Wong L, Liu G. Protein interactome analysis for countering pathogen drug resistance. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 1– Jan. 2010

Protein Interactome Analysis for Countering Pathogen Drug Resistance

Limsoon Wong and Guimei Liu

School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417

E-mail: wongls@comp.nus.edu.sg; liugm@comp.nus.edu.sg

Received October 22, 2009; revised November 16, 2009.

Abstract Drug-resistant varieties of pathogens are now a recognized global threat. Insights into the routes for drug resistance in these pathogens are critical for developing more effective antibacterial drugs. A systems-level analysis of the genes, proteins, and interactions involved is an important step to gaining such insights. This paper discusses some of the computational challenges that must be surmounted to enable such an analysis; viz., unreliability of bacterial interactome maps, paucity of bacterial interactome maps, and identification of pathways to bacterial drug resistance.

Keywords protein complexes, protein interactomes, drug resistance, infectious diseases

1 Introduction

It is critical to address the emergence of drug-resistant varieties of pathogens for infectious diseases, especially those that have spread globally, such as drug-resistant tuberculosis[1]. Approaches to counter drug resistance have so far achieved limited success[2]. This lack of success may be due to a lack of understanding of how resistance emerges in bacteria upon drug treatment. It has been proposed that a systems-level analysis of the genes, proteins, and interactions involved is the key to gaining insights into routes required for drug resistance[3].

One of pre-requisites of such an analysis is the existence of a comprehensive protein interactome of the relevant pathogen. For example, let us assume that a comprehensive protein interactome of *Mycobacterium tuberculosis* is available. Then one can identify a minimal set of proteins (or protein interactions) whose inhibition would disconnect all essential pathways in *M. tuberculosis*. Alternatively, one can trace the interaction route of the known targets of a drug to various efflux pump proteins and drug-modifying enzyme proteins[4].

Two yeast interactome maps[5-6] based on yeast two-hybrid technology were published eight years ago. Today, the quantity and variety of protein interaction data have increased rapidly. In particular, two-hybrid-based interactome maps have been generated for model organisms such as *C. elegans*[7], *D. melanogaster*[8], and human[9-10]. Proteome-scale interactome maps have also been generated for yeast by TAP-MS experiments[11-13]. However, very few bacterial species[14-16] have been analyzed at the proteome level for protein interactions. Furthermore, the quality of these interactome maps has much to be improved[17-18]. For example, out of nearly 10 000 interactions surveyed in [19], less than 25% were detected in two different assays. This implies high false positive and false negative rates in the underlying biological assays.

Hence a systems-level analysis of proteins and their interactions in pathogens of infectious diseases for identifying drug resistance pathways is difficult. This is a worthy problem for computational biologists. So, we propose the following challenges in the systems-level analysis of proteins and interactions in pathogens of infectious diseases for identifying drug resistance pathways:

- Unreliability of bacterial interactome maps;
- Paucity of bacterial interactome maps;
- Identification of candidate pathways to drug resistance in a bacterium.

Furthermore, we suggest using *M. tuberculosis* as a test case. These challenges and some approaches are discussed in the remainder of this paper.

2 Unreliability of Bacterial Interactome Maps

Comprehensive and high-quality bacterial protein interactome maps are needed to support reliable inference of pathways to drug resistance in bacteria.

Survey

This work was supported in part by Singapore National Research Foundation under Grant No. NRF-G-CRP-2997-04-082(d).

Unfortunately, the quality of available bacterial protein interactome maps is likely to be low, as several studies show that existing interactome maps have a lot of noise[17]. For example, while real protein interactions are expected to be generally reproducible, it has been found that no more than 25% of protein interactions reported in various high-throughput experiments can be detected in two or more different high-throughput assays[18-19]. As another example, while real protein interactions are expected to be generally between proteins in the same cellular compartments, no more than 55% of protein interactions in the DIP yeast protein interaction database[20] are between proteins in the same cellular compartments. Thus it is critical to develop techniques to assess the reliability of protein interactions in bacterial protein interactome maps. Also, these interactome maps are still essentially an *in vitro* scaffold and thus do not directly reflect, e.g., *in vivo* protein complexes[21]. It is critical to investigate techniques for improving the reliability of interactome maps, as well as techniques for inferring protein complexes from interactome maps.

Protein reliability assessment has been made based on the sharing of a common cellular localization or a common functional role[19,22]. It has also been made based on the reproducibility and non-randomness of the observation of an interaction[23]. Related to the ideas of functional homogeneity, localization coherence and observational reproducibility are a large number of other approaches[24] for estimating the reliability of protein interactions based on the use of additional information, such as protein annotation, or the use of information from multiple assays. Current state-of-the-art approaches generally integrate multiple information types into the reliability assessment process[25].

However, the additional information required by these approaches may be unavailable, as is likely to be the case in *M. tuberculosis*. Therefore, reliability indices that do not require such information are critical. Recent work on reliability indices based purely on the topology of the interactome map may be suitable for this purpose[21,26]. These indices are based on the simple "guilt by association" idea: two proteins sharing a large number of common partners are more likely to be co-located and to participate in the same cellular processes; and thus they are more likely to interact[21]. The most direct formulation of this idea is the adjusted *CD*-distance index[26],

$$AdjustCD(u,v) = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

where N_u and N_v are respectively the sets of neighbours of proteins u and v, and λ_u and λ_v are used to penalized proteins with very few neighbours. An edge (u,v) is more likely to be a real interacting pair if $AdjustCD(u,v)$ is a high value. This idea can be applied along with an expectation maximization process to achieve even better cleansing performance[26]. Specifically, define

$$w^{k+1}(u,v) = \frac{\sum_{x \in N_u \cap N_v} (w^k(x,u) + w^k(x,v))}{\sum_{x \in N_u} w^k(x,u) + \lambda_u^k + \sum_{x \in N_v} w^k(x,v) + \lambda_v^k}$$

where $w^k(x,u)$ is the score of (x,u) in the k-th iteration; and $w^0(x,u) = 1$ when there is an edge (x,u) in the original (unclean) protein interaction network and $w^0(x,u) = 0$ otherwise. It is easy to see that $w^1(u,v) = AdjustCD(u,v)$.

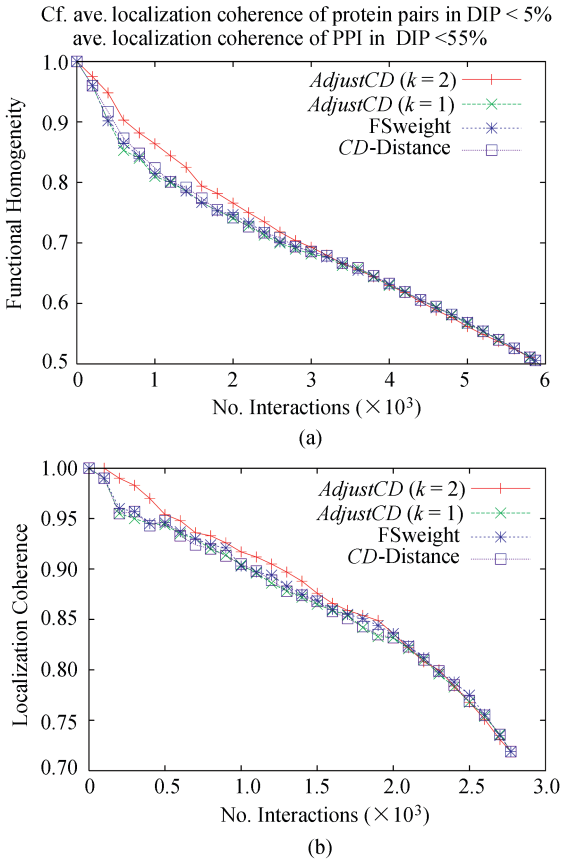

Fig.1. Edges in the DIP yeast interaction database are ranked by *AdjustCD* and other reliability indices. The figure shows the correlation of these indices to functional homogeneity (a) and localization coherence (b).

In spite of its simplicity, this idea works very well. Fig.1 shows the correlation of *AdjustCD* and a few other related indices to functional homogeneity and localization coherence in the DIP yeast interaction database[20].

Note that DIP is quite noisy and the probability of two proteins in an edge in DIP being in the same cellular location is less than 55%. It can be seen from Fig.1 that proteins in the 3000 top-ranked edges by these indices have more than 70% probability of being in the same cellular locations and thus are much more likely to be capable of real interactions *in vivo* than other edges in DIP.

Nonetheless, there is still much room to further improve reliability assessment indices for cleaning protein interactions. For example, *AdjustCD* has inherent limitation when applied on a sparse partially protein interactome map, as its formulation implicitly requires the proteins being considered to have sufficient number of partners.

3 Paucity of Bacterial Interactome Maps

We have earlier suggested *M. tuberculosis* as the target bacterial species because the emergence of drug-resistant tuberculosis is posing a major threat to global tuberculosis eradication programs[3]. However, very few bacterial species have been analyzed at the proteome level for protein interactions. In particular, only a small partially inferred *M. tuberculosis* interactome map is currently available[3]. To deal with this challenge, we need to develop techniques for (i) transferring conserved interactions from other bacterial interactome maps, (ii) predicting *de novo* protein interactions and protein complexes, and (iii) mining protein interactions from biological literature.

3.1 Transfer Conserved Interactions

It is necessary to integrate as many bacterial interactome maps as possible to form a more comprehensive interactome map for *M. tuberculosis*. Possible interactome maps that can serve as starting points include *C. jejuni*[15] and *E. coli*[16].

A major issue in inferring conserved protein interactions is the determination of orthologs whose function and interaction are highly likely to be conserved from one bacterium to another. Candidate orthologs can be obtained through a pre-computed data source such as OMA[27]. However, the mapping is not guaranteed to be unique as OMA is primarily based on evolutionary distance derived from pairwise sequence comparison[28]. It is probably necessary to use conserved gene clusters to disambiguate orthologs that are not uniquely mapped[29]. The inference of conserved gene cluster is non-trivial. Perhaps the recently developed approach based on the idea of gene team tree[30] is helpful here. The article by Tao Jiang[31] in this special issue provides further insights and discussions on this topic.

3.2 Infer Protein Interactions

It is important to infer some protein interactions *de novo* to supplement the paucity of known protein interactions information. There already exists a variety of protein-protein interaction prediction techniques, including domain-domain interactions[32], interaction motifs[33], paralogous interactions[34], protein function similarity[35], protein coevolution information[36], as well as data mining technique to identify domain or functional combination pairs associated with interacting proteins to derive complex interaction rules[25].

Interestingly, the *AdjustCD* index described in the previous section can be made into a technique for *de novo* protein interaction prediction based on topology of protein interactome maps! In particular, one can predict proteins (u, v) to interact if the score $AdjustCD(u, v)$ is high. Fig.2 shows that the top 1000 new interactions in yeast (which are edges missing in DIP) predicted by *AdjustCD* have more than 60% probability of being functionally homogeneous and more

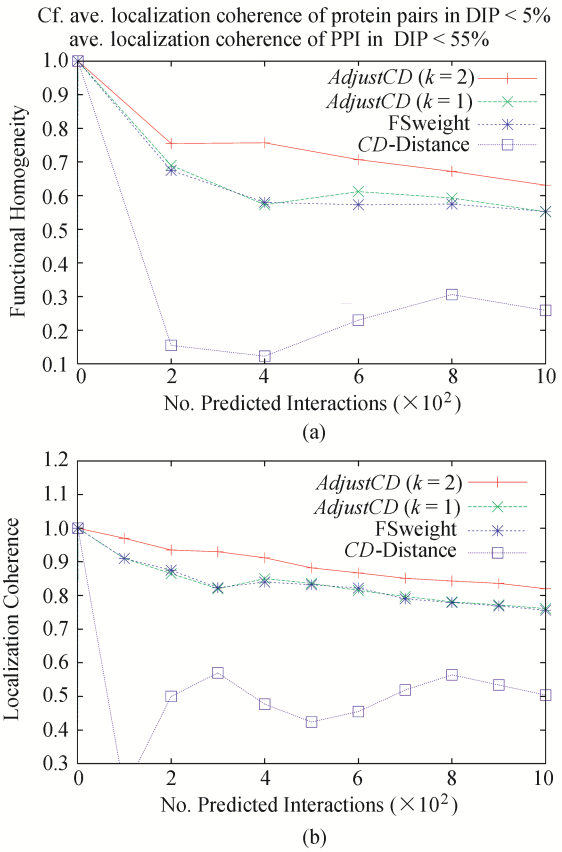

Fig.2. New yeast protein interactions predicted using *AdjustCD* and other reliability indices based on the DIP yeast protein interaction database. The figure shows these new edges have high functional homogeneity (a) and localization coherence (b).

than 80% probability of being in the same cellular compartment, suggesting their strong likelihood to be real interactions[37].

Nevertheless, there is considerable room for improvement as current best-in-class approaches, which already integrate many of the techniques mentioned, are only achieving an ROC score of just above 63%[25].

3.3 Infer Protein Complexes

It is desirable to investigate techniques for predicting protein complexes and functional modules from an interactome map to derive higher-level information for the interactome map. Several algorithms based on graph clustering, dense region finding, or clique finding have been developed to discover protein complexes from protein interaction networks, including MCL[38], RNSC[39], DPClus[40], CFinder[41], PCP[42], and CMC[26]. These algorithms are based on the assumption that proteins in a complex have much denser mutual interactions between themselves than with proteins outside of the complex. However, these algorithms have low sensitivity and low precision[42-43].

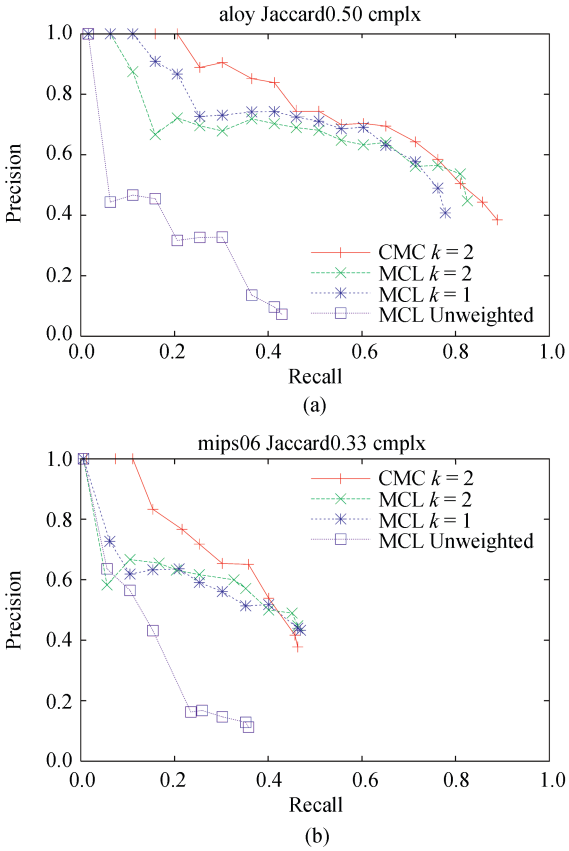

Fig.3. Precision and recall of CMC and MCL on Aloy (a) and MIPS (b) datasets are improved significantly after the input protein interactome map is cleansed using *AdjustCD*.

One reason for this low sensitivity and precision is the noise in protein interactome maps. This is confirmed by an analysis[26], which shows that cleansing the input interactome map has a large positive impact on algorithms for inferring protein complexes from the interactome map. For example, as shown in Fig.3, the popular MCL algorithm[38] nearly doubles its precision and recall on the Aloy[44] and MIPS[45] datasets of yeast protein complexes, after the input protein interactome map is cleansed using *AdjustCD*. Thus the earlier suggested challenge on increasing the reliability of protein interactome should directly benefit protein complex prediction.

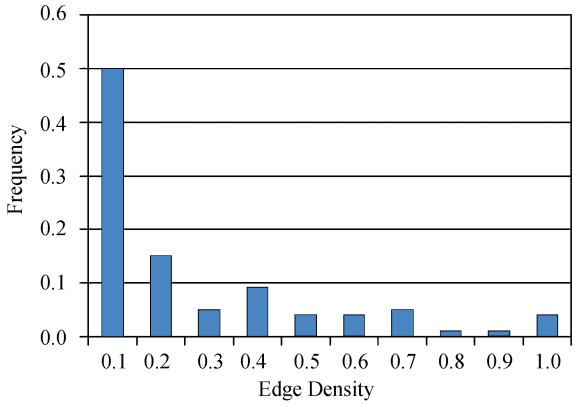

Fig.4. Edge density of yeast protein complexes computed based on protein interactomes from the popular BioGRID database.

Another cause for the low sensitivity and low precision of protein complex prediction algorithms is that they cannot predict complexes that are small or have low edge density, due to their assumption on high edge density. However, complexes having low density of protein interactions are by far more common than high-density ones! As evident in Fig.4, most yeast complexes in MIPS[45] have edge density no more than 30% with respect to protein-protein interactions in BioGRID[46]. So protein complex prediction needs new ideas that do not rely on assuming dense interactions in the protein interaction network.

3.4 Mine Biological Literature

There is a limit to the number of new protein interactions and protein complexes that can be predicted reliably. At the same time, new protein interactions and complexes are also being reported in the scientific literature. It is important that such new information be captured into an integrated database of bacterial interactome maps as soon as possible. Such a database should also be enhanced with information such as the known *M. tuberculosis* drug targets and the known proteins

involved in drug resistance.

This calls for effective tools for mining literature abstracts in Pubmed, reports in US FDA website, and databases like PharmGKB[47]. Such tools should include ability for automated extraction of drug-enzyme and protein-protein interactions from literature.

Recent progress on protein interaction extraction from literature[48] is promising, achieving sensitivity and precision above 80% on a small set of sentences. However, the performance of these literature mining tools on the whole-Pubmed scale has not been satisfactorily evaluated to date. The extraction of drug-enzyme interaction information from literature is even less mature. The article by Hong-Jie Dai *et al.* in this special issue[49] is a good resource for further insights and discussions on this topic.

4 Identify Candidate Pathways to Drug Resistance in Bacteria

Removal or suppression of critical proteins on an essential pathway or complex is expected to disrupt the pathway or complex and prohibit the pathogen from performing a vital function[50]. Thus disconnecting multiple pathways should effectively disrupt the survival of the bacteria. This brings us to the concept of co-targets[3], whereby a combination of drugs is used to simultaneously suppress several critical proteins, to reduce the emergence of drug resistance. This approach to suppressing drug resistance in bacteria requires techniques for (i) identifying essential pathways to drug resistance in a bacterium, and (ii) finding effective co-targets on these pathways to disrupt.

4.1 Identify Pathways

The general known mechanisms for drug resistance in bacteria include efflux pumps to transport drugs out of the cell, cytochromes-like enzymes that chemically modify the drug, and horizontal gene transfer that imports a detoxifying protein from the environment[4,51]. For each of these mechanisms, there are at least two pathways to be identified. The first is the pathway by which the specific efflux pump, drug-modifying enzyme, or detoxifying protein is produced and activated. The second is the pathway through which the drug or its toxic downstream metabolites come into contact with the specific efflux pump, drug modifying enzyme, or detoxifying protein.

The identification of the first type of pathways is by now an established problem in computational biology. It involves classical topics such as discovering transcription factors and their binding sites[52] on the genes for the efflux pump, drug-modifying enzyme, or detoxifying protein.

The identification of the second type of pathways is one of the interesting new challenges in countering bacterial drug resistance. Although there is no clear formulation yet, an extensive use of gene regulatory network and protein interaction network information appears inevitable[18]. For example, Raman and Chandra[3] formulated it as a search for shortest paths of interacting proteins, in the relevant bacterial protein interactome map, that connect the direct targets of a given drug to efflux pumps, drug modifying enzymes, or detoxifying proteins. These paths are then further culled by correlating with gene expression data obtained from the bacteria after exposure to the drug.

However, this formulation has an obvious weakness. It assumes that the shortest paths are the only routes to escape or counter the effect of the drug. Nature is known to be full of "back-up". If there are two disjoint paths between a drug target and (say) an efflux pump, disrupting the shorter path may simply channel the flux to the longer one and thus the drug may still be pumped out of the bacterial cell.

Also, the genome of the drug-resistant strain and non-drug-resistant strain of the pathogen should be compared to identify extra genes in the former. The protein products of these extra genes are worth considering as components of drug-resistance pathways.

4.2 Select Co-Targets

Proteins and interactions on such paths described in the previous subsection are candidate co-targets. A combination of complementary drugs that inhibit them should disrupt pathways that confer resistance to the main drug; thus allowing the main drug to kill the bacteria. For practical purposes, it is preferable to limit the set of complementary drugs to be as small or as cheap as possible.

If, for each path, there is at least one known drug that inhibits a co-target on that path, choosing the smallest or cheapest set of complementary drugs is related to the vertex cover problem, which is NP-complete[53]. If, for some path, there is no known drug that inhibits any co-target on that path, the problem becomes related to the multicut problem or minimum bisection problem, which are NP-hard[54-55]. In either case, the challenge promises to be interesting.

5 Closing Remarks

The emergence of drug resistance in bacteria is a serious global threat. Even though the repertoire of bacterial drug resistance mechanisms is still limited[4], there has not been much success in countering these drug resistance mechanisms. It is hoped that new insights to bacterial drug resistance can be gained by a

systems-level analysis of bacterial gene regulation and protein interaction networks[3,18].

The unreliability and paucity of bacterial interactome maps are key obstacles to such a systems-level analysis. In this paper, we have outlined some current solutions and/or suggested possible approaches to these issues from the computational perspective. In particular, we have described the need to develop techniques to assess reliability of protein interactions that can work on a sparse input protein interactome map without requiring much annotations on the proteins. We have pointed out the need to develop techniques to predict protein complexes from protein interactome maps that can identify protein complexes that have low edge density. We have also highlighted the need to formulate new ways to identify pathways through which a bacterium achieves resistance to a drug. Much work remains to be done indeed!

References

[1] Antimicrobial Resistance Interagency Task Force 2007 Annual Report. CDC USA. 2008.

[2] Johnson R *et al.* Drug resistance in *mycobacterium tuberculosis*. *Curr. Issues Mol. Biol.*, 2006, 8(2): 97-111.

[3] Raman K, Chandra N. *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. *BMC Microbiol.*, 2008, 8: 234.

[4] Nguyen L, Thompson C J. Foundations of antibiotic resistance in bacteria physiology: The mycobacterial paradigm. *Trends Microbiol.*, 2006, 14(7): 304-312.

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

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

[7] Li S, Armstrong C M, Bertin N *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science*, 2004, 303(5657): 540-543.

[8] Giot L, Bader J S, Brouwer C, Chaudhuri A *et al.* A protein interaction map of *drosophila melanogaster*. *Science*, 2003, 302(5651): 1727-1736.

[9] Rual J F, Venkatesan K, Hao T *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 2005, 437(7062): 1173-1178.

[10] Stelzl U, Worm U, Lalowski M, Haenig C *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 2005, 122(6): 957-968.

[11] Gavin A C, Aloy P, Grandi P, Krause R *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006, 440(7084): 631-636.

[12] Krogan N J, Cagney G, Yu H *et al.* Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, 2006, 440(7084): 637-643.

[13] Collins S R, Kemmeren P, Zhao X C *et al.* Towards a comprehensive atlas of the physical interactome of *saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, 2007, 6(3): 439-450.

[14] Rain J C, Selig L, De Reuse H *et al.* The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 2001, 409(6817): 211-215.

[15] Parrish J R, Yu J, Liu G, Hines J A *et al.* A proteome-wide protein interaction map for *campylobacter jejuni*. *Genome Biology*, 2007, 8(7): R130.

[16] Su C *et al.* Bacteriome.org—An integrated protein interaction database for *E. coli*. *Nucleic Acid Res.*, 2008, 36(Supplement 1): D632-D636.

[17] Hart G T, Ramani A K, Marcotte E M. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 2006, 7(11): 120.

[18] Bailer S M, Haas J. Connecting viral with cellular interactomes. *Current Opinion in Microbiology*, 2009, 12(4): 453-459.

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

[20] Xenarios I, Salwinski L, Duan X J, Higney P *et al.* DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 2002, 30(1): 303-305.

[21] Chua H N, Wong L. Increasing the reliability of protein interactomes. *Drug Discovery Today*, 2008, 13(15/16): 652-658.

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

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

[24] Ramani A K, Bunescu R C, Mooney R J, Marcotte E M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 2005, 6(5): R40.

[25] Chua H N, Hugo Willy, Liu G, Li X L, Wong L, Ng S-K. A probabilistic graph-theoretic approach to integrate multiple predictions for the protein-protein subnetwork prediction challenge. *Annals of New York Academy of Sciences*, 2009, 1158: 224-233.

[26] Liu G, Wong L, Chua H N. Complex discovery from weighted PPI networks. *Bioinformatics*, 2009, 25(15): 1891-1897.

[27] Schneider A *et al.* OMA Browser—Exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 2007, 23(16): 2180-2182.

[28] Roth A *et al.* Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, 2008, 9: 518.

[29] Pertea M *et al.* OperonDB: A comprehensive database of predicted operons in microbial genomes. *Nucleic Acid Res.*, 2009, 37(Database Issue): D479-D482.

[30] Zhang M, Leong H W. Gene team tree: A compact tree representation of all gene teams. In *Proc. RECOMB Workshop on Comparative Genomics* (*RCG*), Paris, France, October 13-15, 2008, pp.100-112.

[31] Jiang T. Some algorithmic challenges in genome-wide orthology assignment. *Journal of Computer Science and Technology*, 2010, 25(1):

[32] Li X L *et al.* Improving domain-based protein interaction prediction using biologically-significant negative dataset. *International Journal of Data Mining and Bioinformatics*, 2006, 1(2): 138-149.

[33] Li H, Li J, Wong L. Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics*, 2006, 22(8): 989-996.

[34] Mika S, Rost B. Protein-protein interactions more conserved within species than across species. *PLoS Comput Biology*, 2006, 2(7): 379.

[35] Wu X *et al.* Prediction of yeast protein-protein interaction network: Insights from the Gene Ontology and annotations. *Nucleic Acid Res.*, 2006, 34(7): 2137-2150.

[36] Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA*, 2008, 105(3): 934-939.

[37] Liu G, Li J, Wong L. Assessing and predicting protein interactions using both local and global network topological metrics. In *Proc. the 19th Int. Conf. Genome Informatics* (*GIW*), Gold Coast, Australia, December 1-3, 2008, pp.138-149.

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

[39] Przulj N, Wigle D. Functional topology in a network of protein interactions. *Bioinformatics*, 2003, 20(3): 340-348.

[40] Altaf-Ul-Amin M *et al.* Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 2006, 7: 207.

[41] Adamcsek B *et al.* CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, 2006, 22(8): 1021-1023.

[42] Chua H N, Ning K, Sung W-K, Leong H W, Wong L. Using indirect protein-protein interactions for protein complex prediction. *Journal of Bioinformatics and Computational Biology*, 2008, 6(3): 435-466.

[43] Leung H C M *et al.* Predicting protein complexes from PPI data: A core-attachment approach. *J. Comput. Biol.*, 2009, 16(2): 133-164.

[44] Aloy P *et al.* Structure-based assembly of protein complexes in yeast. *Science*, 2004, 303(5666): 2026-2029.

[45] Mewes H W *et al.* MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 2004, 32(Database Issue): D41-D44.

[46] Stark C, Breitkreutz B J, Reguly T, Boucher L *et al.* BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 2006, 34(Database Issue): D535-D539.

[47] Altman R B. PharmGKB: A logical home for knowledge relating genotype to drug response phenotype. *Nature Genet.*, 2007, 39(4): 426.

[48] Chowdhary R, Zhang J, Liu J S. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics*, 2009, 25(12): 1536-1542.

[49] Dai H J, Chang Y C, Tsai R T H *et al.* New challenges for biological text mining in the next decade. *Journal of Computer Science and Technology*, 2010, 25(1):

[50] Strong M, Eisenberg D. The protein network as a tool for finding novel drug targets. *Progress in Drug Research*, 2007, 64: 191-215.

[51] Smith P A, Romesberg F E. Combating bacteria and drug resistance by inhibiting mechanisms of persistence and adaptation. *Nat. Chem. Biol.*, 2007, 3(9): 549-556.

[52] Valouev A, Johnson D S, Sundquist A, Medina C *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 2008, 5(9): 829-834.

[53] Karp R M. Reducibility among combinatorial problems. In *Proc. Symp. Complexity of Computer Computations*, New York, USA, March 20-22, 1972, pp.85-103.

[54] Leighton T, Rao S. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *JACM*, 1999, 46(6): 787-832.

[55] Powers D. Graph partitioning by eigenvectors. *Lin. Alg. Appl.*, 1988, 101: 121-133.

Limsoon Wong is concurrently a professor of computer science and a professor of pathology at the National University of Singapore. He works mostly on knowledge discovery technologies and their application to biomedicine. He serves on the editorial boards of Information Systems (Elsevier), Journal of Bioinformatics and Computational Biology (ICP), Bioinformatics (OUP), and Drug Discovery Today (Elsevier). He is a scientific advisor to Semantic Discovery Systems (UK), Molecular Connections (India), and CellSafe International (Malaysia).

Guimei Liu is a senior research fellow at National University of Singapore School of Computing. She received her Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2005. Her current research interests include frequent pattern mining and its applications, and protein interaction networks mining and analysis.