

Supplemental Appendix A

In the interest of space, we address a number of tangential points in this separate supplementary section, and also provide more detail on our analysis. Specifically, we address issues of novelty of the method, annotation methodology, circularity of the method, and false-positive predictions. Information already included in the main text is not included in this supplement. We also present a useful table illustrating the structure of the Gene Ontology categories and a figure which shows the derivation of the neighborhood weighting method and the determination of the experimental source factor and neighborhood weight filter for the method.

Novelty of the method

Methods of predicting relationships between proteins other than the interolog method have also been used to provide functional annotations. Phylogenetic profiling predicts functional linkages between proteins based on co-inheritance across large phylogenetic distances evolutionary correlation and several groups have reported that proteins linked using this method are more likely to have similar functions (Yanai and DeLisi 2002; Date and Marcotte 2003). These methods have been used to identify novel cellular pathways and to include proteins in known functional pathways (Morett, Korbelt et al. 2003; Bork, Jensen et al. 2004). However, to establish a functional link between two proteins requires examining a large number of evolutionarily related organisms and thus is currently mostly limited to analysis of prokaryotes or individual pathways (Bork, Jensen et al. 2004). Gene fusion analysis and conservation of gene proximity have also been used similarly (Yanai, Derti et al. 2001; Yanai, Mellor et al. 2002), but all these comparative genomics methods are more restrictive in their ability to use data from different sources and provide different types of information than do methods which predict physical interactions such as the interolog method.

Automated annotation methodology

Automated annotations were derived using the following method, which is similar to the Interpro iprscan program (Apweiler, Attwood et al. 2000). The RPS-BLAST (Schaffer, Aravind et al. 2001) program was used to search the BLOCKS database

(Henikoff, Henikoff et al. 1999), PSI-BLAST (Altschul, Gish et al. 1990) was used to search the ProDom database (Corpet, Gouzy et al. 1998), ppsearch (Zdobnov and Apweiler 2001) was used to search for Prosite motifs (Hofmann, Bucher et al. 1999) and hmmer (Eddy 1998) was used to search the Pfam (Bateman, Birney et al. 2000), SMART (Letunic, Goodstadt et al. 2002), Superfamily (Gough and Chothia 2002), and TIGRFAMs (Haft, Loftus et al. 2001) databases. Matches to the databases were combined where possible using the Interpro database (version 7.0; (Apweiler, Attwood et al. 2000)) which provides a mapping from individual family/domain/motif databases to Interpro identifiers, eliminating the considerable redundancy in these databases.

Table I. Gene Ontology (GO) examples

Gene Ontology (GO) path Examples			Level	N
Biological process [GO:0008150]	Cellular component [GO:0005575]	Molecular function [GO:003674]		
Cell growth and/or maintenance [GO:0008151]	Cell [GO:0005632]	Binding [GO:0005488]	3	42
Cell cycle [GO:0007049]	Intracellular [GO:0005622]	Nucleotide binding [GO:0000166]	4	714
M phase [GO:0000279]	Cytoplasm [GO:0005737]	Purine nucleotide binding [GO:0017076]	5	1,069
M phase of mitotic cell cycle [GO:0000087]	Cytoplasmic vesicle [GO:0016023]	Adenyl nucleotide binding [GO:0030554]	6	2,382
Mitosis [GO:0007076]	Coated vesicle [GO:0030135]	ATP binding [GO:0005524]	7	4,361
Mitotic spindle assembly [GO:0007052]	Vesicle coat [GO:0030120]	ATPase [GO:0016887]	8	3,802
Mitotic spindle positioning and orientation [GO:0040001]	Clathrin vesicle coat [GO:0030125]	ATP-binding cassette (ABC) transporter [GO:0004009]	9	3,162
Mitotic spindle orientation [GO:0000132]	Clathrin coat of endocytic vesicle [GO:0030128]	ABC-type efflux porter [GO:00015427]	10	2,149

Example Gene Ontology (GO) paths from least specific (level 3) to highly specific (level 10). Category names and GO identifiers are shown for each main branch of the GO directed acyclic graph (DAG). Level indicates the distance from the GO root and N indicates the number of categories represented at this level.

Circularity of automated annotations

As noted in the main text, automated annotations are performed only by comparison with the family and domain databases listed above. Automated annotation does not incorporate the network-based annotation described in this paper, transfer of

annotation from similar proteins (see below) or any information derived from the interolog determination.

Some computational and manual annotation methods rely on sequence similarity to other proteins for functional assignment (Iliopoulos, Tsoka et al. 2003). For example, protein A may be annotated as a eukaryotic kinase based on weak sequence similarity to protein B from another organism. Using this kind of method, the annotation of protein A is questionable if protein B was annotated through similarity to yet another protein or if the annotation of protein B is itself erroneous. Computational assignment of primary annotations by the Bioverse, and other comparable methods, is based on the detection of sequence similarity to well-curated conserved protein families and motifs (e.g. Pfam, SMART), and therefore avoids this kind of propagation problem. Although prediction of interactions is also based on similarity methods, no functional information is explicitly transferred to the interolog proteins from their experimentally determined counterparts, so primary annotation assignment and interaction prediction are independent of each other.

Computational annotation methods may still have biases and can produce erroneous predictions: The neighborhood weighting and majority-rule methods integrate primary annotations from varying numbers of neighbors and primary annotations may be used more than once in the course of network annotation of an entire network. It is therefore possible that these biases might be amplified and produce artificially inflated results. However, this possibility is not supported by the finding that performance of the majority-rule and/or neighborhood weighting methods applied to yeast and fly networks using their manual annotations which are largely experimentally determined, is comparable to their performance on computationally-assigned annotations.

False positive predictions

The high estimated false positive rate (over 50% (von Mering, Krause et al. 2002)) of experimentally determined interaction data sets means that predicted protein interactions are likely to have much higher false positive rates. Indeed, while developing

the neighborhood weighting method we found that using only interactions predicted by similarity to crystallized complexes provided significantly better precision than using only interactions predicted mainly through large-scale experimental methods (Supplementary Figure I). This indicates that interactions found in such complexes may be more consistent in terms of quality or are more conserved than those determined by methods such as yeast two-hybrid, which may be composed of more organism-specific and/or transient interactions (von Mering, Krause et al. 2002; Bork, Jensen et al. 2004).

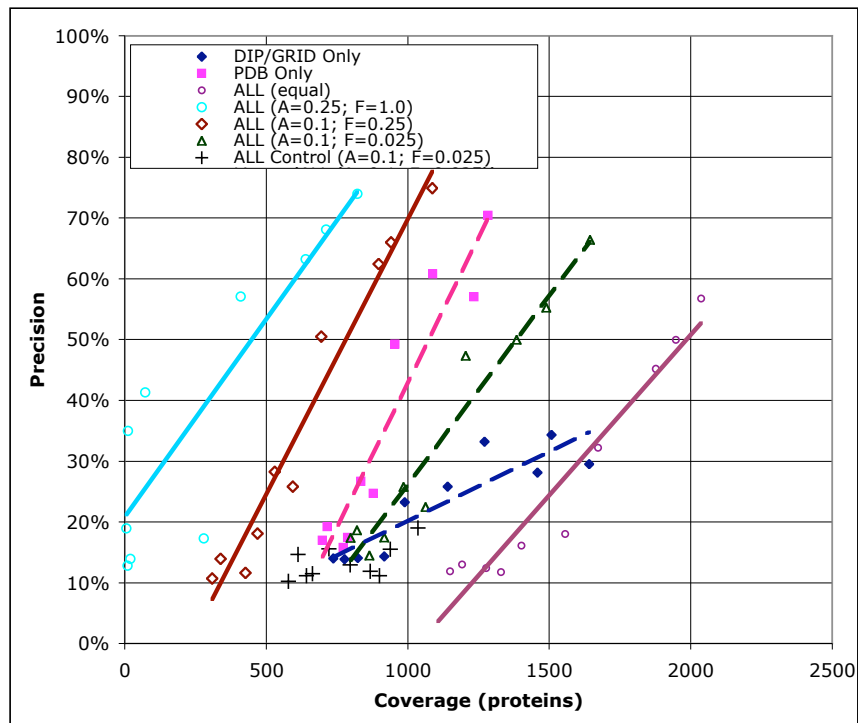
Speed of network annotation

Currently, prediction of protein interaction networks and neighborhood weighting annotation can be performed for 2,000 proteins per hour on our 300+ CPU Linux cluster so even the largest organisms can be annotated in about a day.

Supplementary Figure I. Determination of source contribution and weight

threshold. To determine the best adjustment factor for source data for the neighborhood weighting method a number of different variations on the method were applied to the predicted protein interaction network from *D. melanogaster*. The plot shows the average precision of the variant (ordinate) versus coverage of the variant (abscissa) for the top 10 ranked predictions (data points) for each protein. Predictions are ranked according to their neighborhood weight relative to the other predictions for that protein and so coverage is highest for the top ranked prediction (data point furthest to the right) and lowest for the 10th ranked prediction (data point furthest to the left), since all proteins considered have at least one prediction but not all have 10 predictions. That the top ranked prediction has the highest precision indicates that both the value of the neighborhood weight and the rank of the prediction relative to other predictions for the same protein are important. Using a source factor (A=1.0) and no filter (F=0.0), interactions from the Database of Interacting Proteins (DIP), General Repository for Interaction Data (GRID) and the Protein Databank (PDB) combined (ALL), or the DIP and GRID interactions alone (DIP/GRID; diamonds) or PDB interactions alone (squares)

were considered. Representative variants shown consider all interactions but use a source factor (A) to modify the contribution of interactions derived from the DIP and GRID and exclude all predictions with weights below a filter threshold (F) as described in Methods. The neighborhood weighting variant in which $A=0.1$ and $F=0.025$ (open triangles) provides the best precision and coverage for the method and is used throughout the paper. Also shown are results from application of the neighborhood weighting method to a random control network (crosses). Fitted lines are shown for clarity and dotted lines indicate the PDB alone (left), DIP/GRID alone (right), and the best neighborhood weight variant (middle).



References

- Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-10.
- Apweiler, R., T. K. Attwood, et al. (2000). "InterPro--an integrated documentation resource for protein families, domains and functional sites." Bioinformatics 16(12): 1145-50.
- Bateman, A., E. Birney, et al. (2000). "The Pfam protein families database." Nucleic Acids Res 28(1): 263-6.
- Bork, P., L. J. Jensen, et al. (2004). "Protein interaction networks from yeast to human." Curr Opin Struct Biol 14(3): 292-9.
- Corpet, F., J. Gouzy, et al. (1998). "The ProDom database of protein domain families." Nucleic Acids Res 26(1): 323-6.
- Date, S. V. and E. M. Marcotte (2003). "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages." Nat Biotechnol 21(9): 1055-62.
- Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics 14(9): 755-63.
- Gough, J. and C. Chothia (2002). "SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments." Nucleic Acids Res 30(1): 268-72.
- Haft, D. H., B. J. Loftus, et al. (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." Nucleic Acids Res 29(1): 41-3.
- Henikoff, S., J. G. Henikoff, et al. (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." Bioinformatics 15(6): 471-9.
- Hofmann, K., P. Bucher, et al. (1999). "The PROSITE database, its status in 1999." Nucleic Acids Res 27(1): 215-9.
- Iliopoulos, I., S. Tsoka, et al. (2003). "Evaluation of annotation strategies using an entire genome sequence." Bioinformatics 19(6): 717-26.
- Letunic, I., L. Goodstadt, et al. (2002). "Recent improvements to the SMART domain-based sequence annotation resource." Nucleic Acids Res 30(1): 242-4.
- Morett, E., J. O. Korbel, et al. (2003). "Systematic discovery of analogous enzymes in thiamin biosynthesis." Nat Biotechnol 21(7): 790-5.
- Schaffer, A. A., L. Aravind, et al. (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res 29(14): 2994-3005.
- The Gene Ontology Consortium (2001). "Creating the gene ontology resource: design and implementation." Genome Res 11(8): 1425-33.
- von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature 417(6887): 399-403.
- Yanai, I. and C. DeLisi (2002). "The society of genes: networks of functional links between genes from comparative genomics." Genome Biol 3(11): research0064.
- Yanai, I., A. Derti, et al. (2001). "Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes." Proc Natl Acad Sci U S A 98(14): 7940-5.
- Yanai, I., J. C. Mellor, et al. (2002). "Identifying functional links between genes using conserved chromosomal proximity." Trends Genet 18(4): 176-9.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics 17(9): 847-8.