Protein interaction networks: Protein domain interaction and protein function prediction

Yanjun Qi and William Stafford Noble

Abstract Most of a cell's functional processes involve interactions among proteins, and a key challenge in proteomics is to better understand these complex interaction graphs at a systems level. Because of their importance in development and disease, protein-protein interactions (PPIs) have been the subject of intense research in recent years. In addition, a greater understanding of PPIs can be achieved through the detailed investigation of the protein domain interactions which mediate PPIs. In this chapter, we describe recent efforts to predict interactions between proteins and between protein domains.

We also describe methods that attempt to use protein interaction data to infer protein function. Protein-protein interactions directly contribute to protein functions, and implications about functions can often be made via PPI studies. These inferences are based on the premise that the function of a protein may be discovered by studying its interaction with one or more proteins of known functions. The second part of this chapter reviews recent computational approaches to predict protein functions from PPI networks.

1 Introduction

In recent years, the human and other genome sequencing projects have generated vast amounts of data that identify thousands of new gene products whose functions and interrelationships are not yet known. The overall molecular architecture of all organisms is largely mediated both structurally and functionally through the coordination of protein-protein interactions (PPIs). In particular, the disruption of PPIs

Yanjun Qi
Machine Learning Department, NEC Labs America, e-mail: yanjun@nec-labs.com

William Stafford Noble
Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, e-mail: noble@gs.washington.edu

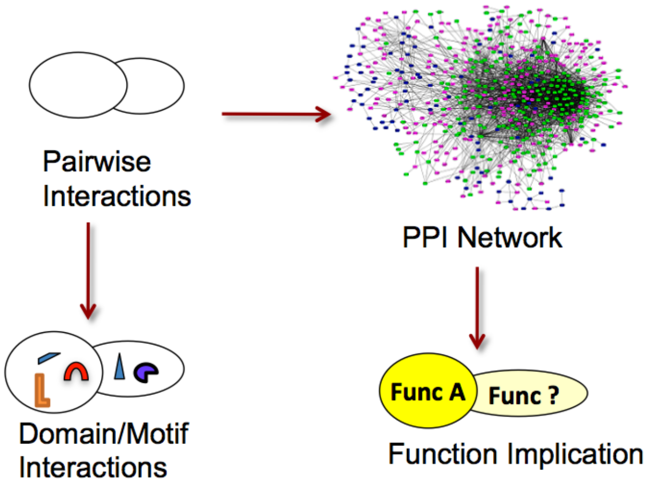


Fig. 1 The framework of contents in this chapter.

may lead to the development of diseases. Thus, correctly identifying the interrelationship between proteins at the systems level is urgent and necessary, since such knowledge would lead to a better understanding of the functional properties that define the behaviors of most complex biological systems.

Experimental techniques [82] to detect PPIs or protein functions have their own limitations, and the resulting data sets are often noisy. Thus, additional approaches are needed to accelerate the recovery of complex protein-interaction systems. Given the vast amount of available biological evidence and the representational power of mathematical models, computational methods are gaining importance. In this chapter, we review three areas to which computational approaches contribute significantly (Figure 1). We first introduce methods targeting protein-protein interaction predictions in Section 2. Then in Section 3 recent advances in identifying domain-domain interactions are presented. Finally, Section 4 reviews various ways to predict protein functions from PPI graphs.

2 Prediction of protein-protein interactions

The term "protein-protein interactions" refers to the association of protein molecules with each other. The associations are interesting from multiple perspectives, including ascertainment of specific biological processes and pathways such as signal transduction pathways, as well as the systems-level studies of networks on the cellular or organism-wide scale. Because direct pairwise PPIs provide the basic building blocks to carry out the myriad of functions in a cell, comprehensively identifying these interactions is essential for understanding the molecular mechanisms underlying biological functions.

Experimental techniques for deciphering protein-protein interactions have been reviewed by [82]. In general, interactions among proteins can take on many forms (e.g., have an impact on functions of one another, or occur in a common pathway), and many proteins only operate in complexes and through physical contact with other proteins. These factors have prompted the development of various complementary experimental methods for detecting protein-protein interactions. Traditionally, PPIs have been studied individually through the use of genetic, biochemical and biophysical experimental techniques (also termed *small-scale* methods). The related experiments are generally time-consuming, sometimes requiring months to detect one PPI. In the last several years, *large-scale* biological PPI experiments have been introduced to directly detect hundreds or thousands of protein interactions at a time. Yeast two-hybrid (Y2H) screens [36, 32, 76, 87] and protein complex purification detection techniques using mass spectrometry [24, 23, 32] are the two most widely used large-scale approaches. However, both methods suffer from high false positive and false negative rates [56]. For the Y2H method, this is due to insufficient depth of screening and misfolding of the fusion proteins. In addition, interaction between ”bait” and ”prey” proteins has to occur in the nucleus, where many proteins are not in their native compartment. The mass spectrometry based complex identification methods [24, 23, 32]) may miss complexes that are not present under the given conditions. In addition, tagging may disturb complex formation and weakly associated components may dissociate and escape detections. In general, the resulting data sets are often incomplete and exhibit high false positive and false negative rates [56, 15, 100]. Consequently, even for well-studied model organisms, most true PPIs have not yet been discovered experimentally.

Computationally, protein-protein interaction networks can be conveniently modeled as undirected graphs, where the nodes are proteins and edges represent physical binding interactions. Initially, this graph is missing many edges (false negatives) and contains many incorrect edges (false positives). To complement and extend experimental methods, a variety of computational methods have been successfully applied to predict protein interactions. These approaches may be categorized on the basis of the types of data they considered when making predictions, as follows:

- Over-represented domain pairs or motif pairs observed in interacting protein pairs have been studied and used to infer PPIs. We provide more details of domain-domain interactions in Section 3. Structural information and sequence evidence about PPI interfaces has been used to predict potential PPIs [21, 13] as well.
- Various genomic methods infer protein interactions based on the conservation of gene neighborhood (Figure 2), conservation of gene order, gene fusion events, or the co-evolution of interacting protein pair sequences [83, 55].
- An attractive alternative approach is to integrate various types of evidence from multiple sources in a statistical learning framework. A number of classification

methods have been explored and multiple ways of using biological evidences have been studied in this framework [6, 8, 38, 102, 97, 68, 72, 79, 61, 99].

- High-throughput PPI experiments for elucidating protein-protein interactions have been applied to model organisms in recent years. Unfortunately the derived data sets are noisy and incomplete [56]. Multiple computational techniques have been proposed to improve the data reliability [5, 85, 10].

In the next sections, we describe in detail methods that fall into the latter three categories.

As mentioned above, interactions among proteins can take on many forms. Most previous computational works either predict direct physical interactions between proteins, or to identify if two proteins operate in the same complex, or to predict if two proteins are functionally linked to each other. The readers should keep this distinction in mind for the following methods. Qi et al. [67] performed a systematic comparison between these tasks and found that the task of identifying co-complex relationship seems to be easier than the other two tasks, with respect to the feature evidence they used.

2.1 Genomic Inference with Context

Accurate and large-scale prediction of protein-protein interactions directly from protein sequences is one of the important challenges in computational biology. Reviewed in [83] as “genomic inference methods” (including gene neighbor, gene fusion, and phylogenetic profile approach), this category uses genomic or protein context to infer functional associations between proteins.

Gene neighborhood: The idea of the gene neighborhood approach is shown in Figure 2. We can see that genes $P1$, $P2$ and $P3$ are neighbors across three different genomes. From this association, we infer that their protein products are likely to associate with one another. The gene neighborhood approach provides strong signals for functional association between gene products within and across species [55], but this approach is arguably less well suited for specifically detecting physical interactions.

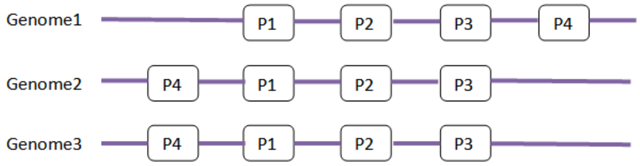


Fig. 2 PPI prediction by gene neighborhood approach (modified from Figure 1 in [83])

Gene fusion: The gene fusion approach [53], infers protein interactions from protein sequences in different genomes. It is based on the observation that some interacting proteins/domains have homologs in other genomes that are fused into one protein chain. Figure 3 gives an example of "gene fusion."

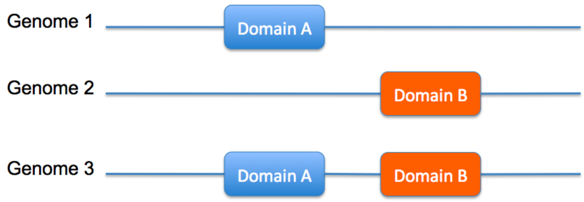
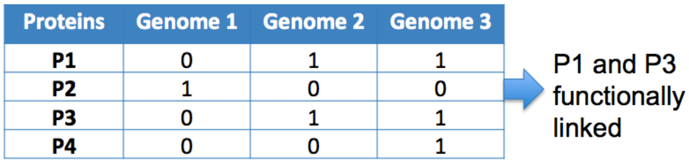


Fig. 3 PPI prediction by gene fusion (modified from Figure 1 in [83])

Phylogenetic profile: The phylogenetic profile method [65] is based on the observation that interacting proteins need to be present simultaneously in order to perform their functions. Therefore, the repeated co-occurrence of a pair of proteins across different organisms provides evidence that they interact. As shown in Figure 4, a phylogenetic profile is constructed for each protein as an N-dimensional vector, where N is the number of genomes under consideration. The presence or absence of a given protein in a given genome is indicated with a 1 or 0 at each position in the profile. Proteins' phylogenetic profiles can then be linked using a bit-distance measure, with linkage indicating physically interaction or functional assocation [65, 83]. This approach can also be used for protein domains, where a profile is constructed for each domain.



Proteins	Genome 1	Genome 2	Genome 3
P1	0	1	1
P2	1	0	0
P3	0	1	1
P4	0	0	1

P1 and P3 functionally linked

Fig. 4 PPI prediction by phylogenetic profile strategy (modified from Figure 1 in [83])

2.2 *Classification from Multiple Types of Evidence*

Studies in this category make use of a classification algorithm to integrate diverse biological datasets (Figure 5). A classifier is trained to distinguish between positive examples of truly interacting protein pairs and negative examples of non-interacting pairs. Many different research groups have independently suggested using supervised learning methods for predicting protein interactions. However, the data sources, approaches and the species they worked on have varied widely. According to these differences, we categorize previous works into four groups: supervised classifiers on protein pairs, kernel based network reconstruction, direct modeling of PPI data sets, and inter-species PPI prediction.

2.2.1 Supervised Classifiers on Each Protein Pair Separately

By transforming multiple biological data sources into a feature vector representing every pair of proteins, the task of predicting pairwise protein interactions can be formalized as a binary classification problem. Each protein pair is encoded as a feature vector where features may represent a particular information source such as related mRNA expressions, domain composition, or evidence coming from experimental methods. There are many possible ways to encode evidence sources into feature attributes and it is an important factor for the reliability of the computational predictions [67]. For instance, pearsons correlation values between two genes could be used as features on selected gene expression sets. Alternatively, feature attributes could describe how likely two proteins interact in other species [55].

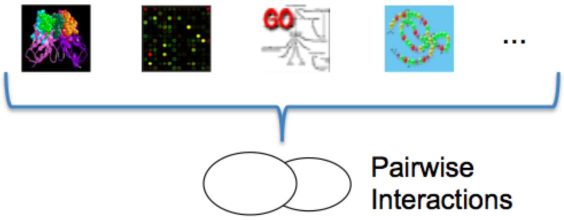


Fig. 5 PPI prediction by classification with multiple evidence

A number of proposed methods belong to this group, including naive Bayes classifiers [38] , decision trees [102], kernel based methods [97, 6], random forests [52, 68], logistic regression [4, 52], and the strategy of summing likelihood ratio scores to predict PPI confidence in human [72, 79, 71] or in yeast [48]. Multiple classifiers were compared for PPI predictions in yeast [67]. Random forests and support vector machines (SVMs) were found to achieve the best performance among them.

These approaches used different types of data, different supervised classifiers and generally treated each protein pair independently for the interaction identification.

The popular STRING database [55] is a successful example of an application of this supervised learning methodology. The authors identified functionally associated protein pairs by computationally integrating known protein-protein associations, coexpression pairs, literature mining and pairs transferred across organisms. The resulting STRING database integrates and ranks predicted PPIs, by benchmarking them against a common reference set with the modified sum of likelihood approach. The most recent version of STRING [41] covers about 2.5 million proteins from 630 organisms. The authors claim that this provides the most comprehensive view of PPIs currently available.

Most of the above scoring methods use a set of likely true positives to train the predictive model. However, a single positive training set may be biased and not representative of true interaction space. To address this concern, Yu et al. [101] demonstrated a method to score protein interactions by using multiple independent sets of training positives to reduce the potential bias inherent in using a single training set. Defining negatives can also be problematic, since the absence of an edge in an observed network does not necessarily imply that the edge does not exist in the true network. Several studies attempt to define a set of high-confidence non-interacting proteins [40, 39]; however, such methods are likely to yield their own biases [7]. Thus, the simpler approach of selecting negatives uniformly at random is generally preferred [28, 6, 103, 69].

2.2.2 Network Reconstruction with Kernel Methods

As mentioned above, multiple data evidence used for PPI predictions are in different formats (e.g. numeric values for gene expression, letter strings for protein sequences). A natural choice for this data integration task is kernel methods [6], which unify the data representation as special matrices called kernels (Figure 6(b)). Kernel methods have been applied successfully on the protein interaction prediction tasks in recent years. The problem of PPI predictions could be framed as the following network reconstruction problem (Figure 6). The input is a graph $G = (V, E, \bar{E})$, where V is a set of nodes representing each protein, and $E, \bar{E} \subset V \times V$ are sets of known edges and non-edges, respectively, corresponding to protein pairs that are known to interact or not. This PPI graph is represented as an adjacency matrix in Figure 6(a) which contains known interactions (black boxes), known non-interactions (white boxes) and pairs with unknown status (gray boxes). In Figure 6(b), kernel methods build kernel matrices (graphs) based on features of proteins or protein pairs. The key question then is to reconstruct those "?" entries in the input PPI graph (gray boxes of Figure 6(a)) based on the kernel graph(s) (Figure 6(b)). Here we describe three interesting papers in this group.

Pairwise kernel between protein pairs: Ben-Hur et al. [6] and Gomez et al. [27] proposed the pairwise kernel approach to use a standard kernel method (such as SVM) for PPI predictions. Treating each protein pair as a data example, a pairwise kernel function computes the similarity between two pairs of proteins. Thus, with n proteins, the resulting kernel matrix (an example in Figure 8(b)) contains n^4 entries. One way to construct such a kernel is to build them on top of an existing kernel between individual proteins. For example, given a kernel matrix K with each entry describing the inner product between two proteins, the pairwise kernel could be built for the four proteins in Figure 8(a) as follows:

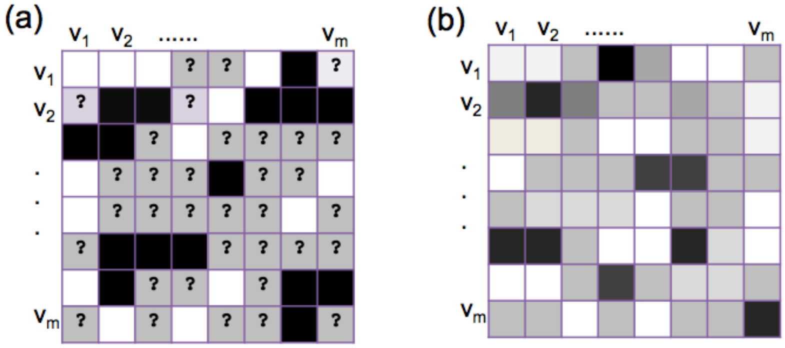


Fig. 6 PPI predictions through kernel methods (modified from Figure 1 of [99]). (a) PPI network is represented as an adjacency matrix which includes: known interactions (black boxes), known non-interactions (white boxes) and pairs with unknown status (gray boxes). (b) Kernel matrix built from a certain feature evidence, with a darker color describing larger value.

$$K'((v_1,v_2),(v_3,v_4)) = K(v_1,v_3)K(v_2,v_4) + K(v_1,v_4)K(v_2,v_3) \qquad (1)$$

The motivation is that protein pair (v_1,v_2) is similar to protein pair (v_3,v_4) if the two proteins v_1 and v_2 are similar to proteins v_3 and v_4, or vice versa. Later, Martin et al. [54] proposed a similar way to make use of protein properties for PPI prediction task, but with a tensor product kernel.

As a continuation of this work, the authors in [70] predicted co-complexed protein pair (CCPP) relationships using kernel methods from heterogeneous data sources. They show that a diffusion kernel [46, 84] based on random walks on the full network topology yields good performance in predicting CCPPs from protein interaction networks (for more details about this kernel, see Section 4.5) . In their setting of direct ranking, a diffusion kernel performs much better than the mutual clustering coefficient. Alternatively, when using SVM classifiers, a diffusion kernel performs much better than a linear kernel. One recent work from Vert et al. [92] explored a closely related approach called the “metric learning pairwise kernel” to convert the problem of direct inference based upon similarities between nodes joined by an edge on the PPI graph to the task of distance metric learning.

Note that the pairwise kernel strategy also belong to the group of methods in Section 2.2.1. Those methods use feature values to describe each protein pair. With an inner product between these features vectors, we could generate a pairwise kernel matrix. Of course, the way to calculate the kernel matrix in equation 1 is more general, since the pairwise kernel could incorporate data from individual proteins (using a pairwise kernel) and protein pairs.

Supervised reconstruction with a kernel between proteins: Because the computational cost for the above pairwise kernel is high, Yip et al. [99] and Yamanishi et al. [97] proposed to work directly with kernels defined on individual proteins. Given such a kernel K (between proteins) and a cutoff t, the method simply predicts interactions for each pair of proteins for which $K(v_i,v_j) \geq t$.

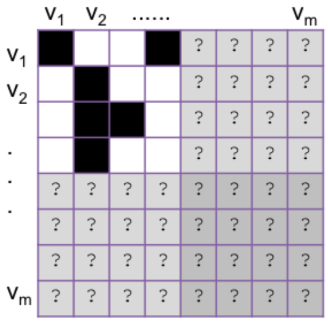


Fig. 7 PPI predictions by the supervised network inference (modified from Figure 1(c) of [99]). Partial complete adjacency matrix required by the supervised reconstruction approach, which needs complete knowledge of a submatrix (upper-left).

To make use of the training examples, supervised algorithms were presented to reconstruct the kernel matrix based on a sub-matrix of known interactions. Assuming that the sub-network of the adjacency matrix is totally known (as shown in Figure 7), the goal is to modify the kernel similarity between proteins (as defined by the kernel) to some values that are more consisitent with the partial sub-matrix. Subsequently, simple thresholding is performed on the resulting similarity values to predict PPIs [99]. Yamanishi et al. [97] presented a method in this style to infer protein interaction networks using a variant of kernel canonical correlation analysis (originated from spectral clustering theory). The goal was to identify features from the input kernel (built from the genomic/proteomic evidence) and features from the diffusion kernel that were derived from the known PPI submatrix, so that two features have the highest correlation under certain smoothness requirements.

Kernel matrix completion: Similar to the above supervised network reconstruction, Kato et al. [43] also assume a partially complete adjacent matrix (Figure 7). They formulated supervised network inference as a kernel matrix completion problem, where the inference of edges boils down to estimation of missing entries of a kernel matrix. The goal is to make the resulting matrix closest to a spectral variant of the kernel matrix as measured by the KL (Kullback-Leibler) divergence. An expectation-maximization algorithm is proposed to simultaneously infer the missing entries of the adjacency matrix and the weights of multiple datasets (a weight is assigned to each type of dataset and thereby to select informative ones). The algorithm iteratively searches for the filled adjacent matrix that is closest to the current spectral variant of the kernel matrix, and at the same time, the spectral variants of the kernel matrix which is closest to the current filled matrix. When convergence is reached, the predictions are thresholded from the final complete adjacency matrix.

Local model: Each of the above approaches builds a global model to predict new edges over the network based on the partial knowledge of the network to be inferred (Figure 8(b)). This single model may not be able to separate all cases of interacting pairs from non-interacting ones, if there are different subgroups of interactions [99]. For instance, protein pairs involved in transient interactions may use a very different

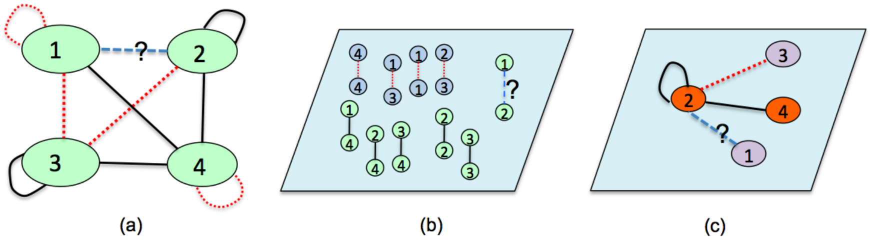


Fig. 8 Global and local modeling for PPI network reconstruction (modified from Figure 2 of [99]). (a) An interaction network, with solid black lines representing known interactions, red dotted edges representing known non-interacting edges and blue dashed lines representing those protein pairs with unknown interaction status. (b) Global model based on pairwise kernel approach, where each edge is treated independently. (c) Local model for protein v_2. Different node colors indicate distinctive evidence status, for instance, different cell compartments that the proteins reside in.

strategy compared with those involved in protein complexes. These two types of interactions may belong two separate subgroups that cannot be fitted by one single model.

Accordingly, Bleakley et al. [8] introduce a novel method that uses a local model to allow for flexible modeling of subgroups of interactions. A local model is built for each protein, using the known interactions and non-interactions of this protein as the positive and negative examples. The resulting classification rule predicts edges associated with a single protein. Thus, each pair of proteins receives two predictions, each from the local model of either protein. In Figure 8(c), the method built a local model for protein v_2. Because node v_1 is similar to node v_3, this local model classified pair (v_2, v_1) as negative. Since each node has its own local model, the approach only needs a kernel defined on proteins, rather than a kernel between pairs of proteins.

Local model with training set expansion: The accuracy of computational techniques proposed for PPI network reconstruction is consistently limited by the small number of high-confidence examples. Specifically, for the local model approach, the uneven distribution of positive examples across the potential interaction space, with some objects having many known interactions and others few, makes it hard to predict new interaction partners for those proteins having very few known interactions reliably. To address this issue, Yip et al. [99] proposed two semi-supervised learning methods by augmenting the limited number of gold-standard training instances with carefully chosen and highly confident auxiliary examples.

- The first method, *prediction propagation* is similar to self-training methods [80] described in the the machine learning community. This method uses highly confident predictions from one local model as the auxiliary examples of another. This propagation strategy uses the learning from information-rich regions in the training network to help make predictions in information-poor regions.

- The second method, *kernel initialization*, takes the most similar and most dissimilar proteins of each protein in a global kernel (between proteins) as the auxiliary examples. Similar to prediction propagation, adding these new examples into the training sets boosts the performance of the local modeling approach.

2.2.3 Inter-species PPI prediction

All of the above studies aim to predict PPIs within a single organism (termed *intra-species PPI prediction*), with most studies focusing on yeast or human. Recently, researchers have begun to extend computational methods to predict PPIs between species (termed *inter-species PPI prediction*).

Of particular interests are host-pathgen PPIs. For any host-pathogen system, it is important to understand the mechanism by which a pathogen can infect its host. One method of infection is via protein interactions, where pathogen proteins target host proteins (as described in Figure 9). Developing computational methods that identify which PPIs enable a pathogen to infect a host has significant implications in identifying potential therapeutical targets.

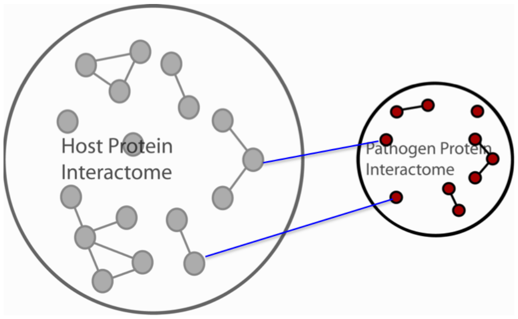


Fig. 9 Protein-protein interactions in host-pathgen systems (modified from Figure 1 of [88])

Davis et al. [16] studied ten host-pathogen protein-protein interactions using structural information with a comparative model: the host/pathogen protein pairs that share similarity to protein complexes with known structures are used to build 3-D structural models of putative complexes, and the modelled pairs are then filtered by functional and genomic experimental information. The technique was applied to ten pathogens and assessed by three independent computational procedures. The results suggest that this method is complementary to experimental efforts in elucidating networks of hostpathogen protein interactions.

Later, Tastan et al. [88] extended the supervised learning framework to predict PPIs between HIV-1 viruses and human proteins. A random forest based classifier was used to integrate multiple biological data types, achieving state-of-the-art performance for this task.

Similar to host-pathgen PPI, several recent papers identify interactions between drugs and target proteins. This is a key area in genomic drug discovery. The authors in [96] formalized the drug-target interaction inference as a supervised learning problem on a bipartite graph, where the model extended the metric embedding approach [1] to integrate chemical and genomic spaces into a unified space.

2.3 Modeling Experimental PPI Data Sets Directly

Genome-wide, high-throughput PPI experiments for elucidating protein-protein interactions have proven to be one of the most important tools in recent years. However the quality of currently available PPI data sets is unsatisfactory, which limits its usefulness to some degree. A crucial step in analyzing proteomics PPI data is to separate the subset of credible interactions from the background noise. Various computational techniques have been proposed for inference of reliable protein-protein interactions directly from experimental interaction results. In the following, several interesting ones are covered.

Von Mering et al. [56] were among the first to discuss the problem of accurately inferring protein interactions from high-throughput data sources. The proposed solution [56], which used the intersection of direct high-throughput experimental results, achieved a very low false positive rate. However, the coverage was also very low. Less than 3 percent of known interacting pairs were recovered using this method.

Later, Bader et al. [5] applied logistic regression to estimate the posterior probability that a pair of proteins will interact. Only statistical and topological descriptors were used to predict the biological relevance of protein-protein interactions obtained from high-throughput PPI screens for yeast. Other evidence, such as mRNA expression, genetic interactions and database annotations, were subsequently used to validate the model predictions. They demonstrated that it is possible to define a quantitative confidence measure based entirely on screening statistics and network topology. The main assumption underlying the confidence measure is that nonspecific interactions are highly likely to be technology-specific [5]. This type of analysis is essential for analyzing the growing amount of genomic and proteomics interation data in model organisms.

Aiming to improve the quality of experimentally available PPI data by identifying erroneous datapoints from PPI experiments, Sontag et al. [85] described a probabilistic approach to estimate errors in yeast-two-hybrid experiments, considering both random and systematic errors. The systematic errors arise from limitations of the Y2H experimental protocol: ideally the reporting mechanism in Y2H should be activated if and only if the two proteins being tested truly interact, but in practice, even in the absence of a true interaction, the reporter may be activated by some proteins - either by themselves or through promiscuous interaction with other proteins. The authors described a probabilistic relational model that explicitly models these two types of errors. They use Markov chain Monte Carlo algorithms for inference. In constrast to previous work, which often models Y2H errors as being independent and random, experimental results showed that this approach could make better use of the available experimental data.

Currently no method exists to systematically and experimentally assess the quality of individual interactions reported in interaction mapping experiments. Braun et al. [10] developed an interaction tool kit consisting of four complementary, high-throughput protein interaction assays and provided a standardized confidence-scoring method. Based on positive and random reference sets consisting of well documented pairs of interacting human proteins and randomly chosen protein pairs, a logistic regression model was trained to combine the assay outputs and calculate the probability that any newly identified interaction pair is a true biophysical interaction once it has been tested in the the four high-throughput PPI assays. This approach allows a systematic and empirical assignment of confidence scores to all individual protein-protein interactions from high throughput interation experiments.

The above approaches have considered protein pairs independently when inferring the presence of PPIs. In contrast, Jaimovich et al. [37] considered the neighborhood interaction pairs together and employed a *relational Markov random field* approach for collective inference of PPIs in yeast. The basic idea is shown in Figure 10:

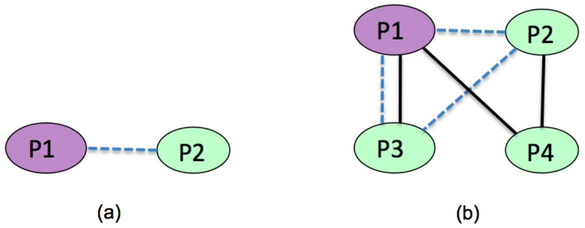


Fig. 10 Improve PPI prediction with dependencies between interactions (modified from Figure 1 in [37]). (a) A possible interaction between proteins P1 and protein P2. They are localized in different cellular positions (indicated with purple and green colors). (b) Two additional proteins P3 and P4 provide extra dependency evidence. Dashed line represents functional association from indirect evidence and solid line describes interactions from experimental interaction sets. The combined evidence gives more support to predict that P1 and P2 interacts.

In this paper [37], the authors view the PPI prediction task as a relational learning problem, where observations about different entities are not independent. The method exploits relational probabilistic models to combine multiple types of features, including protein attributes (e.g., localization of proteins) and protein-protein interactions (e.g., experimental interaction assays). The results demonstrated that modeling the dependencies between interactions leads to significantly better predictions. However, due to the model complexity and the difficulties during inference, this model can currently be applied only to a small set of proteins.

3 Prediction of domain-domain interactions

Many of the experimental and computational approaches described above address the question, “Do these two proteins interact?” In practice, *how* the proteins interact is also of great interest. Protein interactions occur through physical binding of small regions on the surface of proteins. Therefore, insights into the mechanism whereby a protein carries out its function can be obtained by identifying the interaction site where protein binding takes place. Moreover, detailed knowledge about the binding sites at which an interaction takes place can provide insight into the causes of human disease as well as a starting point for drug design [93]. Unfortunately, this type of information is not typically provided in a protein interaction graph and is not revealed by high-throughput experimental methods.

A protein may contain a single domain or multiple domains, each one typically associated with a specific function [89]. The combination of domains determines the function of the protein, such as its subcellular localization and the interactions it is involved in [34]. There exists a certain degree of conservation in the interaction patterns between similar proteins and domains. It has been found that close homologs almost always interact in the same way [82]. Thus, it is interesting to find out what domains are responsible for binding.

Currently little useful data is available from major databases with respect to relations on the domain level [64]. This lack of data makes computational prediction of domain-domain interactions very important. A series of computational approaches have been developed to predict which domains in a protein pair interact given a set of experimental protein interactions [83]. Domain interactions extend the functional significance of proteins and provide a more detailed view of the protein-protein interaction network (Figure 11).

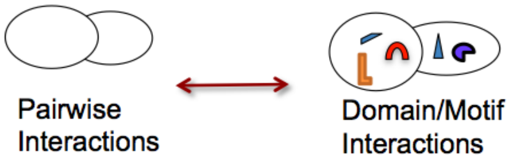


Fig. 11 Prediction of domain-domain interactions

Inferring interactions between domains from protein-protein interactions is a challenging task. Various methods have been proposed to predict domain interactions from protein-protein interaction graphs. Most methods begin by annotating protein sequences with domains that can be defined by Pfam, CDD, or other domain databases. The models are typically trained with certain known protein interactions to identify domain-domain interaction pairs. The predicted domain interactions can be evaluated using structural data or by high quality interaction sets. Moreover, the resulting domain interactions can in turn help in predicting protein-protein interactions. It is worthwhile to mention that some of the approaches mentioned in the last section for protein interaction prediction, such as the sequence co-evolution or phylogenetic profiles (reviewed in [64]) are also applicable to domain interaction prediction [83]. In addition, the following section introduces several methods specifically designed to predict domain-domain interactions from protein interaction data.

Inferences on the interactions among domains can be made by analyzing the domain composition of a set of proteins and their interaction networks.

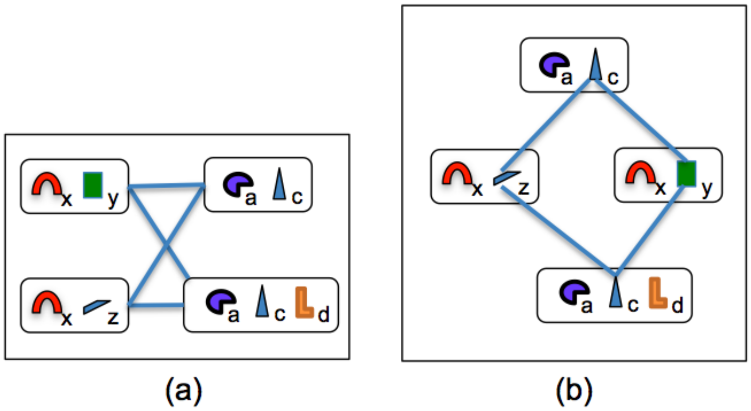


Fig. 12 Two methods to predict domain-domain interactions from PPIs. (a) Association method. The domains *x* and *a* are predicted to interact due to the abundance of domains *x* and *a* in protein interaction pairs, shown as the blue line. (b) As the same PPI dataset in (a), that the actual domain interactions (blue lines) do not include domains *x* and *a*. This shows that accounting for other domains in a protein pair, in addition to *x* and *a*, can result in alternative domain interaction predictions.

Association method: A characteristic domain or structural motifs can be used to distinguish interacting proteins from non-interacting. Association methods [86, 30, 83] use different classifiers for this purpose, and some of them are tuned specifically to identify domains responsible for protein interactions. *Correlated domains* are pairs of domains that are found together more often than expected by chance in known PPI pairs. An association method may predict that two proteins interact if they contain correlated domains, one from each protein, whose association value is greater than a predefined threshold. Because some domain pairs can be found quite often in protein interacting pairs, this simple assocation method can be quite successful in identifying novel PPIs.

An examplar case is given in Figure 12(a). Domain pair (x, a) is the most abundant in all four interacting protein pairs (blue lines) compared with other domain-domain pairs. Taking the domain combination pair as a basic unit, these methods use their frequencies in the interacting and non-interacting sets of protein pairs, for deriving novel protein interactions. For example, Sprinzak et al. [86] use the following score, computed from protein interaction data, to find correlated domains:

$$S(d_m, d_n) = \frac{I_{mn}}{N_{mn}} \tag{2}$$

where I_{mn} is the number of interacting pairs that contain (d_m, d_n), and N_{mn} is the total number of protein pairs that contain (d_m, d_n).

Dyer et al. [20] extended this idea for identify domain interactions in host-pathogen systems. They integrate a number of public intra-species PPI datasets with protein-domain profiles for predicting and studying host-pathogen PPI networks. The model used intra-species PPIs and protein-domain profiles to compute statistics on how often proteins containing specific pairs of domains interact. These statistics can then be used to predict inter-species PPIs in host-pathogen systems.

Maximum Likelihood Estimation: One drawback of the association method is that it ignores other domain-domain interaction information between the protein pairs and, thus, does not make full use of all of the available information. As in Figure 12(a), if domains x and a do not appear in any other proteins, then in the association method this pair is assigned the association score $S(x,a) = 4/4 = 1$. This method ignores other domain-domain interactions among domains b, c, y and z. To infer a domain-domain interaction, other related domain-domain interactions should be taken into account (as shown in Figure 12(b)). To do so, interactions among other proteins containing domains b, c, y or z must be included, and thus, more domains and proteins are involved. Iterating this process, eventually all proteins and all domains are related and need to be taken into account. In addition, the association method ignores experimental errors (normally quite high in current experimental PPI sets) and treats the observed interactions as real interactions. This noise may lead to the impossibility of having a pattern of domain interactions that is compatible with the protein-protein interaction map.

To address the above two issues, Deng et al. [18] develop a global approach using a maximum likelihood estimation (MLE) method that incorporate all available proteins and domains, as well as experimental errors. They used yeast two-hybrid protein interaction data and treated protein sequences as “bags of domains.” The model estimates the probabilities of interactions between every pair of domains. Treating protein-protein interactions and domain-domain interactions as random variables, the two basic assumptions are (1) that two proteins interact if at least one pair of domains of the two proteins interacts and (2) interactions between different domain pairs are independent. Thus, the probability of a potential interaction between a protein pair (i, j) is

$$P(P_{ij} = 1) = 1 - \prod_{(d_m, d_n) \subset (P_i, P_j)} (1 - \lambda_{mn}) \tag{3}$$

where λ_{mn} denotes the probability that domain d_m interacts d_n. The expectation maximization (EM) algorithm is used to find maximum likelihood estimates of unknown parameters by finding the expectation of the complete data consisting of observed and unobserved data in two iterative steps. Here the observed data includes protein-protein interactions and the domain composition of the proteins, and the unobserved data includes all putative domain-domain interactions [83].

The above methods may preferentially identify promiscuous domain interactions, because they focus on those that occur with the highest frequency. Methods are need to detect the low-propensity, high-specificity domain interactions. Thus, Riley et al. [74] proposed the domain pair exclusion analysis (DPEA) method to extend the MLE approach. Riley et al. are specifically interested in extending beyond single proteome prediction to infer domain interactions from the incompletely mapped interactomes of multiple organisms. Their appoach employs a likelihood ratio test to assess the contribution of each potential domain interaction to the likelihood of a set of observed protein interactions from the incomplete interactomes of multiple organisms.

Similarly, Iqbal et al. [35] address the problem of predicting protein domain interactions by using belief propagation, which is a powerful message passing algorithm for probablistic inference. The input to their algorithm is an interaction map among a set of proteins, and a set of domain assignments to the relevant proteins. The output is a list of probabilities of interaction between each pair of domains. The method is able to effectively cope with errors in the protein-protein interaction dataset and systematically resolve contradictions.

Hypothesis test: Nye et al. [62] proposed a statistical method to test the null hypothesis that the presence of a particular domain pair in a protein pair has no effect on whether two proteins interact. The procedure calculates a statistic for each domain pair which takes into account experimental errors and the incompleteness of the dataset. The background distribution is simulated by shuffling domains in proteins so that the network of protein interactions remains fixed. The domain pair with the lowest p value is deemed most likely to interact. The authors point out that, for the majority of test cases, random domain prediction outperforms all methods tested, indicating the low accuracy of all prediction methods of domain interactions.

A set cover approach: Later, Huang et al. [33] proposed an interesting model to map the relationship between interactions of proteins and their corresponding domain architectures to a generalized set cover problem. Figure 13 gives a schematic explanation of the set cover approach. Set Y represents all potential protein pairs, and set X describes all known protein interaction pairs. $F = \{S_i, 1 \leq i \leq t\}$ is a family of subsets of Y. The general set cover problem is to find a subset C of F to cover X, such that $X \subseteq \cup_{S \in C} S$. Often, C is required to satisfy certain conditions. In this case, F is the set of all domain pairs (d_m, d_n). Specifically if a protein interaction pair (P_i, P_j) contains domain pair (d_m, d_n), then (P_i, P_j) belongs to the subset of (d_m, d_n). The goal is to find the collection C to cover X, where C is a subset of F and contains all the domain pairs present in the interaction network. The authors applied a greedy algorithm to identify sets of domain interactions which explain the presence of protein interactions to the largest degree of specificity. Using domain and protein interaction data from *S. cerevisiae*, they claim that this model enables prediction of previously unknown protein interactions.

Prediction with additional information: Recently, researchers started to combine PPIs with a variety of additional types of evidence to predict domain interac-

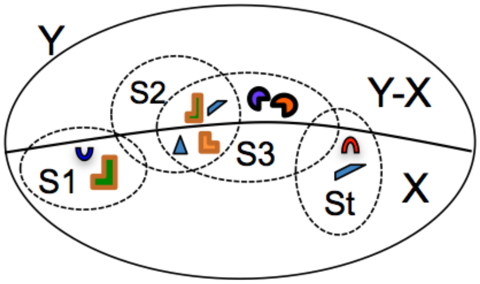


Fig. 13 A set cover approach to predict domain interactions from PPIs. Y set represents all potential protein pairs. X set includes all known protein interaction pairs.

tions. For example, Wang et al. [93] propose a learning method, called InSite, to predict specific regions (domains or motifs) where protein-protein interactions take place. The input includes a library of conserved sequence motifs or domains, a set of protein-protein interactions, and any available indirect evidence on protein-protein interactions and motif-motif interactions, such as expression correlation, gene functional annotation, and domain fusion. InSite makes predictions at the level of individual protein pairs, in a way that takes into consideration the various alternatives for explaining the binding between this particular protein pair. Specifically, this method integrates multiple biological data sets and generates predictions in the form of 'Motif Y on protein P2 binds to protein P5' (as shown in Figure 14). In contrast to previous methods, which predict bindings between pairs of motif types, InSite makes predictions of interactions of particular occurrences of two motifs. Thus, InSite may give the same motif pair different interaction confidences, depending upon the sequence context and the local neighborhood of the PPI network (Figure 14). This approach provides a principal way to integrate all available biological evidence. It also treat PPIs from multiple assays differently, since some of them are noisy and some are indirect.

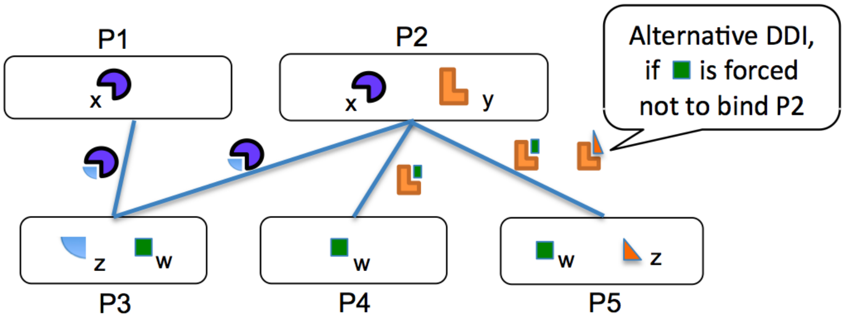


Fig. 14 Basic idea to predict protein interaction sites with the InSite method [93]. This figure is modified from Figure 1 in [93].

As above, we briefly discuss several important approaches to the task of identifying interacting and/or functionally linked domain pairs. These methods exhibit varying levels of success; however, they usually assume that domains interact independently, which is a limitation. Also part of the prediction errors come from incomplete domain assignments, insufficient coverage of domain databases and limited searching ability of domain profiles. In addition, domain interactions are predicted from protein interactions, whose available data is incomplete and noisy at the current stage [83].

There exist a number of important problems related to the domain-domain prediction from PPIs, including the interaction sites' prediction or the docking task. Since they are beyond the scope of this chapter, interested audience could refer to the review paper Zhou et al. [104] for the first task and Ritchie et al. [75] for understanding the second: docking problem.

4 Prediction of protein function from PPI networks

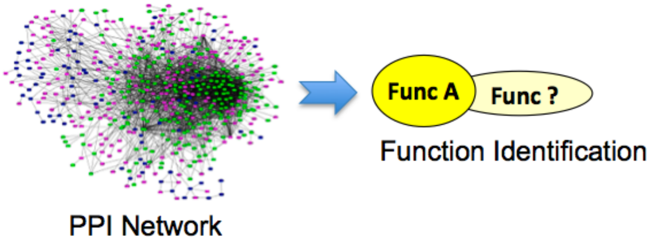


Fig. 15 Prediction of protein function from PPI networks

Proteins are involved in practically every function performed by a cell. However, despite the availability of large amounts of DNA and protein sequence data, the biological function is still unknown for a large proportion of sequenced proteins. Moreover, a given protein may have more than one function, so many proteins that are known to be in one functional class may have as yet undiscovered functionalities [98].

Inferences about function can be made via protein-protein interactions because protein interactions directly contribute to protein function. The premise is that the unknown function of a protein may be discovered through its interaction partners. Besides protein interaction evidence, the function of an unannotated protein can be predicted through various other data sets, including sequence homology, phylogenetic profiles, gene expression and so on. Combining multiple data sources together for protein function prediction is an interesting computational problem [66, 11, 90].

Here we focus on reviewing computational approaches that use protein-protein interaction evidence for protein function inference. It is worth mentioning that the interaction partners for a protein may belong to different functional categories. The problem of functional assignments in the complex protein network of within-function and cross-function interactions remains a difficult task [81].

Previous efforts in this area can be grouped into six categories, which are described in the following sections.

4.1 Simple Statistical Test

The basic assumption of functional annotation is that proteins which lie closer to one another in the PPI network are more likely to have similar functions. Thus, a simple statistical test can be used to assign functions to proteins based on the functions of their interaction partners.

For instance, Schwikowski et al. [78] proposed the neighborhood-counting method to assign k functions to a protein by identifying the k most frequent functional labels among its interacting partners. This strategy is simple and effective, but the full topology of the network is not taken into account in the annotation process, and no confidence scores are created for the annotations.

Another typical technique, referred to as the chi-square method [31], assigns k functions to a protein with the k largest chi-square scores. For a protein p, each function f is assigned a score $\frac{(n_f - e_f)^2}{e_f}$, where n_f is the number of proteins in the n-neighborhood of p that have the function f. The value e_f is the expectation of this number based on the frequency of f among all proteins in the network [81].

Recently Lee et al. [49] extended the neighborhood-counting [78] method to make network-based prediction of loss-of-function phenotypes in Caenorhabditis elegans. For a given phenotype, each gene in the worm proteome was ranked-ordered by the sum of its linkage weight (log-likelihood score of the gene interaction edge) to the "seed" set of genes already known to show that phenotype. The high-scoring genes are most likey to share the given phenotype.

In general, these simple methods lack a systematic mathematical model.

4.2 Graph Topoplogy

Researchers have also explored a variety of graph algorithms for protein functional inference [59, 91, 42]. For instance, Vazquez et al [91] and Karaoz et al. [42] exploit the global topological structure of the interaction network for functional annotation. The basic idea is described with a simple schematic example in Figure 16. This is a subgraph of the protein interaction network in the yeast *Saccharomyces cerevisiae*, with yellow nodes representing unannotated proteins and blue nodes representing annotated ones (the associated functions are listed as numbers in brackets adjacent to the nodes). Given one of these proteins with unknown functions, a simplified

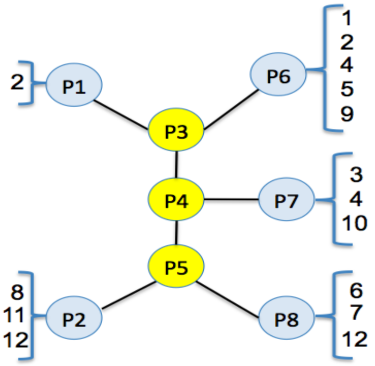


Fig. 16 Functional annotation from graph algorithm on PPI networks. Modified from Figure 1 in [91]. This shows a subgraph of the protein interaction network of the yeast Saccharomyces Cerevisiae. Proteins in yellow are unannotated (unknown function); the others are classified proteins (functions in brackets).

version of the method (proposed in [78]) would predict the function that appears most often in the neighbor proteins of known function. This approach would lead to the following classification result (from top to bottom): P3 (2), P4 (3,4,10) and P5 (12). By contrast, graph algorithms such as the one proposed by Vazquez et al [91], would also consider the interactions among unclassified proteins. Taking into account the interactions among the three unclassified proteins, one more iteration of the "majority rule" would lead to the following classification: P3 (2,4), P4 (3,4,10) and P5 (12). Thus, this extended method determined another possible function for P3.

The approach proposed in [91] assign proteins to functional classes so as to maximize the number of edges that connect proteins (unannotated or previously annotated) assigned with the same function. Precisely, they maximize

$$\sum_{(i,j)\in E'} \delta(\sigma_i,\sigma_j) + \sum_{i\in V} h_i(\sigma_i) \tag{4}$$

where E' is the set of edges between two unannotated proteins, δ is a function that equals 1 if $x = y$ and 0 otherwise, V is the set of nodes (proteins), and $h_i(f)$ denotes the number of neighbors of protein i previously annotated with function f. The first term in the optimization criterion accounts for unannotated proteins, whereas the second term concerns the interactions between unannotated and previously annotated proteins. This optimization problem can be generalized to the computationally hard problem of minimum multiway cut. The authors solved it heuristically using simulated annealing in [91].

Karaoz et al. [42] additionally consider the case where edges in physical interaction networks are weighted using gene expression data. The approach is a generalization of the well-studied multiway k-cut problem. The authors apply a local search strategy in which the state of the vertex is changed according to the majority of the states of its neighbors. Similarly, Nabieva et al. in [59] developed a network flow algorithm that exploits the underlying structure of protein interaction maps in

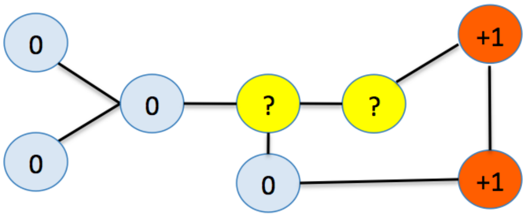


Fig. 17 A schematic illustration of the function prediction task on a protein network. Modified from Figure 1 in [90]. The task is to predict labels of unannotated proteins marked as ”?”. For a specific functions proteins having that function are labeled with ”1” or other wise ”0”.

order to predict protein function. Unlike [91, 42], this method takes advantages of both network topology and a particular measure of locality.

4.3 Graph Clustering

Clustering on protein interaction networks can also be used to predict protein function. For example, Samanta and Liang [77] proposed a network-based statistical measure to represent how many common partners two proteins share. They then use this statistic to hierarchically cluster the proteins in the PPI network. The key idea is that two proteins that share a large number of common partners likely have close functional associations. Arnau et al. [3] also applied hierarchical clustering in the protein-protein interaction network to find functionally consistent clusters. Their similarity measurement is derived from the shortest distance between two proteins in the network. Unlike typical graph clustering, Airoldi et al. explored a generative style of clustering [2]. The authors used a latent mixture membership approach to model the protein-protein interaction network. This approach transforms the function prediction objective into learning of the latent groups.

Sharan et al. [81] recently reviewed current computational approaches on functional annotation of proteins in the context of the protein interaction networks. They split the related papers into two types: (1) direct annotation schemes, which infer the function of a protein based on its connections in the network, and (2) module-assisted schemes, which first identify modules of related proteins and then annotate each module based on the known functions of its members. Methods we cover in other subsections belong to the “direct scheme” category. The current subsection only briefly introduces module-based (we call “graph clustering” based) methods which utilized the modularity assumption of PPI networks. There exist a number of ongoing work that explore this category of strategies for protein function annotation. Readers interested should refer to the overview paper [81] for details. Basically, such methods first attempt to identify coherent groups of genes and then assign functions to all genes in each group. The module-assisted methods differ mainly in their module detection techniques, which include graph clustering, hierarchical clustering, clustering based on network topology, etc. Once a module is obtained, simple methods are usually used for function prediction within the modules.

4.4 Probabilistic Propagation on Belief Networks

Although there exist multiple functional classes, we can approach the functional annotation task one fuction at a time. Figure 17 gives a schematic illustration of this case. For a certain functional class, the proteins assigned this function are labeled “1”. The proteins which are known to not have this function are labeled “0”. The remaining nodes are marked “?”. With this assignment, the protein-protein interaction graph in Figure 17 can be treated as a probabilistic belief network of function annotations. A number of probabilistic approaches to protein function prediction have been suggested. Most such approaches have relied on a Markovian assumption, namely, that the function of a protein is independent of all other proteins given the functions of its immediate neighbors [81] . This global approach takes all the network interactions and the functions of known proteins into consideration, propagating function labels from annotated proteins to unannotated proteins [19, 17, 50, 51].

The Markovian assumption naturally leads to a Markov random field (MRF) model, which was proposed by Deng et al. [19]. In this paper, an MRF was used to assign functions to unknown yeast proteins, with a probability representing the confidence in the prediction. Each protein node is assigned a random variable, with states corresponding to functional annotations in this setting. Thus, the interaction between two known proteins can be classified into one of the three groups: (1,1), (1,0) and (0,0), where numbers describe the involved proteins’ functional annotation. The joint belief can then be represented with a Gibbs distribution by considering the classification of all proteins,

$$Pr(X|PPInet) = \frac{exp[-U(x;\theta)]}{Z(\theta)} \tag{5}$$

where

$$U(x;\theta) = -(\alpha N_1 + \beta N_{11} + \gamma N_{10} + \kappa N_{00}) \tag{6}$$

$U(x;\theta)$ represents the potential function of the PPI network given a functional configuration of all proteins $X = (x_1,...,x_N)$ (discrete states). N_1 is the number of proteins for class “1,” and $N_{ll'}$ is the number of protein interactions between category l and l' in the network. $\theta = (\alpha, \beta, \gamma, \kappa)$ are parameters, where κ is set equal to 1. $Z(\theta)$ is the normalization constant (called the *partition function*), which is calculated by summing over all the configurations,

$$Z(\theta) = \sum_x exp[-U(x;\theta)] \tag{7}$$

Inference in this model is computationally hard. Deng et al. [19] use a quasi-likelihood method to estimate the parameters θ. The posterior probability that an unknown protein has the function of interest given the annotations of its neighbors $P(x_v = 1|x_{N(v)})$ was calculated with a Gibbs sampler.

Letovsky and Kasif [51] assumed a binomial model for local neighbors of a protein annotated with a given term. Also using the MRF propagation this algorithm assigns probabilities for proteins’ functional annotation in the network using loopy

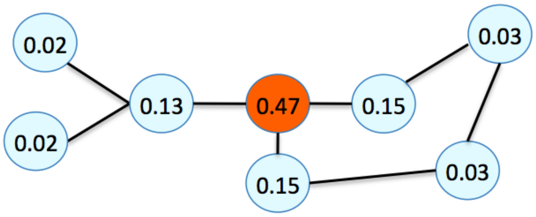


Fig. 18 Actual values of the diffusion kernel for one parameter setting of diffusion parameter β. Modified from Figure 2 in [90]. Each value on a node shows the kernel value between the node and the central node (orange node). The kernel values diffuse through the nodes on the graph.

belief propagation. Leone et al in [50] proposed a belief propagation method on PPI networks in a similar framework.

Later, Wu et al [94] proposed a related probabilisitic model to annotate functions of unknown proteins on PPI networks. Their model is an implicit MRF model that considers all the functions in a single model. This approach allows the model to capture correlations among protein functions. The authors used the conditional distribution and presented a maximum likelihood formulation of the problem. The time complexity of the corresponding learning and inference algorithms is linear in the size of the PPI network.

Mostafavi et al [58] adopted a variation of the Gaussian field label propagation algorithm for gene function prediction. Like the methods described above, this method assigns a score to each node in the network. This score reflects the estimated degree of association that the node has to the seed list defining the given function. The scores can be thresholded to make predictions. Unlike previous approaches using MRFs, the Gaussian field algorithm has a well-defined solution and can be efficiently computed.

4.5 Kernel Method

Kernel machines have been applied extensively for discovering functionally similar proteins within interaction networks. This approach has the ability to integrate multiple types of evidence for functional predictions. For instance, Lanckriet et al. [47] and later Tsuda et al. [90] represent each data type using a matrix of kernel similarity values. These matrices are then combined by learning optimal relative weights for the different kernels.

Here we briefly describe how protein-protein interaction data can be used by a kernel method [90]. Normally, a diffusion kernel [46, 84] is calculated on the graph of proteins connected by interactions. The diffusion kernel is a general method for computing pairwise distances among all nodes in a graph, based on the sum of weighted paths between each pair of nodes. Assume that A is the $n*n$ adjacency marix of a graph, and D is the $n*n$ diagonal matrix such that D_{ii} is the node degree of i-th node. The graph Laplacian matrix is defined as $L = D - A$. The diffusion kernel [46, 84] is then defined as

$$K = exp(-\beta L) \tag{8}$$

where the diffusion parameter $\beta > 0$ determines the degree of diffusion. This kernel can be interpreted in terms of a "lazy" random walk for sufficiently small β. At each step, the next node is randomly chosen from the neighbor nodes according to the transition probabilities. One can also stay at the same node (which is why the random walk is called "lazy". The kernel value K_{ij} is equivalent to the probability that a random walk starting from i will stay at j after infinite time steps. Figure 18 shows the actual values of diffusion kernels with one possible β. When β is large enough, the kernel values among distant nodes can capture the long-range relationships between proteins [90]. Diffusion kernels offer several benefits: (1) these kernels consider similarities among all protein pairs on the graph, not just immediate neighbors, (2) node degrees are taken into account in the kernel calculations, and (3) the parameter β is relatively easy to tune and has a clear meaning.

Lanckriet et al. [47] (and many others) used a diffusion kernel [46, 84] to summarize PPI graph evidence for functional predictions. Later, Tsuda and Noble [90] proposed a locally constrained variant of the diffusion kernel. They showed that computing the diffusion kernel is equivalent to maximizing the von Neumann entropy, subject to a global constraint on the sum of the Euclidean distances between nodes. This global constraint allows for high variance in the pairwise kernel distances. Thus, the authors proposed an alternative, locally constrained diffusion kernel and demonstrated that the resulting kernels allow for more accurate support vector machine predictions of protein functional classifications from the metabolic and protein-protein interaction networks.

4.6 Functional Identification Toward Annotation Taxonomy

The above two subsections handle the task of protein function prediction as multiple binary classications, where the methods treat each function at a time and make predictions for each term independently.

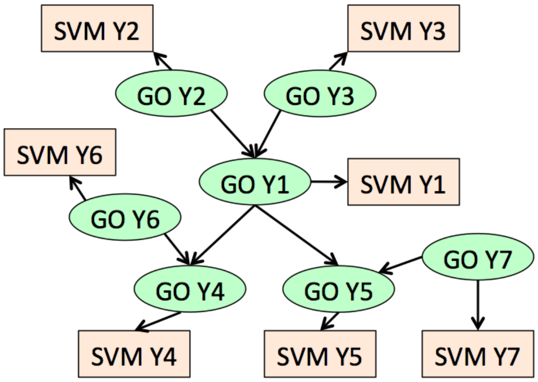


Fig. 19 A simple example of protein function identification considering the annotation taxonomy. Modified from Figure 2 in [29]. SVM classifier is represented with light red node and GO terms are described with green. Here single SVM classifiers (with one SVM per function term) were combined through Bayesian networks to correct their predictions based on the hierarchical relationship between GO [14] terms.

A more general approach to protein function prediction uses labels that follow a directed acyclic graph taxonomy as defined by the Gene Ontology (GO) [14]. The GO defines a set of terms to which any given protein may be annotated. In GO representation, the parent-child relationship among terms implies that the child term is either a special case of the parent term or describes a process or component that is part of the parent process/component. In either case, there is a clear directional dependency. Specifically, a protein positively annotated to a child term is, by definition, also positively annotated to the parent term(s), but not vice versa. As a logical consequence, a protein that is negatively annotated to a parent term is also negatively annotated to the child term(s). A negative annotation indicates that a protein has been experimentally verified not to be involved in a particular function.

Researchers proposed a variety of methods for systematically predicting protein function considering its taxonomy structures at the same time. Here we list three representative approaches as following:

Markov Random Field Extension: A MRF model was extended to chain graphs in [11] to directly incorporate the structure of the Gene Ontology into the graphical representation for protein classification. The authors presented a method in which each protein is represented by a replicate of the Gene Ontology structure, effectively modeling each protein in its own annotation space. Belief propagation was used to make predictions at all ontology terms.

Ensemble Framework: Guan et al. [29] describe an ensemble framework based on SVMs that considers correlation between multiple function terms (see Figure 19). A single SVM is used to predict a certain function for an unknown protein by integrating diverse datasets. In the context of the Gene Ontology hierarchy, single SVM classifiers are combined through Bayesian networks to correct their predictions based on the hierarchical relationship between GO terms in the GO directed acyclic graph. For each GO term, the method included all neighboring nodes in its Markov blanket to construct the Bayesian network. Shown in Figure 19, $Y1$ is the GO node of interest in this example. Thus this Bayesian network was constructed with the local Markov blanket surrounding $Y1$.

Reconciliation Method: Similar to the above paper, Obozinski et al. [63] proposed to predict GO terms using an ensemble of discriminative classifiers. This paper focused on *reconciliation* methods for combining independent predictions to obtain a set of probabilistic predictions that are consistent with the topology of the ontology. Eleven distinct reconciliation methods were investigated: three heuristic methods; four variants of a Bayesian network; an extension of logistic regression to the structured case; and three novel projection methods including isotonic regression and two variants of a Kullback-Leibler projection method. The authors found that many apparently reasonable reconciliation methods yield reconciled probabilities with significantly lower precision than the original, unreconciled estimates. On the other hand, the isotonic regression method seems to be able to use the constraints from the GO network to its advantage, usually performing better than the underlying, unreconciled predictions.

Recently, in a special issue of *Genome Biology*, several research groups [66] used GO annotation as a benchmark to compare methods of protein function predictions with GO hierarchy structure being considered. Readers could refer to [66] for more discussion.

5 Related General Topics

All sub-problems covered in this chapter are instances of more general tasks like "link prediction", "entity labeling", "structural output learning" or "graph mining" in the machine learning, data mining, and social network analysis communities. Methods proposed in related research fields have great potentials to be used for protein-protein interaction prediction, protein function identification or domain-domain interaction detection in the near future. As the literature on these topics is vast, this section will briefly discuss just a few related studies as a guide.

Statistical Relational Learning (SRL) As an area of growing interest in machine learning, statistical relational learning [25, 26] takes an object oriented approach to clearly distinguish between entities, relationships and their respective attributes in a probabilistic setting. Unlike most previous learning algorithms that assume all training examples are mutually independent, SRL methods try to capture complex relations among examples. A simple example of a relational system is a recommendation system: based on the attributes of two entities, i.e. of the user and the item, one wants to predict relationships like the preference (rating, willingness to purchase, ...) of this user for this item. One can exploit the known relationship attributes and the attributes of entities to predict unknown entity or relationship attributes [95]. This case is quite similar to protein-protein interaction prediction where we want to find the interaction preference of one protein to another. Various paradigms of SRL have been proposed in recent years, including probabilistic relational models, Bayesian logic programs, relational dependency networks, Markov logic networks, infinite relational model [44], infinite hidden relational model [95] and etc (surveyed in [60, 25, 26]). Several methods have software package available online, for instance, the open-source Alchemy system [45] provided a series of algorithms for statistical relational learning and probabilistic logic inference, based on the Markov logic representation [73]. It has been applied to problems in entity resolution, link prediction, information extraction and others [45].

Graph-Based Semi-Supervised Learning Semi-supervised learning (SSL) [12] occupies the middle ground, between supervised learning (in which all training examples are labeled) and unsupervised learning (in which no label data is given). In application domains where unlabeled data are plentiful, such as bioinformatics, SSL got growing interests in recent years. One category of SSL algorithms consider dependencies between the labels of nearby examples on a constructed graph [105, 9] to perform joint inference. These models train to encourage nearby data points to have the same class labels, which is exactly protein function detection aims for. The graph-based SSL can obtain impressive performance using a very small amount of labeled data [12]. As we know from above, for a large number of protein functional categories, there exist very few annotated genes from experimental tests. Graph-based SSL might make better functional predictions for these classes. Mostafavi et al. [58] made some attemps in this direction.

Mining of Entity-Relation Graphs In the data mining research community, relational or semi-structured data is naturally represented in a graph schema, where nodes denote entities and edges between nodes represent the relations between entities [22]. Such graphs are heterogeneous, since they include different types of nodes and different types of edges [57]. Many social networks could be described as entity-relation graphs. Using email system as an example, the graph inludes email-message, from-to-person, email-address and time entities which are inter-connected via relations derived from textual and structural information residing in a corporate database or a personal computer [57]. Similarly, protein interaction network could be converted to this schema easily where proteins, protein function annotations or domain compositions could be treated as different types of entities. Given an entity-relation graph, a popular question of interest is how to determine the nature of relationship between two entities that are not directly connected in the graph. The classical strategy [22] proposed in the literature performs random "lazy" graph walks on the entity-relation network to measure entity similarities. This strategy is closely related to graph-based SSL methods where "labels" (or "similarity") from a start node propogate through edges in the graph, e.g. ccumulating evidence of relatedness over multiple connecting paths. The problem of "entity proximity" has connections to all three tasks we covered in this chapter. For instance, protein function prediction could be treated ("implicitly") as a task of finding how similar an unknown protein is, to a known protein in terms of a specific functional category.

6 Summary

Biology relies on the concerted action of a number of biomolecules organized in networks, including proteins, small molecules, DNA and RNA. A key challenge is to understand the interactions among these molecules. The role of computational research on protein-protein interactions includes not only prediction, but also understanding the nature of the interactions and their binding residues on interaction interfaces. This chapter surveys recent efforts to predict interactions between proteins and between protein domains.

Predicting protein functions is one of the most important challenges of current computational biology research. A large number of computational techniques have been suggested for functional annotation using interaction networks; we have reviewed a few typical approaches in this chapter.

References

1. Abraham, I., Bartal, Y., Neimany, O.: Advances in metric embedding theory. In: STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing, pp. 271–286. ACM, New York, NY, USA (2006). DOI http://doi.acm.org/10.1145/1132516.1132557
2. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. **9**, 1981–2014 (2008)
3. Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics **21**(3), 364–78 (2005)
4. Bader, G.D., Hogue, C.W.: Analyzing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology **20**(10), 991–997 (2003)
5. Bader, J., Chaudhuri, A., Rothberg, J., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nature Biotechnology **22**(1), 78–85 (2004)
6. Ben-Hur, A., Noble, W.: Kernel methods for predicting protein-protein interactions. Bioinformatics (Proceedings of the Intelligent Systems for Molecular Biology Conference) **21**, i38–i46 (2005)
7. Ben-Hur, A., Noble, W.: Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics **20**(Suppl 1), S2 (2006)
8. Bleakley, K., Biau, G., Vert, J.P.: Supervised reconstruction of biological networks with local models. Bioinformatics **23**(13), i57–i65 (2007). DOI 10.1093/bioinformatics/btm204. URL http://dx.doi.org/10.1093/bioinformatics/btm204
9. Blum, A.: Semi-supervised learning using randomized mincuts. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning (2004)
10. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S., Venkatesan, K., Rual, J.F., Vandenhaute, J., Cusick, M.E., Pawson, T., Hill, D.E., Tavernier, J., Wrana, J.L., Roth, F.P., Vidal, M.: An experimentally derived confidence score for binary protein-protein interactions. Nat Methods **6**(1), 91–97 (2009). DOI 10.1038/nmeth.1281. URL http://dx.doi.org/10.1038/nmeth.1281
11. Carroll, S., Pavlovic, V.: Protein classification using probabilistic chain graphs and the gene ontology structure. Bioinformatics **22**, 1871–78 (2006)
12. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning (Adaptive Computation and Machine Learning). MIT Press (2006)
13. Chia, J.M., Kolatkar, P.R.: Implications for domain fusion protein-protein interactions based on structural information. BMC Bioinformatics **5**, 161 (2004)
14. Consortium, T.G.O.: Gene ontology: tool for the unification of biology. Nature Genet. **25**, 25–9 (2000)
15. Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P., Vidal, M.: Literature-curated protein interaction datasets. Nat Methods **6**(1), 39–46 (2009). DOI 10.1038/nmeth.1284. URL http://dx.doi.org/10.1038/nmeth.1284
16. Davis, F.P., Barkan, D.T., Eswar, N., McKerrow, J.H., Sali, A.: Host pathogen protein interactions predicted by comparative modeling. Protein Sci **16**(12), 2585–2596 (2007). DOI 10.1110/ps.073228407. URL http://dx.doi.org/10.1110/ps.073228407
17. Deng, M., Chen, T., Sun, F.: An integrated probabilistic model for functional prediction of proteins. J Comput Biol. **11**(2-3), 463–75 (2004)
18. Deng, M., Mehta, S., Sun, F., Chen, T.: Inferring domain-domain interactions from protein-protein interactions. Genome Res. **12**(10), 1540–8 (2002). Their method is actually an EM-based MLE
19. Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F.: Prediction of protein function using protein-protein interaction data. J Comput Biol. **10**(6), 947–60 (2003)
20. Dyer, M.D., Murali, T.M., Sobral, B.W.: Computational prediction of host-pathogen protein-protein interactions. Bioinformatics **23**(13), i159–i166 (2007). DOI 10.1093/bioinformatics/btm208. URL http://dx.doi.org/10.1093/bioinformatics/btm208

21. Espadaler, J., Romero-Isart, O., Jackson, R., Oliva, B.: Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics. **21**(16), 3360–8 (2005)
22. Faloutsos, C., Miller, G., Tsourakakis, C.: Large graph-mining: Power tools and a practitioner's guide. KDD 09 Tutorial (2009)
23. Gavin, A., Aloy, P., Grandi, P., et al., Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery. Nature **440**(7084), 631–6 (2006)
24. Gavin, A.C., Bosche, M., Krause, R., et al., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**(6868), 141–7 (2002). URL http://dx.doi.org/10.1038/415141a
25. Getoor, L., Diehl, C.: Link mining: A survey. SIGKDD Explorations **7**(2), 3–12 (2005)
26. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press (2007)
27. Gomez, S., Noble, W., Rzhetsky, A.: Learning to predict protein-protein interactions from protein sequences. Bioinformatics **19**(15), 1875–81 (2003). Protein-protein pair interaction probability takes as only the single most informative related domain-domain probability,
28. Gomez, S.M., Noble, W.S., Rzhetsky, A.: Learning to predict protein-protein interactions. Bioinformatics **19**, 1875–1881 (2003)
29. Guan, Y., Myers, C.L., Hess, D.C., Barutcuoglu, Z., Caudy, A.A., Troyanskaya, O.G.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biol **9 Suppl 1**, S3 (2008). DOI 10.1186/gb-2008-9-s1-s3. URL http://dx.doi.org/10.1186/gb-2008-9-s1-s3
30. Han, D., Kim, H.S., Seo, J., Jang, W.: A domain combination based probabilistic framework for protein-protein interaction prediction. Genome Inform **14**, 250–259 (2003)
31. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast **18**(6), 523–531 (2001). DOI 10.1002/yea.706. URL http://dx.doi.org/10.1002/yea.706
32. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., et al., Tyers, M.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415**(6868), 180–3 (2002). URL http://dx.doi.org/10.1038/415180a
33. Huang, C., Morcos, F., Kanaan, S.P., Wuchty, S., Chen, D.Z., Izaguirre, J.A.: Predicting protein-protein interactions from protein domains using a set cover approach. IEEE/ACM Trans Comput Biol Bioinform **4**(1), 78–87 (2007). DOI 10.1109/TCBB.2007.1001. URL http://dx.doi.org/10.1109/TCBB.2007.1001
34. Ingolfsson, H., Yona, G.: Protein domain prediction. Methods Mol Biol **426**, 117–143 (2008)
35. Iqbal, M., Freitas, A.A., Johnson, C.G., Vergassola, M.: Message-passing algorithms for the prediction of protein domain interactions from protein-protein interaction data. Bioinformatics **24**(18), 2064–2070 (2008). DOI 10.1093/bioinformatics/btn366
36. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast proteininteractome. Proc. Natl. Acad. Sci. USA **98**(8), 4569–4574 (2001). URL http://www.pnas.org/cgi/content/full/98/8/4569
37. Jaimovich, A., Elidan, G., Margalit, H., Friedman, N.: Towards an integrated protein-protein interaction network: a relational markov network approach. J Comput Biol. **13**(2), 145–64 (2006)
38. Jansen, R., Gerstein, M.: Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol. **7**, 535–45 (2004). Article
39. Jansen, R., Gerstein, M.: Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Current Opnion in Microbiology **7**, 535–545 (2004)
40. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**, 449–453 (2003)
41. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: String 8–a global view on proteins

and their functional interactions in 630 organisms. Nucleic Acids Res **37**(Database issue), D412–D416 (2009). DOI 10.1093/nar/gkn760. URL http://dx.doi.org/10.1093/nar/gkn760

42. Karaoz, U., Murali, T., Letovsky, S., Zheng, Y., Ding, C., Cantor, C., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA. **101**(9), 2888–93 (2004)
43. Kato, T., Tsuda, K., Asai, K.: Selective integration of multiple biological data for supervised network inference. Bioinformatics **21**(10), 2488–2495 (2005). DOI 10.1093/bioinformatics/bti339. URL http://dx.doi.org/10.1093/bioinformatics/bti339
44. Kemp, C., Tenenbaum, J.B.: Learning systems of concepts with an infinite relational model. In: In Proceedings of the 21st National Conference on Artificial Intelligence (2006)
45. Kok, S., Sumner, M., Richardson, M., Singla, P., Poon, H., Lowd, D., Domingos, P.: The alchemy system for statistical relational ai. Tech. rep., Department of Computer Science and Engineering, University of Washington (2007)
46. Kondor, R.I., Lafferty, J.D.: Diffusion kernels on graphs and other discrete input spaces. In: ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 315–322. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002)
47. Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., Noble, W.: Kernel-based data fusion and its application to protein function prediction in yeast. Pac Symp Biocomput. pp. 300–11 (2004)
48. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M.: A probabilistic functional network of yeast genes. Science **306**, 1555–1558 (2004)
49. Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., Marcotte, E.M.: A single gene network accurately predicts phenotypic effects of gene perturbation in caenorhabditis elegans. Nat Genet **40**(2), 181–188 (2008). DOI 10.1038/ng.2007.70. URL http://dx.doi.org/10.1038/ng.2007.70
50. Leone, M., Pagnani, A.: Predicting protein functions with message passing algorithms. Bioinformatics **21**(2), 239–47 (2004)
51. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics **19 Suppl 1**, I197–204 (2003)
52. Lin, N., Wu, B., Jansen, R., Gerstein, M., Zhao, H.: Information assessment on predicting protein-protein interactions. BMC Bioinformatics **5**, 154 (2004)
53. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. Science **285**, 751–753 (1999)
54. Martin, S., Roe, D., Faulon, J.L.: Predicting protein-protein interactions using signature products. Bioinformatics **21**(2), 218–226 (2005)
55. von Mering, C., Jensen, L., Snel, B., Hooper, S., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M., Bork, P.: STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. **33**, D433–7 (2005)
56. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**(6887), 399–403 (2002)
57. Minkov, E.: Adaptive graph walk based similarity measures in entity-relation graphs. Ph.D. thesis, Carnegie Mellon University (2008)
58. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q.: Genemania: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol **9 Suppl 1**, S4 (2008). DOI 10.1186/gb-2008-9-s1-s4. URL http://dx.doi.org/10.1186/gb-2008-9-s1-s4
59. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21**(S1), i302–10 (2005)
60. Neville, J., Rattigan, M., Jensen, D.: Statistical relational learning: Four claims and a survey. In: the Workshop on Learning Statistical Models from Relational Data, 18th International Joint Conference on Artificial Intelligence (2003)

61. Nguyen, T.P., Ho, T.B.: An integrative domain-based approach to predicting protein-protein interactions. Journal of Bioinformatics and Computational Biology **6**(6), 1115–1132 (2008)
62. Nye, T.M.W., Berzuini, C., Gilks, W.R., Babu, M.M., Teichmann, S.A.: Statistical analysis of domains in interacting protein pairs. Bioinformatics **21**(7), 993–1001 (2005). DOI 10.1093/bioinformatics/bti086. URL http://dx.doi.org/10.1093/bioinformatics/bti086
63. Obozinski, G., Lanckriet, G., Grant, C., Jordan, M.I., Noble, W.S.: Consistent probabilistic outputs for protein function prediction. Genome Biol **9 Suppl 1**, S6 (2008). DOI 10.1186/gb-2008-9-s1-s6. URL http://dx.doi.org/10.1186/gb-2008-9-s1-s6
64. Pagel, P., Strack, N., Oesterheld, M., Stmpflen, V., Frishman, D.: Computational prediction of domain interactions. Methods Mol Biol **396**, 3–15 (2007)
65. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96**(8), 4285–4288 (1999)
66. Pea-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., Krumpelman, C., Tian, W., Obozinski, G., Qi, Y., Mostafavi, S., Lin, G.N., Berriz, G.F., Gibbons, F.D., Lanckriet, G., Qiu, J., Grant, C., Barutcuoglu, Z., Hill, D.P., Warde-Farley, D., Grouios, C., Ray, D., Blake, J.A., Deng, M., Jordan, M.I., Noble, W.S., Morris, Q., Klein-Seetharaman, J., Bar-Joseph, Z., Chen, T., Sun, F., Troyanskaya, O.G., Marcotte, E.M., Xu, D., Hughes, T.R., Roth, F.P.: A critical assessment of mus musculus gene function prediction using integrated genomic evidence. Genome Biol **9 Suppl 1**, S2 (2008). DOI 10.1186/gb-2008-9-s1-s2. URL http://dx.doi.org/10.1186/gb-2008-9-s1-s2
67. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. PROTEINS: Structure, Function, and Bioinformatics. **63**(3), 490–500 (2006)
68. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random forest similarity for protein-protein interaction prediction from multiple sources. Pacific Symposium on Biocomputing **10**, 531–542 (2005)
69. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random forest similarity for protein-protein interaction prediction from multiple sources. In: Proceedings of the Pacific Symposium on Biocomputing (2005)
70. Qiu, J., Noble, W.S.: Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput Biol **4**(4), e1000,054 (2008). DOI 10.1371/journal.pcbi.1000054. URL http://dx.doi.org/10.1371/journal.pcbi.1000054
71. Ramani, A.K., Bunescu, R.C., Mooney, R.J., Marcotte, E.M.: Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol. **6**(5), R40 (2005). Article
72. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., Chinnaiyan, A.M.: Probabilistic model of the human protein-protein interaction network. Nat Biotechnol. **8**, 951–9 (2005). Article
73. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning **62**, 107–136 (2006)
74. Riley, R., Lee, C., Sabatti, C., Eisenberg, D.: Inferring protein domain interactions from databases of interacting proteins. Genome Biol **6**(10), R89 (2005). DOI 10.1186/gb-2005-6-10-r89. URL http://dx.doi.org/10.1186/gb-2005-6-10-r89
75. Ritchie, D.W.: Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci **9**(1), 1–15 (2008)
76. Rual, J.F., Venkatesan, K., et al., Roth, F.P., Vidal, M.: Towards a proteome-scale map of the human protein-protein interaction network. Nature **437**(7062), 1173–8 (2005). 1476-4687 (Electronic) Journal Article
77. Samanta, M., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA. **100**(22), 12,579–83 (2003)
78. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat Biotechnol **18**(12), 1257–1261 (2000). DOI 10.1038/82360. URL http://dx.doi.org/10.1038/82360

79. Scott, M.S., Barton, G.J.: Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics **8**, 239 (2007). 1471-2105 (Electronic) Comparative Study Journal Article Research Support, Non-U.S. Gov't
80. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965)
81. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Mol Syst Biol **3**, 88 (2007). 1744-4292 (Electronic) Journal Article Research Support, Non-U.S. Gov't Review
82. Shoemaker, B.A., Panchenko, A.R.: Deciphering protein-protein interactions. part i. experimental techniques and databases. PLoS Comput Biol **3**(3), e42 (2007). 1553-7358 (Electronic) Journal Article Research Support, N.I.H., Intramural Review
83. Shoemaker, B.A., Panchenko, A.R.: Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. PLoS Comput Biol **3**(4), e43 (2007). 1553-7358 (Electronic) Journal Article Research Support, N.I.H., Intramural Review
84. Smola, A., Kondor, R.: Kernels and regularization on graphs. In: B. Schölkopf, M. Warmuth (eds.) Proceedings of the Annual Conference on Computational Learning Theory and Kernel Workshop, Lecture Notes in Computer Science. Springer (2003)
85. Sontag, D., Singh, R., Berger, B.: Probabilistic modeling of systematic errors in two-hybrid experiments. Pac Symp Biocomput pp. 445–457 (2007)
86. Sprinzak, E., Margalit., H.: Correlated sequence-signatures as markers of protein-protein interaction. Journal of Molecular Biology **311**, 681692 (2001). Use mutual information (average) of two sequence signatures in the interacting protein pairs as signature interact probability; InterPro =¿ sequence signature of protein
87. Stelzl, U., Worm, U., Lalowski, M., et al., Wanker, E.E.: A human protein-protein interaction network: a resource for annotating the proteome. Cell **122**(6), 957–68 (2005). 0092-8674 (Print) Journal Article
88. Tastan, O., Qi, Y., Carbonell, J., Klein-Seetharaman, J.: Prediction of interactions between hiv-1 and human proteins by information integration. Pacific Symposium on Biocomputing (PSB) **14** (2009)
89. Teichmann, S.A.: Principles of protein-protein interactions. Bioinformatics **18 Suppl 2**, S249 (2002)
90. Tsuda, K., Noble, W.: Learning kernels from biological networks by maximizing entropy. Bioinformatics **20 Suppl 1**, I326–I333 (2004)
91. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nature Biotechnology **21**, 697 – 700 (2003)
92. Vert, J.P., Qiu, J., Noble, W.S.: A new pairwise kernel for biological network inference with support vector machines. BMC Bioinformatics **8(Suppl 10):S8** (2007)
93. Wang, H., Segal, E., Ben-Hur, A., Li, Q., Vidal, M., Koller, D.: InSite: A computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. Genome Biology **8**(9), R192.1–R192.18. (2007)
94. Wu, Y., Lonardi, S.: A linear-time algorithm for predicting functional annotations from ppi networks. J Bioinform Comput Biol **6**(6), 1049–1065 (2008)
95. Xu, Z., Tresp, V., Yu, K., Kriegel, H.P.: Infinite hidden relational models. In: Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence (UAI 2006) (2006)
96. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13), i232–i240 (2008). DOI 10.1093/bioinformatics/btn162. URL http://dx.doi.org/10.1093/bioinformatics/btn162
97. Yamanishi, Y., Vert, J.P., Kanehisa, M.: Protein network inference from multiple genomic data: a supervised approach. Bioinformatics **20 Suppl 1**, i363–i370 (2004). DOI 10.1093/bioinformatics/bth910. URL http://dx.doi.org/10.1093/bioinformatics/bth910
98. Yanay, O., Marco, P., Burkard, R.: Tutorial: Function prediction - from high throughput to individual proteins. Pacific Symposium on Biocomputing (2005)

99. Yip, K.Y., Gerstein, M.: Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. Bioinformatics **25**(2), 243–250 (2009). DOI 10.1093/bioinformatics/btn602. URL http://dx.doi.org/10.1093/bioinformatics/btn602
100. Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabsi, A.L., Tavernier, J., Hill, D.E., Vidal, M.: High-quality binary protein interaction map of the yeast interactome network. Science **322**(5898), 104–110 (2008). DOI 10.1126/science.1158684. URL http://dx.doi.org/10.1126/science.1158684
101. Yu, J., Finley, R.L.: Combining multiple positive training sets to generate confidence scores for protein-protein interactions. Bioinformatics **25**(1), 105–111 (2009). DOI 10.1093/bioinformatics/btn597. URL http://dx.doi.org/10.1093/bioinformatics/btn597
102. Zhang, L., Wong, S., King, O., Roth, F.: Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics **5**, 38 (2004)
103. Zhang, L.V., Wong, S., King, O., Roth, F.: Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics **5**(1), 38–53 (2004)
104. Zhou, H.X., Qin, S.: Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics **23**(17), 2203–2209 (2007). DOI 10.1093/bioinformatics/btm323. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/17/2203
105. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML'03: Proceedings of the 20th International Conference on Machine Learning, pp. 912–919 (2003)