

# 1 Computational approaches to predict protein-protein and domain-domain interactions

RAJA JOTHI AND TERESA M. PRZYTYCKA

National Center for Biotechnology Information  
National Library of Medicine  
National Institutes of Health  
8600 Rockville Pike, MD 20894

## 1.1 INTRODUCTION

Knowledge of protein and domain interactions provides crucial insights into their functions within a cell. Various high-throughput experimental techniques such as mass-spectrometry, yeast two-hybrid, and tandem affinity purification have generated a significant amount of large-scale protein interaction data [56, 28, 19, 27, 21, 35, 9, 34]. Advances in experimental techniques are paralleled by rapid development of computational approaches designed to detect protein-protein interactions [45, 11, 15, 49, 36, 44, 24, 47]. These approaches complement experimental techniques and, if proven to be successful in predicting interactions, provide insights into principles governing protein interactions.

A variety of biological information (such as amino acid sequences, coding DNA sequences, 3D structures, gene expression, codon usage, etc.) is used by computational methods to arrive at interaction predictions. Most methods rely on statistically significant biological properties observed among interacting proteins/domains. Most widely used properties include co-occurrence, co-evolution, co-expression and co-localization of interacting proteins/domains.

This chapter is, by no account, a complete survey of all available computational approaches for predicting protein and domain interactions but rather a presentation of a bird's eye view of the landscape of large spectrum available methods. For detailed descriptions, performances, and technical aspects of the methods, we refer the reader to the respective articles.

## ii PROTEIN-PROTEIN INTERACTIONS

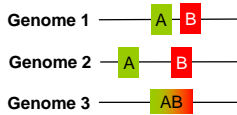
### (a) Phylogenetic profiles

	Genome 1	Genome 2	Genome 3	Genome 4	Genome 5
A	1	0	1	0	1
B	1	0	0	1	1
C	0	0	1	1	0
D	1	0	1	0	1
E	1	0	0	1	1
F	1	0	0	1	1

### Predicted interactions

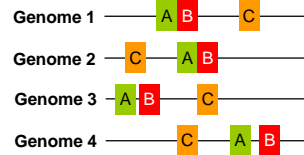


### (b) Gene fusion (Rosetta stone)



Proteins A and B are predicted to interact

### (c) Gene order conservation



**Fig. 1.1** Computational approaches to predicting protein-protein interactions from genomic information. (a) Phylogenetic profiles [18, 49]. A profile for a protein is a vector of 1s and 0s recording presence or absence, respectively, of that protein in a set of genomes. Two proteins are predicted to interact if their phylogenetic profiles are identical (or similar). (b) Gene fusion (Rosetta stone) [36, 15]. Proteins *A* and *B* in a genome are predicted to interact if they are fused together into a single protein (Rosetta protein) in another genome. (c) Gene order conservation [11, 44]. If the genes encoding proteins *A* and *B* occupy close chromosomal positions in various genomes, then they are inferred to interact. Figure reprinted with permission from [?, ??]

## 1.2 PROTEIN-PROTEIN INTERACTIONS

### 1.2.1 Phylogenetic profiles

The patterns of presence or absence of proteins across multiple genomes (phylogenetic or phyletic profiles) can be used to infer interactions between proteins [18, 49]. A phylogenetic profile for each protein  $i$  is a vector of length  $n$  that contains the presence or absence information of that protein in a reference set of  $n$  organisms. The presence or absence of  $i$  in organism  $j$  is recorded as  $P_{ij} = 1$  or  $P_{ij} = 0$ , respectively, which is usually determined by performing a BLAST search [4] with an e-value threshold  $t$ . If the BLAST search results in a hit with e-value  $< t$ , then it is construed as an evidence for the presence of protein  $p$  in  $G$ . Otherwise, it is assumed that  $p$  is absent in  $G$ .

Proteins with identical or similar profiles are inferred to be functionally interacting under the assumption that proteins involved in the same pathway or functional system are likely to have been co-inherited during evolution [18, 49] (Figure 1.1a).

Similarities between profiles can be measured using metrics such as Hamming distance, Jaccard coefficient, mutual information, etc. It has been shown that measuring profile similarity using mutual information rather than metrics such as Hamming distance results in a better prediction accuracy [22]. By clustering proteins based on their profile similarity scores, one can construct functional pathways and interaction network modules [12, 22]. One of the main limitations of the profile comparison approach is the lineage-specific gains and losses of genes, thought to be more pervasive in microbial evolution [38], which could artificially decrease the similarity between functionally interacting genes.

Instead of using an ad-hoc e-value threshold and binary values as originally proposed [49], recent studies have been using  $P_{ij} = -1/\log E_{ij}$  to record the presence/absence information, where  $E_{ij}$  is the BLAST e-value of the top-scoring sequence alignment of protein  $i$  in organism  $j$ . To avoid algorithm-induced artifacts,  $P_{ij} > 1$  are truncated to 1. Notice that a zero (or a one) entry in the profile now indicates the presence (absence, respectively) of a protein. It is being argued using real values for  $P_{ij}$ , instead of binary values, captures varying degrees of sequence divergence, providing more information than the simple presence or absence of genes [36, 12, 32].

For a more comprehensive assessment of the phylogenetic profile comparison approach, we refer the reader to [32].

### 1.2.2 Gene fusion events

There are instances where a pair of interacting proteins in one genome is fused together into a single protein (referred to as the Rosetta Stone protein [36]) in another genome. For example, interacting proteins Gyr A and Gyr B in *Escherichia coli* are fused together into a single protein (topoisomerase II) in *Saccharomyces cerevisiae* [7]. Amino acid sequences of Gyr A and Gyr B align to different segments of the topoisomerase II. Based on such observations, methods have been developed [36, 15] to predict interaction between two proteins in an organism based on the evidence that they form a part of a single protein in other organisms. A schematic illustration of this approach is shown in Figure 1.1b.

### 1.2.3 Gene order conservation

Interactions between proteins can be predicted based on the observation that proteins encoded by conserved neighboring gene pairs interact (Figure 1.1c). This idea is based on the notion that physical interaction between encoded proteins could be one of the reasons for evolutionary conservation of gene order [11]. Gene order conservation between proteins in bacterial genomes has been used to predict functional interactions [11, 44]. This approach's applicability only to bacterial genomes, in which the genome order is a relevant property, is one of its main limitations [57]. Even within the bacteria, caution must be exercised while interpreting conservation of gene order between evolutionarily closely related organisms (for example, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*) as lack of time for genome

rearrangements after divergence of the two organisms from their last common ancestor could be a reason for the observed gene order conservation. Hence, only organisms with relatively long evolutionary distances should be considered for such type of analysis. However, the evolutionary distances should be small enough in order to ensure that a significant number of orthologous genes is still shared by the organisms [11].

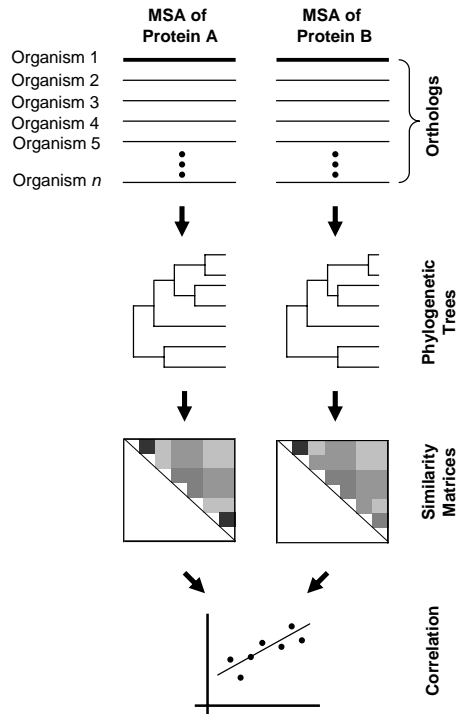
#### 1.2.4 Similarity of phylogenetic trees

It is postulated that the sequence changes accumulated during the evolution of one of the interacting proteins must be compensated by changes in its interaction partner. Such correlated mutations have been subject of several studies [3, 23, 40, 53]. Pazos et al. [45] demonstrated that the information about correlated sequence changes can distinguish right interlocking sites from incorrect alternatives. In recent years, a new method emerged, which, rather than looking at co-evolution of individual residues in protein sequences, measures the degree of co-evolution of entire protein sequences by assessing the similarity between the corresponding phylogenetic trees [24, 25, 45, 47, 50, 31, 46, 52, 30, 33]. Under the assumption that interacting protein sequences and their partners must co-evolve (so that any divergent changes in one partner's binding surface are complemented at the interface by their interaction partner) [39, 6, 45, 29], pairs of protein sequences exhibiting high degree of co-evolution are inferred to be interacting.

In this section, we first describe the basic "mirror-tree" approach for predicting interaction between proteins by measuring the degree of co-evolution between the corresponding amino acid sequences. Next, we describe an important modification to the basic mirror-tree approach, which helps in improving its prediction accuracy. Finally, we discuss a related problem of predicting, based on the co-evolution hypothesis, interaction specificity between two families of proteins (say, ligands and receptors), which are known to interact.

**1.2.4.1 The basic mirror-tree approach** This approach is based on the assumption that phylogenetic trees of interacting proteins are highly likely to be similar due to the inherent need for coordinated evolution [24, 48]. The degree of similarity between two phylogenetic trees is measured by computing the correlation between the corresponding distance matrices, which implicitly contains the evolutionary histories of the two proteins.

A schematic illustration of the mirror-tree method is shown in Figure 1.2. The multiple sequence alignments (MSA) of the two proteins, for a common set of species, are constructed using one of many available MSA algorithms such as ClustalW [55], MUSCLE [14], T-Coffee [42]. The set of orthologous proteins for a MSA is usually obtained by one of the two following ways: (i) a stringent BLAST search with a certain e-value threshold, sequence identity threshold, alignment overlap percentage threshold or a combination thereof, or (ii) reciprocal (bi-directional) BLAST best-hits. In both approaches, orthologous sequences of a query protein  $q$  in organism  $Q$  is searched by performing a BLAST search of  $q$  against sequences in other organisms.



**Fig. 1.2** Schema of the mirror-tree method. Multiple sequence alignments of proteins A and B, constructed from orthologs of A and B respectively from a common set of species, are used to generate the corresponding phylogenetic trees and distance matrices. The degree of co-evolution between A and B is assessed by comparing the corresponding distance matrices using a linear correlation criteria. Proteins A and B are predicted to interact if the degree of co-evolution, measured by the correlation score, is high (or above a certain threshold).

In the former,  $q$ 's best-hit  $h$  in organism  $H$ , with  $e$ -value  $< t$ , is considered to be orthologous to  $Q$ . In the latter,  $q$ 's best-hit  $h$  in organism  $H$  (with no specific  $e$ -value threshold) is considered to be orthologous to  $q$  if and only if  $h$ 's best-hit in organism  $Q$  is  $q$ . Using reciprocal best-hits approach to search for orthologous sequences is considered to be much more stringent than just using unidirectional BLAST searches with an  $e$ -value threshold  $t$ .

In order to be able to compare the evolutionary histories to two proteins, it is required that the two proteins have orthologs in at least a common set of  $n$  organisms. It is advised that  $n$  be large enough for the trees and the corresponding distance matrices to contain sufficient evolutionary informatin. It is suggested that  $n \geq 10$ . Phylogenetic trees from MSA are constructed using standard tree construction algorithms (such as neighbor-joining, UPGMA, etc), which are then used to construct the distance matrices (algorithms to construct trees and matrices from MSAs are available in the ClustalW suite).

The extent of agreement between the evolutionary histories of two proteins is assessed by computing the degree of similarity between the two corresponding distance matrices. The extent of agreement between matrices  $A$  and  $B$  can be measured using Pearson's correlation coefficient, given by

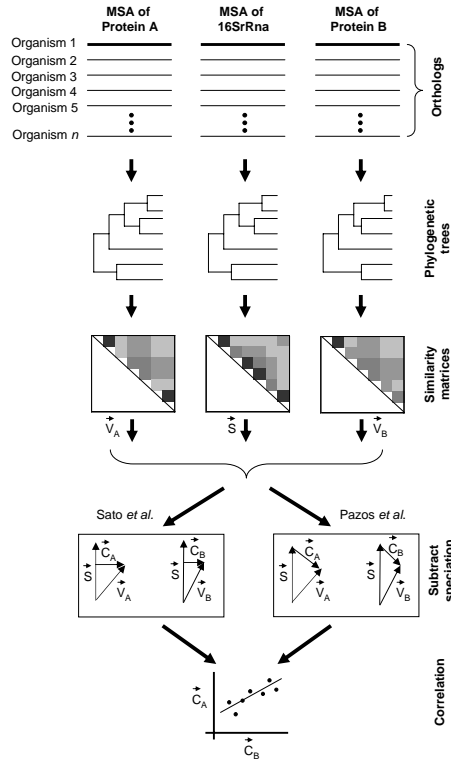
$$r_{AB} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (A_{ij} - \bar{A})(B_{ij} - \bar{B})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (A_{ij} - \bar{A})^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n (B_{ij} - \bar{B})^2}}, \quad (1.1)$$

where  $n$  is the number of organisms (number of rows or columns) in the matrices,  $A_{ij}$  and  $B_{ij}$  are the evolutionary distances between organisms  $i$  and  $j$  in the tree of proteins  $A$  and  $B$ , respectively, and  $\bar{A}$  and  $\bar{B}$  are the mean values of all  $A_{ij}$  and  $B_{ij}$ , respectively. The value of  $r_{AB}$  ranges from -1 to +1. The higher the value of  $r$ , the higher the agreement between the two matrices, and thus the higher the degree of co-evolution between  $A$  and  $B$ .

Pairs of proteins with correlation scores above a certain threshold are predicted to interact. A correlation score of 0.8 is considered to be a good threshold for predicting protein interactions [24, 48]. Pazos et al. [48] estimated that about one third of the predictions by the mirror-tree method are false positives. A false positive in this context refers to a non-interacting pair that was predicted to interact due to their high correlation score. It is quite possible that the evolutionary histories of two non-interacting proteins are highly correlated due to their common speciation history. Thus, in order to truly assess the correlation of evolutionary histories of two proteins, one should first subtract the background correlation that is due to their common speciation history. Recently, it has been observed that subtracting the underlying speciation component greatly improves the predictive power of the mirror-tree approach by reducing the number of false-positives. Refined mirror-tree methods that subtract the underlying speciation signal are discussed in the following subsection.

**1.2.4.2 Accounting for background speciation** As pointed at the end of the previous section, to improve the performance of the mirror-tree approach, the co-evolution due to common speciation events should be subtracted from the overall co-evolution signal. Recently, two approaches, very similar in techniques, have been proposed to address this problem [46, 52].

For an easier understanding of the speciation subtraction process, let us think of the distances matrices used in the mirror-tree method as vectors (i.e., the upper right triangle of the distance matrices is linearized and represented as a vector), which will be referred to as the *evolutionary vectors* hereafter. Let  $\vec{V}_A$  and  $\vec{V}_B$  denote the evolutionary vector computed for a multiple sequence alignment of orthologs of proteins  $A$  and  $B$ , respectively, for a common set of species. Let  $\vec{S}$  denote the canonical evolutionary vector, also referred to as the *speciation vector*, computed in the same way but based on a multiple sequence alignment of 16S rRNA sequences for the same set of species. Speciation vector  $\vec{S}$  approximates the interspecies evolutionary distance based on the set of species under consideration. The differences in the scale



**Fig. 1.3** Schema of the mirror-tree method with a correction for the background speciation. Correlation between the evolutionary histories of two proteins could be due to (i) a need to co-evolve in order to preserve the interaction and/or (ii) common speciation events. To estimate the co-evolution due to the common speciation, a canonical tree-of-life is constructed by aligning the 16 S rRNA sequences. The rRNA alignment is used to compute the distance matrix representing the species tree.  $\vec{V}_A$ ,  $\vec{V}_B$  and  $\vec{S}$  are the vector notations for the corresponding distance matrices. Vector  $\vec{C}_X$  is obtained from  $\vec{V}_X$  by subtracting it by the speciation component  $\vec{S}$ . The speciation component  $\vec{S}$  is calculated differently based on the method being used. The degree of co-evolution between  $A$  and  $B$  is then assessed by computing the linear correlation between  $\vec{C}_A$  with  $\vec{C}_B$ . Proteins  $A$  and  $B$  are predicted to interact if the correlation between  $\vec{C}_A$  and  $\vec{C}_B$  is sufficiently high.

of protein and RNA distance matrices are overcome by re-scaling the speciation vector values by a factor computed based on “molecular clock” proteins [46]. Sato *et al.* considered also alternative method for contraction such speciation vector [52].

A pictorial illustration of the background speciation subtraction procedure is shown in Figure 1.3. The main idea is to decompose evolutionary vectors  $\vec{V}_A$  and  $\vec{V}_B$  into two components: one representing the contribution due to speciation, and the other representing the contribution due to evolutionary pressure related to

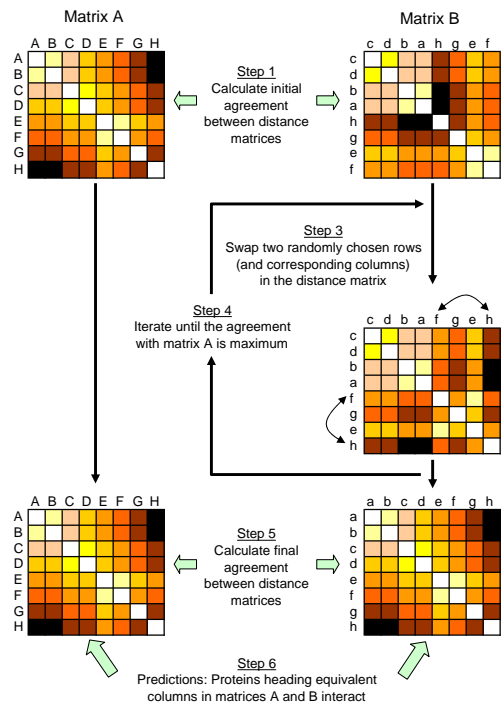
preserving the protein function (denoted by  $\vec{C}_A$  and  $\vec{C}_B$ , respectively). To obtain  $\vec{C}_A$  and  $\vec{C}_B$ , the speciation component  $\vec{S}$  is subtracted from  $\vec{V}_A$  and  $\vec{V}_B$ , respectively. Vectors  $\vec{C}_A$  and  $\vec{C}_B$  are expected to contain only the distances between orthologs that are not due to speciation but to other reasons related to function [46]. The degree of co-evolution between  $A$  and  $B$  is then measured by computing the correlation between  $\vec{C}_A$  and  $\vec{C}_B$  rather than between  $\vec{V}_A$  and  $\vec{V}_B$  as in the basic mirror-tree approach.

The two speciation subtraction methods, due to Pazos et al. [46] and Sato et al. [52], differ in how speciation subtraction is performed (see Figure 1.3). An in-depth analysis of the pros and cons of two methods are provided in [33]. In a nut shell, Sato et al. attribute all changes in the direction of the speciation vector to the speciation process, and thus assume that vector  $\vec{C}_A$  is perpendicular to the speciation vector  $\vec{S}$ , whereas Pazos et al. assume that the speciation component in  $\vec{V}_A$  is constant and independent on the protein family. As a result, Pazos et al. compute  $\vec{C}_A$  to be the difference between  $\vec{V}_A$  and  $\vec{S}$ , which explains the need to re-scale RNA distances to protein distances in the vector  $\vec{S}$ . Interestingly, despite this difference, both speciation correction methods produce similar result [33]. In particular, Pazos et al. report that the speciation subtraction step reduces the number of false positives by about 8.5%.

The abovementioned methods of subtracting of background speciation discussed have been recently complemented by the work of Kann *et al.* who starting from the assumption is that in conserved regions, functional co-evolution is less concealed by speciation divergence, demonstrated that the performance of the mirrortree method can be further improved by restricting the co-evolution analysis to the relatively conserved regions in the protein sequence [33].

**1.2.4.3 Predicting protein interaction specificity** In this section, we address the problem of predicting interaction partners between members of two proteins families that are known to interact [50, 20, 31]. Given two families of proteins, which are known to interact, the objective is to establish a mapping between the members of one family with the members of the other family.

To better understand the protein interaction specificity (PRINS) problem, let us consider an analogous problem, which we shall refer to as the *matching* problem. Imagine a social gathering, which is attended by  $n$  married couples. Let  $H = \{h_1, h_2, \dots, h_n\}$  and  $W = \{w_1, w_2, \dots, w_n\}$  be the sets of husbands and wives attending the gathering. Given that husbands in set  $H$  are married to the wives in set  $W$ , and that the marital relationship is monogamous, the matching problem asks for a one-to-one mapping of the members in  $H$  to those in  $W$  such that each mapping  $(h_i, w_j)$  holds the meaning “ $h_i$  is married to  $w_j$ ”. In other words, the objective is to pair husbands and wives such that all  $n$  pairings are correct. The matching problem has a total of  $n!$  possible mappings, out of which only one is correct. The matching problem becomes much more complex if one were to remove the constraint which requires that the marital relationship is monogamous. Such a relaxation would allow the sizes of sets  $H$  and  $W$  to be different. Without knowing the number of wives (or husbands) each husband (wife, respectively) has, the problem becomes intractable.



**Fig. 1.4** Schema of the column-swapping algorithm. Image reproduced from [50] with permission.

The PRINS problem is essentially the same as the matching problem with the two sets containing proteins instead of husbands and wives. Let  $A$  and  $B$  be the two sets of proteins. Given that the proteins in  $A$  interact with those in  $B$ , the objective is to map proteins in  $A$  to their interaction partners in  $B$ . To fully appreciate the complexity of this problem, let us first consider a simpler version of the problem, which assumes that the number of proteins in  $A$  is the same as that in  $B$ , and the interaction between the members of  $A$  and  $B$  is one-to-one.

Protein interaction specificity (a protein binding to a specific partner) is vital to cell function. In order to maintain the interaction specificity, it is required that it persist through the course of strong evolutionary events such as gene duplication and gene divergence. As genes are duplicated, the binding specificities of duplicated genes (paralogs) often diverge, resulting in new binding specificities. Existence of numerous paralogs for both interaction partners can make the problem of predicting interaction specificity difficult as the number of potential interactions grow combinatorially [50].

Discovering interaction specificity between two interacting families of proteins, such as matching ligands to specific receptors, is an important problem in molecular biology that is largely unsolved. A naive approach to solve this problem would be

to try out all possible mappings (assuming that there is an oracle to verify whether a given mapping is correct). If  $A$  and  $B$  contain  $n$  proteins each, then there are a total of  $n!$  possible mappings between matrices  $A$  and  $B$ . For a fairly large  $n$ , it is computationally unrealistic to try out all possible mappings.

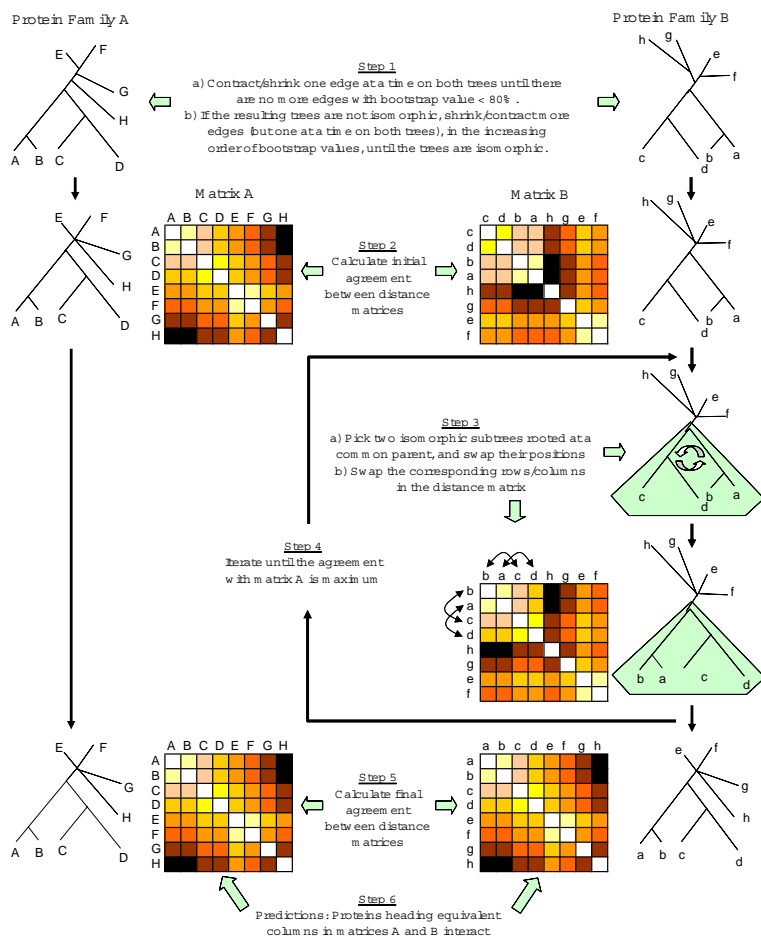
Under the assumption that interacting proteins undergo co-evolution, Ramani and Marcotte [50] and Gertz et al. [20], in independent and parallel works, proposed the “column-swapping” method for the PRINS problem. A schematic illustration of the column-swapping approach is shown in Figure 1.4. Matrices  $A$  and  $B$  in Figure 1.4 correspond to distance matrices of families  $A$  and  $B$ , respectively. In this approach, a Monte Carlo algorithm [37] with simulated annealing is used to navigate through the search space in an effort to maximize the correlation between the two matrices. The Monte Carlo search process, instead of searching through the entire landscape of all possible mappings, allows for a random sampling of the search space in an effort to find the optimal mapping. Each iteration of the Monte Carlo search process, referred to as a “move”, constitutes the following two steps.

1. Chose two columns uniformly at random, and swap their positions (the corresponding rows are also swapped)
2. If, after the swap, the correlation between the two matrices has improved, the swap is kept. Else, the swap is kept with the probability  $p = \exp(-\delta/T)$ , where  $\delta$  is the decrease in the correlation due to the swap, and  $T$  is the temperature control variable governing the simulation process.

Initially,  $T$  is set to a value such that  $p = 0.8$  to begin with, and after each iteration the value of  $T$  is decreased by 5%. After the search process converges to a particular mapping, proteins heading equivalent columns in the two matrices are predicted to interact. As with any local search algorithm, it is difficult to say whether the final mapping is an optimal mapping or a local optima.

The main downside of the column-swapping algorithm is the size of search space ( $n!$ ), which it has to navigate in order to find the optimal mapping. Since the size of the search space is directly proportional to search (computational) time, column-swapping algorithm becomes impractical even for families of size 30.

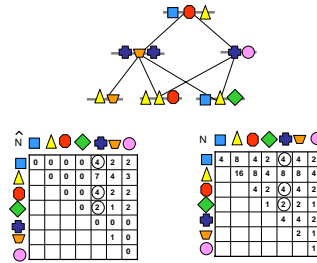
In 2005, Jothi et al. [31] introduced a new algorithm, called MORPH, to solve the PRINS problem. The main motivation behind MORPH is to reduce the search space of the column-swapping algorithm. In addition to using the evolutionary distance information, MORPH uses topological information encoded in the evolutionary trees of the protein families. A schematic illustration of the MORPH algorithm is shown in Figure 1.5. While MORPH is similar to the column-swapping algorithm at the top-level, the major (and important) difference is the use of phylogenetic tree topology to guide the search process. Each move in the column-swapping algorithm involves swapping two random columns (and the corresponding rows), whereas each



**Fig. 1.5** Schema of the MORPH algorithm. Image reprinted from [31] with permission.

move in MORPH involves swapping two isomorphic<sup>1</sup> subtrees rooted at a common node (and the corresponding sets of rows and columns in the distance matrix).

<sup>1</sup>Two trees  $T_1$  and  $T_2$  are isomorphic if there is a one-to-one mapping between their vertices (nodes) such that there is an edge between two vertices in  $T_1$  if and only if there is an edge between the two corresponding vertices in  $T_2$ .



**Fig. 1.6** Three sets of topologically identical (isomorphic) trees. Number of topology preserving mappings of one tree onto another is (a) 8, (b) 8, and (c) 24. Despite the same number of leaves in (a) and (c), the number of possible mappings are different. This is due to the increased complexity of the tree topology in (a) when compared to that in (c). Image reproduced from [31] with permission.

Under the assumption that the phylogenetic trees of protein families  $A$  and  $B$  are topologically identical, MORPH essentially performs a topology preserving embedding (superimposition) of one tree onto the other. The complexity of the topology of the trees play a key role on the number of possible ways that one could superimpose one tree onto another. Figure 1.6 shows three sets of trees, each of which has different number of possible mappings based on the tree complexity. For the set of trees in Figure 1.6a, the search space (number of mappings) for the column-swapping algorithm is  $4! = 24$ , whereas it is only eight for MORPH.

In order to apply MORPH, the phylogenetic trees corresponding to the two families of proteins must be isomorphic. To ensure that the trees are isomorphic, MORPH starts by contracting/shrinking those internal tree edges, in both trees, with bootstrap score less than a certain threshold. It is made sure that equal number of edges are contracted on both trees. If, after the initial edge contraction procedure, the two trees are not isomorphic, additional internal edges are contracted on both trees (in increasing order of the edge bootstrap scores) until the trees are isomorphic. The benefits of edge contraction procedure is two-fold: (i) ensure that the two trees are isomorphic to begin with, and (ii) decrease the chances of less reliable edges (with low bootstrap scores) wrongly influencing the algorithm. Since MORPH relies heavily on the topology of the trees, it is essential that the tree edges are trustworthy. In the worst case, contracting all the internal edges on both trees will leave two star-topology trees (like those in Figure 1.6c), in which case the number of possible mappings considered by MORPH will be the same as that considered by the column-swapping algorithm. Thus, in the worst-case MORPH's search space will be as big as that of the column-swapping algorithm.

After the edge contraction procedure, a Monte Carlo search process similar to that used in the column-swapping algorithm is used to find the best possible superimpo-

sition of the two trees. Like in the column-swapping algorithm, the distance matrix and the tree corresponding to one of the two families is fixed, and transformations are made to the tree and the matrix corresponding to the second family. Each iteration of the Monte Carlo search process constitutes the following two steps.

1. Chose two isomorphic subtrees, rooted at a common node, uniformly at random, and swap their positions (and the corresponding sets of rows/columns)
2. If, after the swap, the correlation between the two matrices has improved, the swap is kept. Else, the swap is kept with the probability  $p = \exp(-\delta/T)$ .

Parameters  $\delta$  and  $T$  are the same as those in the column-swapping algorithm. After the search process converges to a certain mapping, proteins heading equivalent columns in the two matrices are predicted to interact.

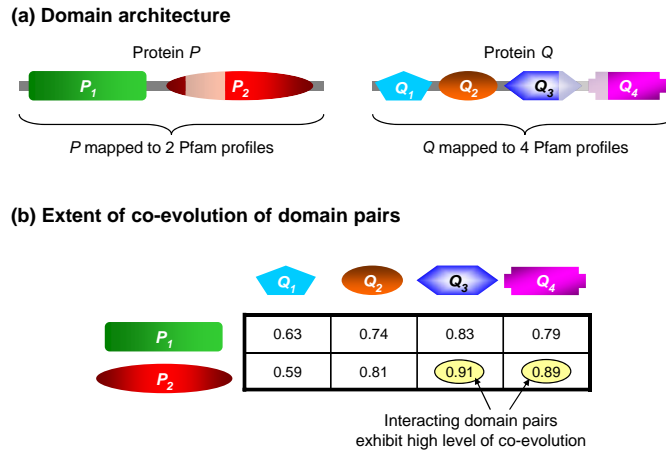
The sophisticated search process used in MORPH reduces the search space by multiple orders of magnitude in comparison to the column-swapping algorithm. As a result, MORPH can help solve larger instances of the PRINS problem. For more details on the column-swapping algorithm and MORPH, we refer the reader to [50, 20] and [31], respectively.

### 1.3 DOMAIN-DOMAIN INTERACTIONS

Recent advances in molecular biology combined with large-scale high-throughput experiments have generated huge volumes of protein interaction data. The knowledge gained from protein interaction networks has definitely helped to gain a better understanding of protein functionalities and inner-workings of the cell. However, protein interaction networks by themselves do not provide insights on interaction specificity at the domain level. Most of the proteins are composed of multiple domains. It has been estimated that about two thirds of proteins in prokaryotes and about four fifths of proteins in eukaryotes are multidomain proteins[5, 10]. Most often, interaction between two proteins involves binding of a pair(s) of domains. Thus, understanding interaction at the domain level is a critical step towards a thorough understanding of the protein-protein interaction networks and their evolution. In this section, we will discuss computational approaches for predicting protein domain interactions. We restrict our discussion to sequence- and network-based approaches.

#### 1.3.1 Relative co-evolution of domain pairs approach

Given a protein-protein interaction, predicting the domain pair(s) that is most likely mediating the interaction is of great interest. Formally, let protein  $P$  contain domains  $\{P_1, P_2, \dots, P_m\}$  and protein  $Q$  contain domains  $\{Q_1, Q_2, \dots, Q_n\}$ . Given that  $P$  and  $Q$  interact, the objective is to find the domain pair  $P_i Q_j$  that is most likely to mediate the interaction between  $P$  and  $Q$ . Recall that under the co-evolution hypothesis, interacting proteins exhibit higher level of co-evolution. Based on this hypothesis, it is only natural and logical to assume that interacting domain pairs

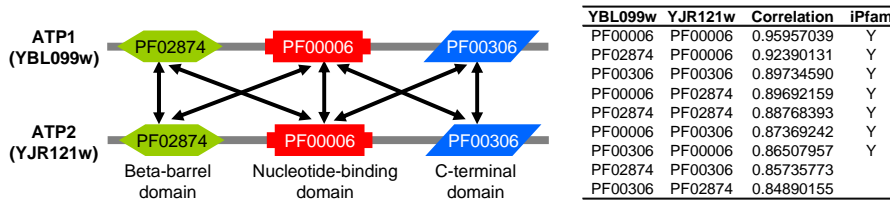


**Fig. 1.7** Relative Co-evolution of domain pairs in interacting proteins. (a) Domain assignments for interacting proteins  $P$  and  $Q$ . Interaction sites in  $P$  and  $Q$  are indicated by light color bands. (b) Correlation scores for all possible domain pairs between interacting proteins  $P$  and  $Q$  are computed using the mirror-tree method. The domain pair with the highest correlation score is predicted to be the one that is most likely to mediate the interaction between proteins  $P$  and  $Q$ .

for given protein-protein interaction exhibit higher degree of co-evolution than the non-interacting domain pairs. Jothi et al. [30] showed that this is indeed the case, and, based on this, proposed the *relative co-evolution of domain pairs* (RCDP) method to predict domain pair(s) that is most likely mediating a given protein-protein interaction.

Predicting domain interactions using RCDP involves two major steps: (i) make domain assignment to proteins, and (ii) use mirror-tree approach to assess the degree of co-evolution of all possible domain pairs. A schematic illustration of the RCDP method is shown in Figure 1.7. Interacting proteins,  $P$  and  $Q$ , are first assigned with domains (HMM profiles) using HMMer [1], RPS-BLAST [2], or other similar tools. Next, MSAs for the two proteins are constructed using orthologous proteins from a common set of organisms (as described in Section 1.2.4.1). The MSA of domain  $P_i$  in protein  $P$  is constructed by extracting those regions in  $P$ 's alignment that correspond to domain  $P_i$ . Then, using the mirror-tree method, the correlation (similarity) scores of all possible domain pairs between the two proteins are computed. Finally, the domain pair  $P_iQ_j$  with the highest correlation score (or domain pairs, in case of a tie for the highest correlation score), exhibiting the highest degree of co-evolution, is inferred to be one that is most likely to mediate the interaction between proteins  $P$  and  $Q$ .

Figure 1.8 shows the domain-level interactions between alpha (YBL099w) and beta (YJR121w) chains of F1-ATPase in *Saccharomyces cerevisiae*. RCDP will correctly predict the top-scoring domain pair (PF00006 in YBL099w and PF00006



**Fig. 1.8** Protein-protein interaction between alpha (ATP1) and beta (ATP2) chains of F1-ATPase in *Saccharomyces cerevisiae*. Protein sequences YBL099w and YJR121w (encoded by genes ATP1 and ATP2, respectively) is annotated with three Pfam [17] domains each: beta-barrel domain (PF02874), nucleotide-binding domain (PF00006), and C-terminal domain (PF00306). The correlation scores of all possible domain pairs between the two proteins are listed (table on the right) in decreasing order. Interchain domain-domain interactions that are known to be true from PDB [8] crystal structures (as inferred in iPfam [16]) are shown using double-arrows in the diagram, and 'Y' in the table. Interacting domain pairs between the two proteins have higher correlation than the non-interacting domain pairs. RCDP will correctly predict the top-scoring domain pair to be interacting.

in YJR121w) to be interacting. In this case, there are more than one domain pair mediating a given protein-protein interaction. Since RCDP is designed to find only the domain pair(s) that exhibits highest degree of co-evolution, it may not be able to identify all the domain level interactions between the two interacting proteins. It is possible that the highest-scoring domain pair may not necessarily be an interacting domain pair. This could be due to what Jothi et al. refer to as the “uncorrelated set of correlated mutations” phenomena, which may disrupt co-evolution of proteins/domains. Since the underlying similarity of phylogenetic trees approach solely relies on co-evolution principle, such disruptions can cause false predictions. RCDP’s prediction accuracy was estimated to be about 64%. A naive random method, which picks an arbitrary domain pair out of all possible domain pairs between the two interacting proteins, is expected to have a prediction accuracy of 55% [30, 43]. RCDP’s prediction accuracy of 64% is significant considering the fact that Nye et al. [43] showed, using a different dataset, that the naive random method performs as well as Sprinzak and Margalit’s association method, Deng et al.’s maximum likelihood estimation approach [13], and their own lowest p-value method, all of which are discussed in the following section. For a detailed analysis of RCDP and its limitations, we refer the reader to [30].

### 1.3.2 Predicting domain interactions from protein-protein interaction network

In this section we describe computational methods to predict interacting domain pairs from an underlying protein-protein interaction network. All interactions in a protein-protein interaction network are assumed to be physical interactions determined through experiments. To begin with, all proteins in the protein-protein interaction

network are first assigned with domains using HMM profiles. Recall that interaction between two proteins is essentially a set of interactions between the domains in the two proteins. A protein-protein interaction is mediated by one or more domain-domain interactions.

We start by introducing notation that will be used in this section. Let  $\{P_1, \dots, P_N\}$  be the set of proteins in the protein-protein interaction network and  $\{D_1, \dots, D_M\}$  be the set of all domains that are present in these interacting proteins. Let  $\mathcal{I} = \{(P_{mn}) | m, n = 1 \dots N\}$  be the set of protein pairs observed experimentally to interact. We say that the domain pair  $D_{ij}$  belongs to protein pair  $P_{mn}$  (denoted by  $D_{ij} \in P_{mn}$ ) if  $D_i$  belongs to  $P_m$  and  $D_j$  belongs to  $P_n$ , or vice-versa. Throughout this section we will assume that all domain pairs and protein pairs are unordered, i.e.,  $X_{ab}$  is the same as  $X_{ba}$ . Let  $N_{ij}$  denote the number of occurrences of domain pair  $D_{ij}$  in all possible protein pairs, and let  $\hat{N}_{ij}$  be the number of occurrences of  $D_{ij}$  only in interacting protein pairs.<sup>2</sup>

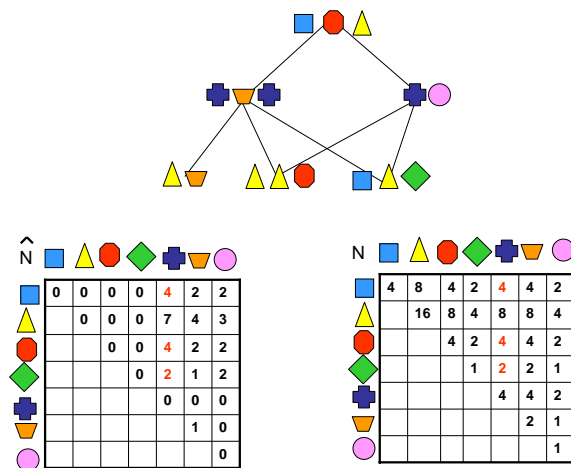
**1.3.2.1 Association Method** Sprinzak and Margalit [54] made the first attempt to predict domain-domain interactions from a protein-protein interaction network. They proposed a simple statistical approach, referred to as the *Association Method* (AM), to identify those domain pairs that are observed to occur in interacting protein pairs more frequently than expected by chance. Statistical significance of the observed domain pair is usually measured by the standard log-odds value  $A$  or probability  $\alpha$ , given by

$$A_{ij} = \log_2 \frac{\hat{N}_{ij}}{N_{ij} - \hat{N}_{ij}}; \quad \alpha_{ij} = \frac{\hat{N}_{ij}}{N_{ij}}. \quad (1.2)$$

The AM method is illustrated using a toy protein-protein interaction network in Figure 1.9. It was shown that among high scoring pairs are pairs of domains that are known to interact, and a high  $\alpha$  value can be used as a predictor of domain-domain interaction.

**1.3.2.2 Maximum likelihood estimation approach** Following the work of Sprinzak and Margalit, several related methods have been proposed [41]. In particular, Deng et al. [13] extended the idea behind the association method and proposed a maximum likelihood approach to estimate the probability of domain-domain interactions. Their expectation maximization algorithm (EM) computes domain interaction probabilities that maximize the expectation of observing a given protein-protein interaction network  $\mathcal{N}et$ . An important feature of this approach is that it allows for

<sup>2</sup>Not all the methods described in this section use unordered pairings. Some of them use ordered pairings, i.e.,  $X_{ab}$  is not the same as  $X_{ba}$ . Depending on whether one uses ordered or unordered pairing, the number of occurrences of a domain pair in a given protein pair is different. For example, let protein  $P_m$  contain domains  $D_x$  and  $D_y$ , and let protein  $P_n$  contain domains  $D_x$ ,  $D_y$ , and  $D_z$ . The number of occurrences of domain pair  $D_{xy}$  in protein pair  $P_{mn}$  is 4 if ordered pairing is used, and 2 if unordered pairing is used.



**Fig. 1.9** Schematic illustration of the association method. The toy protein-protein interaction network is given in the upper panel. The domain composition of each protein in the network is color coded. Lower panels show domain pair occurrence tables  $\hat{N}$  and  $N$ . Each entry  $\hat{N}_{i,j}$  represents the number of times the domain pair  $(i, j)$  occurs in interacting protein pairs, and each entry  $N_{i,j}$  represents the number of times  $(i, j)$  occurs all possible protein pairs. Three domain pairs with maximum scores are encircled.

explicit treatment of missing and incorrect information (in this case, false negatives and false positives in the protein-protein interaction network).

In the EM method, protein-protein interactions and domain-domain interactions are treated as random variables denoted by  $P_{mn}$  and  $D_{ij}$ , respectively. In particular, we let  $P_{mn} = 1$  if proteins  $P_m$  and  $P_n$  interact with each other, and  $P_{mn} = 0$  otherwise. Similarly,  $D_{ij} = 1$  if domains  $D_i$  and  $D_j$  interact with each other, and  $D_{ij} = 0$  otherwise. The probability that domains  $D_i$  and  $D_j$  interact is denoted by  $\mathcal{P}r(D_{ij}) = \mathcal{P}r(D_{ij} = 1)$ . The probability that proteins  $P_m$  and  $P_n$  interact is given by

$$\mathcal{P}r(P_{mn} = 1) = 1 - \prod_{D_{ij} \in P_{mn}} (1 - \mathcal{P}r(D_{ij})). \quad (1.3)$$

Random variable  $\mathcal{O}_{mn}$  is used to describe the experimental observation of protein-protein interaction network. Here  $\mathcal{O}_{mn} = 1$  if proteins  $P_m$  and  $P_n$  were observed to interact (that is  $P_{mn} \in \mathcal{I}$ ), and  $\mathcal{O}_{mn} = 0$  otherwise. False negative rate is given by  $f_n = \mathcal{P}r(\mathcal{O}_{mn} = 0 \mid P_{mn} = 1)$  and false positive rate is given by  $f_p = \mathcal{P}r(\mathcal{O}_{mn} = 1 \mid P_{mn} = 0)$ . Estimations of false positive rate and false negative

rate vary significantly from paper to paper. Deng et al. estimated  $f_n$  and  $f_p$  to be 0.8 and  $2.5E - 4$ , respectively.

Recall that the goal is to estimate  $\mathcal{P}r(D_{ij}), \forall_{ij}$  such that the probability of the observed network  $\mathcal{N}et$  is maximum. The probability of observing  $\mathcal{N}et$  is given by

$$\mathcal{P}r(\mathcal{N}et) = \prod_{(m,n)|\mathcal{O}_{mn}=1} \mathcal{P}r(\mathcal{O}_{mn}=1) \prod_{(m,n)|\mathcal{O}_{mn}=0} \mathcal{P}r(\mathcal{O}_{mn}=0), \quad (1.4)$$

where

$$\begin{aligned} \mathcal{P}r(\mathcal{O}_{mn}=1) &= \mathcal{P}r(P_{mn}=1)(1-f_n) + (1-\mathcal{P}r(P_{mn}=1))f_n \quad (1.5) \\ \mathcal{P}r(\mathcal{O}_{mn}=0) &= 1 - \mathcal{P}r(\mathcal{O}_{mn}=1). \end{aligned} \quad (1.6)$$

The estimates of  $\mathcal{P}r(D_{ij})$  are computed iteratively in an effort to maximize  $\mathcal{P}r(\mathcal{N}et)$ . Let  $\mathcal{P}r(D_{ij}^t)$  be the estimation of  $\mathcal{P}r(D_{ij})$  in the  $t$ -th iteration and let  $D^t$  denote the vector of  $\mathcal{P}r(D_{ij}^t), \forall_{ij}$  estimated in the  $t$ -th iteration. Initially, values in  $D^0$  can all be set the same, or those estimations obtained using the AM method. Note that each estimation of  $D^{t-1}$  defines  $\mathcal{P}r(P_{mn}=1)$  and  $\mathcal{P}r(\mathcal{O}_{mn}=1)$  using equations 1.3 and 1.4. These values are, in turn, used to compute  $D^t$  in the current iteration as follows. First, for each domain pair  $D_{ij}$  and each protein pair  $P_{mn}$  the expectation that domain pair  $D_{ij}$  physically interact in protein pair  $P_{mn}$  is estimated as:

$$E(D_{ij} \in P_{mn}) = \begin{cases} \frac{\mathcal{P}r(D_{ij}^{t-1})(1-f_n)}{\mathcal{P}r(\mathcal{O}_{mn}=1)} & \text{if } (P_m, P_n) \in \mathcal{I} \\ \frac{\mathcal{P}r(D_{ij}^{t-1})f_n}{\mathcal{P}r(\mathcal{O}_{mn}=0)} & \text{otherwise.} \end{cases} \quad (1.7)$$

The vales of  $\mathcal{P}r(D_{ij}^t)$ , for the next iteration are then computed as

$$\mathcal{P}r(D_{ij}^t) = \frac{1}{N_{ij}} \sum_{(m,n)|D_{ij} \in P_{mn}} E(D_{ij} \in P_{mn}). \quad (1.8)$$

Thus, similar to the AM method, the MLE method provides a scoring scheme that measures the likelihood of a given domain pair interacting.

Since our knowledge of interacting domain pairs is limited (only a small fraction of interacting domains pairs have been inferred from crystal structures), it is not clear as to how two methods predicting domain interactions can be compared. Deng et al. [13] compared the performance of their EM method to that of Sprinzak and Margalit's AM method [54] by assessing how well the domain-domain interaction predictions by the two methods can in turn be used to predict protein-protein interactions. For the AM method,  $\mathcal{P}r(D_{ij})$  in equation 1.3 is replaced by  $\alpha_{ij}$ . Thus, rather than performing a direct comparison of predicted interacting domain pairs, they tested which method leads to a more accurate prediction of protein-protein interactions. It was shown that the EM method outperforms the AM method significantly [13]. This is not surprising considering the fact that the values of  $\mathcal{P}r(D_{ij})$  in the EM method are computed

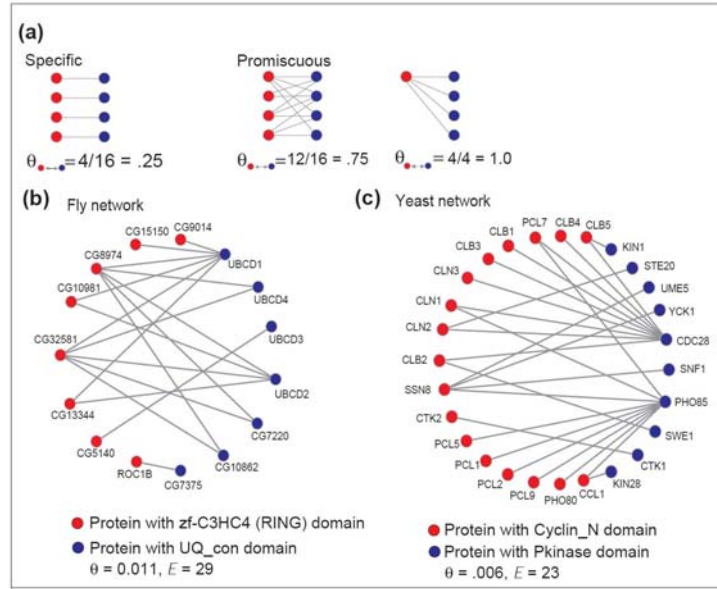
so as to maximize the probability of observed interactions. Comparison of domain interaction prediction methods on the base of how well they predict protein-protein interaction is, however, not very satisfying. Correct prediction of protein interaction does not imply that the interaction domains have been correctly identified. This problem has been recognized by several researchers and we describe other testing techniques in subsequent sections.

**1.3.2.3 Domain Pair Exclusion Analysis (DPEA).** An important problem in inferring domain interactions from protein interaction data using the AM and the EM methods is that is that highest scoring domain interactions tend to be non-specific. The difference between specific and non-specific interactions is illustrated in figure 1.10. Each of the interacting domains can have several paralogs within a given organism - several instances of the same domain. In a highly specific (non-promiscuous) interaction, each such instance of domain  $D_i$  interacts with a unique instance of domain  $D_j$  (see figure 1.10 a). Such specific interactions are likely to receive a low score by methods that detect domain interactions by measuring the probability of interaction of corresponding domains, for example, the AM and the EM methods discussed above. To deal with this problem, Riley et al. [51] introduced a new method called *domain pair exclusion analysis* (DPEA). The idea of the methods is to measure, for each domain pair, how disallowing the given domain-domain interaction reduces the likelihood of the protein-protein interaction network. This is assessed by comparing the results of executing an expectation maximization protocol under the assumption that all pairs of domains can interact and that a given pair of domains cannot interact. The E-value is defined to be the ratio of the corresponding likelihood estimators. For real world examples of very low  $\theta$  score and high E-value see figure 1.10b-c.

The expectation maximization protocol used in the DPEA is similar to the one for the MLE method described above but performed under the assumption that the network is reliable (no false positive or false negatives) and including protein interaction data from multiple organisms.

The DPEA method has been compared to the MLE and the AM methods by the level of retrieval of pairs that are known to interact based on crystal structure evidence recorded in the database of interacting domain pairs, iPFAM [16]. Indeed, the DPEA method outperforms the AM and the EM methods by a significant margin in the number of recovered domain-domain interactions confirmed by crystal structure evidence [51].

**1.3.2.4 Lowest-p-value method** A different, statistical approach, to predict domain-domain interaction was proposed by Nye et al. [43]. The idea their approach is to test, domain pair  $(D_i, D_j)$ , test the null hypothesis  $\mathcal{H}_{ij}$  that presence of the domain pair  $(D_i, D_j)$  in a protein pair  $(P_n, P_m)$  does not affect whether the two proteins interact. They also consider the global null hypothesis  $\mathcal{H}_\infty$  that interaction is entirely unrelated to the domain architectures of proteins. There are two specific assumptions that present in this method that are not made in other approaches. First, each protein



**Fig. 1.10** (a) The difference between promiscuous and specific interactions; (b-c) Examples of two domain-domain interactions scored highly by the E-value method (score E) but missed by the association method (score  $\alpha$ ). Image reprinted from [51] with permission.

interaction is assumed to be mediated by exactly one domain-domain interaction. Second, each occurrence of a domain in a protein sequence is counted separately.

To test the hypothesis, consider first the following two-by-two table:

	$  D_i, D_j  $	remaining domain pairs
interacting domain pairs	$x_{11}^{ij}$	$x_{12}^{ij}$
non-interacting domain pairs belonging to interacting protein pairs	$x_{21}^{ij}$	$x_{22}^{ij}$

The log odds score  $s_{ij}$  is defined as:

$$s_{ij} = \log \frac{x_{11}/x_{21}}{x_{12}/x_{22}} \tag{1.9}$$

Thus large score  $s_{ij}$  signifies that the domain pair  $(D_i, D_j)$  is expected to have larger number of interactions than other domain pairs. Before we show how the values of

the table are computed, we explain the score  $s_{ij}$  is converted into  $p$ -value.  $p$ -value measures the probability that hypothesis  $\mathcal{H}_{ij}$  is true. This is done by estimating how likely score at least this high can be obtained by chance ( under hypothesis  $\mathcal{H}_{\infty}$ ). To compute  $p$ -value, the domain composition within protein is randomized. During the randomization procedure the degree of each node in the protein-protein interaction network remains the same. The discussion of details of the randomization technique exceeds the scope of this chapter and we refer the reader to the original paper [43].

It remains to show how estimate the values in the table. Values  $x_{11}^{ij}$  are computed as the expected number of times domain pair  $(D_i, D_j)$  mediates a protein-protein interaction, under the null hypothesis  $\mathcal{H}_{\infty}$  given the experimental data  $\mathcal{O}$ :

$$E(D_{ij}) = \sum_{P_n, P_m} \mathcal{P}r(D_{ij}(m, n) = 1 | \mathcal{O}) \quad (1.10)$$

where, following the notation from the previous subsection,  $\mathcal{P}r(D_{ij}(m, n) = 1)$  denotes the probability that domain pair  $(D_i, D_j)$  interact in protein pair  $(P_m, P_n)$  Developing the right side of the equation we obtain:

$$E(D_{ij}) = \sum_{P_n, P_m} \mathcal{P}r(P_{mn} = 1 | \mathcal{O}) \mathcal{P}r(D_{ij} = 1 | P_{mn} = 1) \quad (1.11)$$

where  $\mathcal{P}r(P_{mn} = 1 | \mathcal{O})$  can be computed from the approximates of false positive and false negative rates in a way similar as described in the previous subsection, modifying in a natural way the computation of  $\mathcal{P}r(D_{ij} = 1 | P_{mn} = 1)$  so that it takes into account multiple occurrences of the same domain in a protein chain. Namely, let  $N_{ij}^{mn}$  be the number of possible interactions between domains  $D_i$  and  $D_j$  in protein pairs  $P_n, P_m$ .

$$\mathcal{P}r(D_{ij} = 1 | P_{mn} = 1) = \frac{N_{ij}^{mn}}{\sum_{xy} N_{xy}^{mn}} \quad (1.12)$$

Since in the case of  $p$ -value method, the multiple occurrences of domains are counted separately, the value  $\hat{N}_{ij}$ , equal to the number of times domains pair  $(D_i, D_j)$  is counted to occur in interacting protein pairs is, in this case, computed as:

$$\hat{N}_{ij} = \sum_{kt | (P_k, P_t) \in \mathcal{I}} N_{ij}^{kt}$$

Now, the values of the table are estimated naturally as follows:

$$\begin{aligned} x_{ij}^{11} &= E(D_{ij}) \\ x_{ij}^{21} &= \hat{N}_{ij} - E(D_{ij}) \\ x_{ij}^{12} &= \sum_{x, y \neq i, j} E(D_{xy}) \end{aligned}$$

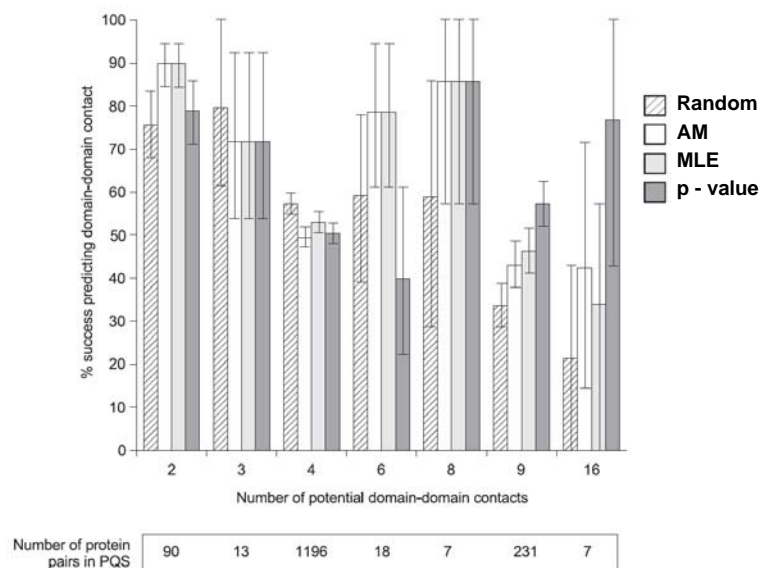
$$x_{ij}^{22} = \sum_{x,y \neq i,j} (\hat{N}_{xy} - E(D_{xy}))$$

Nye et al. [43] pioneered the method of testing correctness of domain interaction prediction method used in section 1.3.1. That is, unlike the approaches described in subsections 1.3.2.1-1.3.2.3, their goal is to predict the most likely pair of domains mediating a given protein interaction, rather than predicting new domain interactions. They predict that within the set of domain pairs belonging to a given interacting protein pair, the domain pair with the lowest p-value is likely to form a contact. To confirm this, they used protein complexes in the PQS database [26] (a data base of quaternary states for structures contained in the Brookhaven Protein Data Bank (PDB) that were determined by X-ray crystallography) restricted to protein pairs that are meaningful in this context (e.g. at least one protein must be multi-domain, both protein contain only domain present in the yeast protein-protein interaction network used in the study etc.). The results of this test for the lowest p-value method compared to random selection (Random) and two of the AM and the EM methods discussed before, are presented on figure 1.11. It is striking from this analysis that the improvement that these method achieve over a random selection is small, a although increasing with the number of possible domain pairs.

**1.3.2.5 Most Parsimonious Explanation (PE).** most parsimonious explanation method Recently, Guimaraes *et al.* introduced a new domain interaction prediction method called Most Parsimonious Explanation [?]. The method relies on the hypothesis that interactions between proteins evolved in a parsimonious way and that the set of correct domain-domain interactions is well approximated by the minimal set of domain interactions necessary to justify a given protein-protein interaction network. The EM problem is formulated as a linear programming optimization problem, where each potential domain-domain contact is a variable that can receive a value ranging between 0 and 1 (called *LP-score*), and each edge of the protein-protein interaction network corresponds to one linear constraint. That is, for each domain pair  $(D_i, D_j)$  that belongs to some interacting protein pair, there is a variable  $x_{ij}$ . The values of  $x_{ij}$  are computed using linear programming (LP):

$$\begin{aligned} \text{minimize} \quad & \sum_{D_i, D_j} x_{ij} & (1.13) \\ \text{subject to:} \quad & \sum_{(D_i, D_j) \in (P_m, P_n)} x_{ij} \geq 1, \text{ where } (P_m, P_n) \in \mathcal{I}. \end{aligned}$$

To account for the noise in the experimental data a set of the linear programming instances is constructed in a probabilistic fashion, where the probability of including an LP constraint in equation 1.13 equals the probability with which the corresponding protein-protein interaction is assumed to be correct. The results of coming from the set of these linear programs are averaged. A different randomization experiment is used to compute p-values and prevent overprediction of interactions between



**Fig. 1.11** Domain-domain contact prediction results. The results are broken down according to the potential number of domain-domain contacts available between protein pairs in the PQS database, and the number of protein pairs within each such category is shown at the bottom of the figure. The proportion of protein pairs for which four different prediction methods correctly predict a domain-domain contact is shown in the main graph. It is often observed in the PQS that several different domain pairs are in contact within each interacting protein pair. Any potential contact picked at random therefore has some probability of being confirmed as a contact in the PQS, and this baseline success rate is shown by the hatched bars. The error bars for the non-random methods correspond to a 90% confidence interval based on a binomial distribution assumption. Image reprinted from [43] with permission.

frequently occurring domain pairs. Guimaraes *et al.* demonstrated that the PE method outperforms the EM and RDCP method significantly [?].

## GLOSSARY

**Co-evolution** Coordinated evolution. It is generally agreed that proteins that interact with each other or have similar function undergo coordinated evolution.

**Gene fusion** A pair of genes in one genome is fused together into a single gene in another genome.

**HMMer** HMMer is a freely distributable implementation of profile HMM (hidden markov model) software for protein sequence analysis. It uses profile HMMs to do sensitive database searching using statistical descriptions of a sequence family's consensus.

**iPfam** iPfam is a resource that describes domain-domain interactions that are observed in PDB crystal structures.

**Ortholog** Two genes from two different species are said to be orthologs if they evolved directly from a single gene in the last common ancestor.

**PDB** The Protein Data Bank (PDB) is a central repository for 3-D structural data of proteins and nucleic acids. This data, typically obtained by X-ray crystallography or NMR spectroscopy, is submitted by biologists and biochemists from around the world, is released into the public domain, and can be accessed for free.

**Pfam** Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

**Phylogenetic profile** A phylogenetic profile for a protein is a vector of 1s and 0s representing the presence or absence of that protein in a reference set organisms.

**Distance matrix** A matrix containing the evolutionary distances of organisms or proteins in a family.

### **Acknowledgments**

This work was funded by the intramural research program of the National Library of Medicine, National Institutes of Health.

## References

1. HMMer. <http://hmmer.wustl.edu>.
2. RPS-BLAST. <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>.
3. D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug. Correlation of coordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol*, 193(4):683–707, 1987.
4. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.
5. G. Apic, J. Gough, and S. A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310(2):311–25, 2001.
6. S. Atwell, M. Ultsch, A. M. De Vos, and J. A. Wells. Structural plasticity in a remodeled protein-protein interface. *Science*, 278(5340):1125–8, 1997.
7. J. M. Berger, S. J. Gamblin, S. C. Harrison, and J. C. Wang. Structure and mechanism of DNA topoisomerase II. *Nature*, 379(6562):225–32, 1996.
8. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, 2000.
9. G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433(7025):531–7, 2005.
10. C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–3, 2003.
11. T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–8, 1998.
12. S. V. Date and E. M. Marcotte. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol*, 21(9):1055–62, 2003.

13. M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10):1540–8, 2002.
14. R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, 2004.
15. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, 1999.
16. R. D. Finn, M. Marshall, and A. Bateman. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–2, 2005.
17. R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51, 2006.
18. T. Gaasterland and M. A. Ragan. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics*, 3(4):199–217, 1998.
19. A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.
20. J. Gertz, G. Elfond, A. Shustrova, M. Weisinger, M. Pellegrini, S. Cokus, and B. Rothschild. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16):2039–45, 2003.
21. L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, Jr. Finley, R. L., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *drosophila melanogaster*. *Science*, 302(5651):1727–36, 2003.
22. G. V. Glazko and A. R. Mushegian. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol*, 5(5):R32, 2004.

23. U. Gobel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–17, 1994.
24. C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–93, 2000.
25. C. S. Goh and F. E. Cohen. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol*, 324(1):177–92, 2002.
26. K. Henrick and J. M. Thornton. PQS: a protein quarternary structure file server. *Trends Biochem Sci*, 23(9):358–61, 1998.
27. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–3, 2002.
28. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–74, 2001.
29. L. jespers, H. R. Lijnen, S. Vanwetswinkel, B. Van Hoef, K. Brepoels, D. Collen, and M. De Maeyer. Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *J Mol Biol*, 290(2):471–9, 1999.
30. R. Jothi, P.F. Cherukuri, A. Tasneem, and T. M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J Mol Biol*, 2006.
31. R. Jothi, M. G. Kann, and T. M. Przytycka. Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, 21 Suppl 1:i241–i250, 2005.
32. R. Jothi, T. M. Przytycka, and L. Aravind. Discovering functional linkages and cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. Unpublished Manuscript, 2007.
33. M. G. Kann, R. Jothi, P. F. Cherukuri, and T. M. Przytycka. Predicting protein domain interactions from co-evolution of conserved regions. (to appear), 2007.

34. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, 440(7084):637–43, 2006.
35. S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan *c. elegans*. *Science*, 303(5657):540–3, 2004.
36. E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–3, 1999.
37. N. Metropolis, A. W. Rosenbluth, A. Teller, and E. J. Teller. Simulated annealing. *J Chem Phys*, 21:1087–92, 1955.
38. B. G. Mirkin, T. I. Fenner, M. Y. Galperin, and E. V. Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 3:2, 2003.
39. W. R. Moyle, R. K. Campbell, R. V. Myers, M. P. Bernard, Y. Han, and X. Wang. Co-evolution of ligand-receptor pairs. *Nature*, 368(6468):251–5, 1994.
40. E. Neher. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*, 91(1):98–102, 1994.
41. S. K. Ng, Z. Zhang, and S. H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–9, 2003.
42. C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17, 2000.
43. T. M. Nye, C. Berzuini, W. R. Gilks, M. M. Babu, and S. A. Teichmann. Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001, 2005.

44. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*, 1(2):93–108, 1999.
45. F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4):511–23, 1997.
46. F. Pazos, J. A. Ranea, D. Juan, and M. J. Sternberg. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol*, 352(4):1002–15, 2005.
47. F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–14, 2001.
48. F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–27, 2002.
49. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–8, 1999.
50. A. K. Ramani and E. M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, 327(1):273–84, 2003.
51. R. Riley, C. Lee, C. Sabatti, and D. Eisenberg. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 6(10):R89, 2005.
52. T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–9, 2005.
53. I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*, 7(3):349–58, 1994.
54. E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–92, 2001.
55. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, 1994.
56. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields,

xxx    **REFERENCES**

- and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–7, 2000.
57. A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–73, 2002.

# Index

- BLAST, ii, v
- HMMer, xiv, xxiv
- MORPH, x, xii–xiii
- MUSCLE, v
- PDB, xv, xxii, xxiv
- PRINS, viii–x, xiii
- RPS-BLAST, xiv
- UPGMA, vi
- Alignment, ii, iv, vii–viii, xiv
- Association method, xvi
- Best-hit, v
- ClustalW, v–vi
- Co-evolution, iv, vi, x, xiii–xv, xxiv
- Column-swapping algorithm, x, xii–xiii
- Distance matrix, vii, xii–xiii, xxiv
- Domain pair exclusion analysis, xix
- Domain-domain interaction, i, xiii, xvi–xvii, xxiii
- E-value, ii, v
- Embedding, xii
- Evolutionary tree, x
- Gene fusion, iii, xxiv
- IPfam, xxiv
- Interaction, i–iv, vi, viii–x, xiii–xxiv
  - domain interaction specificity, xiii
  - functional interaction, ii, iv
  - interaction network, ii, xiii, xvi–xix, xxi–xxii, xxiv
  - physical interaction, iii, xvi, xviii
  - protein interaction specificity, viii–x
- Isomorphism, xii
- Lowest p-value method, xv, xix
- Maximum likelihood estimation, xvi
- Mirror-tree, iv, vi, xiv
- Monte Carlo search, x, xiii
- Most parsimonious explanation method, xxii
- Multiple sequence alignment, iv, vii–viii
- Neighbor-joining, vi
- Ortholog, iv–vii, xiv, xxiv
- Pearson’s correlation coefficient, vi
- Pfam, xxiv
- Phylogenetic profile, ii, xxiv
- Phylogenetic tree, iv, vi, x, xii–xiii, xv
- Protein-protein interaction, i–ii, xiii–xix, xxi–xxiv
- Relative co-evolution of domain pairs approach,
  - xiii
- Superimposition, xii
- T-Coffee, v
- Topology, x, xii
- Tree, iv, vi–vii, x, xii–xiii, xv