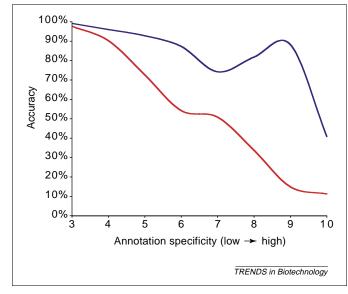**Figure 2**. Comparison of functional annotation accuracy using predicted protein networks. The network-based functional annotation accuracy of both the networks depicted in Figure 1 is shown. For proteins with existing functional annotations provided by Bioverse, the accuracy of the network-based annotation 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.

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.

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).

networks of high quality are obtained, regardless of the method used for defining the network, they can be systematically searched for novel pathways or cellular systems, using the approach we describe [2].

With regard to the relative merits of different methods for calculating networks, the quality of an interaction network can be determined by measuring the accuracy of the interactions described, as well as the number of interactions covered. McDermott and Samudrala [1] compare the quality of networks obtained from phylogenetic profiles with those obtained using Bioverse. In Figure 2 of their article, Bioverse is shown to have a higher accuracy for specific gene ontology categories, whereas the phylogenetic profile network has more than three times the coverage of the Bioverse network. We wish to point out that it is more appropriate to consider coverage and accuracy simultaneously rather than in isolation (as in Figure 2 of the article). These two desirable indicators of quality are inversely related and any method represents a trade-off between them. Instead, the preferred way of estimating the quality of a given network is to consider both parameters at once, an approach known in computer sciences and engineering as 'recall-precision analysis'. The relative accuracy of each method is therefore determined under controlled conditions in which the coverage is held constant. The strength of computational methods for discovering linkages between genes is that each prediction carries with it a measure of confidence. Thus, it is easily possible to perform this test by varying the confidence thresholds and measuring how the coverage and accuracy of each method vary in response – the resulting trends can then be directly compared and the quality of the algorithms assessed.

More important, as pointed out by McDermott and Samudrala, is the fact that phylogenetic profiles and Bioverse represent independent approaches for discovering protein networks, each useful but ultimately limited in scope. We expect to get to truly high quality interaction networks only by integrating information from across the diverse methods that exist for discovering interactions. Such integration promises significant improvements in both accuracy and coverage (e.g. see [3–4]). In practice, this integration requires easy access to the experimental data and computational predictions of many different groups. However, although there have long been community databases archiving experimental sequence, structure and expression data, the functional information extracted from this data is scattered across a myriad of separate publications and web servers. Several model organism databases and open format sequence databases, such as SwissProt (http://www.ebi.ac.uk/swissprot), have made admirable strides towards cataloging this functional data but only a small portion of computational functional analyses are included. Centralization of this information, with uniformity of formats and access, would open up the work of computational biologists to the entire biological community. Most importantly, this would also allow the full weight of evidence for each function and interaction to be examined at once, allowing the consensus view to emerge.

References

1 McDermott, J. Samudrala, R. (2004) Enhanced functional information from predicted protein networks. *Trends Biotechnol.* 22, preceding article in this issue.
2 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
3 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403
4 Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449–453