ELSEVIER

Protein interaction predictions from diverse sources

Yin Liu[1], Inyoung Kim[2] and Hongyu Zhao[3,4]

[1] Department of Neurobiology and Anatomy, University of Texas Medical School at Houston, 6431 Fannin Street, Houston, TX 77030, USA
[2] Department of Statistics, Virginia Tech, 410-A Hutcheson Hall, Blacksburg, VA 24061, USA
[3] Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA
[4] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

Protein–protein interactions play an important role in many cellular processes. The availability of a comprehensive and accurate list of protein interactions can facilitate drug target discovery. Recent advances in high-throughput experimental technologies have generated enormous amounts of data and provided valuable resources for studying protein interactions. However, these technologies suffer from high error rates because of their inherent limitations. Therefore, computational approaches capable of incorporating multiple data sources are needed to fully take advantage of the rapid accumulation of data. In this review, we focus on the computational methods that integrate multiple data sources by combining direct measurements on protein interactions from diverse organisms, and by integrating different types of indirect information from various genomic and proteomic approaches.

Introduction

Protein–protein interactions (PPI) play a critical role in the control of most cellular processes, such as signal transduction, gene regulation, cell cycle control and metabolism. Alterations in protein interactions perturb the normal cellular processes and contribute to many diseases, for example cancer and AIDS. The correct identification of protein–protein interactions can help us assign the cellular functions of novel proteins, investigate the mechanisms of intracellular biochemical pathways and, more importantly, understand the underlying causes of diseases and lead to the development of drugs to treat these diseases. For these reasons, protein–protein interactions represent an important class of targets for human therapeutics and good annotation of protein interactions can greatly help drug discovery [1].

Recent advances in genomics and proteomics technologies have opened the door for rapid biological data acquisition. Within the past decade, genome-wide data on protein interactions in humans and many model species have become available [2–7]. In the meantime, a large amount of indirect biological information on protein interactions, including sequence and functional annotation, protein localization information and gene expression measurements has also become available. These data, however, are far from complete and contain many false negatives and false positives. Current protein interaction information obtained from experimental methods covers only a fraction of the complete PPI networks [8–11]; therefore, there is a great need to develop complementary computational methods, capable of identifying and verifying interactions between proteins. In the past several years, several computational methods have been proposed to predict protein–protein interactions based on various data types. For example, with the genomic information available, the Rosetta stone method predicts the interacting proteins based on the observation that some of single-domain proteins in one organism can be fused into a multiple domain protein in other organisms [12,13]. The phylogenetic profile method is based on the hypothesis that interacting proteins tend to co-evolve so that their respective phylogenetic trees are similar [14,15]. The concept of 'interolog', which refers to homologous interacting protein pairs among different organisms, has also been used to identify protein interactions [16]. The gene neighborhood method is based on the observation that functionally related genes encoding potential interacting proteins are often transcribed as an operon [17,18].

Corresponding author: Zhao, H. (hongyu.zhao@yale.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.01.005

With genome-wide gene expression measurements available, some methods find gene co-expression patterns in interacting proteins [19,20]. Based on protein structural information, the protein docking methods uses geometric and steric considerations to fit multiple proteins of known structure into a bounded protein complex to study interacting proteins at the atomic level [21]. Moreover, some methods analyze protein interactions at the domain level, considering protein domains as structural and functional units of proteins [22–24]. Each of these studies focuses on a specific dataset, either the direct measurement of protein interaction, or the indirect genomics dataset that contains information on protein interaction. With each method, however, focusing only on a single type of data source, it is not surprising to see each method has certain limitations and the detailed reviews of these individual methods can be found elsewhere [25–27]. Considering the biases, overlaps and complementarities among a variety of data sources, we discuss the contribution of these data sources to the protein interaction inference and different approaches that integrate these data sources in this review.

Integrate direct protein interaction information from diverse organisms

Note that only the yeast two-hybrid (Y2H) experimental technique provides direct evidence of physical protein–protein interactions in high-throughput analysis. As a result, we start the review with the methods that use data generated from Y2H technique only. As valuable as it is, this experimental approach suffers from high error rates, because of the limitations of the technique [10,11]. For example, a protein initiating transcription itself, can lead to a false positive result and a protein that cannot be targeted to the yeast nucleus may not yield transcription signals even though it may potentially interact with other proteins, leading to false negative results [28]. The false negative rate is estimated to be above 0.5 [10,11]. Because of the large number of possible non-interacting protein pairs, the false positive rate, defined as the ratio of the number of incorrect interactions observed over the total number of non-interacting proteins, is small and is estimated as 1×10^{-3} or less [29]. But, the false discovery rate, defined as the ratio of the number of incorrect interactions observed over the total number of observed interactions, is much greater and is estimated to be 0.2–0.5 [11], indicating a large portion of the observations from Y2H technique are incorrect. As the Y2H data has become available in many model organisms, such as yeast, worm, fruitfly and humans, several computational methods have been developed to borrow information from diverse organisms to improve the accuracy of protein interaction prediction. Noting that domains are structural and functional units of proteins and are conserved during evolution, these methods aim to identify specific domain pairs that mediate protein interactions, by utilizing domain information as the evolutionary connection among these organisms.

Maximum likelihood-based methods (MLE)

The maximum likelihood estimation (MLE) method, coupled with the expectation-maximization (EM) algorithm, has been developed to estimate the probabilities of domain–domain interactions, given a set of experimental protein interaction data [30]. In this method, a likelihood function that describes the probability of observed protein interactions is maximized through iterative steps, so that the unknown parameters, the domain–domain interaction probabilities for each pair of domains can be estimated and then the estimated domain interaction probabilities were used to estimate the protein interaction probabilities. It was originally used to analyze the Y2H data from a single organism – *Saccharomyces cerevisiae* only. When compared with the association methods, which only consider each domain pair separately [22,24], the MLE method was shown to be among the best performing methods for a single organism. More recently, the MLE method was extended beyond the scope of a single organism by incorporating protein interaction data from three different organisms, *S. cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*, assuming that the probability that two domains interact is the same among all organisms [30]. It was shown that the integrated analysis provides more reliable inference of protein–protein interactions than the analysis from a single organism.

Bayesian methods (BAY)

Unlike the likelihood based approaches reviewed above, the false negative and false positive rates of the observed protein interaction data are treated as unknown in the Bayesian methods, so that the domain interaction probabilities, the false positive rate and the false negative rates of the observed data can be estimated simultaneously [31]. With the prior information incorporated, the posterior distribution of theses parameters can be inferred and the parameters can be estimated from the observed data. We assess the performance of these methods by comparing their sensitivities and specificities in predicting interacting domain pairs, illustrated by their receiver operator characteristic (ROC) curves (Figure 1a). The protein interaction dataset consists of 12, 849 interactions from three organisms [2–5]. Compared to the likelihood-based methods, the Bayesian-based methods may be more efficient in dealing with a large number of parameters and more effective to allow for different error rates across different datasets [31].

Domain pair exclusion analysis (DPEA)

The maximum likelihood methods may preferentially detect domain pairs with high frequency of co-occurrence in interacting protein pairs. To detect more specific domain interactions, Riley *et al.* [32] modified the likelihood-based approach and developed the domain pair exclusion analysis. In this approach, two different measures, the θ values and the *E*-score, were estimated for each pair of domains. The θ value is obtained through the likelihood method, similarly as the one described above, corresponding to the probability that two domains interact with each other. The *E*-score is defined for each domain pair as the log likelihood ratio of the observed protein interactions given that the two domains interact over given that the two domains do not interact. The *E*-score greatly help to identify the specific interacting domain pairs, where the MLE and BAY methods may fail to detect. The approach was applied to all the protein interactions in the database of interacting proteins (DIP) [33], assuming no false positive and false negatives.

Parsimony explanation method (PE)

Using the same dataset constructed by Riley *et al.* [33] from the DIP database, the parsimony approach formulates the problem of

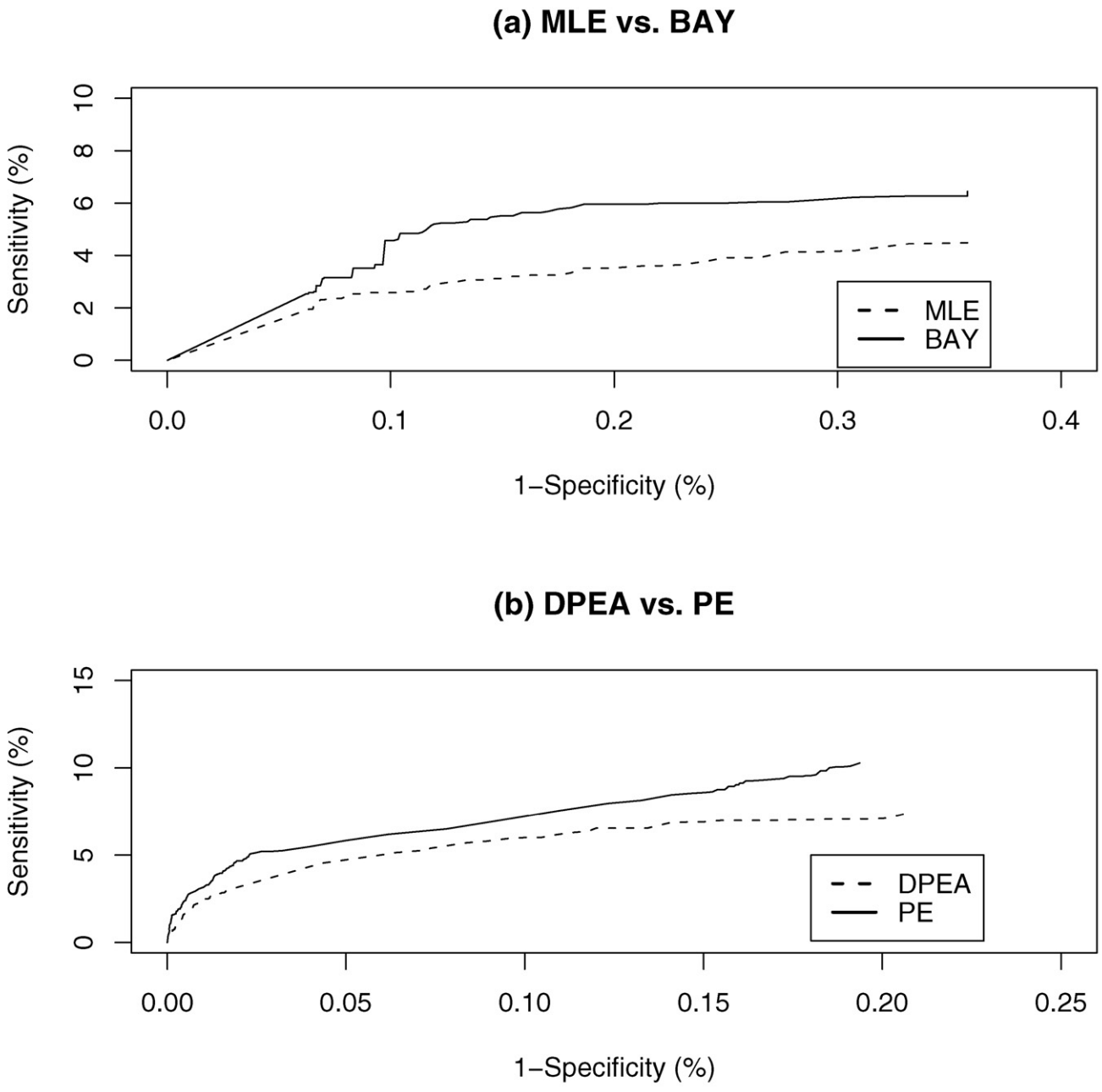

FIGURE 1

ROC curves of domain interaction prediction results from different methods. Comparison of sensitivities and false positive rates (1 – specificity) obtained by different methods. MLE, maximum likelihood-based method [30]; BAY, Bayesian method [31]; DPEA, domain pair exclusion analysis [32]; PE, parsimony explanation method [34]. **(a)** MLE vs. BAY methods; **(b)** DPEA vs. PE methods. The set of protein domain pairs in iPFAM are used as the gold standard set (http://www.sanger.ac.uk/Software/Pfam/iPfam/). Here, the sensitivity is calculated as the number of predicted interacting domain pairs that are included in the gold standard set divided by the total number of domain pairs in the gold standard set. The false positive rate is calculated as the number of predicted interacting domain pairs that are not included in the gold standard set divided by the total number of possible domain pairs not included in the gold standard set. The area under the ROC curve is a measurement of prediction accuracy for each method.

domain interaction prediction as a linear programming (LP) optimization problem [34]. In this approach, the LP score associated with each domain pair was inferred by minimizing the object function subject to a set of constraints describing the protein interaction data. In addition to the LP score, another measure, the pw-score was defined. Similar to the *E*-value in the DPEA method, the pw-score is used as an indicator to remove the promiscuous domain pairs that occur frequently and have few witnesses. Here, the witness of a domain pair is defined as the interacting protein pair only containing this domain pair. We also compare the performance of DPEA method and PE method using their ROC curves (Figure 1b). The reason contributing to the improved performance of the PE method compared to the DPEA method could be that the DPEA method tends to assign higher

BOX 1

Identification of domain interactions under different scenarios

Note: P_1 represents protein 1, D_1 represents domain 1, $P_1 = \{D_1\}$ represents that protein 1 only contain domain 1.
Scenario 1:
Protein-domain relationship:

$P_1 = \{D_1\}$, $P_2 = \{D_2\}$, $P_3 = \{D_1\}$, $P_4 = \{D_2\}$, $P_5 = \{D_1\}$, $P_6 = \{D_2\}$

Observed protein interaction data:

$P_1 \leftrightarrow P_2$, $P_1 \leftrightarrow P_4$, $P_1 \leftrightarrow P_6$, $P_2 \leftrightarrow P_3$, $P_3 \leftrightarrow P_4$, $P_3 \leftrightarrow P_6$, $P_2 \leftrightarrow P_5$, $P_4 \leftrightarrow P_5$, $P_5 \leftrightarrow P_6$

Under this scenario, each protein-containing domain 1 interact with each protein containing domain 2. All four methods MLE, BAY, DPEA, and PE can identify the interaction between domain 1 and domain 2.
Scenario 2:
Protein-domain relationship:

$P_1 = \{D_1\}$, $P_2 = \{D_2\}$, $P_3 = \{D_1\}$, $P_4 = \{D_2\}$, $P_5 = \{D_1\}$, $P_6 = \{D_2\}$

Observed protein interaction data:

$P_1 \leftrightarrow P_2$, $P_3 \leftrightarrow P_4$, $P_5 \leftrightarrow P_6$

Under this scenario, only a small fraction (3/9) of protein pairs containing domain pair 1 and 2 interact. Both MLE and BAY methods may not be able to identify the interaction between domain 1 and 2. But, if the interaction of domain 1 and 2 is excluded, the likelihood of the observed protein interaction data is lower than that under the condition domain 1 and 2 interact, which leads to the high *E*-score of this domain pair in the DPEA method, therefore, DPEA can detect the interaction. For the PE method, the interaction between domain 1 and 2 is the only explanation for the observed data, so it can be detected by the PE method as well.
Scenario 3:
Protein-domain relationship:

$P_1 = \{D_1, D_3\}$, $P_2 = \{D_2, D_4\}$, $P_3 = \{D_1\}$, $P_4 = \{D_2\}$, $P_5 = \{D_1\}$, $P_6 = \{D_2\}$

Observed protein interaction data:

$P_1 \leftrightarrow P2$, $P3 \leftrightarrow P4$, $P5 \leftrightarrow P6$

Under this scenario, as the only protein pair (P_1, P_2) containing domain 3 and 4 interact, both MLE and BAY methods can detect the interaction between domain 3 and 4. While only a small fraction (3/9) of protein pairs containing domain pair 1 and 2 interact, the interaction between domain 1 and 2 may not be detected by MLE and BAY methods. The DPEA method may detect interaction of domain pair (1, 2), and domain pair (3, 4) as well because excluding both domain pairs decreases the likelihood of the observed data. For the PE method, because the interaction between domain pair (1, 2) represents the smallest number of domain pairs to explain the observed data, this interaction is preferred than the interaction between domain pair (3, 4).

probabilities to the infrequent domain pairs in multi-domain proteins, while it is avoided in the PE method [34].

All the methods described in this section focus on estimating domain interaction by pooling protein interaction information from diverse organisms. Box 1 illustrates the difference of these methods in identifying interacting domain pairs. The estimated domain interactions can then be used for the protein interaction prediction by correlating proteins with their associated domains. We expect that the results from these domain interaction prediction methods can be further improved when the domain information is more reliably annotated in the future. The current information on domain annotation is incomplete. For example, only about two-thirds of proteins and 73% of sequence in yeast proteome are annotated with domain information in the latest release of PFAM database [35], as a result, prediction based on domain interaction will only be able to cover a portion of the whole interactome. To overcome this limitation, a support vector machine learning method was developed [36]. Instead of using the incomplete domain annotation information, this method used the signature products, which represent the protein primary sequence only, for protein interactions prediction [36]. In addition to the incomplete domain annotation information, we note that there is another limitation of these domain interaction prediction methods: the accuracy and reliability of these methods highly depend on the protein interaction data. Although the prediction accuracy can be improved by integrating data from multiple organisms, the protein interaction data itself may be far from complete. Therefore, efforts have been made to integrate other types of biological information and these are briefly reviewed next.

Integrate multiple types of biological information

Genomic features

Different genomic features, such as DNA sequence, functional annotation and protein localization information have been used for predicting protein interactions. Each feature provides insights into a different aspect of protein interaction information, thus covers a different subset of the whole interactome. Depending on the data sources, these genomic features can be divided into four categories. First, high-throughput protein interaction data obtained from Y2H and mass spectrometry of co-immunoprecipitated protein complexes (Co-IP/MS) techniques provide direct protein interaction information and protein co-complex membership information, respectively. Second, functional genomic data such as gene expression, gene ontology (GO) annotation provide information on functional relationships between proteins. Third, sequence- and structure-based data reveal the sequence/structure homology and chromosome location of genes. Finally, network topological parameters calculated from Y2H and Co-IP/MS data characterize the topological properties of currently available protein interaction network. For example, the small-world clustering coefficient of a protein pair, calculated as the *p*-value from a hypergeometric distribution, measures the similarity between the neighbors of the protein pair in a network [37]. The effects of these features on the prediction performance depend on how these features are encoded. With many datasets for a single feature under different experiment conditions available, there are two ways to encode the datasets: the 'detailed' encoding strategy treats every experiment separately, while the 'summary' encoding strategy groups all experiments belonging to the same feature together and provides a single value [38]. For example, when all the gene expression datasets are used as one feature, the 'summary' encoding strategy generates a single similarity score for each pair of proteins, but we can obtain multiple similarity scores for each

TABLE 1

Useful features for predicting protein interactions

Category	Feature abbreviation	Feature	Proteome coverage (%)	Data source
Protein interaction data	Y2H	Yeast two-hybrid data	60	[2,3]
	MS	Protein complex data	64	[55,56]
Functional genomic data	GE	Gene expression	100	[57]
	FUN	GO molecular function	62	[58]
	PRO	GO biological process	70	[58]
	COM	GO cellular component	72	[58]
	CLA	MIPS protein class	80	[49]
	PHE	MIPS mutant phenotype	24	[49]
	PE	Protein expression	65	[59]
	ESS	Co-essentiality	67	[49]
	MAE	Marginal essentiality	99	[42]
	GI	Genetic interaction	24	[60]
	TR	Transcription regulation	98	[61]
Sequence/structure information	GF	Gene fusion	19	[42]
	GN	Gene neighborhood	22	[42]
	PP	Pylogenetic profile	29	[42]
	SEQ	Sequence similarity	100	[38]
	INT	Interolog	100	[38]
	DD	Domain–domain interaction	65	[38]
	COE	Co-evolution scores	22	[14]
	THR	Threading scores	21	[42]
	PF	Protein fold	26	[62]
Network topological parameters	CLU	Small-world clustering coefficients[a]	n/a	[44,45]

Coverage, percentage of proteins in *S. cerevisiae* that are annotated by this feature. Data Source, datasets used for the features. Here, only the datasets with the highest coverage are listed. n/a, not applicable.

[a] The clustering coefficients are calculated according to protein interaction data.

protein pair using the 'detailed' encoding strategy, with one score computed from one gene expression dataset. Another challenge of the integrated studies is related to the quality of the data to be integrated. It is well-known that the prediction power and accuracy would be decreased by including irrelevant and 'noisy' features. A list of these features, along with their proteome coverage, is summarized in Table 1. Among all the features, whole-genome gene expression data are currently the largest source of high-throughput genomic information and considered as the most important feature according to the variable importance criterion in a random forest-based method [38]. The Euclidean distance [39] or Pearson correlation coefficients [40] between the expression profiles of genes are calculated to determine the expression similarity of genes and the level of gene expression similarity was applied in all the integrated studies for protein interaction prediction. Following gene expression, function annotation from gene ontology, which covers about 80% of the yeast proteome, is the second most important features according to the importance measure obtained from random forest method. Although previous studies indicate MIPS and GO annotation are more important than gene expression in prediction, this may be because of different number of gene expression datasets used (2 [40–42] vs. 20 [38]) and different encoding styles of the features ('summary' [40–42] vs. 'detailed' [38]). Nonetheless, gene expression, interaction data from Co-IP experiment, function annotation from MIPS and GO are considered the most important features in several studies according to the feature importance analysis [38,40–42]. It also suggests that a small number of important features may be sufficient to predict protein interaction and the prediction performance may not be improved from adding additional weak features. For the weak features, although they all have varying degree of success for the protein interaction prediction task, there are several limitations. First, the datasets associated with the weak features are noisy and not reliable. Second, weak features usually have low proteome coverage, making little contribution to the prediction and leads to biased prediction results. For example, the structural information of proteins is potentially useful for protein interaction prediction, as demonstrated by the 'protein–protein docking' approach which assembles protein complexes based on the three-dimensional structural information of individual proteins. However, the availability of the protein structural information is very limited, and the structures of individual proteins in an unbound state differ from those in a bounded complex, therefore, the prediction of protein interactions based on the structural information only is not reliable [27]. Finally, these weak features may not have strong association with protein interactions. For example, the feature of 'transcription regulation' lists the genes co-regulated by the same set of transcription factors. Although it is shown that the co-regulated genes often function together through protein interactions, the proteins encoded by the co-regulated genes do not necessarily interact [43].

Predicting protein interactions through integrating multiple features

With these genomic features available, many computational approaches, ranging from simple union or intersection of features to more sophisticated machine learning methods, such as random forest and support vector machines (SVM), have been applied to integrate different features for protein interaction predictions. These approaches aim at predicting direct physical interactions,

TABLE 2

Summary of previous methods integrating multiple features for protein interactions prediction

Task	Methods	Features used	Refs
Protein physical interaction	Random forest	Y2H, MS, GE, FUN, PRO, COM, CLA, PHE, PE, ESS, GI, TR, GF, GN, PP, SEQ, INT, DD	[38]
	k-nearest neighbor	Y2H, MS, GE, FUN, PRO, COM, CLA, PHE, PE, ESS, GI, TR, GF, GN, PP, SEQ, INT, DD	[38]
	Logistic regression	Y2H, MS, GE, CLU	[44,45]
Co-complex membership	Random forest	Y2H, MS, GE, PRO, CLA, ESS	[41]
	k-nearest neighbor	Y2H, MS, GE, FUN, PRO, COM, CLA, PHE, PE, ESS, GI, TR, GF, GN, PP, SEQ, INT, DD	[38]
	Support vector machine	Y2H, MS, GE, FUN, PRO, COM, CLA, PHE, PE, ESS, GI, TR, GF, GN, PP, SEQ, INT, DD	[38]
	Logistic regression	GE, PRO, COM, TR, GF, GN, PP, DD, PF	[62]
	Naïve Bayes	GE, PRO, CLA, ESS	[40]
	Decision tree	Y2H, MS, GE, COM, PHE, TR, SEQ, GF, GN, PP	[63]
Functionally linked proteins	Bayes scoring	Y2H, MS, GE, GI, GF, PP	[47]
Domain interactions[a]	Naïve Bayes	Y2H, FUN, PRO, GF	[64]
	Evidence counting	Y2H, FUN, PRO, GF	[64]

Features used, the list of abbreviated features used by each method. There may be several different implementations using different sets of features by each method. When using the same set of features, these methods can be sorted according to their performance, as shown in this table, with the best-performing methods listed at the top. For predicting protein physical interaction and protein co-complex membership, the comparison is based on the precision-recall curves and the receiver operator characteristic (ROC) curves of these methods with the 'summary' encoding [38]. For predicting domain interactions, the comparison is based on the ROC curves of these methods [64].

[a] This task focuses on predicting domain–domain interactions instead of protein interactions. The Y2H data here are obtained from multiple organisms, and the feature GF represents the domain fusion event instead of gene fusion event as used for other tasks.

protein co-complex membership in a complex or proteins functionally linked. For instance, a logistic regression approach, based on functional and network topological features, has been presented to evaluate the confidence of the protein–protein interaction data previously obtained from both Y2H and Co-IP/MS experiments [44]. More recently, Sharan *et al*. [45] also implemented a similar logistic regression model, incorporating different features, to estimate the probability that a pair of proteins interact. Jansen *et al*. [40] was among the first to apply a Naïve Bayes classifier using features including mRNA expression data, localization, essentiality and functional annotation for predicting protein co-complex relationship. Based on the same set of features, Lin *et al*. [41] applied two other classifiers, logistic regression and random forest, and demonstrated that random forest outperforms the other two and logistic regression performs similarly with the Naïve Bayesian method. Kernel methods incorporating multiple sources of data including protein sequences, local properties of the network and homologous interactions in other species have been developed for predicting direct physical interaction between proteins [46]. To predict functionally linked proteins, Lee *et al*. [47] developed a Bayesian approach that integrates multiple functional genomic data. A summary of some published methods along with their specific prediction tasks are listed in Table 2.

Gold standard datasets

In addition to the types of genomic features used and the approaches employed to integrate these features, the gold standard datasets defined for training purposes have effects on the relative performance of different approaches, as discussed on the first DREAM (Dialogue on Reverse Engineering Assessments and Methods) conference [48]. Protein pairs obtained from DIP database [33] and protein complex catalog obtained from MIPS database [49] are two most widely used sets of gold standard positives. While DIP focuses on physically interacting protein pairs, MIPS complexes catalog captures the protein complex membership, which lists proteins pairs that are in the same complex, but not necessarily have physical interactions. Therefore, when predicting physical protein interactions, the set of co-complex protein pairs from MIPS is not a good means of assessing the method accuracy. Two groups of protein pairs have been treated as the gold standard negatives most popularly in the literature: random/all protein pairs not included in the gold standard positives and the protein pairs annotated to be in different subcellular localization. Neither of them is perfect: the first strategy may include true interacting protein pairs, leading to increased false positive, while the second group may lead to increased false negatives when a multifunctional protein is active in multiple subcellular compartments but only has limited localization annotation. Protein pairs whose shortest path lengths exceed the median shortest path for random protein pairs in a protein network constructed from experimental data are treated as negative samples as well [44]. However, this strategy is not quite reliable as the experimental data are considered noisy and incomplete.

Prediction performance comparison

The availability of such a wide range of methods requires a comprehensive comparison among them. One strategy is to validate prediction results according to the similarity of interacting proteins in terms of function, expression, sequence conservation, and so on, as they are all shown to be associated with true protein interactions. The degree of the similarities can be measured to compare the performance of prediction methods [48,50]. However, as the similarity measures are usually used as the input features in the integrated analysis, another most widely applied strategy for comparison uses ROC or precision-recall (PR) curves. ROC curves plot true positive rate vs. false positive rate, while PR curves plot precision (fraction of prediction results that are true positives) vs. recall (true positive rate). When dealing with highly skewed datasets (e.g. the size of positive examples is much smaller than that of negative examples), PR curves can demonstrate differences between prediction methods that may not be apparent in ROC curves [51]. However, because the precision does not

necessarily change linearly, interpolating between points in a PR curve is more complicated than that in a ROC curve, where a straight line can be used to connect points [51]. A recent study evaluated the predictive power and accuracies of different classifiers including random forest (RF), RF-based k-nearest neighbor (kRF), Naïve Bayes (NB), decision tree (DT), logistic regression (LR) and support vector machines and demonstrated in both PR and ROC curves that RF performs the best for both physical interaction prediction and co-complex protein pair prediction problem [38]. Because of its randomization strategy, an RF-based approach can maintain prediction accuracy when data is noisy and contains many missing values. Moreover, the variable importance measures obtained from RF method help determine the most relevant features used for the integrated analysis. One point we need to pay attention to, though, is that RF variable importance measures may be biased in situations where potential variables vary in their scale of measurement or their number of categories [52]. Although NB classifier allows the derivation of a probabilistic structure for protein interactions and is flexible for combining heterogeneous features, it was the worst performer among the six classifiers. The relatively poor performance could be because of its assumption of conditional independence between features, which may not be the case, especially when many correlated features are used. A boosted version of simple NB, which is resistant to feature dependence, significantly outperforms the simple NB, as demonstrated in a control experiment with highly dependent features, indicating the limitation of simple NB. Although a recent study showed no significant correlations between a subset of features using Pearson correlation coefficients and mutual information as measures of correlation, there was no statistical significance level measured in this study [42]. Therefore, we cannot exclude the possibility that genomic features are often correlated with each other, especially when the 'detailed' encoding strategy is used. Logistic regression generally predicts a binary outcome or estimates the probability of certain event, but its relative poor performance with 'detailed' features could be because of the relative small size of gold standard positives currently available for training to the number of features, known as the problem of overfitting [38]. It is shown that when the training size increases, the logistic regression method become more precise and leads to improved prediction. However, RF method still outperforms LR even when a larger training set is used, indicating the superiority of RF method compared to other methods [38].

Conclusions

The incomplete and noisy PPI data from high-throughput techniques requires robust mathematical models and efficient computational methods to have the capability of integrating various types of features for protein interaction prediction problem. Although many distinct types of data are useful for protein interactions predictions, some data types may make little contribution to the prediction or may even decrease the predictive power, depending on the prediction tasks to be performed, the ways the datasets are encoded, the proteome coverage of these datasets and their reliabilities. Therefore, feature (dataset) selection is one of the challenges faced in the area of integrating multiple data sources for protein interaction inference. In addition to this, each method used for integration has its own limitations and captures different aspects of protein interaction information. Moreover, as the gold standard set used for evaluating prediction performance of different methods is incomplete, the prediction results should be validated by small scale experiments as the ultimate test of these methods. For example, when using a set of 98 proteins as the TAP-tagging baits, 424 predicted protein interactions based on integrating different features are validated and 44% of these are gold standard positives [40]. In addition, the Y2H screens of *Treponema pallidum* and *Campylobacter jejuni* confirmed about 33% of predicted interactions based on interologs [53].

Protein interaction prediction allows the construction of protein–protein interaction network, which has revealed global topology that relates to known biological properties [54]. Rather than different individual genomic features, different types of networks, including genetic interaction networks, signal transduction networks and transcription regulation networks, can be integrated as well at the network level. While the interactions supported by multiple types of networks promises to improve the accuracy of prediction and help detect functional modules within the network, it is worth noting that protein interactions are depicted as a multidimensional network instead of simple linear connections, which brings a great challenge in the field of systems biology.

Acknowledgements

This work was supported in part by NIH grants R01 GM59507, N01 HV28286, P30 DA018343, and NSF grant DMS 0714817.

References

1 Arkin, M. and Wells, J. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.* 3, 301–317
2 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
3 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4569–4574
4 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736
5 Li, S. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303, 540–543
6 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178
7 LaCount, D.J. *et al.* (2005) A protein interaction network of the malaria parasite plasmodium falciparum. *Nature* 438, 103–107
8 Von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
9 Hart, G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.* 7, 120
10 Scholtens, D. *et al.* (2008) Estimating node degree in bait-prey graphs. *Bioinformatics* 24, 218–224
11 Huang, H. *et al.* (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* 3, e214
12 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequence. *Science* 285, 751–753
13 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
14 Goh, C.S. and Cohen, F.E. (2002) Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.* 324, 177–192

15 Pazos, F. *et al.* (2003) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* 352, 1002–1015
16 Matthews, L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.* 11, 2120–2126
17 Dandekar, T. *et al.* (1998) Conservation of gene order: a finger print of proteins that physically interact. *Trend Biochem. Sci.* 23, 324–328
18 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. U.S.A.* 96, 2896–2901
19 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein–protein interaction. *Genome Res.* 12, 37–46
20 Ge, H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from *Saccaromyces cerevisiae*. *Nat. Genet.* 29, 482–486
21 Comeau, S. *et al.* (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* 20, 45–50
22 Sprinzak, E. and Margalit, H. (2001) Corrlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* 311, 681–692
23 Deng, M. *et al.* (2002) Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* 12, 1540–1548
24 Gomez, S.M. *et al.* (2003) Learning to predict protein–protein interactions from protein sequences. *Bioinformatics* 19, 1875–1881
25 Shi, T. *et al.* (2005) Computational methods for protein–protein interaction and their application. *Curr. Protein Pept. Sci.* 6, 443–449
26 Shoemaker, B.A. and Panchenko, A.R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PloS Comput. Biol.* 3, e43
27 Bonvin, A.M. (2006) Flexible protein–protein docking. *Curr. Opin. Struct. Biol.* 16, 194–200
28 Shoemaker, B.A. and Panchenko, A.R. (2007) Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PloS Comput. Biol.* 3, e42
29 Chiang, T. *et al.* (2007) Coverage and error models of protein–protein interaction data by directed graph analysis. *Genome Biol.* 8, R186
30 Liu, Y. *et al.* (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* 21, 3279–3285
31 Kim, I. *et al.* (2007) Bayesian methods for predicting interacting protein pairs using domain information. *Biometrics* 63, 824–833
32 Riley, R. *et al.* (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.* 6, R89
33 Salwinski, L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451
34 Guimaraes, K.S. *et al.* (2006) Predicting domain–domain interactions using a parsimony approach. *Genome Biol.* 7, R104
35 Finn, R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.* 34, D247–D251
36 Martin, S. *et al.* (2005) Predicting protein–protein interactions using signature products. *Bioinformatics* 21, 218–226
37 Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4372–4376
38 Qi, Y. *et al.* (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Struct., Funct. Bioinform.* 63, 490–500
39 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 1, 349–356
40 Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453
41 Lin, N. *et al.* (2004) Information assessment on prediction protein–protein interactions. *BMC Bioinform.* 5, 154
42 Lu, L.J. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* 15, 945–953
43 Yu, H. *et al.* (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* 19, 422–427
44 Bader, J.S. *et al.* (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85
45 Sharan, R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1974–1979
46 Ben-Hur, A. and Noble, W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics* 20, 3346–3352
47 Lee, I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science* 302, 1555–1558
48 Stolovitzky, G. *et al.* (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. N. Y. Acad. Sci.* 1115, 1–22
49 Mewes, M. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34, D169–D172
50 Suthram, S. *et al.* (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinform.* 7, 360
51 Davis, J. and Goadrich, M. (2007) The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA. pp. 233–240
52 Strobl, C. *et al.* (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* 8, 25
53 Rajagopala, S. *et al.* (2007) The protein network of bacterial motility. *Mol. Syst. Biol.* 3, 128
54 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113
55 Gavin, A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636
56 Krogan, N.J. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 637–643
57 Demeter, J. *et al.* (2007) The Stanford microarray database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.* 35, D766–D770
58 Gene Ontology Consortim. (2006) The gene ontology (GO) project in 2006. Nucleic Acids Res. 34, D322–326
59 Ghaemmaghami, S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature* 425, 737–741
60 Tong, A.H.Y. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science* 303, 808–813
61 Harbison, C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104
62 Sprinzak, E. *et al.* (2005) Characterization and prediction of protein–protein interactions within and between complexes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14718–14723
63 Zhang, L. *et al.* (2004) Predicting co-complexes protein pairs using genomic and proteomic data integration. *BMC Bioinform.* 5, 38
64 Lee, H. *et al.* (2006) An integrated approach to the prediction of domain–domain interactions. *BMC Bioinform.* 7, 269