Protein Cell 2012, 3(7): 508–520
DOI 10.1007/s13238-012-2945-1

Protein & Cell

REVIEW

Proteome-wide prediction of protein-protein interactions from high-throughput data

Zhi-Ping Liu[1] ✉, **Luonan Chen**[1,2] ✉

[1] Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
✉ Correspondence: zpliu@sibs.ac.cn (Z.-P. Liu), lnchen@sibs.ac.cn (L. Chen)
Received May 12, 2012 Accepted May 30, 2012

ABSTRACT

In this paper, we present a brief review of the existing computational methods for predicting proteome-wide protein-protein interaction networks from high-throughput data. The availability of various types of omics data provides great opportunity and also unprecedented challenge to infer the interactome in cells. Reconstructing the interactome or interaction network is a crucial step for studying the functional relationship among proteins and the involved biological processes. The protein interaction network will provide valuable resources and alternatives to decipher the mechanisms of these functionally interacting elements as well as the running system of cellular operations. In this paper, we describe the main steps of predicting protein-protein interaction networks and categorize the available approaches to couple the physical and functional linkages. The future topics and the analyses beyond prediction are also discussed and concluded.

KEYWORDS proteomics, protein-protein interaction, prediction, systems biology

INTRODUCTION

Protein always performs its biological functions by interacting with other proteins and molecules (Eisenberg et al., 2000; Chen et al., 2009, 2010). Many fundamental and essential biological processes, such as signal transduction, transport, DNA regulatory and alternative splicing, are involved in or mediated by protein-protein interactions (PPIs) (Eisenberg et al., 2000; Barabasi and Oltvai, 2004). It is crucial to build a proteome-wide protein interaction network of one organism for studying its biological functions. There are some traditionally experimental approaches such as co-immunoprecipitation, and newly developed high-throughput techniques such as yeast two-hybrid screening and the combination of large-scale affinity purification with mass spectrometry, to detect protein interactions (Bork et al., 2004). Due to their biological importance, there have been some well-known databases constructed by collecting the available information of protein interactions, such as the reported interactions detected by classical experiments and by the former high-throughput techniques (Bork et al., 2004; Stark et al., 2006). These interactions are currently collected together in some specialized databases for further study and investigation, such as HPRD for human PPIs (Prasad et al., 2009) and IntAct (Aranda et al., 2009) for PPIs of many species. To detect protein interactions, each of the experimental methods has its own advantages and weakness mainly due to high false positive ratio (Valencia and Pazos, 2002). It is still labor tensing and time consuming for detecting the PPIs by these experiments. Moreover, the interactions among proteins are highly related to the environmental conditions and the dynamics of cellular processes (Han et al., 2004a; Bossi and Lehner, 2009). Because of the complexity of massive interactions and the difficulty of detecting them by the traditional methods, it is urgent to develop novel methods to predict protein interactions precisely (Bork et al., 2004; Chen et al., 2009). The computational methods provide promising alternatives for screening and further identifying the relationship between these macromolecules and building their full interaction map. Actually, the emergence of genomics, transcriptomics, proteomics, and metabolomics resources offers new opportunity to infer the protein interaction maps of various species from these high-throughput data (Jansen et al., 2003; Chen et al., 2009, 2010). The inferred interaction networks

are expected to accelerate the interactome research for each particular organism and further lead to a comprehensive understanding of its biological processes.

The prediction of PPIs is based on a general assumption that protein interaction is involved in basic principles and evolutionary implications (Valencia and Pazos, 2002). The rules of how two proteins interact can be extended to other proteins, and the interaction can also be inferred across species. The interaction patterns of known interacting protein pairs are identified, and implemented to predict unknown protein pairs. The conserved interaction mechanisms transferred in one organism can also be detected in other organisms (Bork et al., 2004). Prediction of protein interaction is based on the genomic features representing the two proteins. There are numerous methods have been proposed to predict the interactions between proteins (Valencia and Pazos, 2002; Szilagyi et al., 2005; Skrabanek et al., 2008). We can categorize them into several groups based on their learning processes for interacting features. The genomic features learned from these interacting proteins provide the principles for prediction, such as gene neighborhood (Dandekar et al., 1998), gene fusion (Enright et al., 1999), subcellular localization (Yu et al., 2004; Lee et al., 2008), similarity of evolutionary tracing (Pazos and Valencia, 2001, 2002), gene co-expression (Jansen et al., 2002), and docking complementarily (Aytuna et al., 2005). These features are hypothesized to contribute and determine the events of protein interactions. They are defined and learned from various aspects of genomic perspectives, e.g., sequence, structure, physicochemical attributes as well as evolutionary tracing (Valencia and Pazos, 2002; Jansen et al., 2003). Different methods have succeeded in their own fields and cases and they also provided new alternatives to describe the interacting proteins by encoding them into genomic features.

In this paper, we provide a brief review on the existing methods for inferring protein interaction networks from high-throughput data. Firstly, we introduce a general framework of prediction. We categorize the available methods into two groups, i.e., direct mapping of associated features or elements, and indirect coupling of supervised learning. Then, we summarize the main methods to address the future challenges underlying the prediction. The built interaction maps reflect the framework of performing biological function of thousands of proteins. We also address the advantage in these methods for predicting protein interactions. Last but not least, we provide several popular and future topics regarding to functional analysis.

FRAMEWORK OF PREDICTION

Cell comprises thousands of proteins, which always perform their functions through interacting with each other (Eisenberg et al., 2000). Network presents a powerful model to formulate their complicated relationships which are responsible for cellular functions of collaborative effects of those individual components (Barabasi and Oltvai, 2004). Mathematically, given a network or graph G = (*V*, *E*), where *V* is the node set and *E* is the edge set (Chen et al., 2009), we can represent protein interactions as a protein-protein interaction (PPI) network. The nodes in the PPI network are the interacting proteins and the edges refer to their interactions. To predict proteome-wide PPIs is to infer the interactions between these proteins in a large scale manner, i.e., prediction of *E* for all nodes in *V*. The assumption of prediction is that protein interactions are conserved across different proteins and species (Valencia and Pazos, 2002). Generally, we first extract the knowledge of how proteins are interacted, and then extend to determine which proteins will interact with each other. The framework of predicting PPIs is based on the extracted and learned genomic features underlying the known PPIs.

Substantial efforts have been taken to infer proteome-wide interaction maps in various species (Valencia and Pazos, 2002; Szilagyi et al., 2005; Skrabanek et al., 2008). We can briefly categorize available methods into two major groups by the processes of learning the genomic features from known PPIs. As shown in Fig. 1A, the first group is to directly map the genomic features underlying these known interacting protein pairs into those of predicting ones. Suppose that protein *A* interacts with protein *B* and we want to predict the interaction status between protein *X* and protein *Y*. The direct mapping methods identify the similarity information between the two pairs of proteins by comparisons. For example, *A* is a homolog of *Y* and *B* is a homolog of *X*, then we predict that *X* interacts with *Y*. The similarities of sequence, structure, interface shape, gene expression, subcellular localization, evolutionary information as well as other genomic features have been adopted in these interacting protein pairs and the predicting protein pairs individually (Skrabanek et al., 2008). These elements and features are compared and implemented to predict whether the interaction exists in the predicting protein pairs or not. Clear, the first group directly translocates the information of interacting proteins into the targeted protein pairs. However, the features contributing for protein interaction events are often not identified directly in this prediction. In other words, they can be the knowledge of genomic features about PPIs, but are not necessarily identified from these known interactions. The features detected in the predicting protein pairs are used to predict the existence of interactions (Yu et al., 2004; Lu et al., 2005). The methods in the first group explicitly model the mapping between unknown protein-protein pairs and features extracted from known PPIs. In learning theory, methods in the first group are often based on unsupervised learning processes (Vapnik et al., 1995). The features of known PPIs are mapped into unknown protein pairs by trackable ‘write-box’ process. That is to say, we can track the similar associated elements or features in the predicted interaction as that of known PPI.

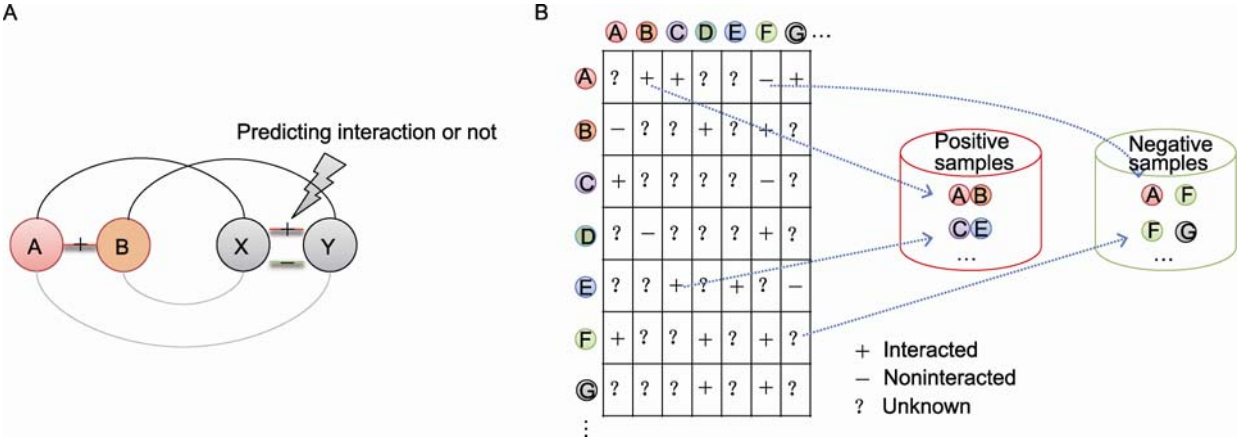

Figure 1. Framework of predicting PPIs. (A) Mapping the interacting proteins to the predicting pairs by directly comparing their similarities (the first group); (B) Constructing positive and negative samples to learn the interaction features for prediction (the second group).

The second group contains a supervised machine learning process which is often based on a classification algorithm. Compared to the methods in the first group, the methods in the second group do not provide the details of relationship between PPIs and individual features. They are 'black-box' learning processes. From the known interactions among some proteins, these methods construct positive samples and negative samples individually. The genomic features of these proteins are identified and encoded into feature vectors (Lu et al., 2005). The labels of interaction or non-interaction are used to supervise the learning algorithm so as to generate the correct predictions (Liu et al., 2012b). The classifier is trained by learning the features of interacting proteins as well as those of non-interacting proteins on the training dataset. Then, the targeted pairs of proteins are predicted as interacted or non-interacted by the trained classifier. As shown in Fig. 1B, the training dataset collects the positive samples referred to these known interactions and negative samples to these non-interactions. Unfortunately, so far there has been a far fewer number of the reported negative protein interactions (Smialowski et al., 2009) in contrast to the positive protein interactions. Thus, it is often to sample randomly some interactions from these unknown interactions between proteins (shown in Fig. 1) and regard them as the negative samples. Some other techniques have been proposed to generate the negative samples (Smialowski et al., 2009; Liu et al., 2012b). For instances, some methods generate the negative samples based on the subcellular location of proteins because proteins in different locations will have low possibility to interact with each other (Lu et al., 2005). Some methods identify the negative samples based on the network theory of six degrees of separation (Chen et al., 2009), i.e., when the shortest distance between two proteins in the network of known interactions is more than six, the two proteins will have weak possibility of interacting with each other (Liu et al., 2012b).

Essentially, predicting a PPI is to make a decision of whether or not there is an interaction in the two proteins. Generally, we predict the interaction by learning the knowledge in these known interacting protein pairs. The genomic information of the two interacting proteins is identified and then compared with those in the predicting protein pairs (Yu et al., 2004). The first group is based on 'write-box' learning, which maps the information of interacting proteins into these unknown pairs by the analog genomic features between the two pairs. The knowledge has directly been implemented to make the prediction. On the other hand, the second group is based on 'black-box' learning. The genomic features of both positive samples and negative samples are encoded into feature vectors and they are implemented to train an algorithm of classification. These machine learning algorithms often transform and reorganize these features underlying these interactions. There are no direct correspondences between features and interactions. As a result, the genomic features are identified and further used indirectly to predict PPIs.

Obviously, the gold standard dataset of protein interactions are very important for the prediction, and also highly affect the effectiveness and efficiency in both kinds of methods (Yu et al., 2004; Lu et al., 2005). The features underlying the interactions in the gold standard dataset are identified, learned and mapped into predicting proteins directly or indirectly. The experimental protein interactions are often selected as the gold standard dataset for learning and testing. Table 1 shows some widely-used available PPI databases for various species. These documented interactions are often selected as positive samples in the gold standard dataset. On the other hand, the collected and designed negative samples are also important for the supervisory design of output in the second group of machine learning algorithms. Generally, these methods all adopt a cross validation process for testing

Table 1. Some available PPI databases

Database	Website	Reference
BioCarta	http://www.biocarta.com/	Biocarta, 2012
BioGrid	http://thebiogrid.org/	Stark et al., 2006
BIND	http://www.bind.ca/	Bader et al., 2003
DIP	http://dip.doe-mbi.ucla.edu/	Salwinski et al., 2004
HPID	http://wilab.inha.ac.kr/hpid/	Han et al., 2004b
HPRD	http://www.hprd.org/	Prasad et al., 2009
I2D	http://ophid.utoronto.ca/	Brown and Jurisica, 2007
IntAct	http://www.ebi.ac.uk/intact/	Aranda et al., 2009
KEGG	http://www.genome.jp/kegg/	Kanehisa et al., 2000
MINT	http://mint.bio.uniroma2.it/	Chatr-aryamontri et al., 2007
MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/	Pagel et al., 2005
Reactome	http://www.reactome.org/	Vastrik et al., 2007
STRING	http://string.embl.de/	Mering et al., 2007

the prediction performance in the gold standard dataset. There are typical measures defined to evaluate the prediction results, e.g., sensitivity (SN), specificity (SP), accuracy (ACC), F-measure, and Matthews correlation coefficient (MCC) (Liu et al., 2010), shown in the following equations:

$$SN = \frac{TP}{TP + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN},$$

$$F\text{-}measure = \frac{2 \times SN \times SP}{SN + SP},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}},$$

where TP, FN, FP and TN are the numbers of true positive, false negative, false positive and true negative protein interactions in the prediction, respectively. The tradeoff of specificity and sensitivity are often presented by the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is also calculated.

THE EXISTING METHODS OF PREDICTION

A number of computational methods have been proposed to address the problem of predicting PPIs in the framework shown in Fig. 1. Table 2 lists some representative methods. Next, we describe their computational procedures of predicting PPIs from high-throughput data. The genomic knowledge found in these interacting protein pairs will be checked in the predicting ones. Based on our categories, the direct mapping methods identify and employ the genomic features underlying the known proteins interactions. These methods map the interaction to the predicting protein pairs by comparing their corresponding features. In these write-box approaches, the features or elements identified in the known PPIs are mapped directly and clearly into the predicting protein pairs. We will highlight the domain association methods based on the domain information in the proteins, which are regarded to be the functional units mediating PPIs. As to the black-box approaches, the supervised machine learning methods will be also introduced in details.

Direct mapping methods

One of the first ideas for predicting PPIs is to identify the genomic contexts highly related to interaction events in the predicting proteins (Valencia and Pazos, 2002). For instance, it is reported that genes which interact physically or functionally will be kept in close physical proximity to each other on the genome (Tamames et al., 1997; Dandekar et al., 1998). The knowledge of gene neighborhood can then be detected to infer the interaction of these predicting proteins.

Gene-neighborhood based method

This method is based on the assumption that genes close in the genome tend to encode functionally related proteins. The neighborhood relationship tends to be more relevant when it is conserved across multiple genomes (Valencia and Pazos, 2002). Co-localization of genes across the genomes often indicates their encoded proteins physically interact, especially in some species such as bacteria (Tamames et al., 1997; Overbeek et al., 1999). The PPIs can be predicted by identifying the contiguity of genes on the chromosome.

Gene fusion based method

Gene fusion is referred to the event of proteins in one organ-

Table 2. Some available methods of predicting PPI networks

Method	Model organism	Major method
Overbeek et al., 1999	Multiple organisms	Gene neighborhood
Enright et al., 1999	Three organisms	Gene fusion
Pazos and Valencia., 2001	*Escherichia coli*	Phylogenetic tree
Pazos and Valencia., 2002	*Escherichia coli*	Phylogenetic profile
Yu et al., 2004	Multiple organisms	Interolog
Jansen et al., 2002	*Saccharomyces cerevisiae*	Co-expression
Aytuna et al., 2005	Multiple species	Docking
Chen et al., 2006	*Saccharomyces cerevisiae*	Domain association
Jansen et al., 2003	*Saccharomyces cerevisiae*	Bayesian integration
Andres et al., 2009	*Escherichia coli*	Five prediction methods
Zhao et al., 2009	*Fusarium graminearum*	Interolog and domain association method
Sapkota et al., 2011	Oryza sativa	Domain association and SVM-based method
Liu et al., 2012b	*Mycobacterium tuberculosis*	Interolog and SVM-based method

ism which have homologs in another genome fused into a single protein (Enright et al., 1999). The highly related gene relationship implies the physical interaction between their corresponding proteins. As to the frequency of gene fusion, Huynen et al. (2000) presented that most of the physical interactions contain the gene fusion events. Tsoka and Ouzounis (2000) reported that metabolic enzymes are frequently involved in gene fusion. The approach based on gene fusion for predicting protein interactions is limited to the shared domains in distinct proteins (Skrabanek et al., 2008). In prokaryotic organisms, its true extent is still unclear (Valencia and Pazos, 2002).

Phylogeny based method

Multiple sequence alignment is effective to grasp the interacting features underlying several protein pairs simultaneously. It is also basically implemented in the methods based on gene fusion and gene neighborhood. BLAST (Altschul et al., 1997) and Clustal series of programs (Chenna et al., 2003) are widely used to detect the sequence similarities. By multiple sequence alignments, the phylogenetic trees of these analyzed proteins can be built. The similarities of interacting proteins are higher than those of noninteracting proteins (Valencia and Pazos, 2002). By analyzing the phylogenetic trees, the co-evolution features of interacting proteins are mapped into the predicting protein pairs (Jothi et al., 2005). Furthermore, the correlation of mutation of interacting protein pairs is higher than that of noninteracting proteins. From multiple sequence alignments, the correlated mutations of intraprotein and interprotein can be identified individually (Valencia and Pazos, 2002). An interaction index can be obtained by calculating the interprotein correlations and with the two intraprotein correlations. The evolutionary information detected by building the phylogenetic tree can be employed to predict the interaction event between proteins (Valencia and Pazos, 2002). Goh et al. (2000) developed a standard measure for detecting the co-evolution of interacting proteins in the phosphoglycerate kinases. The results provided evidence for the efficiency of coupling protein interaction relationship by building the phylogenetic tree. Pazos and Valencia (2001) proposed such a method to predict a large scale PPI network based on the evolutionary distances between the sequences of the associated protein families. Correlated mutations were also used for predicting PPIs (Gobel et al., 1994; Pazos and Valencia, 2002).

The other phylogeny based method is to identify the phylogenetic profile of gene co-occurrences in multiple species. The underlying assumption of phylogenetic profile based method is that functionally related proteins are co-occurred in their evolutionary profiles (Valencia and Pazos, 2002). Each targeted protein can be represented by a binary vector of phylogenetic profile. The vector indicates the conservation status of the represented protein across multiple species (Pazos and Valencia, 2002; Valencia and Pazos, 2002). Often, the proteins with similar profiles are predicted to be interacting protein pairs. Pellegrini et al. (1999) characterized the correlated proteins of one interaction pair by its phylogenetic profile. The proteins with similar profiles have been proved to be functionally linked (Valencia and Pazos, 2002). Wu et al. (2003) extended this method to identify functional linkages between genes by using phylogenetic profiles.

Interolog based method

An interolog is referred to an interaction between a pair of proteins which have interacting homologs based on conser-

vation assumption in and across organisms (Walhout et al., 2000; Yu et al., 2004). Interolog based method of predicting PPIs is to check the homologies between the analyzing proteins and the interacting proteins. Suppose protein *A* and protein *B* are two interacting proteins, and protein *A′* and *B′* are two other proteins. If protein *A* is a homolog of protein *A′* and protein *B* is a homolog of *B′*, the method will predict that there is an interaction between protein *A′* and protein *B′*. The two proteins will be predicted as interacting protein pairs by mapping the homolog information of these existing protein interactions. Walhout et al. (2000) introduced the concept of interologs to be orthologous pairs of interacting proteins in different organisms. Yu et al. (2004) present a large-scale quantitative assessment on the conservation of PPIs between proteins and organisms. Based on the interaction information from four species, i.e., *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Helicobacter pylori*, they verified that PPIs can be transferred when a pair of proteins has a high joint sequence identity. This provided direct evidence for the effectiveness of interolog based method for predicting PPIs.

Co-expression based method

It was reported that proteins in the same complex tend to be co-expressed (Ge et al., 2001). That is to say, proteins with this type of physical interactions often have relevant concentration of gene expression. The principle sheds light on the feasibility of predicting PPIs by the co-expression information between their corresponding genes. There are huge amount of gene expression profiling data stocked in various databases such as GEO (Barrett et al., 2007), and there are often several replicated samples for one experiment. This indicates that we can predict the proteome-wide PPIs by calculating the correlations of gene expression. The high co-expression value between genes indicates the high possibility of interactions between their downstream proteins. Compared with random pairs in yeast, Grigoriev (2001) concluded that the encoded proteins by these co-expressed genes interact with each other more frequently. Moreover, the interacting proteins also tend to be located in the same cluster of gene expression (Jansen et al., 2002). Bhardwaj and Lu (2005) investigated the global relationship of protein interactions with gene expression within and across four evolutionary diverse species. By comparison, they identified that the co-expression of interacting proteins is more conserved than that of random ones. From the importance of evolutionary information during the prediction, they improved the accuracy of predicting PPI by integrating the ortholog information in the correlation calculation (Bhardwaj and Lu, 2005). Obviously, the former methods can be simultaneously implemented to predict proteome-wide PPIs in various species. STRING provides the results of such integrative predictions (Mering et al., 2007).

Docking based method

The methods reviewed in the previous subsections are based only on sequence-related genomic information, while docking based method is a structure based method of predicting PPIs. The docking based method will be developed gradually with the availability of more three-dimensional protein structures. Structure based docking infers not only whether the proteins interact, but also which residues on the protein surfaces interact with each other (Zhou and Shan, 2001; Smith and Sternberg, 2002). The method is to analyze the docking principle between proteins and extend their interacting features to other proteins. The structure complementarity of protein surfaces is a primary principle to analyze the docking (Smith and Sternberg, 2002; Aytuna et al., 2005). The shapes of interacting places are identified and mapped into the similar faces of two proteins. The method computationally represents the protein surface into feature vectors. The choice of representing the protein surface is to encode the structural features of docking between proteins. Some complementary features are also defined to describe the interfaces from electrostatic and hydrophobic (Smith and Sternberg, 2002; Aytuna et al., 2005). The structure features are often combined with machine learning methods, which we regarded as the second major categorized approaches.

It is convenient for inferring a protein interaction map from sequence and genome analysis because they are easily available. On the other hand, to an accurate and detailed understanding of PPI, the structure based methods have the advantage to decipher the interaction mechanisms at the atomic level. The binding residues, the interacting atoms as well as the binding energy of local structures can be analyzed and investigated. With more three-dimensional structures are crystallized, the structure based methods for predicting PPIs will become more and more popular (Aytuna et al., 2005). In particular, there will be more structure templates for building the pool of binding pockets for the prediction (Zhou and Shan, 2001; Valencia and Pazos, 2002).

SCOPPI (Structural Classification of Protein-Protein Interfaces) categorized the types of protein-protein interface from the structural perspective according to the geometry of these interacting domains (Winter et al., 2006). We also quantitatively accessed the ability of predicting protein functions from their local structures of pockets (Liu et al., 2007). These results provided direct evidence for the importance of functional specificities underlying the protein local structures of docking events. Predicting PPIs from docking structures will achieve high accuracy and determine the specific binding structure features. Certain types of residues in protein surface have a major contribution for protein interaction, which are often called ‘hot spots’ (Smith and Sternberg, 2002). It is originated from a binding energy concept and the prior knowledge about them can facilitate the prediction of protein interaction (Szilagyi et al., 2005; Skrabanek et al., 2008). Hot

spots in protein interface also provide crucial information for drug design. In this area, we provided a novel random forest model to identify the hot spots in proteins by extracting hybrid features which incorporate a wide range of information of the target residue and its spatially neighboring residues (Wang et al., 2012).

Domain association methods

Domain is a part of protein sequence and structure which is the basic functional unit of protein (Chen et al., 2006; Wang et al., 2007; Zhao et al., 2010). It is generally believed that two proteins interact with each other if a domain in one protein interacts with a domain in the other protein. Information of domain-domain interaction benefits the detailed understanding of PPI. The association between these domains in the known PPIs can be extended into the predicting protein pairs, i.e., prediction of PPI from domain association (Chen et al., 2006, 2009).

Fig. 2 shows the framework of predicting PPI by associating the domain interactions (Chen et al., 2006, 2009). Firstly, the domains in the interacting proteins are identified (shown in Fig. 2A). Then, as shown in Fig. 2B, the associations between these identified domains are inferred. The domain-domain interaction rules are also learned from these interaction pairs of proteins and domains. The unknown protein pairs are predicted by their involved domains with the learned rules of domain interactions for contributing protein interactions as shown in Fig. 2C. Assume that there are N proteins indicated by $P_1,\cdots,P_N$ and M domains in the proteins represented by $D_1,\cdots,D_M$. Let P_i also denote a set of domains in the protein P_i. A protein P_i may include multiple domains D_j. Let P_{ij} and D_{mn} represent the protein pair (P_i, P_j) and the domain pair (D_m, D_n), respectively. P_{ij} is also used to represent a set of domain pairs in P_i and P_j, i.e., $\{D_{mn} \mid D_m \in P_i, D_n \in P_j\} \subset P$, where P is a multi-set of all protein pairs P_{ij}.

Let an interaction between P_i and P_j or between D_m and D_n be represented by a random variable p_{ij} or d_{mn}. Then, $p_{ij} = 1$ if P_i and P_j interact with each other, otherwise $p_{ij} = 0$. In the same manner, $d_{mn} = 1$ if D_m and D_n interact with each other, otherwise $d_{mn} = 0$. Based on the known protein interaction data, the association method assigns a probability of interaction for domain pair D_m and D_n as (Sprinzak and Margalit, 2001):

$$\lambda_{mn} \equiv \Pr(d_{mn} = 1) = \frac{I_{mn}}{N_{mn}}, \tag{1}$$

where N_{mn} is the total number of protein pairs containing domain pair (D_m, D_n) in the training dataset, and I_{mn} is the number of interacting protein pairs containing domain pair (D_m, D_n) in the training dataset, i.e., $N_{mn} = \sum_{\{P_{ij} | D_{mn} \in P_{ij}\}} 1$ and $I_{mn} = \sum_{\{P_{ij} | D_{mn} \in P_{ij}\}} p_{ij}$. Hayashida et al. (2003) defined the strength ρ_{ij} of interaction between P_i and P_j instead of I_{mn} and defined the probability of interaction between D_m and D_n as

$$\lambda_{mn} \equiv \Pr(d_{mn} = 1) = \frac{\sum_{\{P_{ij} | D_{mn} \in P_{ij}\}} \rho_{ij}}{N_{mn}}, \tag{2}$$

where N_{mn} denotes the number of protein pairs. ρ_{ij} is a confidence ratio of the interaction between proteins P_i and P_j, which is defined as

$$\rho_{ij} = \frac{O_{ij}}{Z}, \tag{3}$$

where O_{ij} is the number of times that proteins P_i and P_j are observed to interact in the experiments, and Z is the total number of the experiments containing domain pairs (D_m, D_n).

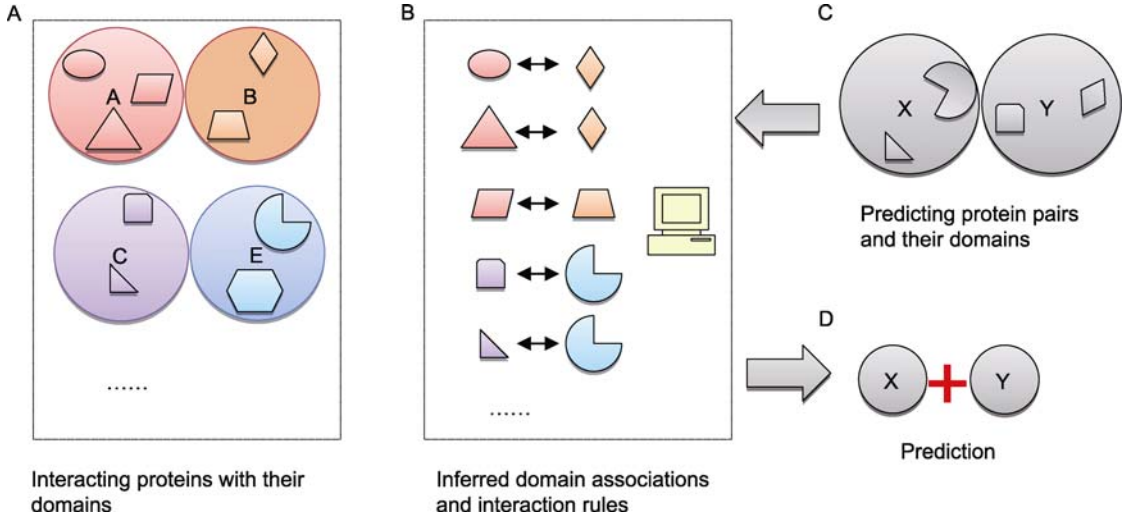

Figure 2. Framework of predicting PPIs by domain associations. (A) Identification of domains in interacting proteins. (B) Association of domain pairs related to protein interactions and learning the domain-domain interaction rules. (C) Prediction of novel interaction by scoring the involved domains.

Therefore, the probability of interaction between P_i and P_j is given by

$$\Pr(p_{ij}=1)=1-\prod_{\{D_{mn}\in P_{ij}\}}(1-\lambda_{mn}). \quad (4)$$

After estimating a set of interacting domain pairs from training protein interactions the interaction probability between a new protein pairs can be predicted.

It is crucial to estimate λ_{mn} accurately from the given interaction data ρ_{ij}. We proposed an algorithm named Association Probabilistic Method (APM) (Chen et al., 2006) to predict PPIs by defining:

$$\lambda_{mn}\equiv\Pr(d_{mn}=1)=\frac{\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}}\left[1-(1-\rho_{ij})^{1/|P_{ij}|}\right]}{N_{mn}}, \quad (5)$$

where $|P_{ij}|$ represents the number of domain pairs in P_{ij}. Obviously, if the ratio ρ_{ij} for each protein pair (P_i, P_j) takes either 0 or 1, (5) is identical to (1) or (2) because of $\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}}\left[1-(1-\rho_{ij})^{1/|P_{ij}|}\right]=\sum_{\{P_{ij}|D_{mn}\in P_{ij}\}}\rho_{ij}=I_{mn}$. λ_{mn} in (5) can be viewed as a reverse function of $\Pr(p_{ij}=1)$ in (4) when all of λ_{mn} in P_{ij} take an identical value. Thus, the protein interaction of APM is obtained by substituting λ_{mn} in (5) into (4). Clearly, both λ_{mn} and $\Pr(p_{ij}=1)$ are straightforwardly equal to ρ_{ij} for $|P_{ij}|$ = 1 (i.e. there is only one domain pair between proteins P_i and P_j. On the other hand, all the domain pairs have the equal opportunity to contribute the interactions between P_i and P_j for $|P_{ij}|>1$ provided that there is no prior information. Our method has shown higher performance in benchmark datasets (Chen et al., 2006). Compared to the former methods, domain association methods define the interacting components in the proteins clearly, and are also associated with the interacting rules quantitatively. We have improved the association method into multiple domain pairs (Wang et al., 2007). Clearly, the associated domains in the predicting proteins can be tracked from the learned domain interaction principles and they are write-box approaches. The domain association methods are often combined with the following machine learning methods (Hayashida et al., 2003).

Machine learning methods

Compared to these write-box methods of direct mapping and domain association, the methods in the second group of PPI prediction methods often employ supervised learning algorithms to mine the features in these interacting protein pairs as well as the noninteracting ones. The features of interacting protein pairs are transferred into the predicting protein pairs by the defined scheme of learning process. The methods transfer the interacting characteristics into these predicting protein pairs without obviously trackable feature mapping. The protein pairs are often encoded by these identified sequence, structure, and various genomic features (Jansen et al., 2003). The label of interaction or non-interaction is used as the sign to supervise the learning. The classifier is trained and can be used to predict the interaction in the encoded proteins. The features are not mapped clearly as those in the direct association or linear mapping methods. We can regard the extension of these features into predicting proteins as an indirect or nonlinear mapping.

Fig. 3 shows the framework of predicting PPIs by machine learning methods. Firstly, the features of interacting proteins and noninteracting proteins including their sequences, structures, physicochemical and other defined features are identified and encoded into feature vectors (shown in Fig. 3A). The proteins are represented by vectors with feature elements. By employing a machine learning algorithm (Vapnik et al., 1995), e.g., naive Bayes (NB), neural network (NN), support vector machine (SVM) and random forest (RF), the predictor is trained by the features with the labels of interacting status. The predictor will be prepared after training (shown in Fig. 3B). When the corresponding features of the predicting proteins are available as shown in Fig. 3C, the trained classifier can predict whether protein *X* interacts with protein *Y* by the learned rules as shown in Fig. 3D.

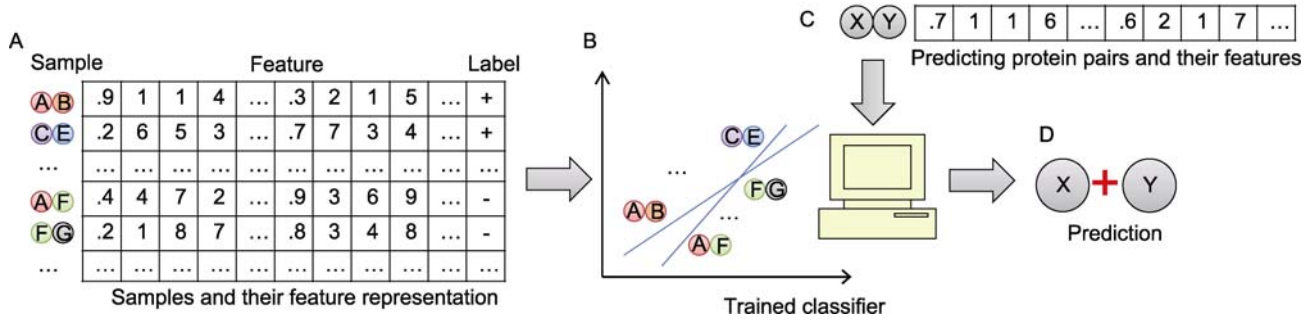

Figure 3. Predicting PPIs by machine learning methods. (A) Encoding positive and negative samples into feature vectors individually. The label of interaction or non-interaction will be used as supervisory signal in the learning process. (B) A classifier is trained to be able to distinguish the interacting protein pairs from noninteracting pairs from the encoded features. (C) Identifying the corresponding features of the predicting proteins. (D) Predicting the interaction between proteins.

Machine learning methods have widely been developed to predict PPIs in various species (Liu et al., 2012b). The various genomic features and physicochemical properties are encoded into vectors for representing the proteins. The contributed feature, the encoding scheme and the employed classification algorithm will affect the prediction accuracy. Among these genomic features, sequence information of gene and protein is relatively easy to be available. In this sense, sequence-based prediction methods have wide application potential or scope. Shen et al. (2007) proposed a sequence-based method for predicting human PPIs. They used a sliding window technique on interacting protein sequences. The frequency of triplet residues in the protein sequences was encoded into features. After training by an SVM algorithm, they predicted a benchmarked protein interaction dataset with high sensitivity and specificity. Guo et al. (2008) also proposed an SVM-based predictor for PPIs. They selected seven physicochemical properties of amino acids to reflect the protein interactions, i.e., hydrophobicity, hydrophilicity, volumes of side chains of amino acids, polarity, polarizability, solvent-accessible surface area and net charge index of side chains of amino acids. A protein sequence was then encoded into a vector by these properties. Due to the different lengths of protein sequences, they implemented an autocorrelation encoding scheme to calculate the auto-covariance variables from these descriptors by taking the effect of the neighboring residues into account. Thus, the interacting protein pairs were represented by concatenating the two vectors of auto-covariance variables. An SVM-based classification algorithm was trained to be the predictor by learning the interacting features. The method achieved a high prediction accuracy in yeast PPIs (Guo et al., 2008).

With the availability of various types of high-throughput data, there are some methods that have been developed for predicting protein interactions by integrating these heterogeneous information (Jansen et al., 2003). Also, we can combine the machine learning methods with the previous direct mapping methods to predict the interactions. Recently, we proposed such an integrative sequence-based method to predict the PPI networks in *Mycobacterium tuberculosis* (*M. tuberculosis*) by the two types of methods, i.e., interologs of direct mapping and indirect coupling of machine learning (Liu et al., 2012b). Firstly, we implemented the interolog method to map the documented protein interactions of other 14 organisms into *M. tuberculosis*. Secondly, we obtained the interaction features of genetic codon underlying these interacting proteins in the relatively well-established interactome of *Escherichia coli* (*E. coli*). The positive and negative sets of protein interactions in *E. coli* were designed to test the performance of our codon-based prediction methods. The genome and proteome of *E. coli* were downloaded and prepared for the interacting sets as well as all known opening reading frames (ORFs) (Cole et al., 1998). The distance of two ORFs in terms of usage of codon *c* is defined as

$$d_{ij}(c) = \left| f_i(c) - f_j(c) \right|,$$

where $f_i(c)$ and $f_j(c)$ are relative frequencies of codon c in ORF i and ORF j. By codon definition, $\sum_k f_i(c_k) = 1$ and $\sum_k f_j(c_k) = 1$ for $k = 1,2,\ldots,64$ in all codons. The cross validation showed the effectiveness and efficiency of our SVM-based predictor. These features of genetic codons of interacting proteins of *E. coli* were mapped to the proteome of *M. tuberculosis* by the trained SVM classifier. Moreover, the available functional genomic information about *M. tuberculosis* was used to evaluate the predicted interactions, i.e., gene co-expression, evolutionary relationship and functional similarity. Multiple high-throughput data were implemented to assess the reliability of these predicted interactions (Liu et al., 2012b).

BEYOND PREDICTION

The PPI network provides a framework of functional relationships among those involved proteins. The global linkage map among proteins will trigger the identification of important mechanisms and highly benefit further researches from the outline of molecular organization (Eisenberg et al., 2000; Barabasi and Oltvai, 2004; Chen et al., 2009). With the emergence and development of high-throughput technologies, it is urgent to build computational methods of reconstructing the interaction networks from these omics data. In the previous sections, we summarized available strategies of predicting protein interactions and categorized them into several groups. Various features and knowledge about protein interaction events were identified and implemented to the prediction. The principles of how one protein interacts with another were learned and extended to those predicting protein pairs. The approaches have been proved to be successful of predicting protein interactions in their own characteristics.

Because of the diversity and complexity of species, proteome-wide PPI networks for many species are still not available (Kerrien et al., 2007). When their genome data are available, the protein interaction networks of the organisms can be generally predicted afterwards by those available methods. STRING has collected and predicted more than 1000 protein interaction networks for different organisms by the various methods (Mering et al., 2007). The recent prediction focuses on some function-specific, tissue-specific protein interactions and virus-host protein interactions. For instance, autophagy is an essential catabolic process to keep the balance of cellular products in the synthesis, degradation and subsequent recycling. Behrends et al. (2010) built a protein interaction map of autophagy in human cells by a proteomic analysis. They provided the global architecture of the autophagy interaction network and revealed those proteins that interact with the core autophagic machinery and related molecules. The related proteins are formed into functional

groups of community and pathway to perform specific functions. Jager et al. (2011) implemented both proteomics and computational experiments to identify host proteins associated with HIV-1 proteins systematically and quantitatively. The interaction map of HIV-1 proteins and host proteins provides the detailed relationship in the host-pathogen system from which new possible targets for drug design will be identified. Based on the reference PPI network of HPRD (Prasad et al., 2009), Wang et al. (2011) established a protein interaction network of human liver. They mapped the interactions in the human liver expression proteins by a yeast two-hybrid technology. In the tissue-specific protein interaction network, they identified the significantly different topology and functional relationships in a liver-specific manner.

Building the interaction map is a milestone for studying an organism at the molecular level. Inferring PPIs is not the goal of making these maps. It is for further extracting biological information, providing valuable insight and deciphering the mechanism from the interaction maps (Skrabanek et al., 2008). After the reconstruction of these networks, they provide the valuable reference resources for further studies of phenotype mechanisms and dysfunctional pathways. Recently, more and more such researches have been available (Ideker et al., 2008). From the interaction map, it is easy to determine the hub proteins and network motifs which will imply essential components (Milo et al., 2002; Han et al., 2004a). Community structure in a protein interaction network indicates the functional characteristics underlying the protein cluster (Newman and Girvan, 2004). We built an optimization algorithm to efficiently detect the community structures in yeast PPI network with a high accuracy (Zhang et al., 2009). With more interaction data available, PPI networks are increasingly serving as tools to reveal the molecular basis of complex diseases (Ideker et al., 2008; Chen et al., 2012; Liu et al., 2012a, 2012c). The topology of the proteins in the network has been investigated with the relationship with diseases (Goh et al., 2007; Ideker et al., 2008). The protein interaction network has been employed to identify new disease related genes (Liu et al., 2012c). Based on these protein interaction maps, there are also some methods which have been developed to identify the active pathways and dysfunctional modules in some diseases. The functional modules often contain biomarker properties which can be applied as network-based biomarkers, i.e., network biomakers and dynamical network biomarkers (Chen et al., 2012), for distinguishing disease samples (He et al., 2011, 2012; Liu et al., 2012a, 2012c) or even pre-disease samples (Chen et al., 2012) from normals. Moreover, the dynamics of protein interaction is a key property of the PPI network (Han et al., 2004a; Liu et al., 2011). There are some methods of scoring schemes which have been provided to annotate the probability of interaction (Bader et al., 2004; Yu et al., 2012) while the predicted interactions are often binary. Two proteins generally interact in some specific conditions and environment (Liu et al., 2011). The spatial and temporal features of these interaction maps will provide deep understanding for the specific and substantial insights into the organism from the systematic perspective (Bossi and Lehner, 2009; Lage et al., 2010; Liu et al., 2011). These features are popular aspects to be considered in disease research and drug discovery (Liu et al., 2012c). For instance, some proteins interact in normal condition, while the interactions disappear or rewired in the disease cases. The dysfunctional interactions definitely shed light on the disease mechanism. The network powers the disease mechanism research and provides new alternatives and resources (Liu et al., 2012c).

There are more and more omics data available at different levels. The ongoing hot research topics include how to combine them together to predict protein interactions and apply them in an integrated framework to solve biological problems. In our method of predicting the PPI network in *M. tuberculosis*, we built a novel framework of integrating these datasets. Firstly, we predicted the protein interaction network by interologs and machine learning based on sequence information. Then we implemented the information of co-expression, co-evolution and co-function to evaluate and access these predicted interactions (Liu et al., 2012b). Moreover, there are many valuable topics need be investigated in addition to the prediction of proteome-wide PPIs, such as how to integrate the reconstructed network with the other gene expression data, RNA-seq data and proteomics data to improve the identification of disease biomarkers, function-specific modules and dysfunctional pathways (Liu et al., 2012c).

ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 31100949, 91029301, 61134013 and 61072149), the Chief Scientist Program of Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS) (No. 2009CSP002) and the Knowledge Innovation Program of SIBS of CAS (No. 2011KIP203), the Shanghai National Science Foundation (No. 11ZR1443100) and the SA-SIBS Scholarship Program. This research was also partially supported by the National Center for Mathematics and Interdisciplinary Sciences of CAS, the Shanghai Pujiang Program, and the FIRST program from JSPS initiated by CSTP. We thank the members of Prof. Xiang-Sun Zhang's research group in the Academy of Mathematics and Systems Science of CAS for helpful discussions.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Andres, L.E., Ezkurdia, I., Garcia, B., Valencia, A., and Juan, D. (2009). EcID. A database for the inference of functional interactions in *E. coli*. Nucleic Acids Res 37, D629–D635.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res 38, D525–D531.

Aytuna, A.S., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21, 2850–2855.

Bader,G.D., Betel, D., and Hogue, C.W. (2003). BIND: the biomolecular interaction network database. Nucleic Acids Res 31, 248–250.

Bader, J.S., Chaudhuri, A., Rothberg, J.M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22, 78–85.

Barabasi, A.L., and Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101–113.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35, D760–D765.

Biocarta. (2012). Available: http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways. Accessed April 7, 2012.

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. Mol Syst Biol 5, 260.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature 466, 68–76.

Bhardwaj, N., and Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 21, 2730–2738.

Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I., and Marcotte, E.M. (2004). Protein interaction networks from yeast to human. Curr Opin Struct Biol 14, 292–299.

Brown, K.R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol 8, R95.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31, 3497–3500.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Res 35, D572–D574.

Chen, L., Liu, R., Liu, Z.P., Li, M., and Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep. 2, 342.

Chen, L., Wang, R.S., and Zhang, X.S. (2009). Biomolecular networks: methods and applications in systems biology (John Wiley & Sons, Hoboken, New Jersey).

Chen, L., Wang, R., Li, C., and Aihara, K. (2010). Modelling biomolecular networks in cells: structures and dynamics. (Springer-Verlag, Berlin).

Chen, L., Wu, L.Y., Wang, Y., and Zhang, X.S. (2006). Inferring protein interactions from experimental data by association probabilistic method. Proteins 62, 833–837.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537–544.

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

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86–90.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000). Protein function in the post-genomic era. Nature 405, 823–826.

Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29, 482–486.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins 18, 309–317.

Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E. (2000). Co-evolution of proteins with their interaction partners. J Mol Biol 299, 283–293.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc Natl Acad Sci U S A 104, 8685–8690.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res 29, 3513–3519.

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res 36, 3025–3030.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P., et al. (2004a). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88–93.

Han, K., Park, B., Kim, H., Hong, J., and Park, J. (2004b). PID: the Human Protein Interaction Database. Bioinformatics 20, 2466–2470.

Hayashida, M., Ueda, N., and Akutsu, T. (2003). Inferring strengths of protein-protein interactions from experimental data using linear programming. Bioinformatics 19, ii58–ii65.

He, D., Liu, Z.P., and Chen, L. (2011). Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. BMC Genomics 12, 592.

He, D., Liu, Z.P., Honda, M., Kaneko, S., and Chen, L. (2012). Co-expression network analysis in chronic hepatitis B and C hepatic lesion reveals distinct patterns of disease progression to hepatocellular carcinoma. J Mol Cell Biol 4, 140–152.

Huynen, M., Snel, B., Lathe, W. 3rd, and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10, 1204–1210.

Ideker, T., and Sharan, R. (2008). Protein networks in disease. Genome Res 18, 644–652.

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li. K., et al. (2011). Global landscape of HIV-human protein complexes. Nature 481, 365–370.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. Genome Res 12, 37–46.

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

Jothi, R., Kann, M.G., and Przytycka, T.M. (2005). Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics 21, i241–i250.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. (2007). IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35, D561–D565.

Lage, K., Mollgard, K., Greenway, S., Wakimoto, H., Gorham, J.M., Workman, C.T., Bendsen, E., Hansen, N.T., Rigina, O., Roque, F.S., et al. (2010). Dissecting spatio-temporal protein networks driving human heart development and related disorders. Mol Syst Biol 6, 381.

Lee, K., Chuang, H.Y., Beyer, A., Sung, M.K., Huh, W.K., Lee, B., and Ideker, T. (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res 36, e136.

Liu, X., Liu, Z.P., Zhao, X.M., and Chen, L. (2012a). Identifying disease genes and module biomarkers with differential interactions. J Am Med Inform Assoc 19, 241–248.

Liu, Z.P., Wang, J., Qiu, Y.Q., Leung, R.K.K., Zhang, X.S., Tsui, S.T.W., and Chen, L. (2012b). Inferring a protein interaction map of Mycobacterium tuberculosis based on sequences and interologs. BMC Bioinformatics 13 (Suppl 7), S6.

Liu, Z.P., Wang, Y., Zhang, X.S., and Chen, L. (2012c). Network-based analysis of complex diseases. IET Syst Biol 6: 22–33.

Liu, Z.P., Wang, Y., Zhang, X.S., Xia, W., and Chen, L. (2011). Detecting and analyzing differentially activated pathways in brain regions of Alzheimer's disease patients. Mol Biosyst 7, 1441–1452.

Liu, Z.P., Wu, L.Y., Wang, Y., Chen, L., and Zhang, X.S. (2007). Predicting gene ontology functions from protein's regional surface structures. BMC Bioinformatics 8, 475.

Liu, Z.P., Wu, L.Y., Wang, Y., Zhang, X.S., and Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics 26, 1616–1622.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005) Assessing the limits of genomic data integration for predicting protein networks. Genome Res 15, 945–953.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. (2007). STRING 7 – recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35, D358–D362.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science 298, 824–827.

Newman, M.E., and Girvan, M. (2004). Finding and evaluating community structure in networks. Phys Rev E 69, 026113.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). Use of contiguity on the chromosome to predict functional coupling. In Silico Biol 1, 93–108.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W., et al. (2005) The MIPS mammalian proteinprotein interaction database. Bioinformatics 21, 832–834.

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

Pazos, F., and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47, 219–227.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285–4288.

Prasad, T.S.K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., et al. (2009). Human Protein Reference Database - 2009 update. Nucleic Acids Res 37, D767–D772.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci U S A 104, 4337–4341.

Skrabanek, L., Saini, H.K., Bader, G.D., and Enright, A.J. (2008). Computational prediction of protein-protein interactions. Mol Biotechnol 38, 1–17.

Smith, G.R., and Sternberg, M.J. (2002). Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12, 28–35.

Sprinzak, E., and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 311, 681–692.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34, D535–D539.

Szilagyi, A., Grimm, V., Arakaki, A.K., and Skolnick, J. (2005). Prediction of physical protein-protein interactions. Phys Biol 2, S1–S16.

Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol 44, 66–73.

Tsoka, S., and Ouzounis, C.A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26, 141–142.

Sapkota, A., Liu, X., Zhao, X.M., Cao, Y., Liu, J., Liu, Z.P., and Chen, L. (2011). DIPOS: database of interacting proteins in Oryza sativa. Mol Biosyst 7, 2615–2621.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. Nucleic Acids Res 32, D449–D451.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D., et al. (2009). The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 38, D540–D544.

Valencia, A., and Pazos, F. (2002). Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 12, 368–373.

Vapnik, V. (1995). The nature of statistical learning theory.

(Springer-Verlag, New York).

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8, R39.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. Science 287, 116–122.

Wang, R.S., Wang, Y., Wu, L.Y., Zhang, X.S., and Chen, L. (2007). Analysis on multi-domain cooperation for predicting protein-protein interactions. BMC Bioinformatics 8, 391.

Wang, J., Huo, K., Ma, L., Tang, L., Li, D., Huang, X., Yuan, Y., Li, C., Wang, W., Guan, W., et al. (2011). Toward an understanding of the protein interaction network of the human liver. Mol Syst Biol 7, 536.

Wang, L., Liu, Z.P., Zhang, X.S., and Chen, L. (2012). Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Eng Des Sel 25, 119–126.

Winter, C., Henschel, A., Kim, W.K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 34, D310–D314.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19, 1524–1530.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14, 1107–1118.

Yu, X., Wallqvist, A., and Reifman, J. (2012). Inferring high-confidence human protein-protein interactions. BMC Bioinformatics 13, 79.

Zhang, X.S., Wang, R.S., Wang, Y., Wang, J., Qiu, Y., Wang, L., and Chen, L. (2009). Modularity optimization in community detection of complex networks. Europhys Lett 87, 38002.

Zhao, X.M., Chen, L., and Aihara, K. (2010). A discriminative approach to identifying domain-domain interactions from protein-protein interactions. Proteins 78, 1243–1253.

Zhao, X.M., Zhang, X.W., Tang, W., and Chen, L. (2009). FPPI: Fusarium graminearum protein-protein interaction database. J Proteome Res 8, 4714–4721.

Zhou, H.X., and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44, 336–343.

