Tel-Aviv University
The Raymond and Beverly Sackler Faculty of Exact Sciences
School of Computer Science

Identification of Conserved Protein Complexes Based on a Model of Protein Network Evolution

This thesis is submitted in partial fulfillment
of the requirements towards the M.Sc. degree
Tel-Aviv University
School of Computer Science

by
Eitan Hirsh

The research work in this thesis has been carried out
under the supervision of Dr. Roded Sharan.

December, 2006

Acknowledgments:

I would like to thank my thesis advisor, Roded Sharan, for the initial idea and direction for this work, and for the continuous and very helpful guidance throughout all research and development stages.
I would also like to thank Trey Ideker for his contribution to the fundamental ideas behind this work, Tomer Shlomi for suggesting the specificity and sensitivity measures and Nir Yosef for supplying the protein-protein interaction data.

Abstract

Data on protein-protein interactions are increasing exponentially. Recent technological advances enable the systematic characterization of protein-protein interaction networks across multiple species. An arising challenge is to organize the accumulating network data into models of cellular machinery. Analysis of proteins taking part in such models can assist in understanding protein function and the dynamics governing inner-cellular processes.

During the past six years several studies focused on finding signaling pathways and protein complexes in protein-protein interaction networks. One of the main challenges these methods face, is the high rate of false positives characterizing the protein-protein interaction data.

As in other biological domains, a comparative approach provides a powerful basis for addressing this challenge. The comparative approach aims to improve the accuracy of protein pathway and complex detection by searching for conserved subnetworks across multiple protein-protein interaction networks. This calls for better understanding of protein-protein interaction network evolution. Two types of processes are considered when studying the evolution of these networks: *link dynamics* and *gene duplication*. The first, models changes in the protein's interaction set due to gene mutation, and the second explains the creation of new proteins through a corresponding gene duplication event. To date, none of the existing methods for finding conserved protein subnetworks uses a probabilistic model that takes evolutionary properties of the network into account.

Here we develop *NetworkBLAST-E*, a new probabilistic approach for identifying protein complexes that are conserved across two species. NetworkBLAST-E describes the evolution of conserved protein complexes from a complex in an ancestral species through link dynamics and gene duplication events. Pairs of extant complexes are scored by their fit to the protein complex model vs. the likelihood that they arise at random. The algorithm we design can be divided into three separate stages. First, the protein-protein interaction and homology data are organized into data models. Second, a search heuristic, which is based on a probabilistic scoring model is executed in order to find potential conserved protein complexes. Finally, a statistical significance filtering stage pinpoints putatively true complexes. These complexes are validated in several ways based on their correspondence to a set of known complexes and based on the functional coherency of their member proteins.

We apply NetworkBLAST-E to search for conserved protein complexes within the protein-protein interaction networks of yeast, fly and human, which are the largest networks in public databases to date. Overall, we detect 1,737 significantly conserved, putatively true, protein complexes that match well known complexes in yeast and are coherent in their functional annotations in yeast, fly and human. In comparison to two previous approaches for protein complex detection NetworkBLAST-E displays higher levels of specificity and sensitivity.

A paper that introduces NetworkBLAST-E together with initial results of applying it to the protein-protein interaction networks of yeast and fly was accepted to ECCB 2006 and will be published in Bioinformatics [22].

Contents

Chapter 1

Introduction

Understanding inner-cellular processes is a critical step in biological and medical research. Proteins take a major part in driving these processes, and through analysis of protein-protein interactions (PPI), researchers can learn much on these processes' dynamics. Recent developments enable wide scale, systematic, characterization of PPI data. Procedures such as yeast two-hybrid [24, 17] and protein co-immunoprecipitation [30] are routinely employed nowadays to generate large-scale PPI networks for human and most model species [19, 23, 25, 39, 49, 52].

An important challenge is to organize the accumulating network data into models of cellular machinery. During the past few years several studies published algorithmic approaches for finding such models [28, 29, 35, 44, 48]. None, however, combine a probabilistic model together with an evolution based scheme.

While PPI detection methods are constantly improving, the data still suffers from high rates of false positives and negatives [16]. As in other biological domains (e.g. biological sequence analysis), a comparative approach provides a powerful basis for addressing this challenge. This approach aims to improve the accuracy of the protein complex search through a comparison among several PPI networks. The idea is that functional network regions are expected to be conserved in evolution. Thus, we expect similar protein complexes to exist in matching regions of PPI networks of different species. Using protein sequence homology, one can match corresponding subnetworks from two or more species and try to find conserved protein subnetworks.

The comparison between several networks calls for better understanding of PPI network evolution. Two types of processes have been invoked to explain the evolution of PPI networks [13, 54]:

link dynamics and gene duplication. The first consists of sequence mutations in a gene that result in modifications of the interface between interacting proteins. Consequently, the corresponding protein may gain new connections (*attachment*) or lose (*detachment*) some of the existing connections to other proteins. The second consists of gene duplication, followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates. The corresponding events in the network are the addition of a protein with the same set of interactions as the original protein, followed by the divergence of their links.

Previous approaches to the problem of identifying protein complexes within PPI networks have shown the utility of comparative analysis. Specifically, Sharan et al. [42, 44] introduced NetworkBLAST, an algorithm that compares PPI networks from multiple species to pinpoint network regions that are conserved in evolution, and have shown that these regions match well known protein complexes in yeast. They used a probabilistic model that scored complexes by their fit to a specific protein complex model vs. the likelihood that they arose at random. However, their scoring scheme treated the networks being compared as independent of one another, and did not take into account the correspondence in interaction patterns between them. Another approach by Koyuturk et al. [29] called MaWish, applied an evolution based scoring scheme, which takes into account duplication and link turnover events. However, the scoring procedure was empirical with no underlying probabilistic model.

Both approaches mentioned above are quite similar. The main difference between them is the scoring model they use in order to calculate the probability that a candidate subnetwork is a true protein complex. They both use the comparative approach and a basic search scheme, which can be divided into three stages:

- **Preprocessing:** PPI data and protein homology data is processed and the required networks are constructed.
- **Search Heuristic:** A greedy search heuristic runs over the networks, scores candidate protein complexes using the specially designed scoring model, and outputs high scoring subnetworks.
- **Post processing:** The output of the previous stage undergoes a statistical filtering process in order to produce a set of significant, non-redundant subnetworks.

In this thesis we adopt the comparative approach and the basic search scheme described above, to develop *NetworkBLAST-E*, a method for detecting conserved protein complexes. Our focus is only on the scoring model, which takes candidate protein subnetworks and tries to asses the

probability they form true conserved protein complexes. We develop a probabilistic model for protein complexes that are conserved across two species, which describes the evolution of conserved protein complexes from an ancestral species through link dynamics and gene duplication events. Pairs of extant complexes are scored by their fit to two distinct models: M_C, a conserved protein complex model and M_N, a null model. The conserved protein complex model assumes the two subnetworks evolved from a single complex in the closest common ancestor of the two species. The null model assumes that every edge in the two subnetworks appears at random, taking vertex degrees into consideration. Finally, the score that is assigned to each pair of subnetworks is the likelihood ratio of these subnetworks being constructed under each of the two models. Thus, our work is the first to formulate a probabilistic model for protein complex conservation that focuses on evolutionary principles.

After publishing our work, a new network alignment method called Graemlin was published [18]. Graemlin, as NetworkBLAST-E, describes the evolution of conserved protein complexes from a hypothetical complex in an ancestral species. However, there are several important differences between the two approaches in terms of the scoring: First, Graemlin uses a probabilistic model in order to score alignment graph edges, which scores the edges independently for each species. In contrast, NetworkBLAST-E uses an evolution based probabilistic model in order to treat edges in the two networks as dependant. Second, Graemlin assumes that all homologous proteins across all examined species are a result of a single protein in the ancestral network, while NetworkBLAST-E does not impose that restriction. Graemlin, unlike NetworkBLAST-E, uses progressive alignment, and thus it is much more scalable and can handle several input networks; while NetworkBLAST-E currently handles only two networks at a time.

An additional issue addressed in this work is the implementation of validation methods for putative conserved complexes. Two references were used: (1) a set of known protein complexes in yeast; and (2) a database of functional annotations of proteins. Several specificity and sensitivity measures are implemented to evaluate the goodness of a set of conserved complexes with respect to these references.

We apply our model to search for conserved protein complexes within the PPI networks of yeast, fly and human. Since the model compares two networks at a time, each of the possible pairs of networks is analyzed separately. Overall, NetworkBLAST-E detects 1,737 significantly conserved subnetworks. More than three-quarters of them were functionally enriched in all three species, serving to validate the biological significance of our findings. When compared to known

complexes in yeast, more than two-thirds significantly matched a known complex, covering more than one fifth of the known complexes. Overall, in comparison to the NetworkBLAST and MaWish, NetworkBLAST-E displays higher levels of specificity, sensitivity and functional enrichment in protein complex detection.

The thesis is organized as follows: Chapter 2 gives general biological and computational background relevant to the issues addressed in this work. Chapter 3 presents the main problem, the basic search scheme and a detailed description of two previous protein subnetwork scoring models. The problems in each of the previous methods are highlighted and serve as the basis for the development of the new method. Chapter 4 describes NetworkBLAST-E our new method for conserved protein complex detection. It includes details on both the probabilistic scoring model and the search and filtering heuristics. Chapter 5 presents the results of applying NetworkBLAST-E to the PPI networks of yeast, fly and human. It also presents a comparison of NetworkBLAST-E's performance to those of two previous approaches. Finally, Chapter 6 gives a brief summary of the thesis and raises open problems for further research.

Chapter 2

Biological and Computational Background

This chapter gives biological background on protein-protein interaction networks, their discovery and evolution. A PPI network represents the physical interactions among proteins of a single species. In this network each node represents a protein and each edge represents an interaction between two proteins. An illustration of the yeast PPI network is given in Figure 2.1.

2.1 High Throughput Protein-Protein Interaction Assays

During the past decade new technological advances enable the systematic characterization of protein-protein interaction networks. The two main methods are yeast two-hybrid (Y2H) [17, 24] and protein co-immunoprecipitation (coIP) [30]. Given two proteins, p_0 and p_1, the Y2H technique tests if they interact. This is done by running an experiment that would cause a reporter gene, g, to be expressed if the two proteins physically interact. The method relies on two protein domains of the yeast GAL4 protein that have specific functions: a DNA-binding domain (BD), capable of binding to a DNA sequence, and an activation domain (AD), capable of activating transcription of the monitored gene. The transcription process of the gene can occur only when both domains are present. Now, two proteins of interest p_0 and p_1 are attached to a binding and activation domains of g, respectively. If p_0 and p_1 interact, an active transcription unit will be formed and g will be expressed, forming a protein product that can be detected and measured (as illustrated in Figure 2.2). The amount of the protein product can be a measure of the interaction between p_0 and p_1.

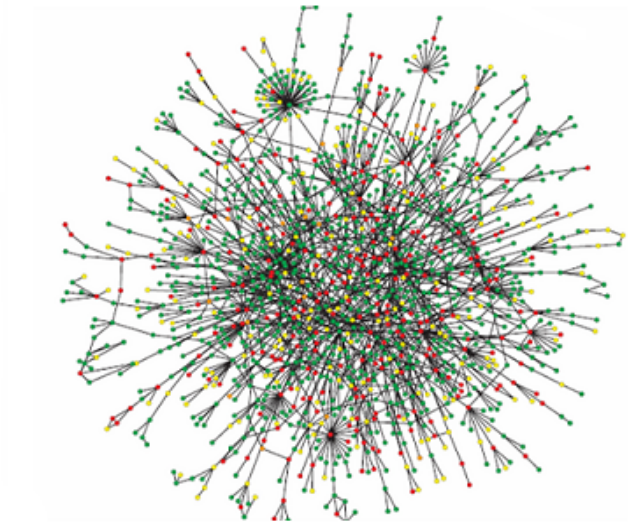

Figure 2.1. Yeast protein-protein interaction network. Presented is the largest cluster of the yeast PPI network, which contains about 78% of all yeast proteins. Figure taken from [26].

In the coIP method, a *bait* protein, p, is marked by a tag. In contrast to the Y2H method, this method does not test if two proteins interact, but rather detects proteins that form an assembly and potentially interact with a specific bait protein. An antibody which recognizes the tag is used in order to trap p, the bait protein, and precipitate it. In the precipitation process any *prey* protein which is in a physical contact or in the same complex with p is precipitated as well. Once a set of all the interacting proteins is found, mass spectrometry is used to identify the prey proteins. See illustration in Figure 2.2.

2.2 Protein-Protein Interaction Networks

Protein-protein interaction data has grown exponentially during the past few years. Procedures, such as Y2H and coIP, described above, are routinely employed to generate large-scale PPI networks. Figure 2.1 illustrate the yeast PPI network. Even though there had been major technological advances in PPI detection, the procedures that are currently employed still suffer from high rates of false positive interactions [53]. Deng et al. [16] estimate the reliability of the Y2H and coIP assays using maximum likelihood estimation on the distribution of gene expression correlation coefficients. PPIs

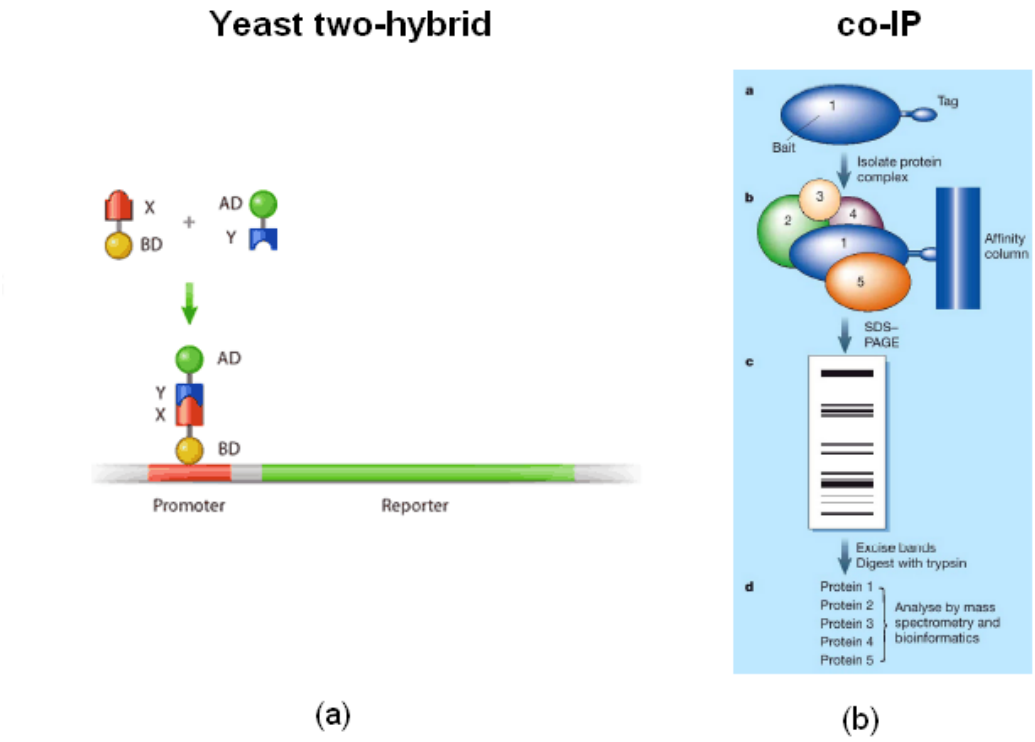

Figure 2.2. High Throughput Protein-Protein Interaction Assays. **(a) The yeast two-hybrid system:** Protein-protein interactions are detected by measuring the expression of a reporter gene. If protein X and protein Y interact, then their DNA-binding domain and activation domain will combine to form a functional transcriptional activator, which will then proceed to transcribe the reporter gene (Figure taken from [46]). **(b) The co-immunoprecipitation assay**: A process divided into four phases. **a)** A specific protein bait is prepared and is attached to an affinity tag that allows the purification of the bait protein and the associated proteins. **b)** The bait protein then interacts with other proteins and is purified. **c)** The purified protein complex is resolved, and discrete protein bands are excited and digested into small peptide fragments. **d)** Peptides are identified using mass spectrometry methods. The identity of a protein associated with a given bait is determined by comparing its peptide fingerprint against known databases (Figure taken from [30]).

detected by coIP [19] and Y2H [52] were shown to have false positives of 42% and 47%, respectively. The reliability of Ito's [25] Y2H PPI data increases as more observations for each interaction are considered, from 83% false positives when considering interactions with a single occurrence, to 4% false positive rate when considering interactions with at least five occurrences (the restriction of the required number of occurrences naturally decreases the amount of interactions detected by the method). False negative rates are more complicated to estimate. A possible method is to estimate the actual number of interactions in a network and calculate the number of missing (unknown) interactions. According to [36] the expected number of interactions in the yeast PPI network is around 30,000 while PPI detection methods (Y2H and coIP) detect about 20,000 interactions. Due to the 50% false positive rate we estimate the false negative rate at 67%.

Due to the low reliability of the PPI data, edges in PPI networks represent noisy observations on the actual interactions. Several authors have suggested methods for evaluating the reliabilities of protein-protein interactions [9, 16, 53]. A common method suggested by Bader et al. [9] and Sharan et al. [42], assigns confidence values to protein interactions using a logistic regression model. The probability of a true interaction between two proteins in a specific species is represented as a logistic function of several observed random variables, such as the number of times an interaction between the proteins was observed in a given experiment, the Pearson correlation coefficient of expression measurements for the corresponding genes and the proteins' small world clustering coefficient.

The PPI network of a species can reveal a lot of information regarding processes that occur inside its cell. Special structures in these networks, such as paths or dense subnetworks, can help analyze protein functions and inner-cellular dynamics. As more PPI networks became available, one of the main challenges was to organize them into models of cellular machinery. There is no specific type of structure we know of as being the most interesting. Two common approaches for modeling interesting protein structures are: (i) *signaling protein pathways*, which are linear chains of interacting proteins that can pass information across different regions of the cell, and are modeled as paths in the PPI network [28, 40, 45, 48]. (ii) *protein complexes*, which are assemblies of proteins that form some cellular machinery, and are modeled as dense protein subnetworks [18, 29, 42, 44].

A study by Steffen et al. [48] identified pathways in PPI networks by applying an exhaustive search procedure to an unweighted interaction graph, considering all interactions equally reliable. The scoring procedure that was used to score a potential path was based on the tendency of its genes to have similar expression patterns. A more advanced method by Scott et al. [40] improves the previous algorithm in two ways. First, by assigning well-founded reliability scores to PPIs, instead

of using an unweighted interaction graph. Second, by exploiting a powerful algorithmic technique by Alon et al. [7], called color coding, to find high-scoring paths efficiently. An additional method by Shlomi et al. [45], called QPath, uses comparative analysis to search for conserved pathways in several PPI networks.

Methods for finding protein complexes were also developed: Sharan et al. [42], Koyuturk et al. [29] and Flannick et al. [18], all search for dense protein subnetworks in PPI networks. A major challenge one encounters when searching for dense protein subnetworks is the high rates of false positives characterizing PPI data. Even the problem of searching for dense subgraphs in a graph when the edges are weighted by 1 and -1 is NP-hard [41]. In the PPI network case, where weights on the edges are defined by probabilities, the search is even harder and calls for an advanced theoretical framework.

A possible approach for dealing with this challenge, is to analyze several PPI networks simultaneously. PPI networks develop according to an evolutionary process, detailed in the next section. Certain regions in PPI networks are expected to be conserved more than others during the course of evolution. It is expected that functional subnetworks in a PPI network, such as protein complexes, would be conserved in evolution. Based on this observation, looking for conserved protein complexes in several PPI networks concurrently, can help overcome the high rates of noise, when searching for protein complexes. In order to compare two or more PPI networks and search for conserved complexes we must first understand the relationship between these networks. This, in turn, calls for a deeper understanding of PPI network evolution.

2.3 Protein-Protein Interaction Network Evolution

Evolution is the change of the properties of a population over many generations. During billions of years evolutionary processes modify existing species and, through the process of speciation, create new ones. Modern understanding of evolution is based on the theory of natural selection, which was presented by Charles Darwin [15].

A widely used model for representing knowledge about the evolutionary relationships between species is the *phylogenetic tree*. In this tree the leaves correspond to extant species and internal nodes correspond to ancestral species. Edge lengths correspond to time estimates. Figure 2.3 demonstrates a rooted phylogenetic tree of life [21]. During the course of evolution, as inner-species DNA

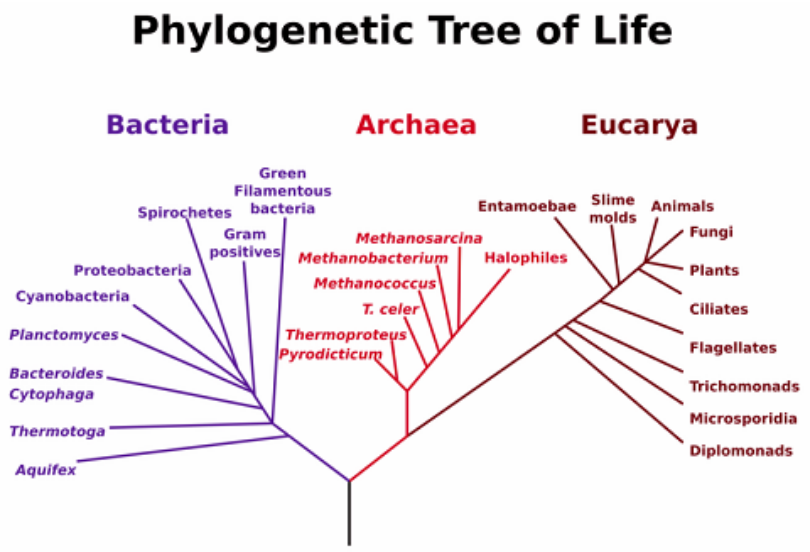

Figure 2.3. Presented is a schematic phylogenetic tree of living things proposed by Carl Woese. It is based on RNA data and shows the separation of bacteria, archaea, and eukaryotes. Figure taken from [33].

sequence undergoes modifications and as new species emerge by speciation events, the PPI networks of the corresponding species also evolve. DNA mutations modify protein interfaces, altering their interaction patterns. Whereas, gene duplication events lead to the creation of new proteins. Species that are closer in the phylogenetic tree are expected to have similar PPI networks and more conserved protein complexes.

A key issue when analyzing PPI networks of multiple species is to understand and take into account the evolutionary processes that produced these networks. Analyzing these processes is a crucial step towards identifying expected relationship between distinct networks and modeling conserved protein complexes.

Two types of processes have been invoked to explain the evolution of PPI networks [13, 54]: link dynamics and gene duplication. The first consists of sequence mutations in a gene that result in modification of the interface between interacting proteins. Consequently, the corresponding protein may gain new connections (*attachment*) or lose (*detachment*) some of the existing connections to other proteins. The second consists of gene duplication, followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates. The corresponding events in the network are the addition of a protein with the same set of interactions as the original protein,

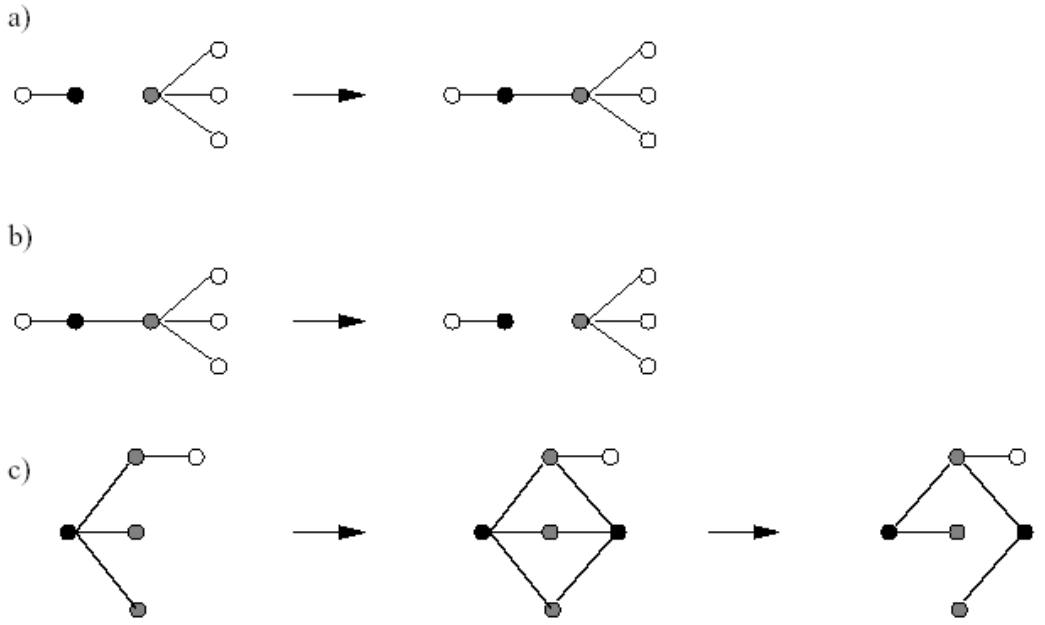

Figure 2.4. The elementary processes of protein network evolution. The progression of time is symbolized by arrows. **a** Link attachment and **b** link detachment occur through nucleotide substitutions in the gene encoding an existing protein. These processes affect the connectivities of the protein whose coding sequence undergoes mutation (shown in black) and of one of its binding partners (shown in gray). Empirical data shows that attachment occurs preferentially towards partners of high connectivity [10]. **c** Gene duplication usually produces a pair of nodes (shown in black) with initially identical binding partners (shown in gray). Empirical data suggests duplications occur at a much lower rate than link dynamics and that redundant links are lost subsequently (often in an asymmetric fashion), which affects the connectivity of the duplicate pair and of all its binding partners [13, 54, 55, 56]. Figure taken from [13].

followed by the divergence of their links. See schematic demonstration in Figure 2.4. Berg et al. [13] estimated the empirical rates of link dynamics and gene duplication in the yeast protein network, finding the former to be at least one order of magnitude higher than the latter. Based on this observation, they proposed a model for the evolution of protein networks in which link dynamics are the major evolutionary forces shaping the topology of the network, while slower gene duplication processes mainly affect its size.

2.4 Comparative Analysis of Networks

Having an understanding of the relationships between PPI networks of different species, we can now design a concurrent search of conserved protein complexes in multiple species. Basically, the use of two or more data sets makes the data more robust and helps overcome the high levels of noise characterizing PPI data [16].

Analysis of PPI networks across multiple species is based on *protein homology*. Two proteins are said to be homologous if they share a common ancestry. Since homologous proteins will tend to be sequence-similar a common approach for detecting them is based on comparing proteins' amino-acid sequences, using, e.g., BLAST [8]. Homology of protein sequences can be of two types: *orthology* or *paralogy*. Homologous proteins across different species are orthologous if they were separated by a speciation event. Homologous proteins within the same species are paralogous if they were separated by a gene duplication event.

The comparative approach has become prominent in the field of PPI network analysis, during the past few years. In their review work Sharan and Ideker [43] give a brief summery of different uses of the comparative approach in this field. They introduce the possibility of comparing whole subnetworks with various structures, which might be conserved across two or more PPI networks. Conserved linear paths, for instance, may correspond to signaling pathways, and conserved connected subnetworks may indicate a conserved protein complex. Even though the problem of finding conserved subnetworks is computationally challenging, heuristic approaches were devised for it, like the one presented by Berg and Lassig [14].

One heuristic approach, called a *network alignment graph*, creates a merged representation of the two networks being compared, facilitating the search for conserved subnetworks. In a network alignment graph, the nodes represent sets of proteins, one from each species, and the edges represent

conserved PPIs across the two species, see example in Figure 3.2. The alignment may consist of one-to-one correspondence between proteins across the two networks; however, in general there may be a many-to-many correspondence between proteins. This scenario can occur, for instance, when a single protein from one species is homologous to multiple proteins from the other species.

A network alignment graph provides the required framework for searching for conserved subnetworks, since these subnetworks will appear as subgraphs with specific structure in the graph. For instance, conserved protein complexes might appear as subgraphs of densely connected nodes. The heuristic was first used by Ogata et al. [35] when searching for correspondences between the reactions of specific metabolic pathways and the genomic locations of the genes encoding the enzymes catalyzing those reactions. Later on, Kelley et al. [28] applied this heuristic to study PPI networks. They translated the problem of finding conserved pathways to that of finding high-scoring paths in the alignment graph. Finally, a network alignment graph, can be used as a basic algorithmic component when searching for conserved protein complexes. This method was used by previous approaches for protein complex search, such as NetworkBLAST [42] and MaWish [29].

Chapter 3

Problem Definition and Previous Work

In this chapter we present previous approaches for finding conserved protein complexes and detail the search scheme and score models they use. We highlight the scenarios under which these methods fail and define the problem our new approach tries to solve.

3.1 Conserved Protein Complex Search Scheme

Detection of conserved protein complexes in two (or more) species can be divided into a five step/module search scheme (as illustrated in Figure 3.1). The first two steps organize the data, one generates protein-protein interaction networks based on Y2H and coIP experimental procedures and the second generates a network alignment graph based on protein homology data, generated by methods such as BLAST. The next two modules execute the actual complex detection algorithm, one performs a search heuristic over the alignment graph and the second supplies a subnetwork scoring model. The fifth module filters the results, leaving only significant and non-redundant conserved protein subnetworks.

3.1.1 PPI Network Data Model

PPI networks are constructed for both studied species. As mentioned earlier, in these networks nodes represent proteins and edges represent pairwise interactions. Since PPI data is very noisy [16], the

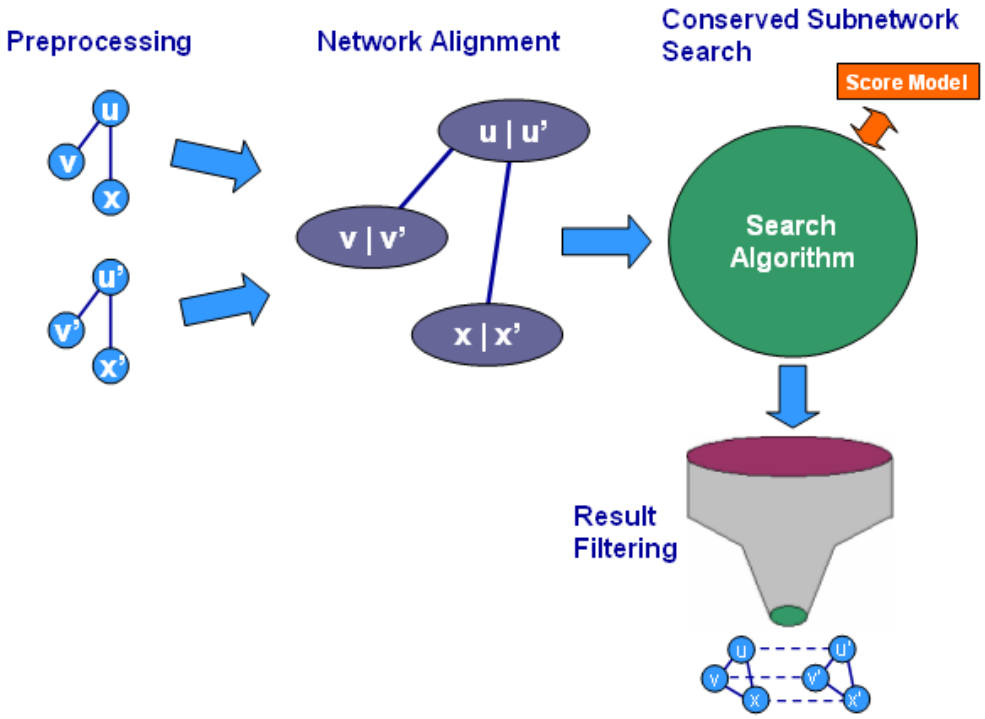

Figure 3.1. Search scheme illustration. A preprocessing phase generates two PPI networks based on PPI data (as detailed in Section 3.1.1). Then, the two networks are aligned and a network alignment graph is constructed, based on homologous relationships between proteins from both networks. Next, a search algorithm, which is based on a subnetwork scoring model, is executed, looking for potential protein complexes (an high-level view of the search algorithm is given in Section 3.1.3). Finally, using several filtering strategies (detailed in Section 3.1.4), only significant and not redundant results are considered.

edges are weighted according to the probability that the two proteins they connect truly interact. We use a method suggested by Sharan et. al. [42] which assigns confidence values to protein-protein interactions using a logistic regression model.

For a given species, the probability of a true interaction between two proteins is defined as a function of the number of times the interaction between these proteins was experimentally observed in each of several different experiments. Indeed, it was shown previously that the number of observations is predictive for the reliability of an interaction [16].

Specifically, given n different experiments, let X_{uv} = (X^1_{uv} ... X^n_{uv}), where X^i_{uv} is the number of observations of an interaction between u and v in experiment i. The probability of a true interaction T_{uv} given X_{uv}, according to the logistic distribution, is:

$$P(T_{uv}|X_{uv}) = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^{n} \beta_i X^i_{uv}}}$$

where β_0 ... β_n are the parameters of the distribution. Given training data, one can optimize the distribution parameters so as to maximize the likelihood of the data using any gradient ascent approach. Following [9], the training data was defined as follows: An observed interaction between protein x and y was considered a positive example if removing it from the network leaves x and y at distance 2. It was considered a negative example if removing it from the network leaves x and y at distance greater than 3. The idea is that in the former case, since there are additional connections between the proteins the chance of the interaction being true is high. Where as in the latter, the interaction is more likely of being a false-positive, since no other connection between the two proteins is observed.

3.1.2 Network Alignment Graph

As discussed earlier, in order to compare between two PPI networks we construct a network alignment graph. In this graph nodes represent sets of proteins, one from each species, and the edges represent conserved PPIs across the two species. The alignment between pairs of distinct proteins from the two species is based on protein homology. Let P_0 and P_1 be the sets of proteins in the PPI networks of species 0 and 1, respectively. For every pair of homologous proteins, $u \in P_0$ and $v \in P_1$, a node $a = (u, v)$ is added to the alignment graph. Edges in the network alignment graph represent *conserved interactions*, which are pairs of observed interactions, one in each species, between corresponding homologous proteins. More precisely, let proteins $u, v \in P_0$ and $u', v' \in P_1$

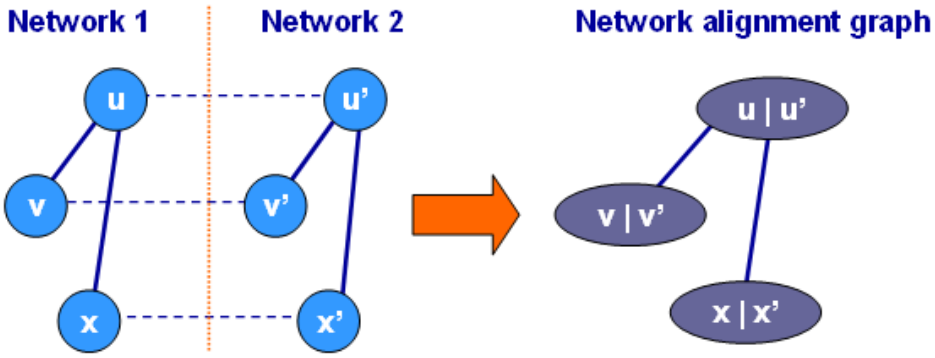

Figure 3.2. Toy example of a network alignment. Network 1 and 2 illustrate PPI networks of two species. Each node represents a protein, solid lines represent PPI and dotted horizontal lines represent homology relationships between proteins from the two species. The alignment graph for the two networks appears on the right. Nodes represent pairs of sequence-similar proteins and edges represent conserved interactions.

take part in two nodes (u, u') and (v, v') in the alignment graph. The two nodes in the alignment graph ((u, u') and (v, v')) are linked if at least one of the pairs $(u, v), (u', v')$ is observed to interact in its PPI network and the second spans proteins of distance at most two in the corresponding PPI network (a protein is considered to be at distance zero from itself). See toy example of this data model in Figure 3.2.

3.1.3 Search Heuristic

The alignment graph is used as a platform for the search of conserved protein complexes across multiple species. By construction, an induced subgraph of the alignment graph corresponds to two species-specific sets of proteins C_0 and C_1, and can be assigned a score (or weight): $Score(C_0, C_1)$. A bottom-up search is performed for heavy (high scoring) subgraphs in the alignment graph, starting with seeds around each of the nodes. These seeds are then expanded by local search, each time adding or deleting a node whose modification increases the weight of the current subgraph the most. Details on the heuristic for finding seeds and the greedy expansion are given in Section 4.2.1.

A key component of the search process is the scoring module, responsible for scoring subnetworks of the alignment graph. A good scoring module should assign high scores to subnetworks

that represent true conserved protein complexes. In this chapter we present two previous scoring models. The next chapter will present NetworkBLAST-E, our new, subnetwork scoring model.

3.1.4 Filtering the Results

Analyzing a large data set and looking for locally high scoring subnetworks yields many subgraphs, some of which may not be true protein complexes. Random sets of proteins may receive locally high scores and be considered by the search heuristic as potentially true. In order to remove these solutions we use a statistical filtering strategy, which assigns significance levels to the results and allows filtering the non-significant subgraphs. The statistical significance of the subgraphs is evaluated by comparing their scores to those obtained on randomized instances of the data. Details on the significance filtering method are given in Section 4.2.2.

Besides the statistical filtering, which discards insignificant subnetworks, the resulting set of putatively true conserved protein complexes may overlap considerably. Several solutions may include the same subnetwork with slight variations. Two subnetworks are said to be *redundant* if their overlap exceeds a predefined threshold. We use a greedy approach to filter redundant subnetworks, the details of which are given in Section 4.2.2.

3.2 Previous Subnetwork Scoring Models

In this section we present in detail two previous models for scoring pairs of aligned subnetworks in two species. We also highlight their shortcomings, motivating the development of a new scoring model. Sharan et al. [42, 44] developed NetworkBLAST, which uses a probabilistic scoring model for scoring PPI subnetworks in several species. Their scoring scheme, however, treated the networks being compared as independent of one another, and did not take into account the correspondence in interaction patterns between them (see detailed discussion of this issue in Section 3.2.1). A second method called MaWish, developed by Koyuturk et al. [29] applied an evolution based scoring scheme, which takes into account duplication and link dynamics events. However, the scoring procedure was empirical with no underlying probabilistic model.

In the following denote by C a subnetwork of interest in the network alignment graph. Let C_0 and C_1 be the two subnetworks induced by C in the PPI networks of the two species 0 and 1, respectively. Let E_0 and E_1 denote the sets of all pairs of proteins in C_0 and C_1, respectively.

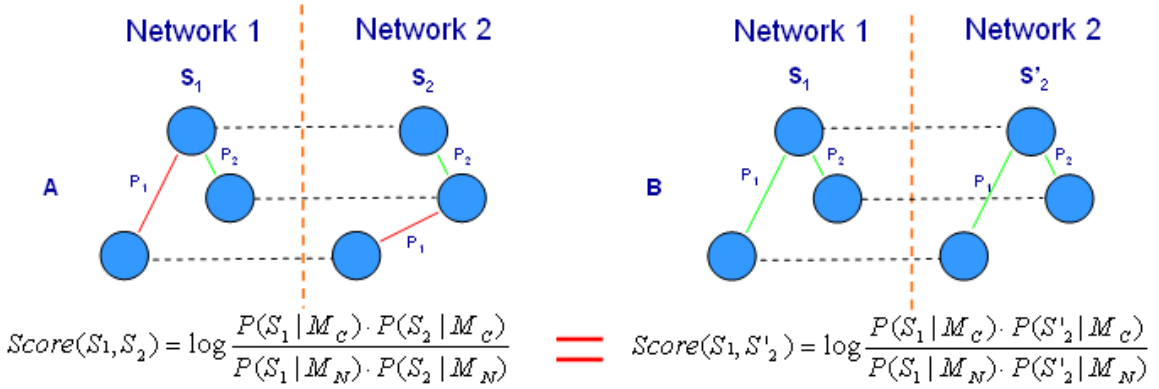

Figure 3.3. NetworkBLAST example. A toy example demonstrating the scoring method of NetworkBLAST and its shortcoming. Shown are three subnetworks: S_1 in Network 1 and S_2, S'_2 in Network 2. Solid lines represent PPIs, and dotted lines represent orthologous relationships. Green solid lines indicate a conserved interaction between two pairs of orthologous proteins and red solid lines indicate a mismatch. Labels on the solid lines indicate the probability for a true interaction. Note that both S_2 and S'_2 are scored the same, since they have similar density properties. So, even though the pair (B) is conserved also at the interaction level, it is given the same total score as pair (A). NetworkBLAST cannot distinguish between the pairs (S_1, S_2) and (S_1, S'_2) since each network is scored independently.

3.2.1 NetworkBLAST

The method considers the two subnetworks C_0 and C_1 as independent. Two models are defined, under which each of the subnetworks could have been created: a protein complex model - M_C, and a null model - M_N. Protein complexes are expected to be dense subnetworks, a property that is formulated in M_C by assuming that every edge appears with some high probability β independently of all other edges. In the null model, M_N, we assume that the subnetwork was randomly selected from the collection of all networks with the same degree sequence. In such a case, an edge between two proteins (u, v) is assumed to appear with some probability r_{uv}, induced by this random model, which depends on the degrees of both u and v (for details on how r_{uv} is calculated see Section 4.1.2).

For a protein pair $(u, v) \in E_0 \cup E_1$, let us denote by T_{uv} the event that these two proteins interact, and by F_{uv} the event that they do not interact (and suppose for now that this information is given to us). Formally, the probability that a given subnetwork C_i was generated by each of the

models is:

$$P(C_i|M_C) = \prod_{(u,v)\in E_i} \beta^{T_{uv}}(1-\beta)^{F_{uv}}$$

$$P(C_i|M_N) = \prod_{(u,v)\in E_i} r_{uv}^{T_{uv}}(1-r_{uv})^{F_{uv}}$$

Using these two models, the score of a subnetwork C is given by a log-likelihood ratio:

$$Score(C) = log\left(\frac{P(C_0|M_C)}{P(C_0|M_N)} \cdot \frac{P(C_1|M_C)}{P(C_1|M_N)}\right)$$

The model aims at distinguishing between a true significantly dense conserved protein complex and a random protein set. The probabilistic model this method is based on allows it to include edge probabilities and additional network properties in its score. For example, this method gives a relatively dense subnetwork that appears in a sparse area of the network a higher score than a dense subnetwork that appears in an area that is generally rich in interactions. However, the disadvantage of the method lies in the fact that it treats the two networks as independent. The evolutionary relations between the networks are not taken into account. Figure 3.3 shows a simple example for this shortcoming. Notice that both pairs of networks (A) and (B) are given the same score, while the subnetworks in (B) are more conserved.

3.2.2 MaWish

This method applies an evolution based scoring model, which takes into account gene duplication and link turnover events. Every set of four proteins, two from each species ($u, v \in P_0$ and $u', v' \in P_1$) is given a weight $W(u, v, u', v')$, based on the probability that the proteins are true orthologs. Specifically, $W(u, v, u', v') = S(u, u') \cdot S(v, v')$, where $S(u, u') \in [0, 1]$ quantifies the likelihood that proteins u and u' are orthologous, and is computed based on their BLAST E-values. When calculating the score for the given subnetwork C, two sets of quadruplets are defined. $M(C)$ - contains all the quadruplets (u, v, u', v') where $u, v \in C_0$ and $u', v' \in C_1$ for which both edges exist ($(u, v) \in E_0$ and $(u', v') \in E_1$). And $N(C)$ - contains all the quadruplets for which an edge exists in one species and not in the other. In addition duplication events are treated as follows: Let $D_0 \in E_0$ and $D_1 \in E_1$ be the sets of pairs of paralogous proteins in species 0 and 1, respectively. Every pair $(u, v) \in D_i$ is assigned a positive/negative duplication factor $d(u, v)$. Due to rapid functional divergence of duplicate proteins, in case the duplication occurred before the speciation event that

split the two examined species, the authors wish to penalize it. Otherwise, in case it occurred after the speciation event, they wish to reward it. The authors employ sequence similarity as a means for distinguishing between events that occurred before and after the speciation event. This is based on the observation that sequence similarity provides a crude approximation for the age of duplication. Finally, the score is calculated by summing over all quadruplets in $M(C)$ and $N(C)$. Conserved interactions increase the score by $\lambda \cdot W(u, v, u', v')$ and non-conserved interactions decrease the score by $\alpha \cdot W(u, v, u', v')$. Duplicate pairs of proteins $(u, v) \in D_0 \cup D_1$ reward/penalize the total weight according to $d(u, v)$. The score is formulated as follows:

$$\begin{aligned} Score(C) \quad = \quad & \sum_{(u,v,u',v') \in M(C)} \lambda \cdot W(u, v, u', v') - \\ & \sum_{(u,v,u',v') \in N(C)} \alpha \cdot W(u, v, u', v') + \\ & \sum_{(u,v) \in D_0} \mu \cdot d(u, v) + \sum_{(u',v') \in D_1} \mu \cdot d(u', v') \end{aligned}$$

where λ, α and μ are parameters of the algorithm (all have positive values).

The strength of this method is that it gives a score to both subnetworks together, taking their topological similarity into account. This evolution based approach helps distinguish between true conserved complexes from random matches of subnetworks in both species. However, since the protein-protein interaction probabilities are not taken into consideration in the model, nor does the local interaction density level around the subnetwork of interest. This method may fail to distinguish between true protein complexes and random subnetworks. Figure 3.4 shows a simple example of using this scoring method. The example shows where the method fails to identify the more conserved dense subnetworks. Both pairs (A) and (B) have the same topological relationship, thus assigned the same score. The method does not take into consideration the fact that S_2 contains proteins with higher degrees than S_2' and that the probabilities on the edges between protein in S_2' are higher than those in S_2, two facts that suggest that the pair (S_1, S_2') is more likely to be a true conserved protein complex.

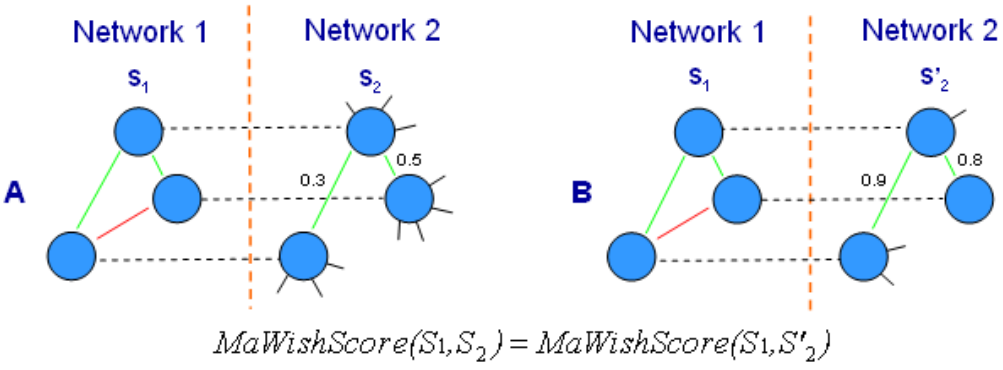

Figure 3.4. MaWish example. A toy example demonstrating MaWish disadvantages. Shown are three subnetworks as in Figure 3.3. Black solid lines indicate the amount of neighbors each node in the PPI network has. Note that both pairs (A) and (B) are scored the same since they share the same matches sequence, without taking PPI probability or amount of neighbors of each protein into account.

3.3 Problem Definition

Discovery of protein complexes in general, and, dense protein subnetworks in particular, is an important step toward understanding cellular processes. It assists in protein function prediction, systematic analysis of cellular machinery, and additional problems at the heart of computational biology. During the last six years, large amounts of PPI data were generated, creating larger and more accurate PPI networks. Data is expected to keep on growing during the next few years. Advanced tools for finding protein complexes are a fundamental requirement at the hands of scientists.

The previous section described previously known methods for this problem and pointed out where these methods fail to identify true conserved protein complexes. As PPI data accumulates, the challenge of pinpointing true protein complexes from random protein sets becomes increasingly important.

The purpose of this research is to develop a new method for identifying conserved protein complexes in two species. We do not alter the basic search scheme used by previous methods, but rather focus on the subnetwork scoring model. Unlike previous methods, we develop a probabilistic model that takes evolutionary properties of the networks into account. This allows us to pinpoint pairs of subnetworks that are both dense and conserved in their interaction patters, thus enhancing the accuracy of the predictions.

Chapter 4

The *NetworkBLAST-E* Algorithm

This chapter gives a detailed description of NetworkBLAST-E, our new method for finding conserved protein complexes. First, a description of the score model is given, followed by a detailed description of the search algorithm. Finally, figures of merit to measure the performance of the algorithm are introduced.

4.1 A Probabilistic Model for Conserved Protein Complexes

In this section we present a probabilistic model for protein complexes that are conserved across two species. The model is based on specifying the pattern of interactions in an unobserved common ancestor of the two species, and on describing the evolutionary events that have yielded the observed subnetworks in each of the species.

We first present the model assuming that the interaction data is accurate and complete, that is, each interaction is true and each non-interaction is also true. We then generalize the model to account for interaction reliabilities. Our full model consists of a conserved protein complex model, M_C, and a null model, M_N. Candidate conserved protein subnetworks, henceforth referred to as *clusters*, are scored by their ratio of likelihoods according to each of the models. In the following we describe these models and their underlying assumptions.

4.1.1 Conserved Protein Complex Model

As suggested by [13] the evolution of a PPI network is shaped by link dynamics and gene duplication events. For a conserved protein complex, consisting of a pair of species-specific complexes, we assume the existence of an ancestral complex in the common ancestor of the two species under study, from which its current forms have evolved through link dynamics and duplication events.

Let the two species under study be indexed by 0 and 1, respectively. Denote their sets of proteins by P_0 and P_1. Denote the set of proteins of a common ancestor of the two species by P. Let $\phi(\cdot)$ be a mapping from proteins in $P_0 \cup P_1$ to P, where $\phi(x) = \phi(y)$ for $x \neq y$ if and only if x and y are homologous. In other words, $\phi(x)$ is the ancestral protein from which x originated. Ways to compute ϕ are described in Section 5.1.3.

Consider a given cluster, and denote by S_0, S_1 and S, the sets of proteins comprising it in species 0, species 1 and the ancestral species, respectively. Our model for the interaction pattern of the ancestral subnetwork is based on the assumption that a protein complex induces a dense subnetwork of protein-protein interactions. This assumption is in agreement with known complexes and has already been used successfully in previous works [44]. Specifically, we assume that within a complex each interaction occurs with high probability β, independently of all other protein pairs in the complex.

The interaction patterns of the extant protein sets, S_0 and S_1, are assumed to have evolved from the ancestral interaction pattern. Let m be the number of protein pairs in the ancestral subnetwork S. For each of these pairs $p_i = (a_i, b_i)$, let I_i be the set of equivalent pairs in S_0 and S_1 under ϕ: $I_i = \{(x, y) \in S_0 : \phi(x) = a_i, \phi(y) = b_i\} \cup \{(x, y) \in S_1 : \phi(x) = a_i, \phi(y) = b_i\}$. We assume that each interaction/ non-interaction relationship between the pairs of proteins in I_i evolved from p_i, independently of all other events. Newly evolved interactions may attach pairs of non-interacting proteins with probability P_A, while detachment of a pair of interacting proteins occurs with probability P_D[1]. Details on the way P_A and P_D are calculated can be found in Section 5.1.4.

To handle duplications in extant species, we have to specify separately our assumption regarding interactions between duplicates, since such interactions did not evolve from an ancestral protein pair as the duplication is assumed to have happened after the speciation event. We choose to treat such interactions in the same manner that we treat interactions in the ancestral species, and assume

[1]Note that P_A and P_D are related: empirical evidence suggests that the overall rate of interaction attachment equals that of interaction detachment [13].

that they occur with probability β independently of all other protein pairs. While there is also information in the number of duplicates of a certain protein that are present in a given subnetwork, empirical evidence suggests that such information is not correlated with complex conservation (see Section 5.1.3). An illustration of the model is given in Figure 4.1.

We are now ready to describe the scoring function. For two proteins x, y, let us denote by T_{xy} the event that these two proteins interact, and by F_{xy} the event that they do not interact. Let $O_{xy} \in \{0, 1\}$ denote the observation on whether x and y interact. Let O_H denote the entire set of observations on the members of H. Let D_{S_i} be the set of duplicate pairs in S_i. The likelihood of a set of observations according to the conserved protein complex model is:

$$P(O_{S_0}, O_{S_1}|M_C) = \prod_{i=1}^{m} P(O_{I_i}|M_C) \cdot \prod_{(x,y)\in D_{S_0}\cup D_{S_1}} P(O_{xy}|M_C)$$

Using the law of complete probability:

$$\begin{aligned} P(O_{I_i}|M_C) &= P(O_{I_i}|T_{a_ib_i})P(T_{a_ib_i}|M_C) + \\ &\quad P(O_{I_i}|F_{a_ib_i})P(F_{a_ib_i}|M_C) \\ &= \beta P(O_{I_i}|T_{a_ib_i}) + (1-\beta)P(O_{I_i}|F_{a_ib_i}) \end{aligned}$$

and

$$\begin{aligned} P(O_{I_i}|T_{a_ib_i}) &= \prod_{(x,y)\in I_i} P(O_{xy}|T_{a_ib_i}) \\ &= \prod_{(x,y)\in I_i} P_D^{[O_{xy}=0]}(1-P_D)^{[O_{xy}=1]} \\ P(O_{I_i}|F_{a_ib_i}) &= \prod_{(x,y)\in I_i} P_A^{[O_{xy}=1]}(1-P_A)^{[O_{xy}=0]} \\ P(O_{xy}|M_C) &= \beta^{[O_{xy}=1]}(1-\beta)^{[O_{xy}=0]} \end{aligned}$$

4.1.2 The Null Model

The null model assumes that each edge in the PPI networks of the two species is present with probability that one would expect if the edges were randomly distributed, preserving vertex degrees.

Formally, for a given PPI network $G = (V, E)$ and a given protein pair (x, y), the probability that x and y interact is defined as the fraction of graphs with the same degree sequence as G that contain an edge between x and y. An approximation to these random interaction probabilities is decribed in a review on networks by Newman [34] and can be computed as follows.

Given the degree sequence of G: d_{v_1} ... $d_{v_{|V|}}$, let us denote by G^F the family of all graphs with the same degree sequence. We can calculate the probability that x and y interact by:

$$P(T_{xy}|M_N) = \frac{1}{1 + \frac{|G^F_{F_{xy}}|}{|G^F_{T_{xy}}|}}$$

where $G^F_{F_{xy}}$ and $G^F_{T_{xy}}$ are the sub-families of all graphs in which x and y do not interact and do interact, respectively. An edge matrix of size $m \times 2$, where $m = \frac{\sum_{i=1}^{|V|} d_{v_i}}{2}$, can be used in order to present each of the possible graphs. The actual amount of possible graphs in each family cannot be calculated analytically but an approximation of the ratio of the sizes can be estimated. We would like to calculate an approximation for the number of possible options to assign d_v distinct occurrences of each vertex $v \in V$ in the matrix, in the two cases: x and y interact or x and y do not interact. Once d_x distinct occurrences of x were already assigned in the matrix we focus on the number of possible assignments of d_y distinct occurrences of y.

In case x and y do not interact, there are: $(m - d_x) \cdot (m - d_x - 1) \cdots (m - d_x + 1) \cdot 2^{d_y}$ options, since the d_y distinct occurrences of y can appear any where except for rows already assigned with x. In case x and y do interact, there are: $d_x \cdot d_y \cdot (m - d_x) \cdot (m - d_x - 1) \cdots (m - d_x + 2) \cdot 2^{d_y - 1}$ options, since one of the distinct occurrences of y needs to be assigned to the same row as one of the distinct occurrences of x and the rest of the $d_y - 1$ distinct occurrences of y are assigned to random rows, as before. In this approximation we do not analyze the amount of possible assignments of the rest of the vertices into the edge matrix. It is impossible to analytically calculate the amount of possible assignments that do not include self loops and parallel edges between all theses vertices. This calculation gives us the approximated ratio:

$$P(T_{xy}|M_N) = \frac{1}{1 + \frac{m - d_x - d_y + 1}{d_x \cdot d_y}}$$

Using this as the probability for interaction between two proteins x and y given the null model, we can calculate $P(O_{xy}|M_N)$ in a similar manner to what we did with the complex model:

$$P(O_{xy}|M_N) = P(T_{xy}|M_N)^{[O_{xy}=1]}(1 - P(T_{xy}|M_N))^{[O_{xy}=0]}$$

This allows us to compute:

$$P(O_{S_0}, O_{S_1}|M_N) = \prod_{x,y \in S_0} P(O_{xy}|M_N) \cdot \prod_{x,y \in S_1} P(O_{xy}|M_N)$$

4.1.3 Noisy observations

The above description assumed that interactions and non-interactions are known. In practice, we have partial, noisy observations on protein-protein interactions. As done in [44], we tackle this problem by generalizing our model to consider the interaction data as noisy observations. To this end, we redefine O_{xy} as the set of experimental observations on whether x and y interact (rather than denoting their status of interaction which is unknown). As before, let T_{xy} and F_{xy} denote the *hidden* events of whether x and y interact or not, respectively. Given the set of experimental observations for a pair of protein (x and y), O_{xy}, we can calculate $P(T_{xy}|O_{xy})$ (the probabilities of the edges of the PPI graph) as described in Section 3.1.1.

We can now use Bayes theorem to compute the likelihood of the observations on an interaction given some model M as follows:

$$\begin{aligned} P(O_{xy}|M) &= P(O_{xy}|T_{xy})P(T_{xy}|M) + \\ & \quad P(O_{xy}|F_{xy})P(F_{xy}|M) \end{aligned}$$

$P(T_{xy}|M)$ and $P(F_{xy}|M)$ are computed as described above (where T_{xy} (F_{xy}) correspond to the event $O_{xy} = 1$ ($O_{xy} = 0$) in the previous notation). $P(O_{xy}|T_{xy})$ and $P(O_{xy}|F_{xy})$ can be computed from interaction reliabilities ($P(T_{xy}|O_{xy})$ and $P(F_{xy}|O_{xy})$) as done by [42] using Bayes theorem:

$$P(O_{xy}|T_{xy}) = \frac{P(T_{xy}|O_{xy})P(O_{xy})}{P(T_{xy})}$$

where $P(T_{xy})$ is the prior probability for a true interaction between two proteins; and $P(O_{xy})$ is the prior probability for observing an interaction between two random proteins. In practice it is not required to calculate $P(O_{xy})$, since it cancels when computing the likelihood ratio.

A straightforward manner for calculating $P(T_{xy})$ would be to sum over the probabilities of all the interactions in the network and divide by the number of protein pairs. However, since there are false-negatives, and we expect the actual amount of true interactions in the network to be f times the amount of the observed interactions. Sharan et al. [42] suggested in their work, to calculate

$P(T_{xy})$ as follows:

$$P(T_{xy}) = \frac{f \cdot \sum_{x,y \in V} P(T_{xy}|O_{xy})}{\frac{|V|(|V|-1)}{2}}$$

where f compensates for false negatives. Here we chose empirically $f = 2$, based also on the estimated total number of interactions in yeast [36].

This modification to the prior probability calls for an updated probability calculation for pairs of proteins for which no interaction was observed. For every pair of proteins x and y, for which x and y were found not to interact ($O_{xy} = \phi$), we assign the updated value:

$$P(T_{xy}|\phi) = \frac{(f-1) \cdot \sum_{(u,v) \in E} P(T_{uv}|O_{uv})}{\frac{|V|(|V|-1)}{2} - |E|}$$

4.1.4 Putting It All Together

Now we can reformulate the previous equation of the **Conserved Complex Model**:

$$P(O_{S_0}, O_{S_1}|M_C) = \prod_{i=1}^{m} P(O_{I_i}|M_C) \cdot \prod_{(x,y) \in D_{S_0} \cup D_{S_1}} P(O_{xy}|M_C)$$

where

$$P(O_{I_i}|M_C) \;=\; \beta P(O_{I_i}|T_{a_i b_i}) + (1-\beta) P(O_{I_i}|F_{a_i b_i})$$

and

$$\begin{aligned} P(O_{I_i}|T_{a_i b_i}) &= \prod_{x,y \in I_i} P(O_{xy}|T_{a_i b_i}) \\ &= \prod_{x,y \in I_i} (P(O_{xy}|T_{xy})(1-P_D) + P(O_{xy}|F_{xy}) P_D) \end{aligned}$$

$$P(O_{I_i}|F_{a_i b_i}) = \prod_{x,y \in I_i} (P(O_{xy}|T_{xy}) P_A + P(O_{xy}|F_{xy})(1-P_A))$$

$$P(O_{xy}|M_C) = P(O_{xy}|T_{xy})\beta + P(O_{xy}|F_{xy})(1-\beta)$$

And the **Null Model**:

$$P(O_{S_0}, O_{S_1}|M_N) = \prod_{x,y \in S_0} P(O_{xy}|M_N) \cdot \prod_{x,y \in S_1} P(O_{xy}|M_N)$$

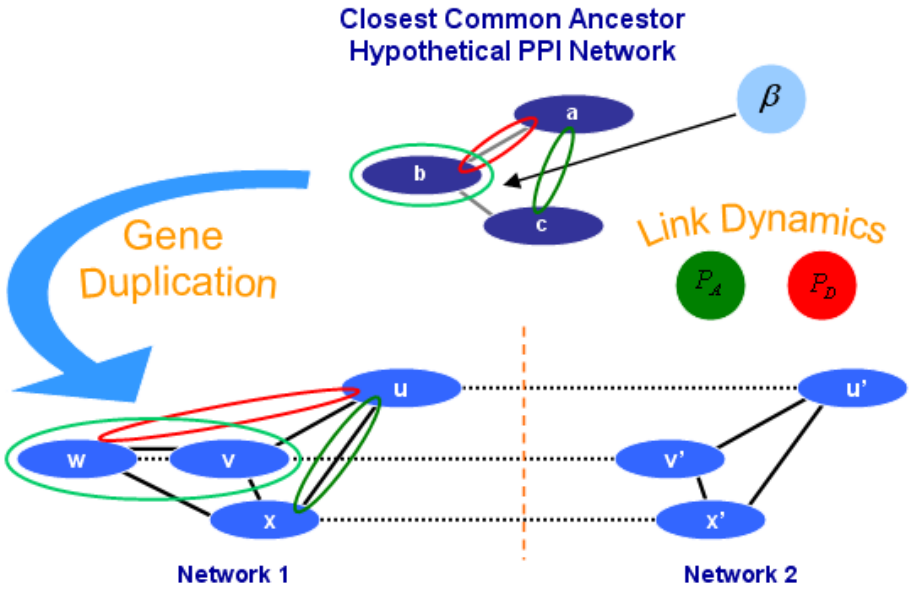

Figure 4.1. Illustration of the new model. Networks 1 and 2 represent two PPI networks. Solid lines represent protein-protein interactions and dotted lines represent homologous relationships between two proteins. As defined by our model, a hypothetical PPI network of the closest common ancestor is considered. Each edge in the ancestor PPI network is expected to appear with probability β. Marked by green and red circles are attachments and detachment events, respectively. These occur between the ancestral network and the observed networks with probability P_A and P_D, respectively. The light green mark indicates a duplication event that occurred in one of the species.

where,

$$P(O_{xy}|M_N) = P(O_{xy}|T_{xy})P(T_{xy}|M_N) + P(O_{xy}|F_{xy})P(F_{xy}|M_N)$$

Finally, the log-likelihood ratio score that we assign to a subgraph is:

$$L(O_{S_0}, O_{S_1}) = log\left(\frac{P(O_{S_0}, O_{S_1}|M_C)}{P(O_{S_0}, O_{S_1}|M_N)}\right)$$

4.2 Searching for Conserved Complexes

As discussed above, a common approach to the problem of identifying conserved protein complexes, is to use an alignment graph. An overview of this approach was given in Section 3.1. We adapt this approach for our search heuristic. This section provides details on the seed construction and greedy search heuristics, and describes result evaluation and filtering methods.

4.2.1 Search Heuristic Details

The seeds are generated based on the following principle: For each node i in the alignment graph choose as a seed the heaviest connected subnetwork of size four that contains i. This is achieved by iterating over all nodes in the alignment graph. For each node an exhaustive local search is executed, to identify all possible connected subnetworks of size four that contain that node. Each of the subnetworks is scored and the highest scoring subnetwork is the one that is added to the set of seeds, with all its subnetworks of size three (that include i).

The seeds are then expanded by local search, each time adding or deleting a node whose modification increases the weight of the current subgraph the most. At each step, the algorithm tries to add one of the neighbors of the nodes that are already part of the cluster or remove one of the cluster members. For each possibility the score of the resulting subnetwork is calculated using the score model and the highest scoring subnetwork is chosen. The greedy search of the subnetwork stops once one of the following conditions occurs:

- The subgraph's size reaches an upper limit (15).
- Every possible change to the current subnetwork will decrease its weight.
- The best possible modification to the subnetwork will yield a non-significant score with respect to the subnetwork's size.

4.2.2 Filtering the Results and Significance Computation

The resulting subgraphs may overlap considerably. We use a greedy algorithm to filter them, so that the intersection of any two subgraphs in their node sets and in their species-specific protein sets is below a threshold (80%; computed with respect to the smaller set). The algorithm iteratively

finds the highest scoring subgraph, adds it to an output list, and removes all the subgraphs that (sufficiently) intersect it from consideration.

The output of the previous stage undergoes further filtering to remove non-significant findings. The statistical significance of the subgraphs is evaluated by comparing their scores to those obtained on randomized instances of the data. These instances are created by shuffling the edges of the two interaction graphs while preserving vertex degrees, as well as shuffling the pairs of homologous proteins while preserving the number of homologous partners per protein. In order to create the degree preserving random PPI networks we use the *swapping* algorithm [32] (see details in Box 1). In order to generate randomized homology data, we use random protein name permutation. Specifically, two random proteins a and b are uniformly selected from one of the species (this procedure is done for both examined species). Then, we switch their names in all the homologous pairs they participate in (see Box 2 for details).

A set of several dozen randomized networks is created using the above procedures. The search algorithm for finding clusters is executed on each of the randomized networks. The results are clustered into groups of clusters with the same size. Only the best result (the one with the highest score) for each cluster size and for each of the runs is recorded. For each cluster size, the score at the 95% percentile is set as the threshold for filtering insignificant clusters. Henceforth, we call the significant clusters, *conserved clusters*.

4.3 Quality Assessment

We used four measures to evaluate the biological significance of the results. The first three quantify the similarity between a given collection of conserved clusters and a reference, putatively true, catalog of protein complexes. As a reference we used known yeast complexes catalogued in the MIPS database [6] (we excluded category 550 which was obtained from high throughput experiments, and retained only manually annotated complexes). The fourth measure assesses the functional coherency of the conserved clusters based on the gene ontology (GO) annotation [51]. These measures are described below.

Specificity and Sensitivity. To measure the level of correspondence between conserved clusters and true complexes, we computed statistically significant matches between the two collections and

Box: 1 - createGeneralizedRandomGraph

Input: G

let $G^*(V, E^*, w) = G(V, E, w)$;

/* n is set to 100 as suggested in [32] */;

for $i = 1$ ... $(n \cdot |E^*|)$ **do**
uniformly choose two random edges (a, b), $(c, d) \subseteq E^*$;

if *a, b, c and d are not all distinct nodes* **then**
continue;

end

randomly choose the switch direction p = cross/parallel;

if *p is cross* **then**

if *edges* (a, d) *and* (b, c) *are not in* E^* **then**
add edges (a, d) and (b, c) to E^*;

set $w(a, d) = w(a, b)$ and $w(b, c) = w(c, d)$;

remove edges (a, b), (c, d) from E^*;

end

else

if *edges* $(a, c), (b, d)$ *are not in* E^* **then**
add edges (a, c) and (b, d) to E^*;

set $w(a, c) = w(a, b)$ and $w(b, d) = w(c, d)$;

remove edges (a, b) and (c, d) from E^*;

end

end

end

return G^*;

Box: 2 - createRandomAlignmentGraph

Input: H - the set of all homologous pairs

/* for each node $a \in H$, let a_i be the protein from species i that takes part in a */;

let $H^* = H$;

for $i = 1$ ... $(100 \cdot |H|)$ **do**
uniformly choose two nodes $a, b \in H^*$;
randomly select a species $s = 0/1$;
switch all occurrences of a_s in H^* to b_s, and vice versa (b_s to a_s);
end

return H^*;

used these matches to evaluate the specificity and sensitivity of the suggested solution. Specifically, for each conserved cluster C we found the true complex A that maximized the hypergeometric overlap score:

$$HG(M, N, T, k) = \sum_{i=k}^{min\{M,N\}} \frac{\begin{pmatrix} N \\ i \end{pmatrix} \begin{pmatrix} T-N \\ M-i \end{pmatrix}}{\begin{pmatrix} T \\ M \end{pmatrix}}$$

where $M = |C|$, $N = |A|$, T is the total amount of proteins in our data set that are spanned by MIPS and $k = |C \cap A|$. In this analysis we consider only proteins that appear both in our protein data set and in at least one MIPS complex.

Significance levels were compared with those obtained for 10,000 random sets of proteins of the same size, and empirical p-values were calculated for each of the conserved clusters. These p-values were further corrected for testing multiple conserved clusters using the false discovery rate (FDR) procedure [12].

Let C be the initial set of conserved clusters, and let $C^* \subseteq C$ be the subset of clusters that had a significant match ($p < 0.05$; only clusters with at least one annotated protein are considered). The *specificity* of the solution is defined as $|C^*|/|C|$. Let M be the set of true complexes, and let $M^* \subseteq M$ be the subset of complexes with a significant match by a conserved cluster. The *sensitivity* of the solution is defined as $|M^*|/|M|$.

Purity. This is an alternative measure for the specificity of the solution. A conserved cluster is called *pure* if there exists a true complex whose intersection with the cluster covers at least 75% of the MIPS annotated proteins in the cluster. Let C be the set of all conserved clusters with at least 3 MIPS annotated proteins, and let $C^* \subseteq C$ be the subset of pure clusters. The *purity* of the solution is defined as $|C^*|/|C|$.

Functional Enrichment. We used the GO [51] process annotation for yeast, fly and human to evaluate the functional coherency of the conserved clusters returned by the algorithm. For each conserved cluster and each GO term, we computed the enrichment of the term in the cluster using a specially-designed hypergeometric score, which takes into account ontology relations between terms. Specifically, since the GO terms are not independent but are rather connected by a directed acyclic graph of parent-child relationship, we computed the enrichment of each term conditioned on the enrichment of its parent term, as done in [42] (see also [20]). Let C be a conserved cluster, let A be a GO term and let T be the union of all of A's parents in the GO hierarchy. The score for each possible match between a conserved cluster and a GO term is calculated by: $HG(|A|, k_p, |T|, k)$, where $k = |C \cap A|$ and $k_p = |C \cap T|$.

For each conserved cluster we chose the term that yielded the highest significance level. We compared this significance level with those obtained for 10,000 random sets of proteins of the same size as the cluster, and derived an empirical p-value for the cluster (in a similar manner as we did with the comparison to MIPS). These p-values were further FDR corrected for multiple testing of conserved clusters. Finally, we report the fraction of functionally enriched clusters ($p < 0.05$; only clusters with at least one GO annotated protein are considered). This procedure is done separately for each of the examined species.

Chapter 5

Experimental Results

We applied NetworkBLAST-E to search for conserved protein complexes in the PPI networks of yeast (*S. Cerevisiae*), fly (*D. Melanogaster*) and human (*H. Sapiens*), which are the three largest networks in public databases. In the following we describe our results and present a comparison to the two existing methods for conserved complex detection described in Section 3.2, NetworkBLAST [42] and MaWish [29].

5.1 Data Description and Parameter Estimation

5.1.1 PPI and Homology Data

We downloaded protein interaction data for yeast, fly and human from the database of interacting proteins (DIP) [3] (July 2005 download). The yeast network contained 15,147 interactions spanning 4,738 proteins. For fly, we complemented the DIP data by interactions from [47], and constructed a network with 23,484 interactions spanning 7,165 proteins. For human, we complemented the DIP data by interactions from [4, 39, 49], and constructed a network with 28,972 interactions spanning 7,915 proteins. We used the previously published logistic regression method [42], detailed in Section 3.1.1, to assign reliabilities to the protein-protein interactions. The reliabilities were based only on the experimental evidence for each interaction.

We analyzed all three network pairs. The network alignment graphs were constructed over pairs of proteins with some interaction information whose BLAST E-value$\leq 10^{-10}$. We used a version of BLAST that was downloaded from [1] with the parameters: b=0, e=1E6, f="C;S" and v=6E5. Overall, the yeast-fly alignment graph contained 890 nodes and 1,070 edges, spanning 482 and 453 distinct proteins in yeast and fly, respectively. The yeast-human alignment graph contained 3,328 nodes and 48,100 edges, spanning 715 and 764 distinct proteins in yeast and human, respectively. The fly-human alignment graph contained 8,308 nodes and 75,636 edges, spanning 1,245 and 869 distinct proteins in human and fly, respectively.

5.1.2 Validation Data

For validation purposes, we downloaded the MIPS complex catalog (December 2005 download). We used complexes at level 3 or lower with at least one protein in the yeast PPI network. Overall, there were 114 such complexes spanning 709 proteins; 68 of these complexes had at least 3 proteins in the network. We also extracted GO process annotations for yeast, fly and human (June 2006 download). There were 4,818, 6,140 and 19,239 annotated proteins for yeast, fly and human, respectively.

5.1.3 Protein Duplication

To determine duplicates and cluster extant proteins according to their ancestral origin, we used the InParanoid algorithm [37]. InParanoid clusters sequence-similar proteins from two species, so that each cluster corresponds to one ancestral protein and contains its present-day descendants and their *inparalogs*, duplicate proteins created after the speciation event. We used InParanoid clustering for yeast-fly, yeast-human and fly-human from the InParanoid public database [5]. For the yeast-fly PPI networks there were 1,128 clusters (December 2005 download). For the yeast-human and fly-human networks there were 889 and 1,268 clusters, respectively (June 2006 download). Note that nodes of the alignment graph may contain pairs of proteins that do not map to the same InParanoid cluster. The inclusion of these nodes reflects a previous observation that functional orthology does not necessarily imply sequence orthology [42].

In order to test whether duplicate pairs of proteins are more/less likely to take part in a conserved protein complex, we employed two statistical tests on the yeast-fly data set:

- Let $u, v \in P_0$ and $u', v' \in P_1$ be two pairs of orthologous proteins ((u, u') and (v, v') are sequence-similar). Assume that conserved interactions are more likely to take part in a conserved complex compared to non-conserved ones. In order to evaluate the tendency of duplicate pairs of proteins to take part in conserved protein complexes, we compared the ratio of duplicate pairs of proteins taking part in conserved interactions and the ratio of duplicate pairs of proteins out of all protein pairs. We found that the ratios are: 0.013 and 0.0034, respectively.
- We compared the ratio of duplicate pairs of proteins in true MIPS complexes and that in random subnetworks. The total amount of duplicate pairs of proteins in known complexes was around 1,300, where as the total amount of duplicate proteins in random clusters, which were generated by a permutation over the proteins in the original MIPS complexes, was around 1,800. Moreover, when comparing the original MIPS complexes and the randomized ones, the amount of complexes that had at least N (ranging from 1 to 5) duplicate pairs of proteins was about 20% higher for the randomized subnetworks.

The statistics show a slight tendency of duplicate pairs of proteins to take part in random protein subnetworks. However, the difference is quite small and does not seem to be a major differentiation factor between the two types of subnetworks. Hence we chose to ignore this factor when scoring protein subnetworks.

5.1.4 Link Dynamics

While previous studies tried to estimate the probabilities P_A and P_D of edge attachment and detachment, respectively [54, 13], these computations were limited to mean estimates over the entire PPI network, and do not directly apply to estimating the rate of these events within conserved complexes. As explained in Section 3.1.1, we expect protein complexes to be dense subnetworks. Our model assumes that each edge occurs with probability β. We expect that during the course of evolution the density property is maintained. Thus, the amount of edges in a conserved complex is expected to remain more or less the same. Let dP_D (dP_A) be the probability that an interaction within a conserved complex is removed (added) in an extant species, within some time unit. We expect dP_D and dP_A to be correlated so that:

$$dP_D \cdot \beta = dP_A \cdot (1 - \beta)$$

Table 1. Comparison of results for different β values.

β	# Complexes	Specificity	Purity	Sensitivity	Functional enrichment	
					Yeast	Fly
0.9	76	76%	66%	11%	88%	63%
0.8	83	77%	70%	12%	87%	60%
0.7	89	75%	63%	13%	88%	53%

Performance measures of NetworkBLAST-E, under different values for the β parameter, when applied to the yeast-fly alignment graph.

Once estimating dP_D and dP_A, we can calculate P_D and P_A using the Jukes-Cantor [27] model. The model is based on a substitution rate matrix. In our case there are two options for each pair of proteins - either they interact or they do not. Given an initial status of interaction between two proteins and the substitution rates for each time unit (dP_D and dP_A), we can calculate the probability for a pair of proteins to interact or not after a certain amount of time t:

$$P_{interact}(t) = P_{interact}(t-1) \cdot (1 - dP_D) + P_{not-interact}(t-1) \cdot dP_A$$

$$P_{not-interact}(t) = P_{interact}(t-1) \cdot dP_D + P_{not-interact}(t-1) \cdot (1 - dP_A)$$

where $P_{interact}(0)$ is set to 1 (0) and $P_{not-interact}(0)$ is set to 0 (1), given the initial interaction status between the two proteins: interacting (non interacting). Knowing the distance, r, of the two examined species from their closest common ancestor, we can calculate the probability for an attachment/detachment event to occur during the evolution process. We set $P_D = P_{not-interact}(r)$, given that the initial status was that the two proteins interacted. And set $P_A = P_{interact}(r)$, given that the initial status was that the two proteins did not interact.

This model for calculating P_D and P_A allows us to estimate only a single parameter, dP_D, and infer the rest. In order to achieve the best performance we enumerated over several values of dP_D and chose the one that gave the highest number of significantly conserved clusters when applying the algorithm to the yeast-fly network. We attained $dP_D = 7 \cdot 10^{-6}$; similar performances were achieved when varying dP_D from $3 \cdot 10^{-6}$ to $4 \cdot 10^{-5}$. These values for dP_D set P_D at values between 0.05 and 0.001, and P_A between 0.2 and 0.004, respectively. The density parameter, β, was set to 0.8 as in a previous work [42]; similar results were obtained when varying β from 0.7 to 0.9, as seen in Table 1.

Table 2. High scoring yeast-fly conserved clusters.

Cluster ID	Size	MIPS category	p-value	Yeast GO process	p-value	Fly GO process	p-value
#4	7	RNA processing	0.03	pre-mRNA splicing factor activity	0.019	RNA binding	0.0031
#214	6	Replication	0.0002	DNA clamp loader activity	0.0001	Nucleotidyl transferase activity	0.009
#222	4	Cytoskeleton	0.0097	Structural constituent of cytoskeleton	0.009	Structural constituent of cytoskeleton	0.015
#313	7	Proteasome	0.0001	Proteasome endopeptidase activity	0.0001	Endopeptidase activity	0.0044

High scoring conserved clusters identified by NetworkBLAST-E when executed on the first data set of the yest-fly networks. For each cluster, shown are its size, best matching MIPS complex (and p-value), and most enriched GO annotations in yeast and fly (and p-values). Specific MIPS categories of the category names mentioned above are as follows: Cytoskeleton - 140.20.20, RNA processing - 440.30.10, Proteasome - 360.10.10 and Replication - 410.40.30.

5.2 Application to Yeast-Fly PPI Network

We applied NetworkBLAST-E to the yeast-fly network alignment graph in search for conserved protein clusters. The algorithm identified 83 significant, non-redundant conserved clusters spanning 155 proteins in yeast and 130 proteins in fly. The sizes of the clusters ranged from 4 to 14, with an average size of 8. Four representative, high scoring conserved clusters are detailed in Table 2 and depicted in Figure 5.1.

We assessed the biological significance of the conserved clusters by comparing them to known MIPS complexes and testing their functional enrichment (see Section 4.3 for a description of the measures we used). 53 of the clusters significantly matched a MIPS complex, yielding a specificity level of 77% and a sensitivity level of 12%. Moreover, 87% of the clusters had an enriched GO annotation in yeast, and 60% were enriched for fly annotations. The enriched annotations in the two species matched in the majority of the cases, as exemplified by the clusters in Table 2. Further information on the identified clusters is given in Table 3.

5.2.1 Comparison to Extant Methods

We compared NetworkBLAST-E with two previously published methods: (1) NetworkBLAST by Sharan et al. [42], which is based on a similar probabilistic model, but treats the two species independently in its score; and (2) MaWish by Koyuturk et al. [29], which is based on evolutionary principles but has no underlying probabilistic model.

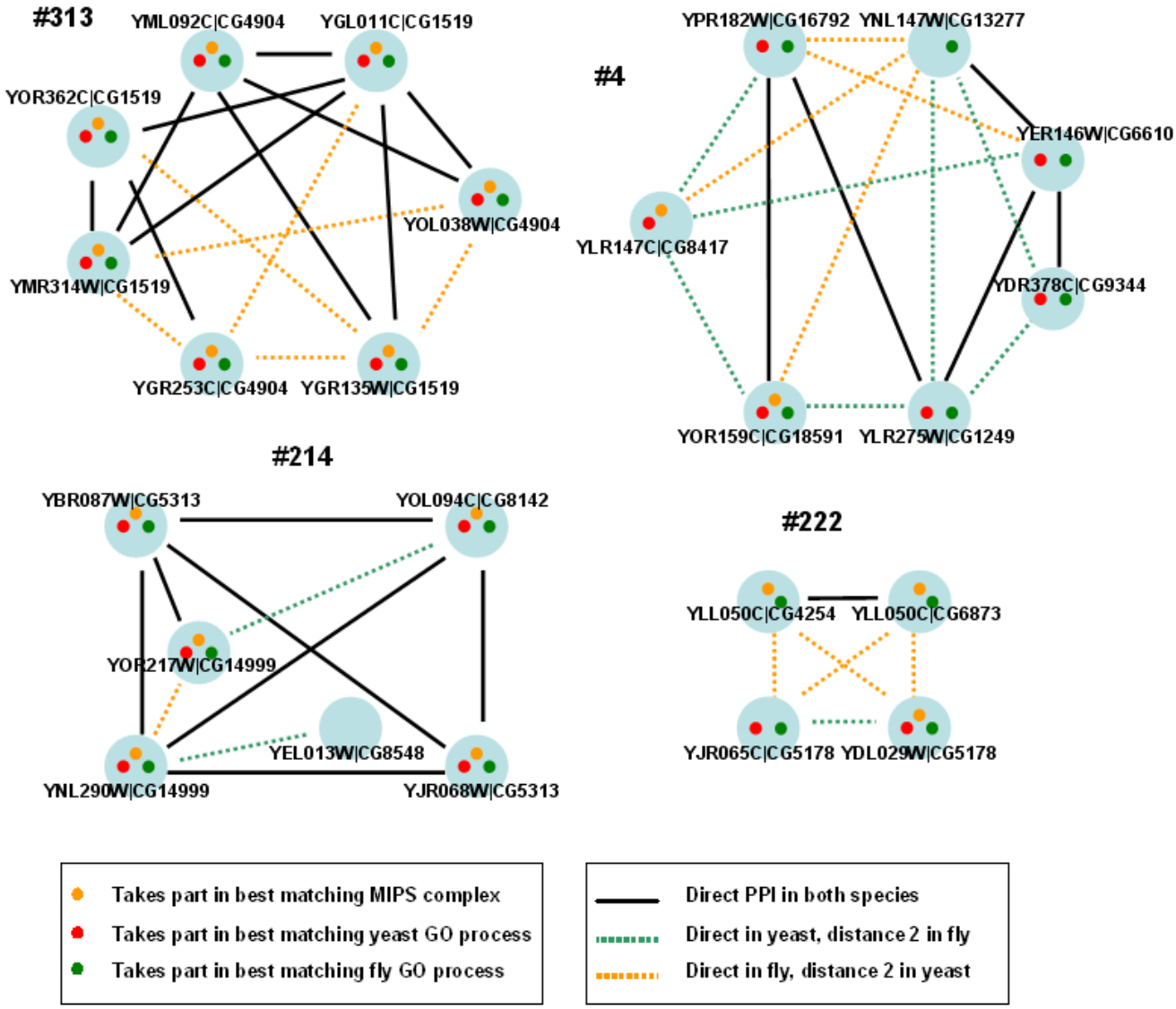

Figure 5.1. Illustration of the four high scoring conserved clusters presented in Table 2. Shown are the alignment subgraphs corresponding to each conserved cluster. Nodes represent pairs of proteins, one from each species. Edges represent conserved (solid) or semi-conserved (dashed; direct in one species and distance 2 in the other) interactions. Edges spanning a direct interaction in one species and the same protein in the other species also appear solid. Colors within nodes indicate whether they participate in the best, significantly matching MIPS complex or GO term.

Table 3. Comparison of results for conserved complex detection.

Algorithm	# Complexes	% Intersection	Specificity	Purity	Sensitivity	Functional enrichment	
						Yeast	Fly
NetworkBLAST-E	83	-	77%	70%	12%	87%	60%
NetworkBLAST [42]	85	92%	77%	60%	11%	87%	58%
MaWish [29]	73	23%	46%	63%	8%	78%	26%

Performance measures of NetworkBLAST-E, NetworkBLAST and MaWish when applied to the yeast-fly alignment graph. The third column specifies the percent of overlapping clusters with NetworkBLAST-E's solution (>80% overlap).

In order to allow a fair comparison we used the same PPI networks, alignment graph, search heuristic and validation methods, thus emphasizing the scoring component of each method. Table 3 summarizes the performances of the three methods when applied to the yeast-fly network. It can be seen that NetworkBLAST-E outperforms MaWish by a significant margin in all measured parameters and that the solutions are very different (23% intersection). In comparison with NetworkBLAST, NetworkBLAST-E has an overall similar performance, which is reflected also in the high overlap between the two solutions (92%). Nevertheless, NetworkBLAST-E exhibits better correspondence with the MIPS catalog, with higher sensitivity and purity levels than those attained by NetworkBLAST.

Due to the overall similarity between the solutions of NetworkBLAST-E and NetworkBLAST, we conducted a more refined analysis of the differences between these approaches. Intuitively, if we consider two species-specific clusters spanning matching sets of proteins, NetworkBLAST will not distinguish between the case that the interactions sets of the two clusters identify and the case they do not (see Figure 3.3). Thus, the key difference between the two approaches is the way they treat conserved interactions within conserved clusters. While the scoring of NetworkBLAST depends only on the total number of interactions within each species, our model distinguishes between a conserved interaction and a pair of species-specific interactions with no match in the other species.

In light of the discussion above, we focused the comparison to NetworkBLAST on clusters containing conserved interactions. We recomputed the quality measures of the two solutions when restricting the computations to conserved clusters that contain at least k conserved interactions, for $k = 1, 2, 3, 4$. This test is motivated by empirical observations on the tendency of interaction conservation across species [31]. The results, summarized in Table 4, demonstrate the superiority of NetworkBLAST-E in this setting, particularly as k grows.

Table 4. Comparison of results for conserved interactions.

Collection	# Complexes	Specificity	Purity	Sensitivity	Functional Enrichment	
					Yeast	Fly
NetworkBLAST-E, $k = 1$	57	70%	52%	12%	82%	67%
NetworkBLAST, $k = 1$	61	69%	42%	11%	80%	59%
NetworkBLAST-E, $k = 2$	34	77%	44%	12%	76%	85%
NetworkBLAST, $k = 2$	38	75%	36%	11%	76%	79%
NetworkBLAST-E, $k = 3$	26	92%	50%	12%	73%	88%
NetworkBLAST, $k = 3$	24	82%	47%	11%	67%	83%
NetworkBLAST-E, $k = 4$	12	90%	50%	11%	67%	92%
NetworkBLAST, $k = 4$	9	71%	25%	10%	44%	78%

Performance measures of NetworkBLAST-E and NetworkBLAST, with respect to conserved clusters containing at least k conserved interactions, for $k = 1, 2, 3, 4$. Columns are as in Table 3.

Moreover, we also applied the two algorithms to a conserved *core* of the network data, obtained by considering only proteins that participate in nodes of the alignment graph that are involved in a conserved interaction. Overall, NetworkBLAST-E's performance is similar to NetworkBLAST's (see Table 5). For comparison purpose, we also detail the performance of MaWish on this data. Evidently, it is less aligned with the MIPS complex data, although displaying high functional enrichment levels.

5.3 Application to Yeast-Human and Fly-Human PPI Networks

Next, we applied NetworkBLAST-E to the yeast-human and fly-human alignment graphs, in search for conserved protein subnetworks. For the yeast-human alignment graph NetworkBLAST-E identified 535 significant, non-redundant conserved clusters spanning 337 proteins in the yeast PPI network and 373 proteins in the human PPI network. The sizes of the conserved clusters detected by the method ranged from 4 to 15, with an average size of 10. When comparing the conserved

Table 5. Comparison of results on a conserved core.

Algorithm	# Complexes	Specificity	Purity	Sensitivity	Functional enrichment	
					Yeast	Fly
NetworkBLAST-E	59	74%	61%	11%	81%	36%
NetworkBLAST	65	77%	51%	11%	82%	32%
MaWish	39	58%	54%	6%	84%	42%

Performance measures of NetworkBLAST-E, NetworkBLAST and MaWish with respect to the conserved core of the yeast-fly alignment graph. Columns are as in Table 3.

clusters with known yeast complexes catalogued in the MIPS database, 150 of the clusters significantly matched a MIPS complex, yielding a specificity level of 61% and a sensitivity level of 32%. Moreover, 86% of the conserved clusters had an enriched GO annotation in yeast, and 87% were enriched for human GO annotations.

When applied to the fly-human alignment graph, NetworkBLAST-E identified 1,119 significant, non-redundant conserved clusters spanning 450 proteins in the human PPI network and 235 proteins in the fly PPI network. The sizes of the conserved clusters detected by the method ranged from 4 to 15, with an average size of 14. The results could not be compared with the MIPS complex data, since it is relevant only for yeast. Thus, specificity, sensitivity and purity measures could not be calculated. The functional enrichment of the results was calculated by a comparison with the GO database. 89% of the conserved clusters had an enriched GO annotation in human, and 88% were enriched for fly GO annotations.

Table 6. Comparison of results for the yeast-human alignment graph.

Algorithm	# Complexes	Specificity	Purity	Sensitivity	Functional enrichment	
					Yeast	Human
NetworkBLAST-E	535	61%	46%	32%	86%	87%
NetworkBLAST	523	58%	41%	33%	89%	89%
MaWish	46	58%	40%	12%	80%	85%

Performance measures of NetworkBLAST-E, NetworkBLAST and MaWish when applied to the yeast-human network. Columns are as in Table 3.

Table 7. Comparison of results for the fly-human alignment graph.

Algorithm	# Complexes	Functional enrichment	
		Fly	Human
NetworkBLAST-E	1,119	88%	89%
NetworkBLAST	1,087	83%	92%
MaWish	214	85%	92%

Performance measures of NetworkBLAST-E, NetworkBLAST and MaWish when applied to the fly-human network. Columns are as in Table 3, MIPS based measures are not applicable for the fly-human data.

For comparison purposes we applied NetworkBLAST and MaWish to both pairs of networks, as detailed in Section 5.2.1. Table 6 and Table 7 summarize the results for applying all three methods to the yeast-human and fly-human networks, respectively. The overall performance of NetworkBLAST-E is very similar to NetworkBLAST, and better than MaWish in most measures. Moreover, NetworkBLAST-E finds 5- to 10-fold more significant conserved clusters than MaWish.

Chapter 6

Conclusions

In this thesis we presented a new method, NetworkBLAST-E, for protein complex detection in PPI networks. Our main contribution is a probabilistic model for the detection of conserved complexes across two species based on the evolutionary processes shaping their networks. Our model has relatively few parameters related to the density of protein complexes, and the determination of link dynamics and gene duplications rates. We applied our approach to study the conservation between the PPI networks of yeast, fly and human. We successfully identified putatively conserved complexes that matched well known complexes in yeast and displayed functional coherency in all three species. Moreover, we have shown that our model aligns with the biological data better than the previous approaches. We expect our model to be more advantageous when comparing evolutionary-closer PPI networks as such become available. In the following we describe open problems and directions for future research.

6.1 Data Integration

A possible approach to overcome the high rates of false negative interactions in current networks is to integrate other data sources [38]. Genetic interaction (GI) data and gene expression data can be used in order to enrich the current networks. We expect that using additional data will yield more conserved interactions and will allow our algorithm to give more accurate results. The main challenge is to integrate the additional data into the protein-protein interaction probability

calculation. This can be achieved by, e.g., adding additional features to the logistic regression calculation of interaction probabilities.

6.2 Protein Subnetwork Model

The current work focused on finding protein complexes, modeled by dense protein subnetworks. It might be interesting to examine other models for protein subnetworks. Paths, for instance, are known to be important structures in PPI networks. Studies like, Kelley et al. [28] and Shlomi et al. [45] concentrate on path finding. Embedding the new evolution based model into these studies may improve their results. Additional models, other than paths, such as trees or other structured subnetworks might be interesting as well. The current scoring model can be modified quite easily to support searching for different subnetwork model types.

6.3 Mapping Proteins to Their Closest Common Ancestor

A very important issue addressed in this work, is the mapping of proteins from the two studied species to a protein in the closest common ancestor. One way to look at this mapping is as a clustering of homologous proteins into sets of proteins that originated from an ancestral protein. We used the InParanoid database of protein clusters as our mapping. The way this mapping is performed has a major effect on the results of our method. Before choosing InParanoid we tried two other methods for creating such a clustering:

- **Single linkage:** In this approach every connected subgraph in the graph of homologous proteins was mapped to a single hypothetical protein in the ancestral species. In this graph, nodes represent proteins from both species and edges link homologous proteins from either different species or the same species.
- **CAST:** CAST is a clustering algorithm designed for clustering gene expression patterns [11]. We adapted this algorithm for clustering homologous proteins. The algorithm supports an affinity parameter, controlling the tightness of the clusters; we tried various values, ranging from 0.1 to 0.7 for this parameter.

InParanoid showed much better performance than the two other approaches and was chosen as our clustering strategy. It might be interesting to check into other clustering approaches and see if they can improve the results. The COG database [50] (can be downloaded form [2]), for instance, can be taken as a candidate clustering.

6.4 Extension to More than Two Species

Sharan et al. [42] demonstrate in their work that the comparative approach can handle more than two species. In the current work we limited ourselves to two species, due to computational limitations, even though using three or more species is expected to improve the results. The key computational challenge lies in controlling the number of alignment graph nodes, which grows in a multiplicative manner with each network added.

In order to avoid this problem, one would have to improve the heuristics for constructing and scanning the alignment graph. Ideas such as progressive network alignment, as done in [18], should be also examined and may be helpful. Adapting the score model to such an approach may be non-trivial, due to the dependency calculation of edge weights among all examined species.

Bibliography

[1] BLAST download site. ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/.

[2] The COG database. http://www.ncbi.nlm.nih.gov/cog/.

[3] The DIP database. http://dip.doe-mbi.ucla.edu/.

[4] Human protein reference database (HPRD). http://www.hprd.org/.

[5] Inparanoid. http://inparanoid.cgb.ki.se/.

[6] The MIPS database. http://mips.gsf.de/.

[7] N. Alon, R. Yuster, and U. Zwick. Color-coding. *Journal of the ACM*, 42:844–856, 1995.

[8] S.F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[9] J.S. Bader, A. Chaudhuri, J.M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature biotechnology*, 22:78–85, 2004.

[10] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[11] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297, 1999.

[12] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57 (1):289–300, 1995.

[13] J. Berg, M. Lassig, and A. Wagner. Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *Bio. Med. Center Evolutionary Biology*, 4:51, 2001.

[14] J. Berg and M. Lassing. Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. USA*, 101:14689–14694, 2004.

[15] C. Darwin. *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray, 1859.

[16] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Eighth Pacific Symposium on Biocomputing*, pages 140–151, 2003.

[17] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, 1989.

[18] J. Flannick, A. Novak, B.S. Srinivasan, H.H. McAdams, and S. Batzoglou. Graemlin:general and robust alignment of multiple large interaction networks. *Genome Research*, 16:1169–1181, 2006.

[19] A.C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

[20] S. Grossmann, S. Bauer, P.N. Robinson, and M. Vingron. An improved statistic for detecting over-represented gene orthology annotations in gene sets. *Research in Computational Molecular Biology*, 3909:85–98, 2006.

[21] S.B. Hedges. The origin and evolution of model organisms. *Nature Genetics*, 3:838–849, 2002.

[22] E. Hirsh and R. Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 2007, in press.

[23] Y. Ho et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[24] T. Ito, T. Chiba, and M. Yoshida. Exploring the yeast protein interactome using comprehensive two-hybrid projects. *Trends Biotechnology*, 19:23–27, 2001.

[25] T. Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98:4569–4574, 2001.

[26] H. Jeong, S.P. Mason, A.L. Barabasi, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–45, 2001.

[27] T. H. Jukes and C. Cantor. Mammalian protein metabolism. *Academic Press*, 1969.

[28] B.P. Kelley et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA.*, 20:11394–9, 2003.

[29] M. Koyuturk et al. Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, 13(2):182–199, 2006.

[30] M. Mann, R.C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem*, 70:437–473, 2001.

[31] L.R. Matthews et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, 11:2120–6, 2001.

[32] R. Milo, N. Kashtan, S. Itzkovski, M.E.J Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv cond-mat/0312028*, 2004.

[33] D. Morrison. Carl woese and new perspectives on evolution. *NASA Astrobiology Institute*, 2003.

[34] M.E.J Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[35] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, 28:4021–4028, 2000.

[36] T. Reguly et al. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of Biology*, 5:11, 2006.

[37] M. Remm, E. Christian, V. Storm, and E.L.L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Molecular Biology*, 314:1041–1052, 2001.

[38] D.R. Rhodes et al. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23:951–959, 2005.

[39] J.F. Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

[40] J. Scott, T. Ideker, R.M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13:133–144, 2006.

[41] R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. *Discrete Applied Mathematics*, 144:173–182, 2004.

[42] R. Sharan et al. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA.*, 102:1974–1979, 2005.

[43] R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006.

[44] R. Sharan, T. Ideker, B.P. Kelley, R. Shamir, and R.M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Journal of Computational Biology*, 12:835–846, 2005.

[45] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath:a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7:199, 2006.

[46] S. Sobhanifar. The yeast two-hybrid assay: an exercise in experimental eloquence. *The science creative quarterly*, 2, 2003.

[47] C.A. Stanyon et al. Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biology*, 5:R96, 2004.

[48] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G.Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3:34–44, 2002.

[49] U. Stelz et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[50] R.L. Tatusov, M.Y. Galperin, D.A. Natale, and E.V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28:33–36, 2000.

[51] The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–9, 2000.

[52] P. Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

[53] von C. Mering et al. Comparative assessment of large-scale data sets of protein-protein interaction. *Nature*, 417:399–403, 2002.

[54] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18:1283–1292, 2001.

[55] A. Wagner. Asymmetric functional divergence of duplicate genes. *Mol. Biol. Evol.*, 19:1760–1768, 2002.

[56] A. Wagner. How the global structure of protein interaction networks evolves. *Proc. Roy. Soc. London*, 270:457–466, 2002.