OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

# Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners

**Benjamin A. Shoemaker, Anna R. Panchenko***

*A Tutorial in PLoS Computional Biology*

Recent advances in high-throughput experimental methods for the identification of protein interactions have resulted in a large amount of diverse data that are somewhat incomplete and contradictory. As valuable as they are, such experimental approaches studying protein interactomes have certain limitations that can be complemented by the computational methods for predicting protein interactions. In this review we describe different approaches to predict protein interaction partners as well as highlight recent achievements in the prediction of specific domains mediating protein–protein interactions. We discuss the applicability of computational methods to different types of prediction problems and point out limitations common to all of them.

## Introduction

In our companion review published in the last issue [1], we outlined the experimental techniques for the identification and characterization of protein interactions. We showed that high-throughput experimental methods produce a large amount of data which needs to be analyzed and verified. Despite this, interactomes of many organisms are far from complete. The low interaction coverage along with the experimental biases toward certain protein types and cellular localizations reported by most experimental techniques call for the development of computational methods to predict whether two proteins interact. These methods can be very useful for choosing potential targets for experimental screening or for validating experimental data (see [1]) and can provide information about interaction details (in the case of domain prediction methods) which might not be apparent from the experimental techniques. Many methods use combinations of experimental and computational techniques to different extent (for example, gene co-expression and synthetic lethality methods were covered among experimental approaches in our companion paper [1]) and do not predict physical interactions directly but rather infer the functional associations between potentially interacting proteins.

In this review, we report on several methods to predict protein or domain interaction partners. Some computational methods are based on the co-localization of potentially interacting genes in the same gene clusters or protein chains (gene cluster, gene neighborhood, and Rosetta stone methods), on co-evolution patterns in interacting proteins (sequence co-evolution methods), and on the co-expression of genes. Some methods find patterns of co-occurences in interacting proteins, protein domains, and phenotypes (phylogenetic profiles and synthetic lethality methods), while others use the presence of sequence/structural motifs characteristic only for interacting proteins (classification methods, association methods). To analyze interaction specificity at the domain level, in this second paper of the review we describe methods that are aimed at identifying specific domains mediating interactions in an interacting protein pair.

## Methods for Predicting Protein Interaction Partners

Table 1 lists different protein interaction methods, and Figure 1 illustrates their main ideas. We start the review with the genomic inference methods [2] (gene neighbor, gene cluster, Rosetta stone, and phylogenetic profile) that use genomic/protein context to infer functional associations. Gene neighbor and gene cluster methods are referred to as GN.

**Gene neighbor and gene cluster methods.** Genes with closely related functions encoding potentially interacting proteins are often transcribed as a single unit, an operon, in bacteria and are co-regulated in eukaryotes. Different methods have been developed trying to predict operons based on intergenic distances [2–6] (Figure 1A). Despite the effect of neutral evolution which tends to shuffle gene order between distantly related organisms, gene clusters or operons encoding for co-regulated genes are usually conserved; and operons found by gene neighbor methods can provide

**Editor:** Fran Lewitter, Whitehead Institute, United States of America

**Citation:** Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3(4): e43. doi:10.1371/journal.pcbi.0030043



**Abbreviations:** GN, gene neighbor and gene cluster; MLE, Maximum Likelihood Estimation; PP, phylogenetic profile; RFD, Random Forest Decision

Benjamin A. Shoemaker and Anna R. Panchenko are with the Computational Biology Branch of the National Center for Biotechnology Information in Bethesda, Maryland, United States of America.

* To whom correspondence should be addressed. E-mail: panch@ncbi.nlm.nih.gov

**Table 1.** Different Prediction Methods

| Method Name | Protein/Domain Interaction | Physical Interaction/ Functional Association |
|---|---|---|
| Gene co-expression | P | F |
| Synthetic lethality | P | F |
| Gene cluster and gene neighbor | P | F |
| Phylogenetic profile | P, D | F |
| Rosetta Stone | P | F |
| Sequence co-evolution | P, D | F |
| Classification | P, D | P |
| Integrative | P, D | P |
| Domain association | D | P |
| Bayesian networks | P, D | F, P |
| Domain pair exclusion | D | P |
| *p*-Value | D | P |

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions).
Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).
doi:10.1371/journal.pcbi.0030043.t001

additional evidence about functional linkage between their constituent genes [2,7–10]. Analysis of gene order conservation within three bacterial and archaeal genomes found that 63%–75% of co-regulated genes interact physically [7,11]. Similar results were obtained from two eukaryotes, yeast and worm [12]. Moreover, it was found that GN methods have higher coverage (about 37%) compared with other genomic inference methods [11]. An interesting example of GN involves the prediction of archael exosome by comparing gene order in archaeal and eukaryotic genomes [13]. The predicted archaeal exosomal superoperon was confirmed later by the experiment [14] and was shown to encode among other proteins two protein subunits of RNase P. This suggested a possible interaction between RNase P and the exosome in eukaryotes, a connection that was not reported earlier.

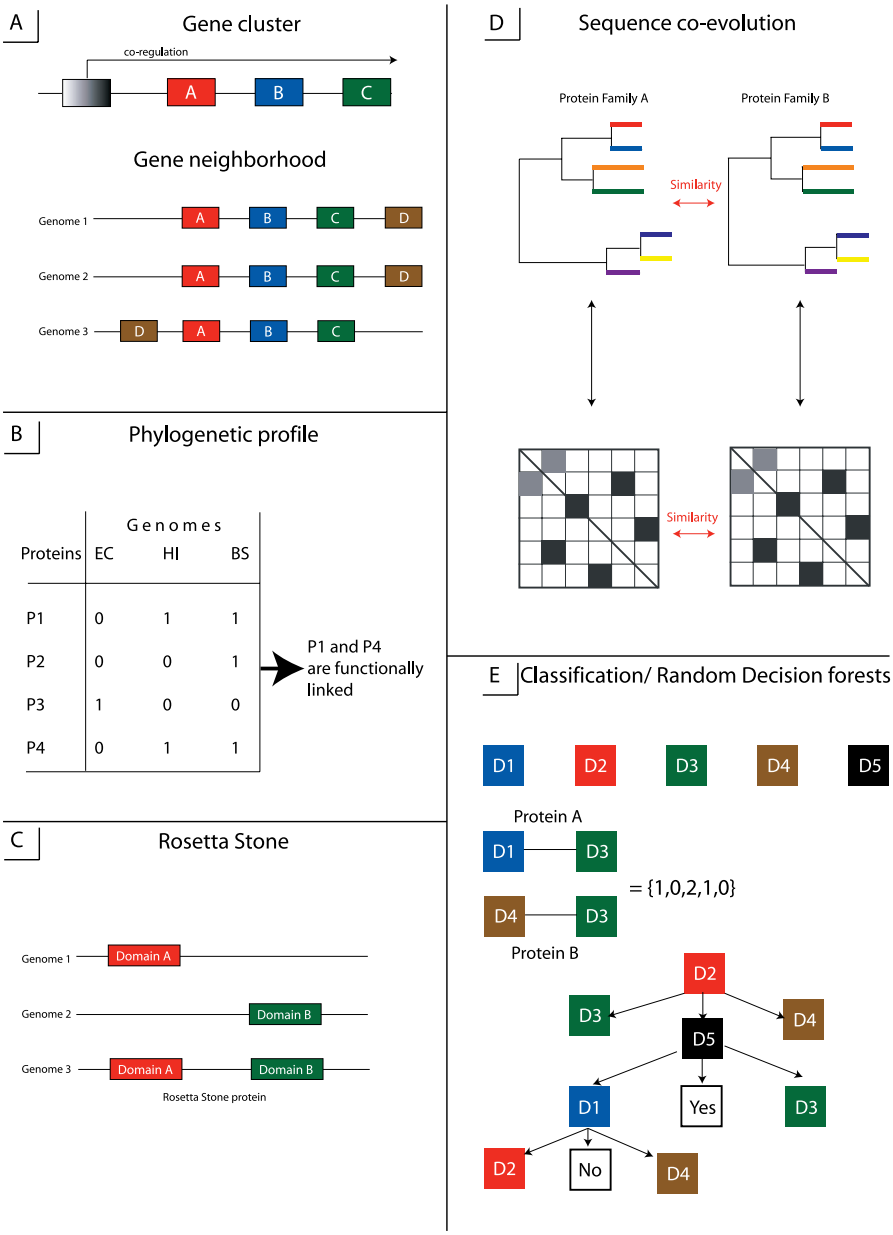
**Phylogenetic profile methods.** The phylogenetic profile (PP) method is based on the hypothesis that functionally linked and potentially interacting nonhomologous proteins co-evolve and have orthologs in the same subset of fully sequenced organisms [9,15–19]. Indeed, components of complexes and pathways should be present simultaneously in order to perform their functions. A phylogenetic profile is constructed for each protein, as a vector of $N$ elements, where $N$ is the number of genomes (Figure 1B). The presence/absence of a given protein in a given genome is indicated as "1" or "0" at each position of a profile. Proteins or their profiles can then be clustered using a bit-distance measure, and those proteins from the same cluster are considered functionally related. Higher-order relationships between several proteins also can be identified using extensions of PP [20,21]. Phylogenetic profiles can also be identified for protein domains instead of entire proteins [22]. A profile is constructed for each domain and the presence/absence of the domain in different genomes is recorded which in turn can give information about domain interactions.

Some drawbacks of PP include its high computational cost, its dependence on high information profiles, and homology detection between distant organisms. For example, ubiquitous unlinked proteins present in all genomes (profiles with all "1"s) will be counted by PP as correlated. The same is true for proteins that are specific to a given genome (profiles with all, but one, "0"s). Shared phylogenetic relationships between two proteins can also produce false correlations between profiles. This issue has recently been addressed by incorporating the phylogenetic trees in the analysis of correlated gains and losses of pairs of proteins [23].

**Rosetta Stone method.** The Rosetta Stone approach infers protein interactions from protein sequences in different genomes [24–27]. It is based on the observation that some interacting proteins/domains have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein (Figure 1C). Gene fusion apparently occurs to optimize co-expression of genes encoding for interacting proteins. In *Escherichia coli,* the Rosetta Stone method found 6,809 potentially interacting pairs of nonhomologous proteins; both proteins from each pair had significant sequence similarity to a single protein from some other genome. Analysis of pairs found by this approach revealed that for more than half of the pairs both members were functionally related [24]. Comparison with the experimental data on protein interactions from the DIP database showed that about 6.4% of all experimental interactions can be linked by Rosetta Stone proteins.

**Sequence-based co-evolution methods.** As was mentioned earlier, interacting proteins very often co-evolve so that changes in one protein leading to the loss of function or interaction should be compensated by the correlated changes in another protein. The orthologs of coevolving proteins also tend to interact, thereby making it possible to infer unknown interactions in other genomes [28]. It has been argued that co-evolution can be reflected in terms of the similarity between phylogenetic trees of two non-homologous interacting protein families (Figure 1D). The similarity between phylogenetic trees can be quantified by calculating the correlation coefficient between distance matrices used to construct the trees with large values indicating co-evolution between two protein families [29,30] or domain families [31]. Correspondence between the elements of two matrices or branches of two trees is required to calculate the correlation coefficient, but such information is not always available. To address this issue, several algorithms have been developed to identify specific interaction partners between two interacting families that are especially useful when families contain paralogs with different binding specificities [32–34]. Given a pair of protein families, their distance matrices are aligned to minimize the difference between their elements, and interactions are predicted as those corresponding to aligned columns of two matrices. It was noticed earlier that most methods cannot perform an alignment search successfully if the size of families is large (more than 30 proteins in a family) [32]. One way to reduce the search space is to use the information encoded in phylogenetic trees [34].

The similarity between two phylogenetic trees is influenced by the speciation process, and therefore there is a certain "background" similarity between trees of any proteins, no matter if they interact or not. Different statistical techniques have been developed to account for "phylogenetic subtraction" [35]. Simplified versions of this approach were introduced recently to account for the background similarity in protein interaction prediction [36–38]. According to one




doi:10.1371/journal.pcbi.0030043.g001

**Figure 1.** Different Methods of Protein Interaction Prediction

(A) Gene cluster and gene neighborhood methods, different boxes showing different genes.
(B) Phylogenetic profile method, showing the presence/absence of four proteins in three genomes.
(C) Rosetta Stone method.
(D) Sequence co-evolution method looking for the similarity between two phylogenetic trees/distance matrices
(E) Classification methods shown with the example of RFD method, where five different features/domains are used and each interacting protein pair is encoded as a string of 0, 1, and 2. The decision trees are constructed based on the training set of interacting protein pairs and decisions are made if proteins under the question interact or not ("yes" for interacting, "no" for non-interacting).

of them [36], the "background" tree is constructed from the 16S rRNA sequences and is considered to be a canonical tree of life. The final distance matrices are obtained by subtracting the rescaled rRNA-based distances from the evolutionary distances obtained from the original phylogenetic trees. It has been shown that this method finds 50% of real interacting proteins at a 6.4% false positive rate compared with the 16.5% false positive rate obtained using methods which do not take into account evolutionary distances and the "background" canonical tree [29,30].

One example of how co-evolution studies could be used in confirming and predicting putative interaction partners is the case of DNA colicins and their immunity proteins [39]. Colicins consist of an N-terminal domain participating in translocation across the membrane of the target cell, the central domain which specifically binds to the extracellular surface receptor, and the C-terminal domain responsible for the toxic activity of colicin. Each DNase colicin has a specific immunity protein, which binds to the toxic domain and inhibits its cytotoxic activity. Co-evolution studies showed

that there is a significant correlation between the two families of DNA colicins and their immunity proteins (with the correlation coefficient of 0.67), with weaker correlation between Im2, Im8, and Im9 immunity proteins and their corresponding binding partners. Experimental studies indicated that there is indeed a cross-reactivity between colicin E9 and Im8 and Im2 proteins [40].

**Classification methods.** Different classification methods have been successfully applied to the prediction of protein and domain interactions [41–54]. These methods use various data sources to train a classifier to distinguish between positive examples of truly interacting protein/domain pairs from the negative examples of non-interacting pairs. Kernel methods are particularly useful in this respect as they provide a vectorial representation of data in the feature space through the set of pairwise comparisons [54]. Each protein or protein pair can be encoded as a feature vector where features may represent a particular information source on protein interactions, domain compositions, or evidence coming from various experimental methods. As a result of a comparison of different classifiers, it has been shown that Random Forest Decision (RFD) consistently ranks as a top classifier, with Support Vector Machines being in second place [55].

Figure 1E shows an example of the use of RFD to predict protein interactions. RFD builds decision trees based on the domain composition of interacting and non-interacting proteins, explores all possible combinations of interacting domains, and predicts at the end if a given pair of proteins interacts [43]. Each protein pair is represented as a vector of length N, where N is the number of different domain types (features), and each feature can have values 2, 1, or 0 depending if this domain is found in both proteins, in one of them, or not found in the protein pair. Given an experimental training set of interacting protein pairs, the method constructs a decision tree (or many trees) which defines the best splitting feature at each node from a randomly selected feature subspace. The best feature is selected based on the measure of "goodness of fit," which estimates how well this feature can discriminate between two classes of interacting and non-interacting pairs. The method stops growing the tree as soon as all pairs at a given node are well-separated into two classes. Traversing along the tree provides a classification for an unknown protein pair.

Multiple sources of direct and indirect data on protein–protein interactions can be combined in a supervised learning framework or integrative scoring scheme to predict protein–protein and domain–domain interactions [47,53,56–60]. It has been shown that the prediction accuracy is improved when several sources of data are used, and, in addition, integrative approaches can provide means to justify the confidence of inferred interactions.

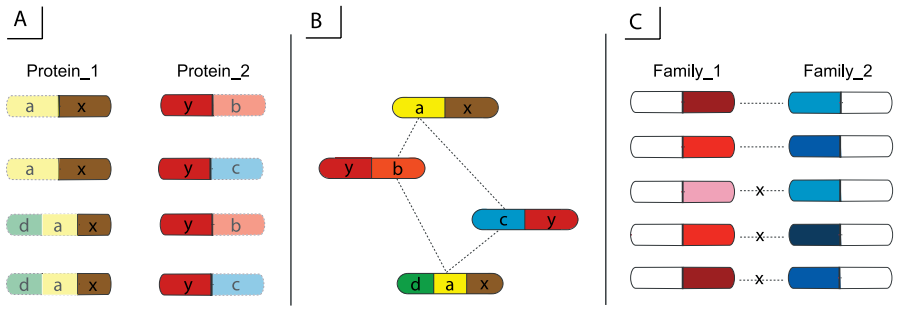
## Predicting Domain Interactions from Protein Interactions

By far the most coverage of experimental data describing protein interaction networks comes from high-throughput experiments giving us the identity of interacting protein pairs (see our previous review [1]). Unfortunately, these experiments reveal no structural details about the interaction interfaces and the formation of protein complexes. To deal with these limitations, several approaches have been developed to predict which domains in a protein pair interact given a set of experimental protein interactions; some of them focus on interactions involving specific mediating domains/peptides (SH2, SH3, PDZ domains) [61,62].

The following section gives an overview of domain prediction methods which are listed in Table 1. Note that some of the approaches already mentioned for protein interaction prediction, namely the sequence co-evolution, phylogenetic profiles, and classification methods, are also applicable to domain interaction prediction. Most methods begin by annotating protein sequences with domains that can be defined by Pfam, SCOP, CDD, or other domain databases [63–65]. The methods are typically trained on high-throughput protein interaction data. Predicted domain interactions are evaluated using structural data or by higher quality interaction sets such as MIPS [66]. Moreover, accounting for domains in proteins and domain interaction networks can in turn help in predicting protein interactions [67–70].

**Association methods.** This group of methods looks for the characteristic sequence or structural motifs which distinguish interacting proteins from non-interacting [71–74]. For this purpose association methods can use different classifiers (see previous section), and some of them are tuned specifically to identify domains responsible for protein interactions. For example, as shown in Figure 2A, correlated sequence signatures, or domains, that are found together more often than expected by chance can be used as markers to predict a new type of protein interaction [71]. In this case, protein interaction data is used to compute log-odds scores and to find correlated domains. The log-odds score is computed as: $\log_2(P_{ij}/P_iP_j)$, where $P_{ij}$ is the observed frequency of domains $i$ and $j$ occurring in one protein pair; $P_i$ and $P_j$ are the background frequencies of domains $i$ and $j$ in the data. In this approach predicted domain interactions have been defined as those having positive log-odds scores and having several instances of occurrence of a given domain pair in the database. Using this method, it was found that certain domains can be found quite often in protein interacting pairs and can be used for protein interaction prediction.

**Bayesian network models and maximum likelihood methods.** The association method which uses correlated sequence signatures [71] considers each pair of interacting domains separately, ignoring other domains in a given pair of interacting proteins (Figure 2A and 2B). Moreover, many association methods do not explicitly take into account the missing and incorrect interaction data which can be treated by using the Bayesian network methods [67,75,76]. To estimate the parameters of Bayesian models, the Maximum Likelihood Estimation method (MLE) [76] can be used. MLE maximizes the probability of interaction of all putative domain pairs and incorporates the experimental errors of protein interaction data into the scoring scheme (Figure 2B). The likelihood function is a function of parameters $\theta(\lambda_{ij}, f_p, f_n)$, where $\lambda_{ij}$ is the probability that domains $i$ and $j$ interact, $f_p$ is the false positive rate, and $f_n$ is the false negative rate derived from experimental data. It is difficult to maximize the likelihood function directly because of the large number of parameters (large number of different types of interacting domains). To solve this problem, the Expectation Maximization algorithm is used to find maximum likelihood




doi:10.1371/journal.pcbi.0030043.g002

**Figure 2.** Strategies to Predict Domain Interactions from Protein Interactions

(A) Shows that due to the abundance of domains *x* and *y* in protein interaction pairs shown on the same line, the domains *x* and *y* are predicted to interact.

(B) Illustrates the same dataset revealing that the actual domain interactions (dotted lines) do not include domains *x* and *y*. It shows that accounting for other domains in a protein pair in addition to *x* and *y* can result in alternative predictions.

(C) Considers the case of several paralogous protein pairs (from Family_1 and Family_2) containing the same two domains. In this case each paralog from one domain family (represented by a shade of red for Family_1) interacts with only one specific paralog (represented by a shade of blue for Family_2) of the other domain family. While there are examples of specific interacting domains (shown by dotted line), there are even more cases where they do not interact (shown with an "X"), meaning that the larger abundance of non-interacting examples can mask the few, specific interacting cases.

estimates of unknown parameters θ by finding the expectation of the complete data consisting of observed data and unobserved data in two iterative steps. The observed data includes protein–protein interactions and the domain composition of the proteins, and the unobserved data includes all putative domain–domain interactions.

**Domain pair exclusion analysis.** The domain pair exclusion analysis method extends the previously described MLE method and can detect specific domain interactions (see Figure 2C) which are hard to detect using MLE [77]. MLE and other methods emphasize nonspecific promiscuous domain interactions which are detected as those having large θ values. On the contrary, specific, rare interactions between certain members of two domain families can be neglected. The domain pair exclusion analysis method accounts for this by estimating an $E_{ij}$ score which measures the evidence that domains $i$ and $j$ interact and is defined as the logarithm of a ratio of two probabilities. The numerator corresponds to the probability that two proteins interact given that domains $i$ and $j$ interact. The denominator corresponds to the probability that proteins interact given that domains $i$ and $j$ do not interact. To compute E-scores for a given domain pair, the probability in the numerator is calculated with the Expectation Maximization procedure (similar to the one described in the previous section). For the probability in the denominator, the procedure is repeated where the probability for a given pair of domains to interact is set to zero. This allows the competing domains to maximize $\theta_{ij}$.

A high E-score value shows the high propensity of two domains to interact, while a low value indicates that competing domains from the same protein pair are more likely to be responsible for this interaction. Therefore, specific domain interactions can be found by screening for low θ values and high E-scores. Although this model does not account for false positives and negatives in the experimental data, it was shown that the E-scores perform better than its constituent quantities, finding 2.9 times more true positives than random assignment; for comparison, θ values yield 1.4 times more true positives than random assignments [77] .

***p*-Value method.** The *p*-value method tests a null hypothesis that the presence of a particular domain pair in a protein pair has no effect on whether two proteins interact [78]. To test this hypothesis, a statistic is calculated for each domain pair which takes into account experimental error (fraction of false positives) and incompleteness of the dataset (fraction of false negatives). The reference distribution is simulated by shuffling domains in proteins so that the network of protein interactions remains fixed. Obtained *p*-values show the reliability of domain interactions given that two proteins interact, and the domain pair with the lowest *p*-value is most likely to interact. The *p*-value method performs reasonably well when there are nine or more domains on a protein pair. However, interestingly enough, for the majority of test cases, random domain prediction outperforms all methods tested, pointing to the low accuracy of all prediction methods of domain interactions.

The methods of domain interaction prediction described in this section all have varying degrees of success, but have limitations common to most of them. First, domains are assumed to interact independently, although their interactions can depend on other domains in a protein pair. Second, incomplete domain assignments, due to insufficient coverage of domain databases and limited searching ability of domain profiles, can lead to false positive and negative interaction predictions. Finally, protein interaction data is not complete, whereas domain prediction methods are based on this data.

In this paper, we reviewed various computational methods to predict protein and domain interaction partners [79]. All of these methods use experimental data sources, some of them to a larger extent (gene co-expression, synthetic lethality) than others. As a result, they all suffer from the limitations of experimental approaches and incompleteness of observed data. Despite the fact that there is a certain circularity in testing the computational methods on experimental data, their prediction accuracy proved to be increasing, which makes them useful for the validation and analysis of diverse protein interactomes. The majority of presented prediction methods do not rely on protein structures and potentially can be applied on the genome-wide

scale; while structural analysis can provide further details of protein–protein and domain–domain interfaces and give clues on their modeling. ■

## Acknowledgments

The authors thank Elena Zotenko and Teresa Przytycka for helpful discussions and Robert Yates for graphic design of the figures.

**Author contributions.** BAS and ARP analyzed the data and wrote the paper.

**Funding.** This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health of the US Department of Health and Human Services.

**Competing interests.** The authors have declared that no competing interests exist.

#### References

1. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part I. Experimental techniques and databases. PLoS Comp Biol 3: e42.
2. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, et al. (2004) Prolinks: A database of protein functional linkages derived from coevolution. Genome Biol 5: R35.
3. Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. Nucleic Acids Res 29: 1216–1221.
4. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18: S329–S336.
5. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D (2003) Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: A combined computational approach. Genome Biol 4: R59.
6. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: Genomic analyses and predictions. Proc Natl Acad Sci U S A 97: 6652–6657.
7. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324–328.
8. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896–2901.
9. Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 18: 609–613.
10. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, et al. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30: 2212–2223.
11. Huynen M, Snel B, Lathe W III, Bork P (2000) Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Res 10: 1204–1210.
12. Teichmann SA, Babu MM (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol 20: 407–410.
13. Koonin EV, Wolf YI, Aravind L (2001) Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. Genome Res 11: 240–252.
14. Evguenieva-Hackenberg E, Walter P, Hochleitner E, Lottspeich F, Klug G (2003) An exosome-like complex in *Sulfolobus solfataricus*. EMBO Reports 4: 889–893.
15. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
16. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405: 823–826.
17. Date SV, Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol 21: 1055–1062.
18. Koonin EV, Galperin MY (2002) Sequence—Evolution—Function. In: Computational approaches in comparative genomics. Boston: Kluwer Academic Publishers. 488 p.
19. Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics 7: 420.
20. Bowers PM, O'Connor BD, Cokus SJ, Sprinzak E, Yeates TO, et al. (2005) Utilizing logical relationships in genomic data to decipher cellular processes. Febs J 272: 5110–5118.
21. Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. Science 306: 2246–2249.
22. Pagel P, Wong P, Frishman D (2004) A domain interaction map based on phylogenetic profiling. J Mol Biol 344: 1331–1346.
23. Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1: e3.
24. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein–protein interactions from genome sequences. Science 285: 751–753.
25. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90.
26. Marcotte CJ, Marcotte EM (2002) Predicting functional linkages from gene fusions with confidence. Appl Bioinformatics 1: 93–100.
27. Yanai I, Derti A, DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A 98: 7940–7945.
28. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, et al. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 287: 116–122.
29. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299: 283–293.
30. Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 14: 609–614.
31. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J Mol Biol 362: 861–875.
32. Ramani AK, Marcotte EM (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol 327: 273–284.
33. Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, et al. (2003) Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 19: 2039–2045.
34. Jothi R, Kann MG, Przytycka TM (2005) Predicting protein–protein interaction by searching evolutionary tree automorphism space. Bioinformatics 21: i241–250.
35. Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford: Oxford University Press.
36. Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352: 1002–1015.
37. Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482–3489.
38. Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2006) Predicting protein domain interactions from co-evolution of conserved regions. Proteins: Structure Function Genetics. In press.
39. Goh CS, Cohen FE (2002) Co-evolutionary analysis reveals insights into protein–protein interactions. J Mol Biol 324: 177–192.
40. Wallis R, Leung KY, Pommer AJ, Videler H, Moore GR, et al. (1995) Protein–protein interactions in colicin E9 DNase–immunity protein complexes. 2. Cognate and noncognate interactions that span the millimolar to femtomolar affinity range. Biochemistry 34: 13751–13759.
41. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302: 449–453.
42. Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2005) Random forest similarity for protein–protein interaction prediction from multiple sources. Pac Symp Biocomput: 531–542.
43. Chen XW, Liu M (2005) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21: 4394–4400.
44. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22: 78–85.
45. Gilchrist MA, Salter LA, Wagner A (2004) A statistical framework for combining and interpreting proteomic datasets. Bioinformatics 20: 689–700.
46. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
47. Yamanishi Y, Vert JP, Kanehisa M (2004) Protein network inference from multiple genomic data: A supervised approach. Bioinformatics 20: I363–I370.
48. Zhang LV, Wong SL, King OD, Roth FP (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5: 38.
49. Fariselli P, Pazos F, Valencia A, Casadio R (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269: 1356–1361.
50. Koike A, Takagi T (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 17: 165–173.
51. Bradford JR, Westhead DR (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 21: 1487–1494.
52. Albert I, Albert R (2004) Conserved network motifs allow protein–protein interaction prediction. Bioinformatics 20: 3346–3352.
53. Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein–protein interactions. Bioinformatics 21: i38–46.
54. Scholkopf B, Tsuda K, Vert JP, editors (2004) Kernel methods in computational biology. Cambridge, Massachusetts: MIT Press.
55. Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of different

biological data and computational classification methods for use in protein interaction prediction. Proteins 63: 490–500.
56. Liu Y, Liu N, Zhao H (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21: 3279–3285.
57. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198–207.
58. Ng SK, Zhang Z, Tan SH (2003) Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19: 923–929.
59. Huttenhower C, Hibbs M, Myers C, Troyanskaya OG (2006) A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics 22: 2890–2897.
60. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, et al. (2004) Combining biological networks to predict genetic interactions. Proc Natl Acad Sci U S A 101: 15682–15687.
61. Reiss DJ, Schwikowski B (2004) Predicting protein–peptide interactions via a network-based motif sampler. Bioinformatics 20: I274–I282.
62. Lehrach WP, Husmeier D, Williams CK (2006) A regularized discriminative model for the prediction of protein–peptide interactions. Bioinformatics 22: 532–540.
63. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, et al. (2004) SCOP database in 2004: Refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226–D229.
64. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: Clans, web tools and services. Nucleic Acids Res 34: D247–D251.
65. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, et al. (2002) CDD: A database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30: 281–283.
66. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, et al. (2006) MPact: The MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436–D441.
67. Gomez SM, Rzhetsky A (2002) Towards the prediction of complete protein–protein interaction networks. Pac Symp Biocomput: 413–424.
68. Wuchty S (2006) Topology and weights in a protein domain interaction network—A novel way to predict protein interactions. BMC Genomics 7: 122.
69. Moon HS, Bhak J, Lee KH, Lee D (2005) Architecture of basic building blocks in protein and domain structural interaction networks. Bioinformatics 21: 1479–1486.
70. Wuchty S, Almaas E (2005) Evolutionary cores of domain co-occurrence networks. BMC Evol Biol 5: 24.
71. Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol 311: 681–692.
72. Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17: 455–460.
73. Espadaler J, Romero-Isart O, Jackson RM, Oliva B (2005) Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics 21: 3360–3368.
74. Martin S, Roe D, Faulon JL (2005) Predicting protein–protein interactions using signature products. Bioinformatics 21: 218–226.
75. Gomez SM, Noble WS, Rzhetsky A (2003) Learning to predict protein–protein interactions from protein sequences. Bioinformatics 19: 1875–1881.
76. Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain–domain interactions from protein–protein interactions. Genome Res 12: 1540–1548.
77. Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biol 6: R89.
78. Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA (2005) Statistical analysis of domains in interacting protein pairs. Bioinformatics 21: 993–1001.
79. Panchenko AR, Shoemaker BA (2006) Protein–protein interactions: Structure and systems approaches to analyze diverse genomic data. Available: http://www.ncbi.nlm.nih.gov/CBBresearch/Panchenko/ismb__tutorial2006.ppt. Accessed 16 February 2007.