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*,

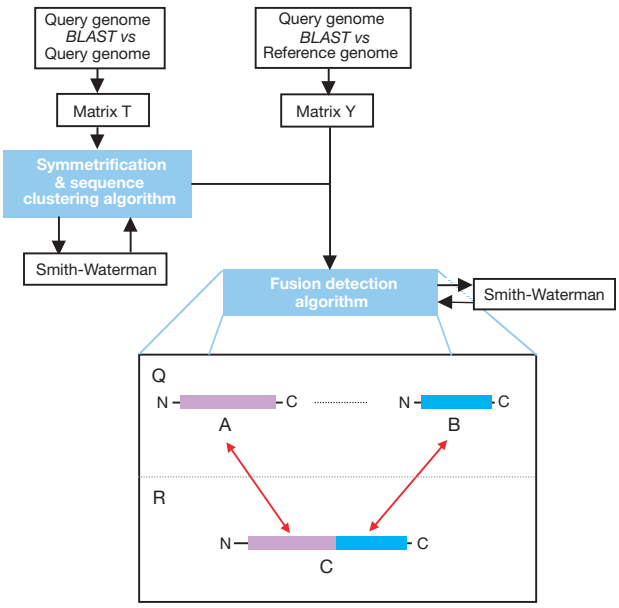

Figure 1 Flowchart of the algorithm. All similarities within the query genome Q detected using BLAST[23] are stored in matrix T. For all nonsymmetrical hits, an additional Smith–Waterman comparison[26] is used to resolve false negatives. The query genome is then compared with the reference genome (or 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.

***Haemophilus influenzae* and *Methanococcus jannaschii* are involved in 64 unique fusion events. The approach is general, and can be applied even to genes of unknown function.**

Previously, the only computational approach to the problem of protein–protein interactions has been the study of subunit interfaces, using information from the protein structure database[7–9]. Recently, a number of studies have exploited the ordering of related genes in genome sequences as a basis for the identification of interacting proteins[10,11]. This approach relies on the notion that gene proximity is a result of selective pressure to associate genes that are co-regulated and possibly interacting[10,12]. These analyses result in a number of false predictions, however, because the constraint of proximity is not strong, and interactions between products of distantly located genes are not identifiable. In addition, this

Table 1 The 64 fusion events in the genomes of *E. coli*, *H. influenzae* and *M. jannaschii* detected on the basis of composite proteins in these three genomes and the genome of *S. cerevisiae*.

Case	Component	Component	Composite	EC	HI	MJ	SC	N
1	GalE	GalM	GAL10	▲▼	▲▼		●1	2
2	AccC	B0712-hypothetical	DUR1,2	▲▼	▲▼		●1	2
3	Hypothetical	Hypothetical	PYC2,PYC1			▲▼	●2	1
4	HisH	HisF	HIS7	▲▼	▲▼	▲▼	●1	3
5	HisI(E)	HisD	HIS4	▲▼	▲▼	▲▼	●1	3
6	RpoA′	RpoA″	RPO21,RPO31,RPA190			▲▼	●3	1
7	GltB	GltD	GLT1	▲▼			●1	1
8	AroB/AroA/AroK/AroD/AroE	Multiple fusion	ARO1	▲▼▲▼▲	▲▼▲□□	□▼□□▲	●1	3
9	Aconitase subunit	Aconitase subunit	LYS4			▲▼	●1	1
10	ArgA	ArgC	ARG5,6	▲▼			●1	1
11	LeuC	LeuD	LEU1	▲▼	▲▼	▲▼	●1	3
12	TrpA	TrpB	TRP5	▲▼	▲▼	▲▼	●1	3
13	PurD	PurM	ADE5,7	▲▼	▲▼	▲▼	●1	3
14	PurL	PurQ	ADE6	●1	●1	▲▼	●1	1
15	CarA/CarB/PyrB	Multiple fusion	URA2	▲▼▲		▲▼▲	●1	2
16	B1378	CysI	ECM17	▲▼			●1	1
17	TrpG	TrpC	TRP3	▲▼	▲▼	▲▼	●1	3
18	AgaG	AgaF	HyuA,HuyB			▲▼	●1	1
19	IlvG_1	IlvG_2	ILV2	▲▼	●1	●2	●1	1
20	GmpA	GmpB	GUA1	●1	●1	▲▼	●1	1
21	GyrB,ParE	GyrA,ParC	TOP2	▲▲▼▼	▲▲▼▼		●1	4
22	FolK	FolP	Folate biosynthesis (probable)	▲▼	▲▼▼		●1	3
23	PabA	PabB	ABZ1	▲▼	▲▼		●1	2
24	PurK	PurE	ADE2	▲▼	▲▼	▲▼	●1	3
25	RpoB″	RpoB′	RPB2,RET1,A135	●1	●1	▲▼	●3	1
26	ThiE	ThiM	THI6		▲▼		●1	1
27	ThiD	TenA	thi21,thi20,thi22		▲▼		●3	1
28	TklA	TklB	TKL1,TKL2	●2	●1	▲▼	●2	1
29	LysC	hom	ThrA,MetL	●2	●1	▲▼		1
30	ABC transporter	Hypothetical	B0879	●1		▲▲▼▼		4
31	Hypothetical	Putative methyltransferase	B0948	●1		▲▼▼		2
32	TrpG	TrpD	TrpD	●1	▲▼	▲▼		2
33	FumA	FumB	FumA	●1		▲▼		1
34	Hypothetical tkt	PheA	PheA	●1	●1	▲▼		1
35	FprA	Rubredoxin	B2710	●1		▲▲▲▼▼		6
36	TrxM	Hypothetical	B0492	●1	▲▼			1
37	Hypothetical	Hypothetical	B1816,B2063	●2	▲▼▼▼			3
38	CpxR,YgiX	TyrR	AtoC,YfhA,GlnG,HydG	●4	▲▲▼			2
39	Hypothetical	Multiple fusion	B2324	●1	▲▼▲			1
40	Hypothetical	Hypothetical	B2474	●1	▲▼			1
41	Hypothetical	Hypothetical	SufI	●1	▲▼			1
42	HemX	Hypothetical	HemX	●1	▲▼			1
43	TrpC	TrpF	TrpC		●1	▲▼		1
44	CitX	CitG	CitG	▲▼	●1			1
45	SbmA	Hypothetical	ABC transporter/ATP-binding	▲▼▼▼▼▼▼▼	●1			7
46	B3777	B3776	Hypothetical	▲▼	●1			1
47	B2612	YfjD	Hypothetical	▲▼	●1			1
48	YgfQ	YgfR	Hypothetical	▲▼	●1	●1		1
49	YabK	B0263	Hypothetical	▲▼	●1			1
50	YhaQ	YhaP	SdaA	▲▼	●1			1
51	YbfH	YbfG	Hypothetical	▲▼	●1			1
52	PurF	YhfN	GlmS	▲▼	●1			1
53	FrwB,FrwD	FrwC,B2386	FruA	▲▲▼▼	●1			4
54	UgpC	YtfS	RbsA,MglA	▲▼	●2			1
55	B1515,B1899	AraH	RbsC,MglC	▲▲▼	●2			2
56	NrfF	NrfG	NrfF	▲▼	●1			1
57	MsrA	B1778	MsrA	▲▼	●1			1
58	SbmA	Hypothetical	ABC transporter/ATP-binding	▲▼▼▼▼▼▼▼	●1			7
59	YhgK	YhgJ	Probable RNA cyclase	▲▼		●1		1
60	FrdB	GlpC	Iron-sulfur-binding reductase	▲▼		●1		1
61	RffH,RfbA,GalF,GalU	B0359	Glucose-1-P thymidyltransferase	▲▲▲▲▼		●1		4
62	B3016	B3015	Hypothetical	▲▼		●1		1
63	LeuS	YgjH	MetS	▲▼		●1		1
64	TopB	YrdD	TopA	▲▼	▲▼	●1		2
								122

The component gene/protein names (or identifiers) and the composite (fusion) gene/protein names (or identifiers) are listed. Columns EC, HI, MJ and SC correspond to *E. coli*, *H. influenzae*, *M. jannaschii* and *S. cerevisiae*, respectively; N lists the maximum number of possible pairwise interactions taking into account paralogy in the query genomes (multiple-component cases are counted as a single interaction). Symbols represent a corresponding component or composite genes/proteins: triangle pairs, ▲▼, a pair of component proteins in the query genome predicted to interact based on their similarity to a composite protein in the reference genome; alternating triangles, ▲▼▲/▲▼▲▼▲, multiple-component genes/proteins (cases 8, 15 and 39); open squares, □, absence of a component from a multiple-fusion event (case 8); consecutive triangles, ▲▲…/▼▼…, the exact number of detected paralogous component genes/proteins in the query genome; filled circles, composite genes/proteins, the number represents the number of paralogous composite genes/proteins in the reference genome. The sort order follows the three species against the composite-protein sequence identifiers for the yeast genome, and then the other three species in succession. Genes are named where possible; where none is available, the sequence identifier is used instead. All fusions were confirmed by reverse BLAST searches using the composite protein as query, which identified all the component proteins. Note that functional annotation is not necessary but frequently useful in resolving paralogous cases (for example, case 21). Predictions imply functional associations and not necessarily direct molecular interactions. For gene/protein identifiers and references for the known cases, see Supplementary Information.

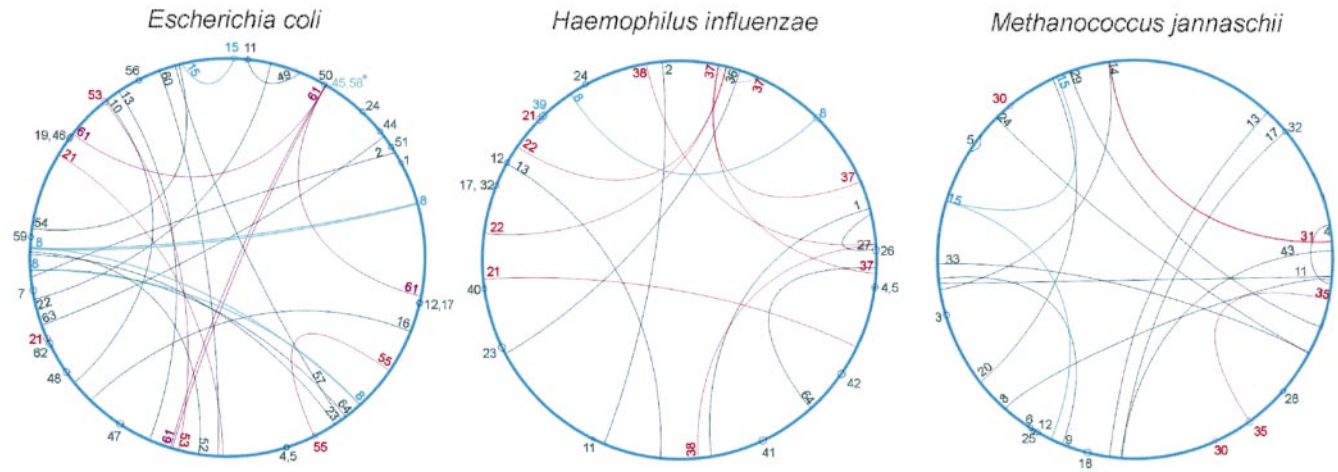

Figure 2 Representation of protein interaction maps for the most likely interactions predicted for *E. coli*, *H. influenzae* and *M. jannaschii*. In the large blue circles, which represent the three genomes, 0° corresponds to the first base pair, and 360° the last base pair of the genome. Predicted interactions are indicated by linking the circular map positions of the genes involved. In cases of neighbouring genes (<5°), a small circle indicates the predicted interaction between two genes at that region; otherwise, an arc links the two genes in question. Multiple interactions are not cross-labelled. Some paralogous cases are resolved and only the most likely case is indicated by an arc. All cases are numbered according to Table 1. Predictions are colour coded: black, pairwise interactions; blue, multiple interactions; red/purple, cases where, due to paralogy, more than one pairwise interaction is possible (red, two possibilities; purple, more than two possibilities); green (marked by asterisk), because of a large number of paralogues, no interaction can be easily resolved. The source of the prediction (composite protein from a given species) is not indicated.

approach may not be applicable to eukaryotes, because the co-regulation of genes is not imposed at the genome structure level. Another approach, based on the mere presence/absence of genes in different species, also attempts to identify functionally associated genes[13].

Our method overcomes some of these problems by exploiting the fact that certain protein families in a given species consist of fused domains that usually correspond to single, full-length proteins in other species. We define these proteins as 'composite' proteins (otherwise known as fusion proteins) and their individual domains as 'component' proteins. We rely primarily on complete genome sequences for the identification of fusion events because we can detect 'orthologous' (equivalent) proteins across species. The underlying assumption is that if a composite protein is uniquely similar to two component proteins in another species, which may not necessarily be encoded by neighbouring genes, the component proteins are most likely to interact (Fig. 1; see Methods). We use the term 'interaction' to imply either direct physical interaction or an indirect functional association (for example, involvement in the same biochemical pathway or similar gene regulation).

The situation is comparable to experimental approaches that make use of artificial constructs of fused genes for biochemical analysis and protein-purification technology[14,15]. The process has also been observed in evolution, perhaps the most widely known example being the fusion of tryptophan synthetase α- and β-subunits from bacteria to fungi[16].

We identified 215 component proteins in the 3 query genomes that are involved in 88 fusion events (of which 64 are unique) (Fig. 2; triangles in Table 1). There are 39 fusion events in *E. coli*, 24 in *H. influenzae* and 25 in *M. jannaschii*, with 2.44 proteins per fusion event on average (due to both multiple-fusion events, for example, case 8, and paralogous genes, for example, case 21, Table 1).

We identified 94 composite proteins in the 4 reference genomes (representing 77 fusion cases or protein families, circles in Table 1; there are 17 paralogous composite proteins). The *E. coli* genome contains only 24 composite proteins (18 fusion-protein families) with reference to the component proteins, as opposed to the much smaller *H. influenzae* genome which contains 25 composite proteins (23 fusion-protein families). In contrast, the *M. jannaschii* genome has only 9 composite proteins (8 fusion-protein families). The remaining 36 composite proteins (28 fusion-protein families) are detected in the genome of *Saccharomyces cerevisiae*. Because query genomes were in turn used as reference genomes (see Methods), we detected eight cases shared between them: *E. coli* and *H. influenzae* share six; *E. coli* and *M. jannaschii* none; and *H. influenzae* and *M. jannaschii* two (Table 1).

There are only three multiple-fusion events (case 8 deriving from the yeast ARO1 gene[17], case 15 deriving from the yeast URA2 gene[18] and case 39 based on the *E. coli* gene 2282 of unknown function) (Table 1). The total number of possible pairwise interactions can be as many as 122 (column N in Table 1), depending on the degree of paralogy for certain proteins, which introduces some uncertainty in the prediction. Paralogy in the component proteins increases the number of possible interactions (for example, case 45), thereby decreasing the certainty of the prediction. Conversely, paralogy in the composite proteins increases the certainty that the component proteins interact (for example, case 27 (ref. 19)), because the fusion event is repeatedly observed within (or even across) genomes.

Notably, as many predicted interactions occur between products of distant genes as between products of neighbouring genes (Fig. 2): there are only 8 interactions between products of neighbouring genes in *M. jannaschii*, 14 in *H. influenzae* and 18 in *E. coli* (Fig. 2). This underlines the potential of our method to identify interacting proteins in complete genome sequences beyond the simple proximity constraint[10,11]. Some of the interactions between products of neighbouring genes are cases of well-known gene pairs (for example, case 6, the DNA-dependent RNA polymerase A′/A″ pair in *M. jannaschii*), although there are also cases that may be sequencing artefacts (for example, case 19 in Table 1, which is 'intact' in all species except *E. coli*).

The predicted interactions display some complex patterns of distribution along the genome sequences (Fig. 2). For example, certain regions seem to contain a disproportionate number of genes or proteins involved in fusion events (for example, cases 12, 17 and 32 representing the tryptophan biosynthesis gene cluster). Also, some intricate symmetries occur, deriving from multiple-fusion events or paralogous proteins (for example, cases 8 and 61 in *E. coli*, respectively).

Our method identifies a number of well-known interacting partners (for example, cases 4–13). In total, 26 out of the 64 cases

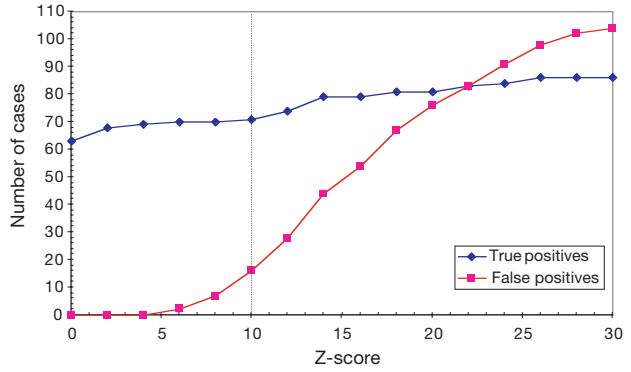

Figure 3 The dependence of the number of true and false positive hits with respect to the Z-score threshold used. Dashed line indicates the threshold used. True positives are 'unrelated' genes that are involved in a fusion event and are candidates for pairwise interactions; false positives are 'related' genes that are also accepted. The decision of whether two genes are related or not (in terms of sequence similarity) depends on the Z-score value: the higher that value is, the more permissive the criterion becomes for two related genes to be considered. True positives are all cases listed in Table 1. False positives were eliminated manually by inspection of the alignment overlap between the two component genes. For Z-scores <10, all cases have been manually and automatically verified. For Z-scores >10, all cases have been only automatically checked; therefore, these values represent an upper bound for true positives.

(40%) listed in Table 1 are involved in the same protein complex or biochemical process (for references, see Supplementary Information). A number of unconfirmed cases (for example, cases 31, 57 and 60, Table 1) constitute some interesting, testable predictions; for example, case 60 represents a predicted interaction between gene products FrdB (fumarate reductase) and GlpC (heterodisulfide reductase) in *E. coli*. We obtained 85 fusion events of which 64 are unique and 21 are false positives (Fig. 3); thus, the precision of the method can be estimated as 75% (64/85) with the current parameter settings (see Methods).

Coverage cannot easily be estimated, as we do not know in advance how many proteins potentially interact within the query genome. We also lack a standard set of fusion events to estimate coverage for the given query genomes. We have tried to estimate coverage by counting the number of false negatives (11, Table 1; see Supplementary Information). Thus, the maximum estimate for the coverage is as high as 95% (215/226) with the current parameter settings and the above assumption. Another false-negative case, fatty acid synthase (and its bacterial homologues)[20], is caused by low sequence similarity relationships. Precision and coverage can be controlled by modifying the cut-off scores in the post-processing of the homology searches (see Methods).

From a total of 7,768 protein sequences in the 3 complete query genomes, the minimum number of components involved in a multidomain-fusion event is 215, or 2.8%. Most of these proteins of known function appear to be metabolic enzymes, an effect possibly due to metabolic channelling of substrates[21]. Our results, together with another study[22], provide the first estimates for the numbers of gene-fusion events based on complete-genome comparisons. By adding more reference genomes, or using the complete database, this number is expected to increase. We are in the process of repeating this analysis to identify fusion events and possible interactions for the genomes of *S. cerevisiae* and *Caenorhabditis elegans*. With sufficient computational power, this analysis could be performed for all complete genomes. The exactness and high efficiency of the method makes it applicable to proteomics research, and complementary to continuing experimental approaches for the identification of protein interactions. We believe that this general approach, if used wisely, will find a number of applications in genomics and computational biology. □

Methods

Genome sequences

The genome sequences (and total number of open reading frames (ORFs)) used were *E. coli* (4290; ftp://genetics.wisc.edu/pub/sequence/m52p.fap.gz); *H. influenzae* (1707; ftp://ftp.tigr.org/pub/data/h_influenzae/GHI.pep.gz); *M. jannaschii* (1771; ftp://ftp.tigr.org/pub/data/m_jannaschii/GMJ.pep.gz); and *S. cerevisiae* (6243; ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/orf_trans.fasta.Z).

Genome comparison

The most reliable prediction of protein–protein interactions is that within complete genomes, and we only considered query databases that have this property. In other words, we did not address the general problem of protein interaction predictions within any arbitrary database. Both query and reference databases are always considered to be complete-genome databases. Generalizations of this formalism are possible and query and reference databases may be interchangeable. The advantage of performing this analysis only within complete genomes is that these data are both comprehensive (no better candidate may be identified) and unbiased (no cases occur in which more fusions will be detected because of experimental sampling). Each of the three complete genomes of *E. coli*, *H. influenzae* and *M. jannaschii* were used in turn as the query genome Q; Q was compared with the other two genomes, plus the genome of the yeast *S. cerevisiae*, as reference genomes.

Sequence analysis

The query database is compared against itself using BLASTP (v 2.0)[23], after masking compositionally biased regions (using the CAST algorithm[24]; and V. Promponas *et al.*, manuscript in preparation), and all pairwise sequence similarities are recorded in a binary matrix T (Fig. 1). The matrix is symmetrified (similarly to that in ref. 25), using a Smith–Waterman[26] dynamic-programming alignment algorithm[27], which is executed only for nonsymmetrical pairs. The query database is also compared against the reference database, as above, and similarities are recorded in binary matrix Y (Fig. 1). For all entries C in the reference database, entry pairs (A,B) from the query database similar to reference entry C are collected (Fig. 1). Every pair (A,B) similar to C is looked up in the self-comparison matrix T; if dissimilar, it is further checked for similarity using a second dynamic-programming alignment pass to eliminate the possibility that it was a false-negative case during the initial self-comparison phase. If not (see Fig. 1 for details), then A and B in the query database (usually a complete genome) are candidates for a fusion event and can be predicted to interact. It is apparent that, although the coverage of the query database may strongly depend on the reference database, the precision can be very high. Unfortunately, no pertinent data set exists to estimate precision. Precision and coverage are controlled by a Z-score parameter setting (Fig. 3).

Received 15 July; accepted 15 September 1999.

1. Mendelsohn, A. R. & Brent, R. Protein interaction methods—toward an endgame. *Science* **284,** 1948–1950 (1999).
2. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17,** 121–127 (1999).
3. Phizicky, E. M. & Fields, S. Protein–protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59,** 94–123 (1995).
4. Luban, J. & Goff, S. P. The yeast two-hybrid system for studying protein–protein interactions. *Curr. Opin. Biotechnol.* **6,** 59–64 (1995).
5. Lecrenier, N., Foury, F. & Goffeau, A. Two-hybrid systematic screening of the yeast proteome. *BioEssays* **20,** 1–5 (1998).
6. Lakey, J. H. & Raggett, E. M. Measuring protein–protein interactions. *Curr. Opin. Struct. Biol.* **8,** 119–123 (1998).
7. Janin, J. & Rodier, F. Protein–protein interaction at crystal contacts. *Proteins* **23,** 580–587 (1995).
8. Jones, S. & Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA* **93,** 13–20 (1996).
9. Larsen, T. A., Olson, A. J. & Goodsell, D. S. Morphology of protein–protein interfaces. *Structure* **6,** 421–427 (1998).
10. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a finger-print of proteins that physically interact. *Trends Biochem. Sci.* **23,** 324–328 (1998).
11. 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).
12. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44,** 66–73 (1997).
13. 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).
14. Bülow, I. Preparation of artificial bifunctional enzymes by gene fusion. *Biochem. Soc. Symp.* **57,** 123–133 (1990).
15. Wales, M. E. & Wild, J. R. Analysis of structure–function relationships by formation of chimeric enzymes produced by gene fusion. *Methods Enzymol.* **202,** 687–706 (1991).
16. Burns, D. M., Horn, V., Paluh, J. & Yanofsky, C. Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains. *J. Biol. Chem.* **265,** 2060–2069 (1990).
17. Duncan, K., Edwards, R. M. & Coggins, J. R. The *Saccharomyces cerevisiae* ARO1 gene. An example of the co-ordinate regulation of five enzymes on a single biosynthetic pathway. *FEBS Lett.* **241,** 83–88 (1988).
18. Denis-Duphil, M. Pyrimidine biosynthesis in *Saccharomyces cerevisiae*: the ura2 cluster gene, its multifunctional enzyme product, and other structural or regulatory genes involved in *de novo* UMP synthesis. *Biochem. Cell Biol.* **67,** 612–631 (1989).
19. Ouzounis, C. A. & Kyrpides, N. C. ThiD-TenA: a gene pair fusion in eukaryotes. *J. Mol. Evol.* **45,** 708–711 (1997).

20. Smith, S. The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. *FASEB J.* **8,** 1248–1259 (1994).
21. Welch, G. R. & Easterby, J. S. Metabolic channeling versus free diffusion: transition-time analysis. *Trends Biochem. Sci.* **19,** 193–197 (1994).
22. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285,** 751–753 (1999).
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
24. Andrade, M. A. *et al.* Automated genome sequence analysis and annotation. *Bioinformatics* **15,** 391–412 (1999).
25. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl Acad. Sci. USA* **95,** 6239–6244 (1998).
26. Smith, T. F. & Waterman, M. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
27. Pearson, W. R. Effective protein sequence comparison. *Methods Enzymol.* **266,** 227–258 (1996).

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

Acknowledgements

We thank M. Carroll and S. Searle for technical advice. This work was supported by the European Molecular Biology Laboratory and the TMR Programme of the European Commission DGXII (Science, Research and Development). Patent application filed on behalf of the European Molecular Biology Laboratory.

Correspondence and requests for materials should be addressed to C.A.O (e-mail: ouzounis@ebi.ac.uk).

Glycoproteins form mixed disulphides with oxidoreductases during folding in living cells

Maurizio Molinari & Ari Helenius

Swiss Federal Institute of Technology Zurich, Institute of Biochemistry, Universitätstrasse 16, CH-8092 Zurich, Switzerland

The formation of intra- and interchain disulphide bonds constitutes an integral part of the maturation of most secretory and membrane-bound proteins in the endoplasmic reticulum[1,2]. Evidence indicates that members of the protein disulphide isomerase (PDI) superfamily are part of the machinery needed for proper oxidation and isomerization of disulphide bonds[3–6]. Models based on *in vitro* studies predict that the formation of mixed disulphide bonds between oxidoreductase and substrate is intermediate in the generation of the native intrachain disulphide bond in the substrate polypeptide[7]. Whether this is how thiol oxidoreductases work inside the endoplasmic reticulum is not clear. Nor has it been established which of the many members of the PDI superfamily interacts directly with newly synthesized substrate proteins, because transient mixed disulphides have never been observed in the mammalian endoplasmic reticulum during oxidative protein folding[7,8]. Here we describe the mechanisms involved in co- and post-translational protein oxidation *in vivo*. We show that the endoplasmic-reticulum-resident oxidoreductases PDI and ERp57 are directly involved in disulphide oxidation and isomerization, and, together with the lectins calnexin and calreticulin, are central in glycoprotein folding in the endoplasmic reticulum of mammalian cells.

To determine whether maturation of proteins in the endoplasmic reticulum (ER) of live cells is accompanied by formation of mixed disulphides containing ER oxidoreductases, the folding of viral glycoproteins E1 and p62 expressed in Semliki Forest virus (SFV)-infected mammalian cells was monitored. Four hours after infection, Chinese hamster ovary (CHO) cells were pulse-labelled for 2 min with ^{35}S-methionine and -cysteine (0.5 mCi per dish). After various chase times, cells were flooded with a membrane-permeable alkylating agent, *N*-ethylmaleimide (NEM, 20 mM in PBS, pH 6.8), to prevent post-lysis oxidation of free cysteines, and to trap mixed disulphides. NEM has been extensively used as a rapidly acting, *in vivo* alkylating agent of cysteines[9], and its efficiency has been confirmed[10]. Similar results were obtained when we solubilized pulse-labelled cells at acidic pH (detergent and 20% formic acid) as an alternative method to block cysteine reactivity. Under these conditions, the SH form, which predominates at low pH, is not capable of attacking existing disulphides[11]. Postnuclear supernatants (PNS) were prepared with nonionic detergent essentially as described[9], and analysed using a sensitive two-dimensional (2D) gel electrophoresis technique, in which the first dimension was run under nonreducing conditions in a 1.5-mm diameter capillary tube. For the second dimension, the gel extruded from the glass capillary was boiled for 10 min in reducing sample buffer and placed on a standard 8% SDS polyacrylamide gel electrophoresis (PAGE) gel[12,13].

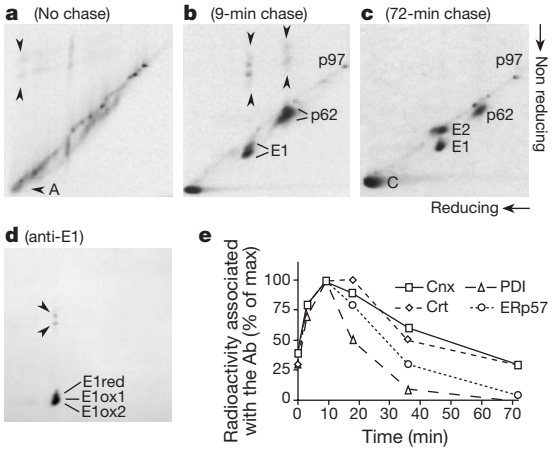

Figure 1 Cotranslational folding of SFV glycoproteins analysed by 2D SDS–PAGE. **a**, Viral growing nascent chains are visible on, above and below the gel diagonal. Mixed disulphides and intramolecular disulphide bonds formed cotranslationally (arrowheads indicate the shortest viral nascent chains engaged in intermolecular disulphide-bonded complexes; arrow A indicates the shortest viral nascent chains acquiring intramolecular disulphide bonds). **b**, **c**, Upon chase, viral proteins are terminated and radioactivity condenses in full-length viral proteins. **d**, Postnuclear supernatants from **b** were precipitated with a monoclonal antibody to E1. E1 in three different oxidation states (E1red, E1ox1 and E1ox2), as well as E1-containing mixed disulphides (arrowheads), are indicated. **e**, Kinetics of viral glycoprotein binding/release from Cnx, Crt, PDI and ERp57. Ab, antibody.

The 2D gels shown in Fig. 1a–c indicate the progression of synthesis, disulphide bond formation and processing of SFV proteins. They allow identification of intermolecular disulphide-bonded complexes, which run above the gel diagonal, and proteins with intrachain disulphides which run below. The simplest pattern is seen in Fig. 1c, after 72 min of chase. Five major radioactive spots are visible. All of these are viral proteins, because SFV induces a complete block in host protein synthesis[14]. One corresponds to the capsid protein (C) and one to p97, a nontranslocated side product of viral glycoprotein synthesis[15]. Capsid and p97 are cytosolic, and their positions on the gel diagonal indicate that they do not contain intrachain disulphide bonds. The three other spots, located below the diagonal, correspond to type I membrane glycoproteins E1 (16 cysteine residues, 1 N-linked glycan), p62 (22 cysteine residues, 4 N-linked glycans) and E2, a cleavage product of p62 formed in the Golgi compartment. E1 and p62 are the only proteins in the infected cells that are translocated into the ER and processed by the ER folding machinery.

After 72 min of chase, folding of SFV glycoproteins was completed and E1, p62 and E2 had acquired a conformation that made