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*

tionally 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

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

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

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						
Experimentall	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

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).

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*,