Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W83–W88***
DOI: 10.1093/nar/gkh411

PathBLAST: a tool for alignment of protein interaction networks

Brian P. Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan[1], Brent R. Stockwell and Trey Ideker[2,*]

Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA, [1]University of California at Berkeley, Berkeley, CA 94720, USA and [2]Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093-0412, USA

Received February 15, 2004; Revised and Accepted April 1, 2004

ABSTRACT

PathBLAST is a network alignment and search tool for comparing protein interaction networks across species to identify protein pathways and complexes that have been conserved by evolution. The basic method searches for high-scoring alignments between pairs of protein interaction paths, for which proteins of the first path are paired with putative orthologs occurring in the same order in the second path. This technique discriminates between true- and false-positive interactions and allows for functional annotation of protein interaction pathways based on similarity to the network of another, well-characterized species. PathBLAST is now available at http://www.pathblast.org/ as a web-based query. In this implementation, the user specifies a short protein interaction path for query against a target protein–protein interaction network selected from a network database. PathBLAST returns a ranked list of matching paths from the target network along with a graphical view of these paths and the overlap among them. Target protein–protein interaction networks are currently available for *Helicobacter pylori*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Just as BLAST enables rapid comparison of protein sequences between genomes, tools such as PathBLAST are enabling comparative genomics at the network level.

INTRODUCTION

A major challenge of post-genomic biology is to understand how genes, proteins and small molecules interact to form signaling and regulatory networks. Recent progress in high-throughput technologies has enabled us to characterize these networks more directly than ever before, using procedures such as the two-hybrid assay (1), co-immunoprecipitation (2) or the chIP-chip approach (3,4) to screen for protein–protein or protein–DNA interactions. To date, these technologies have generated large interaction networks for bacteria (5), yeast (6–10), nematode worm (11) and fruit fly (12).

The enormous amount of data now available on protein interaction networks raises new questions about network evolution and function. These data also introduce a number of technical challenges: how to separate true protein–protein and protein–DNA interactions from false positives (13); how to annotate interactions with functional roles; and, ultimately, how to organize large-scale interaction data into models of cellular signaling and regulatory machinery (14). As is often the case in biology, an approach based on cross-species comparisons may provide a valuable framework for addressing these challenges. By comparing networks drawn from different species or conditions (15–17), it is possible to reinforce the common signal present in both networks while reducing the independent noise (i.e. false-positive interactions). Moreover, network comparisons can be used systematically to catalog all of a cell's conserved network regions, each representing a functionally homologous mechanism or pathway.

We have recently devised a method called PathBLAST (18) to enable comparative network biology of this type (Figure 1). Just as BLAST is used to perform rapid alignment of protein sequences (19), PathBLAST is based on alignment of protein networks. Specifically, PathBLAST searches for high-scoring *pathway alignments* between two paths, one from each network, in which proteins of the first path are paired with putative orthologs occurring in the same order in the second path. Pathway alignments are scored by the degree of protein sequence similarity at each pathway position and by the quality of the protein interactions they contain. To account for experimental error and evolutionary variation between networks, the method also allows for 'gaps' in the pathway

*To whom correspondence should be addressed. Tel: +1 858 822 4558; Fax: +1 858 534 5722; Email: trey@bioeng.ucsd.edu

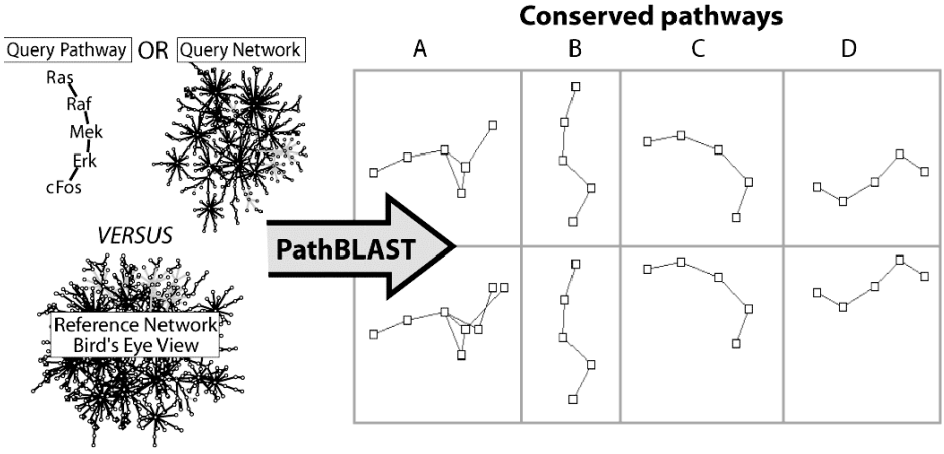

Figure 1. Identifying conserved protein interaction pathways with PathBLAST. PathBLAST operates in two modes, depending on whether the query is a single pathway or a whole network. In the first mode, single user-defined pathways are queried against a reference network of observed protein interactions from bacteria, yeast, fly or worm. In the second mode, two large protein networks are aligned against each other to enumerate all of the pathways that are conserved between them. High-scoring pathway matches (A–D) are ranked by score and indicate pathways that are potentially conserved over evolution. The current focus of the PathBLAST website is on the first (more common) mode of query.

alignment. A gap occurs when interacting proteins in one path are aligned against orthologous proteins in the other path that do not interact directly but are connected at distance two (i.e. both interact via a common protein). PathBLAST implements an efficient search through all possible alignments between two networks to identify the highest scoring pathway alignments overall.

Since reporting on basic algorithmic methods for identifying conserved protein interaction paths (18) or complexes (20), our focus has been on making these methods accessible to the biological community at large. Here, we report development of a server-side PathBLAST query tool available at http://www.pathblast.org/. This tool enables short user-defined pathway queries (paths $\leqslant 5$ proteins) against the current database of observed protein interactions from bacteria, yeast, fly or worm. High-scoring pathway matches are extracted from the interaction database and ranked by score. This search is general, such that the query may consist of proteins and protein interactions from any arbitrary pathway and species provided that the protein sequences are available.

NETWORK QUERIES AND THE PathBLAST SERVER

The core PathBLAST algorithm, as previously reported (18), operates on two protein networks to identify their significant pathway alignments. The available website implementation focuses on the special but practical case in which the first network is a single protein interaction path of interest (Figure 1, top left) and the second network is a complete set of protein–protein interactions that has been experimentally observed for an organism of choice (Figure 1, bottom left). Referring to the first network as the *query* and to the second network as the *target*, PathBLAST outputs all paths in the target that form high-scoring alignments with the query.

Input of query and target

Query submission is modeled strongly after the interface developed by the NCBI for submitting sequence queries via BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). This interface is a well-accepted and intuitive means of extracting information from large bioinformatics databases. The PathBLAST front page prompts users to specify both the pathway query and the target network (Figure 2). The query pathway is specified by entering a sequence of two to five proteins (left column of fields labeled A, B, C in Figure 2). Proteins are identified either by protein ID or by direct input of an amino acid sequence in FASTA format. Recognized protein IDs are the common names for proteins from yeast, bacteria (*Helicobacter pylori*), fruit fly or nematode worm (species for which the target protein networks are available—see below). Systematic open reading frame (ORF) designations are also recognized for these species. Alternatively, protein IDs may be specified as DIP reference numbers used by the Database of Interacting Proteins (21). Direct entry of FASTA sequences is useful for more general pathway queries based on the proteins of species not included in the above list. Several sample pathways are available as a tutorial and for test of the query system.

The target network is specified from a pull-down menu system in the lower left-hand corner of the PathBLAST front page. Target protein–protein interaction networks are drawn from the DIP database (21) and currently include *Saccharomyces cerevisiae* (6,7,9,10), *H.pylori* (5), *Drosophila melanogaster* (12) and *Caenorhabditis elegans* (11). Partial protein–protein interaction networks are also available for *Homo sapiens* and *Mus musculus*.

Putative homologous proteins between the query and the target are defined by specifying a BLAST expected value (*E*-value) threshold. This threshold reduces the potential search space by disregarding all pairs of homologs with a higher-than-specified *E*-value. This does not however, mean, that a protein will necessarily align against the homolog with the best absolute BLAST *E*-value overall (the alignment being constrained by the protein interactions that are present in the network). To calibrate *E*-values of protein sequence homology across different species, a single composite protein sequence database is used for all query pathways, independent of the target species.

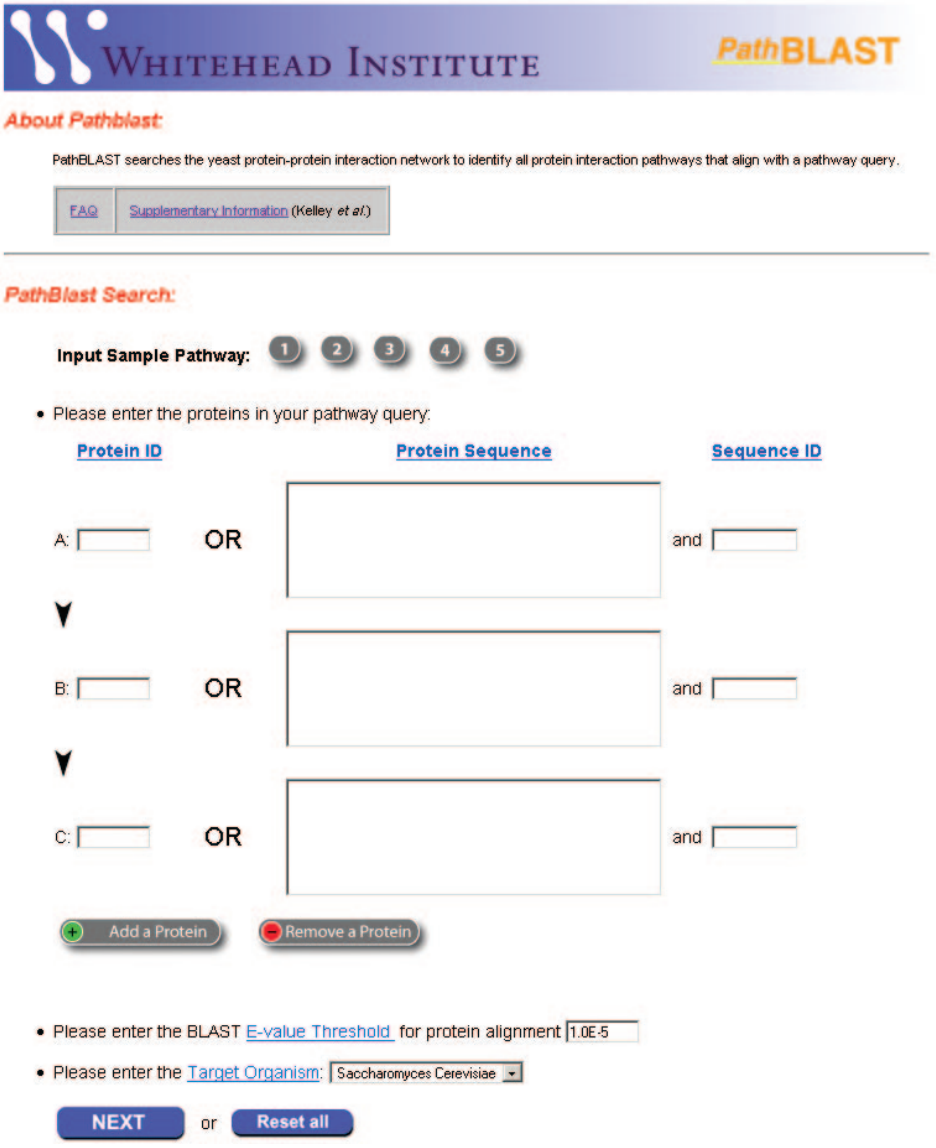

Figure 2. PathBLAST front page. To define the pathway query, users enter a series of protein IDs (DIP number, common name, or systematic ORF designation), or a series of FASTA-format protein sequences, each with a corresponding sequence identifier. The length of the query pathway can be varied between two and five proteins by using the 'Add a Protein' and 'Remove a Protein' buttons. Users must specify the BLAST *E*-value threshold for protein sequence similarity (used to determine which protein pairs should be considered as putative orthologs) as well as the target network for comparison. As part of a brief tutorial, users can evaluate the approach on several example pathways.

Before the PathBLAST search begins, each protein ID is indexed against a local sequence database to obtain the corresponding protein sequence. A dialog appears asking the user to confirm that the correct protein sequences have been matched to each ID. This step identifies erroneous IDs (failure to identify a matching sequence) or ambiguous IDs (multiple matching sequences) before submitting the query to PathBLAST. In the case that FASTA sequences were supplied directly, this sequence lookup is not necessary.

PathBLAST search and output

After confirming the protein sequences corresponding to each ID, the query is submitted to PathBLAST for processing. Query times typically fall in the range of 45–80 s. Query results are returned via a text report and a graphical display. As shown in Figures 3 and 4, the overall report follows a look and feel similar to the manner in which BLAST sequence alignments are returned at the NCBI website. Interestingly, most of the PathBLAST computation time involves construction and layout of the graphical display; the actual PathBLAST alignment and search operation typically completes within a second.

The PathBLAST text report lists the best matching paths (high-scoring pathway alignments) in order of score. In the graphical display, these paths are represented as a network of nodes (proteins) and edges (interactions). High-scoring paths that have one or more proteins in common are merged and shown as overlapping paths in the network. Proteins are color coded to indicate the relative pathway ranking of each aligned

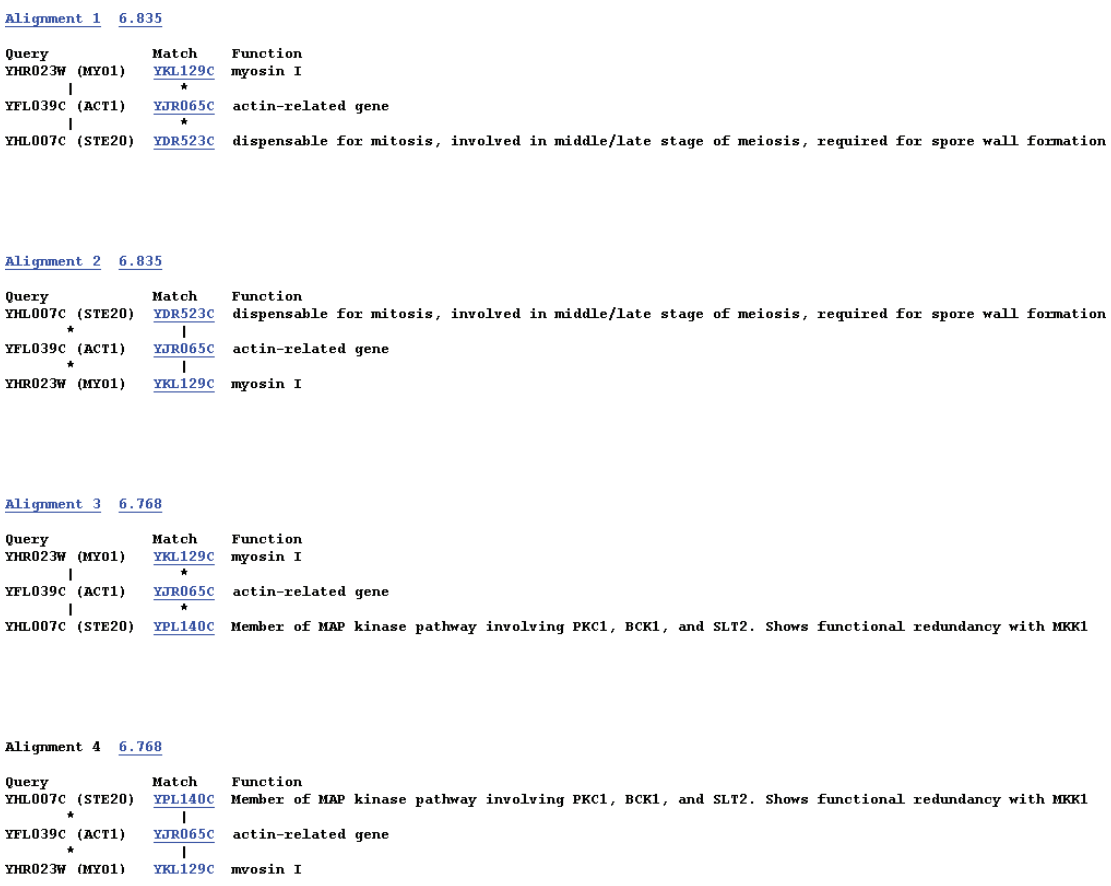

Figure 3. Example of PathBLAST search result. The pathway (Ste20-Act1-Myo1) was used to query the yeast protein–protein interaction database for high-scoring pathway alignment matches. Each matching protein is linked to a functional annotation, if available.

pathway match, with the best scoring match shown in red. In cases in which a protein is involved in more than one high-scoring pathway alignment, the highest scoring pathway is used to determine the node color. When available, each aligned protein is hyperlinked to its corresponding functional annotations drawn from the relevant genome database [i.e. the Saccharomyces Genome Database (22)] for queries against the yeast network, and so on). Together, the PathBLAST report allows users to traverse from individual pathway alignments (Figure 3) to a view of how these alignments overlap with other high-scoring pathway alignments in the target network (Figure 4), to obtaining functional information about the protein members of each pathway (links to the genome databases).

Pathway alignment scoring

The score of each pathway alignment is also reported with each textual and graphical alignment result. As previously described (18), the score is a product of independent probabilities for each aligned protein pair and for each protein interaction. The probability of each protein pair is based on its BLAST *E*-value of sequence alignment, whereas the probability of each protein interaction is based on the false-positive rates associated with interactions in the target network.

Ideally, the score would be calibrated against the expected score distribution for random (permuted) target networks. This calibration would allow derivation of a *P*-value of significance for each pathway alignment. However, computing the expected score distribution empirically is not possible within an acceptable time frame for user-initiated pathway queries (e.g. within a few minutes). Accordingly, *P*-values are not currently generated at the PathBLAST web interface. We are working on techniques to overcome this limitation and expect that as more organisms are added to the PathBLAST network database, it will be possible to derive an approximate formula for obtaining calibrated significances.

DISCUSSION

Beyond the scoring limitation, several directions for PathBLAST website development are of immediate interest. First, although the current focus of the web server is on protein–protein interaction networks, large protein–DNA (transcriptional) networks for *Escherichia coli* (23) and yeast (8) have also recently become available and need to be added to the network database. Handling protein–DNA networks will require changes in network semantics and representation. For instance, for a protein–protein interaction (*a*, *b*), the direction of biological information flow (whether *a* precedes *b*, or vice versa) is typically ambiguous or unknown. Conversely, a protein–DNA interaction (*a*, *b*) typically indicates that protein *a* transcriptionally regulates the gene encoding protein *b*, a non-symmetric relationship (8).

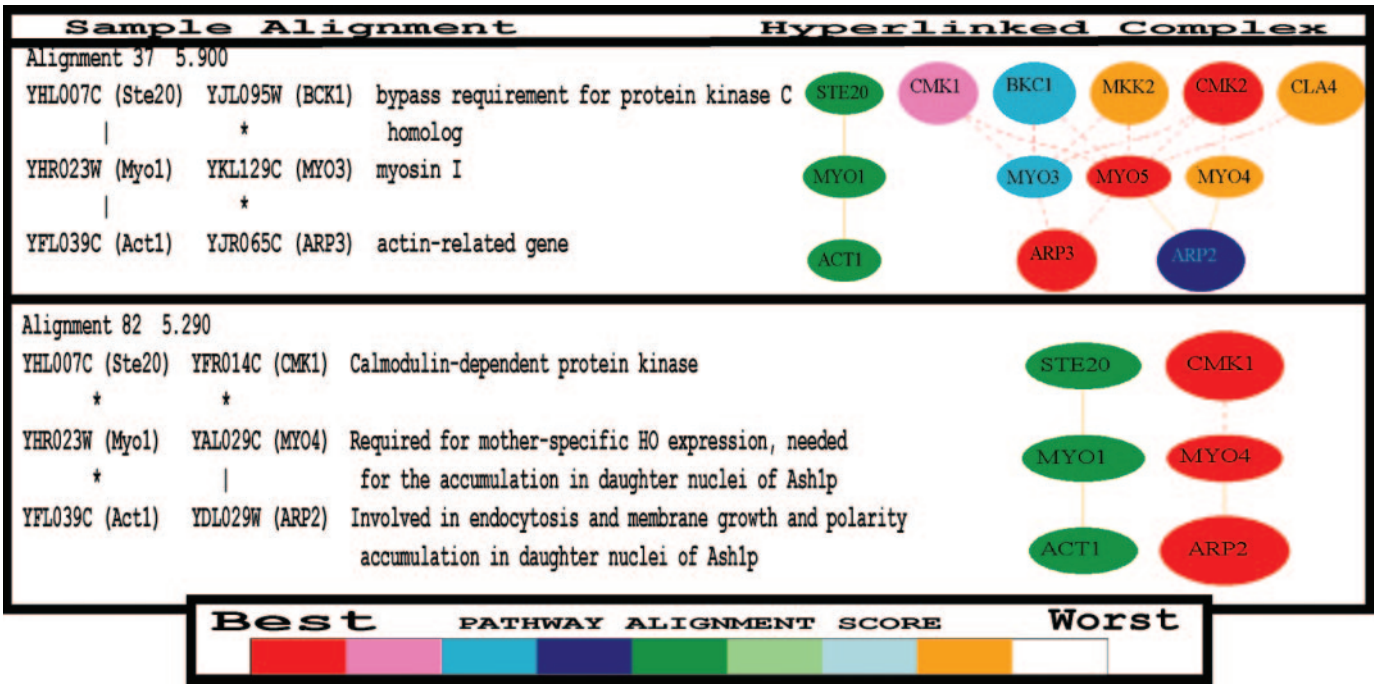

Figure 4. Linked graphical representation. Each pathway alignment is hyperlinked to a graphical representation which, if applicable, shows other high-scoring aligned paths that overlap the present alignment. These overlapping paths are connected in the target network and may shed light on higher order pathway bifurcations or protein complexes of interest. The query pathway (Ste20-Myo1-Act1) is shown on the left in green. Matches from the target network are shown on the right, with relative score encoded by color so that the best scoring match is red, the second best is purple and so on according to the score key at bottom. For instance, the best scoring match in the upper display is (Cmk2-Myo5-Arp3). Solid edges indicate direct protein–protein interactions in the network database, whereas dashed lines indicate gapped alignments (see text).

Second, we are preparing an implementation of PathBLAST that functions as a plug-in to the Cytoscape software environment (24) (http://www.cytoscape.org) for network visualization and modeling. While the website enables alignment of a single path against a larger network database, the PathBLAST plug-in will be optimized for alignment of two large protein networks against each other to catalog all of their conserved pathways. As a plug-in, PathBLAST will leverage Cytoscape's core functionality for network layout, data integration and visualization as well as interact with other plug-in analyses and with the PathBLAST web server. To facilitate this level of integration, we have registered an XML MIME type for transfer of PathBLAST pathway alignments between the website and Cytoscape. This will enable users to hyperlink from a pathway alignment on the website to a Cytoscape session based on the same alignment.

Finally, we will make available a recently published variant of PathBLAST for which the queries are not paths but protein complexes (20). Here, a complex is defined as a 'clique' of proteins in the network, i.e. a set of proteins for which all pairwise interactions are present. The protein complex provides an alternative model for searching for conserved structure within a large protein network.

As protein interaction datasets and the methods for analyzing them mature, we envision that these technologies will be instrumental in extending comparative molecular biology from the level of DNA and protein sequences to the level of protein interaction networks. The focus of the present PathBLAST tool has been on discovery and functional annotation of protein interaction pathways based on similarity to the networks of other, well-characterized species. However, because recurrence of protein interactions across several networks suggests that they are biologically significant, network comparisons are also proving useful for distinguishing between true- and false-positive interactions. Perhaps most attractive of all is the potential impact systematic network comparisons could have on the study and treatment of disease, for instance, by directing drugs to pathways that are present in a pathogenic organism but absent from its human host.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following funding sources: National Cancer Institute Grant 1R01CA97061-01 (B.P.K. and B.R.S.); NSF ITR Grant CCR-0121555 (R.S.); Career Award at the Scientific Interface from the Burroughs Wellcome Fund (B.R.S.); and NCRR grant 1P41RR018627-01 (T.I.).

REFERENCES

1. Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.
2. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
3. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
4. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
5. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The

protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
6. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
7. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
8. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
9. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
10. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
11. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
12. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
13. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
14. Ideker,T. and Lauffenburger,D.A. (2003) Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol*., **21**, 255–262.
15. Forst,C.V. and Schulten,K. (1999) Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J. Comput. Biol*., **6**, 343–360.
16. Ogata,H., Fujibuchi,W., Goto,S. and Kanehisa,M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*., **28**, 4021–4028.
17. Dandekar,T., Schuster,S., Snel,B., Huynen,M. and Bork,P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J*., **343**, 115–124.
18. Kelley,B.P., Sharan,R., Karp,R.M., Sittler,T., Root,D.E., Stockwell,B.R. and Ideker,T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci., USA*, **100**, 11394–11399.
19. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
20. Sharan,R., Ideker,T., Kelley,B.P., Shamir,R. and Karp,R. (2004) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology—RECOMB*, San Diego, CA, pp. 282–289.
21. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.
22. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32**, D311–D314.
23. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
24. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.