Constructing More Reliable Protein-Protein Interaction Maps

Limsoon Wong
National University of Singapore
email: wongls@comp.nus.edu.sg

1 Introduction

Progress in high-throughput experimental techniques in the past decade has resulted in a rapid accumulation of protein-protein interaction data [27, 17, 28]. High-quality protein-protein interaction maps are useful for a deeper understanding of how proteins may together to carry out specific functions. However, high-throughput methods are known to yield a non-negligible rate of false positives, and to miss a fraction of existing interactions [28, 26, 10]. As a result, further carefully-focused small-scale experiments are often needed to complement the large-scale methods to validate the detected interactions. Therefore computational analysis techniques for assessing and ranking the reliability of a protein-protein interaction are highly desirable.

I describe here our work in assessing and improving the reliability of protein-protein interactions from these high-throughput experiments. I also show the impact of more reliable protein interaction data on recognition of protein complexes.

2 Functional Homogeneity, Localization Coherence

One of the earliest ideas for assessing the reliability of protein interaction experiments is to consider supporting evidence from the biological perspective. For example, a pair of interacting proteins are generally expected to be localized to the same cellular component or to have a common cellular role [26, 6]. Therefore one rough estimate of the reliability of a protein interaction assay is the proportion of interacting protein pairs reported by that assay that have a common cellular role or are localised to the same cellular component [20, 10]. Figure 1 contains such rough estimates for various protein interaction assays, computed based on common cellular role [10].

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

Figure 1: Estimated reliability for each protein interaction assay in the GRID dataset [4], computed based on functional homogeneity [10].

Protein functional annotations and subcellular localization annotations are generally incomplete; and not all protein pairs localized to the same cellular compartment or participating cellular process interact in reality. So a more elaborate scheme [26] can be conceived as follows. The proportion D of interacting pairs reported by the assay in question is contributed by (a) the proportion I of true interacting pairs that are co-localized or have common cellular role that are correctly detected by the assay, and by (b) the proportion R of random pairs that are co-localized or have common cellular role that are falsely detected by the assay. More formally, $D = TP * I + (1 - TP) * R$. Thus the true-positive rate TP of the assay can be derived as $TP = (D - R)/(I - R)$. The proportion I can be estimated from a large-enough gold standard protein interaction data set.

3 Information Fusion

Besides estimating reliability of a protein interaction assay as a whole, it is also desirable to assess reliability of an individual reported protein interacting pair. An early idea for this is that of repeatability. An interaction observed in two or more separate experiments is obviously more reliable than one observed only in one experiment.

Suppose the reliability r_i of each protein interaction assay i is known or has been estimated as in previous section. Assume 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 taken as the probability that at least one of the assays involved is reliable [10, 20]. 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 common technique for assessing the reliability of an interacting pair from multiple assays is to compute a p-value based on the hypergeometric distribution [16]. Suppose a total of h interactions are reported. 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 $P(k|n, m, h) = \binom{h}{k}\binom{h-k}{n-k}\binom{h-n}{m-k}/\binom{h}{n}\binom{h}{m}$. Then the p-value for x and y being reported to interact in k_0 experiments is $\sum_{k\geq k_0} P(k|n, m, h)$.

4 Topology of Interactions

A large number of more sophisticated approaches exist for estimating error rates of protein interaction assays [3, 13, 21] and for ranking individual protein interacting pairs [15, 24, 25, 19]. These approaches generally require the use of additional information such as annotations on proteins or the use of information from multiple assays. By contrast, Saito et al [23, 22], Chen et al [6, 7, 8], and Albert and Albert [1] are much more interesting in the sense 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".

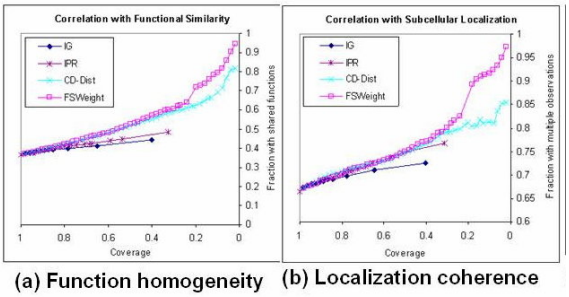
For example, the "interaction generality index" (*IG*) of Saito et al [22] uses the property of two-hybrid assay that a large number of false positives in two-hybrid assay are due to self activators and "sticky" proteins that transactivate the reporter gene without actually interacting with their partners. A characteristic of these self activators and sticky proteins is that they appear to have a large number of interaction partners in experiment, but these partners typically do not interact with each others. Thus the *IG* on a pair of reported interacting proteins (u, v) is simply a count of the number of isolated interaction partners that they have. The larger this count is, the more unlikely that (u, v) is interacting.

The "interaction pathway reliability index" (*IPR*) of Chen et al [7] relies on the assumptions that a biological function is generally performed by a highly interconnected network of interactions and that evolution favours adding interactions that shorten the pathways of the function. Therefore, a pair of proteins that are connected by a short path of reliable interactions are likely to directly interact. Thus the *IPR* on a pair of candidate interacting proteins (u, v) is defined 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 [7] uses *IG* as the rough estimate of the reliability of an individual interaction.

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 are generally expected to have a common cellular role. It has been observed that proteins that have common interaction partners have a high chance of sharing a common function [10]. 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 in-

dex is the Czekanowski-Dice distance, $CD\text{-}Dist_{u,v} = 2|N_{u,v}|/(|N_u| + |N_v|)$, 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 *FSWeight* measure, $FSWeight_{u,v} = (2|N_{u,v}|/(|N_u - N_v| + 2|N_{u,v}| + \lambda_{u,v}))(2|N_{u,v}|/(|N_v - N_u| + 2|N_{u,v}| + \lambda_{v,u}))$, where $\lambda_{u,v}$ is a pseudo count to penalize similarity weights between protein pairs when any of the proteins has too few interacting partners. Both were originally used for the purpose of protein function prediction from protein interaction graphs [5, 10].

The effectiveness of these indices can be gauged by their correlation with functional homogeneity and localization coherence. For example, as shown in Figure 2, 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 between these indices and the gene expression correlation of highly ranked candidate interacting proteins, as well as between these indices and number of times highly ranked candidate pairs are observed in multiple protein interaction assays. See my GIW2006 keynote paper for further details [6].



(a) Function homogeneity (b) Localization coherence

Figure 2: Comparison of *IG* [22], *IPR* [7], *CD-Dist* [5], and *FSWeight* [10] indices on their correlation with (a) function homogeneity and (b) localization coherence. This comparison was performed in [6] using data on 19452 interactions in yeast from the GRID database [4]. 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.

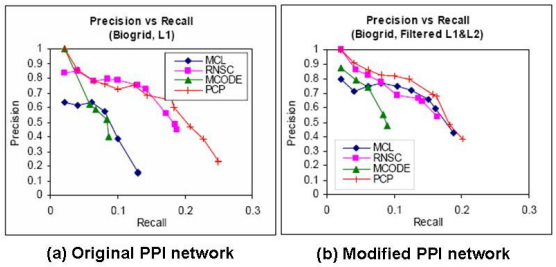
5 Protein Complex Prediction

Protein complexes are useful for understanding principles of cellular organizations [12]. *MCODE* [2], *RNSC* [18], and *MCL* [14] are three better-known approaches to protein complex prediction. *PCP* [9] is probably the latest approach to this problem.

Figure 3(a) shows the performance of *MCODE*, *MCL*, *RNSC*, and *PCP* on the BioGRID dataset [11]. *RNSC* and *PCP* correctly predict, by analysing the yeast protein interaction network in BioGRID, 10% (20%) of the known yeast protein complexes at 80% (50%) precision. *MCODE* and *MCL* do not perform as well on this dataset. There is thus much to be improved in the capability of these methods.

Protein interaction datasets contain a lot of noise [26]. If such noise can be reduced in the input protein interaction network, *MCODE*, *MCL*, *RNSC*, and *PCP* should improve in performance. We have shown earlier that *FSWeight* is a good index of the reliability of an interaction. Thus we can preprocess the protein interaction network by computing the *FSWeight* of each interaction and retaining only the high-scoring interactions. Also, while proteins within a complex interact to perform some common functions, they need not have full mutual interactions. As shown in [10], a high *FSWeight* between two indirectly interacting proteins is an excellent indication of function sharing, and thus the two proteins are likely to be indirect interaction partners within a complex. As *MCODE*, *MCL*, *RNSC*, and—to a lesser extent—*PCP* rely on direct interaction partners, their performance can be boosted if we can modify the input network by augmenting it with direct edges between indirect interaction proteins that have high *FSWeight*.

Figure 3(b) shows the impact of these two types of preprocessing of the input protein interaction network on *MCODE*, *MCL*, *RNSC*, and *PCP*. For example, the precision of *PCP* on the BioGRID dataset is increased by almost 10% in the 10–20% sensitivity range. *MCL* benefits even more significantly, with sensitivity increasing from 10% to 20% and precision



(a) Original PPI network (b) Modified PPI network

Figure 3: Comparison of *MCODE* [2], *RNSC* [18], *MCL* [16], and *PCP* [9]. The comparison was performed in [9] on (a) the original BioGRID [11] yeast interaction data and (b) the modified BioGRID yeast interaction data.

improving by over 20% in the entire sensitivity range. For more details, please see my CSB2007 paper [9].

6 Summary and Thanks

I have provided an overview of methods for estimating the reliability of protein-protein interaction experiments. I have shown that it is possible to rank the reliability of an individual reported interaction by the topology of its local interaction network. I have also demonstrated the beneficial effect of cleansing a protein interaction network on the problem of protein complex prediction.

I am grateful for a Singapore MOE AcRF Tier 1 grant that has supported this work in part. I also thank the organizer of *Bioinformatica Indica 2008* for inviting me to present this work. Lastly, I thank my collaborators who have made many contributions to the results described here: Jin Chen, Kenny Chua, Wynne Hsu, Mong Li Lee, Hon Wai Leong, See-Kiong Ng, Rintaro Saito, and Wing-Kin Sung.

References

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

[2] G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[3] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.

[4] B.-J. Breikreutz, C. Stark, and M. Tyers. The GRID: The general repository for interaction datasets. *Genome Biology*, 4:R23, 2003.

[5] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6, 2003.

[6] J. Chen, H. N. Chua, W. Hsu, M.-L. Lee, S.-K. Ng, R. Saito, W.-K. Sung, and L. Wong. Increasing confidence of protein-protein interactomes. In *Proceedings of 17th International Conference on Genome Informatics*, pages 284–297, Yokohama, Japan, December 2006.

[7] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng. Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22:1998–2004, 2006.

[8] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng. 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*, pages 106–115, Philadelphia, PA, August 2006.

[9] H. N. Chua, K. Ning, W.-K. Sung, H. W. Leong, and L. Wong. Using indirect protein-protein interactions for protein complex prediction. In *Proceedings of 6th Annual International Conference on Computational Systems Bioinformatics*, pages 97–110, San Diego, California, August 2007.

[10] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623–1630, 2006.

[11] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue):D535–539, 2006.

[12] U. de Lichtenberg, L. J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, 2005.

[13] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: Two methods for assessment of the reliability of high-throughput observations. *Molecular and Cellular Proteomics*, 1:349–356, 2002.

[14] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[15] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, et al. A protein interaction map of *drosophila melanogaster*. *Science*, 302:1727–1736, 2003.

[16] G. T. Hart, I. Lee, and E. M. Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):236, 2007.

[17] 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.

[18] A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[19] S. Martin, D. Roe, and J.-L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21(2):218–226, 2005.

[20] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl. 1):i302–i310, 2005.

[21] A. Patil and H. Nakamura. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6:100, 2005.

[22] R. Saito, H. Suzuki, and Y. Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, 30:1163–1168, 2002.

[23] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, 2003.

[24] M. P. Samanta and S. Liang. Predicting protein functions from redundancies in large-scale protein interaaction networks. *Proc. Acad. Sci. USA*, 100(22):12579–12583, October 2003.

[25] T. Schlitt, K. Palin, J. Rung, S. Dietmann, M. Lappe, E. Ukkonen, and A. Brazma. From gene networks to gene function. *Genome Research*, 13:2568–2576, 2003.

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

[27] 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.

[28] C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.