Enhanced functional information from predicted protein networks

Jason McDermott and Ram Samudrala

Computational Genomics Group, Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195, USA

Experimentally derived genome-wide protein interaction networks have been useful in the elucidation of functional information that is not evident from examining individual proteins but determination of these networks is complex and time consuming. To address this problem, several computational methods for predicting protein networks in novel genomes have been developed. A recent publication by Date and Marcotte describes the use of phylogenetic profiling for elucidating novel pathways in proteomes that have not been experimentally characterized. This method, in combination with other computational methods for generating protein-interaction networks, might help identify novel functional pathways and enhance functional annotation of individual proteins.

The advent of the 'genomic age' in biology has brought about several new challenges, particularly to the area of computational biology. The vast amount of information already present and becoming available daily is driving the need for new techniques used to derive useful hypotheses from genomic sequence data, even in the absence of experimental data from the particular organism. Protein networks – the representation of the functional, contextual or physical linkages between all proteins in an organism – have been useful in the prediction of function for proteins that cannot be annotated (i.e. assigned a function) by conventional means [1–3]. Date and Marcotte [4] used a phylogenetic profile method to predict functional linkage networks for several organisms and then use the networks to find and describe previously uncharacterized cellular pathways. This approach is one of several new network-based techniques for improving the functional annotation of novel genomes [5,6] and highlights some of the challenges facing the field. Here, we provide an outline of this and similar methods and compare the results of this method to networks predicted by the Bioverse [7] computational framework (http://bioverse.compbio.washington.edu).

Utility of protein-interaction networks

Several experimental techniques have been used to derive protein-interaction networks for yeast and *Helicobacter pylori* [8–10] and these networks exhibit a specific topology and functional modularity [2,11]. The interactions between complexes in specific pathways are highlighted and many previously uncharacterized proteins can be associated with known pathways. Other features of the networks are interesting for biologists, including the observation that highly connected proteins in the yeast network correlate with essential proteins [12].

Prediction of protein networks

A number methods based on evolutionary and/or contextual sequence information have been developed to predict protein–protein interaction and functional relationship networks in novel genomes [5,6,13–16]. Contextual methods include examining patterns of domain fusion across genomes, operon association and gene-order analysis [5,6]. Evolutionary methods include experimental similarity methods (i.e. the identification of pairs of proteins encoded by a target genome similar to pairs of proteins experimentally determined to interact [13,14]) and the phylogenetic profiling methods used by Date and Marcotte [15,16]. Phylogenetic profiling involves the construction of a homolog profile, which measures the occurrences of homologous proteins across a number of genomes for a particular protein. A score describing the co-occurrence of pairs of genes across multiple genomes (mutual information score) is used to predict functional linkages on the assumption that proteins in the same pathway or complex are more likely to be inherited together in the course of evolution. Whereas sequence-similarity methods (and to a certain extent contextual methods) provide predictions of physical protein interactions, phylogenetic profiling provides functional linkages between proteins.

Functional annotation using protein-context networks

Several methods have been described for providing functional annotation for uncharacterized proteins using protein networks [2,6,11,17]. Function prediction based on protein-interaction networks assumes that interacting proteins are likely to share similar functions. The 'majority rule' method annotates a protein by surveying the functions of all the proteins predicted to interact with it and choosing the most frequently occurring function [2]. A more sophisticated method designed for use on predicted protein interaction networks provides a confidence score for each function based on the scores of the functional annotations and the score of the predicted interaction (McDermott and Samudrala, unpublished). Other methods use global network properties [17] or

Corresponding author: Ram Samudrala (ram@compbio.washington.edu).

probability-based models [18] to provide accurate functional annotations.

The phylogenetic profiling prediction method clusters proteins with similar functions in the same area of the network. Date and Marcotte used a predominantly manual method to derive functions for the unknown proteins in their networks. Clusters of proteins in the network with no clear function are identified and extended to include proteins with linkages below the selected threshold or proteins found in the same operon. The function of unknown proteins is then predicted from their location in the network.

Network comparison

Figure 1a shows the largest predicted *E. coli* network generated using Date and Marcotte's phylogenetic profile linkages [Date–Marcotte (DM) network] consisting of 1751 proteins and 12 874 linkages [4, supplementary information]. Figure 1b is the *E. coli* network predicted by the Bioverse (511 proteins; 4075 interactions), based on similarity to experimentally derived interactions. Proteins are colored by broad gene ontology (GO) [19] categories and the 220 proteins shared by both networks are shown with a blue outline. The Date–Marcotte network has a significantly higher average number of connections per protein (15.6 with DM versus 3.8 with Bioverse), similar to that observed in predicted eukaryotic networks (e.g. *C. elegans* network; http://bioverse.compbio.washington.edu). The Bioverse-generated network was more accurate for more specific functional categories but provided fewer annotations (Figure 2).

Implications of network-based functional annotation

Date and Marcotte described a technique for predicting genomic-scale protein networks based on evolutionary information and they have used it to elucidate novel, uncharacterized pathways from genomes. In prokaryotes, these networks provide more coverage than networks predicted by similarity to experimentally determined interactions but the similarity-derived network contains 291 proteins not included in the DM network. In addition the functional resolution of the DM networks is less specific than that in the similarity-derived networks. These factors suggest that the two methods could be combined both to improve the quality of the networks and annotations and to expand their coverage [20].

Identification of uncharacterized conserved pathways is important for providing new insights into cellular function. The method described by Date and Marcotte provides results independent of experimental data suggesting the utility of integrating experimental similarity data with the functional linkages to provide a more complete picture of proteomic-scale networks. The functional linkages provide groupings that can be further resolved by examining predicted protein-interaction networks generated via other methods. Improved prediction of protein networks will allow rapid and accurate functional annotation of newly sequenced genomes and provide a convenient

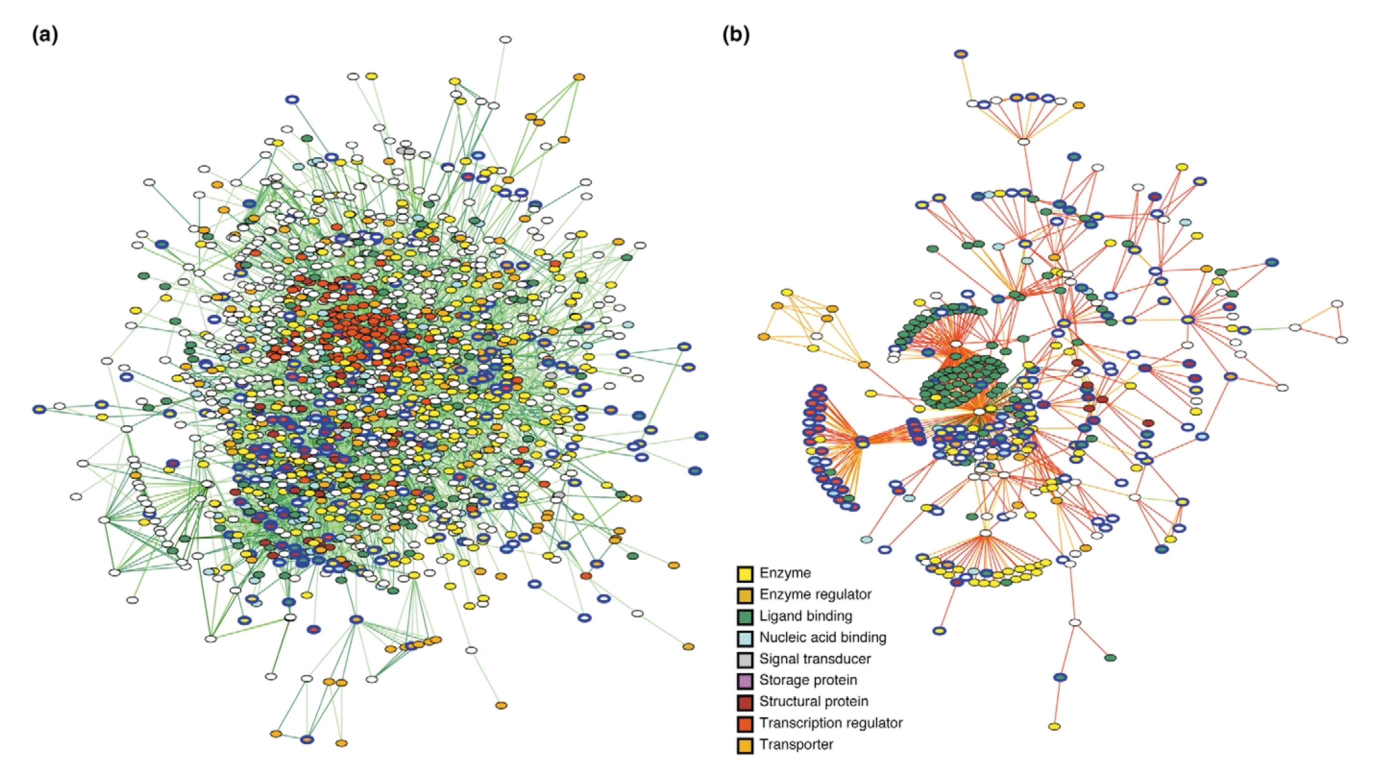

Figure 1. Comparison of predicted protein networks for *E. coli*. (a) Protein pairs and their mutual information scores based on phylogenetic profiling were used to generate a network for *E. coli*. Figure generated using data from [4, supplementary information] (b) Protein interactions were predicted using Bioverse [7] based on finding pairs of proteins similar in sequence to proteins from a database of experimentally determined interactions. Figure generated using data from Bioverse (http://bioverse.compbio.washington.edu). For both networks, nodes representing proteins are colored based on their gene ontology (GO) [19] category and the 220 proteins present in both networks are outlined in blue. Edges represent the predicted relationships between proteins [functional linkages in (a) and protein interactions in (b)] and are colored by confidence (a) or mutual information score (b).

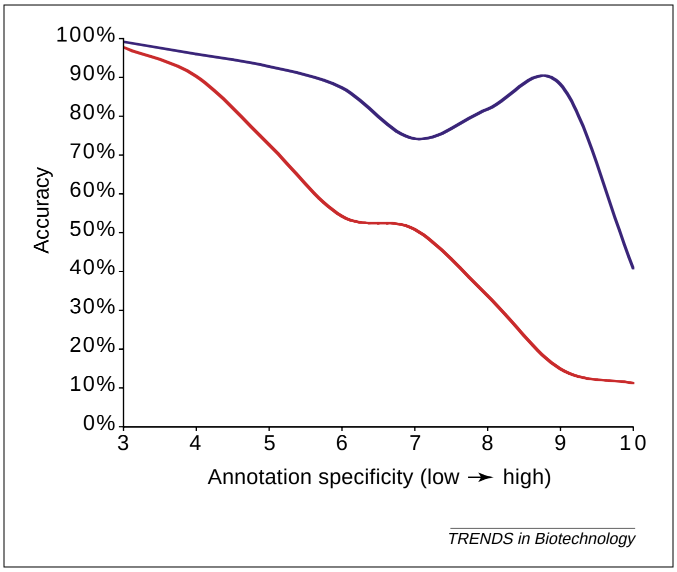

Figure 2. Comparison of functional annotation accuracy using predicted protein networks. The network-based functional annotation accuracy of both the networks depicted in Figure 1 is shown. For proteins with existing functional annotations provided by Bioverse, the accuracy of the network-based annotation was assessed by comparing the existing annotations with the network-based annotations at varying levels of functional specificity. The gene ontology (GO) [19] vocabulary was used because it provides a structured, hierarchal description of protein function. Accuracy of the method on the Bioverse network (blue) or the phylogenetic-profile network (red) is plotted against the specificity of GO category, from broadest (level 3, 47 categories) to most specific (level 8, ~7000 categories). Both methods provide highly accurate functional annotation but the Date and Marcotte networks provide greater genomic coverage than the Bioverse (40% versus 12%, respectively). Figure generated using data from Bioverse (http://bioverse.compbio.washington.edu).

framework for performing functional and evolutionary comparisons between organisms that have not been extensively studied experimentally.

Acknowledgements

We acknowledge support from the following sources: a Searle Scholar Award (to R.S.), NSF Grant DBI-0217241 and the University of Washington's Advanced Technology Initiative in Infectious Diseases.

References

1 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147

2 Schwikowski, B. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261

3 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403

4 Date, S.V. and Marcotte, E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* 21, 1055–1062

5 Huynen, M. *et al.* (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.* 10, 366–370

6 Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10, 359–365

7 McDermott, J. and Samudrala, R. (2003) Bioverse: functional, structural and contextual annotation of proteins and proteomes. *Nucleic Acids Res.* 31, 3736–3737

8 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574

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

10 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627

11 Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1128–1133

12 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42

13 Wojcik, J. *et al.* (2002) Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.* 323, 763–770

14 Matthews, L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.* 11, 2120–2126

15 Ramani, A.K. and Marcotte, E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* 327, 273–284

16 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

17 Vazquez, A. *et al.* (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.* 21, 697–700

18 Letovsky, S. and Kasif, S. (2003) Predicting protein function from protein–protein interaction data: a probabilistic approach. *Bioinformatics* 19 (Suppl. 1), I197–I204

19 The GO Consortium, (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433

20 Yanai, I. and DeLisi, C. (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol.* 3, research0064.1–research0064.12.

0167-7799/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tibtech.2003.11.010

Research Focus Response

Response to McDermott and Samudrala: Enhanced functional information from predicted protein networks

Shailesh V. Date and Edward M. Marcotte

Institute for Cellular and Molecular Biology, 2500 Speedway, MBB 3.232, University of Texas at Austin, Austin, TX 78712, USA

McDermott and Samudrala [1] describe a different, but interesting, approach for protein network reconstruction, than the one described in our recent paper [2]. This underscores the fact that a large number of computational approaches appear to be suitable for recreating protein–protein interactions on a genome-wide scale. Once

Corresponding author: Edward M. Marcotte (marcotte@intron.icmb.utexas.edu).