BIOINFORMATICS

Vol. 19 no. 12 2003, pages 1524–1530
DOI: 10.1093/bioinformatics/btg187

Identification of functional links between genes using phylogenetic profiles

Jie Wu[1], Simon Kasif[1, 2] and Charles DeLisi[1, 2]

[1]Department of Biomedical Engineering and [2]Bioinformatics Graduate Program, Boston University, 44 Cummington St., Boston, MA, 02215, USA

Received on November 22, 2002; revised on February 14, 2003; accepted on February 15, 2003

ABSTRACT

Motivation: Genes with identical patterns of occurrence across the phyla tend to function together in the same protein complexes or participate in the same biochemical pathway. However, the requirement that the profiles be identical (i) severely restricts the number of functional links that can be established by such phylogenetic profiling; (ii) limits detection to very strong functional links, failing to capture relations between genes that are not in the same pathway, but nevertheless subserve a common function and (iii) misses relations between analogous genes. Here we present and apply a method for relaxing the restriction, based on the probability that a given arbitrary degree of similarity between two profiles would occur by chance, with no biological pressure. Function is then inferred at any desired level of confidence.

Results: We derive an expression for the probability distribution of a given number of chance co-occurrences of a pair of non-homologous orthologs across a set of genomes. The method is applied to 2905 clusters of orthologous genes (COGs) from 44 fully sequenced microbial genomes representing all three domains of life. Among the results are the following. (1) Of the 51 000 annotated intrapathway gene pairs, 8935 are linked at a level of significance of 0.01. This is over 30-fold greater than the 271 intrapathway pairs obtained at the same confidence level when identical profiles are used. (2) Of the 540 000 interpathway genes pairs, some 65 000 are linked at the 0.01 level of significance, some 12 standard deviations beyond the number expected by chance at this confidence level. We speculate that many of these links involve nearest-neighbor path, and discuss some examples. (3) The difference in the percentage of linked interpathway and intrapathway genes is highly significant, consistent with the intuitive expectation that genes in the same pathway are generally under greater selective pressure than those that are not. (4) The method appears to recover well metabolic networks. This is illustrated by the TCA cycle which is recovered as a highly connected, weighted edge network of 30 of its 31 COGs. (5) The fraction of pairs having a common pathway is a symmetric function of the Hamming distance between their profiles. This finding, that the functional correlation between profiles with near maximum Hamming distance is as large as between profiles with near zero Hamming distance, and as statistically significant, is plausibly explained if the former group represents analogous genes.

Contact: delisi@bu.edu

1 INTRODUCTION

In recent years several high throughput computational methods have been developed in an effort to begin closing the sequence-function gap. The methods include the application of machine learning algorithms to microarray perturbation experiments (Eisen *et al.*, 1998; Southern *et al.*, 1999); domain fusion (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Yanai *et al.*, 2001); chromosomal proximity (Overbeek *et al.*, 1999) and phylogenetic profiling (Huynen and Bork, 1998; Gaasterland and Ragan, 1998; Pellegrini *et al.*, 1999). Phylogenetic profiling is especially intriguing because it is broadly applicable, permitting inference about human as well as microbial genomes. In its simplest form however, when only identical profiles pairs are used to draw inferences, the method has low coverage.

The profile of a gene is the pattern of occurrence of its orthologs across a set of genomes. In particular, it is a vector of binary digits; a 1 denoting the presence of the gene in a given genome, a 0, absence. The restriction to genes with identical profiles can be relaxed in a number of ways. One way is to calculate the probability that the relation observed between two non-identical profiles could have been obtained purely by chance. Functional links are then assigned in accordance with the criterion used to reject the null hypothesis. The restriction can also be lifted by calculating a correlation coefficient between two vectors, or by calculating their mutual information. In order to provide perspective, we apply these methods as well, and discuss them in terms of our background distribution.

We apply the method to clusters of orthologous genes from 44 fully sequenced microbial genomes representing 41 lineages in the three domains of life. We find 51 thousand annotated intrapathway gene pairs of which 8935 are linked at a level of confidence of 99%. We also identify 540 thousand interpathway genes pairs, of which some 65 000 are linked at the 99% level of confidence. This latter number is some 12 standard deviations beyond the number expected by chance at this confidence level, suggesting that the large majority of the genes thus identified subserve a common function, even though they are not in the same pathway. We speculate that many of these links involve intersecting paths (i.e. paths sharing one of more genes) and discuss some examples. At any confidence level, the difference between the percentage of interpathway genes that are linked and intrapathway genes that are linked is highly significant. In addition, as confidence level increases, the ratio of intrapathway to interpathway linked pairs increases. These results are consistent with the intuitive expectation that genes in the same pathway are generally under greater selective pressure than those that are not.

2 METHODS

2.1 The data set

We adhere closely to the conventions of the COG database (http://www.ncbi.nlm.nih.gov/COG/; Tatusov *et al.*, 2001) and construct profiles only for genes that occur in at least three lineages. All paralogs are collapsed; i.e. a set of closely related genes in a given lineage is treated as a single entity. The collapse of paralogous genes substantially reduces the number of links. For example, we can detect 469 129 linked gene pairs among 4289 *E.coli* genes at a confidence level of 99% when the paralogous genes are not collapsed, as opposed to 158 080 pairs from 1612 COGs, when they are collapsed. However, the additional links add little information, since paralogs typically have related functions.

We also eliminated from consideration all sets of 20 or more COGs that share the same profile. There are 261 COGs with seven distinct profiles in this group, leaving $3166 - 261 = 2905$ COGs. For example housekeeping genes have a common profile (present in all or almost all genomes), but they also have diverse functions; they and their profiles are therefore eliminated from consideration at the outset. Such prior screening is important when the inference criterion is based on identical profiles. In the more general situation considered here, we will see that housekeeping and other genes having unusual profiles are eliminated naturally by the method itself, rather than by an a priori screen.

Let $N = 41$ be the total number of *lineages* over which we construct profile vectors for R COGs, denoted by X_1, $X_2 \ldots X_R$. Each gene X_i is therefore represented by an N-component vector describing its pattern of occurrence across the set of lineages, a component having a value of 1 when the gene is present and 0 when it is not. Define the profile vectors $\vec{x}$ and $\vec{y}$ for genes X and Y, and let x and y be the number of lineages in which (orthologs of) gene X and Y occur (i.e. x and y are the sums of the components of their profile vectors, sometimes called the Hamming weight). We further define the variable z as the number of lineages in which X and Y co-occur. We will be comparing average or first order properties of these vectors: the chance probability of a given number of co-occurrences, properly conditioned as described below; the Pearson correlation coefficient and the Mutual Information. First-order properties mean that we omit, as in all previous work, correlations between the genomes themselves (i.e. variations in phylogenetic distance between genomes).

2.2 Chance co-occurrence probability distribution (P)

The distribution of interest is $P(z|N, x, y)$, the probability of observing by chance (i.e. no functional pressure) z co-occurrences of genes X and Y in a set of N lineages, given that X occurs in x lineages, and Y occurs in y.

To be specific in the development of the formalism, we will define x to be the smaller of the two variables (x, y). Therefore, by definition

$$z_{\max} = x. \tag{1}$$

In addition, the minimum value of z is constrained by

$$z_{\min} = \begin{cases} x + y - N & x + y \geqslant N \\ 0 & x + y < N \end{cases}. \tag{2}$$

Define w_z as the number of ways to distribute z co-occurrences over the N lineages, and $\bar{w}_z$ as the number of ways of distributing the remaining $x - z$ and $y - z$ genes over the remaining $N - z$ lineages. $P(z|N, x, y)$ is the number of ways in which x and y can be distributed over N genomes, given that there are z co-occurrences, divided by the total number of ways x and y can be distributed without restriction. The total number of ways in which z co-occurrences can occur, given N, x and y is $w_z \times \overline{w}_z$ where

$$\begin{aligned} w_z = \binom{N}{z} \text{ and } \overline{w}_z &= \binom{N-z}{x-z}\binom{N-x}{y-z} \\ &= \frac{(N-z)!}{(N+z-x-y)!(x-z)!(y-z)!}. \end{aligned} \tag{3}$$

The number of ways, W, of distributing X and Y over N lineages without restriction is

$$W = \binom{N}{x}\binom{N}{y}. \tag{4}$$

Therefore,

$$P = \frac{w_z \overline{w}_z}{W} \tag{5}$$

where the arguments of P have been omitted for notational simplicity.

For future reference we define $p_1(X)$ and $p_2(X)$ as the fraction of lineages in which gene X is present and absent, respectively.

$$p_1(X) \equiv \bar{x} = x/N,\ p_2(X) \equiv 1 - p_1(X) = 1 - \bar{x}. \quad (6)$$

2.3 Other measures of correlation between profiles

To provide perspective we consider three standard measures of profile similarity, the Hamming distance (D), the Pearson correlation coefficient (r) and mutual information (I).

Hamming Distance (D)

Hamming distance as a function of (x, y, z) is

$$D = x + y - 2z. \quad (7)$$

Pearson Correlation coefficient (r)

With no correlation between phyla, we have

$$r = \frac{Nz - xy}{\sqrt{(Nx - x^2)(Ny - y^2)}}. \quad (8)$$

If $x = y = z$, then $D = 0$ and $r = 1$ for all (x, y, z). This expresses the fact that the Pearson correlation coefficient cannot discriminate between identical profiles. The statement may seem obvious, but other measures, in particular mutual information (below) and the probability of chance occurrence, do make such a distinction. In a similar manner we find that when $D = 1$, the range of correlation coefficients is highly restricted.

The actual number of variables on which r depends can be seen more clearly by defining mean occurrence probabilities $f_z = z/N$, $f_y = y/N$, $f_x \equiv p_1(X) = x/N$. Then

$$r = \frac{f_z - f_x f_y}{\sqrt{(f_x - f_x^2)(f_y - f_y^2)}}. \quad (9)$$

Mutual Information (I)

The mutual information is

$$I(X, Y) = \sum_{i,j} p_{ij}(X, Y) \log_2 \frac{p_{ij}(X|Y)}{p_i(X)}$$
$$= \sum_{i,j} p_{ij}(X, Y) \log_2 \frac{p_{ij}(X, Y)}{p_i(X) p_j(Y)}. \quad (10)$$

To make the connection with Equation (9) more explicit, we carry out the sums in Equation (10). In particular, define

$$I_1(X, Y) \equiv f_z \log_2 \frac{f_z}{f_x f_y};$$
$$I_2(X, Y) \equiv (f_x - f_z) \log_2 \frac{(f_x - f_z)}{f_x(1 - f_y)};$$
$$I_3(X, Y) \equiv (f_y - f_z) \log_2 \frac{(f_y - f_z)}{f_y(1 - f_x)};$$
$$I_4(X, Y) \equiv (1 - f_x - f_y + f_z) \log_2 \frac{(1 - f_x - f_y + f_z)}{(1 - f_x)(1 - f_y)}.$$

Then

$$I(X, Y) = I_1(X, Y) + I_2(X, Y) + I_3(X, Y) + I_4(X, Y). \quad (11)$$

For identical profiles, the mutual information is zero as the co-occupancy of lineages becomes complete $(z = N)$ or non-existent $(z = 0)$, and reaches a maximum of 1 when the co-occupancy of the set of lineages is 50%. Thus there is a great deal of variation in the mutual information of identical profile pairs. In general, r, I and P are all functions of x, y and z given N but not unique functions of one another. For completeness we note that combinatorial expression for $P(z|x, y, N)$ (Equation 6) which is exact for all N, is approximated by a binomial provided all factorials become large, and provided $f_z = f_x f_y$. In that case Equation (6) becomes the natural background distribution for the Pearson correlation coefficient r.

3 RESULTS AND DISCUSSION

3.1 The relation between different measures of correlation

We have developed an expression for the probability that the two profiles match by chance, as a function of x, y, z and N. For comparison we also discuss the utility of the Pearson correlation coefficient and mutual information, which are functions of the normalized variables x/N, y/N and z/N. Ultimately we would like to be able to comment on the extent to which the value of a correlate provides information about the relation between a pair of genes. There is of course a certain level of arbitrariness in choosing the level of biological function at which such a correlation is to be used. In this paper we discuss functional linkage at two levels: in terms of 126 biochemical pathways as given in the *Kyoto Encyclopedia of Genes and Genomes* (http://www.genome.ad.jp/kegg/kegg2.html, Kanehisa and Goto, 2000) and in terms of the 18 coarse grained functional categories compiled by the *National Center of Biotechnology Information* (http://www.ncbi.nlm.nih.gov/COG/). Although these ontologies are useful, they provide only a small glimpse at all the relations between genes they contain since many genes in different pathways, or in different COG categories, subserve common functions. For example, of the 42 thousand

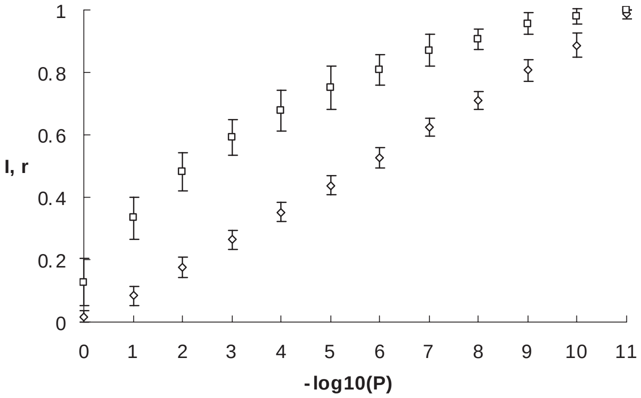

Fig. 1. The distribution of mutual information (I: diamond) and Pearson correlation coefficient (r: square) as a function of probability of chance co-occurrence. For any particular value of probability, there is a distribution of values of the other two measures of correlation. The standard deviations in these distributions are shown by the error bars.

COG pairs that are in the same COG functional category and annotated in KEGG, 31 thousand are in different pathways. We will return to this later. First we comment on the relation between the different correlation measures.

Profile pairs in a relatively narrow probability interval will have disparate combinations of (x, y, z). A particular set of combinations corresponding to a common probability does not, however, correspond to a common value of Pearson correlation or mutual information; i.e. the chance occurrence probability of a profile-pair does not uniquely specify either the Pearson correlation coefficient or the mutual information (Fig. 1). Analogous statements are true when r and I are the independent variables (data not shown). The three measures of correlation are nevertheless strongly coupled.

With mutual information and Pearson correlation as a function of probability of chance co-occurrence (Fig. 1), the mutual information increases more or less logarithmically as probability decreases, and its dispersion is relatively invariant and tight. The correlation coefficient (r) on the other hand has a larger dispersion, rises relatively rapidly with $-\log P$, and is relatively insensitive to changes in probability for correlation coefficients above 0.8.

Profiles with r below 0.4 are not significantly correlated and the dispersion in the relation between I (or probability) and r increases as the correlation coefficient increases (data not shown). High correlations are of course significant in terms of probability, but the range of probability values corresponding to a given level of correlation spans several orders of magnitude; i.e. r essentially loses its ability to discriminate between profiles whose levels of significance varies widely. This has implications (to be developed elsewhere) for the specificity with which genes can be allocated to neighboring pathways. The dynamic range of the Pearson correlation coefficient is therefore smaller than it is for the mutual information.

3.2 Degenerate profiles

A number of genes have identical profiles. When the number of genes that share a particular profile is small, such identity is a strong indication that the genes are related. To be specific, seven profiles have 20 or more genes in common. Five of the seven represent housekeeping genes; i.e. genes that occur in all or nearly all genomes. One of them, having a total number of 70 genes, occurs in all 41 lineages. An application of Equation (6) shows immediately that pair wise profile patterns formed by the profile of these genes and any other profile is entirely insignificant. They therefore do not need to be screened out a priori, but are properly eliminated by the method. Two others are more interesting. One occurs in nine out of 41 lineages; the other in 30 of 41 lineages. The probability of pairing these two profiles, taken across all genomes, with others can be significant. However, the former which includes 20 genes, occurs in archaea only (9/9), and the latter (21 genes) occurs in bacteria (30/30) only. When paired with *any* profiles *within* their life domains the patterns are insignificant.

The remaining two profiles occur in archaea (9/9) + eukaryotes (2/2) only and bacteria (30/30) + eukaryotes (2/2) only (53 and 28 genes respectively). Here too, there is little or no chance that a function would be incorrectly assigned.

Two additional profiles remain and they do cause error, because they are related to genomic correlations. One of them includes 39 genes that only occur in three archaea genomes, *Archaeoglobus fulgidus*, *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*, and forms a cluster on the evolutionary tree. These genes perform functions ranging from cell division, ribosomal structure and biogenesis, to energy production and conversion and amino acid transport and metabolism. The other promiscuous profile contains 30 genes exclusively from four closely related bacterial genomes: *Escherichia coli*, *Haemophilus influenzae*, *Vibrio cholerae* and *Pasteurella multocida*. None of the 30 genes have pathway annotation in KEGG; 10 of them are annotated, and these are distributed over three COG functional categories: DNA replication, recombination and repair; cell envelope biogenesis; cell division and chromosome partitioning. The significance of the intra-group profiles for the 39 gene archaea group and the 30 gene bacterial group are 10^{-5}, 10^{-6} respectively when all 41 lineages are used, and 10^{-2}, 10^{-5} after life domain adjustment. In this instance we clearly draw the wrong inference—the correlation is significant not because of similarity between genes, but because of similarity between genomes.

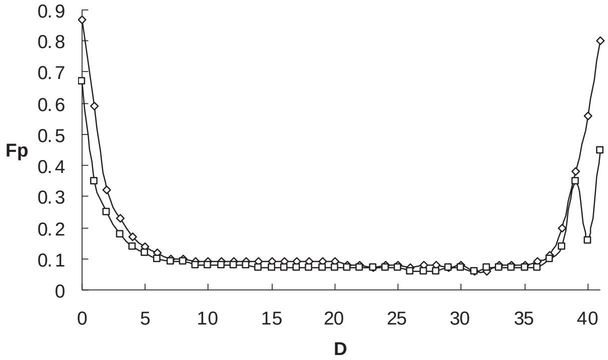

Fig. 2. F_P: fraction of gene pairs that have a profile correlation of Hamming Distance D, and share a common pathway in KEGG (diamond), or fall into the same functional category in COG (square).

3.3 Correlations as a function of Hamming distance (D)

When profiles are identical ($D = 0$), the fraction of annotated genes sharing at least one pathway is approximately 0.9 and, over a range of seven decades, it is nearly independent of the probability that the profiles are related by chance. As D increases (Fig. 2) the fraction of genes that share a pathway decreases rapidly until D of about 5, remains relatively constant at 0.08 between 5 and 35, and then increases again. Thus the likelihood that pairs will have a similar function as defined by the COG and KEGG ontologies is independent of the distance between profiles, except for profiles that are very similar (fewer than three or four differences), or very dissimilar (fewer than three or four similarities).

For $D > 36$, the two profiles in a pair are nearly symmetric, rather than nearly identical. They therefore have a high negative correlation, and as a result their functional correlation increases at large D in a way that is symmetric to its decline for small D. In other words at large values of D, when gene X is present in a given genome, gene Y tends to be absent, and vice versa, but X and Y occur in the same set of orthologous pathways (that is, pathways in different organisms, in analogy to orthologous genes), as though they are substituting for one another. It therefore appears plausible that X and Y are analogs; i.e. they are different genes with different sequences, which appear to be playing a similar role. Such patterns can arise by non-orthologous gene displacement, in which one gene is replaced by analogous genes in one or more genomes (Koonin *et al.*, 1996). For example, a class I lysyl-tRNA (COG1384) can replace the non-homologous but analogous class II gene (COG1190) in function, leaving each genome able to use either of the two (Galperin and Koonin, 1999). Both genes participate in the *Lysine biosynthesis* and *Aminoacyl-tRNA biosynthesis* pathway while having a profile $D = 41$. What is important as far as this paper is concerned is that (Equation 5) symmetric or nearly symmetric profiles have a low chance co-occurrence probability. If we were limited to using only perfect matches to draw functional correlation, the significance of symmetric or nearly symmetric profile pairs would be missed entirely.

3.4 Functional linkage

The NCBI database provides 1.54×10^6 annotated pairs of COGs, 113 380 are in the same functional categories; for the Kyoto encyclopedia of genes and genomes, 51 643 of the 594 595 annotated gene pairs are in the same pathway. To be specific about the meaning of this last set of numbers, define n_k as the number of pathways with k proteins. Such pathways have $k(k-1)/2$ pairs and $\sum n_k k(k-1)/2 = 51\,643$, the total number of intrapathway pairs.

We refer to a gene pair as linked if the correlation in their profiles exceeds some specified value. Thus profiles having a chance occurrence probability below 10^{-2} are said to be linked at a significance level of 10^{-2}; those whose probability of chance occurrence is below 10^{-3} are said to be linked at that level, etc. For any specified degree of linkage, gene pairs can be in one of four mutually exclusive and collectively exhaustive categories: in the same pathway and either linked or not linked, and in different pathways and either linked or not linked. We are interested in the distribution of genes pairs and their correlations, in these categories.

We characterize this distribution with the 2×2 matrix $C_{ij}(i = 1, 2;\ j = 1, 2)$, where i labels pathway state ($i = 1$, same pathway; $i = 2$, different pathway) and j labels linkage state ($j = 1$, linked; $j = 2$, not linked) (Table 1). Because there are many more interpathway than intrapathway gene pairs, the number of links is actually larger in the former group than in the latter when the confidence level is 99%. In particular, of the more than 51 000 pairs that share a pathway, 8935 are linked at the 0.01 level; i.e. the probability that the relation between the profiles could have been a chance happening with no evolutionary pressure is less than 0.01 (Fig. 3). Many of the gene pairs in the same pathway are not sufficiently correlated to rule out chance occurrence, even at the significance level of 10^{-2}. For example, the gene pair of pyruvate dehydrogenase (COG2609) and phosphoglucomutase (COG0033) both participate in the glycolysis pathway while their profiles chance co-occurrence probability is more than 10^{-2}.

On the other hand, for the complementary population of 542 952 pairs that do not share a pathway, 65 422 are correlated at a confidence level of 99% (Table 1). This is more than 10 standard deviations above the number expected by chance at the 0.01 level of significance. As noted earlier, when we commented on the connection between COG

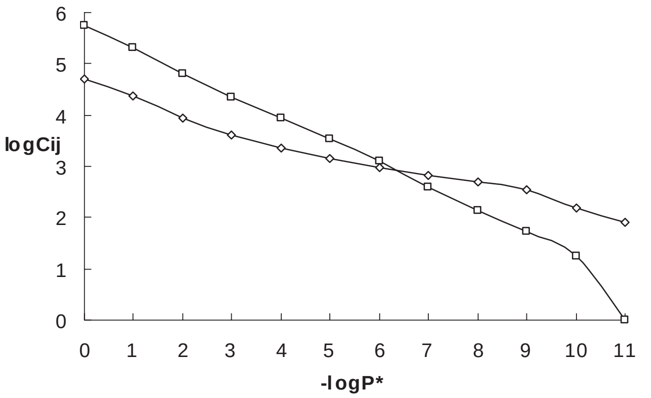

Fig. 3. Number of gene pairs as a function of confidence level. C_{11}(diamond) is the number of gene pairs that share a common pathway in KEGG: intrapathway linked pairs. C_{21} (square) is the number of gene pairs that do not share pathway: interpathway linked pairs.

Table 1. C_{ij} is the number of pairs with link state at the confidence level of P^*, ($j = 1$, linked, $j = 2$, not linked), and pathway state ($i = 1$, share a common pathway or fall in the same functional category (COG); $i = 2$, not share a common pathway (KEGG) or in different functional categories (COG))

$-\log_{10} P^*$	C_{11}	C_{12}	C_{21}	C_{22}
0	51643	0	542952	0
1	23252	28391	208550	334402
2	8935	42708	65422	477530
4	2236	49407	8586	534366
6	961	50682	1250	541702
8	493	51150	133	542819
10	156	51487	18	542934
11	79	51564	1	542951

and KEGG, many of interpathway linked genes can fall into the same functional category. For example, NADH (COG0623) and acetate kinas (COG0282) have profile pairs significant at the 10^{-2} level. The former participates only in the Fatty acid biosynthesis pathway, while the latter is in three pathways: Glycolysis, Pyruvate metabolism and Propanoate metabolism pathway. However, the fatty acid biosynthesis pathway neighbors both the Pyruvate metabolism and Propanoate metabolism pathway, in the sense that they share common genes, which in this case are Acetyl-CoA carboxylase beta subunit (COG0777) and Acetyl-CoA carboxylase alpha subunit (COG0825). Another example is Lysophospholipase (COG2267) and Glycerol 3-phosphate dehydrogenase (COG0240) which participate in the Phospholipid degradation and Glycerolipid metabolism pathway respectively, while having a profile pair linked at a significance level of 10^{-2}. These two pathways are also neighbors, sharing Glycerophosphoryl diester phosphodiesterase (COG0584) and an outer membrane phospholipase A (COG2829).

These examples of interpathway functional correlations not withstanding, intrapathway genes are significantly more related than interpathway genes. Thus while 12% of the interpathway genes (65 422 out of 542 952) are linked at the 0.01 level of significance, 17.3% of the intrapathway genes are linked at that significance level. If instead we use probabilities of 10^{-4} and 10^{-6} as measures of significance, the corresponding numbers are 1.5 and 4.3% in the first case, and 0.2 and 1.8% in the second. These differences are all highly significant. More generally, we summarize in Table 1 C_{ij} as a function of significance level of linkage. The differences are significant at a *Chi-square* level of 867 with one degree of freedom. This number can be placed in perspective by recalling that a *Chi-square* value of seven corresponds to a probability of chance occurrence between 10^{-2} and 10^{-3}.

An exact calculation using fisher's test indicates that the probability that percentages are drawn from the same population is essentially zero. In particular under the null hypothesis, the probability of interest is the number of ways in which at least C_{11} linked genes can be distributed between intra and interpathway pairs, divided by the total number of ways in which all linked genes can be

$$\frac{\displaystyle\sum_{k=C_{11}}^{\min(C_{11}+C_{21},C_{11}+C_{12})}\binom{C_{11}+C_{12}}{k}\binom{C_{21}+C_{22}}{C_{11}+C_{21}-k}}{\dbinom{C_{11}+C_{12}+C_{21}+C_{22}}{C_{11}+C_{21}}}.$$

In addition to significantly different percentages, plots of the number of intrapathway and interpathway linked genes as a function of level of significance cross at a $P^* \approx 10^{-6}$ (Fig. 3); in other words, the more stringent the criterion, the more likely the linked genes will belong to the same pathway. This again is consistent with the idea that interpathway genes are much more constrained to co-function and therefore to co-evolve.

These relations not withstanding, there are a large number of pairs whose profiles are significantly correlated but are not in the same pathway. A significant statistical correlation therefore tends to suggest the same pathway, with the likelihood increasing with stringency of confidence level, but the same pathway is not guaranteed at any level of confidence. Nevertheless these findings, and in particular the increase in pathway co-occurrence with threshold, have implications for the allocation of genes of unknown function to locations on a graph of a cellular protein network, a subject that will be developed in future publications.

As noted previously (Yanai and DeLisi, 2002), phylogenetic links uncover networks. Restricted profiling however, only uncovers fully connected clades. By

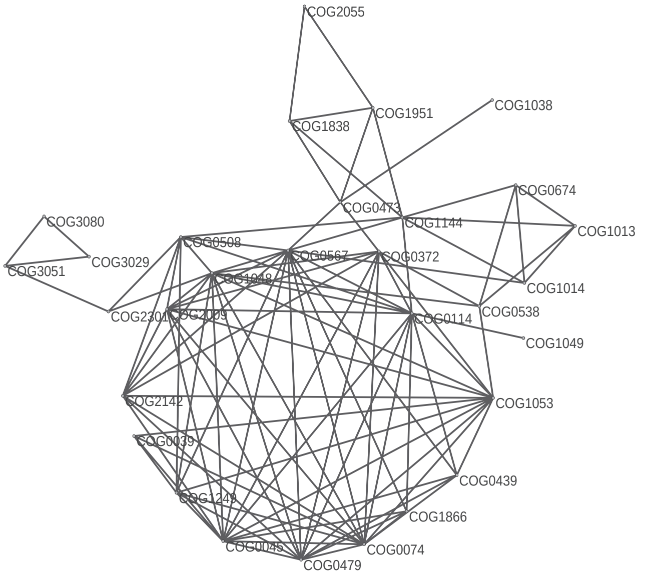

Fig. 4. 30 out of 31 TCA genes are uncovered as a weighted edge network at the 99% confidence level (weights not shown).

relaxing the restriction, we also remove the constraint on complete network connectivity and thereby recover many more clusters. It is not our intention to develop this complex subject here; we show only one example to illustrate the difference between graphs obtained by unrestricted and those obtained by restricted profiling as presented previously (Yanai and DeLisi, 2002). In particular at a 99% confidence level we uncover 30/31 COGs in the TCA cycle as a network with weighted edges (Fig. 4). This compares favorably with the best results obtained previously, when chromosomal proximity, domain analysis and restricted phylogenetic profiling were used in conjunction with one another (Yanai and DeLisi, 2002).

ACKNOWLEDGEMENTS

We thank Itai Yanai, Vladimir Pavlovic and Joseph Mellor for helpful discussions. S.K.is supported in part by NSF (KDI). C.D. acknowledges grant support from NSF (D90-9870710) and NIH (A08_POGM66401).

REFERENCES

Aravind,L. (2000) Guilt by association: contextual information in genome analysis. *Genome Res.*, **8**, 1074–1077.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Eisen,J.A. and Wu,M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.*, **61**, 481–487.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Bouzoukis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.

Galperin,M.Y. and Koonin,E.V. (1999) Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica*, **106**, 159–170.

Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.

Huynen,M.A. and Snel,B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.*, **54**, 345–379.

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

King,R.D., Karwath,A., Clare,A. and Dehaspe,L. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, **17**, 445–454.

Koonin,E.V., Mushegian,A.R. and Bork,P. (1996) Non-orthologous gene displacement. *Trends Genet.*, **12**, 334–336.

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.

Marcotte,E.M., Pelligrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

McEliece,R.J. (2002) *The theory of information and coding*. Cambridge University Press.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pellegrini,M. (2001) Computational methods for protein function analysis. *Curr. Opin. Chem. Biol.*, **5**, 46–50.

Pellegrini,M., Marcotte,E.M., M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nat. Genet.*, **21**, 5.

Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov,R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.

Wolf,Y., Rogozin,I., Grishin,N. and Koonin,E. (2002) Genome trees and the tree of life. *Trends Genet.*, **18**, 472.

Yanai,I. and DeLisi,C. (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol.*, **3**, research 0064.1–0064.12.

Yanai,I., Derti,A. and 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**, 7940–7945.