Predicting functional linkages from gene fusions with confidence

Cynthia J Verjovsky Marcotte[1] and Edward M Marcotte[2]

[1]Department of Mathematics, St Edwards University, Austin, Texas, USA; [2]Department of Chemistry and Biochemistry, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA

Abstract: Pairs of genes that function together in a pathway or cellular system can sometimes be found fused together in another organism as a Rosetta Stone protein – a fusion protein whose separate domains are homologous to the two functionally-related proteins. The finding of such a Rosetta Stone protein allows the prediction of a functional linkage between the component proteins. The significance of these deduced functional linkages, however, varies depending on the prevalence of each of the two domains. Here, we develop a statistical measure for the significance of predicted functional linkages, and test this measure for proteins of *E. coli* on a functional benchmark based on the KEGG database. By applying this statistical measure, proteins can be linked with over 70% accuracy. Using the Rosetta Stone method and this scoring scheme, we find all significant functional linkages for proteins of *E. coli*, *P. horikshii* and *S. cerevisiae*, and measure the extent of the resulting protein networks.

Keywords: functional genomics, networks, protein function, gene fusions

Introduction

The genome sequencing revolution has led to the discovery of literally thousands of previously unknown genes, many broadly conserved across multiple species but of entirely uncharacterised function. The field of functional genomics has emerged to discover the functions of many of these uncharacterised genes, and to determine how the genes work together in systems and pathways. To address these problems, computational methods known as non-homology methods (Marcotte 2000; Eisenberg et al 2000; Huynen et al 2000a) have been developed to reconstruct gene pathways and protein interactions. Using non-homology methods, ‘functional linkages’ are predicted between genes operating in the same pathways by matching the genes that are found in similar contexts. For example, pathways can be reconstructed by identifying genes occurring in the same operon (Dandekar et al 1998; Overbeek et al 1999; Salgado et al 2000), finding genes present and absent from the same sets of organisms (Pellegrini et al 1999), or finding pairs of genes fused together as a single gene in another organism (Marcotte et al 1999a; Enright et al 1999; Huynen et al 2000b). It is critical that all predictions are accompanied by appropriate statistical analyses surrounding the significance of the results. Here, we address the problem of establishing and testing the significance of functional linkages derived from the gene fusion method.

The essence of the gene fusion method is illustrated in Figure 1. Two separate genes in one organism, such as the *E. coli* L-tartrate dehydrogenase A and B subunit proteins at

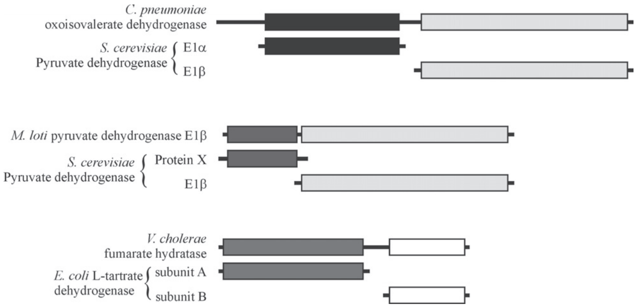


Figure 1 Three examples of the Rosetta Stone gene fusion analysis. In each example, the functional linkage between a pair of proteins is inferred from the occurrence of a third ‘Rosetta Stone’ protein, drawn directly above the pair, which is composed of component sequences similar to each of the separated proteins. The top example shows two yeast proteins, the E1 alpha component (Pda1) and E1 beta component (Pdb1) of the pyruvate dehydrogenase complex. Pda1 and Pdb1 are correctly predicted to work together because of their sequence homology to distinct regions of a third protein, the oxoisovalerate dehydrogenase from *C. pneumoniae*. In the second example, the E1 beta component is additionally linked to the protein X component (Pdx1) of the mitochondrial pyruvate dehydrogenase. In the third example, two *E. coli* proteins, subunits A and B of the L-tartrate dehydrogenase, are correctly linked via the presence of the *V. cholerae* Rosetta Stone protein fumarate hydratase.

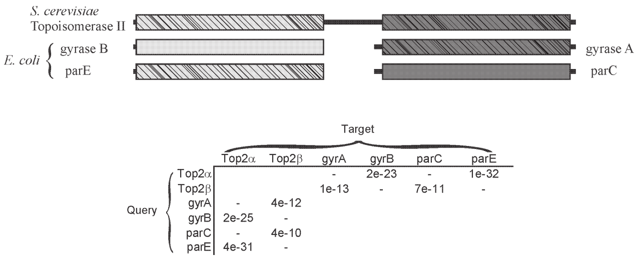
Correspondence: EM Marcotte, Institute for Cellular and Molecular Biology, MBB3.23, 2500 Speedway, University of Texas at Austin, Austin, TX 78712, USA; tel +1 512 471 5435;
fax +1 512 232 3432; email marcotte@icmb.utexas.edu

the bottom of Figure 1, can occasionally also be found fused into a single protein, such as the *Vibrio cholerae* protein drawn above the two *E. coli* proteins. The presence of this fusion protein allows us to infer a functional linkage between the two component proteins (Marcotte et al 1999a; Enright et al 1999). The fusion protein is termed a 'Rosetta Stone protein' for decoding this relationship (Marcotte et al 1999a). Pairs of proteins linked by such a Rosetta Stone relationship often occur in the same cellular pathway (Marcotte et al 1999b; Yanai et al 2001).

It is clear that not all fusion events are equally valuable for inferring functional linkages between genes. For example, certain 'promiscuous domains' (Marcotte et al 1999a), such as ATP binding cassettes and SH3 domains, are known to participate in fusions with more than a hundred other domains. It would be essentially uninformative to link every protein containing an SH3 to every protein containing a kinase domain due to the presence of fusion proteins with both kinase and SH3 domains. For this method to become a widely applicable functional genomics technique, we therefore require a way of identifying only high quality Rosetta Stone linkages and avoiding the promiscuous domain-induced linkages.

The limitations of orthology

Intuitively, one might expect that accurate predictions of Rosetta Stone linkages could be found by considering only orthologous relationships. However, the applicability of orthology for this purpose is limited. An example, diagrammed in Figure 2, will serve to illustrate this point. In this example, the gyrase A and B, and parE and C topoisomerase proteins, are linked due to the presence of the yeast topoisomeraseII (Top2) protein, which is a fusion of both gyrase A-like and gyrase B-like domains. To apply orthology to predict precise linkages between these proteins, the fusion protein must first be separated into domains, and the sequence of each domain must be compared to the separate proteins to identify orthologues.



		Target					
		Top2α	Top2β	gyrA	gyrB	parC	parE
Query	Top2α			-	2e-23	-	1e-32
	Top2β			1e-13	-	7e-11	-
	gyrA	-	4e-12				
	gyrB	2e-25	-				
	parC	-	4e-10				
	parE	4e-31	-				

Figure 2 An example of the limits of orthology for discovering Rosetta Stone fusion relationships. Yeast topoisomeraseII (Top2) is a protein consisting of a gyrB/parE-type domain fused to a gyrA/parC-type domain. As defined by the bidirectional best hit criterion, parE is the orthologue of the first domain of yeast Top2, and gyrase A is the orthologue of the second domain of yeast Top2, as indicated in the figure by cross-hatching. The precise BLASTP expectation scores obtained from sequence comparisons are shown in the table. Thus, the orthology-based method incorrectly pairs parE with gyrA, and omits the correct linkages of gyrA with gyrB, and parE with parC. Using homology rather than orthology would ensure both correct linkages are discovered, but also introduces two incorrect linkages. Therefore, homology increases the coverage, while potentially decreasing the accuracy, while orthology decreases the coverage without guaranteeing enhanced accuracy.

A common heuristic method for identifying orthologues is to identify 'bidirectional best hits' (for example, see Overbeek et al 1999): if the most similar sequence to protein A in genome 2 is B, and if the most similar sequence to protein B in genome 1 is A, then A and B are bidirectional best hits, and are operationally considered to be orthologues. In the case of Figure 2, parE is the bidirectional best hit of the first domain of Top2 (Top2α), and gyrase A is the bidirectional best hit of the second domain of Top2 (Top2β). Therefore, this approach would incorrectly predict a functional linkage between parE and gyrase A, and would miss the correct linkages between gyrase A and B, and between parE and C. Although it may be argued that the bidirectional best hit method is inadequate, alternative high-throughput methods for establishing orthology, such as the COGS database (Tatusov et al 1997), are not useful here either—both gyrase B and parE belong to the same COG family, as do gyrase A and parC.

Beyond difficulties in establishing orthologues in a scaleable fashion, the orthology-based method suffers from an additional shortcoming. Even in a situation where the fused orthologues are the correct partners, the method predicts only one linkage per fusion protein, and makes no prediction about the remaining proteins, even in simple cases such as that in Figure 2. For example, even if gyrase A and B had actually been the orthologues of Top2β and α, respectively, the method would still omit the linkage between parC and parE. Thus, the orthology-based method: (1) suffers from considerably lower coverage than the homology-based method, and (2) does not guarantee accurate linkages. Instead, we have opted to use homology, but provide a quantitative measure of confidence in each prediction. In this manner, the method's coverage is maximised, and the predictions can be easily combined with functional linkages derived by other methods.

Measuring the significance of Rosetta Stone linkages

To account for the promiscuous-domain induced ambiguities in Rosetta Stone linkages, we have developed a statistical measure that takes into account the prevalence of the two

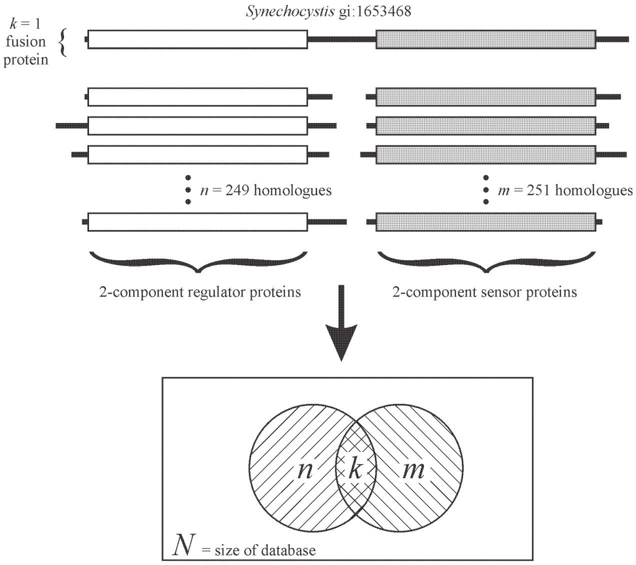


Figure 3 An example of promiscuous domains generating uninformative Rosetta Stone linkages. Each of the two component regulator proteins is linked with each of the two component sensor proteins because of the presence in the sequence database of a single Rosetta Stone protein from *Synechocystis* (GenBank identifier 1653468) that encodes fused regulator and sensor domains. As there are a large number of sensor and regulator proteins in the database (as well as in any given organism), it is uninformative to link every sensor to every regulator protein, and because only a single fusion is observed, one's confidence in the fusion may be low. To calculate the significance of such a case, the number n of regulator sequence homologues in the database of N sequences is measured, the number m of sensor sequence homologues in the database is measured, and the number k of sensor/regulator fusion proteins is measured. Then, the probability of observing k or more fusions by chance due to the prevalence of the two domains is calculated. As illustrated in the Venn diagram at the bottom of the figure, this is equivalent to the probability of drawing a set of n objects from a bag containing N, replacing the n objects, then drawing m from the bag and observing k that were also seen on the first draw. Here, N = 145 579; n = 249; m = 251; and k = 1, and p(Number of fusions $\geq k \mid n = 249, m = 251, N = 145579) = 0.35$. This probability is multiplied by a correction term accounting for the within-genome ambiguity arising from the large number of homologues of sensors (S_A) and regulators (S_B) in *E. coli*. For this example, this probability term equals $(1\text{-}1/\max(S_A,S_B)) = (1\text{-}1/23) = 0.96$. The product of these probabilities (0.96 x 0.35 = 0.34) is the probability of this fusion occurring by random chance. In this example, a probability of ~1/3 is quite high, so functional linkages generated between the sensor and regulator proteins could easily be achieved by random chance and are not deemed significant.

separated proteins and favors predictions made between otherwise rare domains. The approach is illustrated in Figure 3.

Figure 3 shows an example of promiscuous domains in bacterial genomes, the two-component signalling protein sensor and regulator domains. Again, it is clear that although the domains have a close functional relationship, it is uninformative to link every sensor to every regulator protein due to the presence of the *Synechocystis* fusion protein. There is an ambiguity as to which of the sensors should be linked to which of the regulators.

Two proteins, A and B, are predicted to be functionally linked due to the presence of a third AB fusion protein, where the sequences A and B are similar to non-overlapping regions of AB and show no significant similarity to each other. The number of similar sequences of each of these proteins is counted in a database of N protein sequences, where a sequence is considered similar if its amino acid sequence aligns with the query protein with a statistically significant BLASTP expectation score ($E < 1 \times 10^{-6}$; Altschul et al 1997) using default BLASTP 2.1.2 parameters. If n represents the number of A homologues in the database, m represents the number of B sequence homologues and k represents the number of distinct AB fusion proteins linking A and B, then based on the hypergeometric distribution we can calculate the probability of observing exactly k fusions between domains occurring n and m times in the sequence database:

$$p(k \mid n,m,N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}} = \frac{n!(N-n)!m!(N-m)!}{(n-k)!k!(m-k)!(N-n-m+k)!N!} \qquad (1)$$

This probability corresponds to the number of unique ways that n and m proteins could be chosen with exactly k fusions, divided by the total number of unique ways that n and m proteins could be chosen from the database. In practice, to avoid underflow errors in computing the probability, the log of the probability is calculated:

$$\begin{aligned}\ln p(k \mid n,m,N) = {} & \ln(n!) + \ln(N-n)! + \ln(m!) + \ln(N-m)! \\ & - \ln(n-k)! - \ln(k!) - \ln(m-k)! \\ & - \ln(N-n-m+k)! - \ln(N!)\end{aligned} \qquad (2)$$

and large factorials are calculated with Sterling's approximation. To convert this instantaneous probability into the probability of observing k or more fusions by random chance, we subtract from one the probability of observing from 0 to k-1 fusions:

$$p(\text{Number of fusions} \geq k \mid n,m,N) = 1 - \sum_{i=0}^{k-1} p(i \mid n,m,N) \qquad (3)$$

where i is merely a counter for the summation. This score now gives us a measure of the significance of the functional linkage between two proteins that are linked by a fusion event. The lower the probability, the less likely it is that the fusions were observed by chance, and the more significant the linkage is.

Intuitively, this score captures the properties we desired: it favours fusions between rare domains (small n and m) over those between common domains (large n and m), unless those proteins are virtually always observed as fusions (large k), in which case the rare separated domains are predicted to be linked. However, this score fails to account for a second source of ambiguity: the *within-genome* ambiguity introduced by multiple homologues of the A or B proteins within the query genome for which we are predicting functional linkages. For example, given an A protein in *E. coli* and a B protein in *E. coli* (with fusion proteins in other genomes), it is reasonable to link A to B. Given two A proteins and two B proteins in *E. coli*, four possible links can be made (A1–B1, A1–B2, A2–B1, A2–B2), only two of which are likely to be correct, as in the example in Figure 2. So, given S_A homologues of A and S_B homologues of B in the query genome, the probability that any given linkage between an A protein and a B protein is correct is given by:

$$p(\text{A, B are functionally linked}) = \frac{\min(S_A, S_B)}{S_A \times S_B} = \frac{1}{\max(S_A, S_B)} \qquad (4)$$

This probability term will be equal to 1 if there is only one A and one B protein in the query genome. Thus, all unambiguous linkages are automatically included. By assuming independence between this within-genome ambiguity and the whole-database ambiguity discussed earlier, the final significance score is merely the product of the probabilities of the two events:

$$\begin{aligned} &p(\text{A, B linked by random chance}) \\ &= p(\text{A, B are not functionally linked}) \\ &\qquad \times p(\text{Number of fusions} \geq k \mid n, m, N) \\ &= \left(1 - \frac{1}{\max(S_A, S_B)}\right) \times \left(1 - \sum_{i=0}^{k-1} p(i \mid n, m, N)\right) \end{aligned} \qquad (5)$$

The assumption of independence between these two effects is certainly incorrect with rare genes that occur only in the query genome; but for most genes, this assumption holds, and will only become stronger as the number of genome sequences grows.

Construction of a benchmark for testing functional prediction

Having established a scoring scheme, it is necessary to test this measure to ensure that it correctly predicts the quality of the functional linkages. Ideally, we expect the functional relatedness of linked proteins to increase as the probability of observing the fusion by chance decreases. To test this functional similarity, a benchmark will be required consisting of a set of proteins known to operate in a set of cellular pathways or systems.

Such benchmarks have previously utilised the keywords associated with proteins in sequence databases such as SwissProt (Bairoch and Apweiler 2000) or the hierarchical functional categorisations of proteins, such as the KEGG database (Kanehisa and Goto 2000) or MIPS database (Mewes et al 2002), in which curators manually assign proteins into functional categories. The SwissProt database keywords, while convenient and available for many proteins, are limited in their usefulness for this purpose (Devos and Valencia 2000), and are often only indirectly related to the functions of the proteins, referring instead to motifs or subcellular locations.

However, the KEGG database (http://www.genome.ad.jp/KEGG) organises proteins into pathways and cellular systems, and this functional hierarchy seems intrinsically well-suited for serving as a functional benchmark. The database categorises 1283 *E. coli* proteins into 24 main pathways, including metabolic pathways (such as carbohydrate and energy metabolism), and general cellular systems (such as transcription, translation, and sorting and degradation). Within each main category, proteins are further grouped into subcategories; 147 of these specific categories are described, of which 117 apply to *E. coli*. Example KEGG categories are shown in Table 1.

Unfortunately, the KEGG functional categories also suffer from certain limitations. Primarily, they are not defined according to objective criteria—some categories represent not pathways, but homologous protein families (such as two component signalling proteins or ATP binding cassette systems). For the purposes of a functional benchmark, we are willing to accept a broad definition of pathways encompassing metabolic pathways, signal transduction cascades and cellular systems. However, it would be incorrect to suggest that all two-component sensor proteins operate in the same specific pathway. Some processing must therefore be done to select a consistent set of functional categories from the KEGG database for use as a functional benchmark.

Each KEGG category in KEGG release 22.0 was evaluated for its suitability to describe pathway relationships rather than homologous protein families. First, the set of *E. coli* proteins were selected in a given KEGG category (from the set of 117 specific categories). The amino acid sequences of these proteins were compared to each other using BLASTP 2.1.2 to identify proteins with related sequences. The proteins from the KEGG category were then modelled as a graph, with each protein represented by a vertex in the graph, and each significant sequence similarity

Table 1 An example of the functional annotation of 3 pairs of linked proteins from *E. coli*. The KEGG database (Kanehisa and Goto 2000) annotation for each protein is listed. In this example, the functional similarity between the linked proteins ttdA and ttdB would be 100%, but the functional similarity between the unlinked proteins ttdA and phoB would be 0%. The functional similarity between phoB and envZ, linked with a poor significance score and known to act in different specific pathways (phosphate starvation response and osmolarity sensing, respectively), would be scored as 0% due to removal of the 'two-component system' KEGG category during construction of the functional benchmark.

Protein	*General KEGG category*	*Specific KEGG category*	*Link significance score*	*Functional similarity*
ttdA L-tartrate dehydratase subunit A	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism	$p = 0$	100%
ttdB L-tartrate dehydratase subunit B	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism		
fucOoxidoreductase	Carbohydrate metabolism	Pyruvate metabolism		
		Glyoxylate and dicarboxylate metabolism	$p = 0$	50%
aldA aldehyde dehydrogenase	Carbohydrate metabolism	Pyruvate metabolism		
phoB regulator	Signal transduction	Two-component system	$p = 0.06$	0%
envZ sensor	Signal transduction	Two-component system		

(BLASTP expectation score < $1{\times}10^{-6}$) represented by an edge in the graph connecting the similar proteins. The largest family of homologous proteins was identified as the largest cluster in this graph. The number of proteins in this family serves to indicate the nature of the KEGG category.

As shown in Figure 4, more than 95% of the KEGG categories describe pathways composed of diverse proteins sharing little sequence similarity to each other. However, a small number of KEGG categories were composed not of pathway members, but of homologous sequence families, and are unsuitable for use as a functional benchmark. Categories containing more than 10 proteins from a single protein family (KEGG categories 2020 and 2010) were removed, and the functional benchmark was composed of the remaining KEGG categories.

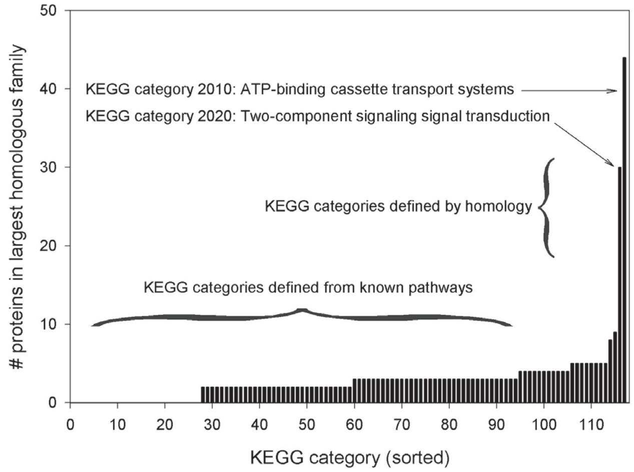


Figure 4 Construction of a functional prediction benchmark. The number of *E. coli* proteins from the single largest protein family in a given KEGG category is plotted. More than 95% of the KEGG categories can be seen to reflect pathway relationships, but a small number of KEGG categories, primarily categories 2010 and 2020, are defined instead by homology. When these categories are removed, the remaining categories provide a functional benchmark for the proteins of *E. coli*.

Testing the linkage scoring scheme

Given the functional benchmark, the scoring scheme could now be tested. We identified all possible Rosetta Stone links between proteins of the bacterium *E. coli* in the following manner. First, we constructed a database consisting of 145 579 protein sequences from the completed genomes of 50 organisms, including the genome of every organism discussed in this article. Using the program BLASTP 2.1.2 and default search parameters, we then compared the sequence of every *E. coli* protein to every protein sequence in this database. Rosetta Stone linkages were defined between pairs of *E. coli* proteins where the proteins showed no sequence similarity to each other (using a BLASTP expectation value threshold of $1{\times}10^{-4}$), but were each homologous to non-overlapping segments of a third protein from the database. Of the 4289 proteins in the *E. coli* genome (Blattner et al 1997), we found 4613 functional linkages between 1124 proteins of *E. coli*, corresponding to predictions of functionally linked proteins for about 26% of the genome of *E. coli*.

To then quantitatively measure protein function, we examined the subset of *E. coli* proteins of known function, as defined in our functional benchmark. For all Rosetta Stone linked pairs of *E. coli* proteins assigned to functional categories in the benchmark, we measured the extent of functional similarity as the Jaccard coefficient of their functional category annotations:

$$\text{Functional similarity} = 100 \times \frac{|\mathrm{KEGG_A} \cap \mathrm{KEGG_B}|}{|\mathrm{KEGG_A} \cup \mathrm{KEGG_B}|} \quad (6)$$

where KEGG_x is the set of specific KEGG categories in which protein x is known to participate, and $|\mathrm{KEGG}_x|$ is the size of the set. This measure represents the normalised intersection of the set of specific categories to which proteins A and B belong, and results in a similarity of 100% for proteins occurring in the identical set of categories and 0% for proteins in entirely distinct categories. We calculated the functional similarity for all pairs of linked annotated proteins, and binned the similarities according to the significance scores of the linkages, plotting the mean of the functional similarity for links within a given range of significance scores. The results of this genome-wide test are shown in Figure 5.

Several results are notable from this test. Primarily, the significance scores described earlier correctly predict the functional relatedness of the linked proteins. The pathway similarity of the two proteins shows an exponential dependence upon the linkage probability score, with Rosetta Stone linked protein pairs with significance scores equal to $1{\times}10^{-6}$ corresponding to ~50% functional similarity, and pairs linked with significance scores equal to $1{\times}10^{-10}$ corresponding to ~70% functional similarity. A linear regression fit of the data (R = 0.88) provides the following relationship:

$$S = 12.7 - 5.5 \times \log p(\text{A, B linked by random chance}) \quad (7)$$

where S is the functional similarity, and p(A, B linked by random chance) is the significance score. The linkage scoring scheme therefore establishes an objective measure of how much confidence to place in each functional linkage.

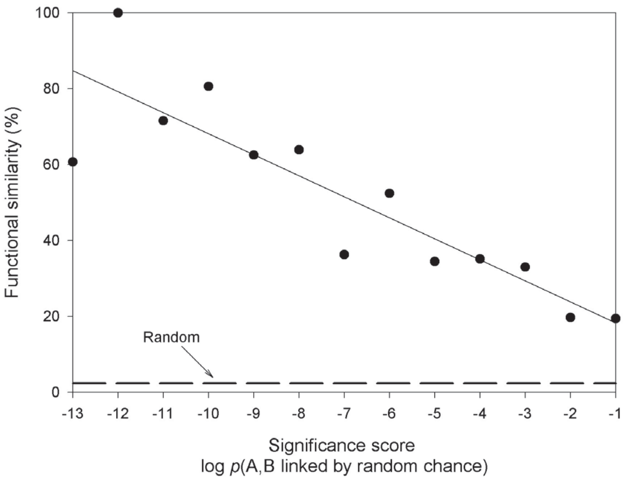


Figure 5 The statistical measure of confidence in Rosetta Stone linkages, log *p*(A,B fused by random chance), correlates well with known functional relationships among proteins of *E. coli*, as shown by the closed circles and their fit by linear regression (solid line). All Rosetta Stone linkages, even those with poor significance scores, associate genes together whose functions are considerably more similar than random pairs of *E. coli* proteins (dashed horizontal line).

Calculating networks of functionally linked proteins

To measure the extent of high confidence fusions, we examined the number of functional linkages predicted at different significance scores in several different genomes. Table 2 summarises the total number of links and the number of links with significance scores below $1{\times}10^{-6}$ and $1{\times}10^{-10}$, corresponding to the ~50% and ~70% functional similarity cutoffs, in the bacterium *E. coli*, the archaeon *Pyrococcus horikoshii* (Kawarabayasi et al 1998) and the eukaryote *Saccharomyces cerevisiae* (Goffeau et al 1997). As seen in Table 2, hundreds of very high confidence predictions can be generated for proteins in a genome. Specifically, links at the ~50% functional similarity level can be predicted for more than 14% of the proteins of *E. coli*, 10% of the proteins of yeast and 8% of the proteins of the archaeon *P. horikoshii*.

Because these linkages join proteins into networks, it is reasonable to try and visualise these resulting networks. Figure 6 shows the network of 475 *E. coli* proteins connected by 854 Rosetta Stone linkages with significance scores better than $1{\times}10^{-10}$. The networks generated only by Rosetta Stone linkages are sparse, but begin to define cellular systems, many of which are labeled in Figure 6, such as the system of pyruvate metabolism and aromatic amino acid biosynthesis, and the fructose and fructose-like phosphotransferase system.

Metabolic proteins are known to be common participants in gene fusions (Tsoka and Ouzounis 2000), and indeed many of the proteins pictured in Figure 6 work in metabolic pathways, such as aroA, aroB and aroD in shikimate

Table 2 The extent of high confidence functional linkages found in complete genomes. The numbers of links predicted with significance scores below a given threshold are measured for all proteins from organisms from the three domains of life. The percentage of proteins in the genome for which some functional link can be predicted is shown in parentheses in the last column.

Organism	*Significance score threshold*	*Number of functional links*	*Number of proteins (% of genome)*
E. coli	1 (all links)	4613	1124 (26%)
	1×10^{-6}	1121	583 (14%)
	1×10^{-10}	854	475 (11%)
P. horikoshii	1 (all links)	653	384 (19%)
	1×10^{-6}	135	165 (8%)
	1×10^{-10}	107	137 (7%)
S. cerevisiae	1 (all links)	9382	1547 (24%)
	1×10^{-6}	1473	611 (10%)
	1×10^{-10}	918	406 (6%)

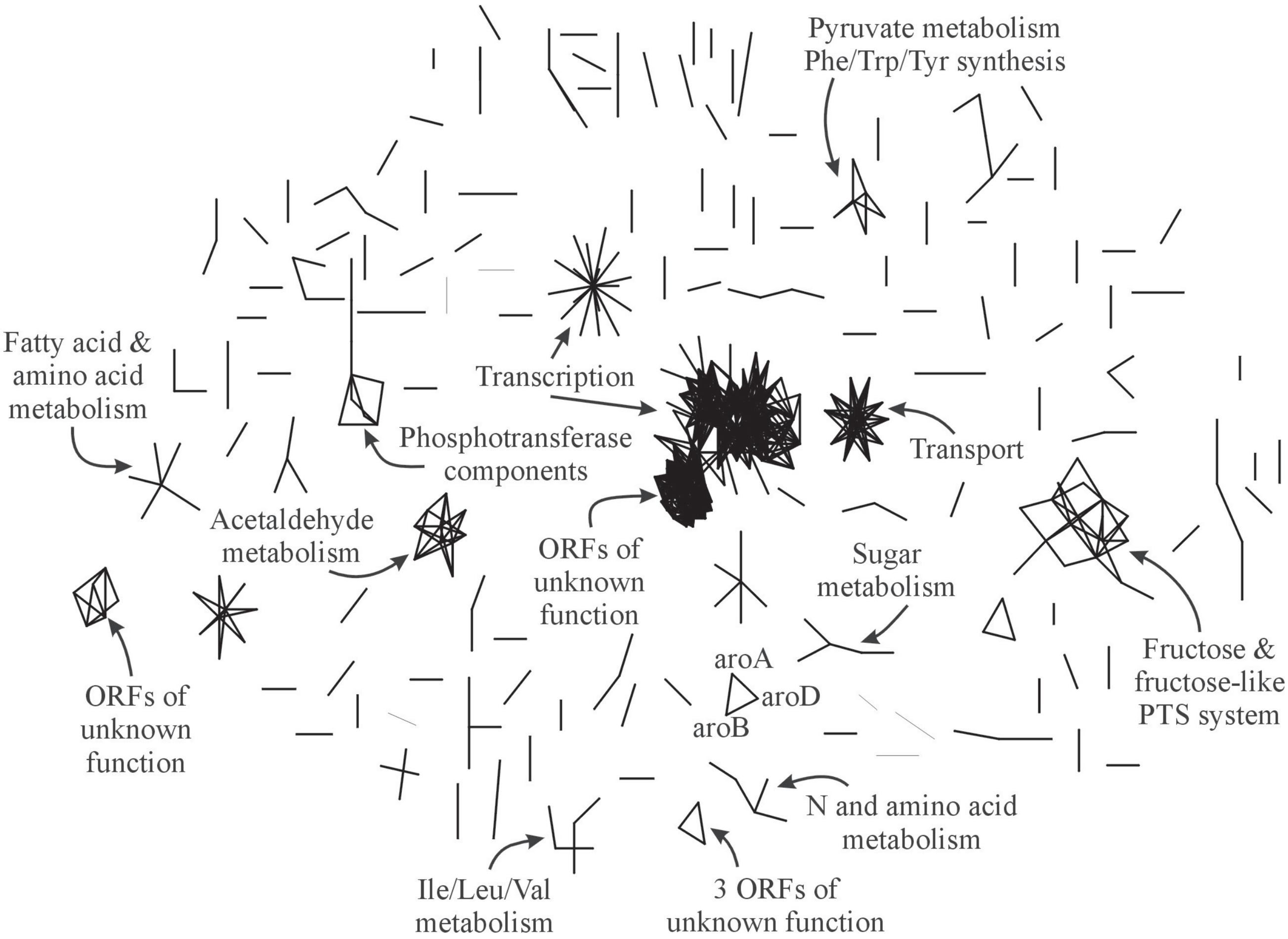


Figure 6 The 854 high confidence functional linkages between 475 *E. coli* proteins define a subset of the genome-wide gene network. Proteins are drawn as points. Rosetta Stone links with significance scores better than 1×10^{-10} are drawn as lines connecting the linked proteins. This figure was generated as in Marcotte 1999b. Essentially, each protein was represented as a point with random coordinates in the plane, and linked proteins were represented as being connected by springs. Following iterative cycles of moving the proteins to minimise the spring energies, linked proteins sit close to each other on the page.

synthesis. Non-metabolic proteins are also linked by fusion events, such as proteins in signalling pathways and transcriptional regulation. Many open reading frames (ORFs) of unknown function are included in the network, and their linkage to proteins of known function should allow preliminary assignment of the general function of many of these ORFs. Other uncharacterised ORFs are linked only to each other, such as the yedL, yedN, and yedM genes linked to each other (labeled '3 ORFS of unknown function'), suggesting only that the genes associate in the same pathway.

Discussion

It is an open question if this method will be scaleable to human proteins. We have empirically observed that higher eukaryotes have large numbers of duplicate genes. The effect of these duplicate genes is to exacerbate both the promiscuous-domain induced ambiguity illustrated in Figure 3 and the within-genome ambiguity that arises from multiple human homologues of each of the linked proteins. The statistical measure we have described here should help in this regard. Early tests on the set of *C. elegans* genes (*C. elegans* genome sequencing consortium 1998) resulted in many linkages being created between adjacent genes. This trend suggested that such pairs of genes were incorrectly annotated as being separate genes; instead, each pair more likely represented two halves of a single gene. As the set of human genes is currently even more fragmented than the worm genes (Lander et al 2001; Venter et al 2001), it is likely that this method will initially allow the identification of many such cases of poor annotation. As the annotation improves, the method should gain power for identifying functionally linked human genes, as well as enable the use of the human genes for linking proteins from other organisms. The method therefore works well on prokaryotes and lower eukaryotes, and is likely to work on higher eukaryotes as gene annotation improves.

In conclusion, we have analysed the use of gene fusion events as a tool to predict functional linkages between proteins. We have developed a statistical measure that appears to account for the largest sources of error. Development of a benchmark for testing functional predictions, and comparison of the significance score against the benchmark shows that the significance score correlates well with the degree of functional relatedness of the linked proteins. The significance score will now also enable calculation of the significance of a link predicted by multiple methods, provided each has a measure of significance. With such a measure of linkage significance in hand, it is possible to calculate partial protein networks. These partial networks can logically be combined with functional linkages inferred by other functional genomics information, such as linkages from phylogenetic profiles (Pellegrini et al 1999), operon predictions (Dandekar et al 1998; Overbeek et al 1999; Salgado et al 2000), and mRNA coexpression patterns (Eisen et al 1998; Marcotte et al 1999b). By combining functional linkages inferred from many disparate sources of information, it should soon be possible to reconstruct large portions of the genetic networks of an organism.

Acknowledgements

We would like to thank Shailesh Date for constructing the sequence database. This work was supported by grant # F-1515 from the Welch Foundation and a Camille and Henry Dreyfus New Faculty Award to EM Marcotte.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–402.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28:45–8.

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science*, 277:1453–74.

*C. elegan*s genome sequencing consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282:2012–18.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23:324–8.

Devos D, Valencia A. 2000. Practical limits of function prediction. *Proteins*, 41:98–107.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863–8.

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. 2000. Protein function in the post-genomic era. *Nature*, 405:823–6.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90.

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D et al. 1997. The yeast genome directory. *Nature*, 387(6632 Suppl):5–6.

Huynen M, Snel B, Lathe W, Bork P. 2000a. Exploitation of gene context. *Curr Op Struct Biol*, 10:366–70.

Huynen M, Snel B, Lathe W, Bork P. 2000b. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10:1204–10.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28:27–30.

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res*, 5:55–76.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–3.

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999b. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–6.

Marcotte EM. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol*, 10:359–65.

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B. 2002. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30:31–4.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA*, 96:2896–901.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 96:4285–8.

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. 2000. Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci USA*, 97:6652–7.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science*, 278:631–7.

Tsoka S, Ouzounis CA. 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet*, 26:141–2.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. *Science*, 291:1304–51.

Yanai I, Derti A, DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci USA*, 98:7940–5.