Identification of functional links between genes using phylogenetic profiles

Jie Wu¹, Simon Kasif¹,² and Charles DeLisi¹,²

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

ABSTRACT

Motivation: Genes with identical patterns of occurrence across the phyla tend to function together 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, 9335 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 gene 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; Pelegrini 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 interpathway to intrapathway 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 more COGs that share the same profile. There are 261 typically have related functions.

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

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

Therefore,

$$ P = \frac{w_z \overline{w}_z}{W}. $$
where the arguments of \( P \) have been omitted for notational simplicity.

For future reference we define \( p_1(X) = \frac{x}{N} \) and \( p_2(X) \) as the fraction of lineages in which gene \( X \) is present and absent, respectively.

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

### 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

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

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

\[
I_1(X, Y) = f_z \log_2 \frac{f_z}{f_x f_y}; \quad I_2(X, Y) = (f_x - f_z) \log_2 \frac{(f_x - f_z)}{f_x(1 - f_y)}; \\
I_3(X, Y) = (f_y - f_z) \log_2 \frac{(f_y - f_z)}{f_y(1 - f_x)}; \quad I_4(X, Y) = (1 - f_z - f_x - f_y) \log_2 \frac{(1 - f_x - f_y + f_z)}{(1 - f_x)(1 - f_y)}. \quad (11)
\]

Then

\[
I(X, Y) = I_1(X, Y) + I_2(X, Y) + I_3(X, Y) + I_4(X, Y).
\]

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, \( r \) 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 \( f_x = f_x f_y \). 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
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. Of 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.
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 several 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, 113380 are in the same functional categories; for the Kyoto encyclopedia of genes and genomes, 51643 of the 594595 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 \( \sum n_k(k−1)/2 = 51643 \), 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 51000 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 542952 pairs that do not share a pathway, 65422 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
and KEGG, many of interpathway linked genes can fall into the same functional category. For example, NADH (COG0623) and acetate kinases (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

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

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

$$\sum_{k=C_{11}}^{\min(C_{11}+C_{12}+C_{11}+C_{12})} \frac{C_{11} + C_{12}}{C_{11} + C_{12} + C_{11} + C_{21} + C_{22}} \frac{C_{21} + C_{22}}{C_{11} + C_{21} + C_{22}}$$

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-evolve. These relations notwithstanding, 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 uncover fully connected clades. By
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


