Increasing the reliability of protein interactomes

Hon Nian Chua[1] and Limsoon Wong[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore
[2] National University of Singapore, School of Computing, Singapore 117590, Singapore

Protein interactions are crucial components of all cellular processes. An in-depth knowledge of the full complement of protein interactions in a cell, therefore, provides insight into the structure, properties and functions of the cell and its components. An accurate and comprehensive protein interaction network is, thus, an invaluable framework to study protein regulation in disease. Although the amount of protein–protein interaction data has grown significantly because of advances in high-throughput experimental techniques, these high-throughput methods are highly susceptible to noise. Therefore, computational techniques for assessing the reliability of a protein–protein interaction are highly desirable. We review here computational techniques for assessing and improving the reliability of protein–protein interaction data from these high-throughput experiments.

Introduction

Progress in the development of mass spectrometry, two-hybrid methods, genetic studies and other technologies has resulted in a rapid accumulation of data that provide a global description of the whole network of protein interactions – the interactome – for a given organism [1–4]. Protein interactions are crucial components of all cellular processes. Therefore, analysis on the interactome is expected to produce several types of useful information. For example, recent works have explored the use of protein interactome information to infer protein function [5–9]. Similarly, some recent work has attempted using protein interactome information to derive protein complexes and functional modules [10–15]. There are also many papers that propose using protein interactome information to analyze and predict drug targets [16–18]. These and other works provide insight into the structure, properties and functions of the cell and its components. Thus, an accurate and comprehensive protein interactome is an invaluable framework to study protein regulation in disease.

Although data on the protein interactome are being accumulated rapidly, several surveys and analyses [4,8,19–22] have revealed that interaction data obtained by high-throughput protein interaction assays, such as yeast two-hybrid experiments, contain a significant proportion of false positives and false negatives. There is, thus, a need to prioritize the protein–protein interactions reported in such assays for further validation by carefully focused small-scale experiments. Computational analysis techniques for assessing and ranking the reliability of protein–protein interactions are, hence, highly desirable. Many such techniques have emerged [23–38].

In this paper, we review some of these studies on increasing the reliability of protein interactome information. Specifically, the paper is organized as follows. The second section describes the idea of ranking the reliability of protein interactions based on the sharing of a common cellular localization or a common cellular role [20,32,39]. The following section describes the idea of ranking the reliability of protein interactions based on the reproducibility and nonrandomness of the observation of an interaction [8,39,40].

Related to the ideas of functional homogeneity, localization coherence and observational reproducibility are a large number of other approaches for estimating the reliability of protein interactions [23–29] based on the use of additional information, such as protein annotation, or the use of information from multiple assays. The fourth section describes the most general among them [38]. As the additional information required by these approaches may be unavailable, in the fifth section we describe some interesting and straightforward-to-use reliability indices that are based

Corresponding author: Wong, L. (wongls@comp.nus.edu.sg), (dcswls@nus.edu.sg)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.05.004

solely on the topology of the neighborhood of an interacting pair of proteins in the interactome [30–34].

Methods for ranking the reliability of reported protein interactions are, basically, the methods for detecting false positives from protein interaction assays. The detection of false negatives is considerably more difficult and is equivalent to the problem of predicting new protein interactions. The final section provides a review of approaches to this problem, including ideas such as gene-fusion events [41–43], interacting domains [44–47], interacting motifs [48–50], coevolution of proteins or residues [51–54] and the topology of protein–protein interaction networks [33,55,56].

Functional homogeneity and localization coherence

An early idea for assessing the reliability of an interacting protein pair reported by a high-throughput experiment is to use supporting evidence from the biological perspective. In particular, a pair of interacting proteins is generally expected to be localized to the same cellular component or to have a common cellular role [20,32]. Therefore, interacting protein pairs can be categorized into four groups. Those having both a common cellular localization and a common cellular role are considered most reliable. Those having no common cellular role and no common cellular localization are considered least reliable. Those having a common cellular localization or a common cellular role but not both are considered intermediate in reliability.

This rough grouping can be fine-tuned by a global estimate of the reliability of the entire protein interaction assay based on known annotation information. In particular, the reliability of an interacting pair of proteins that are both localized to a cellular compartment c can be estimated as the ratio of the number of interacting protein pairs in the assay that are both localized to c to the number of interacting protein pairs in the assay where at least one of the proteins in the pair is localized to c [8,39].

Naturally, we can set c above to be the union of all cellular compartments or the union of all cellular roles to obtain an estimate of the overall reliability of a protein interaction assay based on common cellular localization or common cellular role. Table 1 contains such an estimate for various protein interaction assays, computed based on common cellular role [8].

There are several shortcomings of using common cellular localization and common cellular role for assessing protein interaction reliability. Firstly, there may not be sufficient pairs of proteins having subcellular localization (or cellular role) c for a good estimate that is statistically reliable, if the granularity of c is too fine. Secondly, protein functional annotations and subcellular localization annotations are often incomplete. Thirdly, even if a pair of proteins localizes to the same cellular compartment or participate in the same cellular process, they may not interact in real life. In fact, there is a limit to the resolution of common cellular localization and common functional role. For example, if 20% of the proteins in an organism are localized to a common cellular compartment on average, then two proteins may have a non-negligible, though <20%, chance of not interacting even when they have a common cellular localization.

In addition to the idea of common cellular roles and common cellular localization, proteins having coexpressed genes are also more likely to interact with each other than with random proteins [57,58]. However, coexpression of genes usually does not correlate directly with physical interaction; genes are often coexpressed simply because they are activated during a similar phase of the cell cycle, or because they belong to similar pathways, or they participate in the regulation of one another through transcription factor interactions. Indeed, it has been observed that, without the use of additional information such as conserved coexpression in a second genome, a mere 18–28% of the most highly coexpressed human gene pairs correlate with protein interactions [36].

TABLE 1

Estimated reliability for each protein interaction assay in the GRID dataset [78], computed based on common cellular role [8]

Assay	Reliability
Affinity chromatography	0.82
Affinity precipitation	0.46
Biochemical assay	0.67
Dosage lethality	0.50
Purified complex	0.89
Reconstituted complex	0.50
Synthetic lethality	0.37
Synthetic rescue	1.00
Two hybrid	0.27

Observational reproducibility

An early idea to overcome the shortcomings of using common cellular localization and common cellular role for assessing protein interaction reliability is that of reproducibility. The idea of reproducibility is based on the reasonable assumption that an interaction that is observed in two or more separate experiments is more reliable than one that is observed in just one experiment.

Suppose the reliability, r_i, of each protein interaction assay, i, is known, or has already been estimated, as in the previous section. Assume also that a set $E_{u,v}$ of protein interaction assays that report an interacting pair of proteins (u, v) are independent. Then the reliability, $r_{u,v}$, of the interaction of (u, v) can be optimistically taken as the probability that at least one of the assays involved is reliable [8,39]. More formally:

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)^{n_{i,u,v}},$$

where $n_{i,u,v}$ is the number of times the pair (u, v) is observed to interact in assay i.

Another technique for assessing the reliability of an interacting pair from multiple assays is to compute a P-value based on the hypergeometric distribution [40]. Let there be a total of h interactions reported by various assays. Suppose proteins u and v are reported to participate in m and n interactions, respectively. Then the probability for u and v being reported to interact in k experiments at random is

$$P(k|n, m, h) = \frac{\binom{h}{k}\binom{h-k}{n-k}\binom{h-n}{m-k}}{\binom{h}{n}\binom{h}{m}}.$$

Then the *P*-value for u and v being reported to interact in k_0 experiments is

$$\sum_{k>k_0} P(k|n,m,h).$$

Two proteins that are reported to have interactions with a lot of proteins in a protein interaction assay have a naturally higher random chance to be reported to interact with each other. The *P*-value also serves as an estimate of this random chance.

The main shortcoming of using repeatability to assess the reliability of protein interactions is that multiple experiments must have been performed on the proteins. This shortcoming is a significant issue. For example, Sprinzak *et al.* [20] have shown that out of the 9347 interactions reported from a large number of experiments, a mere 570 are in the intersection of 4 different experiments, 1212 are in 3 different experiments, and 2360 are in 2 different experiments. Thus, using repeatability to assess reliability may cause a very large proportion of true positives to be mistakenly declared as false positives.

Fusion of multiple information types

There are a large number of other approaches for estimating error rates of protein interaction assays [23–25] and for ranking individual protein interacting pairs [26–29]. These approaches generally require the use of additional information, such as annotations on proteins or the use of information from multiple assays. The simplest among these approaches is the log-likelihood ratio [38]. Let D denote the event that a pair of proteins is 'linked' based on the chosen additional information. Then the log-likelihood ratio is defined as

$$\mathrm{LLR} = \ln\left(\frac{P(D|I)}{P(D|\sim I}\right),$$

where $P(D|I)$ and $P(D|\sim I)$ are the probability of observing the event D conditioned on protein pairs that are known to interact and known not to interact, respectively. As an example, D may be the event that the pair of proteins contains a pair of known interacting domains. As another example, D may be the event that the pair of proteins has gene expression profiles that are highly correlated.

As these approaches require the use of additional information, they inherit the basic weaknesses of using common cellular localization, common cellular role and/or the reproducibility of observations. Nevertheless, it is possible to overcome these weaknesses by a more effective fusion of additional information, especially information derived from multiple organisms based on evolutionary conservation principle [36,37]. A recent study by Ramani *et al.* [36] is an excellent illustration. They show that a significantly higher log-likelihood ratio can be obtained by considering coexpression that is conserved in a second genome. In particular, 49–59% of the 7000 most highly coexpressed human gene pairs that have coexpressed orthologs in a second organism correlates with protein interactions, while only 18–28% of the 7000 most coexpressed human gene pairs correlates with protein interactions when information on the coexpression of orthologs is ignored [36].

Topology of interactions

The family of methods from Saito *et al.* [30,31], Chen *et al.* [32–34] and Albert and Albert [35] avoid the weaknesses of using coherence of annotations and repeatability of observations by taking an entirely different approach that is more thought-provoking and, yet, more easily applied. They do not use annotations on proteins or information from multiple assays. Instead, they rank the reliability of an interaction between a pair of proteins primarily using the topology of the interactions between that pair of proteins and their neighbors within a short 'radius'.

The 'interaction generality index' (IG) of Saito *et al.* [31] is perhaps the earliest proposal of using the topology of interactions in the immediate neighborhood of a pair of interacting proteins to assess the reliability of that pair of proteins. It is based on the property of two-hybrid assays that a large number of false positives in two-hybrid assays are because of selfactivators and 'sticky' proteins that transactivate the reporter gene without actually interacting with their partners [1]. A characteristic of these self-activators and sticky proteins is that they usually appear to have a large number of interaction partners in the experiment; but these partners generally do not interact with each other. So Saito *et al.* [31] define the IG on a pair of reported interacting proteins (u, v) to be the number of isolated interaction partners that (u, v) have in the experiment. The larger this count is, the more unlikely that (u, v) is interacting.

In contrast to IG, which is based on a defect of two-hybrid assays, the 'interaction pathway reliability index' (IPR) of Chen *et al.* [33] relies on a set of assumptions on biological networks. They hypothesize that a biological function is generally performed by a highly interconnected network of interactions and that evolution favors adding interactions that shorten the pathways of the function. Therefore, a pair of proteins that is connected by a short alternate path of reliable interactions is likely to interact directly. Thus, Chen *et al.* [33] define the IPR on a pair of candidate interacting proteins (u, v) as the maximum reliability of the shortest nonreducible indirect path connecting (u, v). By assuming independence, the reliability of a nonreducible indirect path can be computed as a product of the rough estimates of the reliability of individual interactions in the path. Chen *et al.* [33] use IG as the rough estimate of the reliability of an individual interaction. IPR is applicable to a wider range of protein interaction assays, as its underlying assumption is based on a more general theory than IG.

Newer examples are indices that exploit a topological consequence of the functional homogeneity expected of true interacting protein pairs. As we have mentioned earlier, a pair of real interacting proteins is generally expected to have a common cellular role. It has been proposed that a pair of proteins having many common interaction partners has a high chance of sharing a common cellular role [8], because these two proteins must share some physical or biochemical characteristics that allow them to bind to these common interaction partners. The more proteins that they interact with in common, the higher is the chance that they belong to the same protein complex. Therefore, a reliability index for a pair of reported interacting proteins can be formulated in terms of the proportion of interaction partners that two proteins have in common.

A simple and direct formulation of such an index is the Czekanowski–Dice distance [7]. Czekanowski–Dice distance is defined as

$$\text{CD-Dist}_{u,v} = \frac{2|N_{u,v}|}{|N_u| + |N_v|},$$

where $N_{u,v}$ is the set of interaction partners shared by u and v, and N_u and N_v are, respectively, the set of interaction partners of u and v. Another example is the 'functional similarity weight' measure [8]. Functional similarity weight is defined as

$$\text{FSWeight}_{u,v} = \left(\frac{2|N_{u,v}|}{|N_u - N_v| + 2|N_{u,v}| + \lambda_{u,v}}\right) \times \left(\frac{2|N_{u,v}|}{|N_v - N_u| + 2|N_{u,v}| + \lambda_{v,u}}\right),$$

where $\lambda_{u,v}$ is a pseudocount to penalize similarity weights between protein pairs when any of the proteins has too few interacting partners. $\text{FSWeight}_{u,v}$ essentially refines $\text{CD-Dist}_{u,v}$ by giving the neighborhood of u and v equal weight.

Although these indices do not make use of either annotation or repeatability information, they are surprisingly effective [31–33]. The effectiveness of these indices can be gauged by their correlation with functional homogeneity and localization coherence. For example, as shown by Chen *et al.* [32], in Fig. 1, over 80% (70%) of the top 10% of protein interactions ranked by FSWeight (CD-Dist) have a common cellular role and over 90% (80%) of them have a common subcellular localization. Similar strong correlations are observed [32] between these indices and the gene expression correlation of highly ranked candidate interacting proteins, as well as between these indices and the number of times highly ranked candidate pairs are observed in multiple protein interaction assays.

Although IPR, CD-Dist, and FSWeight are defined purely in terms of the topology of the neighborhood of the protein pairs, it is possible to incorporate additional information, such as functional annotations and multiple experiments, if such additional information is available. For example, let $r_{u,v}$ be a rough estimate of the reliability of an interaction (u, v). We can interpret $r_{u,v}$ as the probability that u and v actually interact. Assuming independence, the probability of u and v having a common interaction partner w is thus $r_{u,w}r_{w,v}$. Then CD-Dist incorporating this information is

$$\text{CD-Dist}_{u,v} = 2\frac{\sum_{w \in N_u \cap N_v} r_{u,w}r_{w,v}}{\sum_{w \in N_u} r_{u,w} + \sum_{w \in N_v} r_{v,w}}.$$

IPR and FSWeight incorporating similar information can be similarly derived [8].

The main shortcoming of using indices like IG, IPR, CD-Dist and FSWeight to assess the reliability of protein interactions is that their performance edge becomes less significant when the input interaction network is sparse. This is because the number of direct and indirect interactions is much lower for sparser networks due to limited connectivity. Given the rapid growth of protein interaction data, these indices are expected to become increasingly more effective.

Predicting new protein interactions

The methods for ranking the reliability of protein interactions reported by high-throughput assays described in the previous sections are essential methods for detecting false positives in these assays. However, these assays are also known to produce a large number of false negatives. The identification of false negatives is equivalent to the problem of predicting new protein interactions. Many computational approaches have also been proposed to predict new protein interactions [59]. Various information have been used for this purpose, including protein primary structures and associated physicochemical properties [60], interacting domains [44–47], interacting motifs [48–50], gene-fusion events [41–43], coevolution of proteins or residues [51–54] and the topology of protein–protein interaction networks [33,55,56].

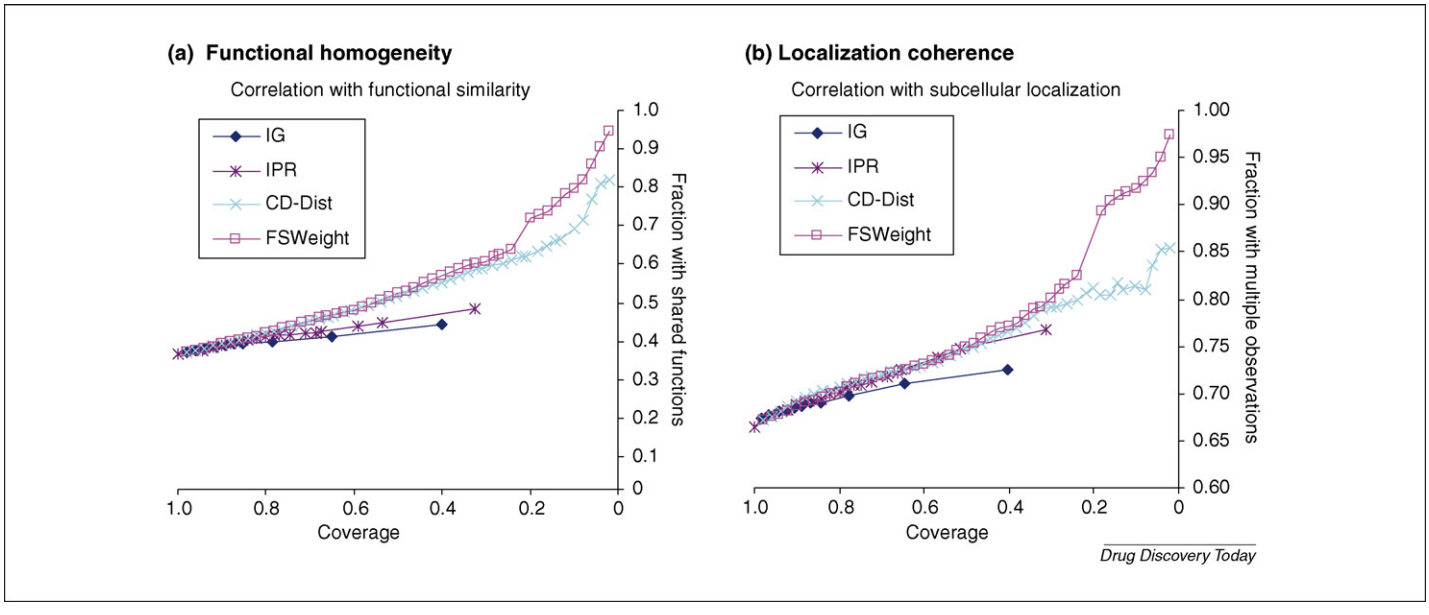

FIGURE 1

Comparison of IG [31], IPR [33], CD-Dist [7] and FSWeight [8] indices on their correlation with **(a)** function homogeneity and **(b)** localization coherence. The horizontal axis indicates the proportion of reported interacting protein pairs that satisfy a given IG, IPR, CD-Dist or FSWeight threshold. The vertical axis indicates the proportion of reported interacting protein pairs that share a common function or a common cellular localization at that threshold. This comparison was performed in [32] using data on 19,452 interactions in yeast from the GRID database [78]. We can see, for example, over 80% of the top 10% of interacting protein pairs ranked by FSWeight have a common cellular role and over 90% of them have a common subcellular localization.

A gene-fusion event refers to an observation of two genes that are separate in one species, but are fused as a single gene in a second species. Such a fusion event is hypothesized as an indication that the products of the two genes interact. Several studies have used this approach to predict interactions between proteins [41–43] with some degree of success. A major limitation of this gene-fusion approach is its poor sensitivity, because not many pairs of interacting proteins have fused homologs.

Approaches based on interacting domains [44–47] or interacting motifs [48–50] are more direct. The central idea of these approaches is a lock-and-key model [61] whereby a pair of proteins interacts through their complementary binding domains. Interactions of the complementary binding domains are in turn mediated by short sequences of residues that form the contact interfaces between the two complementary binding domains [62,63]. Furthermore, there are only a limited number of interaction types, such that two proteins are likely to interact whenever an interaction type occurs in the protein pair [64]. For example, Li *et al.* [50] first use frequent pattern mining technique to enumerate bicliques in a protein interaction network. They then use PROTOMAP to extract a pair of motifs from sequences in the two vertex sets of each biclique. If a pair of such motifs is statistically overrepresented in known interacting protein pairs, it is then proposed as a binding motif pair for identifying additional interacting proteins.

Another major family of approaches for predicting interacting protein pairs is that of coevolution [51–53]. The coevolution of an interacting protein pair is based on the hypothesis that the interaction sites of these proteins are under pressure to coevolve [65]. That is, the mutations in one protein must be compensated by the mutations in the other protein [66]. Therefore, the corresponding phylogenetic trees of the interacting proteins should show a greater degree of similarity than noninteracting proteins are expected to show. Typically, the similarity of the phylogenetic trees of two candidate interacting proteins is quantified based on the correlation between the distance matrices used to construct the trees [51–53]. To obtain a good distance matrix, a high-quality multiple alignment of the orthologs of the protein is necessary. Hence, a shortcoming of these approaches is the need for such a high-quality multiple alignment of orthologs from the same species for the two proteins under consideration.

Furthermore, a given protein may interact with many others. Then it must coevolve with all of them. Consequently, its phylogenetic tree is a composite of the influence of all of its interaction partners. This issue further limits the performance of methods [51–53] that rely on the similarity of the phylogenetic trees of the two proteins. Juan *et al.* [54] outline an exciting recent refinement, addressing this weakness of approaches based on the coevolution principle. For each protein u, Juan *et al.* compute a vector S_u of the pairwise similarities of the phylogenetic tree of u with all other proteins. Then for each pair of proteins (u, v), they determine the correlation of S_u to S_v. If the correlation is significant, the pair (u, v) is predicted to interact. This refined coevolution approach provides drastically better accuracy and coverage than earlier coevolution approaches [54].

Another interesting group of approaches rely on the topology of protein–protein interaction networks [33,55,56]. The simplest methods from this group are those that identify a subgraph in the network that is basically a clique with a small number of missing edges and propose those missing edges to be the new interactions [55]. The basis for these methods is the assumption that, when two proteins have a lot of partners in common, they should participate in the same complex with these partners. The more sophisticated methods from this group are those that realize that the topology-based reliability measures mentioned in Topology of Interactions section – viz., IPR, CD-Dist and FSWeight – can be applied to a pair of proteins (u, v), even when (u, v) are not reported by a protein interaction assay. For example, Pei and Zhang [56] and Chen *et al.* [33] apply two variations of the IPR index on every pair of proteins (u, v) and declare those that score well to be interacting.

Concluding remarks

The quantity and variety of protein interaction data have increased rapidly since the publication of two yeast interactome maps based on the yeast two-hybrid technology eight years ago [2,67]. In particular, two-hybrid-based interactome maps have been generated for model organisms such as *C. elegans* [68], Drosophila [26], bacteria [69–71] and human [72,73]; and proteome-scale interactome maps have also been generated for yeast by TAP-MS experiments [74–76]. Nevertheless, their quality has much to be improved [4,8,19–22].

We have discussed, in depth, several approaches – based on principles, such as functional homogeneity; localization coherence; observational repeatability and topology of interactions – for assessing the reliability of protein–protein interactions and thus detecting false positives reported by various high-throughput experiments. In particular, we have highlighted that it is possible to rank the reliability of a protein interaction pair by the local topology around the pair in the protein interaction network [30–34].

We have also surveyed several approaches for predicting new protein interactions and, thus, detecting false negatives in protein interaction assays. The approaches reviewed include those that are based on principles, such as interacting domains [44–47], interacting motifs [48–50], gene-fusion events [41–43], coevolution of proteins or residues [51–54] and the topology protein–protein interaction networks [33,55,56].

The ability of these computational approaches to identify false-positive and false-negative protein interactions is quite remarkable. For example, as shown in Table 1, a good experimental assay such as affinity chromatography has a reliability of 82% – and thus an estimated false-positive rate of 18% – using common cellular role as the yardstick. As shown in Fig. 1, this is matched by the FSWeight index [32] that is based solely on the topology of the local neighborhood of interacting proteins. Similarly, with respect to false negatives, Juan *et al.* [54] has reported impressive results that the top 100 interacting protein pairs predicted by their coevolution method are completely correct and up to 40% of their top 500 predictions are correct.

It is also interesting to mention the outcome of the protein–protein subnetwork challenge under the DREAM2 Project held in late 2007 [77]. In this challenge, the participants were asked to predict the protein–protein interaction network involving 47 yeast genes. In parallel, a very stringent series of yeast two-hybrid experiments were repeated three times to obtain a gold standard

positive and negative sets of interactions involving these 47 genes. The gold standard positive set comprises interactions that were observed in all three repetitions. The gold standard negative set comprises interactions that were observed in none of the three repetitions. The top team – relying on an integration of many of the techniques described in this paper – achieved a performance of 0.63 area under the ROC curve, which is a very competitive performance compared to popular experimental techniques.

Although these computational approaches to identify false positives and false negatives protein interactions are competitive, there is still room for improvement. For example, the availability of large-scale interactome information on multiple model organisms and a variety of corollary events that accompany *in vivo* protein interactions have not been exploited fully. Also, most of the resulting interactome maps are still essentially an *in vitro* scaffold. Further progress in computational analyses techniques and experimental methods is needed to reliably deduce *in vivo* protein interactions, to distinguish between permanent and transient interactions, to distinguish between direct protein binding from membership in the same protein complex and to distinguish protein complexes from functional modules.

Acknowledgements

This work is supported in part by a Singapore MOE AcRF Tier 1 grant (Wong) and an A*STAR NGS scholarship (Chua).

References

1 Ng, S-K. and Tan, S-H. (2004) Discovering protein–protein interactions. *J. Bioinform. Comput. Biol.* 1, 711–741
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 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
5 Schwikowski, B. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261
6 Hishigaki, H. *et al.* (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* 18, 521–523
7 Brun, C. *et al.* (2003) Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. *Genome Biol.* 5, R6
8 Chua, H.N. *et al.* (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* 22, 1623–1630
9 Chua, H.N. *et al.* (2007) Using indirect protein interactions for the prediction of gene ontology functions. *BMC Bioinformatics* 8 (Suppl. 4), S8
10 Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12123–12128
11 Krause, R. *et al.* (2003) A comprehensive set of protein complexes in yeast: mining large scale protein–protein interaction screens. *Bioinformatics* 19, 1901–1908
12 Bader, G.D. and Hogue, C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2
13 King, A.D. *et al.* (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* 20, 3013–3020
14 Brohee, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics* 7, 488
15 Chua, H.N. , Ning, K. , Sung, W.K. , Leong, H.W. and Wong, L. (2008) Using indirect protein-protein interactions for protein complex prediction. *J Bioinform Comput Biol* 6 (3), 435–466
16 Schachter, V. (2002) Protein-interaction networks: from experiments to analysis. *Drug Discov. Today* 7, S48–S54
17 Ruffner, H. *et al.* (2007) Human protein–protein interaction networks and the value for drug discovery. *Drug Discov. Today* 12, 709–716
18 Guimera, R. *et al.* (2007) A network-based method for target selection in metabolic networks. *Bioinformatics* 23, 1616–1622
19 Legrain, P. *et al.* (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.* 17, 346–352
20 Sprinzak, E. *et al.* (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.* 327, 919–923
21 Deng, M. *et al.* (2002) Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* 12, 1540–1548
22 Hart, G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.* 7, 120
23 Bader, J.S. *et al.* (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85
24 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
25 Patil, A. and Nakamura, H. (2005) Filtering high-throughput protein–protein interaction data using a combination of genomic features. *BMC Bioinformatics* 6, 100
26 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736
27 Samanta, M.P. and Liang, S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12579–12583
28 Schlitt, T. *et al.* (2003) From gene networks to gene function. *Genome Res.* 13, 2568–2576
29 Martin, S. *et al.* (2005) Predicting protein–protein interactions using signature products. *Bioinformatics* 21, 218–226
30 Saito, R. *et al.* (2003) Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics* 19, 756–763
31 Saito, R. *et al.* (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res.* 30, 1163–1168
32 Chen, J. *et al.* (2006) Increasing confidence of protein–protein interactomes. In *Proceedings of 17th International Conference on Genome Informatics*, Yokohama, Japan, December 2006. pp. 284–297
33 Chen, J. *et al.* (2006) Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics* 22, 1998–2004
34 Chen, J. *et al.* (2006) NeMoFinder: dissecting genome wide protein–protein interactions with repeated and unique network motifs. In *Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, August 2006. pp. 106–115
35 Albert, I. and Albert, R. (2004) Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics* 20, 3346–3352
36 Ramani, A.K. *et al.* (2008) A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. *Mol. Syst. Biol.* 4, 180
37 Liu, Y. *et al.* (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* 21, 3279–3285
38 Ramani, A.K. *et al.* (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* 6, 40
39 Nabieva, E. *et al.* (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21 (Suppl. 1), i302–i310
40 Hart, G.T. *et al.* (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* 8, 236
41 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
42 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
43 Tsoka, S. and Ouzounis, C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat. Genet.* 26, 141–142
44 Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein–protein interactions. *J. Mol. Biol.* 311, 681–692
45 Kim, W.K. *et al.* (2002) Large scale statistical prediction of protein–protein interaction by potentially interacting domain (pid) pair. In *Proceedings of 13th International Conference on Genome Informatics (GIW2002)* pp. 42–50
46 Han, D.S. *et al.* (2004) PreSPI: a domain combination based prediction system for protein–protein interaction. *Nucleic Acids Res.* 32, 6312–6320
47 Wojcik, J. and Schächter, V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17 (Suppl. 1), S296–S305

48 Tong, A.H.Y. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321–324

49 Aytuna, A.S. *et al.* (2005) Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21, 2850–2855

50 Li, H. *et al.* (2006) Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics* 22, 989–996

51 Goh, C.S. *et al.* (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* 299, 283–293

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

53 Pazos, F. *et al.* (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* 352, 1002–1015

54 Juan, D. *et al.* (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U. S. A.* 105, 934–939

55 Yu, H. *et al.* (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* 22, 823–829

56 Pei, P. and Zhang, A. (2005) A topological measurement for weighted protein interaction network. In *Proceedings of 4th International Computational Systems Bioinformatics Conference*, Stanford, CA, August 2005. pp. 268–278

57 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

58 Fraser, H.B. *et al.* (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9033–9038

59 Shoemaker, B.A. and Pachenko, A.R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.* 3, e43

60 Bock, J.R. and Gough, D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics* 17, 455–460

61 Morrison, J.L. *et al.* (2006) A lock-and-key model for protein–protein interactions. *Bioinformatics* 22, 2012–2019

62 Sheu, S.H. *et al.* (2005) PRECISE: a database of predicted and consensus interaction sites in enzymes. *Nucleic Acids Res.* 33 (Database Issue), D206–D211

63 Keskin, O. *et al.* (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* 345, 1281–1294

64 Aloy, P. and Russell, R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* 22, 1317–1321

65 Fryxell, K.J. (1996) The coevolution of gene family trees. *Trends Genet.* 12, 364–369

66 Pazos, F. *et al.* (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.* 271, 511–523

67 Ito, T. *et al.* (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1143–1147

68 Li, S. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303, 540–543

69 Rain, J.C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215

70 Parrish, J.R. *et al.* (2007) A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol.* 8, R130

71 Rajagopala, S.V. *et al.* (2007) The protein network of bacterial motility. *Mol. Syst. Biol.* 3, 128

72 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178

73 Stelzl, U. *et al.* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968

74 Gavin, A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636

75 Krogan, N.J. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 637–643

76 Collins, S.R. *et al.* (2007) Towards a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 6, 439–450

77 The results of the DREAM2 Challenge are not published yet. However, the reader may visit http://www.wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project for more information

78 Breikreutz, B-J. *et al.* (2003) The GRID: the general repository for interaction datasets. *Genome Biol.* 4, R23