Research

Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions

Anton J Enright and Christos A Ouzounis

Address: Computational Genomics Group, European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK.

Correspondence: Christos A Ouzounis. E-mail: ouzounis@ebi.ac.uk

Published: 28 August 2001

Genome ***Biology*** 2001, **2(9)**:research00341–0034.7

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/research/0034

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 19 March 2001
Revised: 29 June 2001
Accepted: 2 July 2001

Abstract

Background: It has recently been shown that the detection of gene fusion events across genomes can be used for predicting functional associations of proteins, including physical interaction or complex formation. To obtain such predictions we have made an exhaustive search for gene fusion events within 24 available completely sequenced genomes.

Results: Each genome was used as a query against the remaining 23 complete genomes to detect gene fusion events. Using an improved, fully automatic protocol, a total of 7,224 single-domain proteins that are components of gene fusions in other genomes were detected, many of which were identified for the first time. The total number of predicted pairwise functional associations is 39,730 for all genomes. Component pairs were identified by virtue of their similarity to 2,365 multidomain composite proteins. We also show for the first time that gene fusion is a complex evolutionary process with a number of contributory factors, including paralogy, genome size and phylogenetic distance. On average, 9% of genes in a given genome appear to code for single-domain, component proteins predicted to be functionally associated. These proteins are detected by an additional 4% of genes that code for fused, composite proteins.

Conclusions: These results provide an exhaustive set of functionally associated genes and also delineate the power of fusion analysis for the prediction of protein interactions.

Background

Recent progress in genome analysis has shown that it is possible to predict protein interactions or, more generally, functional associations of proteins using genome sequences alone [1-3]. These powerful methods rely on the observation that pairs of genes encoding proteins of known function (usually interacting or forming a complex) tend to be found in other species as a fused gene encoding a single multifunctional protein [4]. This type of event is known as gene fusion and is a well-known process in molecular evolution [5]. Many of these gene fusion events appear to be selectively advantageous by decreasing the regulational load in the cell for a particular process [1,3,5]. Therefore, the detection of gene fusions in one genome (defined as 'composite' proteins) allows the prediction of functional associations between homologous genes that remain separate in another genome (defined as 'component' proteins).

Although gene fusion events appear to be relatively rare, the accurate detection of a gene fusion event in one genome allows interactions to be predicted between many proteins in other genomes. It is this kind of one-to-many relationship that makes this method unique for discovering possible interactions or functional associations between proteins, even for those of unknown function. Unlike previous methods that rely on gene proximity to predict functional

coupling [6], this robust method can also detect distal genes within a genome that may be involved in the same process. Furthermore, we have previously demonstrated [1] the high precision of our algorithm, which with an additional constraint of minimum alignment overlap has now increased to over 86% (see Materials and methods). This family of sequence-based methods is analogous with and complementary to the experimental approaches for the detection of protein interaction [7].

In order to predict functional associations of proteins through the dynamics of gene fusion events, we have applied our algorithm to 24 entire genome sequences that were available from a variety of species (Table 1). We define the genome where we seek component proteins as the 'query' genome and all genomes from which we obtain composite proteins as 'reference' genomes. A 'fusion event' is therefore defined as any pair of component proteins that are detected as a fused, composite protein in a reference genome. For simplicity, we do not attempt to attach directionality to fusion events. In other words, some of these fusion cases (for example, fused in bacteria but split in metazoa) may represent gene 'fission' events.

Table 1

Genomes used in the present analysis

Organism name (strain)	Number of ORFs	ID
Aeropyrum pernix (K1)	2,694	aerpe
Aquifex aeolicus (VF5)	1,522	aquae
Archaeoglobus fulgidus (DSM4304)	2,409	arcfu
Bacillus subtilis (168)	4,100	bacsu
Borrelia burgdorferi (B31) + plasmids	1,639	borbu
Caenorhabditis elegans	19,099	caeel
Chlamydia pneumoniae (CWL029)	1,052	chlpn
Chlamydia trachomatis (serovar D)	894	chltr
Drosophila melanogaster	13,710	drome
Escherichia coli (K12- MG1655)	4,290	escco
Haemophilus influenzae (KW20)	1,707	haein
Helicobacter pylori (26695)	1,577	help2
Helicobacter pylori (J99)	1,495	helpj
Methanococcus jannaschii (DSM 2661)	1,773	metja
Methanobacterium thermoautotrophicum (delta)	1,871	metth
Mycoplasma genitalium (G-37)	479	mycge
Mycoplasma pneumoniae (M129)	677	mycpn
Mycobacterium tuberculosis (H37Rv)	3,924	myctu
Pyrococcus horikoshii (shinkaj OT3)	2,061	pyrho
Rickettsia prowazekii (Madrid E)	837	ricpr
Saccharomyces cerevisiae (S288C)	6,305	sacce
Synechocystis sp. (PCC 6803)	3,168	synsp
Thermotoga maritima (MSB8)	1,849	thema
Treponema pallidum (Nichols)	1,030	trepa

The species/strain names, the number of ORFs and the species name abbreviation used in all figures are given. References for each genome can be found elsewhere [19].

Our algorithm was applied individually for each of the 24 genomes, against the remaining 23 genomes which are used as references (see also Materials and methods). Paralogy in the query genome makes it difficult to determine precisely the actual number of possible associations. As we have previously pointed out, paralogy in the query genome increases uncertainty, while paralogy in the reference genome increases the fidelity of the predictions [1]. It is for this reason that detected component and composite proteins from all genomes are subsequently clustered according to sequence similarity [8]. Each cluster should therefore indicate a distinct family of component or composite proteins. The analysis of the distribution of these gene fusion classes among genomes allows us to investigate the dynamics and distribution of this evolutionary process and to assess the extent of the predictive power of the approach.

Results

The detection of gene fusion events yielded 132,812 component and 66,406 composite proteins in an all-against-all genome comparison, but these values represent multiple occurrences of the same proteins across species. Of these, there are 7,224 component and 2,365 composite unique proteins across the 24 species (a 18- and 28-fold reduction respectively). The multiple detection of these cases within or across genomes signifies that the majority of components and composites are observed more than once and therefore represent genuine cases (as opposed to sequencing artifacts, which are usually isolated cases).

The high precision of the method allows the prediction of 39,730 unique pairwise functional associations of the components with reference to the composite protein set. Eighty-six percent of the 66,406 predicted associations obtained from the total number of composite proteins yield a *Z*-score value of less than 3 (Figure 1), previously shown to result in virtually no false-positive cases [1]. This increased precision is due to the introduction of an additional constraint that does not permit any overlap between the component proteins (see Materials and methods). All the above results are available on the worldwide web (see Materials and methods). Some of these associations are known, but we estimate that more than half of them are newly detected cases, testable by using techniques from functional genomics [7].

Currently, the only species for which predictions can be extensively validated is the yeast *Saccharomyces cerevisiae*, given the ongoing work on transcript profiling [9] and two-hybrid technology [10]. For yeast, there are 440 distinct component cases (predicted by all other genomes as reference,

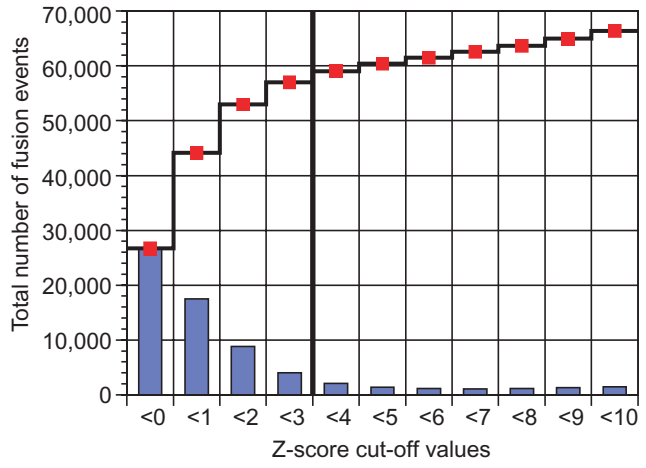

Figure 1
Z-scores for component proteins. The graph illustrates the Z-score (blue bars) distribution and its cumulative sum (step function, with red rectangles) between components, for all detected fusion events (66,406 in total). The Z-score is a statistical measure of similarity for each pair of components. Components that have a Z-score similarity of less than 10, and both exhibit similarity to the same composite protein are detected as fusion events. In general, fusion events where the Z-score between components is less than 3 (marked by a vertical line) result in fewer false-positive fusion detections.

excluding some highly paralogous *Drosophila melanogaster* homologs) involved in 706 predicted interactions, most of which are detected by their homology to composite proteins from *Caenorhabditis elegans* and *D. melanogaster*. Two examples of predicted protein pairs that are known to interact are CPA1 (YOR303w) with CPA2 (YJR109c) [11] and MET3 (YJR010w) with MET14 (YKL001c) [12], both derived from *C. elegans* homologs.

We have attempted to test the validity of our predictions by comparing the set of components to a list of potentially interacting gene products, using results from a large-scale two-hybrid experiment [10]. However, there is only one case shared between the 1,004 proteins involved in 957 putative interactions detected by the two-hybrid system and the complete set of 706 pairs in this analysis: YIL033C (SRA1) and YKL166C (TPK3) matching the *C. elegans* protein C09G4.2 and *D. melanogaster* protein CT10911. This very low count of common pairs may be expected by the sampling biases of the two rather independent methodologies, given that each approach can only detect a very small subset of the total number of actual interacting pairs in yeast. Interestingly, based on a simple conditional probability calculation, an estimate for the total number of detectable interactions in the yeast cell may be of the order of 675,000.

Another validation procedure for the *S. cerevisiae* predictions was obtained by comparing all 706 component pairs against their expression profiles from publicly available gene expression data. We have found that at least 20% of our predictions exhibit very strong correlations across gene expression experiments. For each of the pairs, a profile from 87 experiments involving cell cycle [13], sporulation [14] and diauxic shift [9] was used to determine whether expression data corroborated our predictions for the association of the component proteins (see Materials and methods). The detected pairs of components from fusion analysis clearly exhibit similar patterns of expression for the above mentioned experiments (Figure 2, inset). Despite the noise levels for the gene expression data as a result of the limited number of experimental conditions available, with some random pairs exhibiting significant correlations, there are twice as many predicted associations than random, above the threshold of average correlation value of 0.5. With a higher threshold of 0.55, precision is increased, with four times as many predicted associations over the random background. Above this value, 92 predicted functional associations (20% of 536 available pairs, see Materials and methods) exhibit high correlation across all experiments (Figure 2). Below that threshold, it is very difficult to estimate the precision rate of our predictions, because of the high amount of noise and the rather limited number of publicly available gene expression data sets. This comparison between fusion detection and transcript profiling contrasts with previous approaches [2], where expression data was used as a filtering step for the detection of functional associations, and not as a validation criterion.

We have analyzed the *S. cerevisiae* predictions and detected many interesting cases, which appear to be hitherto

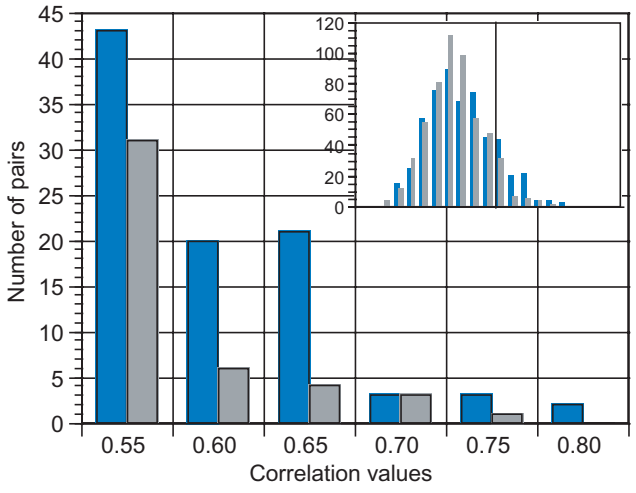

Figure 2
Correlation of gene expression between component pairs. The graph illustrates the distributions of average correlation values of gene expression between component pairs (blue bars) and randomly selected pairs (gray bars), above a threshold value of 0.5. Inset: Distributions of average correlation values for both predicted and random associations (vertical line indicates the cut-off value of 0.5).

undetected functional associations between yeast proteins. Two of these are discussed in some detail here. First, MXR1 (peptide methionine sulfoxide reductase, involved in anti-oxidative processes) [15] and YCL033C (function unknown) are predicted to be functionally associated by virtue of gene fusion in three other species - *Helicobacter pylori* (both strains), *Haemophilus influenzae* and *Treponema pallidum*. This observation is supported by experimental results [16]. MXR1 is 39% identical to the amino terminus of the *H. pylori* composite proteins and YCL033C is 38% identical to the carboxyl terminus of these proteins. It appears that YCL033C is a selenoprotein, also homologous to the human SelX protein, which may be involved in scavenging reactive oxygen species [16]. These two proteins may be associated to protect the yeast cell from oxidative damage.

Second, another interesting observation involves yeast proteins MSS4 (phosphatidylinositol 4-phosphate kinase), which is involved in a signaling pathway responsible for the cell-cycle-dependent organization of actin cytoskeleton [17], and CCT3 (cytoplasmic chaperonin subunit gamma) which is involved in microtubule and actin assembly [18]. A central domain of CCT3 is 25% identical to a large domain of *C. elegans* protein VF11C1L.1 and the carboxy-terminal domain of MSS4 is 29% identical to its carboxyl terminus. Thus, these two proteins are predicted to cooperate in cell-cycle-dependent cytoskeleton organization and assembly.

The distribution of components and composites differs dramatically between species. There are 7,224 component cases, with an average of 350 cases per genome, exhibiting significant variation (Figure 3a, blue bars). The query genome sequences detected 2,365 composite cases, with an average of 115 cases per genome (Figure 3b, blue bars). Interestingly, we have observed some relatively small genomes containing composite proteins, which may yield predictions for components of higher organisms. For instance, there are 71 proteins (forming 30 families) in the *C. elegans* genome that match a fused protein gene in *Mycobacterium tuberculosis*. Two such examples are the component pair T06C10.1/C49H3.7 matching composite Rv0957 and the component pair W04C9.1/Y65B4B_12.b matching both composites Rv1272c and Rv1273c. Another clear prediction is the *S. cerevisiae* component pair YER052c/YJR139c (encoding HOM3/HOM6 respectively) matching composite MetL (3847.PRO) from *Escherichia coli* and other species.

This has been a key observation that dictated the all-against-all genome comparison in this analysis. In other words, when species A is used as a query against species B, the resulting set of component and composite proteins is different from that with the reverse comparison, when species B is used as a query against species A. The three principal factors in gene fusion during evolution appear to be paralogy, genome size and phylogenetic distance. For instance, larger genomes have more composite, possibly paralogous,

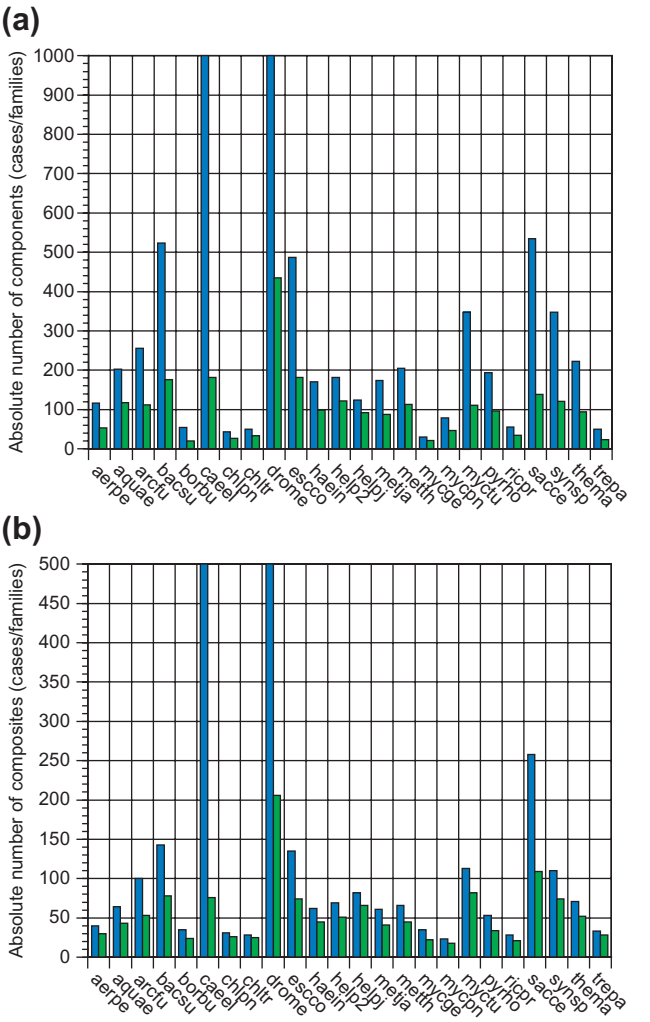

Figure 3
Numbers of component and composite proteins. Absolute number of **(a)** component and **(b)** composite proteins as individual cases (blue bars) and protein families (green bars), by species. Species name abbreviations as in Table 1. Data for *C. elegans* and *D. melanogaster* are clipped (1,973 and 1,981 components, 567 and 559 composites, respectively).

proteins. At the same time, closely related species evidently show similar patterns of gene fusion. The results below address each of these factors in turn and examine their relative contribution to the gene fusion process and their effects on the prediction of functional association of proteins.

For every genome, both sets of component and composite proteins were subsequently clustered [8], to detect the degree of paralogy for these proteins (Figure 3, green bars; see also sequence clustering in Materials and methods). There are 2,534 component families, with an average of 105 families per genome (Figure 3a, green bars) and 1,323 composite families, with an average of 55 families per genome (Figure 3b, green bars). Comparing these numbers with the number of unique cases, it is evident that there is a paralogy

level of two- to threefold per genome for the composite and component proteins, respectively. As mentioned above, this effect contributes to the confidence of the predictions, depending on whether paralogy is observed in the query or the reference genome.

Another characteristic of this process is the redundancy of both sets of component and composite cases: the number of instances of these may be high but they are widely present across species, falling into well-defined protein families. When all components and composites are clustered as a single set (as opposed to within species, above), sequence clustering results in 1,287 single component families and 621 single composite families (as represented in the current analysis for the 24 species). Comparing these numbers with the number of families per species, it is apparent that there is a further twofold reduction for both sets. This result indicates that gene fusion is widespread in evolution but forms a finite set. Different species may contain a common core of composite families, but also provide new families that are used to predict functional association. For instance, *D. melanogaster* provides far more composite families (more than 200) compared to *C. elegans* (fewer than 100) (Figure 3b, green bars). Genomes with unique composite families, such as *D. melanogaster*, contribute strongly to the majority of predicted interactions. It may also be that only certain classes of proteins are involved in gene fusion and that there is an upper limit for the predictive power of this approach obtainable from (currently available) 621 families.

Evidently, the number of component and composite proteins detected in each species is also dependent on genome size (Figure 4). When the above numbers for unique cases and families of components (Figure 4a) and composites (Figure 4b) are normalized by the number of open reading frames (ORFs) for the species examined, the patterns of distribution are significantly altered. For instance, *Aquifex aeolicus* and *Thermotoga maritima* appear to have a large number of components involved in gene fusion (more than 12% of their genes are involved in this process) (Figure 4a), whereas the absolute numbers are low (Figure 3a). This is also the case for composites, where, for example, *S. cerevisiae* yields as many cases as *D. melanogaster* in relative terms (4% of the genome) (Figure 4b), while the absolute counts are dramatically different (Figure 3b).

Finally, when the factors of paralogy and genome size are removed by sequence clustering and normalization, respectively, the effect of phylogenetic distance between species can be detected. A distance measure based on shared composite families has been devised (see Materials and methods) and was used to identify relationships between the 24 species examined. The fact that the tree based on this distance measure (Figure 5) does not significantly contradict other trees based on sequence alignments is a strong indication that our hypotheses about the factors involved in gene

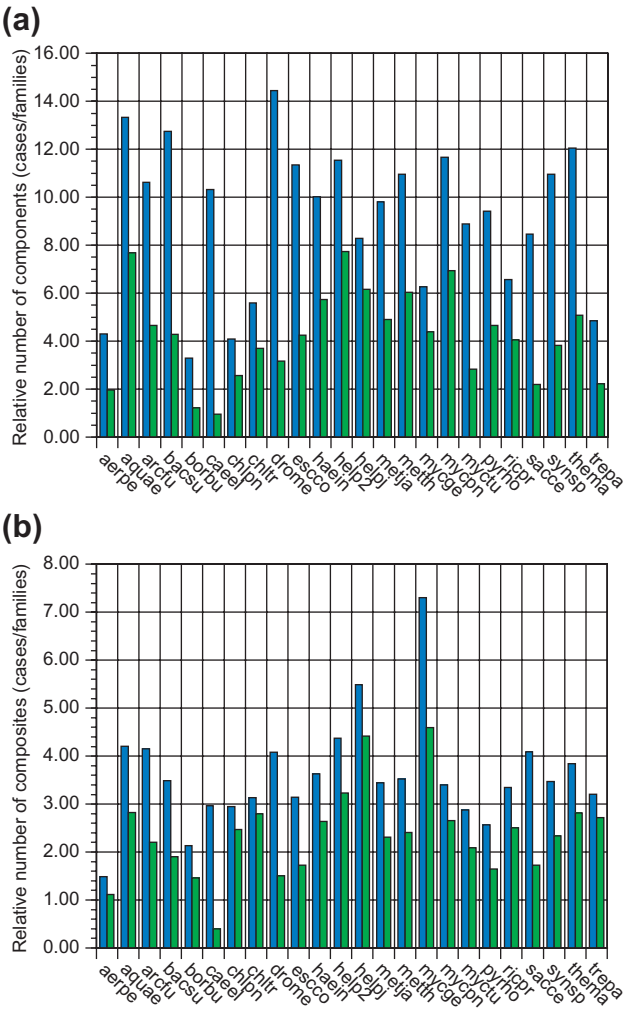

Figure 4
Numbers of component and composite proteins relative to genome size. Relative numbers of **(a)** component and **(b)** composites per species, as individual cases (blue bars) and protein families (green bars), normalized by total genome size (number of ORFs). Species name abbreviations as in Table 1. Average values per genome are 9% for components and 4% for composites.

fusion are valid. This result also indicates that certain types of fusion events appear to be confined to specific phylogenetic groups, such as the Archaea, various bacterial clades and the Eukarya (Figure 5).

Discussion

The exhaustive detection of gene fusion events in entire genome sequences allows the prediction of functionally associated components based merely on genome structure. The all-against-all species comparison is a necessary step because we have repeatedly observed fused, composite proteins in taxonomically lower organisms. The landscape of

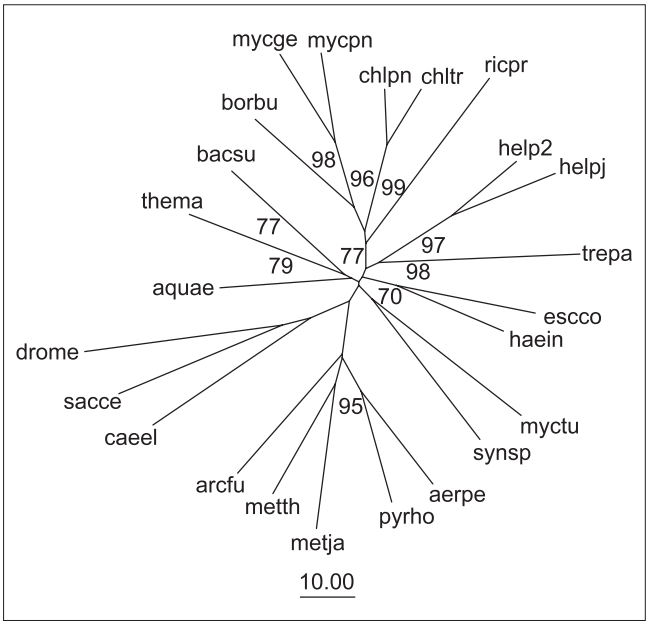

Figure 5
Neighbor-joining dendrogram representing the phylogenetic proximity of each of the 24 species in terms of detected gene fusion events. The distance measure is derived from the count of composite families (see Materials and methods). Scale bar is set to indicate a distance of 10 (ranging from 0 to 100). Species name abbreviations as in Table 1. Only bootstrap values less than 100 are shown.

gene fusions appears to be a complex one, affected by paralogy, genome size and phylogenetic distance.

Although gene fusion is widely present across various phylogenetic groups, it is a process that may involve only certain types of proteins. Yet, this approach for the prediction of functional associations of proteins results in robust predictions for physical interactions, pathway involvement, complex formation and other types of functional associations of protein molecules.

With the present analysis, we delineate the available universe of fusion events and detect a set of 621 composite protein families from which predictions may be obtained. This approach results in 39,730 pairs of functionally associated proteins across 24 species, with high precision and coverage. This novel set of predictions is made available to the scientific community for the first time, and we believe that many of these cases can be subsequently verified by experimental methods.

Materials and methods

Genome sequences

Complete genome sequences for the 24 species were obtained from their original sources [19]. The species names, number of ORFs and the identifiers used throughout this study are listed in Table 1.

Genome comparison

All 24 genomes were filtered using the CAST compositional bias filtering algorithm [20], then compared against themselves and each of the other 23 genomes using the BLASTp [21] sequence similarity searching algorithm with a cut-off E-value of 1 x 10^{-10}. The DifFuse algorithm [1] was then applied automatically to each genome in turn as a query against the other 23 (reference) genomes. Using other protein databases as reference yields fewer composite cases (for example, the well-known case of the TopA/TopB pair appears multiple times in this analysis), showcasing the extreme bias of annotated databases, such as SwissProt (data not shown). Performing the same computation using the non-redundant sequence database (nrdb) is prohibitively expensive in terms of computation time for an analysis of this size. The detected gene fusion results for each of the 552 comparisons were further automatically filtered for significant overlap of the BLAST alignments of the component proteins. In this case, component proteins that overlap by more than 10% of their total length when aligned together with the composite protein. This step avoids the detection of 'promiscuous domains' [3] and gene prediction errors, which result in false-positive fusion detection cases. The detected component and composite proteins are far fewer in number than for the two previous reports of *E. coli* [3] and *S. cerevisiae* [2], due to the much stricter criteria employed in the present analysis and the multi-step protocol we have developed. This analysis was fully automatic and carried out in parallel over a period of four weeks on 11 SUN UltraSPARC CPUs running Solaris 7.

Expression profile analysis

Gene expression ratios for all experiments were transformed into log-odds values so that induction and repression measurements are directly comparable (positive and negative values, respectively). The log-odd values were then normalized across all timepoints for each experiment, using Z-score values for each timepoint. The Z-score values for all time points of each experiment thus allow cross-comparison of gene expression across separate experiments [9].

Our predicted functional associations for *S. cerevisiae* with available expression data represent 536 component pairs in total. For each pair of proteins, a Pearson correlation coefficient was calculated between two corresponding experiments and averaged over all experiments. To estimate noise in these data, a control set of 536 randomly selected *S. cerevisiae* proteins was taken and treated as above (Figure 2).

The distribution of averaged Pearson correlation coefficients for the predicted functional associations was compared against the distribution of coefficients for the control set using a *t*-test for mean values (where the null hypothesis is that the two means are equal). The test results in a *t*-value of 3.6 (critical *t*-value is 1.64), which is highly significant (*P*-value is 0.000173), indicating that there is a higher

average correlation of expression profiles for the predicted functional associations against the background.

Sequence clustering

All proteins involved in gene fusion events as either component or composite proteins were identified automatically from the results of the fusion analysis. From these data we can obtain raw counts of the number of gene fusion events detected and the number of proteins involved in these events as either composite or component proteins. These figures are skewed however, due to the presence of homology in both the query and reference sets. Proteins involved in gene fusion events as either component or composite genes are then assembled into two lists. These lists are then used to generate two sequence databases, the first one containing all component sequences from the 24 genomes and the second containing all composite sequences.

These sequence databases of component and composite proteins are then compared against themselves using the BLASTp (version 2.0) sequence similarity searching algorithm [21] (cut-off E-value 1 x 10^{-10}), then clustered according to their similarity using the RAGE algorithm [8]. The RAGE algorithm lists all composite and component proteins in clusters according to similarity and domain structure. Homologous proteins with similar domain structure were clustered together. Each cluster in this case indicates a distinct class of fusion event and cluster members indicate which proteins involved in this type of event from different genomes. These clusters are used to calculate the number of unique fusions detected within and across genomes. This is done by examining how many distinct types of fusion are present in any given genome.

Dendrogram calculation

All composite proteins were clustered into 621 families and a distance measure δ was derived according to the sharing of clusters between the 24 species examined. This pairwise distance measure is calculated as $\delta = (1 - S_{A,B}/T_{A,B}) \times 100$, where $S_{A,B}$ is the number of shared composite clusters and $T_{A,B}$ is the average of the composite cluster counts from the two species. This measure is reminiscent of a recent genome-wide "ortholog" analysis [22]. This measure was used to calculate a nearest-neighbor dendrogram for the 24 species. Bootstrap values were generated using a 'delete-half' jack-knife procedure.

Data availability

All results of the present analysis are available from the Computational Genomics Group website [23].

Acknowledgements

We thank John Aach (Harvard Medical School), Despina Alexandraki (University of Crete and IMBB, Heraklion) and members of the Computational Genomics Group at the EBI for discussions. This work was fully supported by the European Molecular Biology Laboratory (EMBL). C.O. acknowledges further support from the European Commission DGXII (Science, Research and Development), the Medical Research Council (UK) and IBM Research. Patent application filed on behalf of EMBL.

References

1. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
2. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402:**83-86.
3. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
4. Sali A: **Functional links between proteins.** *Nature* 1999, **402:**23-26.
5. Doolittle RF: **Do you dig my groove?** *Nat Genet* 1999, **23:**6-8.
6. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
7. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.** *Proc Natl Acad Sci USA* 2000, **97:**1143-1147.
8. Enright AJ, Ouzounis CA: **GeneRAGE: a robust algorithm for sequence clustering and domain detection.** *Bioinformatics* 2000, **16:**451-457.
9. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
10. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**623-627.
11. Lim AL, Powers-Lee SG: **Requirement for the carboxyl-terminal domain of *Saccharomyces cerevisiae* carbamoyl-phosphate synthetase.** *J Biol Chem* 1996, **271:**11400-11409.
12. Blaiseau PL, Isnard AD, Surdin-Kerjan Y, Thomas D: **Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism.** *Mol Cell Biol* 1997, **17:**3640-3648.
13. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
14. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast [Erratum: Science 1998, 282:1421].** *Science* 1998, **282:**699-705.
15. Moskovitz J, Berlett BS, Poston JM, Stadtman ER: **The yeast peptide-methionine sulfoxide reductase functions as an antioxidant in vivo.** *Proc Natl Acad Sci USA* 1997, **94:**9585-9589.
16. Lescure A, Gautheret D, Carbon P, Krol A: **Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif.** *J Biol Chem* 1999, **274:**38147-38154.
17. Helliwell SB, Howald I, Barbet N, Hall MN: **TOR2 is part of two related signaling pathways coordinating cell growth in *Saccharomyces cerevisiae*.** *Genetics* 1998, **148:**99-112.
18. Stoldt V, Rademacher F, Kehren V, Ernst JF, Pearce DA, Sherman F: **The Cct eukaryotic chaperonin subunits of *Saccharomyces cerevisiae* and other yeasts.** *Yeast* 1996, **12:**523-529.
19. Kyrpides NC: **Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide.** *Bioinformatics* 1999, **15:**773-774.
20. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA: **CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts.** *Bioinformatics* 2000, **16:**915-922.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
22. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nature Genet* 1999, **21:** 108-110.
23. **Computational Genomics Group** [http://www.ebi.ac.uk/research/cgg/diffuse]