cluster for the synthesis of a class of cell wall lipids unique to pathogenic mycobacteria. *J. Biol. Chem.* **272,** 16741–16745 (1997).
9. Mathur, M. & Kolattukudy, P. E. Molecular cloning and sequencing of the gene for mycocerosic acid synthase, a novel fatty acid elongating multifunctional enzyme, from *Mycobacterium tuberculosis* var. bovis Bacillus Calmette-Guerin. *J. Biol. Chem.* **267,** 19388–19395 (1992).
10. Azad, A. K., Sirakova, T. D., Rogers, L. M. & Kolattukudy, P. E. Targeted replacement of the mycocerosic acid synthase gene in *Mycobacterium bovis* BCG produces a mutant that lacks mycosides. *Proc. Natl Acad. Sci. USA* **93,** 4787–4792 (1996).
11. Cole, S. T. *et al.* Deciphering the biology of *Mycobactrium tuberculosis* from the complete genome sequence [see comments]. *Nature* **393,** 537–544 (1998).
12. Fitzmaurice, A. M. & Kolattukudy, P. E. An acyl-CoA synthase (acoas) gene adjacent to the mycocerosic acid synthase (mas) locus is necessary for mycocerosyl lipid synthesis in *Mycobacterium tuberculosis* var. bovis. BCG. *J. Biol. Chem.* **273,** 8033–8039 (1998).
13. Fitzmaurice, A. M. & Kolattukudy, P. E. Open reading frame 3, which is adjacent to the mycocerosic acid synthase gene, is expressed as an acyl coenzyme A synthase in *Mycobacterium bovis* BCG. *J. Bacteriol.* **179,** 2608–2615 (1997).
14. Bystrykh, L. V. *et al.* Production of actinorhodin-related "blue pigments" by *Streptomyces coelicolor* A3(2). *J. Bacteriol.* **178,** 2238–2244 (1996).
15. Pierce, C. H. & Dubos, R. J. Differential characteristics *in vitro* and *in vivo* of several substrains of BCG. II. Morphologic characteristics *in vitro* and *in vivo. Am. Rev. Tuberc. Pulm. Dis.* **74**, 667–682 (1956).
16. Middlebrook, G., Dubos, R. J. & Pierce, C. Virulence and morphological characteristics of mammalian tubercle bacilli. *J. Exp. Med.* **86,** 175–183 (1947).
17. Rainwater, D. L. & Kolattukudy, P. E. Synthesis of mycocerosic acids from methylmalonyl coenzyme A by cell-free extracts of *Mycobacterium tuberculosis* var. bovis BCG. *J. Biol. Chem.* **258,** 2979–2985 (1983).
18. Hunter, S. W. & Brennan, P. J. Further specific extracellular phenolic glycolipid antigens and a related diacylphthiocerol from *Mycobacterium leprae. J. Biol. Chem.* **258,** 7556–7562 (1983).
19. Besra, G. S. & Chatterjee, D. in *Tuberculosis: Pathogenesis, Protection, and Control* (ed. Bloom, B. R.) 285–306 (American Society for Microbiology, Washington DC, 1994).
20. Bardarov, S. *et al.* Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis. Proc. Natl Acad. Sci. USA* **94,** 10961–10966 (1997).
21. Ochman, H., Gerber, A. S. & Hartl, D. L. Genetic applications of an inverse polymerase chain reaction. *Genetics* **120,** 621–623 (1988).
22. Folch, J., Lees, M. & Stanley, G. H. S. A simple method for the isolation and purification of total lipids from animal tissues. *J. Biol. Chem.* **226,** 497–509 (1957).
23. McAdam, R. A. *et al. In vivo* growth characteristics of leucine and methionine auxotrophic mutants of *Mycobacterium bovis* BCG generated by transposon mutagenesis. *Infect. Immun.* **63,** 1004–1012 (1995).

Supplementary information is available from *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank P. Brennan for valuable discussions and for purified PDIM; M. Glickman, J. McKinney, R. Morbidoni and P. Draper for helpful discussions; P. Walter, M. Glickman and J. Blanchard for critical review of the manuscript and encouragement; D. Chatterjee, J. Torrelles, D. Dick and M. Scherman for mass spectrum and NMR analysis; R. Russell for valuable discussions and assistance with histopathology; and R. McAdam and S. Quan for the inverse PCR protocol. J.S.C. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research.

Correspondence and requests for materials should be addressed to W.R.J. Jr (e-mail: jacobs@aecom.yu.edu).

A combined algorithm for genome-wide prediction of protein function

Edward M. Marcotte*†, Matteo Pellegrini†‡, Michael J. Thompson*‡, Todd O. Yeates* & David Eisenberg*

** Molecular Biology Institute, UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, University of California, PO Box 951570, Los Angeles, California 90095, USA*
† These authors contributed equally to this work

The availability of over 20 fully sequenced genomes has driven the development of new methods to find protein function and interactions. Here we group proteins by correlated evolution[1], correlated messenger RNA expression patterns[2] and patterns of domain fusion[3] to determine functional relationships among the 6,217 proteins of the yeast *Saccharomyces cerevisiae*. Using these methods, we discover over 93,000 pairwise links between functionally related yeast proteins. Links between characterized and uncharacterized proteins allow a general function to be assigned to more than half of the 2,557 previously uncharacterized yeast proteins. Examples of functional links are given for a protein family of previously unknown function, a protein whose human homologues are implicated in colon cancer and the yeast prion Sup35.

The historical method of finding the function of a protein involves extensive genetic and biochemical analyses, unless the amino-acid sequence of the protein resembles another whose function is known. With complete genome sequences and total mRNA expression patterns, new strategies become available. We show that the general biochemical functions of proteins can be inferred by associating proteins on the basis of properties other than the similarity between their amino-acid sequences. These properties associate proteins that are functionally related, which we define as proteins that participate in a common structural complex, metabolic pathway, biological process or closely related physiological function. Here we combine three independent methods of prediction[1–3] with available experimental data[3–5] to create a large-scale prediction of protein function which does not rely upon direct sequence homology. These methods allow us to establish many thousands of links between proteins of related function in the yeast *S. cerevisiae*[6].

One might arbitrarily expect each protein coded by a genome to be functionally linked, and therefore to perform closely related functions in the cell, with perhaps 5–50 other proteins, giving 30,000–300,000 biologically meaningful links in yeast. We find 20,749 protein–protein links from correlated phylogenetic profiles[1], 26,013 links from correlated mRNA expression patterns, and 45,502 links from Rosetta Stone sequences of the domain fusion method[3].

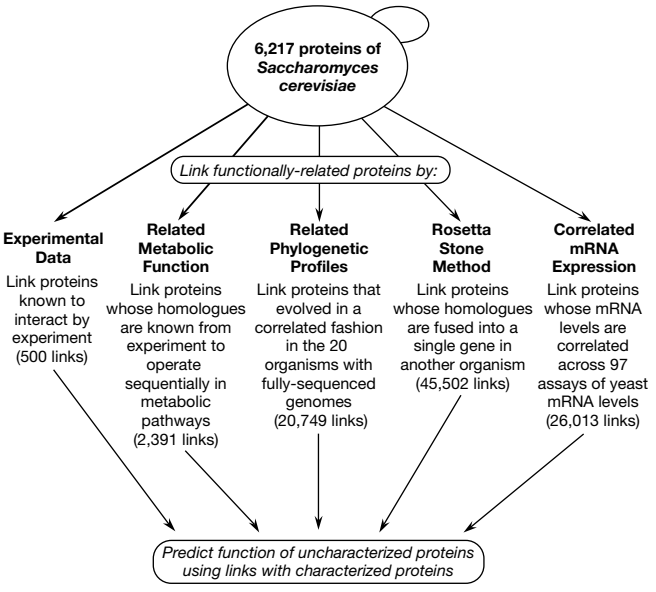

Figure 1 Strategies used to link functionally related yeast proteins, showing the number of links provided by each method. 4,130 very high confidence links were constructed from experimental data, between proteins of related metabolic function, and from links generated by at least two of the prediction methods. 19,521 high confidence links were constructed from the phylogenetic profile method when unconfirmed by other techniques. The remaining links were considered to be of lower confidence. Links associate proteins of related function and can therefore be used to predict the function of many uncharacterized proteins. The phylogenetic profile method[1] identifies functionally related proteins by assuming that proteins that are always inherited together operate together. The fused domain method[3] identifies functionally related proteins by assuming that two proteins function together if they appear in some other organism on the same polypeptide chain, called a Rosetta Stone sequence.

‡ Present address: Protein Pathways, 1145 Gayley Avenue, Ste 304, Los Angeles, California 90024, USA

As shown in Fig. 1, these links were combined with an additional 500 experimentally derived protein–protein interactions from the Database of Interacting Proteins (DIP)[3] and the MIPS yeast genome database[4], and 2,391 links among yeast proteins that catalyse sequential reactions in metabolic pathways[5].

Of the total of 93,750 functional links found among 4,701 (76%) of the yeast proteins, we define 4,130 links to be of the 'highest confidence' (known to be correct by direct experimental techniques or validated by two of the three prediction techniques); 19,521 others are defined as 'high confidence' (predicted by phylogenetic profiles), and the remainder were predicted by either domain fusions or correlated mRNA expression, but not both.

We evaluated the quality of the links by recovery of known protein functions by prediction: we assume that if we link a protein, A, to a group of functionally related proteins, the shared functions of these other proteins provide a clue to the general function for A. Where the function of A is already known, we can test the predicted function. For this test, we chose the standardized keyword annotation of the Swiss-Prot database[7] and systematically compared the known functions of all characterized yeast proteins with the function predicted by our methods. As one example chosen from the many yeast proteins tested, the Swiss-Prot keywords for the enzyme SAICAR synthetase (ADE1), which catalyses the seventh step of *de novo* purine biosynthesis, are 'purine biosynthesis' and 'ligase'. Based upon the frequencies with which keywords appear in the annotation of the 18 proteins linked to ADE1, we predict the general function of ADE1 to be purine biosynthesis (13.6%), lyase (13.6%), transferase (11.4%) and ligase (6.8%). Therefore, we can use this analysis to predict the general biological process that a protein (such as ADE1) participates in, as well as to link the protein to many other proteins of closely related function. We note that, for this example, none of the 18 proteins linked to ADE1 (which include ADE2, ADE5/7, ADE6, ADE8, ADE12, ADE13 and ADE16) shares any sequence similarity to ADE1, and only two pairs are similar to each other. Our results of systematic keyword analyses are listed in Table 1, along with confidence levels, data coverage and comparisons with random trials. The links verified by any two independent prediction techniques predict protein function with the same reliability as experimental interaction data and with over eight times the accuracy of random trials.

Functional links provide a means to characterize proteins of unknown function. There are 2,557 uncharacterized proteins in yeast[4], proteins not studied experimentally and with no strong homologues of known function. Of these, 374 (or 15%) can be assigned a general function from the high and highest confidence functional links and 1,589 (or 62%) can be assigned a general function using all links.

A specific example of the assignment of function is shown in Fig. 2 for yeast open reading frame YGR021W, a member of a highly conserved protein family of unknown function. On the basis of the methods described here and the functional links they uncover to 28 other yeast proteins, this family can now be assigned a function related to mitochondrial protein synthesis. Two of the functional partners of YGR021W are also proteins in conserved families of unknown function: the gidA family and the *Caenorhabditis elegans* M02F4.4 family. These families, too, can now be associated with mitochondrial (or bacterial) protein synthesis. The link to triose-phosphate isomerase (Fig. 2) is of interest, considering the human myopathy in which a deficiency of this enzyme is correlated with grossly altered mitochondrial structure[8].

Two additional examples of links are to the yeast prion Sup35 (ref. 9), and to MSH6, the yeast homologue of human colon-cancer related genes[10,11]. In both cases, a general function is already known, but our method also predicts previously unknown functional links. In Fig. 3, the yeast prion Sup35, which acts as a translation release factor in its non-prion state, is linked with many proteins involved

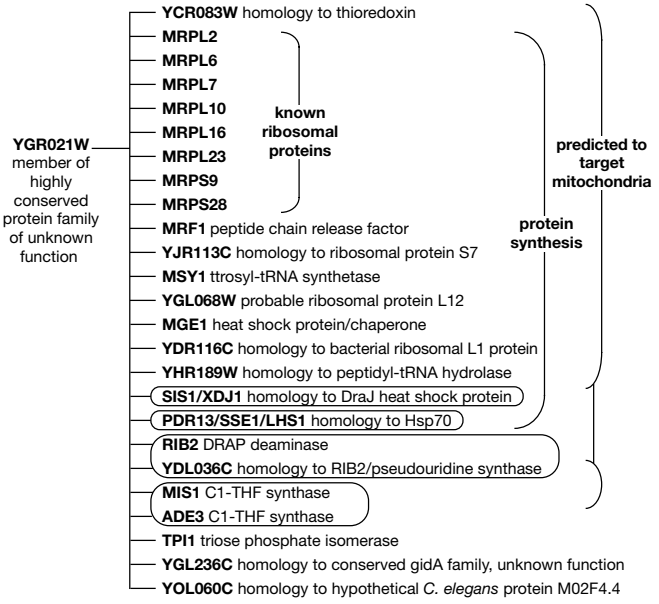

Figure 2 High confidence functional links found by phylogenetic profiles for the yeast protein YGR021W, a member of a protein family conserved in many organisms but of entirely unknown function. The links suggest that members of the YGR021W family operate in mitochondrial protein synthesis (brackets on the right). Of the 28 yeast proteins predicted to have a closely related function, 18 are also independently predicted to be mitochondrial[26], as is YGR021W, supporting the functional assignment. In Figs 2–4, only links from correlated evolution are given unless otherwise noted. Each protein is labelled by its genetic name (for example, MRPL7) or, if absent, by its yeast open reading frame code (for example, YGR021W), and a short description. Families of proteins with related sequences are circled. We note that the 28 linked proteins are all from yeast and are not homologues of YGR021W or (if not circled) of each other; instead, they are functionally linked.

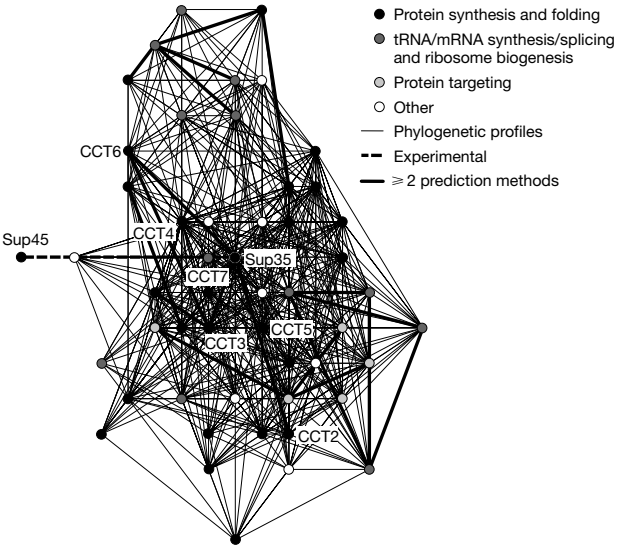

Figure 3 An illustration of the network of high (thin lines) and highest (bold lines) confidence links discovered among the proteins (circles) linked to the yeast prion and translation termination factor Sup35 (double circle). The functions linked to Sup35 are overwhelmingly related to ribosomes and protein synthesis, and include proteins with roles in protein folding, sorting, modification and targeting. For display purposes, proteins were modelled as points and links as springs, thereby positioning functionally related proteins close together.

Table 1 Reliability of functional assignments assessed by recovery of known protein function by prediction

	Number of proteins	Number of functional links	False positive rate* (%)	Ability to predict known function† (%)	Ability in random trials‡ (%)	Signal to noise ratio§
Individual prediction techniques						
Experimental‖	484	500	6.5	33.2	4.0	8.3
Metabolic pathway neighbours	188	2,391	2.5	20.3	4.5	4.5
Phylogenetic profiles	1,976	20,749	29.5	33.1	7.4	4.5
Rosetta Stone method	1,898	45,502	36.4	26.5	7.7	3.4
Correlated mRNA expression	3,387	26,013	35.8	11.5	6.9	1.7
Combined predictions						
Links made by ⩾2 prediction techniques	683	1,249	16.1	55.6	6.9	8.1
Highest confidence links	1,223	4,130	4.8	40.9	5.5	7.4
High confidence links	1,930	19,521	30.6	30.8	7.4	4.2
High and highest confidence links	2,356	23,651	21.8	32.0	6.8	4.7
All links	4,701	93,750	33.1	20.7	7.2	2.9

* The reliability of individual links was calculated as the percentage of pairwise links found between proteins of known function but having no functional categories in common (as tabulated in the MIPS database[4], ignoring the functional categories 'unclassified' and 'classification not clear cut'). This estimate of false positives assumes complete knowledge of protein function and is therefore an upper limit. By this test, random links achieve a false positive rate of ~47%.

† The predictive power of individual techniques and combinations of techniques was evaluated by automated comparison of annotation keywords. By the methods listed, each protein is linked to one or more neighbour proteins. For characterized proteins ('query' proteins), the mean recovery of known Swiss-Prot keyword annotation by the keyword annotation of linked neighbours was calculated as:

$$\langle \text{keyword recovery} \rangle = \frac{1}{A} \sum_{i=1}^{A} \sum_{j=1}^{x} \frac{n_j}{N}, \quad (1)$$

where A is the number of annotated proteins, x is the number of query protein Swiss-Prot keywords, N is the total number of neighbour protein Swiss-Prot keywords, and n_j is the number of times query protein keyword j occurs in the neighbour protein annotation. Because functional annotations typically consist of multiple keywords, both specific and general, even truly related proteins show only a partial keyword overlap (for example, ~35%).

‡ Mean recovery of Swiss-Prot keyword annotation for query proteins of known function by Swiss-Prot keyword annotation of randomly chosen linked neighbours, calculated as in equation (1) for the same number of links as exist for real links (averages of 10 trials).

§ Calculated as ratio of known function recovered by real links to that recovered by random links. Although individual links have only moderate accuracy, combining information from many links significantly enhances prediction of function.

‖ Experimentally observed yeast protein–protein interactions contained in the DIP[3] and MIPS[4] databases.

in protein synthesis, consistent with the primary role of Sup35, which interacts with the ribosome to release the newly synthesized peptide chain[12,13]. Also linked to Sup35 are protein sorting and targeting proteins, consistent with an accessory role in guiding nascent proteins to their final cellular destinations. Sup35 shows both correlated evolution and correlated mRNA expression with the CCT chaperonin system, a yeast chaperonin system believed to aid folding of newly synthesized actin and microtubules[14].

New links are also established for MSH6, a DNA mismatch repair protein[15] whose human homologues, when mutated, cause most hereditary nonpolypoid colorectal cancers[16]. MSH6 is linked in Fig. 4 to the sequence-unrelated PMS1 DNA mismatch repair protein family, mutations of which are also tied to human colorectal cancer[17]. MSH6 is in turn linked via homologue MSH4 to the purine biosynthetic pathway by methylenetetrahydrofolate dehydrogenase[18], to two RNA modification enzymes, and to an uncharacterized protein family, which can now be investigated by taking nucleic acid repair or modification into consideration.

The methods of phylogenetic profiles[1] and domain fusions[3], as well as gene localization[19,20], although they indirectly rely on sequence matching, discover links between non-homologous proteins, which provide much new information about functional relationships and hence go beyond the capabilities of traditional sequence matching. Analyses of mRNA co-expression, and potentially protein co-expression, are entirely independent of sequence matching and therefore provide information about proteins unique to the organism examined; this information is entirely inaccessible from sequence-matching techniques. The analysis of protein relationships by the methods we describe, especially when enhanced by other sources of functional relationships, should substantially speed the discovery of protein function. Functional predictions are available at http://www.doe-mbi.ucla.edu.

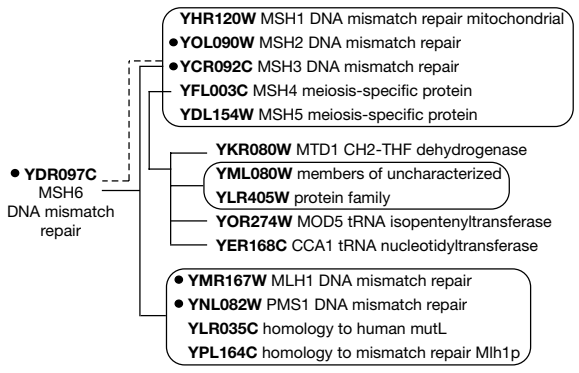

Figure 4 High and highest confidence functional links found for the yeast DNA repair protein MSH6, which is similar in sequence to colorectal cancer causing proteins in humans[10,11]. Families of proteins with related sequences (circled), proteins implicated in human cancers (bullet points), and homology of MSH6 to the MSH DNA mismatch repair family (dashed line) are shown. MSH6 is also linked to MSH2 by correlated mRNA expression.

Methods

Experimental interactions

Pairwise links were created between yeast proteins, known from experimental literature to interact, using techniques such as co-immunoprecipitation and two-hybrid methods. Experimental observations include both direct *in vitro* measurements as well as indirect measurements of protein–protein interactions such as yeast two-hybrid data. We combined interaction data from the MIPS database[4] and the DIP[3], a community-developed database of protein–protein interactions. At the time this work was carried out, DIP contained 179 interactions between yeast proteins.

Linking of metabolic pathway neighbours

Yeast homologues of *Escherischia coli* proteins were found by BLAST[21] homology searches. Pairwise links were defined between yeast proteins whose *E. coli* homologues catalyse sequential reactions (or one reaction step further away) in metabolic pathways, as defined in the EcoCyc database[5].

Calculation of correlated evolution

Phylogenetic profiles were constructed for each yeast protein encoding the appearance of homologous amino-acid sequences in other organisms, as described in ref. 1, with some modifications. Instead of describing the presence or absence of yeast homologues in the full sequences genomes of 19 other organisms by a single bit, a real number expressing the evolutionary distance between the homologues (to be described elsewhere) was substituted, allowing functionally related proteins to be clustered by short euclidean distances between profile strings. The 20 completely sequenced genomes encoded in this way are available at the TIGR web site (http://www.tigr.org/tdb/mdb/mdb.html).

Calculation of correlated mRNA expression

Results of 97 individual publicly available DNA chip yeast mRNA expression data sets[22–25] were encoded as a string of 97 numbers associated with each yeast open reading frame (ORF) describing how the mRNA of that ORF changed levels during normal growth, glucose starvation, sporulation and expression of mutant genes. This string is the analogue within one organism of a phylogenetic profile[1]. The mRNA levels for each of the 97 experiments were normalized, and only genes that showed a two-standard-deviation change from the mean in at least one experiment were accepted, thereby ignoring genes that showed no change in expression levels for any experiment. Open reading frames with correlated expression patterns were grouped together by calculating the 97-dimensional euclidean distance that describes the similarity in mRNA expression patterns. Open reading frames were considered to be linked if they were among the 10 closest neighbours within a given distance cut-off, conditions that maximized the overlap of ORF annotation between neighbours.

Calculation of domain fusions

Proteins were linked by Rosetta Stone patterns as in ref. 3. Alignments were found with the program PSI-BLAST[21]. □

Received 7 May; accepted 23 August 1999.

1. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96,** 4285–4288 (1999).
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
3. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285,** 751–753 (1999).
4. Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.* **26,** 33–37 (1998).
5. Karp, P., Riley, M., Paley, S. & Pellegrini-Toole, A. EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **26,** 50–53 (1998).
6. The yeast genome directory. *Nature* **387** (suppl), 1–105 (1997).
7. Bairoch, A. & Apewiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27,** 49–54 (1999).
8. Bardosi, A., Eber, S. W., Hendrys, M. & Pekrun, A. Myopathy with altered mitochondria due to a triosephosphate isomerase (TPI) deficiency. *Acta Neuropathol. (Berl.)* **79,** 387–394 (1990).
9. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264,** 566–569 (1994).
10. Miyaki, M. *et al.* Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature Struct. Biol.* **17,** 271–272 (1997).
11. Fishel, R. *et al.* The human mutator gene homologue MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75,** 1027–1038 (1993).
12. Kushirov, V. V. *et al.* Nucleotide sequence of the Sup2(Sup35) gene of *Saccharomyces cerevisiae*. *Gene* **66,** 45–54 (1988).
13. Stansfield, I. *et al.* The products of the SUP45 (eRF1) and SUP35 genes interact to mediate translation termination in *Saccharomyces cerevisiae*. *EMBO J.* **14,** 4365–4373 (1995).
14. Chen, X., Sullivan, D. S. & Huffaker, T. C. Two yeast genes with similarity to TCP-1 are required for microtubule and actin function *in vivo*. *Proc. Natl Acad. Sci. USA* **91,** 9111–9115 (1994).
15. Johnson, R. E., Kovvali, G. K., Prakash, L. & Prakash, S. Requirement of the yeast MSH3 and MSH6 genes for MSH2-dependent genomic stability. *J. Biol. Chem.* **271,** 7285–7288 (1996).
16. Lynch, H. T., Fusaro, R. M. & Lynch, J. F. Cancer genetics in the new era of molecular biology. *Ann. NY Acad. Sci.* **833,** 1–28 (1997).
17. Papadopolous, N. *et al.* Mutations of a MutL homolog in hereditary colon cancer. *Science* **263,** 1625–1629 (1994).
18. West, M. G., Horne, D. W. & Appling, D. R. Metabolic role of cytoplasmic isozymes of 5,10-methylenetetrahydrofolate dehydrogenase in *Saccharomyces cerevisiae*. *Biochemistry* **35,** 3122–3132 (1996).
19. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23,** 324–328 (1998).
20. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96,** 2896–2901 (1999).
21. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
22. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297 (1998).
23. DeRisi, J. L., Iyer, V. R. & Brown, P. O Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686 (1998).
24. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).
25. Myers, L. C., Gustafsson, C. M., Hayashibara, K. C., Brown, P. O. & Kornberg, R. D. Mediator protein mutations that selectively abolish activated transcription. *Proc. Natl Acad. Sci. USA* **96,** 67–72 (1999).
26. Horton, P. & Nakai, K. Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *Intell. Sys. Molec. Biol.* **5,** 147–152 (1997).

Acknowledgements

This work was supported by a Department of Energy/Oak Ridge Institute for Science and Education Hollaender postdoctoral Fellowship (E.M.), a Sloan Foundation/Department of Energy postdoctoral fellowship (M.P.), and grants from the DOE.

Correspondence and requests for materials should be addressed to D.E. (e-mail: david@mbi.ucla.edu).

Protein interaction maps for complete genomes based on gene fusion events

Anton J. Enright, Ioannis Iliopoulos, Nikos C. Kyrpides* & Christos A. Ouzounis

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

** Integrated Genomics Inc., 2201 West Campbell Park Drive, Chicago, Illinois 60612, USA*

A large-scale effort to measure, detect and analyse protein–protein interactions using experimental methods is under way[1,2]. These include biochemistry such as co-immunoprecipitation or crosslinking, molecular biology such as the two-hybrid system or phage display, and genetics such as unlinked noncomplementing mutant detection[3]. Using the two-hybrid system[4], an international effort to analyse the complete yeast genome is in progress[5]. Evidently, all these approaches are tedious, labour intensive and inaccurate[6]. From a computational perspective, the question is how can we predict that two proteins interact from structure or sequence alone. Here we present a method that identifies gene-fusion events in complete genomes, solely based on sequence comparison. Because there must be selective pressure for certain genes to be fused over the course of evolution, we are able to predict functional associations of proteins. We show that 215 genes or proteins in the complete genomes of *Escherichia coli*,

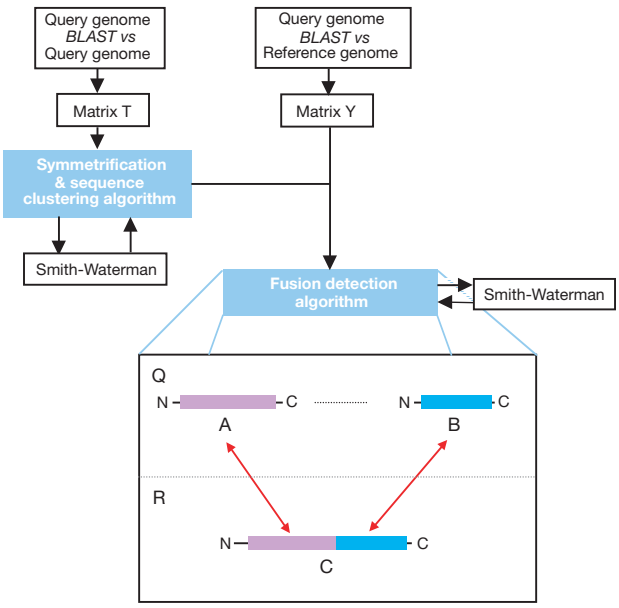

Figure 1 Flowchart of the algorithm. All similarities within the query genome Q detected using BLAST[23] are stored in matrix T. For all nonsymmetrical hits, an additional Smith–Waterman comparison[26] is used to resolve false negatives. The query genome is then compared with the reference genome (or database), and similarities are stored in matrix Y. The fusion-detection algorithm identifies cases of the form depicted in the inset, where query (Q) proteins A and B exhibit similarity to reference (R) protein C by checking matrix Y, but not to each other, by checking matrix T (which is further confirmed by an additional Smith–Waterman comparison). Both Smith–Waterman runs are executed an additional 25 times, with randomization of the sequences, and a Z-score is obtained: if the Z-score is higher than a certain threshold, the similarity is accepted as significant. The 'key' abstraction is that a candidate pair (A,B) of query proteins can either represent a false-negative, or a component pair matching the composite protein C. Total computation time is ~4 h on an SGI Octane two-processor workstation.