Increasing Confidence of Protein-Protein Interactomes

Jin Chen[1] chenjin@comp.nus.edu.sg
Hon Nian Chua[2] g0306417@nus.edu.sg
Wynne Hsu[1] whsu@comp.nus.edu.sg
Mong-Li Lee[1] leeml@comp.nus.edu.sg
See-Kiong Ng[3] skng@i2r.a-star.edu.sg
Rintaro Saito[4] rsaito@sfc.keio.ac.jp
Wing-Kin Sung[1] ksung@comp.nus.edu.sg
Limsoon Wong[1,2] wongls@comp.nus.edu.sg

[1] National University of Singapore, School of Computing, 3 Science Drive 2, Singapore 117543
[2] National University of Singapore, Graduate School for Integrated Sciences and Engineering, 10 Medical Drive, Singapore 117597
[3] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[4] Keio University, Institute for Advanced Biosciences, 14-1 Baba-cho, Tsuruoka, Yamagata, Japan 997-0035

Abstract

High-throughput experimental methods, such as yeast-two-hybrid and phage display, have fairly high levels of false positives (and false negatives). Thus the list of protein-protein interactions detected by such experiments would need additional wet laboratory validation. It would be useful if the list could be prioritized in some way. Advances in computational techniques for assessing the reliability of protein-protein interactions detected by such high-throughput methods are reviewed in this paper, with a focus on techniques that rely only on topological information of the protein interaction network derived from such high-throughput experiments. In particular, we discuss indices that are abstract mathematical characterizations of networks of reliable protein-protein interactions—e.g., "interaction generality" (*IG*), "interaction reliability by alternatve pathways" (*IRAP*), and "functional similarity weighting" (*FSWeight*). We also present indices that are based on explicit motifs associated with true-positive protein interactions—e.g., "new interaction generality" (IG_2) and "meso-scale motifs" (*NeMoFinder*).

Keywords: Protein-protein interaction network, graph mining, network motif.

1 Introduction

Progress in high-throughput experimental techniques in the past decade has resulted in a rapid accumulation of protein-protein interaction data [33, 13, 21]. However, recent surveys [16, 21, 32, 9] have revealed that interaction data obtained by the popular yeast-two-hybrid assay may contain as much as 50% false positives and false negatives. As a result, further carefully-focused small-scale experiments are often needed to complement the large-scale methods to validate the detected interactions. However, the vast interactomes require much more scalable and inexpensive approaches.

We review in this paper several methods for assessing the reliability of a protein-protein interaction, given a graph derived from high-throughput protein interaction experiments. The pairs of "interacting" proteins that are ranked highly by these methods are shown more likely to be true positive interacting pairs. Conversely, the pairs of proteins that are ranked lowly by these methods are shown more likely to be false positives. The most interesting feature of these methods is that they are able to rank the reliability of an interaction between a pair of proteins using only the topology of the interactions between that pair of proteins and their neighbours within a short "radius".

The methods considered in this paper can be roughly divided into two groups. The first group comprises the "interaction generality" (*IG*), "interaction reliability by alternate pathways" (*IRAP*), and "functional similarity weighting" (*FSWeight*) indices. This group of indices attempt to provide abstract mathematical characterizations of networks of reliable protein-protein interactions. They are based on the hypothesis that true-positive interactions are likely to be characterized by dense cross connections in the derived interaction graph. The second group comprises the "new interaction generality" (IG_2) and "meso-scale motifs" (*NeMoFinder*) indices. This group of indices attempt to provide explicit motifs of network connections that are associated with reliable protein interactions. They also demonstrate how to systematically use such motifs for assessing the reliability of protein-protein interaction data.

The remainder of this paper is organized as follows. Section 2 introduces *IG* [29] and IG_2 [30]. They are among the first generation of indices for assessing reliability of protein interactions based on topological information of the derived interaction graph. We also present here a framework for evaluating the performance of these and other indices based on repeatability of observations, function homogeneity, localization coherence, and gene expression correlation. Section 3 describes *IRAP* [5], which is a significant improvement to *IG*. It goes beyond the immediate neighbourhood of an interacting pair of proteins in the interaction graph, as it considers also possibly long alternate paths linking that pair. Section 4 presents *FSWeight*. Although *FSWeight* is devised originally for protein function prediction [8], it improves upon *IRAP* in the context of assessing protein interaction reliability in two ways: it is more accurate when the interaction graph is dense; it is also more efficient to compute. Section 5 presents meso-scale motifs (*NeMoFinder*) [7]. These motifs are "meso scale" in the sense that they generally involve 5–25 proteins. Their discovery is computationally demanding. They are a significant improvement of the small-scale motifs used in IG_2. Section 6 contains a summary of the paper and some comments on application of these indices for false negative detection, as well as some historical remarks.

2 Interaction Generality

There is no better place to start our review on the detection of false positives in protein-protein interactomes, based on topology of derived interaction graphs, than the two really remarkable papers of Saito, Suzuki, and Hayashizaki [29, 30]. These two papers have introduced four important ideas. Firstly, they postulate that it is possible to evaluate the reliability of a protein interaction using solely information based on the topology of the derived interaction graph. This is an important conceptual contribution, because previous approaches have all used information beyond the derived interaction graph. Secondly, they demonstrate how to use a simple topological measure—the interaction generality (*IG*) index—to capture a biologically intuitive characteristic of false protein-protein interactions in high-throughput yeast-two-hybrid experiments. Thirdly, they also show how to use statistical learning to identify simple topological motifs—the new interaction generality (IG_2) index—that are associated with true protein-protein interactions. Lastly, they have established a rigorous framework—based on repeatability of observations, function homogeneity, localization coherence, and gene expression correlation—for validating such protein-protein interaction reliability indices. We discuss these important concepts in this section.

Let us start with the first interaction generality index (*IG*) [29].

Definition 2.1 (Interaction Generality Index, *IG*) *The "interaction generality" index $IG^G(X,Y)$, on a pair of potentially interacting proteins X and Y in an interaction graph G, is defined as the number of proteins that directly interact with exactly one of X or Y and nothing else in G. That is,*

$$IG^G(X,Y) = 1 + |\{\{X',Y'\} \in G \mid X' \in \{X,Y\},\ Y' \notin \{X,Y\},\ deg^G(Y') = 1\}|$$

where $deg^G(U) = |\{V \mid \{U,V\} \in G\}|$ is the degree of the node U in the undirected graph G.

The intuition underlying the *IG* index is a simple one. In a yeast-two-hybrid experiment, two target proteins are fused separately with a DNA-binding domain and a transactivation domain of a transcription factor; the expression of the reporter gene is then revealed if the two target proteins interact [26]. This method has been applied to construct the yeast protein-protein interactome by Uetz et al. [33] and Ito et al. [13] in particular. There are a large number of false positives in such experiments, due to self activators and "sticky" proteins that transactivate the reporter gene without actually interacting with their partners. A characteristic of these self activators and sticky proteins is that they seem to have a large number of interaction partners in the yeast-two-hybrid experiment, but these partners typically do not interact with each others. Thus one of the two proteins in such a false-positive interaction tends to be a "hub" that has many partners that do not interact with each other nor with the second protein in the pair. Furthermore, the higher the number of such non-mutually interacting partners that an interacting pair has, the more likely the pair is false positive. The *IG* index is a clear and direct exploitation of this characteristic; and the smaller the *IG* index value is, the more likely the interaction.

On the other hand, true-positive interactions are likely to be characterized by dense cross connections in the derived interaction graph. A direct use of the idea of "dense cross connections" is to treat a pair of interacting proteins that are in a clique-like subnetwork to be true positive [11, 34]. However, given the constraint of limited available interaction area on the surface of proteins, it is unlikely that two proteins having a large number of interaction partners can directly interact with each other. So, real interactions tend to involve two proteins where one has a large number of other partners and one has only a few other partners [20], as opposed to a clique-like subnetwork. This "many-few" property can be thought of as a form of "motifs" that characterize true interacting pairs. This inspires Saito et al. [30] to propose a number of small local motifs and use them in a new interaction generality index (IG_2) to characterize true-positive interactions.

Definition 2.2 (New Interaction Generality Index, IG_2) *The new interaction generality index $IG_2^G(X,Y)$, on a pair of potentially interacting proteins X and Y in an interaction graph G, is defined as a weighted sum of the 5 local topological configurations τ_1, ..., τ_5 shown in Figure 1:*

$$IG_2^G(X,Y) = \sum_{i=1}^{5} \lambda_i * |\{X' \mid \{X',Y'\} \in G,\ Y' \in \{X,Y\},\ \tau_i^G(X',X,Y)\}|$$

where λ_i is the weight for configuration τ_i, and $\tau_i^G(X',X,Y)$ means X' is in configuration τ_i in graph G with respect to the pair X and Y.

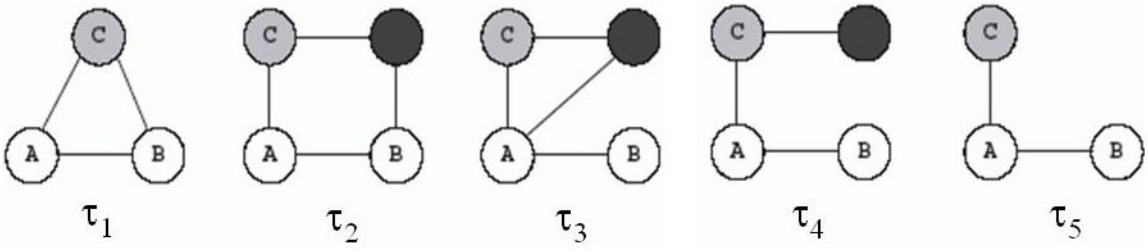

Figure 1: The 5 small local motifs proposed by Saito et al. [30]. These motifs are based on the possible topological configurations of a protein C that interacts with the target pair of interacting proteins A and B.

Specifically, Saito et al. [30] propose 5 small local motifs (τ_1, ..., τ_5) shown in Figure 1. These motifs are based on the possible topological configurations of a third protein that interacts with the

target pair of interacting proteins. Of these 5 motifs, Saito et al. postulate that those that form a closed loop (τ_1, τ_2, τ_3) are likely to be associated with true positives, and those that do not form a closed loop (τ_4, τ_5) with false positives. Saito et al. further use a statistical learning approach to determine the weights (λ_1, ..., λ_5) of these motifs in distinguishing true- vs false-positive protein interactions. Furthermore, the more negative the IG_2 index value is, the more likely the interaction.

How effective are IG and IG_2 as indices for assessing the reliability of a protein-protein interaction? A thorough evaluation based on laboratory experiments is not feasible due to constraints on cost, materials, and time. Saito et al. [29, 30] therefore propose an alternate rigorous framework based on repeatability of observations, function homogeneity, localization coherence, and gene expression correlation. We discuss this framework below.

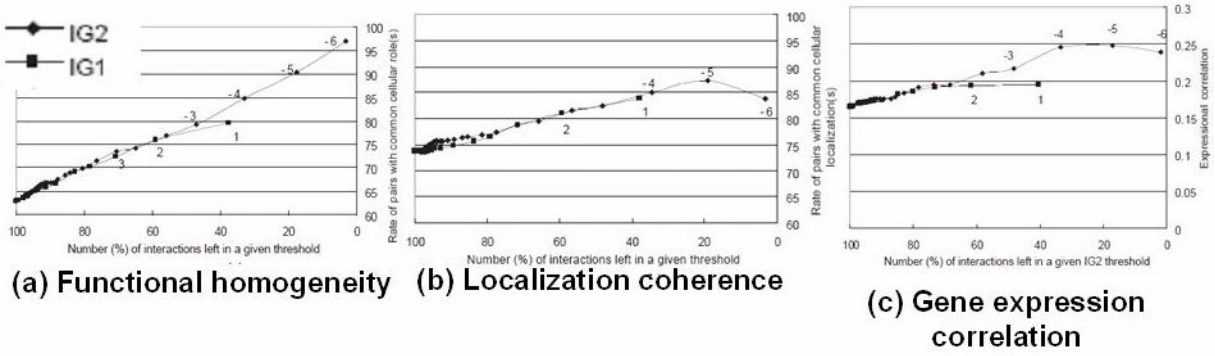

Figure 2: Comparison of IG and IG_2 indices on their correlation with (a) function homogeneity, (b) localization coherence, and (c) gene expression correlation. The comparison is performed by Saito et al. [30] using the yeast protein interaction data of Ito et al. [13], Uetz et al. [33], and MIPS (2000) [22].

A protein-protein interaction that is observed in several different experiments is obviously more reliable than one that is observed in only one experiment. Thus a good reliability index for protein-protein interactions should correlate with the likelihood that a protein-protein interaction is observed in multiple experiments. Indeed, both IG and IG_2 show such a correlation [29, 30]. For example, while 30% of those proteins interactions observed by Ito et al. [13] but not by Uetz et al. [33] achieve the best IG index value of 1, 50% of protein interactions observed by both Ito et al. and Uetz et al. achieve the best IG index value of 1 [29].

The cellular functions of proteins participating in a genuine biological interactions are likely to be similar [32]. Thus a good reliability index for protein-protein interactions should correlate with the likelihood of two interacting proteins having a common cellular function. Indeed, both IG and IG_2 show such a correlation [29, 30]. For example, as can be seen in Figure 2(a), while 63% of all interacting proteins from Ito et al. and Uetz et al. share a common general cellular function, 90% of those having an IG_2 index value of –5 share a common function.

With the exception of proteins involved in pathways such as signaling and transportation, the cellular localizations of proteins participating in a true interaction are expected to be the same [32]. Thus a good reliability index for protein-protein interactions should correlate with the likelihood of two interacting proteins having a common cellular localization. Indeed, both IG and IG_2 show such a correlation [29, 30]. For example, as can be seen in Figure 2(b), while 75% of all interacting proteins from Ito et al. and Uetz et al. share a common cellular localization, 87% of those having an IG_2 index value of –5 share a common cellular localization.

The average correlation coefficient of gene expression profiles that corresponds to interacting protein pairs is higher than those of random protein pairs [12]. Thus a good reliability index of protein-protein interactions should correlate with the likelihood of two interacting proteins having highly

correlated gene expression profiles. Indeed, both IG and IG_2 show such a correlation [29, 30]. For example, as can be seen in Figure 2(c), while 16% of all interacting proteins from Ito et al. and Uetz et al. have correlated gene expression profiles, 25% of those having an IG_2 index value of –5 have correlated gene expression profiles.

The correlations of IG and IG_2 to repeatability of observations, function homogeneity, localization coherence, and gene expression correlation support the hypothesis that they correlate to the reliability of protein-protein interactions. However, it is also clear from Figure 2 that the correlation to reliability is far from perfect, and there is much ground for improvement. We review several such improvements in the next three sections.

3 Interaction Reliability by Alternate Pathways

Recall that the self activators and sticky proteins in a yeast-two-hybrid experiment have a large number of interaction partners that do not interact with each others. A consequence is that such a false-positive interaction is "orphaned" in the sense that there is no other indirect path of interactions between the two proteins in the interaction. The *IRAP* and *FSWeight* indices to be discussed in this section and in Section 4 are both exploitations of the contra-positive of this consequence—there tend to be some alternative paths connecting a pair of real interacting proteins.

Thus, Chen et al. [5] define the *IRAP* index to link the reliability of an interacting pair to the "confidence" of the strongest irreducible alternate path connecting the pair. Here, a path ϕ connecting a pair of proteins X and Y is irreducible if there is no shorter path ϕ' connecting X and Y that shares some common intermediate nodes with the path ϕ. The confidence of a path can be estimated from the confidence $conf^G(U,V)$ of individual edges (U,V) in the path in a number of ways. One possibility is to assume independence, and define the confidence of a path as the product of the confidence of individual edges in the path, as given in the definition of the *IRAP* index below. We can also think of this definition of the *IRAP* index as a way to refine the confidence of the edges in an interaction graph based on an initial estimate of the confidence of the edges.

Definition 3.1 (Interaction Reliability by Alternate Pathways Index, *IRAP*) *The "interaction reliability by alternate pathways" index $IRAP^G(X,Y)$, on a pair of proteins X and Y in an interaction graph G, is defined as the maximum probability of a reliable irreducible alternate path connecting X and Y. Specifically,*

$$IRAP^G(X,Y) = \max_{\phi \in \Phi^G(X,Y)} \prod_{\{U,V\} \in \phi} conf^G(U,V)$$

where $\Phi^G(X,Y)$ is the set of all posible non-reducible paths between X and Y, but excluding the direct edge connecting X and Y; and $conf^G(U,V)$ is an estimated confidence of the edge $\{U,V\}$ in the interaction graph G. Thus, the higher the IRAP index value that an edge has, the more likely the interaction.

But what should we use as an initial estimate of the confidence of the edges? Recall from the previous section that edges having high IG index values tend to be unreliable. Thus the IG index value of an edge normalized by the maximum possible IG index value of the interaction graph can be thought of as a kind of probability that the edge is unreliable. Then taking the complement of this value gives us a kind of probability that the edge is reliable. This leads Chen et al. [5] to define

$$conf^G(U,V) = \left(1 - \frac{IG^G(U,V)}{IG^G_{\max}}\right)$$

where $IG^G_{\max} = \max\{IG^G(U,V) \mid \{U,V\} \in G\}$ is the maximum IG index value in G.

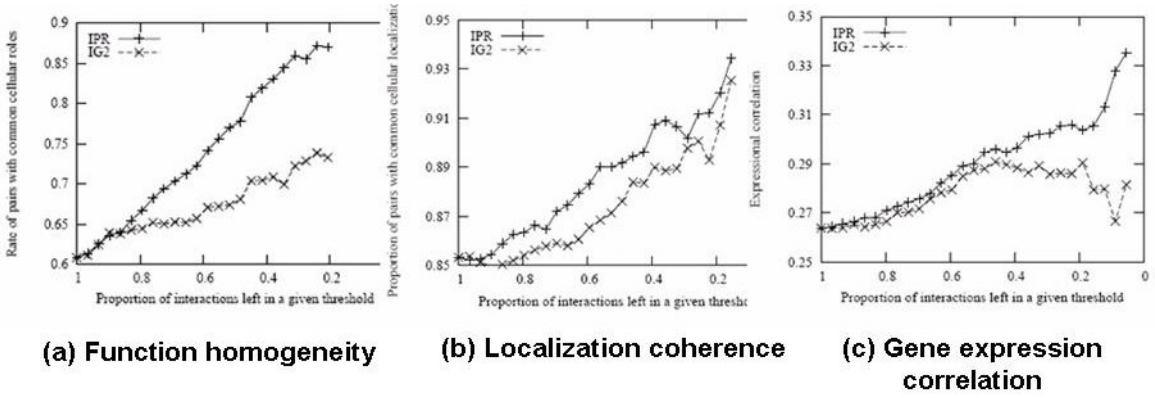

Figure 3: Comparison of IG_2 and *IRAP* (denoted as IPR here) on their correlation with (a) function homogeneity, (b) localization coherence, and (c) gene expression correlation. The comparison is performed by Chen et al. [5] using the yeast protein interaction data of Ito et al. [13], Uetz et al. [33], and MIPS (2002) [23].

Figure 3 compares the effectiveness of *IRAP* to IG_2 as indices for assessing the reliability of a protein-protein interaction. In terms of function homogeneity, it can be seen from Figure 3(a) that *IRAP* is significantly better than IG_2. For example, nearly 90% of the top 20% of interacting pairs ranked according to *IRAP* are annotated to have a common cellular role; in contrast, less than 75% of the top 20% of interacting pairs ranked according to IG_2 are annotated to have a common cellular role. Similarly, though to a lesser extent, *IRAP* exhibits better localization coherence and higher gene expression correlation than IG_2, as shown in Figures 3(b) and (c).

There are a number of possible variations to the way *IRAP* is formulated. For example, instead of the confidence of the strongest irreducible path, one can consider the confidence of the strongest shortest alternate path, the combined confidence of all shortest alternate paths, or even the combined confidence of all alternate paths. Another example, instead of using *IG* to define $conf^G(U, V)$, one can consider IG_2, or one can base it on the type of biological experiments that were performed to detect the interaction represented by the edge, as different types of such experiments have different known range of errors [32]. A notable use of these variations is Pei and Zhang [27], who use the combined confidence of all alternate paths up to a length k, and who base $conf^G(U, V)$ on the type of biological experiments performed to detect the interaction represented by the edge (U, V).

While *IRAP* [5] and variations such as that of Pei and Zhang [27] are more effective than *IG* and IG_2, there is still considerable room for improvement in terms of their accuracy for assessing reliability of protein-protein interactions. Furthermore, they are many orders of magnitude more computationally complex and time consuming to obtain than simple indices like *IG* and IG_2.

4 Functional Similarity Weighting

The computational inefficiency of *IRAP* [5] and its variantions such as that of Pei and Zhang [27] is largely due to the need to consider many alternative paths. In fact, for practical reason, Pei and Zhang consider only alternative paths of length up to 5. This motivates us to look for more efficient alternatives that are based on a similar principle. The obvious possibility is to look at alternative paths that go through only one intermediate node. The number of such paths between an interacting pair of proteins is equal to the number of common interaction partners between the two proteins. The existence of a common interaction partner between two proteins X and Y actually makes X simultaneously both an immediate partner of Y and an indirect partner of Y. As shown in Chua et

al. [8], such a pair of proteins are highly likely to be functionally linked; thus they are highly likely to be a true positive interacting pair.

Therefore a reliability index for protein-protein interactions 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 (*CD-Dist*), originally used for the purpose of protein function prediction from protein interaction graphs [4].

Definition 4.1 (Czekanowski-Dice Distance Index, *CD-Dist*) *The Czekanowski-Dice distance index $CD\text{-}Dist^G(X,Y)$, on a pair of proteins X and Y in an interaction graph G, is defined as the proportion of partners that the two proteins have in common. Specifically,*

$$CD\text{-}Dist^G(X,Y) = \frac{2 * |N^G(X) \cap N^G(Y)|}{|N^G(X)| + |N^G(Y)|}$$

where $N^G(U) = \{V \mid \{U,V\} \in G\}$ is the set of neighbours of U in the graph G.

As two proteins may have very different numbers of partners, it is useful to refine the *CD-Dist* index to account for this situation. Chua et al. [8] introduce a variation called functional similarity weight (*FS*), also for the purpose of protein function prediction from protein interaction graphs:

Definition 4.2 (Functional Similarity Weight Index, *FS*) *The "functional similarity weight" index $FS^G(X,Y)$, on a pair of proteins X and Y in an interaction graph G, is defined as*

$$\begin{aligned} FS^G(X,Y) &= \frac{2 * |N^G(X) \cap N^G(Y)|}{|N^G(X)| + |N^G(X) \cap N^G(Y)| + \lambda} * \frac{2 * |N^G(X) \cap N^G(Y)|}{|N^G(Y)| + |N^G(X) \cap N^G(Y)| + \lambda} \\ &= \frac{2 * |N^G(X) \cap N^G(Y)|}{|N^G(X) - N^G(Y)| + 2 * |N^G(X) \cap N^G(Y)| + \lambda} * \\ &\quad \frac{2 * |N^G(X) \cap N^G(Y)|}{|N^G(Y) - N^G(X)| + 2 * |N^G(X) \cap N^G(Y)| + \lambda} \end{aligned}$$

where λ is a weight to correct for the situation where the proteins have too few neighbours.

It is well known that different experimental sources of deriving protein-protein interaction may have different reliability. An approach [25] to estimate the overall reliability of an experimental source is to find the fraction of interaction pairs from each source that shares at least one function. Then for a pair of proteins X and Y, the reliability of their interaction can be optimistically estimated as $r(X,Y) = 1 - \prod_{i \in E(X,Y)}(1 - r(i))^{n(i,X,Y)}$, where $r(i)$ is the overall reliability of experimental source i, $E(X,Y)$ is the set of experimental sources in which an interaction between X and Y is observed, and $n(i,X,Y)$ is the number of times the interaction between X and Y is observed from experimental source i.

Chua et al. [8] further refine the *FS* index to take into account the reliability of each interaction, when predicting the function of a protein from a protein interaction graph, by substituting the estimated reliability $r(U,V)$ for each interaction $\{U,V\}$ into the definition of *FS*. This gives us the following index:

Definition 4.3 (Functional Similarity Weight with Reliability Index, *FSWeight*) *The "functional similarity weight with reliability" index $FSWeight^G(X,Y)$, on a pair of protein X and Y in an interaction graph G, is defined as*

$$FSWeight^G(X,Y) = \frac{A^G(X,Y)}{B^G(X,Y) + A^G(X,Y) + \lambda} * \frac{A^G(X,Y)}{B^G(Y,X) + A^G(X,Y) + \lambda}$$

where $A^G(X,Y) = 2 * \sum_{W \in N^G(X) \cap N^G(Y)} r(X,W) * r(W,Y)$, and $B^G(X,Y) = \sum_{W \in N^G(X)} r(X,W) + \sum_{W \in N^G(X) \cap N^G(Y)} r(X,W) * (1 - r(W,Y))$.

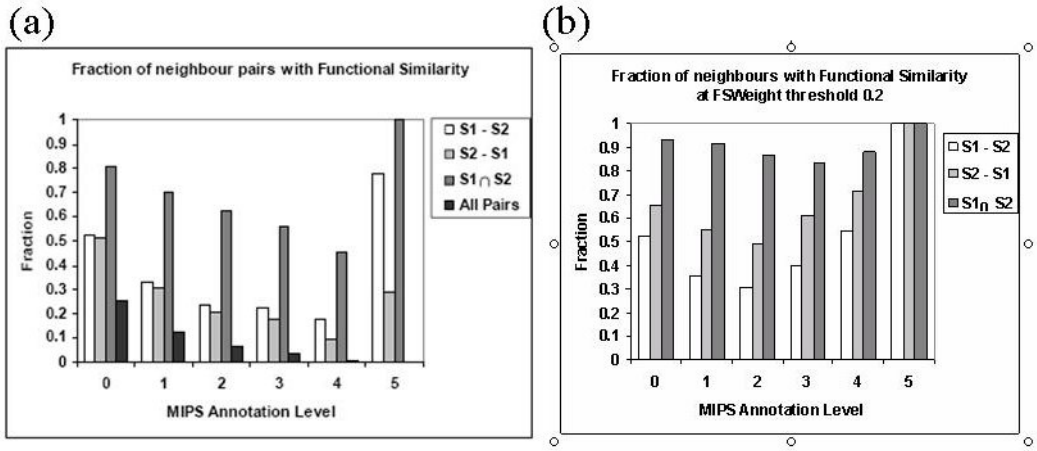

Figure 4: Effectiveness of *FSWeight* as an index for predicting the functional similarity of proteins, taken from Chua, Sung, and Wong [8]. (a) The fraction of direct and indirect neighbours of a protein that have at least one function in common with that protein. (b) The fraction of direct and indirect neighbours of a protein, at *FSWeight* index value threshold of 20%, that have at least one function in common with that protein. Here S1 refers to direct neighbours and S2 refers to indirect neighbours. Thus we can see that there is significant enrichment of shared functions at all MIPS annotation levels for both direct and indirect interaction pairs at *FSWeight* threshold of 20%.

Although *FSWeight* is originally devised for the purpose of protein function prediction, it is also suitable as an index for assessing the reliability of protein-protein interactions. First, common interaction partners are explicitly used in *FSWeight*; this is a special case of the hypothesis expressed in Section 3 that there tend to be some alternative paths connecting a pair of real interacting proteins. Second, two proteins having many interaction partners in common are likely to have similar physical and biochemical characteristics; this makes the two proteins likely to interact. Third, as shown in Chua et al. [8] and in Figure 4, a protein pair having a high *FSWeight* value are likely to share a common function; this is likely to be a consequence of the protein pair interacting.

As noted in the previous section, *IRAP* correlates relatively well with functional homogeneity, localization coherence, and gene expression correlation in the smaller and older interaction datasets (e.g., MIPS 2000, 2003, and 12-08-2003). However, it seems to be less effective with some newer and bigger interaction datasets. With more interactions in the larger datasets, few proteins have neighbours with only one interaction neighbour. Hence the *IG* value for a large fraction of the proteins is 1 (the best *IG* value). This limits the usefulness of *IG* as a reliability indicator since the range of *IG* values becomes limited. As *IRAP* is derived from IG, a similar limitation applies.

FSWeight and *CD-Dist*, on the other hand, seem to be better indicators of interaction reliability for larger interaction datasets. In particular, while *FSWeight*, *CD-Dist*, and *IRAP* exhibit similar performance on older versions of datasets in MIPS, *FSWeight* and *CD-Dist* outperform *IRAP* on newer datasets that contain more interaction data. For example, Figure 5 compares *FSWeight* and *CD-Dist* to *IG* and *IRAP* on the 18-04-2005 version of GRID [3]. In terms of function homogeneity, it can be seen from Figure 5(a) that about 50% of the top 30% of interacting pairs ranked according to *IRAP* have a common function; in contrast, over 60% of the top 30% of interacting pairs ranked by *FSWeight* and *CD-Dist* share a common function; and *FSWeight* further outperforms *CD-Dist* in this aspect when the top 20% of interacting pairs are considered. Similarly, as shown in Figures 5(b) and (c), *FSWeight* and *CD-Dist* both exhibit better localization coherence and higher gene expression correlation than *IRAP*, with *FSWeight* further outperforming *CD-Dist*.

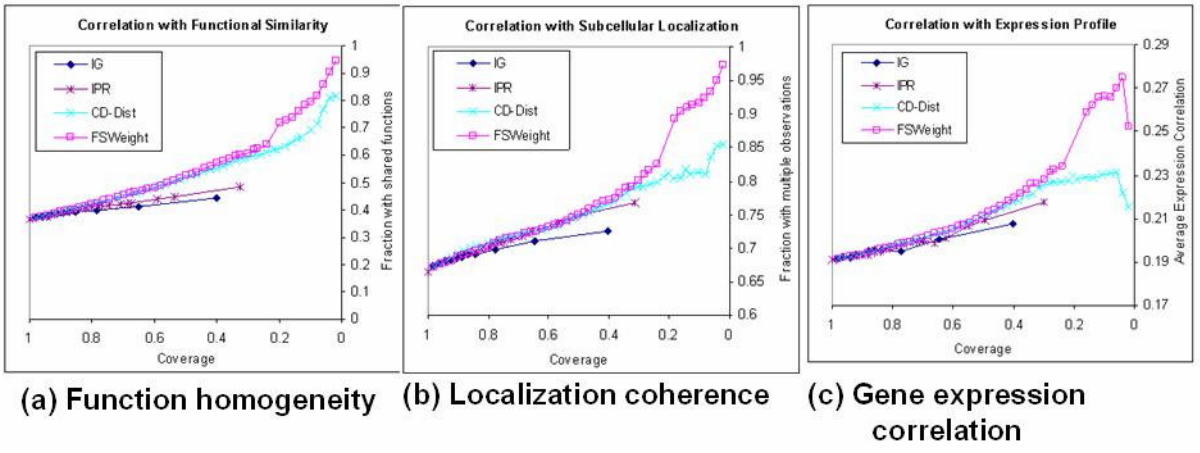

Figure 5: Comparison of *IG*, *IRAP* (denoted as IPR here), CD distance, and *FSWeight* indices on their correlation with (a) function homogeneity, (b) localization coherence, and (c) gene expression correlation. This comparison is performed here using data on 19452 interactions from the General Repository for Interaction Datasets (18-04-2005 version) [3].

5 Meso-Scale Motifs

We have seen in the two preceding sections how *IRAP* [5] and *FSWeight* [8] improve upon *IG* [29] as models of protein interaction networks. However, these improved models are abstract characterizations of networks of reliable interactions. They do not present explicit motifs associated with reliable protein interactions like IG_2 [30]. IG_2 is basically a set of 5 manually derived network motifs of size 3 or 4 for detecting false positives in protein-protein interaction data [30]. Similarly, 4 small motifs have been used by Albert and Albert [1] to predict protein-protein interactions. But many processes in biological networks are meso scale in the sense that they involve 5–25 genes or proteins. In this section, we discuss explicit meso-scale network motifs of Chen et al. [7] that can be systematically exploited for assessing the reliability of protein-protein interaction data.

Interesting network motifs are typically repeated and unique [24]. That is, these motifs appear repeatedly in protein-protein interaction networks but do not appear in randomized networks. Thus, Chen et al. [7] define network motifs as follows:

Definition 5.1 (Network Motifs) *A "network motif" g in a protein-protein interaction graph G is a connected, unlabelled, and undirected topological pattern of inter-connections that is "repeated" and "unique" in G. A topological pattern g is said to be repeated in an interaction graph G if $f(g, G)$ exceeds a specified threshold, where $f(g, G)$ is the number of times g occurs in G. A topological pattern g is said to be unique in an interaction graph G if $s(g, G)$ exceeds a specified threshold, where $s(g, G) = |\{G_{rand_i} \mid 1 \leq i \leq n, f(g, G_{rand_i}) < f(g, G)\}|/n$ and each G_{rand_i} is a randomization of G that preserves the number of nodes and edges.*

We write $g(X, Y)$ for the topological pattern g that is partially labelled with a protein interaction $\{X, Y\}$. For convenience, we also called such a partially instantiated topological pattern a network motif. The definitions of $f(g, G)$ and $s(g, G)$ are extended to these partially instantiated network motifs in the obvious manner.

We illustrate some of the concepts mentioned in the definition above in Figure 6. In part (a) of the figure, an example graph G is shown. In part (b), a number of topological patterns of size 2 to size 5 are displayed. In part (c), the occurrences of the pattern identified as t_{4_2} in G in part (b) of the figure are shown.

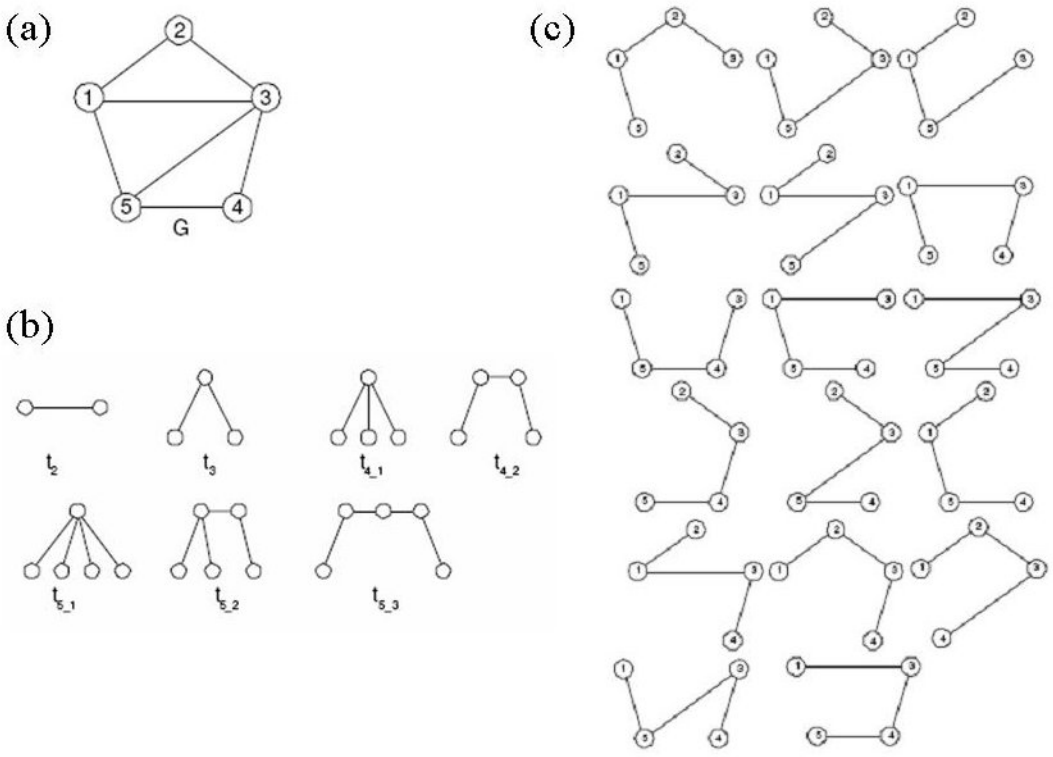

Figure 6: (a) An example graph G. (b) Some example topological patterns of size 2 to size 5 in G. (c) Occurrences of the topological pattern t_{4_2} in G.

There are a number of computational challenges that Chen et al. [7] have to solved to provide an efficient mining of such meso-scale network motifs. We briefly describe the challenges now. It is common for proteins and their interactions in complex biological networks to participate in multiple functional modules. So during the subgraph counting process in accordance to Definition 5.1, we must consider patterns with arbitrary overlaps of vertices and edges. This results in a computationally complex problem as the degree of "repeatedness" and "uniqueness" of network motifs is not downward closed,[1] and thus the usual pruning heuristics for frequent pattern search cannot be applied. In particular, when a motif g extends or reduces to its supergraph or subgraph, the decrease or increase of $f(g, G)$ and $f(g, G_{rand_i})$ is nondeterministic. Thus we cannot directly infer whether the supergraphs or subgraphs of a network motif g are unique. In fact, even when we have found a non-unique motif, we still have to generate its supergraphs and check for their frequencies and uniqueness. So determining the uniqueness of a motif is computationally costly.

Chen et al. [7] propose an efficient algorithm called NeMoFinder to discover repeated and unique meso-scale motifs in a large protein-protein interaction graph. As shown in Figure 7, NeMoFinder is many times more efficient, as well as more complete, than existing network motif discovery algorithms such as naive enumeration [24], sampling [14], and FPF [31]. Furthermore, the network motifs discovered by NeMoFinder have very good coverage. For example, 96% of the protein interactions in the MIPS CYGD dataset are covered by at least one network motif discovered by NeMoFinder.

A strategy similar to IG_2 can be employed to rank the reliability of a protein interaction by meso-scale motifs; viz., we sum the "strength" of all meso-scale network motifs containing the interaction

[1] "Downward closed" is an important concept in the data mining of frequent patterns. In the context of this paper, a motif is "downward closed" means that the degree of "repeatedness" and "uniqueness" of each of its submotifs is as high as that of the motif itself.

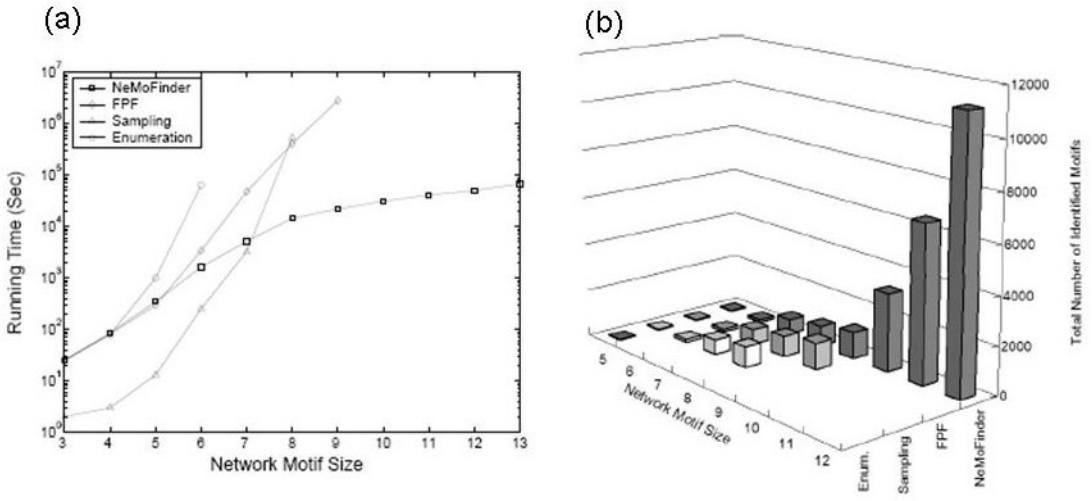

Figure 7: Performance of NeMoFinder. (a) Comparison of computational times to find network motifs of varying sizes in the protein interaction data of Uetz et al [33]. (b) Comparison of size and number of network motifs found in MIPS CYGD (2005) [23].

in question in the derived interaction graph. Specifically, Chen et al. [7] define the following index:

Definition 5.2 (Meso-Scale Motif Interaction Reliability Index, *NeMoFinder***)** *The meso-scale motif interaction reliability index* $NeMoFinder_k^G(X,Y)$*, on a pair of potentially interacting proteins* X *and* Y *in an interaction graph* G *based on network motifs of size up to* k*, is defined as the sum of the "strength" of all network motifs of size up to* k *in* G *that contain the interaction* $\{X,Y\}$*:*

$$NeMoFinder_k^G(X,Y) = \sum_{i=2}^{k} \sum_{g \in M_i(X,Y,G)} \frac{i \times s(g,G) \times f(g,G)}{\max_i(G)}$$

where $M_i(X,Y,G)$ *is the set of network motifs of size* i *in* G *that are partially instantiated with—i.e., contain—the interaction* $\{X,Y\}$*, and* $\max_i(G)$ *is the maximum value of* $s(h,G) \times f(h,G)$ *of all network motifs of size* i *in* G*.*

Figure 8 compares the effectiveness of *NeMoFinder* to *IG*, IG_2, and *IRAP* as indices for assessing the reliability of protein-protein interactions. In terms of functional homogeneity, it can be seen from Figure 8(a) that *NeMoFinder* is generally better than *IRAP*, and is significantly better than *IG* and IG_2. For example, using network motifs of size up to 8 (12), over 82% (87%) of the top 20% of interacting pairs ranked according to *NeMoFinder* are annotated to have a common cellular role; in contrast only about 77% (74%, 65%) of the top 20% of interacting pairs ranked according to *IRAP* (IG_2, *IG*) are annotated to have a common cellular role. Similarly, *NeMoFinder* exhibits better localization coherence and higher gene expression correlation than *IRAP*, IG_2, and *IG*, as shown in Figures 8(b) and (c). These results demonstrate the positive effect of using a more comprehensive set of actual network motifs against a small number of simple predefined motifs. Additionally, Chen et al. [7] have also compared the performance of using motifs of different sizes. They have shown that using network motifs of size up to 12 consistently show superior performance over using motifs of size up to 8. Thus it is advantageous to include the larger motifs, justifying the need for discovering meso-scale network motifs.

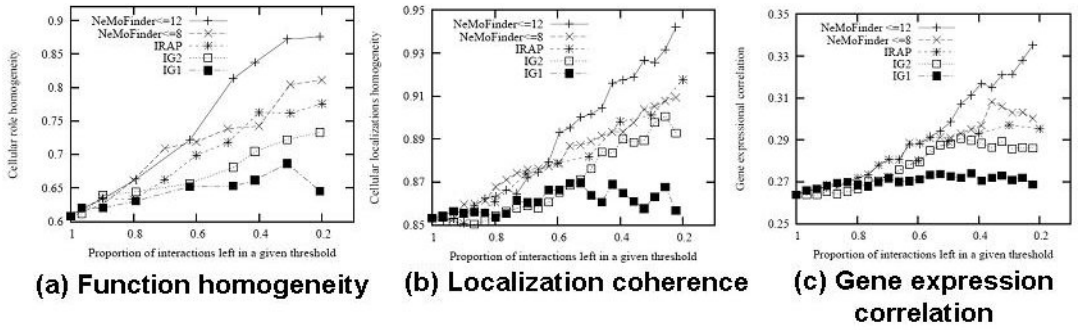

Figure 8: Comparison of meso-scale motifs of size up to 8 ($NeMoFinder = 8$), meso-scale motifs of size up to 12 ($NeMoFinder = 12$), $IRAP$, IG, and IG_2 indices on their correlation with (a) function homogeneity, (b) localization coherence, and (c) gene expression correlation. This comparison is performed here using data on 10199 nonredundant yeast protein interactions from MIPS CYGD (14-11-2005) [23]. The threshold for $f(\cdot,\cdot)$ is 50 and the threshold for $s(\cdot,\cdot)$ is 0.95, following [24].

6 Closing Remarks

In this paper, we have presented an overview of approaches for assessing the reliability of high-throughput protein-protein interaction data that rely solely on the topological properties of the derived protein interaction graph. In particular, we have focused on three generations of indices—IG and IG_2 in the first generation, $IRAP$ in the second generation, and $FSWeight$ and $NeMoFinder$ in the third generation—which show increasing promise in detecting false positives.

Due to space and time constraints, we have not discussed other approaches [2, 10, 11, 19, 15, etc.] to detect false positives. We have also not discussed approaches to detect false negatives. However, we note that indices like $IRAP$, $FSWeight$, and their variants can be used for detecting false negatives. In particular, one can postulate a pair $\{X, Y\}$ to be interacting if $IRAP^G(X, Y)$ or $FSWeight^G(X, Y)$ are high, even if $\{X, Y\}$ are not observed to be interacting in the graph G [6, 27].

We note that the detection of false positives here has been intentionally presented using only the topological information that are mathematically derived from the underlying interaction graphs. This allows us to clearly demonstrate the potential usefulness of the indices discussed. However, in practice, other biological information—such as gene expression correlation, functional homogeneity, localization coherence, presence of binding motifs, etc.—should be incorporated to improve the quality of reliability assessment. As a future work, it will be interesting to investigate how other topological measures, as well as additional biological information, can be incorporated for increasing the reliability of protein-protein interactome.

We now close this paper by relating some "history". The last author (Wong) first learned, at GIW 2002, of the possibility of ranking the reliability of protein-protein interactions reported in high-throughput yeast-two-hybrid assays from Saito (the 6th author), who was showing a poster of his works on IG [29] and IG_2 [30]. Wong was so impressed with the poster that, upon returning to Singapore, he told his colleagues Ng (the 5th author) and Soon-Heng Tan about it. Ng subsequently followed up on the idea with his collaborators Chen (the 1st author), Hsu (the 3rd author), and Lee (the 4th author); and developed improvements such as $IRAP$ [5], $IRAP^*$ [6], and $NeMoFinder$ [7]. Soon-Heng Tan did not follow up on the idea, though he was inspired to work on identification of protein-protein binding motifs [17]. Wong followed up on the paper of Tan, and co-authored with Haiquan Li and Jinyan Li a paper on binding motifs [18], which contained a figure showing SH3 and SH3-binding proteins in way that places SH3 proteins on one side and SH3-binding proteins one another side of the figure, as shown in Figure 9. Wong showed this figure to Chua (the 2nd author), who remarked

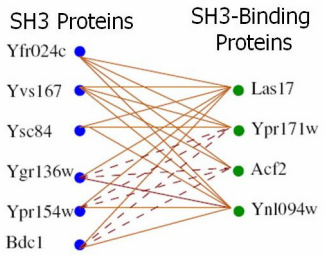

Figure 9: The figure of SH3 and SH3-binding proteins from [18] that inspired the work on *FSWeight* [8].

that the figure implied that a protein might share functions with its indirect neighbours even though it might share no function with its direct neighbours. A paper with Sung (the 7th author) soon followed on using indirect neighbours to infer protein function [8], in which the *FSWeight* measure was defined. Subsequently, we realised the possibility of *FSWeight* as a means to rank the reliability of protein interactions. As we can see, the discussion of Wong and Saito at GIW 2002 has lead to a very fruitful chain of results. We wish to thank the organizers of GIW 2006 for inviting Wong to make a keynote presentation. It is our great pleasure to take this opportunity to summarise these results in this keynote paper at GIW 2006.

Acknowledgements

This work is supported in part by an A*STAR AGS scholarship (Chua), and the I^{2}R-SOC Joint Lab on Knowledge Discovery from Clinical Data (Chen, Hsu, Lee, Sung, Wong).

References

[1] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, 2004.

[2] Joel S. Bader, Amitabha Chaudhuri, and others. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.

[3] Bobby-Joe Breikreutz, Chris Stark, and Mike Tyers. The GRID: The general repository for interaction datasets. *Genome Biology*, 4:R23, 2003.

[4] Christine Brun, Francois Chevent, David Martin, and others. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6, 2003.

[5] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artificial Intelligence in Medicine*, 35(1–2):37–47, 2005.

[6] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22:1998–2004, 2006.

[7] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. NeMoFinder: Dissecting genome wide protein-protein interactions with repeated and unique network motifs. In *Proceedinsg of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 106–115, Philadelphia, PA, 2006.

[8] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623–1630, 2006.

[9] M. Deng, S. Mehta, and others. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12(10):1540–1548, 2002.

[10] Aled M. Edwards, Bart Kus, and others. Bridging structural biology and genomics: Assessing protein interaction data with known complexes. *Trends in Genetics*, 18(10):529–536, 2002.

[11] Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA*, 100(8):4372–4376, 2003.

[12] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 29(17):3513–3519, 2001.

[13] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, 2001.

[14] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[15] Insuk Lee, Shailesh V. Date, and others. A probabilistic functional networks of yeast genes. *Science*, 306:1555–1558, 2004.

[16] P. Legrain, J. Wojcik, and J. M. Gauthier. Protein-protein interaction maps: A lead towards cellular functions. *Trends in Genetics*, 17(6):346–352, 2001.

[17] Haiquan Li, Jinyan Li, Soon-Heng Tan, and See-Kiong Ng. Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. In *Pacific Symposium on Biocomputing*, pages 312–332, 2004.

[18] Haiquan Li, Jinyan Li, and Limsoon Wong. Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.

[19] Nan Lin, Baolin Wu, and others. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5:154, 2004.

[20] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[21] C. von Mering, R. Krause, and others. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[22] H. W. Mewes, D. Frishman, C. Gruber abd B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil. MIPS: A database for genomes and protein sequences. *Nucleic Acids Research*, 28:37–40, 2000.

[23] H. W. Mewes, D. Frishman, U. Guldener, et al. MIPS: A database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–34, 2002.

[24] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[25] Elena Nabieva, Kam Jim, and others. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl. 1):i302–i310, 2005.

[26] S.-K. Ng and S.-H. Tan. Discovering protein-protein interactions. *Journal of Bioinformatics and Computational Biology*, 1(4):711–741, 2004.

[27] Pengjun Pei and Aidong Zhang. A topological measurement for weighted protein interaction network. In *Proceedings of 4th International Computational Systems Bioinformatics Conference*, pages 268–278, Stanford, CA, August 2005.

[28] Arun K. Ramani, Razvan C. Bunescu, and others. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6:R40, 2005.

[29] Rintaro Saito, Harukazu Suzuki, and Yoshihide Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, 30:1163–1168, 2002.

[30] Rintaro Saito, Harukazu Suzuki, and Yoshihide Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, 2003.

[31] F. Schreiber and H. Schwobbermeyer. Frequency concepts and pattern detection for the analysis of motifs in networks. *Transactions on Computational Systems Biology*, 3(LNBI 3737):89–104, 2005.

[32] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.

[33] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[34] Haiyuan Yu, Alberto Paccanaro, Valery Trifonov, and Mark Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, 2006.