}) * (\sum_{i=1}^{m} |\mathcal{P}_i| - m * (k-1)) \\ &= p(\mathcal{M}) * (|\mathbb{P}| - m * (k-1)) \end{aligned} \tag{4.2}$$

Nicodeme et al. (2002) also showed that for most motifs,

$$\sigma(\mathcal{M}, \mathbb{P}) \approx \sqrt{exp(\mathcal{M}, \mathbb{P})} \tag{4.3}$$

From formula (4.2) and (4.3), we can see that after one pass of pre-computation for the number of residues in $\mathbb{P}$ and the number of proteins m, the expectation and standard deviation of any motif can be calculated approximately in linear time with respect to the number of positions in the motif, *i.e.* in $O(k)$ time.

4.3.2 Significance Measurements for Motif Pairs

The significance measurements for motif pairs are more complicated that those of single motifs. We first review the Definition 3.3 of the support of a motif pair and discuss several candidate measurements directly related to it. Then we define the contributive support of a motif pair and introduce the concept of P-scores, followed by their computational issues.

Definition 4.3. [Support for a motif pair] *The absolute support of a motif pair* $MPr = \{\mathcal{M}_L, \mathcal{M}_R\}$ *in* $\mathbb{D}$ *is defined as the number of interacting protein pairs in* $\mathbb{D}$ *that contain* MPr*, denoted by* $\pi(MPr, \mathbb{D}) = |\{PPr^i \in \mathbb{D} \mid MPr \subseteq PPr^i\}| = |s_{\mathbb{D}}^{MPr}|$.

Emerging Significance

The significance of motif pairs can be measured straightforwardly by the ratio of frequency in positive and negative dataset providing the negative data is available, which is often referred as to emerging significance (Dong and Li, 1999).

Definition 4.4. ***[Emerging Significance]*** *Suppose we have a dataset $\mathbb{D}$ consisting of sequence pairs $\mathbb{D}$=$\{(\mathcal{P}_1^i, \mathcal{P}_2^i)|1 \leq i \leq n\}$, the frequency of a motif pair MPr with respect to D is defined as: $Freq(MPr, D) = \frac{\pi(MPr,\mathbb{D})}{|\mathbb{D}|}$. Suppose $\mathbb{D}$ is further divided into a positive dataset $\mathbb{D}_{Pos}$ and a negative dataset $\mathbb{D}_{Neg}$. The* **emerging significance** *of MPr with respect to $\mathbb{D}_{Pos}$ and $\mathbb{D}_{Neg}$ is defined as: $ratio(MPr, \mathbb{D}_{Pos}, \mathbb{D}_{Neg}) = \frac{Freq(MPr,\mathbb{D}_{Pos})}{Freq(MPr,\mathbb{D}_{Neg})}$.*

As mentioned in Chapter 2, negative data are usually unavailable. In that case, we can define it against the random expectations as follows:

Definition 4.5. ***[Emerging Significance against Randomness]*** *The emerging significance of a motif pair $MPr = \{\mathcal{M}_L, \mathcal{M}_R\}$ in $\mathbb{D}$ against randomness is defined as the ratio between the support of the motif pair and the production of the support of its two component motifs, denoted by $ratio(MPr, \mathbb{D}) = \frac{\pi(MPr,\mathbb{D})}{\pi(\mathcal{M}_L,\mathbb{P})*\pi(\mathcal{M}_R,\mathbb{P})}$.*

We tried the first measurement and Tan et al. (2004) applied the second measurement to evaluate the significance of motif pairs. Both measurements are insufficiently effective to filter out insignificant singleton and random motif pairs. Therefore, we seek for more efficient measurements.

P-scores

The problem of emerging significance is its reliance on the support of a motif pair. From a biology perspective, we know that not every occurrence of a motif pair will result an interaction. Hence, we define the contributive support of motif pairs to reflect the true contributor of motif pairs for an interaction. Our definition of P-scores is based on the concept.

Definition 4.6. [**Contributive support for a motif pair**] *The contributive support of a motif pair MPr in* $\mathbb{D}$ *is the number of protein pairs in* $\mathbb{D}$ *whose interaction is partially contributed by MPr, denoted by* $\pi^c(MPr, \mathbb{D}) = |\{PPr^i \in \mathbb{D} | MPr \subseteq PPr^i, MPr \text{ contributes } PPr^i\}|$, *or simply denoted by* $\pi^c(MPr)$.

The contribution in above definition means that the occurrence segments of the motif pair are in the interaction sites if the structural data of protein complexes is available. If the data is not available, contributive support is only a theoretical concept. Later on, we will show how to estimate contributive support values based on a sequence dataset $\mathbb{D}$ of interacting protein pairs and a set of motif pairs.

Corresponding to Z-scores (Atteson, 1998) to measure the significance of single motifs with regard to $\mathbb{P}$, we define P-scores to measure the significance of motif pairs. Given an $MPr = \{\mathcal{M}_L, \mathcal{M}_R\}$ and a protein interacting sequence dataset $\mathbb{D}$,

$$p_s(MPr, \mathbb{D}) = \frac{\pi^c(MPr, \mathbb{D})}{exp(MPr, \mathbb{D})} \tag{4.4}$$

where $exp(MPr, \mathbb{D})$ is expectation support of random co-occurrences of MPr in $\mathbb{D}$.

Based on the Z-scores of single motifs and P-scores of motif pairs, now we define significant motif pairs:

Definition 4.7. [**Significant motif pairs**] *A motif pair* $MPr = \{\mathcal{M}_L, \mathcal{M}_R\}$ *is significant in a protein interacting sequence dataset* $\mathbb{D}$ *and the corresponding protein set* $\mathbb{P}$ *if* $z_s(\mathcal{M}_L, \mathbb{P}) \geq \tau_L$, $z_s(\mathcal{M}_R, \mathbb{P}) \geq \tau_R$, *and* $p_s(MPr, \mathbb{D}) \geq \tau_B$, *where* $\tau_L \geq 0, \tau_R \geq 0, \tau_B \geq 1$ *are pre-set thresholds.*

This definition emphasizes that the observations should be far away from the expectation values.

Computation of P-scores

Computationally calculating P-scores is not straightforward because the accurate contributive support is almost impossible to be obtained without wet-experimental examina-

tion. So, we present an approximate solution. First, assume $\mathcal{M}_L$ and $\mathcal{M}_R$ are independent, the expectation can be calculated as follows:

$$exp(MPr, \mathbb{D}) = n * \frac{\pi(\mathcal{M}_L)}{m} * \frac{\pi(\mathcal{M}_R)}{m} \tag{4.5}$$

where m is the number of unique proteins in $\mathbb{D}$. Therefore, the P-score can be re-written as

$$p_s(MPr, \mathbb{D}) = \frac{m^2 * \pi^c(MPr)}{n * \pi(\mathcal{M}_L) * \pi(\mathcal{M}_R)} \tag{4.6}$$

Assume an interaction contains only one binding motif pair, then, the contribution of a motif pair to a protein pair is influenced by other motif pairs. Given a sufficiently large set of motif pairs S_{MPr}, we can estimate the contributive support using the following

$$\begin{aligned} \pi^c(MPr) &= \lim_{|S_{MPr}|\to\infty} \sum_{i=1}^{n} \frac{p_s(MPr,\mathbb{D})}{\sum_{MPr' \in S_{MPr}, MPr' \subseteq PPr^i} p_s(MPr',\mathbb{D})} \delta_i(MPr) \\ \delta_i(MPr) &= \begin{cases} 1 & if\ MPr \subseteq PPr^i \\ 0 & otherwise \end{cases} \end{aligned} \tag{4.7}$$

It can be seen that for a motif pair, the supports of its two contained motifs are fixed values in a given protein set $\mathbb{P}$. So, when handling a large motif pair set S_{MPr}, formula (4.6) and (4.7) will consist of a large group of equations with two types of variables: the P-scores and the contributive supports of the motif pairs.

To solve this group of equations, we explore the use of iterative programming. First we set an identical initial value for the P-score of every motif pair. Then we use the current P-scores to calculate the contributive support for all motif pairs by formula (4.7). We can thereafter get new P-scores using formula (4.6) for each motif pair and start a new round of calculation, until the changes of most variables are less than a threshold.

Given a set of motif pairs S_{MPr} and a protein interacting sequence dataset $\mathbb{D}$, the convergence of P-scores for the motif pairs in S_{MPr} is a problem. We use the overall score difference between continuous rounds to evaluate the convergence, where the overall

difference between the j-th and the $(j-1)$-th iteration is calculated by an index $\Delta(j)$

$$\Delta(j) = \frac{2 * \sum_{MPr^i \in S_{MPr}} (p_s(MPr^i, \mathbb{D})_j - p_s(MPr^i, \mathbb{D})_{j-1})^2}{\sum_{MPr^i \in S_{MPr}} ((p_s(MPr^i, \mathbb{D})_j)^2 + (p_s(MPr^i, \mathbb{D})_{j-1})^2)} \tag{4.8}$$

.

We observe that the P-scores of most motif pairs ($> 90\%$) in the iterative process are convergent, either monotonically increasing or monotonically decreasing. But we also observe some motif pairs with waved P-scores. Therefore, the convergence is not guaranteed, from both theoretical and practical aspects.

4.4 Algorithm and Results Overview

4.4.1 Overall Algorithm of the Fixed Point Model

The overall flow of our heuristic fixed point model with significance evaluation is summarized as follows:

Input: A sequence dataset $\mathbb{D}$ of interacting protein pairs, a complex dataset $\mathcal{T}$

Output: A set of stable and significant motif pairs (fixed points) S_{MPr}

for all complex CPL in $\mathcal{T}$ **do**

for all protein pair $\mathcal{P}_a$ and $\mathcal{P}_b$ in CPL **do**

compute contact sites, and then find the set of maximal contact segment pairs S_{SPr};

end for

end for

for all maximal contact segment pair SPr in S_{SPr} **do**

generalize SPr to produce a starting motif pair MPr

end for

for all starting motif pair MPr **do**

transform MPr to either a stable motif pair MPr' or an emptyset by $f_{\mathbb{D}}$.

end for

for all stable motif pair MPr' **do**

filter those stable motif pairs MPr' which are not significant

end for

4.4.2 Data and Parameters

We use two interaction datasets to test the algorithm: a sequence dataset of interacting protein pairs collected by von Mering et al. (2002), and a protein complex dataset derived from PDB (`http://www.rcsb.org/pdb/`). The sequence dataset consists of 78390 non-redundant interactions, containing almost all the latest interacting protein pairs in yeast genome produced by various experimental and high confident computational methods. The protein complex dataset was generated from the PDB on the 9th of June, 2003, containing 1533 entries that have at least two chains, by using online search tools in the PDB-REPRDB (`http://mbs.cbrc.jp/pdbreprdb-cgi//reprdb_query.pl`). In this complex dataset, the maximum pairwise sequence identity between any two complexes is 30% and each complex has a structure resolution of 2.0 or higher.

In the computation of contact residues in a complex, we set the distance threshold as $5\mathring{A}$, that is, every residue/atom pair which have a distance less than $5\mathring{A}$ is regarded to be in contact. In the computation of maximal contact segment pairs, we required that every contact segment should contain at least four residues. In the generation from maximal contact segment pairs to starting motif pairs, we set different λ thresholds for local alignment based on the segment lengths: λ was set strictly for short segments but loosely for long segments. Actual λ values used in this study is referred to Figure 4.3.

After obtaining starting motif pairs from the complex dataset, we conducted the transformation process to find stable motif pairs from the sequence dataset of interacting protein pairs. For a motif pair MPr, to discover $f_{\mathbb{D}}(MPr)$—the consensus pattern—and subsequently $f_{\mathbb{D}}(\cdots f_{\mathbb{D}}(f_{\mathbb{D}}(MPr)))$ until a stable state, we computed a latter cluster based

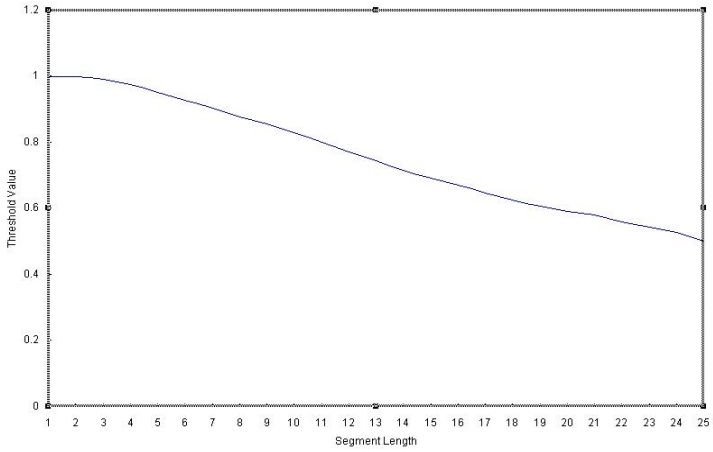


Figure 4.3: The threshold for local alignment with respect to different segment lengths.

on its previous cluster instead of the whole dataset (according to Proposition 3.1). The efficiency was therefore greatly improved.

After obtaining a set of stable motif pairs from the starting motif pairs and the refinement, we filtered the insignificant ones. The thresholds for the significance indices were set as: $\tau_L = 0, \tau_R = 0, \tau_B = 1$. The computation of the supports and Z-scores are straightforward according to our algorithm. However, the computation of the P-scores is an iterative process. The initial P-score for every motif pair was set as 1.0 in this work. We use Δ defined in Equation 4.8 to evaluate the overall score difference between adjacent iterations. If $\Delta < 0.01$, we stop the iterative process. For most sets of motif pairs, the process could stop within four iterations.

4.4.3 Results Overview

In total, we discovered 765 binding (stable and significant) motif pairs from the sequence dataset of interacting protein pairs using 1403 maximal contact segment pairs identified from the protein complex dataset. Table 4.1 provides these results and other related results such as the support information.

The P-sore values of the 765 stable and significant motif pairs differ very much from

Table 4.1: The overall results of our fixed point model

Num of Contact Segment Pairs	Num of Starting Motif Pairs	Num of Stable Motif Pairs	Num of Stable & Significant Motif Pairs	Support of Stable Motif Pairs	Support of Stable & Significant Motif Pairs
1403	1222	913	765	122193	107028

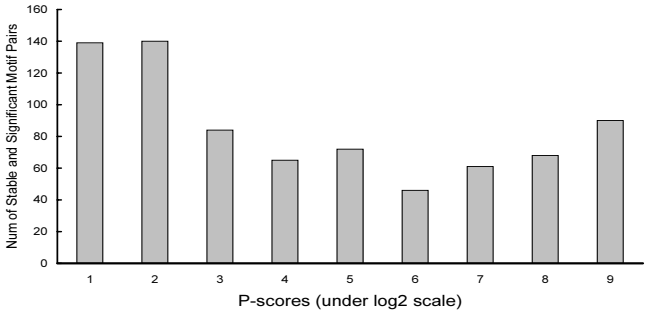


Figure 4.4: The distribution of the P-scores (under $\log_2$) for our 765 stable and significant motif pairs.

one another. Figure 4.4 shows the distribution of these P-scores (under $\log_2$ scale). It can be seen that our algorithm can discover motif pairs with both high and low P-scores (larger than a threshold).

Besides P-scores, another important information is the support. The distribution of the support values (under $\log_2$ scale) of the 765 stable and significant motif pairs is depicted in Figure 4.5. It can be seen that our algorithm preferred to discover motif pairs with relatively low supports. This is an advantage to our algorithm as the support of many real binding motif pairs is quite possible to be low in an incomplete dataset. The

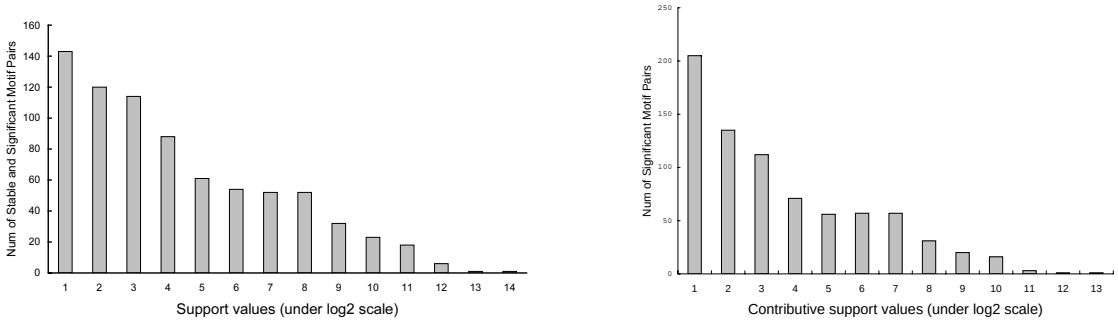


Figure 4.5: The distribution of the absolute support values and contributive support values (under $\log_2$ scale) of our 765 stable and significant motif pairs.

distribution of the estimated contributive support values for our discovered motif pairs exhibits almost the same shape as that of absolute support values, depicted in Figure 4.5.

To evaluate the lengths of our discovered motif pairs, we used *information content* (Tompa, 1999) as the index. Assume each residue has equal distribution, the information content of a motif $M = \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k$ can be computed by:

$$I(M) = k\log_{10}|\Sigma| - \sum_{i=1}^{k}\log_{10}|\mathcal{A}_i| \tag{4.9}$$

For a motif pair $MPr = \{M_L, M_R\}$, we define

$$I(MPr) = I(M_L) + I(M_R) \tag{4.10}$$

So, the information content largely reflects the length of a motif. The distribution of the information contents of the 765 motif pairs is shown in Figure 4.6. It can be seen that most of the motif pairs have an information content between 10 and 20, except for very few cases. Therefore, these motif pairs roughly have residues between 10 and 20.

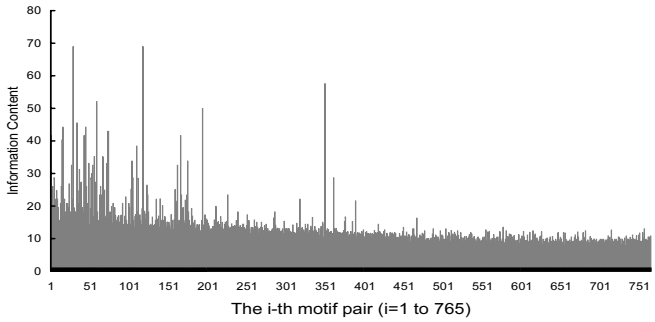


Figure 4.6: The distribution of information content of our discovered stable and significant motif pairs.

4.5 Effectiveness Comparison with Random Patterns

To demonstrate that our discovered stable and significant motif pairs are credible and also to illustrate the generalization from our maximal contact segment to the starting motif pairs makes benefits to the discovery, we conduct a comprehensive computational comparison between our patterns and random patterns. These experiments include (1) the comparison between our 913 stable motif pairs and 10 random sets each consisting of 913 random motif pairs; and (2) the comparison between our 1222 starting motif pairs and 10 random sets each consisting of 1222 random starting motif pairs; and (3) the comparison between our 1403 maximal contact segment pairs and 10 random sets each consisting of 1403 random segment pairs.

A random motif pair or a random segment pair is generated by substituting every residue in our pattern with a random residue. Therefore, the random pattern has the same length as ours. The distribution of the randomly generated residues follows the same distribution of all the residues in the contact sites of our complex dataset. [In fact, it has no significant difference between this distribution and that in the whole yeast genome (Fariselli et al., 2002)].

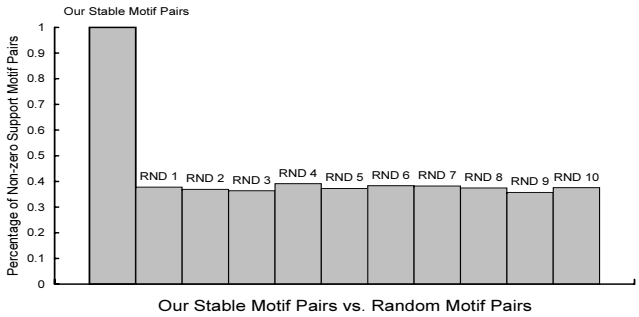


Figure 4.7: The percentage of non-zero support motif pairs in our discovered stable motif pairs and those in 10 sets of equal size of random motif pairs.

First, we compare our 913 stable motif pairs with the 10 sets of random motif pairs of equal size to see how much percentage of them are significant. We observed that

- About two-thirds of the random motif pairs have a zero-support in the interaction dataset $\mathbb{D}$, namely $\pi(MPr^{random}, \mathbb{D}) = 0$. However, for every MPr of our 913 stable motif pairs, $\pi(MPr, \mathbb{D}) \neq 0$. Figure 4.7 shows the percentage of random patterns having non-zero support for the 10 rounds of random experiments.

- Only about one-ninth of the random motif pairs are significant. However, 84% of our 913 stable motif pairs are significant. Complete results are shown in Figure 4.8.

- The total support of our stable and significant motif pairs is much larger than that of significant random motif pairs, which is shown in Figure 4.9.

These results indicate that our discovered stable motif pairs are much more statistically significant than random patterns. Therefore, they are most likely to be potential binding motif pairs.

Second, we substitute our 1222 starting motif pairs with random starting motif pairs to see how much percentage of stable motif pairs can be discovered, and how much percentage

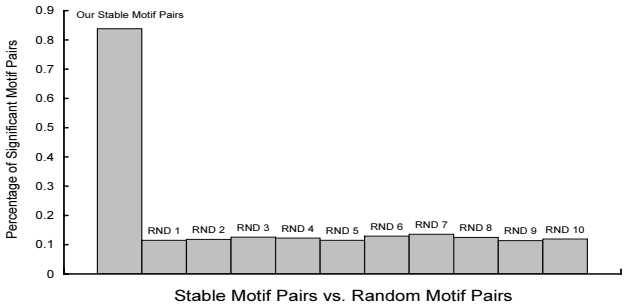


Figure 4.8: The percentage of significant motif pairs for our discovered stable motif pairs and those for 10 sets of equal size of random motif pairs.

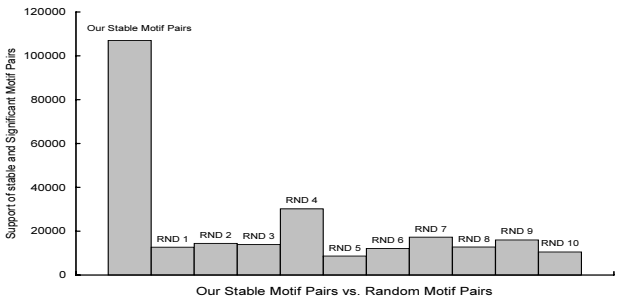


Figure 4.9: The total support of our discovered stable and significant motif pairs and those for 10 sets of equal size of random motif pairs.

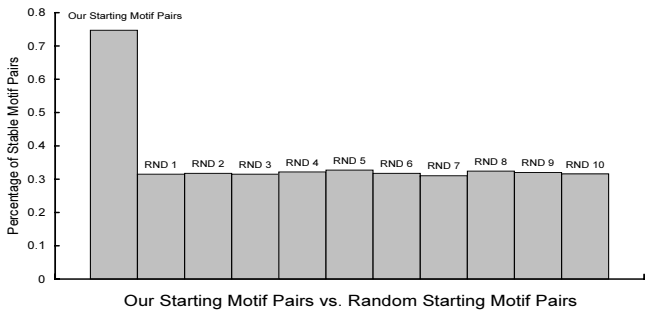


Figure 4.10: The percentage of stable motif pairs derived from our starting motif pairs and those derived from 10 sets of equal size of random starting motif pairs.

of stable and significant can be discovered. Such substitution is repeated for 10 times. We observed that

- Our starting motif pairs can lead to 75% (913) of stable points, but those random starting points in each round lead to $< 33\%$ of stable motif pairs. Complete results are shown in Figure 4.10.

- Our starting motif pairs can lead to $\sim 63\%$ of stable and significant motif pairs, but $< 18\%$ of those random starting points can lead to stable and significant motif pairs. Figure 4.11 shows complete results.

From these comparisons, we conjecture that the generalization from maximal contact segment pairs to our starting motif pairs is a useful method because it contributes much more number of stable and significant motif pairs than the random method does.

Third, we substitute, repeatedly 10 times, our 1403 maximal contact segment pairs with random segment pairs and then perform the transformation to find stable and significant motif pairs. We found that

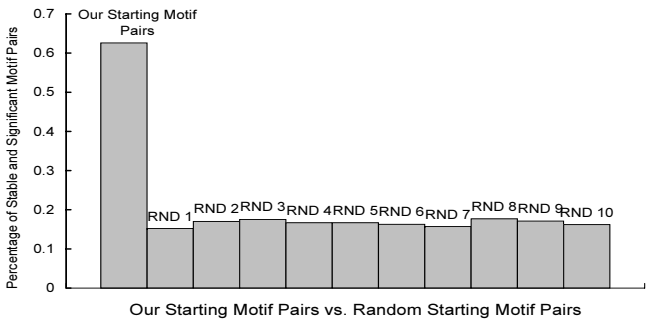


Figure 4.11: The percentage of stable and significant motif pairs derived from our starting motif pairs and those derived from 10 sets of equal size of random starting motif pairs.

- The support of our discovered significant and stable motif pairs is nearly 40% larger than that of random segment pairs. And no set among the 10 rounds of random experiments can lead to more support than our maximal contact segment pairs. Complete results are shown in Figure 4.12.

This comparison means that the maximal contact segment pairs have more supporting evidence in the interaction dataset $\mathbb{D}$ than random segment pairs. This also confirms our assumption that using the complex dataset is a good way to get some real biological support.

From these various randomization experiments, we can see that the stable and significant motif pairs that we discovered are far way from random expectation, which benefits from the choice of maximal contact segment pairs. Therefore, it is reasonable that they provide much information to find real binding motif pairs. This is also confirmed by our literature searching results reported in the next section.

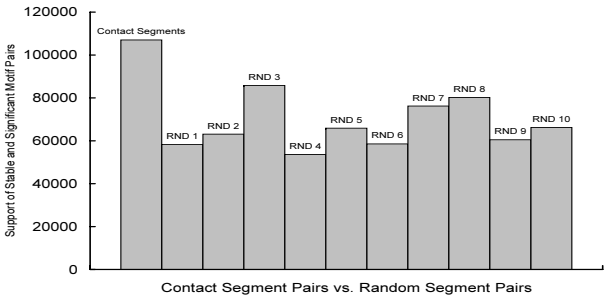


Figure 4.12: The total support of stable and significant motif pairs derived by our maximal contact segment pairs and those from random segment pairs.

4.6 Literature Validation

To demonstrate the biological significance of our discovered patterns, ideally, they should be validated by wet-experimental methods. Unfortunately, there are few well-known wet-experimental methods that can determine the two sides of binding motif pairs simultaneously. As reviewed in Chapter 2, current available technique such as phage display (Smith, 1985a) and NMR (Takahashi et al., 2000) can only determine one side of the interaction sites and produce protein–peptide/motif binding pairs or domain–motif interacting pairs. As a result, there is still limited data about the interaction sites, mostly spanning across various individual literature, without an integrative and comprehensive database available, which makes our validation even harder.

Nevertheless, we still find some evidences to show the biological significance of our discovered patterns. First, we check the coincidence of the *individual* motifs in our motif pairs with the reported binding motifs determined by various wet-experimental methods. For example, using key words 'binding motif OR site AND mutagenesis', we extracted 202 binding motifs from the abstracts of NCBI PUBMED; 89 of them have at least three positions compatible to ours and 40% overall similarity. Of these 89 binding motif pairs, 42 motifs are highly similar to our discovered motifs, having at least four positions compatible

and 50% overall similarity. We show in Table 4.2 the top 5 matches in comparison with mutagenesis. We show in Table 4.3 the top 5 matches in comparison with phage display with key words 'binding motif OR site AND phage display'.

Table 4.2: Motif coincidence with the mutagenesis method

Our Motif	Mutagenesis Motif	PMID of Mutagenesis Motif
{g}{s}{g}{k}{t}	{g}*{g}{k}{t}	10464259
{a}{l}{e}{t}{s}	{l}{e}{t}{s}	11435317
{p}{iv}{d}{l}{s}	{p}{v}{d}{l}{s}	11373277
{l}{dn}{l}{l}	{l}{l}{d}{l}{l}	11451993
{k}{de}{k}{ek}	{k}{e}{k}{e}	10748065

Table 4.3: Motif coincidence with the phage display method

Our Motif	Phage Display Motif	PMID of Phage Display Motif
{g}{ly}{d}{iy}{iv}	{y}{d}{y}{v}	11389136
{g}{iv}{g}{fi}{iv}	{k}{v}{g}{i}{v}	12110480
{s}{dgh}{ek}{d}	{i}{s}{h}{k}{d}{m}{q}{l}{g}	9373320
{g}{s}{g}{k}{t}	{g}{h}{n}{g}{s}{g}{k}{s}{t}{l} {a}{k}{t}{i}{n}	12110480
{a}{iv}{a}{g}	{e}{l}{s}{g}{g}{q}{m}{r}{r}{v}{a}{i}{a}{g}{v}	12110480

Second, we check our discovered motif pairs with protein–peptide/motif binding pairs determined mainly by phage display. First, we identify the individual motifs in our population of discovered motif pairs that match closely with a binding peptide/motif in the literature. Then, for each of such matched motifs, we verify whether the motif on the other side of the corresponding motif pairs can be found in the protein known to bind the particular peptide/motif.

Table 4.4: The coincidence between our motif pairs and motif-actin binding pairs

Actobindin Motif	Left Motif	Confirmed Right Motif in Actin
{v}{th}{v}{k}{k}{v}	{iv}{t}{iv}{k}	{a}{ek}{iv}{fl}{g}{kr}
{v}{th}{v}{k}{k}{v}	{iv}{ek}{k}{flv}{de}	{ek}{il}{l}{p}
{v}{th}{v}{k}{k}{v}	{ek}{iv}∅{ilv}{ek}	{g}{k}{k}{il}{v}{s}

We describe three examples to explain the biological significance of our discovered motif pairs using this validation method. As the first example, Vancompernolle et al. (1991) reported that protein actobindin contains an actin-binding motif $\{v\}\{th\}\{v\}\{k\}\{k\}\{v\}$. From our discovered 913 stable motif pairs, we observed that there are three motif pairs containing motifs that are similar to the actin-binding motif $\{v\}\{th\}\{v\}\{k\}\{k\}\{v\}$. The left side and right side of the three motif pairs are listed in the second and third column of Table 4.4 respectively. A more interesting observation is that the three right-side motifs are all contained in the sequence of the protein actin or its associated proteins.

The second example is shown in Table 4.5. Tumbarello et al. (2002) studied the interaction sites of protein paxillin and its interacting proteins. The interaction site of paxillin is in the form of {l}{d}*{l}{l}**{l}, shown in the first column of Table 4.5. Our method discovered three similar motifs as shown in the second column of Table 4.5. The other side of the corresponding motif pairs is shown in the third column of the table. All of them have been found to exist in the interacting proteins reported in the literature (Tumbarello et al., 2002), which are shown in the last column of the table with the matched segments.

As the final example, Kay et al. (2000) had a study on the interaction of proline-rich motifs in signaling proteins with their cognate domains. Four binding motifs [called binding consensus sequences in (Kay et al., 2000)] are listed in the first column of Table 4.6. From our discovered binding motif pairs, we observed that there are four motif pairs containing a motif that is similar to one of the four binding motifs. The four motif pairs are listed in the second and third columns of Table 4.6. Another observation is that our right-side motifs are all contained in the proteins in the last column of Table 4.6 which

Table 4.5: The coincidence between our discovered motif pairs and the interaction sites between paxillin and its binding proteins

Paxillin Motif	Left Motif	Right Motif	Confirmed Proteins
{l}{d}*{l}{l}**{l}	{d}{il}{l}{il}	{st}{d}{ek}{a}	Vinculin,FAK
{l}{d}*{l}{l}**{l}	{il}{dg}{iv}{l}{d}	{d}{ek}{e}{g}{i}	PYK2({d}{ek}{e}{g})
{l}{d}*{l}{l}**{l}	{l}{fl}{v}{l}{k}	{l}{fl}{v}{l}{k} PYK2({l}{fl}{v}{l})	Vinculin({l}{fl}{v}{l})

Table 4.6: The coincidence between our motif pairs and peptide-protein binding pairs

Binding motif	Left motif	Right Motif	Confirmed Binding Protein
{p}*{l}{p}*{kr}	{p}{ek}*{p}	{g}{v}{fi}{s}	CRK A
{rkh}{p}{p}{ailvp}{p}{ailvp}{k}{p}	{p}{iv}{ep}{iv}{a}	{a}{a}{s}{fi}	Cortactin
{r}{l}{p}*{l}{p}	{p}{ek}*{p}	{g}{v}{fi}{s}	Synaptojanin I
{rkh}{p}{p}{ailvp}{p}{ailvp}{k}{p}	{p}{iv}{dp}{p}{fv}	{p}{iv}{dp}{p}{fv}	Shank

are reported to bind to the corresponding binding motif in the first column (Kay et al., 2000).

In the remainder of the section, we check our discovered motif pairs in more details with domain–motif interacting pairs in the literature as they are our most similar patterns. Similarly, we identify the individual motifs in our population of discovered 765 stable and significant motif pairs that match closely with a binding motif in the literature. Then, for each of such matched motifs, we verify whether the motif on the other side of the corresponding motif pairs can match in the domains known to bind the particular motif. We give full details to see how they are discovered, where their origins are, and what the biological significance is.

The first example motif pair is

$$MPr_{example1} = \{\{l\}\{dn\}\{l\}\{l\}, \{ek\}\{lv\}\{g\}\{d\}\{g\}\}.$$

Its origin is located at the so-called pdb3daa protein complex. Specifically, the motif

$\{l\}\{dn\}\{l\}\{l\}$ is evolved from the segment $lnll$ at the chain A of the pdb3daa complex, indexed from 147 to 150 residues in the chain A. The motif $\{ek\}\{lv\}\{g\}\{d\}\{g\}$ is rooted at the segment $yqfgdg$ at the chain B of the pdb3daa complex, indexed from 24 to 29 residues in the chain B. See Figure 4.13.

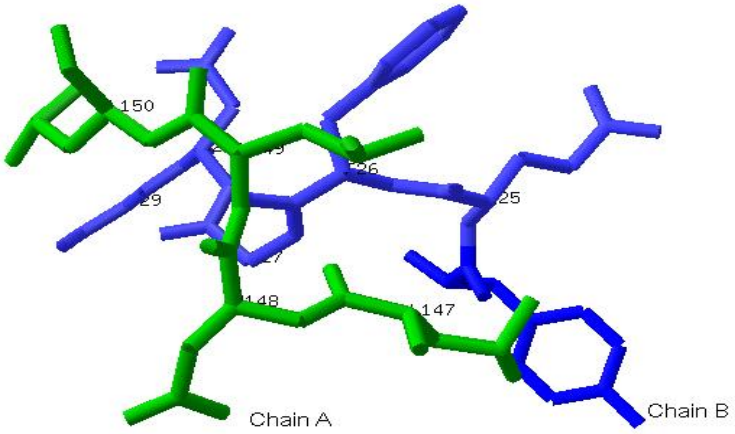


Figure 4.13: Three-dimensional structure of an interaction site in the pdb3daa protein complex, a D-amino acid aminotransferase in species thermophilic bacterium ps3. Chain A is in green color, Chain B is in blue color.

The segment pair, (lnll,yqfgdg) between chain A and chain B, is a maximal contact segment pair. We use Figure 4.14, abstracted from Figure 4.13, to demonstrate it.

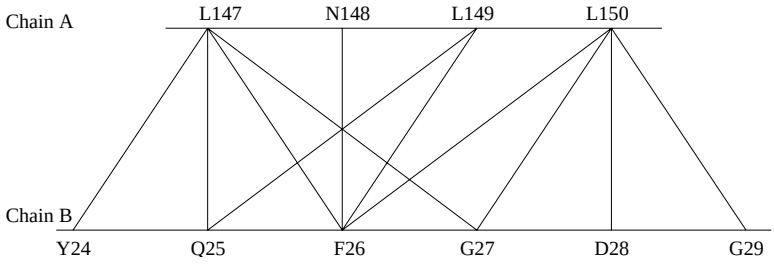


Figure 4.14: A maximal contact segment pair discovered from the pdb3daa complex. A line between Chain A and Chain B represents that the two corresponding amino acids are close in distance.

This maximal contact segment pair (lnll,yqfgdg) is then generalized to the following

starting motif pair MPr_{start1}

$$MPr_{start1} = \{\{l\}\{dn\}\{l\}\{l\}, \{ek\}\{lv\}\{g\}\{d\}\{g\}\}$$

for the function $f_{\mathbb{D}}$. During the generalization, the residues in the segment pair may be extended to their structure interchangeable residues (or be removed in the margins). More specifically, the residue n is extended to residue d, the residue y is removed it is not conserved at this position, the residue q is replaced by e and k and the residue f is replaced by l and v (Azarya-Sprinzak et al., 1997). After generalization, the overall structures of the mutants are usually maintained even with the considerable changes of sequences, thus, the interactions between the mutated segments are still expected to maintain[1].

Interestingly, we found that $f(MPr_{start1}) = MPr_{start1} = MPr_{example1}$. That is, this starting motif pair MPr_{start1} itself is a stable motif pair.

We found that this stable motif pair $MPr_{example1}$ is statistically significant after examining its support level and P-score against random motif pairs. The support of motif $\{l\}\{dn\}\{l\}\{l\}$ is 265 in $\mathbb{P}$, the support of motif $\{ek\}\{lv\}\{g\}\{d\}\{g\}$ is 13 with respect to the same protein set $\mathbb{P}$. The support of $MPr_{example1}$ as a pair is 58 in the protein interaction sequence data set $\mathbb{D}$. The P-score of $MPr_{example1}$ as a pair is 6.15 with respect to the data set $\mathbb{D}$. Then, we generated 1000 random motif pairs according to $MPr_{example1}$, using the method described in Section 4.5. For these 1000 random motif pairs, the average support of the random motifs corresponding to $\{l\}\{dn\}\{l\}\{l\}$ is 32.91, the average support of the random motifs corresponding to $\{ek\}\{lv\}\{g\}\{d\}\{g\}$ is 4.41. The average support for those 1000 motif pairs is 1.83 in the protein interaction sequence data set $\mathbb{D}$. From these results, we can see that $MPr_{example1}$ has occurrence much more than its random expectation either in single motifs or in pairs. Hence, the stable motif pair $MPr_{example1}$ is not a random result indeed.

[1]As partial evidence, segment pair (ldll,evgdg) occurs at eight complexes (pdb1jez, pdb1k8c, pdb1mi3, pdb1r38, pdb1sm9, pdb1ye4, pdb1ye6, pdb1z9a) in the current PDB database, moreover, the segment ldll locates at the interaction sites of these eight complexes.

We also found much biological significance of the motif pair $MPr_{example1}$. Doray and Kornfeld (2001) found a protein motif $M_{DK} = \{l\}\{l\}\{d\}\{l\}\{l\}$, a functional variant of the $\{l\}\{l\}\{n\}\{l\}\{d\}$ motif within the beta 1 subunit of AP-1, was biologically confirmed to bind to the terminal domain of the clathrin heavy chain. From the sequence of this terminal domain, we find that there exists a segment $elgd$ near the end part of this domain. Comparing these biological results and our computational results, we can see that

- $M_{DK} = \{l\}\{l\}\{d\}\{l\}\{l\}$ is similar to the left motif $\{l\}\{dn\}\{l\}\{l\}$ of our motif pair $MPr_{example1}$.
- The segment $elgd$ matches well with our right motif $\{ek\}\{lv\}\{g\}\{d\}\{g\}$ of $MPr_{example1}$. The precise position of the segment $elgd$ is from positions 462 to 465 at the end of the globular terminal domain (from 1th to 479th) of clathrin heavy chain 1 of human.
- Besides, our left motif $\{l\}\{dn\}\{l\}\{l\}$ is similar to $\{l\}\{l\}\{d\}\{l\}\{l\}$ and $\{l\}\{l\}\{n\}\{l\}\{d\}$ both of which share the same functions.

The second example motif pair is

$$MPr_{example2} = \{\{g\}\{ly\}\{d\}\{iy\}\{iv\}, \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}\{l\}\}.$$

Its origin is located at the so-called pdb1ors protein complex (Jiang et al., 2003). Specifically, the motif $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is evolved from the segment $gydyf$ at the chain B of the pdb1ors complex, indexed from 99 to 103 residues. The motif $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is rooted at the segment $aglglfrl$ at the chain C of the pdb1ors complex, indexed from 111 to 118 residues. See Figure 4.15.

The segment pair, ($gydyf$,$aglglfr$) between chain B and chain C, is a maximal contact segment pair. We use Figure 4.16, abstracted from Figure 4.15, to demonstrate it. The maximal segment pair is then generalized to the following starting motif pair X,

$$X = \{\{g\}\{ly\}\{d\}\{fiy\}\{fiv\}, \{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{dr\}\{fi\}\}$$

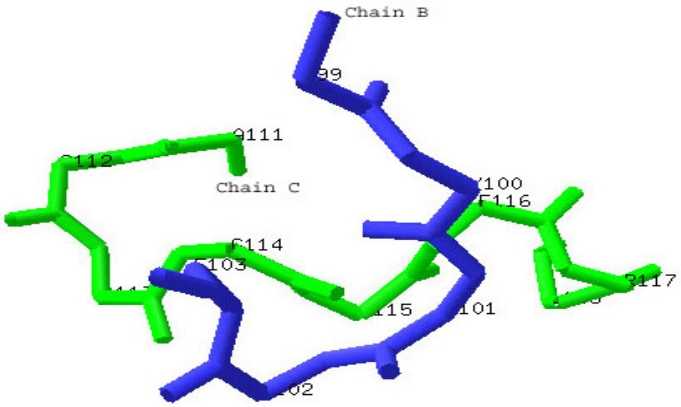


Figure 4.15: Three-dimensional structure of an interaction site in the pdb1ors protein complex, a complex between the kvap potassium channel voltage sensor and an fab in species mouse and E. Coli., where Chain B is in blue color, and Chain C is in green color.

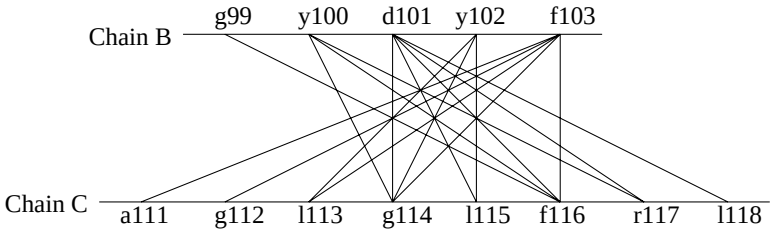


Figure 4.16: A maximal contact segment pair discovered from the pdb1ors complex. A line between Chain B and Chain C represents that the two corresponding amino acids are close in distance.

for the function $f_{\mathbb{D}}$. After one step of transformation by $f_{\mathbb{D}}$, this starting motif pair X becomes the fixed point $MPr_{example2}$, i.e. $f_{\mathbb{D}}(X) = MPr_{example2}$.

We found that this motif pair $MPr_{example2}$ is statistically significant after examining its support level against random motif pairs. The support of motif $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is 15 in yeast protein set $\mathbb{P}$, and the support of motif $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is 2 with respect to the same protein set $\mathbb{P}$. The support of $MPr_{example2}$ as a pair is 6 in the protein interaction sequence dataset $\mathbb{D}$. Then, we generated 1000 random motif pairs according to $MPr_{example2}$, using the same method describe in Section 4.5. For these 1000 random motif pairs, the average support of the random motifs corresponding to $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ is

11.14, the support of every random motif corresponding to $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r\}\{f\}$ is 0. Consequently, the support for any of those 1000 motif pairs is also 0 in the protein interaction sequence dataset $\mathbb{D}$. From these results, we can see that $MPr_{example2}$ has occurrence much more than its random expectation in single motifs or in pairs. Therefore, the stable motif pair $MPr_{example2}$ is not a random result indeed.

We also found some biological significance of the motif pair $MPr_{example2}$. Pellicena and Miller (2001) studied a protein motif $M_{PM} = \{y\}\{d\}\{y\}\{v\}$ within the protein p130Cas of v-Src transformed cells. This motif was biologically confirmed to bind to the Src homology 2 (SH2) domain that is a protein domain with about 100 amino-acid residues in many intracellular signal-transducing proteins (Russell et al., 1992). We had the following observations after comparing these biological literature results with our computational results:

- $M_{PM} = \{y\}\{d\}\{y\}\{v\}$ is similar to the left motif $\{g\}\{ly\}\{d\}\{iy\}\{iv\}$ of our motif pair $MPr_{example2}$.
- The segment $lvrf$ in the SH2 domain partially matches to our right motif $\{r\}\{g\}\{l\}\{g\}\{l\}\{v\}\{r$ of $MPr_{example2}$. The precise location of the segment $lvrf$ is from positions 118 to 121 at the SH2 domain of the protein $SH2A\_HUMAN$, and from positions 139 to 142 at the SH2 domain of the protein $SH2A\_MOUSE$. At the left side of the matched segments in the SH2 domain, there is a segment $qgcy$ from 114 to 117 in $SH2A\_HUMAN$. The residue q at position 114 of this segment is a structure interchangeable residue of r (Azarya-Sprinzak et al., 1997); the residue g at position 115 exactly matches with the second residue in our motif; at position 116, both residue c and l are hydrophobic residues that imply some structure similarity; at position 117, both residue y and residue g are surface residues (charged/polar residues). Similarly, we find a segment gcy from 136 to 138 in $SH2A\_MOUSE$. Hereby, the right motif of $MPr_{example2}$ has five positions which are exact matches and two positions which are compatible with the biological protein sequences (from a domain of 92 residues).

- There are total 295 proteins containing SH2 domains, where the segment lvr occurs in 139 of them. (This can be seen from the prosite: `http://tw.expasy.org/prosite/`.) Moreover, the segment lvr locates near the most conserved region in the domain, where the most conserved region is just between g—the second residue and r—the last second residue. (See `http://tw.expasy.org/cgi-bin/aligner?psa=PS50001&color=1&maxinsert=10&linelen=0`). This implies that the motif pair we discovered is likely to be the most critical factor for the binding between the $\{y\}\{d\}\{y\}\{v\}$ motif in p130Cas and SH2 domain.

These literature validations indicate that the stable and significant motif pairs discovered by our fixed-point based method would possess a strong biological meaning. An important implication of this is that our discovered binding motif pairs are likely to be real biological interaction sites. Therefore, this computational method would have a potential guidance role to play for the identification of real biological interaction sites.

4.7 Discussions

To tackle the computational difficulties in the fixed point model, we propose a heuristic approach in this chapter. The approach can be regarded as an in-silico mutagenesis. In both approaches, experimentally determined interaction sites are used as origins. They are mutated and their binding affinities after mutations are examined. Consensus patterns of affinity mutants are then summarized. In both approaches, the original interaction sites are mutated either randomly or into structure-similar mutants. The difference is that the binding affinities of mutants in mutagenesis are obtained directly through assays, while the binding affinities of our motif pairs are evaluated as the stabilities or resistances upon a transformation function.

In the heuristic approach, we use maximal contact segment pairs to represent interaction sites in protein complexes. We require each residue in one segment having at

least one contact residue in the other segment. Biologically, it is unnecessary because contact segment pairs are still valid even if a few residues among them are not in contact. Computationally, however, our top-down recursive algorithm for finding maximal contact segment pairs will no longer be valid without this constraint. Therefore, more work should be carried out to explore the relaxation of this constraint while retaining the efficiency of the algorithm, to reveal more significant binding motif pairs.

We call the maximal contact segment pairs with some relaxation as maximal gapped contact segment pairs, while the previous ones are maximal continuous contact segment pairs. To generate such gapped segment pairs, an extension strategy is possible. The strategy is based on the principle that a maximal gapped contact segment pair most likely contains some shorter maximal continuous contact segment pairs. Hereby, we can first work out all maximal continuous contact segment pairs, using the algorithm described in Section 4.2.2, then, extend them in both directions, examining whether the relaxed constraints still hold, and stop until we fail to make further extensions.

The method to generate starting motif pairs using gapped maximal contact segment pairs is similar to that of maximal continuous contact segment pairs, except that all gaps in the pairs are regarded as a '*' character in the local alignment. The consensus searching on the aligned sequence is carried out as before including the gapped positions, to capture additional patterns.

In the heuristic approach, we examine the significance of the stable motif pairs by our proposed measurements (Z-scores and P-scores). Alternatively, traditional measurements like maximum-likelihood and expectation maximization can be applied to evaluate the significance, as done by Wang et al. (2005) and Deng et al. (2002). However, they are too complicated and time-consuming to be applied in our study. Hence, we propose our own, to reduce the computing time while keeping the efficiency.

Another alternative to evaluate the significance is to compare our motif pairs with the equal-length random patterns and to check the difference of occurrences in the same

dataset, as carried out in Section 4.5. However, the method is too computationally expensive since we need to go through a huge number of random patterns to guarantee that the significance values are accurate.

Note that our measurements have several limitations. Recall that we assume each interaction only contains one binding motif pair. In real circumstance, it is somewhat simplified. In most cases, multiple binding motif pairs cooperate with each other to build up a single interaction. Taking this into consideration, our computation will obviously underestimate the values of the contributive supports. Another limitation of our measurements is that we assume a large number of binding motif pairs before evaluating the computation, which is not always available. As a result, the computation of significance for a motif pair will be a little bit inaccurate.

4.8 Summary

Finally, we summarize this chapter. It is a computationally challenging problem to work out a complete set of stable motif pairs from a large sequence dataset of interacting protein pairs for the fixed point model proposed in the previous chapter. To overcome the computational difficulty, we present a heuristic approach that is guided by a biologically reliable protein complex structural dataset. The key idea in the heuristic approach is using maximal contact segment pairs to represent interaction sites in protein complexes and generalizing the segment pairs into our crucial patterns—starting motif pairs that lead to stable motif pairs by our transformation function. We have formulated the extraction of maximal contact segment pairs from protein complexes as a novel computational search and optimization problem, and have provided an efficient algorithm for the problem.

The stable motif pairs derived by the heuristic approach may still be insignificant. To filter out insignificant ones, we present two measurements: Z-scores and P-scores, for single motifs and motif pairs respectively. Z-scores are widely accepted measurements but P-scores are novel measurement proposed by us. The significance of P-scores of motif

pairs is their unexpected contributive frequency in the same sequence dataset comprising known interacting protein pairs. We have presented methods to compute them efficiently.

We have applied the heuristic approach to a huge real-life dataset and found many biologically interesting motif pairs. Our comprehensive comparison results have shown that our discovered binding motif pairs are much more statistically significant than random motif pairs, a result from the choice of maximal contact segment pairs. We validate the discovered binding motif pairs with binding patterns in literature which were determined by experimental methods, both for single motifs and motif pairs. The validations show good matches between our motif pairs and protein/domain–motif interacting pairs reported in the literature, which demonstrates the strength of our model.

As future work of this model, we intend to collect a comprehensive database about experimentally determined interaction sites or interacting patterns and perform a systematic validation for our discovered binding motif pairs. Meanwhile, we may also collaborate with biologists to confirm some of our results using wet experiments. Besides, we may also work on different functions $f_{\mathbb{D}}$ to see whether it can be optimized.

Chapter 5

Interacting Protein Group Pairs

5.1 Introduction

In Chapter 3 and Chapter 4, we present a fixed point model to discover binding motif pairs from protein interaction sequence data and protein complex structural data. Although the model is effective, it has several deficiencies: (1) The transformation function is crucial in the model, but the current function $f_{\mathbb{D}}$ is still too simple to emulate the real evolution; (2) It is computationally difficult to find a complete solution even for the simple transformation function $f_{\mathbb{D}}$; (3) The strategy proposed in Chapter 4 highly depends on the incomplete protein complex structural data. Hence, we propose a new approach in this chapter.

The new approach discovers binding motif pairs only from protein interaction sequence data. In the new approach, protein interactions in cells are modeled by a graph, namely a *protein interaction network* (Schwikowski et al., 2000), in which a vertex is a protein and an edge is the interaction between a protein pair. By observations, we find that there exist many *most-versus-most* and even *all-versus-all* interaction subnetworks between two groups of proteins in a protein interaction network. We term the latter that exhibit an all-versus-all interaction as a pair of *interacting protein groups*. Such protein group pairs often

imply some biological meanings such as shared functions or shared binding mechanisms among proteins within the same groups due to the common interacting partners.

It is a challenging problem to discover all pairs of interacting protein groups from a proteome-wide protein interaction network because the number of combinations of proteins is exponential. From a graph theory perspective, interacting protein group pairs are similar to *maximal complete bipartite subgraphs* which also represent a kind of full connectivity between two vertices sets in a graph. The only difference between them is that interacting protein group pairs may have inner interactions within protein groups but maximal complete bipartite subgraphs strictly exclude such edges within the same vertices sets in some studies. Even with the difference, the two problems have the same level of time complexity through an easy transformation. Hereby, they are equivalent in complexity since the transformation is not computationally dominant. Due to the equivalence, several propositions about mining interacting protein groups can be obtained directly from the studies of listing all maximal complete bipartite subgraphs in a graph (see Eppstein (1994) for a review). For example, all pairs of interacting protein groups in a protein interaction network can be enumerated in time $O(\alpha^3 2^{2\alpha} m)$, where α is the arboricity and m is the number of proteins of the protein interaction network. Even though the algorithm has a linear complexity to the number of proteins, it is not practical for large protein interaction networks due to the large constant overhead (α can easily be around 10-20 in practice) (Zaki and Ogihara, 1998).

Since there are very few algorithms that are efficient for all graphs in graph theory, we study the problem practically from a data mining perspective. We interpret the adjacency matrix of a protein interaction network into *a transactional database* and associate our interacting protein groups with a data mining concept called *closed patterns*. Our main contribution in this chapter is to find the enumerating all interacting protein group pairs of a protein interaction network is equivalent to mining of closed patterns from the adjacency matrix of the protein interaction network. Since the mining of closed patterns has been extensively studied, there are many existing algorithms and implementations that can be used to solve this problem (Bastide et al., 2000; Goethals and Zaki, 2003; Grahne and

Zhu, 2003; Nicolas et al., 1999; Uno et al., 2004; Wang et al., 2003; Zaki and Hsiao, 2002). The data structures of those algorithms are efficient and the mining speed is tremendously fast in practice.

The rest of this chapter is organized as follows: In Sections 5.2 and 5.3, we provide the basic definitions and implications about interacting protein groups and closed patterns. In Section 5.4, we prove that there is a one-to-one correspondence between closed pattern pairs and interacting protein group pairs for any simple protein interaction network. In Section 5.6 and 5.6, we discuss some related works and conclude this chapter.

5.2 Definition of Interacting Protein Group Pairs

A protein interaction network $\mathbf{G} = \langle \mathbb{P}, \mathbb{D} \rangle$, which is a graph, is comprised of a set of proteins (vertices) $\mathbb{P}$ and a set of interactions (edges) $\mathbb{D} \subseteq \mathbb{P} \times \mathbb{P}$. Throughout this chapter, we assume $\mathbf{G}$ is an undirected graph without any self-loop, called a simple protein interaction network. In other words, we assume that (i) there is no interaction $(\mathcal{P}, \mathcal{P}) \in \mathbb{D}$ and (ii) for every $(\mathcal{P}, \mathcal{Q}) \in \mathbb{D}$, $(\mathcal{P}, \mathcal{Q})$ can be replaced by $(\mathcal{Q}, \mathcal{P})$—that is, $(\mathcal{P}, \mathcal{Q})$ is an unordered pair. Note that it is just to simplify the problem. A protein may interact with itself in some cases, which forms a self-loop.

Two proteins $\mathcal{P}, \mathcal{Q}$ of a protein interaction network $\mathbf{G}$ are said to be adjacent if $(\mathcal{P}, \mathcal{Q}) \in \mathbb{D}$—that is, there is an interaction between $\mathcal{P}$ and $\mathcal{Q}$ in $\mathbf{G}$. The **neighborhood** $\beta(\mathcal{P})$ of a protein $\mathcal{P}$ in a protein interaction network $\mathbf{G}$ is the set of all proteins in $\mathbf{G}$ that are adjacent to $\mathcal{P}$—that is, $\beta(\mathcal{P}) = \{\mathcal{Q} \mid (\mathcal{Q}, \mathcal{P}) \text{ or } (\mathcal{P}, \mathcal{Q}) \in \mathbb{D}\}$. The neighborhood $\beta(\mathsf{X})$ for a subset X of proteins in a protein interaction network $\mathbf{G}$ is the set of common neighborhood of the proteins in X—that is, $\beta(\mathsf{X}) = \cap_{\mathcal{P} \in \mathsf{X}} \beta(\mathcal{P})$. To make $\beta(\mathsf{X})$ well-defined, we define $\beta(\mathsf{X}) = \mathbb{P}$ in case $\mathsf{X} = \emptyset$.

If a protein interacts with all proteins in X, it must be in the neighborhood set $\beta(\mathsf{X})$. However, if a protein interacts with all proteins in $\beta(\mathsf{X})$, it may not be in X. In this

case, the subset X can be expanded by adding the protein $\mathcal{P}$, while maintaining the same neighborhood. Where to stop the expansion? We use the following definition of interacting protein group pairs.

Definition 5.1. *Let* $\mathsf{X}, \mathsf{Y} \subseteq \mathbb{P}$ *are two subset of proteins of* $\mathbf{G}$. *If* $\beta(\mathsf{X}) = \mathsf{Y}$ *and* $\beta(\mathsf{Y}) = \mathsf{X}$, *then we call* X *and* Y *a* **pair of interacting protein groups**.

If two protein groups X and Y are a pair of interacting protein groups, then every protein in one set (X or Y) interacts with all proteins in the other set, and vice versa. Also, a pair of interacting protein groups X and Y of $\mathbf{G}$ such that $\beta(\mathsf{X}) = \mathsf{Y}$ and $\beta(\mathsf{Y}) = \mathsf{X}$ is maximal in the sense that there is no other pairs of interacting protein group X' and Y' of $\mathbf{G}$ with $\mathsf{X} \subseteq \mathsf{X}'$ and $\mathsf{Y} \subseteq \mathsf{Y}'$ such that $\beta(\mathsf{X}') = \mathsf{Y}'$ and $\beta(\mathsf{Y}') = \mathsf{X}'$. To appreciate this notion of maximality, we prove the proposition below.

Proposition 5.1. *Let* X *and* Y, X' *and* Y' *be two interacting protein group pairs of* $\mathbf{G}$ *such that* $\mathsf{X} \subseteq \mathsf{X}'$ *and* $\mathsf{Y} \subseteq \mathsf{Y}'$. *Then* $\mathsf{X} = \mathsf{X}'$ *and* $\mathsf{Y} = \mathsf{Y}'$.

Proof. Suppose X and Y, X' and Y' are two interacting protein group pairs of $\mathbf{G}$ such that $\mathsf{X} \subseteq \mathsf{X}'$ and $\mathsf{Y} \subseteq \mathsf{Y}'$. Since $\mathsf{X} \subseteq \mathsf{X}'$ and $\mathsf{Y} \subseteq \mathsf{Y}'$, we have $\beta(\mathsf{X}') \subseteq \beta(\mathsf{X})$ and $\beta(\mathsf{Y}') \subseteq \beta(\mathsf{Y})$. Using the definition of interacting protein group pairs, we derive $\mathsf{Y}' = \beta(\mathsf{X}') \subseteq \beta(\mathsf{X}) = \mathsf{Y}$ and $\mathsf{X}' = \beta(\mathsf{Y}') \subseteq \beta(\mathsf{Y}) = \mathsf{X}$. Then $\mathsf{X} = \mathsf{X}'$ and $\mathsf{Y} = \mathsf{Y}'$ as desired. □

Note that not every protein set is an interacting protein group because the partner interacting protein group may not exist. Also note that not all interacting protein group pairs are equally interesting. We would probably not be very interested in two groups with a small size containing a single protein or just a few. In contrast, we would probably be considerably more interested if one of the groups is large, or both of the groups are large, to derive significant patterns from the groups. Hence, we can introduce the notion of density on interacting protein group pairs.

Definition 5.2. *A pair of interacting protein groups* X *and* Y *in a protein interaction network* $\mathbf{G}$ *is said to be* (τ_1, τ_2)*-dense if* $|\mathsf{X}|$ *or* $|\mathsf{Y}|$ *is at least* τ_1, *and the other is at least* τ_2.

Problem statement: Let a protein interaction network $\mathbf{G} = \langle \mathbb{P}, \mathbb{D} \rangle$. The problem is to find all pairs of frequent interacting protein groups X and Y that are (τ_1, τ_2)-dense, where both τ_1 and τ_2 are positive integers.

5.3 Closed Patterns of Adjacency Matrices

The adjacency matrix of a protein interaction network is important in this study. Let $\mathbf{G}$ be a simple protein interaction network with $\mathbb{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m\}$. The **adjacency matrix A** of $\mathbf{G}$ is the $m \times m$ matrix defined by

$$\mathbf{A}[i,j] = \begin{cases} 1 & \text{if } (\mathcal{P}_i, \mathcal{P}_j) \in \mathbb{D} \\ 0 & \text{otherwise} \end{cases}$$

Since $\mathbf{G}$ is a simple network (undirected without self-loop), $\mathbf{A}$ is a symmetric matrix and every entry on the main diagonal is 0. Also, $\{\mathcal{P}_j \mid \mathbf{A}[k,j] = 1, 1 \leq j \leq m\} = \beta(\mathcal{P}_k) = \{\mathcal{P}_j \mid \mathbf{A}[j,k] = 1, 1 \leq j \leq m\}$.

The adjacency matrix of a protein interaction network can be interpreted into a **transactional database** (*DB*) (Agrawal and Srikant, 1994). To define a *DB*, we first define a **transaction**. Let $\mathbf{I}$ be a set of **items**. Then a transaction is defined as a subset of $\mathbf{I}$. For example, assume $\mathbf{I}$ to be all items in a supermarket, a transaction by a customer is the items that the customer bought. A *DB* is a non-empty multi-set of transactions. Each transaction $\mathbf{T}$ in a *DB* is assigned a unique identity $id(\mathbf{T})$. A **pattern** is defined as a set of items of $\mathbf{I}$. A pattern may be or may not be contained in a transaction. Given a *DB* and a pattern X, the number of transactions in *DB* containing X is called the **support** of X, denoted $\pi^{DB}(\mathsf{X})$. We are often interested in patterns that occur sufficiently frequent in a *DB*. Those patterns are called **frequent** patterns—that is, patterns X satisfying $\pi^{DB}(\mathsf{X}) \geq \tau$, for a threshold $\tau > 0$. So, by a pattern of a *DB*, we mean that it occurs in *DB* at least once.

Closed patterns are a type of interesting patterns in a *DB*. In the last few years, the problem of efficiently mining closed patterns from a large *DB* has attracted a lot of researchers in the data mining community (Bastide et al., 2000; Goethals and Zaki, 2003; Grahne and Zhu, 2003; Nicolas et al., 1999; Uno et al., 2004; Wang et al., 2003; Zaki and Hsiao, 2002). Let $\mathbf{I}$ be a set of items, and *DB* be a transactional database defined on $\mathbf{I}$. For a pattern $\mathsf{X} \subseteq \mathbf{I}$, let $f^{DB}(\mathsf{X}) = \{\mathbf{T} \in DB \mid \mathsf{X} \subseteq \mathbf{T}\}$—that is, $f^{DB}(\mathsf{X})$ are all transactions in *DB* containing the pattern X. For a set of transactions $DB' \subseteq DB$, let $g(DB') = \bigcap_{\mathbf{T} \in DB'} \mathbf{T} = \bigcap DB'$—that is, the set of items which are shared by all transactions in DB'. Using these two functions, we can define the notion of **closed patterns**. For a pattern X, $CL^{DB}(\mathsf{X}) = g(f^{DB}(\mathsf{X}))$ is called the **closure** of X. A pattern X is said to be **closed** with respect to a transactional database *DB* iff $CL^{DB}(\mathsf{X}) = \mathsf{X}$.

Let $\mathbf{G}$ be a protein interaction network with the protein set $\mathbb{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m\}$. If each protein in $\mathbb{P}$ is defined as an item, then the neighborhood $\beta(\mathcal{P}_i)$ of $\mathcal{P}_i$ is a transaction. Thus,

$$\{\beta(\mathcal{P}_1), \beta(\mathcal{P}_2), \ldots, \beta(\mathcal{P}_m)\}$$

is a *DB*. Such a special *DB* is denoted by $DB_{\mathbf{G}}$. The identity of a transaction in $DB_{\mathbf{G}}$ is defined as the protein itself—that is, $id(\beta(\mathcal{P}_i)) = \mathcal{P}_i$. Note that $DB_{\mathbf{G}}$ has the same number of items and transactions. Note also that $\mathcal{P}_i \notin \beta(\mathcal{P}_i)$ since we assume $\mathbf{G}$ to be a simple network (undirected without self-loop).

$DB_{\mathbf{G}}$ can be represented as a binary square matrix. This binary matrix $\mathbf{B}$ is defined by

$$\mathbf{B}[i,j] = \begin{cases} 1 & \text{if } \mathcal{P}_j \in \beta(\mathcal{P}_i) \\ 0 & \text{otherwise} \end{cases}$$

Since $\mathcal{P}_j \in \beta(\mathcal{P}_i)$ iff $(\mathcal{P}_i, \mathcal{P}_j) \in \mathbb{D}$, it can be seen that $\mathbf{A} = \mathbf{B}$. So, "a pattern of $DB_{\mathbf{G}}$" is equivalent to "a pattern of the adjacency matrix of $\mathbf{G}$".

We define the **occurrence set** of a pattern X in *DB* as $occ^{DB}(\mathsf{X}) = \{id(\mathbf{T}) \mid \mathbf{T} \in DB, \mathsf{X} \subseteq \mathbf{T}\} = \{id(\mathbf{T}) \mid \mathbf{T} \in f^{DB}(\mathsf{X})\}$. It is straightforward to see that $id(\mathbf{T}) \in occ^{DB}(\mathsf{X})$

iff $\mathbf{T} \in f^{DB}(\mathsf{X})$. There is a connection between the notion of occurrence sets and closed patterns in any transaction database *DB*.

Proposition 5.2. *Let* C_1 *and* C_2 *be two closed patterns of DB. Then* $\mathsf{C}_1 = \mathsf{C}_2$ *iff* $occ^{DB}(\mathsf{C}_1) = occ^{DB}(\mathsf{C}_2)$.

Proof. The left-to-right direction is trivial. To prove the right-to-left direction, let us suppose that $occ(\mathsf{C}_1) = occ(\mathsf{C}_2)$. It is straightforward to see that $id(\mathbf{T}) \in occ(\mathsf{X})$ iff $\mathbf{T} \in f(\mathsf{X})$. Then we get $f(\mathsf{C}_1) = f(\mathsf{C}_2)$ from $occ(\mathsf{C}_1) = occ(\mathsf{C}_2)$. Since C_1 and C_2 are closed patterns of *DB*, it follows that $\mathsf{C}_1 = g(f(\mathsf{C}_1)) = g(f(\mathsf{C}_2)) = \mathsf{C}_2$, and finishes the proof. □

We discuss in the next section the relationships between the closed patterns of $DB_{\mathbf{G}}$ and the interacting protein groups of $\mathbf{G}$.

5.4 Relationship between Interacting Protein Groups and Closed Patterns

5.4.1 Relationships among Neighborhood, Occurrence Sets and Closed Patterns

The occurrence set of a closed pattern C in $DB_{\mathbf{G}}$ plays a key role in the interacting protein group pairs of $\mathbf{G}$. We introduce below some of its key properties.

There is a tight connection between the notions of neighborhood in a protein interaction network $\mathbf{G}$ and occurrence in the corresponding transactional database $DB_{\mathbf{G}}$.

Proposition 5.3. *Given a protein interaction network* $\mathbf{G}$ *and a pattern* X *of* $DB_{\mathbf{G}}$. *Then* $\beta(\mathsf{X}) = occ^{DB_{\mathbf{G}}}(\mathsf{X})$.

Proof. If $\mathcal{P} \in \beta(\mathsf{X})$, then $\mathcal{P}$ is adjacent to every protein in X. So, $\beta(\mathcal{P}) \supseteq \mathsf{X}$. Therefore, $\beta(\mathcal{P})$ is a transaction of $DB_{\mathbf{G}}$ containing X. So, $\mathcal{P} \in occ(\mathsf{X})$.

If $\mathcal{P} \in occ(\mathsf{X})$, then $\mathcal{P}$ is adjacent to every protein in X. Therefore, $\mathcal{P} \in \beta(\mathcal{P}')$ for each $\mathcal{P}' \in \mathsf{X}$. That is, $\mathcal{P} \in \bigcap_{\mathcal{P}' \in \mathsf{X}} \beta(\mathcal{P}') = \beta(\mathsf{X})$.

□

There is also a nice connection between the notions of neighborhood in a protein interaction network and that of closure of patterns in the corresponding transactional database.

Proposition 5.4. *Given a protein interaction network* $\mathbf{G}$ *and a pattern* X *of* $DB_{\mathbf{G}}$*. Then* $\beta(\beta(\mathsf{X})) = CL^{DB_{\mathbf{G}}}(\mathsf{X})$*. Thus* $\beta \circ \beta$ *is a closure operation on patterns of* $DB_{\mathbf{G}}$*.*

Proof. By Proposition 5.3, $\beta(\beta(\mathsf{X})) = \beta(occ(\mathsf{X})) = \bigcap_{(\mathcal{P}=id(\mathbf{T})) \in occ(\mathsf{X})} (\beta(id(\mathbf{T})) = \mathbf{T}) = \bigcap_{\mathbf{T} \in f(\mathsf{X})} \mathbf{T} = g(f(\mathsf{X})) = CL(\mathsf{X})$. □

The occurrence sets in $DB_{\mathbf{G}}$ have a specific property, that is, they are all closed patterns in $DB_{\mathbf{G}}$, with respect to the protein interaction network $\mathbf{G}$.

Lemma 5.1. *Let* $\mathbf{G}$ *be an undirected protein interaction network without self-loop. Let* C *be a closed pattern of* $DB_{\mathbf{G}}$*. Then* $f^{DB_{\mathbf{G}}}(occ^{DB_{\mathbf{G}}}(\mathsf{C})) = \{\beta(\mathcal{P}) \mid \mathcal{P} \in \mathsf{C}\}$.

Proof. As C is a closed pattern, by definition, then $\{\mathcal{P} \mid \mathcal{P} \in \mathsf{C}\}$ are all and only items contained in every transaction of $DB_{\mathbf{G}}$ that contains C. This is equivalent to that $\{\mathcal{P} \mid \mathcal{P} \in \mathsf{C}\}$ are all and only proteins of $\mathbf{G}$ that are adjacent to every protein in $occ(\mathsf{C})$. This implies that $\{\beta(\mathcal{P}) \mid \mathcal{P} \in \mathsf{C}\}$ are all and only transactions that contain $occ(\mathsf{C})$. In other words, $f(occ(\mathsf{C})) = \{\beta(\mathcal{P}) \mid \mathcal{P} \in \mathsf{C}\}$. □

Proposition 5.5. *Let* $\mathbf{G}$ *be a protein interaction network and* C *a pattern of* $DB_{\mathbf{G}}$*. Then* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *is a closed pattern of* $DB_{\mathbf{G}}$*.*

Proof. By Lemma 5.1, $f(occ(\mathsf{C})) = \{\beta(\mathcal{P}) \mid \mathcal{P} \in \mathsf{C}\}$. So $CL(occ(\mathsf{C})) = g(f(occ(\mathsf{C}))) = \bigcap f(occ(\mathsf{C})) = \bigcap_{\mathcal{P}\in\mathsf{C}} \beta(\mathcal{P}) = \beta(\mathsf{C})$. By Proposition 5.3, $\beta(\mathsf{C}) = occ(\mathsf{C})$. Thus $occ(\mathsf{C})$ is a closed pattern. □

From Proposition 5.3 and 5.5, for every pattern C of $DB_{\mathbf{G}}$, $\beta(\mathsf{C})$ is a close pattern of $DB_{\mathbf{G}}$.

5.4.2 Number of Closed Patterns in Adjacency Matrices

Proposition 5.6. *Let* $\mathbf{G}$ *be a protein interaction network and* C *a pattern of* $DB_{\mathbf{G}}$*. Then* C *and its occurrence set has empty intersection. That is,* $occ^{DB_{\mathbf{G}}}(\mathsf{C}) \cap \mathsf{C} = \emptyset$.

Proof. Let $\mathcal{P} \in occ(\mathsf{C})$. Then $\mathcal{P}$ is adjacent to every protein in C. Since we assume $\mathbf{G}$ is a protein interaction network without self-loop, $\mathcal{P} \notin \mathsf{C}$. Therefore, $occ^{DB_{\mathbf{G}}}(\mathsf{C}) \cap \mathsf{C} = \emptyset$. □

The above propositions above give rise to a couple of interesting corollaries below.

Corollary 5.1. *Let* $\mathbf{G}$ *be a protein interaction network. Then the number of closed patterns that appear at least once in* $DB_{\mathbf{G}}$ *is even.*

Proof. Suppose there are k closed patterns in $DB_{\mathbf{G}}$, denoted as C_1, C_2, ..., C_k. As per Proposition 5.5, $occ(\mathsf{C}_1)$, $occ(\mathsf{C}_2)$, ..., $occ(\mathsf{C}_k)$ are all closed patterns of $DB_{\mathbf{G}}$. As per Proposition 5.2, $occ(\mathsf{C}_i)$ is different from $occ(\mathsf{C}_j)$ iff C_i is different from C_j. So every closed pattern can be paired with a distinct closed pattern by $occ(\cdot)$ in a bijective manner. Furthermore, as per Proposition 5.6, no closed pattern is paired with itself. This is possible only when the number k is even. □

Corollary 5.2. *Let* $\mathbf{G}$ *be a protein interaction network. Then the number of closed patterns* C*, such that both* C *and* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *appear at least* τ *times in* $DB_{\mathbf{G}}$*, is even.*

Proof. As seen from the proof of Corollary 5.1, every closed pattern C of $DB_{\mathbf{G}}$ can be paired with $occ^{DB_{\mathbf{G}}}(\mathsf{C})$, and the entire set of closed patterns can be partitioned into such pairs. So a pair of closed patterns C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ either satisfy or do not satisfy the condition that both C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ appear at least τ times in $DB_{\mathbf{G}}$. Therefore, the number of closed patterns C, satisfying that both C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ appear at least τ times in $DB_{\mathbf{G}}$, is even. □

Note that this corollary does not imply the number of frequent closed patterns that appear at least τ times in $DB_{\mathbf{G}}$ is always even. A counter example is given below.

Example 5.1. *Consider a $DB_{\mathbf{G}}$ given by the following matrix:*

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$
$\beta(\mathcal{P}_1)$	*0*	*1*	*1*	*0*	*0*
$\beta(\mathcal{P}_2)$	*1*	*0*	*1*	*1*	*1*
$\beta(\mathcal{P}_3)$	*1*	*1*	*0*	*1*	*1*
$\beta(\mathcal{P}_4)$	*0*	*1*	*1*	*0*	*0*
$\beta(\mathcal{P}_5)$	*0*	*1*	*1*	*0*	*0*

We list its closed patterns, their support, and their $occ(\cdot)$ counterpart patterns below:

support of X	*close pattern* X	$\mathsf{Y} = occ(\mathsf{X})$	*support of* Y
3	$\{\mathcal{P}_2, \mathcal{P}_3\}$	$\{\mathcal{P}_1, \mathcal{P}_4, \mathcal{P}_5\}$	*2*
4	$\{\mathcal{P}_2\}$	$\{\mathcal{P}_1, \mathcal{P}_3, \mathcal{P}_4, \mathcal{P}_5\}$	*1*
4	$\{\mathcal{P}_3\}$	$\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_4, \mathcal{P}_5\}$	*1*

Suppose we take $\tau = 3$. Then there are only 3 closed patterns—an odd number—that occur at least τ times, viz. $\{\mathcal{P}_2, \mathcal{P}_3\}$, $\{\mathcal{P}_2\}$, and $\{\mathcal{P}_3\}$.

5.4.3 One-to-one Correspondence between Interacting Protein Groups and Closed Patterns

Finally, we demonstrate our main result on the relationship with interacting protein group pairs and closed patterns. In particular, we discover that every pair of a closed pattern C and its occurrence set $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ yields a distinct pair of interacting protein groups of $\mathbf{G}$.

Theorem 5.1. *Let* $\mathbf{G}$ *be a simple protein interaction network (undirected without self-loop). Let* C *be a closed pattern of* $DB_{\mathbf{G}}$*. Then* C *and* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *is a pair of interacting protein groups of* $\mathbf{G}$*.*

Proof. By Proposition 5.6, $\mathsf{C} \cap occ^{DB_{\mathbf{G}}}(\mathsf{C}) = \emptyset$. By Proposition 5.3, we have $occ^{DB_{\mathbf{G}}}(\mathsf{C}) = \beta(\mathsf{C})$. By Proposition 5.4, $\mathsf{C} = \beta(\beta(\mathsf{C}))$. By Proposition 5.3, we derive $\mathsf{C} = \beta(occ^{DB_{\mathbf{G}}}(\mathsf{C}))$. Therefore, C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ is a pair of interacting protein groups. □

Theorem 5.2. *Let* $\mathbf{G}$ *be a simple protein interaction network (undirected without self-loop). Let* X *and* Y *be a pair of interacting protein groups of* $\mathbf{G}$*. Then,* X *and* Y *are both a closed pattern of* $DB_{\mathbf{G}}$*,* $occ^{DB_{\mathbf{G}}}(\mathsf{X}) = \mathsf{Y}$ *and* $occ^{DB_{\mathbf{G}}}(\mathsf{Y}) = \mathsf{X}$*.*

Proof. Since X and Y is a pair of interacting protein groups of $\mathbf{G}$, then $\beta(\mathsf{X}) = \mathsf{Y}$ and $\beta(\mathsf{Y}) = \mathsf{X}$. By Proposition 5.4, $CL(\mathsf{X}) = \beta(\beta(\mathsf{X})) = \beta(\mathsf{Y}) = \mathsf{X}$. So, X is a closed pattern. Similarly, we can get Y is a closed pattern. By Proposition 5.3, $occ(\mathsf{X}) = \beta(\mathsf{X}) = \mathsf{Y}$ and $occ(\mathsf{Y}) = \beta(\mathsf{Y}) = \mathsf{X}$, as required. □

The above two theorems indicate that interacting protein group pairs of $\mathbf{G}$ are all in the form of X and Y, where X and Y are both a closed pattern of $DB_{\mathbf{G}}$. Also, for every closed pattern C of $DB_{\mathbf{G}}$, C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ is a pair of interacting protein groups of $\mathbf{G}$. So, there is a one-to-one correspondence between interacting protein group pairs and closed pattern pairs.

We can also derive a corollary linking support threshold of $DB_{\mathbf{G}}$ to the density of interacting protein group pairs of $\mathbf{G}$.

Corollary 5.3. *Let* $\mathbf{G}$ *be a simple protein interaction network (undirected without self-loop). Then* C *and* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *is a* (τ_1, τ_2)*-dense interacting protein group pairs of* $\mathbf{G}$ *iff* C *is a closed pattern such that* C *or* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *occurs at least* τ_1 *times in* $DB_{\mathbf{G}}$ *and the other occur at least* τ_2 *times in* $DB_{\mathbf{G}}$.

The corollary above has the following important implication.

Theorem 5.3. *Let* $\mathbf{G}$ *be a simple protein interaction network (undirected without self-loop). Then* C *and* $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ *is a* (τ_1, τ_2)*-dense interacting protein group pair of* $\mathbf{G}$ *iff* C *is a closed pattern such that* C $(occ^{DB_{\mathbf{G}}}(\mathsf{C}))$ *occurs at least* τ_1 *times in* $DB_{\mathbf{G}}$ *and* $|\mathsf{C}| \geq \tau_2$ $(|occ^{DB_{\mathbf{G}}}(\mathsf{C})| \geq \tau_2)$.

Proof. Suppose C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ is a (τ_1, τ_2)-dense pair of interacting protein groups of $\mathbf{G}$. By Theorem 5.2, $\mathsf{C} = occ(occ(\mathsf{C}))$. By definition of $occ(\cdot)$, $\pi(occ(\mathsf{C})) = |occ(occ(\mathsf{C}))| = |\mathsf{C}|$. Substitute this into Corollary 5.3, we get C and $occ^{DB_{\mathbf{G}}}(\mathsf{C})$ is a (τ_1, τ_2)-dense pair of interacting protein groups of $\mathbf{G}$ iff C is a closed pattern such that C $(occ^{DB_{\mathbf{G}}}(\mathsf{C}))$ occurs at least τ_1 times in $DB_{\mathbf{G}}$ and $|\mathsf{C}| \geq \tau_2$ $(|occ^{DB_{\mathbf{G}}}(\mathsf{C})| \geq \tau_2)$ as desired. □

Theorems 5.1 and 5.2 show that algorithms for mining closed patterns can be used to extract interacting protein group pairs of undirected protein interaction networks without self-loop. Such data mining algorithms are usually significantly more efficient at higher support threshold. Thus Theorem 5.3 suggests an important optimization for mining (τ_1, τ_2)-dense interacting protein group pairs. To wit, assuming $\tau_1 > \tau_2$, it suffices to mine closed patterns at support threshold τ_1, and then get the answer by filtering out those patterns of length less than τ_2.

5.5 Discussions

As proved in the last section, the mining of interacting protein groups can be transformed into the mining of close patterns. We use efficient closed pattern mining algorithms such

as FPCLOSE (Grahne and Zhu, 2003) or LCM (Uno et al., 2004) to achieve this goal. First, we work out the set of closed patterns with respect to the large threshold. Then, we compute the neighborhoods (occurrence sets) for those closed patterns with lengths no smaller than the small threshold. To avoid duplications, we record the processed closed patterns and their neighborhoods in a set and check the set whenever a new closed pattern is processed.

Even with the above optimization, an interacting protein group pair may be traveled twice. It is inefficient as one closed pattern implies the other. Moreover, many small closed patterns need to be gone over before the larger ones are obtained, which is a waste since the small ones will be filtered out finally. It is an interesting issue to prune the small closed patterns earlier using the constraints on both closed patterns and their occurrence sets. The close patterns which are impossible to extend some frequent interacting protein group pairs can be pruned as early as possible. These issues are being studied by our co-workers.

Recall that mining interacting protein group pairs is equivalent to listing all maximal complete bipartite subgraphs (or maximal bipartite cliques in some literatures). There are some recent works in the area. First, Makino and Uno (2004) investigated the problem of enumerating all maximal bipartite cliques from a *bipartite graph*. Since their work is limited to enumerating from only bipartite graphs while our work can work on any simple protein interaction networks, ours is more general. Second, Zaki and Ogihara (1998) observed that a transactional database DB can be represented by a bipartite graph $\mathbf{G}$, and also a relation that closed patterns (wrongly stated as maximal patterns in Zaki and Ogihara (1998)) of DB one-to-one correspond to maximal bipartite clique of $\mathbf{G}$. However, our work is to convert a graph (a protein interaction network) $\mathbf{G}$, including a bipartite graph, into a special transactional database $DB_{\mathbf{G}}$, and then to discover all closed patterns from $DB_{\mathbf{G}}$ for enumerating all maximal bipartite subgraphs (interacting protein group pairs) of $\mathbf{G}$. Furthermore, the occurrence set of a closed pattern in Zaki's work may not be a closed pattern, but that of ours is always a closed pattern.

The mining of maximal complete bipartite subgraphs is a model problem in graph theory, so it has many applications in mathematics, electrical engineering, computer programming, business administration, sociology, economics, marketing, biology, and networking and communications. The theoretical proof between maximal complete bipartite subgraphs and closed patterns in this chapter can also benefit those applications. For example, suppose there is a set of customers in a mobile communication network. Some people have a wide range of contact, while others have few. Which groups of customers (with a maximal number) have a full interaction with another group of customers? This situation can be modeled by a graph where a mobile phone customer is a vertex and a communication is an edge. Thus, a maximal bipartite subgraph of this graph corresponds to two groups of customers between whom there exist a full communication. Another similar example is studied in web mining (Andrei et al., 2000; Kumar et al., 1999; Murata, 2004) where web communities are modeled by bipartite cores. Using the results in this chapter, these problems can all be solved efficiently by the mining of closed patterns.

5.6 Summary

Finally, we summarize the results achieved in this chapter. We have studied the problem of listing all interacting protein group pairs from a protein interaction network. We proved that this problem is equivalent to the mining of all closed patterns from the adjacency matrix of the protein interaction network. More specifically, we prove, for a simple protein interaction network $\mathbf{G}$ (undirected without self-loop): (i) that the number of closed patterns in the adjacency matrix of $\mathbf{G}$ is even; (ii) that the number of the closed patterns is precisely double the number of interacting protein group pairs of $\mathbf{G}$; (iii) that for every pair of interacting protein groups, there always exists a unique pair of closed patterns that matches the two protein sets. Therefore, we can enumerate all interacting protein group pairs using efficient algorithms for mining closed patterns, which have been extensively studied in the data mining field. As the major focus, we demonstrate the usage of interacting protein groups to discover binding motif pairs in the next chapter.

Chapter 6

Binding Motif Pairs from Interacting Protein Group Pairs

6.1 Introduction

In the previous chapter, we observe that there exist many interacting protein group pairs in a protein interaction network that exhibit an all-versus-all interaction between two groups in the pairs, we called them *interacting protein group pairs*. We demonstrated that the mining of interacting protein group pairs from a protein interaction network can be transformed into the mining of closed patterns in the adjacency matrix of the protein interaction network. In this chapter, we show why and how those interacting protein groups are applied to discover binding motif pairs at interaction sites on a proteome-wide scale.

Figure 6.1 shows a typical example of such an interacting protein group pair from a yeast interaction network (Schwikowski et al., 2000; Tong et al., 2002). The two protein groups correspond to SH3 proteins and SH3-binding proteins. It reveals a binding pair between SH3 domain and SH3-binding motifs. Generally, if a large enough subnetwork

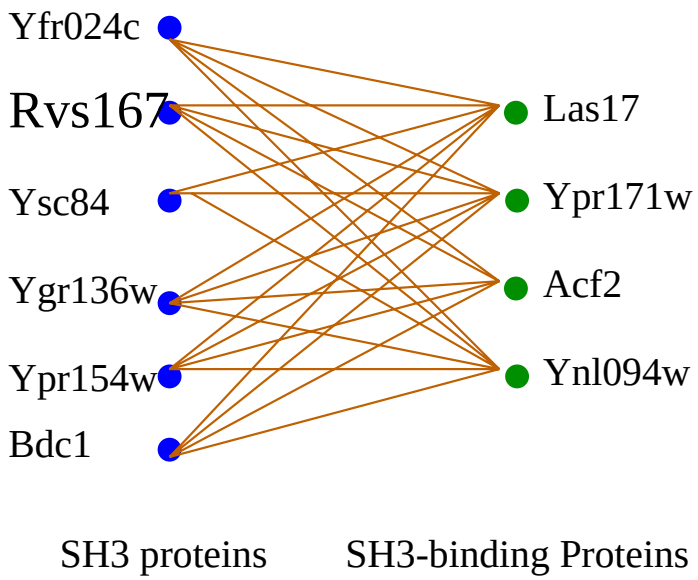


Figure 6.1: An all-versus-all predicted interaction subnetwork (most are confirmed by experiments) consisting of two groups of proteins, where one group contains six proteins with SH3 domains and the other contains four proteins with SH3-binding motifs. The data is from (Tong et al., 2002).

with most-versus-most or even all-versus-all interactions between two protein groups is found in a protein network, a binding motif pair at interaction sites of the interacting protein group pair can be discovered. That is because most proteins only contain a small number of interaction sites (usually, $2 \sim 6$ for typical proteins (Liang et al., 1998)). Due to the constraints of all-versus-all interactions between these two groups, it is expected that there exists two groups of interaction sites from these two protein groups which interact with each other for at least some occurrences. The interaction sites within the same group should hold similar structures and possibly have a sequence motif as they have similar interaction partners. These two groups of interaction sites and their corresponding motifs can be easily identified using standard motif discovery methods from the sequence data of the corresponding protein group. Then, a binding motif pair is formed, with which to represent the corresponding interaction sites of the protein group pair.

More specifically, the binding motif pair from an interacting protein group pair is the conserved pattern at the interaction sites of an interaction type. And most likely,

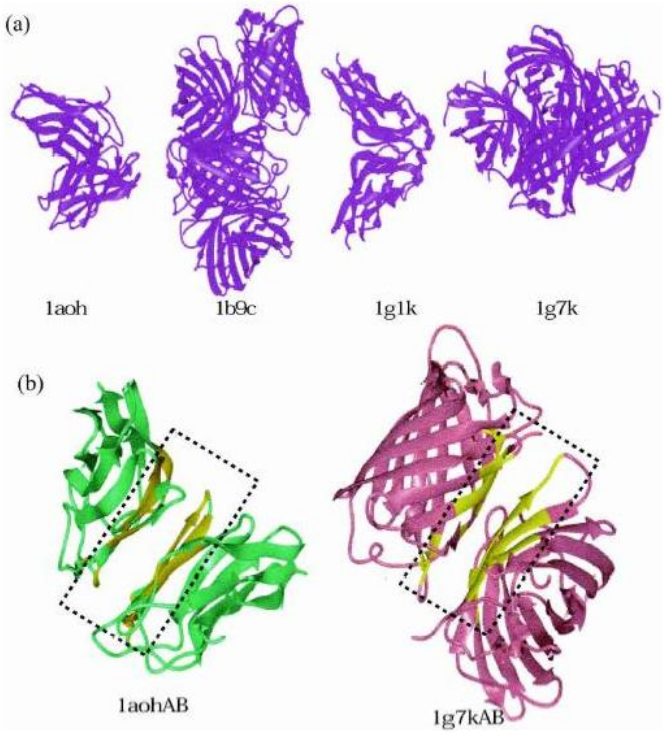


Figure 6.2: The example of an interaction type, figure from (Keskin et al., 2004).

an interacting protein group pair belongs to such an interaction type. As mentioned in Chapter 2, an interaction type is a set of conserved interaction sites sharing the common binding mechanism (Aloy and Russell, 2004). Those conserved interaction sites are favorable interfacial scaffolds that have been repeatedly used in the evolution process by proteins with different sequence, structure and function (Keskin et al., 2005; Keskin and Nussinov, 2005). An example can be seen from *cipa* (PDB code 1aoh) and *Dsred* (PDB code 1g7k), two complexes which have similar interfaces between their component chains A and B (Keskin et al., 2004), but which have dissimilar global structures and functions (See Figure 6.2). Whenever the conserved pattern occurs in a novel protein pair regardless of their homology, the two proteins are likely to interact—this principle has been used by Tong et al. (2002) and Aytuna et al. (2005) to predict protein interactions with an acceptable performance.

To assess the performance of the method using closed pattern mining and motif discovery to discover binding motif pairs, we propose a systematic validation experiment on

comprehensive domain databases and domain–domain interaction databases. We compare our single motifs with the domains in specific domain databases to study the relationship between our motifs and domains. Even more importantly, we study the relationship between binding motif pairs and interacting domain pairs, by mapping our binding motif pairs into domain–domain interacting pairs and analyzing the amount of overlaps between our mapped domain pairs and those in domain–domain interaction databases.

The organization of the chapter is as follows. In Section 6.2, we describe the detailed algorithm in addition to implementation issues. In Section 6.3, we depict the overall results. In Section 6.4, we examine our results carefully by systematic validation experiments. We report a case study in Section 6.5 and conclude this chapter in the final section.

6.2 Generating Binding Motif Pairs from Interacting Protein Group Pairs

6.2.1 Algorithm Issues

Our algorithm consists of two steps: The first step is to find all interacting protein group pairs from a protein interaction network where this problem is transformed into the mining of *closed patterns* using efficient algorithms such as FPCLOSE (Grahne and Zhu, 2003) (see the previous chapter for details); The second step is to identify binding motif pairs from those interacting protein group pairs.

Given a protein group and its associated sequences, we can get a motif (possibly with flexible gaps) using standard motif discovery algorithms such as PROTOMAT (Henikoff and Heinikoff, 1991) and MEME (Bailey and Elkan, 1995). So, we can easily obtain a motif pair from a pair of interacting protein groups by executing the motif discovery algorithm twice. In this chapter, we choose PROTOMAT (Henikoff and Heinikoff, 1991)

as the motif discovery algorithm because it is believed to be a good method to find local conserved regions from a group of related proteins. PROTOMAT is also a key method to construct BLOCKS database (Pietrokovski et al., 1996)—a comprehensive database of highly conserved regions for homologous protein groups (domains).

The PROTOMAT method consists of two steps. In the first step, a modified version of MOTIF program (Smith et al., 1990) looks for the presence of all spaced triplets (also called motifs) in at least a subset of sequences from the given group of proteins. All parameters required by the program can be determined automatically. In the second step, a graph theory-based method called MOTOMAT (Henikoff and Heinikoff, 1991) is used to determine the best set of blocks from this group of proteins. MOTOMAT works in this way: (i) it merges overlapping candidate blocks which are alignments for the motifs discovered by the MOTIF program; (ii) it refines the blocks by extending them in both directions until similarity falls off, generating blocks with maximum scores; and (iii) it determines the best set of blocks which are in the same order and do not overlap for a critical number of sequences using well-known techniques in graph theory (Henikoff et al., 1995).

6.2.2 Implementations

To assess the performance of our proposed method for mining binding motif pairs, we performed several experiments on a PC with a CPU clock rate of 3.2GHz and 2GB of main memory. As there are many physical protein interaction networks corresponding to different species, here we take the simplest and most comprehensive yeast physical interaction network as an example. The protein interaction network used in the experiments was downloaded from DIP (database of interacting proteins) on Oct. 23, 2005, consisting of 17511 experimentally determined interactions in *Saccharomyces cerevisiae* (yeast) among 4959 proteins. We select 10640 physical interactions by excluding 6871

interactions determined only by complex level experiments[1]. The adjacency matrix of the protein interaction network is a transactional database with 4447 items and transactions, with average transaction (neighborhood) size 4.79. To discover frequent closed patterns from this database, we use FPClose* (Grahne and Zhu, 2003), a state-of-the-art algorithm for mining closed pattern, for enumerating the interacting protein group pairs. Default parameters are used for PROTOMAT (Henikoff and Heinikoff, 1991). To facilitate our validations, we further term the motifs induced from closed patterns as left motifs (left blocks), while the ones induced from the occurrence sets of the closed patterns as right motifs (right blocks).

6.3 Results Overview

The results of protein groups (closed patterns) with respect to different support thresholds are reported in Table 6.1 (5 is chosen as the final threshold), where the second column shows the total number of **frequent** closed patterns whose support level is at least the threshold number in the column one. The third column of this table shows the number of closed patterns whose cardinality and support are both at least the support threshold; all such closed patterns are termed qualified closed patterns. Only these qualified closed patterns can be used to form interacting protein group pairs such that both groups meet the thresholds. From the table, we can see:

- The number of all closed patterns (corresponding to those with the support threshold of 1) is even. Moreover, the number of qualified closed patterns with cardinality no less than any support level is also even, as expected from Corollary 5.2.
- The algorithm runs fast—the algorithm program can complete within 10 seconds for all situations reported here. This indicates that enumerating all interacting protein

[1]immunoprecipitation, co-purification, tandem affinity purification (TAP), interaction adhesion assay, genetic, electron microscopy, immunostaining, immunofluorescence, transient co-expression and mass spectrometry

group pairs from a large protein interaction network can be practically solved by using algorithms for mining closed patterns.

- A so-called "many-few" property (Maslov and Sneppen, 2002) of protein interactions is observed again in our experiment results. The "many-few" property says that: a protein that interacts with a large number of proteins tends *not* to interact with another protein which also interacts with a large number of proteins (Maslov and Sneppen, 2002). In other words, highly connected proteins are separated by low-connected proteins. This is most clearly seen in Table 6.1 at the higher support thresholds. For example, at the support threshold 12, there are 4981 protein groups that have full interactions with at least 12 proteins. But there are only 8 groups, as seen in the third column of the table, that each contain at least 12 proteins and that have full mutual interaction.

We choose $\tau = 5$ (both τ_1 and τ_2), the average number of interactions per protein in the yeast genome (Grigoriev, 2003), as the threshold for both protein groups in the pairs. Under the threshold, the FPClose* algorithm outputs a total of 5349 non-redundant pairs of interacting protein groups, by taking 4.35 seconds on our machine (including the transformation). The mining based on the transformation idea is very efficient compared with a naive search method which needs $\sim$ 33 minutes (455-fold more than the efficient approach) to find all the protein groups. The implementation of the naive search contains some optimization techniques.

The homology property within a group is an interesting issue. It can be estimated simply by the sequence identity within the group—A value $< 15\%$ is often considered as a good indicator for non-homology (Doolittle, 1981). We calculate all pairwise sequence identities within a same protein group using CLUSTAL W package with default parameters (Thompson et al., 1994). Then we use the average value of these pairwise sequence identities as the sequence identity within the group. The distribution of the sequence identities within the 10698 groups is shown in Figure 6.3. The expected value of the sequence identities within the groups is 7.48%, with a standard deviation 1.33%. This is

Table 6.1: Closed patterns in a yeast protein physical interaction network

support threshold	# of frequent closed patterns	# of qualified closed patterns	time in sec.
1	31642	31642	9.18
2	28771	26004	7.9
3	24316	20220	6.71
4	20747	15712	6.59
5	17377	10698	4.35
6	14390	6892	3.15
7	12001	4362	2.27
8	10066	2688	1.66
9	8465	1412	1.19
10	7155	538	0.87
11	6008	106	0.68
12	4981	8	0.62
13	4039	0	0.6

a good value indicating the non-homology within these groups. Therefore, these groups and their underlying sequence motifs are unlikely to detect by standard methods based on sequence homology (Sauder et al., 2000).

The PROTOMAT method outputs 5343 binding motif pairs from these 5349 pairs of interacting protein groups by taking 3 hours. Of the protein groups 85% generate two or three blocks. (Note that a group in BLOCKS contains 6.91 blocks on average.) Only four left groups and two right groups failed to produce any valid motif, with a failure rate $< 0.2\%$. Totally, there are 11948 left blocks and 13004 right blocks. The average length of these blocks is 11.05, with a standard deviation 5.06. Compared with BLOCKS where the average length of blocks is 25.337 and the standard deviation is 12.897, our blocks are more specific and match better with current knowledge about interaction sites, i.e., 10-20

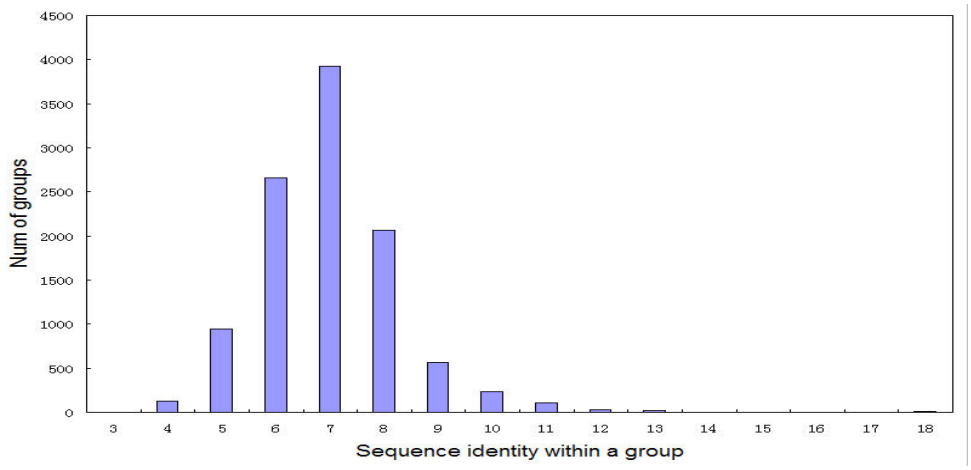


Figure 6.3: The distribution of the sequence identities within our 10698 groups.

residues in length (Sheu et al., 2005).

We treat the whole set of blocks generated by PROTOMAT from a protein group rather than each individual block as a motif to reflect the cooperation among these blocks. We expect that some interactions happen among the blocks from different sides of the motif pair, but do not study the detailed interactions among these blocks in this chapter. In our results, the average number of blocks per motif is 2.33, with a standard deviation 0.73, with details in Figure 6.4. The average number of proteins per motif is 7.01, with a standard deviation 2.59, see Figure 6.5 for details.

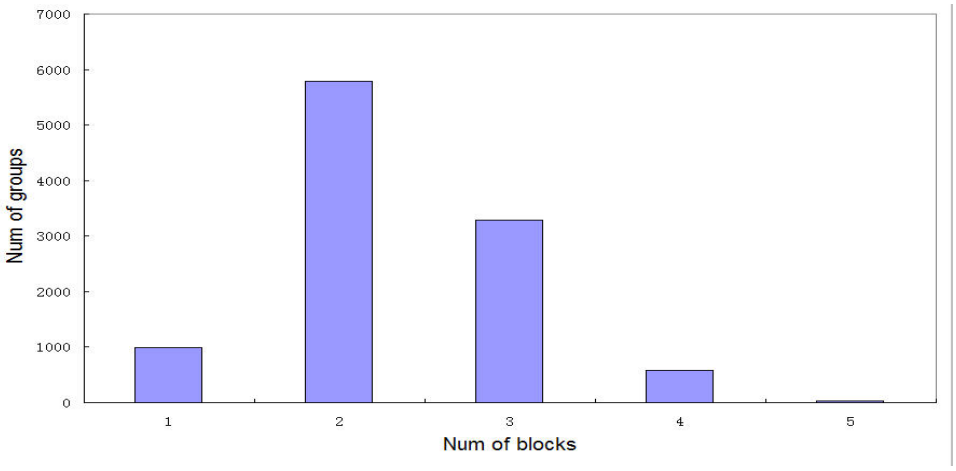


Figure 6.4: The distribution of the block numbers within our 10698 groups.

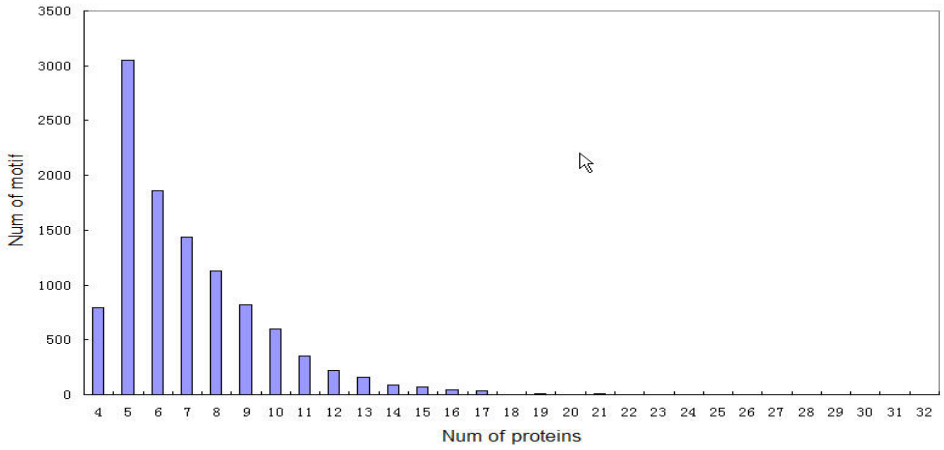


Figure 6.5: The distribution of the protein numbers within our 10698 motifs.

6.4 Validations

Currently, comprehensive databases for motif–motif interactions (binding motif pairs) are hard to find but there are a handful of databases for domain–domain interactions such as iPfam (Finn et al., 2005), 3did (Stein et al., 2005) and InterDom (Ng et al., 2003). Since domains are known to involve in protein interactions and are closely related to motifs, we compare our binding motif pairs with these domain pairs. The following two steps are employed to illustrate the effectiveness of our algorithm.

- Compare all single motifs in our discovered binding motif pairs with all domains in specific domain databases to obtain overall matches, i.e. to determine the number of motifs that can be mapped to these domains and the overall correlation in the portions that are mapped.

- Map our binding motif pairs into domain–domain interacting pairs to determine the number of overlaps between our mapped domain pairs and those in the domain–domain interaction database.

Table 6.2: Databases used in our validation experiments

	BLOCKS	PRINTS	Pfam	iPfam
Version	14.0	37.0	16.0	18.0
Num. of domains	4944	1850	7677	2145
Num. of entries	24294	11170	7677	3045

6.4.1 Validations of Single Motifs

As our motifs are in the form of blocks, we need domain databases also in the form of blocks for comparison. Currently, there are two major domain databases in the form of blocks: BLOCKS (Pietrokovski et al., 1996) and PRINTS (Attwood and Beck, 1994). Some information of these two domain databases are shown in the first two columns of Table 6.2, where an entry corresponds to a block.

The comparison is conducted by a program called *Local Alignment of Multiple Alignments* (LAMA) (Pietrokovski, 1996) which is an effective tool to determine local similarities between pairs of blocks. In the process, the method first transforms blocks into position-specific scoring matrices (PSSMs) (Gribskov et al., 1987), which specify the possibilities for all residues in each column of the blocks. Then it utilizes Smith-Waterman algorithm (Smith and Waterman, 1981) to determine the optimal local alignments for pairs of PSSMs of the corresponding blocks, utilizing Pearson's correlation coefficient (Pearson and Lee, 1903) as a metric to measure the similarity between two columns. To estimate the alignment scores with different lengths and to filter out the coincidental matches, LAMA uses the *Z-score* as a significance measurement, where a Z-score between a pair of PSSMs is defined as the number of standard deviations away from the mean score generated by millions of shuffled blocks in the BLOCKS database.

In our study, we used the default threshold 5.6 for Z-score in LAMA to compare our blocks with those in BLOCKS and PRINTS. If 95% of the positions of a block are in the optimal alignment between this block and another block and the Z-score is no less than

the threshold, we say there is a *mapping* from the former block to the latter one. If there is a mapping from some blocks of a motif to some blocks of a domain, we say the motif can be mapped to the domain. We have following results from this experiment:

- On average, each of our blocks maps to ~ 3.08 blocks in the BLOCKS or PRINTS databases. See more detailed report in the columns 2 and 3 of Table 6.3.
- The average correlation between the columns of our blocks and the columns from the database in the optimal alignments is as high as 53.88%. See column 4 of Table 6.3 for details.
- Our motifs can be mapped to 4221 domains out of a total of 6794 domains in these two databases, having a coverage of 62%. See Table 6.4. This result is interesting as our blocks can only be mapped to 8582 blocks out of the total 35464 blocks in these two databases, having a coverage $< 24\%$. The interpretation from a biological perspective is that most domains have about 40% of blocks as their interaction sites, while others may be related to folding.
- Although only 59% (14620 out of 24952) of our blocks can be mapped to blocks in BLOCKS and PRINTS, as high as 86% (9153 out of 10686) of motifs can be mapped to domains in these two databases. See Table 6.5 for details.

Note that our groups and groups in BLOCKS and PRINTS are constructed in quite different ways and their homology properties are also different. However, our comparison results reveal high correlation between their resulted blocks. This correlation may originate from the common involvement of interactions for both our motifs and their domains. This confirms the effectiveness of our method in some way.

6.4.2 Validations of Binding Motif Pairs

To assess whether our discovered binding motif pairs are indeed interaction sites, we compare them with domain–domain interacting pairs. If our binding motif pairs represent

Table 6.3: Statistics of mappings from our blocks to blocks in the BLOCKS and PRINTS databases

	# of our blocks	# of mappings to BLOCKS blocks	# of mappings to PRINTS blocks	Average correlation
Left blocks	11948	29357	8632	54.31
Right blocks	13004	30220	8738	53.42

Table 6.4: Statistics of blocks or domains in the BLOCKS or PRINTS databases that can be mapped from our blocks or motifs

	Mapped / total # in BLOCKS	Mapped / total # in PRINTS	Mapped / total # in ANY
Blocks	6408 / 24294	2174/ 11170	8582 / 35464
Domains	3128 / 4944	1093/ 1850	4221 / 6794

Table 6.5: Statistics of blocks or motifs in our binding motif pairs that can be mapped to blocks or domains in BLOCKS or PRINTS databases

	total #	# mapped to BLOCKS	# mapped to PRINTS	# mapped to ANY
Blocks	24952	13859	8010	14620
Motifs	10686	8879	6464	9153

interaction sites, they should be mapped to some domain–domain interacting pairs in some databases. We choose iPfam (Finn et al., 2005) for this purpose. It consists of 3045 interacting pairs among 2145 Pfam domains derived from protein complexes in PDB.

The cross-links between our binding motif pairs and the domain–domain pairs in iPfam is complicated. A reason is that the domain–domain pairs are represented by Pfam entries. To find the cross-links, (1) we first map our motifs to domains (protein groups) in the BLOCKS or PRINTS database, as shown in Section 6.4.1; (2) we then map a protein group of BLOCKS to a protein group of InterPro (Apweiler et al., 2001) as there exists a one-to-one mapping between an entry of BLOCKS and an entry of InterPro; (3) then we use existing cross-links between protein groups of InterPro and domains of Pfam to determine the cross-link between our motifs and Pfam domains. By this roadmap, we can map our binding motif pairs into domain–domain pairs with Pfam domain entries. Note that the association between PRINTS and Pfam is clear. Also note that the cross-linking mapping between binding motif pairs and domain–domain pairs is not a one-to-one mapping.

Using the above cross-link mapping, we compared our 5343 binding motif pairs with the 3045 domain-domain pairs in the iPfam database, 47 binding motif pairs can be mapped to 18 distinct domain pairs among 22 domains occurring in PDB complexes for 172 times (totally 105 distinct protein complexes).

Though the overlapping proportion seems modest, we assert that the result is significant because of the following:

- We read only interacting protein sequence pairs, while some predictions about interaction sites can be confirmed by domain–domain interactions in PDB complexes.
- iPfam is a rather incomplete database, containing merely 3045 pairs among 2145 domains. Moreover, only 997 out of 4221 of our mapped domains are studied in iPfam, as shown in Table 6.6.

Table 6.6: Occurrences of our mapped domains in different databases

	BLOCKS	PRINTS	Combined
BLOCKS/PRINTS domains	3128	1093	4221
Pfam domains	2305	144	2338
iPfam domains	975	87	997

- The binding motif pairs we discovered are taken only from the yeast genome while iPfam covers a variety of species.
- Comparing with Interdom with 30037 putative interacting domain pairs (Ng et al., 2003), our binding motif pairs can be mapped to 203 domain pairs, including 94 high-confidence ones.

6.5 A case study

The 5343 binding motif pairs that we discovered can be ranked according to their correlation score in the mapping. Most of top-ranked binding motif pairs can be confirmed by protein complexes. Here we report details of one such pair. Our purpose is to check whether some block pairs in the motif pair can be aligned with a segment pair in a complex containing the mapped domain pair, and then check whether the segment pair has some contacts among their residues.

This motif pair is generated from the first pair of interacting protein groups. This protein group pair generates three blocks on the left and one block on the right. The first left block 1xxxxxxA contains 24 positions, while the right block 1xright contains 36 positions, as shown in Table 6.7.

Through the approach depicted in section 6.4.2, we map the block pair (1xxxxxxA, 1xright) into domain pair (PF01423,PF01423) in iPfam. Pfam database indicates that

Table 6.7: Left block 1xxxxxxA aligning with the chain A and right block 1xright aligning with the chain B of complex 1mgq, where capital letters are well aligned and lowercase letters are skipped in the alignment

AC 1xxxxxxA; distance from previous block=(18,243)

BL LLE motif=[4,0,17] motomat=[1,80,-10] width=24 seqs=4

DIP:1330N (58) LRDGRMLFGVLRTFD QY A NLI LQD

DIP:2570N (206) TLE GRE I MIRNLSTE LL D ENLLRE

DIP:848N (19) LKNGE I I QG ILT NVD NWM NLT LSN

DIP:883N (244) LQS GRR SKRDLS PEE QR R LQI RHA

pdb1mgq_A(30) LKg dRE f r GVLk SFD Lh M NLv Ln D

AC 1xright; distance from previous block=(6,52)

BL GNL motif=[3,0,17] motomat=[1,80,-10] width=36 seqs=5

DIP:1417N (12) IDK TI N QKVLI V LQS NRE FEG TLV GFD DFV NVI LED

DIP:1418N (53) LSD I I G KTV NVK LAS GLL YSG RLE S I D GFM NVA LSS

DIP:1419N (22) LAK YKD SK I RVK LMGGKL VI G VLK GYD QLM NLV LDD

DIP:794N (7) FKT LVD QEV VVE LKN DI E I KG TLQ SVD QFL NLK LDN

DIP:903N (24) LKD YLN KRV V I I KVDGEC LI A SLN GFD KNT NLF I TN

pdb1mgq_A(18) Lg n s LN S p Vi I KLKG DRE Fr G VLK SFD l h M NLV Ln D

PF01423 is a LSM domain, and iPfam shows that one LSM domain interacts with another LSM domain densely in 20 complexes such as pdb1mgq, pdb1h64. We take the complex pdb1mgq as an example to explain what we found. It has 7 chains each containing a LSM domain. The three-dimensional structure of these 7 chains and their interactions can be found in Figure 6.6 with reference (`http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1mgq`). We observed the following details:

- Our left block 1xxxxxxA can be well aligned at positions 30 to 53 within the LSM domain of the chain A at the complex pdf1mgq, and our right block 1xright can be well aligned at positions 18 to 53 of the chain B also within the LSM domain at the same complex. See Table 6.7 for alignment details.

- The residue 47M (residue M at position 47) of the chain A interacts with residue 48N of the chain B in pdb1mgq; another pair between residue 46H of the chain A and residue 48N of the chain B is also spatially close. See Figure 6.7 for details about the interactions between this segment pair (`http://www.sanger.ac.uk/cgi-bin/Pfam/detailed_interaction_view.pl?acc=PF01423&partner=PF01423&pdb=1mgq`).

- The interaction pair (47M,48N) is well conserved in the complex pdb1mgq—it occurs in seven chain interactions out of a total of nine chain interactions. The seven interactions are between chain A and chain B, between chain B and chain C, ..., and between chain G to chain A. Interestingly, this residue interaction located in the middle of the domain is also highly conserved in other complexes containing LSM domains, e.g. in the complex pdb1h64.

6.6 Discussion and Summary

Our interacting protein group pairs are structurally similar to interacting domain profile pairs proposed by Wojcik and Schachter (2001). But each of their domain profiles is the

Figure 6.6: Three-dimensional structure of the pdb1mgq complex.

summarization of a domain cluster, which is a set of domains sharing significant sequence similarity and interacting with the same region of a certain protein. This approach replies on protein-protein interactions with domain interaction annotations, which are not widely available.

In our model, we require that pairs of interacting protein groups should always have an all-versus-all relationship. This is a bit strict as it is vulnerable to handle incomplete dataset. For the example in the Section 6.1, this strong requirement missed one SH3 protein as it binds to only three out of the four proteins in the other group (Tong et al., 2002). Therefore, this strong requirement may miss many significant pairs, make the discovered groups smaller, and may decrease the significance of the motifs. As a future direction, we will consider *most-versus-most* relationship.

Other future works include new evaluation methods. For example, the predicted interaction sites in the blocks of binding motif pairs can be compared with known interaction sites in some protein-protein interaction databases (Rain et al., 2001) or compared with interaction sites in interface databases (Keskin et al., 2004). Also our binding motif pairs can be compared with those learned from non-interacting protein pairs or from random

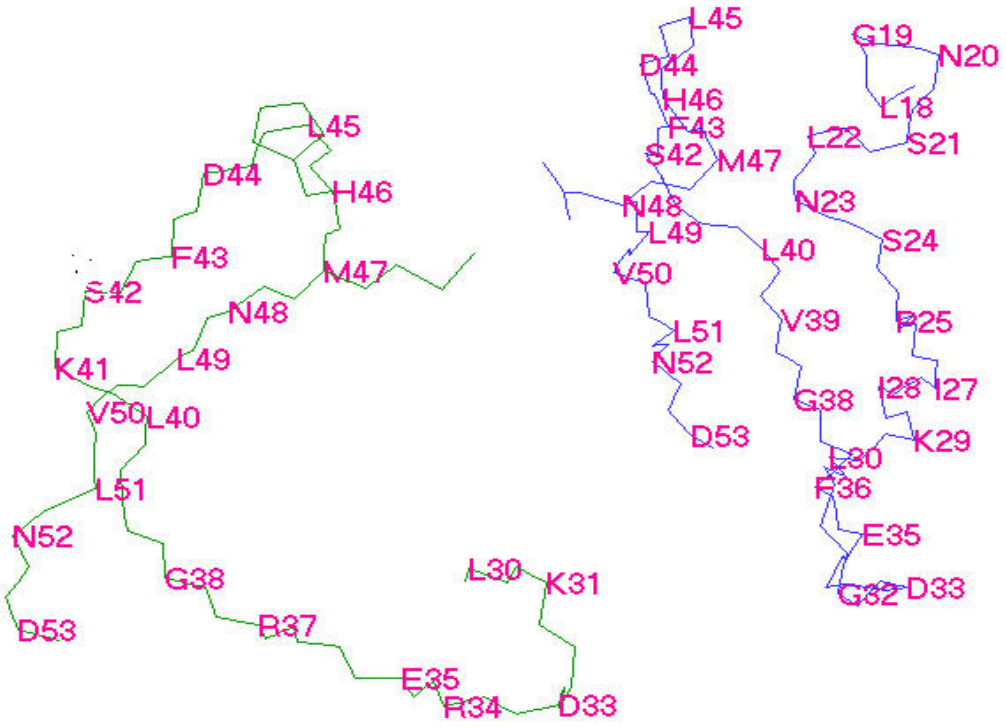


Figure 6.7: Interactions between segment [30L, 53D] of the chain LSM A and segment [18L,53D] of the chain LSM B in the pdb1mgq complex (showing only the backbone).

protein pairs, to study their statistical significance, as shown in Section 4.5.

Finally, we summarize the main results achieved in the group based method. To discover binding motif pairs only from the sequence data of interacting protein pairs, we have proposed the new concept of interacting protein group pairs, where a protein group may share a common interaction motif and a pair of protein groups may share a binding motif pair at their interaction sites. We transformed the mining of interacting protein groups into the mining of frequent closed patterns and then used standard motif discovery algorithms onto these discovered interacting protein groups to generate binding motif pairs in form of blocks. The high efficiency of this two-step approach is because of: (1) In the discovery of interacting protein groups, we examine only interacting protein pairs without checking their sequences, thereby dramatically reduce the complexity of the problem; (2) By producing protein groups first, the discovery of interaction motifs is greatly accelerated as we need not execute the NP-hard motif discovery algorithm on

insignificant candidates of protein sets.

The systematic validation results of the discovered motif pairs indicate that our discovered motifs have high correlation with domains in the existing domains databases. Our discovered motif pairs can also be mapped into the domain–domain interacting pairs in an experimentally validated domain–domain database with good matches. A few case studies on the high-confidence motif pairs confirm that our method is effective.

Chapter 7

Conclusions

As shown in the Chapter 2, current techniques to determine protein interaction sites, either experimental or computational ones, are still in a preliminary stage. So, in this dissertation, we rethink the problem and define new patterns for the prediction of protein interaction sites. Our hypothesis is that the correlation between the two sides of interaction sites should be more important than the compositions on each side. We use *binding motif pairs* to represent the correlated patterns of protein interaction sites and propose two different methods to discover them from various protein-protein interaction data. In Section 7.1, we summarize the two methods along with the major results achieved in the dissertation. In Section 7.2, we point out the limitations of the methods. The future research issues for motif pair discovery are also discussed at the end of this chapter.

7.1 Summary of Results

We have proposed *a fixed point model* in Chapter 3 and Chapter 4 to discover binding motif pairs from protein interaction sequence data and protein complex structural data. In Chapter 3, we defined a point consisting of two traditional motifs and a transformation function closely related to a sequence dataset of interacting protein pairs. We concluded

that the function, especially with percent thresholds, is suitable to simulate the pattern evolution in the interaction sites of protein-protein interactions.

To reduce the huge search space in the search for a complete set of fixed points, we proposed a heuristic method in Chapter 4. It, at the first step, extracted continuous interaction sites from protein complexes, formalized as *maximal contact segment pairs*, and then generalized them to biologically significant starting points for the fixed point model. To evaluate the stable motif pairs (fixed points) we discovered, we proposed P-scores as significance measurements, together with traditional Z-scores, to evaluate both motif pairs and their single motifs. We have presented an efficient method to compute the measurements. In the remainder of the chapter, we reported the overall results of the heuristic fixed point model and confirmed the effectiveness of our model through random experiments and literature validations. Through a series of random experiments, we showed that the choice of maximal contact segment pairs and the choice of starting motif pairs led to the statistical significance of the discovered stable motif pairs. Through a careful comparison between our motif pairs and those in literatures, we illustrated that our motif pairs are promising to be real interaction sites.

To tackle the constraints of limited complex data in the fixed point model, we proposed another method in Chapter 5 and 6. The method discovered binding motif pairs from a novel concept named *interacting protein group pairs*, which reflects an all-versus-all interaction between two protein groups from a protein interaction network. The all-versus-all interaction implies a shared binding mechanism within the groups. In Chapter 5, we found the mining of interacting protein group pairs from a protein interaction network is equivalent with the mining of closed patterns from the adjacency matrix of the interaction network.

In Chapter 6, we carried out systematic validations between our discovered motif pairs and domain-domain interacting pairs. The validations revealed that our motifs have high correlations with domains and our motif pairs match well with experimentally examined domain-domain pairs. The results indicated that our method based on interacting protein

groups is efficient to discover patterns at protein interaction sites.

Overall, the two methods sound reasonable and promising. Various results have confirmed our initial hypothesis.

7.2 Limitations

Although the fixed point model combines different categories of protein interaction data, the current solution is incomplete, highly depending on the limited protein complex data. Moreover, the current transformation function may be too simple to emulate the real evolution in interaction sites.

Although the method based on interacting protein groups is theoretically complete and makes full use of the abundant interacting sequence data, the all-versus-all interaction may not model protein groups with same binding mechanism perfectly and completely, because protein-protein interactions are regulated by complicated mechanism among multiple pockets or interaction sites, thus, all-versus-all interaction may not be achieved always. On the other hand, the coverage of this method is also a problem because of the strict constraint. Consequentially, some proteins may not be covered by any protein group. This prevents the revealing of the interaction sites of the proteins. Besides, the biological significance for the discovered protein groups should be carefully examined, to avoid spurious patterns caused by some imperfect experiments such as TAP (Puig et al., 2001).

Since both methods have their own limitations along with their advantages, a promising method is to combine their strength. In the combined method, interacting protein groups are identified first, then the fixed point model is applied to refine the motif pairs, using a more comprehensive transformation function.

7.3 Further Research Issues

The two methods presented in this thesis are some pioneering work in the research problem of discovering motif pairs, and I think there are several follow-up research works worth of trying in the future.

1. **Complete searching of fixed points**

   As shown in Chapter 3, even the transformation function is in the simplest form, the search space is still huge. We leave this complicated problem in the future, to search a full set of fixed points under the current transformation, or even, under some other more advanced transformation functions.

2. **Mining most-versus-most relationship**

   As mentioned, all-versus-all relationship may not accurately model the interacting protein group pairs. Most-versus-most relationship is believed to be more appropriate to model the protein groups with common binding mechanism (interaction sites) (the relationship is called by our colleagues as quasi-bicliques or quasi-bipartite). The mining of most-versus-most relationship is much more computationally challenging and significant in data mining, bioinformatics and even in graph theory, because it supports many other fields such as communication networks and webs.

3. **Extending the concept of interacting protein groups**

   The concept of interacting protein groups can be extended for more biological significance and applications such as function prediction. Moreover, we can check the biological significance of the generators (key proteins in the closed patterns) in the interacting protein groups. Beside, the interacting protein groups can be extended to protein-DNA interaction networks to examine the significance.

4. **Comparative study of the two methods**

   Since the two methods are able to discover binding motif pairs from the same protein interaction data set, the results should be cross-examined for the overlapping. Even

when the formats are different, motif pairs in form of regular expressions can be transformed into position weight matrices if amino acids in each position are treated equally. Thus, the overall correlation can be evaluated through local alignment of two motif pairs using tools like LAMA (Pietrokovski, 1996).

Note this method has obvious deficiencies. The reason is that our current fixed point model is an incomplete solution. Therefore, the comparison between fixed point model and the one based on interacting protein groups will lead to a poor coincidence rate. Meanwhile, the all-versus-all model for interacting protein groups may miss some motif pairs also, which causes further problems. Hence, we did not perform such validations in the dissertation, while leaving it to the future when complete solutions are found for both methods.

5. **Predicting protein-protein interactions from binding motif pairs**

   A major intention for motif pair discovery is to predict protein-protein interactions. However, it is not a trivial work and the prediction need many careful considerations in the future, such as sufficient coverage of motif pairs, a complicated binding model (whether and how a single motif pair or a group of motif pairs determines an interaction) and the definition of the occurrence of motif pairs. See Aytuna et al. (2005) for some aspects of these issues.

Bibliography

Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th Int'l Conference on Very Large Databases* (pp. 487–499). Chile.

Aloy, P., Pichaud, M., and Russell, R. (2005). Protein complexes: structure prediction challenges for the 21st century. *Current Opinion in Structural Biology*, *15*, 15–22.

Aloy, P., and Russell, R. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, *22*, 1317–1321.

Andrei, Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. L. (2000). Graph structure in the web. *Computer Networks*, *33*, 309–320.

Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Birney, E., and et al. (2001). The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, *29*, 37–40.

Atteson, K. (1998). Calculating the exact probability of language-like patterns in biomolecular sequences. *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB)* (pp. 17–24).

Attwood, T., and Beck, M. (1994). Prints–a protein motif fingerprint database. *Protein Engineering*, *7*, 841–848.

Aytuna, A., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions

by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, *21*, 2850–2855.

Azarya-Sprinzak, E., Naor, D., Wolfson, H. J., and Nussinov, R. (1997). Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Engineering*, *10*, 1109–1122.

Bailey, T., and Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, *21*, 51–80.

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. *Proceedings of the First International Conference on Computational Logic* (pp. 972–986).

Bender, A., and Pringle, J. R. (1991). Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in saccharomyces cerevisiae. *Molecular Cell Biology*, *11*, 1295–305.

Berchanski, A., Shapira, B., and Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. *Proteins*, *56*, 130–142.

Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, *5*, 279–305.

Brunger, A. (1997). X-ray crystallography and nmr reveal complementary views of structure and dynamics. *Nature Structural Biology*, 862–865.

Burdick, D., Calimlim, M., and Gehrke, J. (2001). Mafia: A maximal frequent itemset algorithm for transactional databases. *Proceedings of the Int'l Conference on Data Engineering (ICDE)* (pp. 443–452). Heidelberg, Germany.

Carter, P., Lesk, V., Islam, S., and Sternberg, M. (2005). Protein-protein docking using 3d-dock in rounds 3, 4, and 5 of capri. *Proteins*, *60*, 281–288.

Chakrabarti, P., and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, *47*, 334–343.

Chen, R., Mintseris, J., Janin, J., and Weng, Z. (2003). A protein-protein docking benchmark. *Proteins*, *52*, 88–91.

Chen, R., and Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, *47*, 281–294.

Clemmons, D. (2001). Use of mutagenesis to probe igf-binding protein structure/function relationships. *Endocrine Reviews*, *22*, 800–817.

Comeau, S., Gatchell, D., Vajda, S., and Camacho, C. (2004). Cluspro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Research*, *32*, W96–W99.

Connolly, M. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, *221*, 709–713.

Cunningham, B., and Wells, J. (1989). High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science*, *244*, 1081–1085.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, *23*, 324–328.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, *12*, 1540–1548.

Dominguez, C., Boelens, R., and Bonvin, A. (2003). Haddock: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*, 1731–1737.

Dong, G., and Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. *ACM SIGKDD 1999 International Conference on Knowledge Discovery and Data Mining* (pp. 43–52). San Diego,USA.

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science*, *214*, 149–159.

Doray, B., and Kornfeld, S. (2001). Gamma subunit of the ap-1 adaptor complex binds clathrin: implications for cooperative binding in coated vesicle assembly. *Molecular Biology Cell*, *12*, 1925–1935.

Dziembowski, A., and Seraphin, B. (2004). Recent developments in the analysis of protein complexes. *FEBS Letters*, *556*, 1–6.

Edelsbrunner, H., Facello, M., and Liang, J. (1996). On the definition and the construction of pockets in macromolecules. *Pacific Symposium on Biocomputing* (pp. 272–87). Hawaii.

Ehrlich, L., Nilges, M., and Wade, R. (2005). The impact of protein flexibility on protein-protein docking. *Proteins*, *58*, 126–133.

Eppstein, D. (1994). Arboricity and bipartite subgraph listing algorithms. *Information Processing Letter*, *51*, 207–211.

Evans, J., and Levine, B. (1979). Protein-protein interaction sites of the troponin complex. *Biochemical Society Transactions*, *7*, 701–702.

Fariselli, P., Pazos, F., and et al. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, *269*, 1356–1361.

Fernandez-Gacio, A., Uguen, M., and Fastrez, J. (2003). Phage display as a tool for the directed evolution of enzymes. *Trends in Biotechnology*, *21*, 408–414.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2002). Screened charge electrostatic model in protein-protein docking simulations. *Pacific Symposium on Biocomputing* (pp. 552–563). Hawaii.

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2003). Icm-disco docking by global energy optimization with fully flexible side-chains. *Proteins*, *52*, 113–117.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, *340*, 245–246.

Figeys, D., McBroom, L., and Moran, M. (2001). Mass spectrometry for the study of protein-protein interactions. *Methods*, *24*, 230–239.

Finn, R., Marshall, M., and Bateman, A. (2005). ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, *21*, 410–412.

Fryxell, K. (1996). The coevolution of gene family trees. *Trends in Genetics*, *12*, 364–369.

Gabb, H., Jackson, R., and Sternberg, M. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, *272*, 106–120.

Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology*, *302*, 917–926.

Garman, S., Wurzburg, B., Tarchevskaya, S., Kinet, J., and Jardetzky, T. (2000). Structure of the fc fragment of human ige bound to its high-affinity receptor fc epsilonri alpha. *Nature*, *406*, 259–266.

Gavin, A., Bosche, M., Krause, R., Grandi, P., and et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*, 141–147.

Ge, H., Liu, Z., Church, G., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nature Genetics*, *29*, 482–486.

Glaser, F., Steinberg, D., Vakser, I., and Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, *43*, 89–102.

Goethals, B., and Zaki, M. J. (2003). Fimi03: Workshop on frequent itemset mining implementations. *The Third IEEE International Conference on Data Mining Workshop on Frequent Itemset Mining Implementations* (pp. 1–13).

Grahne, G., and Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. *Workshop on Frequent Itemset Mining Implementations(FIMI)*. USA.

Gribskov, M., McLachlan, A., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of National Academy of Sciences USA*, *84*, 4355–4358.

Grigoriev, A. (2003). On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, *31*, 4157–4161.

Hadley, C., and Jones, D. (1999). A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure*, *7*, 1099–1112.

Halperin, I., Ma, B., and et al. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, *47*, 409–443.

Han, J., and Kamber, M. (2000). *Data mining: Concepts and techniques*. Data Management Systems. Morgan Kaufmann Publishers.

Heifetz, A., Katchalski-Katzir, E., and Eisenstein, M. (2002). Electrostatics in protein-protein docking. *Protein Science*, *11*, 571–587.

Henikoff, S., and Heinikoff, J. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, *19*, 6565–6572.

Henikoff, S., Henikoff, J., Alford, W., and Pietrokovski, S. (1995). Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, *163*, 17–26.

Higgins, D., and Sharp, P. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, *73*, 237–244.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G., and et al. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*, 180–183.

Hoofnagle, A., Resing, K., and NG, A. (2003). Protein analysis by hydrogen exchange mass spectrometry. *Annual Review of Biophysics and Biomolecular Structure*, *32*, 1–25.

Hoogenboom, H., and Chames, P. (2000). Natural and designer binding sites made by phage display technology. *Immunology Today*, *21*, 371–378.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., and Tropsha, A. (2004). Mining protein family specific residue packing patterns from protein structure graphs. *In Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* (pp. 308–315).

Huerta, M., and et al. (2000). *Working definition of bioinformatics and computational biology* (Technical Report). National Institute of Mental Health, USA.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., and et al. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of National Academy of Sciences*, *98*, 4569–4574.

Janin, J. (2001). Welcome to capri: A critical assessment of predicted interactions. *Proteins*, *47*, 257.

Janin, J., Henrick, K., and et al. (2003). Capri: a critical assessment of predicted interactions. *Proteins*, *52*, 2–9.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A. Snyder, M., Greenblatt, J., and Gerstein, M. (2003). A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, *302*, 449–453.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B., and Mackinnon, R. (2003). X-ray structure of a voltage-dependent k+ channel. *Nature*, *423*, 33–41.

Joel, R., and David, A. (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics*, *17*, 455–460.

Jonassen, I. (1997). Efficient discovery of conserved patterns using a pattern graph. *Computer Applications in Biosciences*, *13*, 509–522.

Jonassen, I., Collins, J. F., and G., H. D. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Science*, *4*, 1587–1595.

Jonassen, I., Eidhammer, I., Conklin, D., and Taylor, W. (2001). Structure motif discovery and mining the pdb. *Bioinformatics*, *18*, 362–367.

Jones, S., and Thornton, J. (1997). Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, *272*, 133–143.

Jones, S., and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of National Academy of Sciences USA*, *93*, 13–20.

Kainosho, M. (1997). Isotope labelling of macromolecules for structural determinations. *Nature Structural Biology*, 858–861.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., and Vakser, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of National Academy of Sciences USA*, *89*, 2195–2199.

Kay, B. K., Williamson, M. P., and Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB Journal*, *14*, 231–241.

Kellogg DR, M. D. (2002). Protein- and immunoaffinity purification of multiprotein complexes. *Methods in Enzymology*, *351*, 172–183.

Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, *345*, 1281–1294.

Keskin, O., and Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Engineering Design & Selection*, *18*, 11–24.

Keskin, O., Tsai, C., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Science*, *13*, 1043–1055.

Kretzschmar, T., and von Ruden, T. (2002). Antibody discovery: phage display. *Current Opinion in Biotechnology*, *13*, 598–602.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Computer Networks*, *31*, 1481–1493.

Kumar, S., Ma, B., Tsai, C., Sinha, N., and Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science*, *9*, 10–19.

Lanman, J., and Prevelige, P. J. (2004). High-sensitivity mass spectrometry for imaging subunit interactions: hydrogen/deuterium exchange. *Current Opinion in Structural Biology*, *14*, 181–188.

Leibowitz, N., Fligelman, Z., and Nussinov, R, W. H. (2001). Automated multiple structure alignment and detection of a common substructural motif. *Proteins*, *43*, 235–45.

Li, H., and Li, J. (2005a). Discovery of stable and significant binding motif pairs from pdb complexes and protein interaction datasets. *Bioinformatics*, *21*, 314–324.

Li, H., Li, J., Tan, S., and Ng, S. (2004). Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. *Proceedings of the 9th Pacific Symposium on Biocomputing (PSB)* (pp. 312–323). Hawaii.

Li, H., Li, J., and Wong, L. (2006a). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, *22*, 989–996.

Li, H., Li, J., Wong, L., Feng, M., and Tan, Y. P. (2005a). Relative risk and odds ratio: A data mining perspective. *Proceedings of 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2005)* (pp. 368–377). Baltimore, Maryland, USA.

Li, J., and Li, H. (2005b). Using fixed point theorems to model the binding in protein–protein interactions. *IEEE transactions on Knowledge and Data Engineering*, *17*, 1079–1087.

Li, J., Li, H., Soh, D., and Wong, L. (2005b). A correspondence between maximal complete bipartite subgraphs and closed patterns. *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)* (pp. 146–156). Porto, Portugal.

Li, J., Li, H., Wong, L., Pei, J., and Dong, G. (2006b). Minimum description length (mdl) principle: Generators are preferable to closed patterns. *Proceedings of 21th National Conference on Artificial Intelligence (AAAI-06)* (pp. ??–??). Boston, USA.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science*, *7*, 1884–1897.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, *285*, 2177–2198.

Lu, L., Lu, H., and Skolnick, J. (2002). Multiprospector:an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, *49*, 350–364.

Lupyan, D., Leo-Macias, A., and Ortiz, A. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, *21*, 3255–3263.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of National Academy of Sciences USA*, *100*, 5772–5777.

MacBeath, G., and Schreiber, S. (2000). Printing proteins as microarrays for high-throughput function determination. *Science*, *289*, 1760–1763.

Makino, K., and Uno, T. (2004). New alogrithms for enumerating all maximal cliques. *Proceedings of the 9th Scandinavian Workshop on Algorithm Theory (SWAT 2004)* (pp. 260–272). Springer-Verlag.

Marcotte, E., Pellegrini, M., Ho-Leung, N., Rice, D., Yeates, T., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequence. *Science*, *285*, 751–753.

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, *296*, 910–913.

Matthews, L., Vaglio, P., Reboul, J., Ge, H., Davis, B., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs. *Genome Research*, *11*, 2120–2126.

Mattos, C., and Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature Biotechnology*, *14*, 595–599.

McKay, R., Pearlstone, J., Corson, D., Gagne, S., Smillie, L., and Sykes, B. (1998). Structure and interaction site of the regulatory domain of troponin-c when complexed with the 96-148 region of troponin-i. *Biochemistry*, *37*, 12419–12430.

Mendez, R., Leplae, R., De Maria, L., and Wodak, S. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, *52*, 51–67.

Mendez, R., Leplae, R., Lensink, M., and Wodak, S. (2005). Assessment of capri predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, *60*, 150–169.

Meng, S. W., Zhang, Z., and Li, J. (2004). Twelve c2h2 zinc finger genes on human chromesone 19 can be each translated into the same type of protein after frameshifts. *Bioinformatics*, *20*, 1–4.

Miller, S. (1990). Protein-protein recognition and the association of immunoglobulin constant domains. *Journal of Molecular Biology*, *216*, 965–973.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005). Protein-protein docking benchmark 2.0: an update. *Proteins*, *60*, 214–216.

Mohamed, A. K., and William, A. K. (2001). *An introduction to metric spaces and fixed point theory*. John Wiley & Sons.

Muller, D., Schindler, P., and et al. (2001). Isotope-tagged cross-linking reagents. a new tool in mass spectrometric protein interaction analysis. *Analytical Chemistry*, *73*, 1927–1934.

Mullis, K. (1990). Target amplification for dna analysis by the polymerase chain reaction. *Annales de Biologie Clinique (Paris)*, *48*, 579–582.

Murata, T. (2004). Discovery of user communities from web audience measurement data. *The 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004)* (pp. 673–676).

Mustard, D., and Ritchie, D. (2005). Docking essential dynamics eigenstructures. *Proteins*, *60*, 269–274.

Nakanishi, T., Miyazawa, M., Sakakura, M., Terasawa, H., Takahashi, H., and Shimada, I. (2002). Determination of the interface of a large protein complex by transferred cross-saturation measurements. *Journal of Molecular Biology*, *318*, 245–249.

Nevill-Manning, C. G., Wu, T. D., and Brutlag, D. L. (1998). Highly specific protein sequence motifs for genome analysis. *Proceedings of National Academy of Sciences*, *95*, 5865–5871.

Ng, S., Zhang, Z., and Tan, S. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, *19*, 923–929.

Nicodeme, P., Salvy, B., and Flajolet, P. (2002). Motif statistics. *Theoretical Computer Science*, *287*, 593–618.

Nicolas, P., Yves, B., Rafik, T., and Lotfi, L. (1999). Discovering frequent closed itemsets for association rules. *Proceedings of the 7th International Conference on Database Theory* (pp. 398–416). Israel.

Nietlispach, D., Mott, H., Stott, K., Nielsen, P., Thiru, A., and Laue, E. (2004). *Protein nmr techniques*, vol. 278 of *Methods in Molecular Biology*, chapter Structure determination of protein complexes by NMR, 255–288. second edition.

Ofran, Y., and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Letters*, *544*, 236–239.

Oyama, T., Kitano, K., Satou, K., and Ito, T. (2002). Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, *18*, 705–714.

Pages, S., Belaich, A., Belaich, J., Morag, E., Lamed, R., Shoham, Y., and Bayer, E. (1997). Species-specificity of the cohesin-dockerin interaction between clostridium thermocellum and clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins*, *29*, 517–527.

Paterson, Y., Englander, S., and Roder, H. (1990). An antibody binding site on cytochrome c defined by hydrogen exchange and two-dimensional nmr. *Science*, *249*, 755–759.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, *271*, 511–523.

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

Pearson, K., and Lee, A. (1903). On the laws of inheritance in man. i. inheritance of physical characters. *Biometrika*, *2*, 357–462.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. (1999). Assigning protein functions by comparative genome analysis:protein phylogenetic profiles. *Proceedings of National Academy of Sciences*, *96*, 4285–4288.

Pellicena, P., and Miller, W. (2001). Processive phosphorylation of p130cas by src depends on sh3-polyproline interactions. *Journal of Biological Chemistry*, *276*, 28190–28196.

Peters, K., Fauck, J., and Frommel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of Molecular Biology*, *256*, 201–213.

Phizicky, E., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiology Reviews*, *59*, 94–123.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, *24*, 3836–3845.

Pietrokovski, S., Henikoff, J., and Henikoff, S. (1996). The blocks database–a system for protein classification. *Nucleic Acids Research*, *24*, 197–200.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, *24*, 218–229.

Rain, J., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of helicobacter pylori. *Nature*, *409*, 211–215.

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. *Nature Biotechnology*, *17*, 1030–1032.

Rigoutsos, I., and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, *14*, 55–67.

Ringe, D. (1995). What makes a binding site a binding site? *Current Opinion in Structural Biology*, *5*, 825–829.

Roberts, L., Davenport, R., Pennisi, E., and Marshall, E. (2001). A history of the human genome project. *Science*, *291*, 1195.

Rossmann, M., and Argos, P. (1978). The taxonomy of binding sites in proteins. *Molecular Cell Biochemistry*, *21*.

Russell, R., Breed, J., and Barton, G. (1992). Conservation analysis and structure prediction of the sh2 family of phosphotyrosine binding domains. *FEBS Letters*, *304*, 15–20.

Sauder, J., Arthur, J., and Dunbrack, R. J. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, *40*, 6–22.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, *270*, 467–470.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. (2005). Geometry-based flexible and symmetric protein docking. *Proteins*, *60*, 224–231.

Schueler-Furman, O., Wang, C., and Baker, D. (2005). Progress in protein-protein docking: atomic resolution predictions in the capri experiment using rosettadock with an improved treatment of side-chain flexibility. *Proteins*, *60*, 187–194.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, *18*, 1257–1261.

Shatsky, M., Nussinov, R., and Wolfson, H. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, *56*, 143–156.

Sheu, S., Lancia, D. J., Clodfelter, K., Landon, M., and Vajda, S. (2005). Precise: a database of predicted and consensus interaction sites in enzymes. *Nucleic Acids Research*, *33*, D206–D211.

Shimada, I. (2005). Nmr techniques for identifying the interface of a larger protein-protein complex: cross-saturation and transferred cross-saturation experiments. *Methods Enzymol*, *394*, 483–506.

Sidhu, S. S., Fairbrother, W. J., and Deshayes, K. (2003). Exploring protein-protein interactions with phage display. *Chembiochem*, *4*, 14–25.

Smith, G. (1985a). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, *228*, 1315–1317.

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

Smith, H., Annau, T. M., and Chandrasegaran, S. (1990). Finding sequence motifs in groups of functionally related proteins. *Proceedings of National Academy of Sciences*, *87*, 826–830.

Smith, M. (1985b). In vitro mutagenesis. *Annual Review of Genetics*, *19*, 423–462.

Smith, T., and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*, 195–197.

Song, J., and Markley, J. (2001). Nmr chemical shift mapping of the binding site of a protein proteinase inhibitor: changes in the (1)h, (13)c and (15)n nmr chemical shifts of turkey ovomucoid third domain upon binding to bovine chymotrypsin a(alpha). *Journal of Molecular Recognition*, *14*, 166–171.

Sonnhammer, E., Eddy, S., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*, 405–420.

Spalholz, B., Byrne, J., and Howley, P. (1988). Evidence for cooperativity between e2 binding sites in e2 trans-regulation of bovine papillomavirus type 1. *Journal of Virology*, *62*, 3143–3150.

Sparks, A. B., Rider, J. E., and et al. (1996). Distinct ligand preferences of src homology 3 domains from src, yes, abl, cortactin, p53bp2, plcgamma, crk, and grb2. *Proceedings of National Academy of Sciences USA*, 1540–1544.

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

Stein, A., Russell, R., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research*, *33*, D413–D417.

Stults, J. (1995). Matrix-assisted laser desorption/ionization mass spectrometry (maldi-ms). *Current Opinion on Structural Biology*, *5*, 691–698.

Swanson, R., Lowry, D., Matsumura, P., McEvoy, M., Simon, M., and Dahlquist, F. (1995). Localized perturbations in chey structure monitored by nmr identify a chea binding interface. *Nature Structural Biology*, *2*, 906–910.

Takahashi, H., Nakanishi, T., Kami, K., Arata, Y., and Shimada, I. (2000). A novel nmr method for determining the interfaces of large protein-protein complexes. *Nature Structural Biology*, *7*, 220–223.

Tan, S. H., Sung, W. K., and Ng, S. K. (2004). Discovering novel interacting motif pairs from large protein-protein interaction datasets. *Proceedings of the 4th IEEE Symposium of Bioinformatics and Bioengineering (BIBE2004)* (pp. 568–575). taipei.

Terwilliger, T. (2004). Structures and technology for biologists. *Nature Structural Molecular Biology*, *11*, 296–297.

Thompson, J., Higgins, D., and Gibson, T. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*, 4673–4680.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB)* (pp. 262–271).

Tong, A. H., Drees, B., Nardelli, G., Bader, G., Brannetti, B., Castagnoli, L., and et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, *295*, 321–324.

Tumbarello, D. A., Brown, M. C., and Turner, C. E. (2002). The paxillin ld motifs. *FEBS Letters*, *513*, 114–118.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., and et al. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, *403*, 623–627.

Uno, T., Kiyami, M., and Arimura, H. (2004). Lcm ve. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. *IEEE ICDM'04 Workshop FIMI'04 (International Conference on Data Mining, Frequent Itemset Mining Implementations)*.

Vajda, S. (2005). Classification of protein complexes based on docking difficulty. *Proteins*, *60*, 176–180.

Vajda, S., and Camacho, C. (2004). Protein-protein docking: is the glass half-full or half-empty? *Trends in Biotechnology*, 110–116.

Vancompernolle, K., Vandekerckhove, J., Bubb, M. R., and Korn, E. D. (1991). The interfaces of actin and acanthamoeba actobindin. identification of a new actin-binding motif. *Journal of Biological Chemistry*, *266*, 15427–15431.

Vasilescu, J., Guo, X., and Kast, J. (2004). Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. *Proteomics*, *4*, 3845–3854.

von Mering, C., Krause, R., and et al. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, *417*, 399–403.

Wagner, C. R., and Benkovic, S. J. (1990). Site directed mutagenesis: a tool for enzyme mechanism dissection. *Trends in Biotechnology*, 263–270.

Walhout, A., Sordella, R., Lu, X., Hartley, J., Temple, G., Brasch, M., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, *287*, 116–22.

Wand, A., and Englander, S. (1996). Protein complexes studied by nmr spectroscopy. *Current Opinion in Biotechnology*, *7*, 403–408.

Wang, H., Segal, E., Ben-Hur, A., Koller, D., and Brutlag, D. (2005). Identifying protein-protein interaction sites on a genome-wide scale. *Advances in Neural Information Processing Systems 17* (pp. 1465–1472). USA.

Wang, J., Han, J., and Pei, J. (2003). Closet+: Searching for the best strategies for mining frequent closed itemsets. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)* (pp. 236–245). Washington, DC. USA.

Wiehe, K., Pierce, B., and et al. (2005). Zdock and rdock performance in capri rounds 3, 4, and 5. *Proteins*, *60*, 207–213.

Wilkins, M., Sanchez, J., Gooley, A., Appel, R., Humphery-Smith, I., and Hochstrasser, D. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology & Genetic Engineering Reviews*, *13*, 19–50.

Wodak, S., and Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of Molecular Biology*, *124*, 323–342.

Wojcik, J., and Schachter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, *17*, S296–S305.

Yan, C., Dobbs, D., and Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, *20*, I371–I378.

Yan, X., Yu, P. S., and Han, J. (2005). Substructure similarity search in graph databases. *Proceedings 2005 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'05)* (pp. 766–777). Baltimore, Maryland.

Zacharias, M. (2005). Attract: protein-protein docking in capri using a reduced protein model. *Proteins*, *60*, 252–256.

Zaki, M., and Ogihara, M. (1998). Theoretical foundations of association rules. *Proceedings of 3rd SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 71–78). Seattle, WA, USA.

Zaki, M. J., and Hsiao, C.-J. (2002). Charm: An efficient algorithm for closed itemset mining. *Proceedings of the second SIAM International Conference on Data Mining*.

Zhou, H., and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, *44*, 336–343.