Deciphering Protein Network Organization Using Phylogenetic Profile Groups

Jie Wu[1] **Joseph C. Mellor**[2] **Charles DeLisi**[1,2]
jiewu@bu.edu mellor@bu.edu delisi@bu.edu

[1] Department of Biomedical Engineering, Boston University, 24 Cummington St. Boston, MA, 02215, USA
[2] Graduate program of Bioinformatics, Boston University, 24 Cummington St. Boston, MA, 02215, USA

Abstract

Phylogenetic profiling is now an effective computational method to detect functional associations between proteins. The method links two proteins in accordance with the similarity of their phyletic distributions across a set of genomes. While pair-wise linkage is useful, it misses correlations in higher order groups: triplets, quadruplets, and so on. Here we assess the probability of observing co-occurrence patterns of 3 binary profiles by chance and show that this probability is asymptotically the same as the mutual information in three profiles. We demonstrate the utility of the probability and the mutual information metrics in detecting overly represented triplets of orthologous proteins which could not be detected using pairwise profiles. These triplets serve as small building blocks, *i.e.* motifs in protein networks; they allow us to infer the function of uncharacterized members, and facilitate analysis of the local structure and global organization of the protein network. Our method is extendable to N-component clusters, and therefore serves as a general tool for high order protein function annotation.

Keywords: probability, mutual information, logic analysis, protein functional linkage network

1 Introduction

The advent of fully sequenced genomes provides the possibility of systematically annotating large numbers of genes by comparative genomics. High throughput algorithms that explore genomic context have been developed to infer functionally linked proteins and consequently to transfer functionality of annotated proteins to their unannotated partners. Among them are phylogenetic profiling methods [3, 14, 16], which link proteins based on the high correlation between their phyletic distribution across a suitable set of fully sequenced genomes; the domain fusion method [4, 19], which identifies protein pairs that fuse into a single peptide in another organism; and chromosomal proximity method [1, 2], which links proteins based on their co-localization on the chromosome in multiple genomes. These methods annotate unknown proteins by detecting pair-wise protein functional associations including co-occurrence in complexes and pathways. However, pair-wise associations do not adequately describe the full complexity of cellular networks which consist of hierarchies of correlated clusters [11, 12].

Phylogenetic profiling links two proteins based on the correlation between their profiles, i.e. the binary strings that encode the absence or presence of proteins across a set of genomes. The performance of phylogenetic profiling in inferring protein functional linkages is quantitatively comparable to, if not better than, Yeast 2 hybrid [3, 6, 17]. However, since previous work has been limited to evaluating correlation between protein pairs information on the higher order organization of proteins in the cell is overlooked.

In this work, we assess two correlates: the probability of observing a specified similarity by chance, and the mutual information in the triplet, and show that the former approaches the latter asymptotically. We briefly discuss new logical relations among triplets and their potential biological implications.

2 Method and Results

Binary Relations The phylogenetic profile of a protein is a binary string recording the presence (1) or absence (0) of a gene across a suitable set of genomes. If the correlation between the profiles of two proteins, X and Y, is much greater than would be expected by chance, then they are assumed to be functionally related. This correlation can be evaluated using the exact probability of observing such a co-occurrence pattern by chance, or by a Pearson Correlation or Mutual information. Protein pairs with a profile correlation greater than a pre-selected threshold are considered functionally linked [1, 16]. In this section, instead of identifying protein pairs, we develop two metrics to detect statistically significant groups of triplets–an exact probability metric and mutual information metric. We show that the latter is a special case of the former, for very long strings, and discuss their utility in deciphering higher order organization of protein networks.

2.1 Probability Metric of Phylogenetic Profiles Groups

Pairs

Let N be the number of genomes over which the profiles are defined, with protein X occurring in x genomes, Y occurring in y genomes, and both occurring in N_{xy} genomes. Then $P(N_{xy}|N,x,y)$, the probability of observing N_{xy} co-occurrences purely by chance, given N, x and y, can be evaluated by hyper-geometric distribution [16]

$$P(N_{xy}|N,x,y) = \frac{\begin{pmatrix} N-x \\ y-N_{xy} \end{pmatrix}\begin{pmatrix} x \\ N_{xy} \end{pmatrix}}{\begin{pmatrix} N \\ y \end{pmatrix}} = \frac{(N-x)!(N-y)!x!y!}{(N+N_{xy}-x-y)!(x-N_{xy})!(y-N_{xy})!N_{xy}!N!} \ . \quad (1)$$

Note that for the two string case, there are only four independent variables: the length N, which is the same for each string; the number of ones in each string (x and y), and the relation between the two strings, for example, the number of positions at which both strings are equal to one (N_{xy}). The metric assumes independence between different genomes.

Ternary Relations

Eight independent variables are needed to characterize the co-occurrence pattern of a triplet of proteins X, Y and Z which occur respectively in x, y and z out of N genomes. In fact for a group of M proteins, 2^M independent variables are needed. For the triplet, we choose N, x, y, z, N_{xy}, N_{yz}, N_{xz} and N_{xyz} where N_{ij} denotes the number of genomes in which genes i and j co-occur and N_{xyz} is the number genomes in which all three proteins co-occur. We then ask a general question: given N, x, y, and z, what is $P(N_{xy}, N_{yz}, N_{xz}, N_{xyz}|N,x,y,z)$, the probability of observing the co-occurrence pattern of the three strings by chance?

To evaluate this probability, we decompose the formation of the co-occurrence pattern of the three strings into 7 steps ($w_1 - w_7$). (i) Distribute the x 1s over the N organisms (w_1). (ii) Distribute the N_{xy} over x organisms (w_2). (iii) Distribute the remaining $y - N_{xy}$ 1s for y over $N - x (w_3)$. x, y are now fixed and N_{xy} is satisfied. To distribute z over N while satisfying N_{yz}, N_{xz}, N_{xyz}, (iv) distribute N_{xyz} out of z over $N_{xy} (w_4)$. Then (v) distribute the remaining $N_{yz} - N_{xyz}$ 1s over $y - N_{xy} (w_5)$ and (vi) $N_{xz} - N_{xyz}$ 1s over $x - N_{xy} (w_6)$. (vii) Complete the process by distributing the remaining $z - N_{xyz} - (N_{yz} - N_{xyz}) - (N_{xz} - N_{xyz}) = z - N_{yz} - N_{xz} + N_{xyz}$ 1s over $N - x - y + N_{xy}$ organisms (w_7). The desired probability can be written as

$$P(N_{xyz}, N_{xz}, N_{xy}, N_{yz}|x,y,z,N) = \frac{\prod_{k=1}^{7} w_k}{W} \ , \quad (2)$$

where

$$W = \begin{pmatrix} N \\ x \end{pmatrix} \begin{pmatrix} N \\ y \end{pmatrix} \begin{pmatrix} N \\ z \end{pmatrix}$$

is the number of ways of distributing x, y and z over N organisms without restriction. The terms in the numerator are

$$\begin{aligned} w_1 &= \begin{pmatrix} N \\ x \end{pmatrix}, w_2 = \begin{pmatrix} x \\ N_{xy} \end{pmatrix}, w_3 = \begin{pmatrix} N-x \\ y-N_{xy} \end{pmatrix}, w_4 = \begin{pmatrix} N_{xy} \\ N_{xyz} \end{pmatrix}, \\ w_5 &= \begin{pmatrix} y-N_{xy} \\ N_{yz}-N_{xyz} \end{pmatrix}, w_6 = \begin{pmatrix} x-N_{xy} \\ N_{xz}-N_{xyz} \end{pmatrix}, w_7 = \begin{pmatrix} N-x-y+N_{xy} \\ z-N_{xz}-N_{yz}+N_{xyz} \end{pmatrix}. \end{aligned} \tag{3}$$

$P(N_{xy}, N_{yz}, N_{xz}, N_{xyz}|N, x, y, z)$ is the number of ways in which x and y and z can be distributed over N genomes, given that there are N_{xy}, N_{yz}, N_{xz} co-occurrences of X and Y, Y and Z, X and Z respectively and N_{xzy} for all of them, divided by the total number of ways x and y and z can be distributed without restriction. To write Eqn. (3) in terms of factorials define

$$\begin{aligned} T &\equiv x!y!z!(N-x)!(N-y)!(N-z)! \quad \text{and} \\ B &\equiv (N_{yz}-N_{xyz})!(N_{xz}-N_{xyz})!(N_{xy}-N_{xyz})! \\ &\quad (y-N_{xy}-N_{yz}+N_{xyz})!(x-N_{xy}-N_{xz}+N_{xyz})!(z-N_{xz}-N_{yz}+N_{xyz})! \\ &\quad (N-x-y-z+N_{xz}+N_{yz}+N_{xy}-N_{xyz})!N!N!N_{xyz}! \ . \end{aligned} \tag{4}$$

Then

$$P(N_{xyz}, N_{xz}, N_{xy}, N_{yz}|x, y, z, N) \equiv T/B,$$

which is symmetric in x, y, z. Eqn. (1) is a special case of Eqn. (4) with $z = 0$.

Eqn. (4) serves as a general metric for identification of statistically significant triplets, i.e. the members of any triplet whose probability is less than a pre-defined threshold are considered functionally related and are subject to subsequent analysis. This probability can identify statistically significant triplets for which none of the three pairs are significantly related and can therefore uncover relationships that are not readily identified using pair-wise profiling.

2.2 Mutual Information in Phylogenetic Profiles

Mutual information $MI(X, Y)$, which measures the reduction in the uncertainty in a string given the information in another string, also provides a useful and computationally less expensive measure of correlation between binary strings of phylogenetic profiles.

$$MI(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{5}$$

Here $H(X)$ and $H(X|Y)$ are the entropy and conditional entropy of X.

Substituting expressions for the entropy and conditional entropy into Eqn. (5) and manipulating the sums gives the standard expression for mutual information:

$$MI(X, Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

The expression on the right can also be recognized as the relative entropy (*Kullback - Leibler* distance) between the joint distribution $p(X, Y)$ and marginal distributions $p(X)$ and $p(Y)$.

$$D(p(X, Y)||p(X)p(Y)) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{6}$$

The result is directly generalizable for M strings. In particular the mutual information $I(S_1, S_2, \ldots, S_M)$ is

$$I(\overrightarrow{S}) = D\left(p(S_1, S_2, \ldots, S_M)||\prod_{i=1}^{M} p(S_i)\right) = \sum_{s_1, s_2, \ldots, s_M} p(s_1, s_2, \ldots, s_M) \log \frac{p(s_1, s_2 \ldots, s_M)}{\prod_{i=1}^{M} p(s_i)} \tag{7}$$

which for triplets is

$$I(X, Y, Z) = D(p(X, Y, Z)||p(X)p(Y)p(Z)) = \sum_x \sum_y \sum_z p(x, y, z) \log \frac{p(x, y, z)}{p(x)p(y)p(z)}. \tag{8}$$

To show that Eqn. (4) approaches Eqn. (8) asymptotically, we write the joint probabilities in terms of x/N, y/N, z/N, N_{xy}/N, N_{yz}/N, N_{xz}/N, N_{xyz}/N. In particular

$$\begin{aligned}
p(1,1,1) &= p_{xyz} = \frac{N_{xyz}}{N}, \\
p(1,1,0) &= p_{xy} - p_{xyz} = \frac{N_{xy} - N_{xyz}}{N}, \\
p(1,0,1) &= p_{xz} - p_{xyz} = \frac{N_{xz} - N_{xyz}}{N}, \\
p(0,1,1) &= p_{yz} - p_{xyz} = \frac{N_{yz} - N_{xyz}}{N}, \\
p(1,0,0) &= p_x - p_{xy} - p_{xz} + p_{xyz} = \frac{N_x - N_{xy} - N_{xz} + N_{xyz}}{N}, \\
p(0,1,0) &= p_y - p_{xy} - p_{yz} + p_{xyz} = \frac{N_y - N_{xy} - N_{yz} + N_{xyz}}{N}, \\
p(0,0,1) &= p_z - p_{yz} - p_{xz} + p_{xyz} = \frac{N_z - N_{yz} - N_{xz} + N_{xyz}}{N}, \\
p(0,0,0) &= 1 - p_x - p_y - p_z + p_{xy} + p_{yz} + p_{xz} - p_{xyz} = \frac{N - x - y - z + N_{xy} + N_{yz} + N_{xz} - N_{xyz}}{N},
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
p_x &\equiv p(X=1) = \frac{x}{N}, \qquad p(X=0) \equiv 1 - p_x = 1 - \frac{x}{N}, \\
p_y &\equiv p(Y=1) = \frac{y}{N}, \qquad p(Y=0) \equiv 1 - p_y = 1 - \frac{y}{N}, \\
p_z &\equiv p(Z=1) = \frac{z}{N}, \qquad p(Z=0) \equiv 1 - p_z = 1 - \frac{z}{N}, \\
p_{xy} &\equiv p(X=1, Y=1) = \frac{N_{xy}}{N}, \\
p_{yz} &\equiv p(Y=1, Z=1) = \frac{N_{yz}}{N}, \\
p_{xz} &\equiv p(X=1, Z=1) = \frac{N_{xz}}{N}, \\
p_{xyz} &\equiv p(X=1, Y=1, Z=1) = \frac{N_{xyz}}{N}.
\end{aligned} \tag{10}$$

The limiting form of Eqn. (4) can be obtained using Sterling's approximation

$$n! \approx \sqrt{2\pi n} n^n e^{-n}$$

for all the factorials. Then

$$
\begin{aligned}
&-\lim_{N\to\infty}\frac{1}{N}\log P(N_{xy},N_{yz},N_{xz},N_{xyz}|N,x,y,z)\\
&=(-p_x)\log p_x-(1-p_x)\log(1-p_x)\\
&\quad+(-p_y)\log p_y-(1-p_y)\log(1-p_y)\\
&\quad+(-p_z)\log p_z-(1-p_z)\log(1-p_z)\\
&\quad+(p_{xy}-p_{xyz})\log(p_{xy}-p_{xyz})\\
&\quad+(p_{xz}-p_{xyz})\log(p_{xz}-p_{xyz})\\
&\quad+(p_{yz}-p_{xyz})\log(p_{yz}-p_{xyz})\\
&\quad+(p_x-p_{xy}-p_{xz}+p_{xyz})\log(p_x-p_{xy}-p_{xz}+p_{xyz})\\
&\quad+(p_y-p_{xy}-p_{yz}+p_{xyz})\log(p_y-p_{xy}-p_{yz}+p_{xyz})\\
&\quad+(p_z-p_{xz}-p_{yz}+p_{xyz})\log(p_z-p_{xz}-p_{yz}+p_{xyz})\\
&\quad+(1-p_x-p_y-p_z+p_{xy}+p_{xz}+p_{yz}-p_{xyz})\log(1-p_x-p_y-p_z+p_{xy}+p_{xz}+p_{yz}-p_{xyz})\\
&=H(X)+H(Y)+H(Z)-H(X,Y,Z)=I(X,Y,Z)
\end{aligned}
$$

Hence,

$$P\sim 2^{-N\times I(X,Y,Z)}=2^{-N\times I(\overrightarrow{S})} \tag{11}$$

Define a general metric of correlation

$$C\equiv-\frac{1}{N}\log_2 P \qquad 0\le C\le 1 \tag{12}$$

The upper bound on C is obtained at $P=2^{-N}$, the lower bound at $P=1$. The *MI* approximation can be safely used when N, x, y, z all exceed 7 and N_{xyz} less than 52.

The generalization of Eqn. (11) for an arbitrary number of strings is direct, and involves utilizing Eqn. (7) for the mutual information.

2.3 Relational Logic and Higher Order Organization

In this section, we briefly discuss the identification of overrepresented triplets, each of whose three pairs has a low correlation, and the relational logic of the triplets so identified.

Identification of Statistically Significant Triplets

To identify the statistically significant triplets from the 4873 orthologous groups in the *COG* database [15], we apply our probability metric (Eqn. (12)) to all triplets that do not contain a protein pair with pair-wise correlation (C) greater than 0.40. This pre-screen step not only confines our set to triplets whose subsets cannot be identified using the conventional profiling method but also substantially decreases the computation time needed. We then filter out those triplets with a correlation less than 0.65. Table 1 shows a list of pair-wise probability based correlations between 5 *COG*s as an example. The resulting matrix is symmetric, and of the 10 non redundant pairs only C-D (yellow) exceeds the threshold, with pairs D-E and C-E (green) marginal. Note that using the traditional profiling method, if only C were annotated, D could be inferred with a function, and perhaps E, but not A and B since neither A nor B is correlated with other pathway proteins here. The phylogenetic profiles of the five *COGs* are shown in Figure 1. Table 2 lists all possible triplets between these five orthologs. Seven out of the 10 have a correlation greater than 0.65; A B and C now appear as part of a very

highly correlated triplet, and the triplet C-D-E is unambiguously correlated ($C = 0.96$). However, 2 triplets (B-C-D and C-D-E, in gray) contain C-D and should be filtered out. The remaining triplets are subject to further analysis, which is discussed in detail in the next section.

Table 1: Pair-wise C of five *COG*s.

	Pair-wise C				
A(COG1685)		0.27	0.05	0.05	0.04
B(COG0703)	0.27		0.29	0.29	0.18
C(COG0128)	0.05	0.29		0.56	0.39
D(COG0082)	0.05	0.29	0.56		0.39
E(COG1605)	0.04	0.18	0.39	0.39	
	A	B	C	D	E

Table 2: Probability (C) of triplets from Table 1.

Triplets	C
A,B,C	0.85
A,B,D	0.85
A,B,E	0.69
A,C,D	0.61
A,C,E	0.46
B,C,D	0.86
B,C,E	0.70
C,D,E	0.96
A,D,E	0.46
B,D,E	0.70

Analysis of Logical Relationships of the Statistically Significant Triplets

There are in general 16 possible logical relationships between an arbitrary set of 3 profiles: AND, OR, XOR, NEGATE AND *etc.* Figure 2 shows two examples; one being protein C is present only if A or B is present, but not both A and B (exclusive OR); the other being C is present only if either A or B or both are present (inclusive OR).

Determination of logical relationships among groups of strings has been studied previously. Recently Mellor and DeLisi [10] have determined tens of thousands of significant logical relations between Transcription factors (TFs) and targets by combining gene expression data with protein-DNA protein interaction data, using a new entropy minimization method to determine the set of experiments under which the relations hold. Bowers *et al.* [2] have analyzed logical relations in general based on phylogenetic profiling using a mutual information method applied to all triplets. With significant triplets determined using the method in the previous section, logical relations can be obtained relatively rapidly using a simple Hamming distance metric. A preliminary test using the triplets identified in the previous section finds the same logical relationships obtained by Bowers.

Going beyond triplets, one can also examine the logical relations for quadruplets, e.g. E is present if any pair drawn from the three genes A, B, C is present. Finding logical relations with high reliability for groups of size greater than size 3 will require algorithmic advances since computational time is a

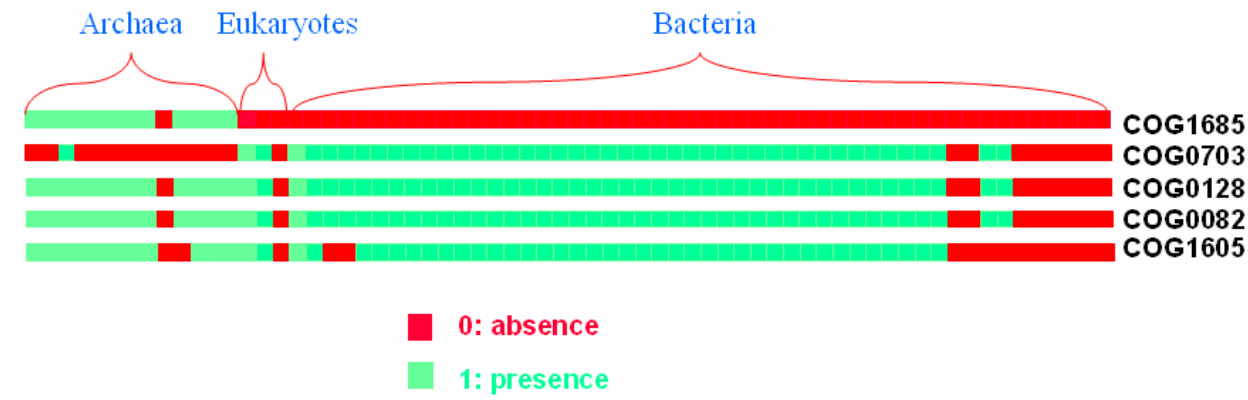

Figure 1: Phylogenetic profiles of the five *COG*s in Table 1 and 2.

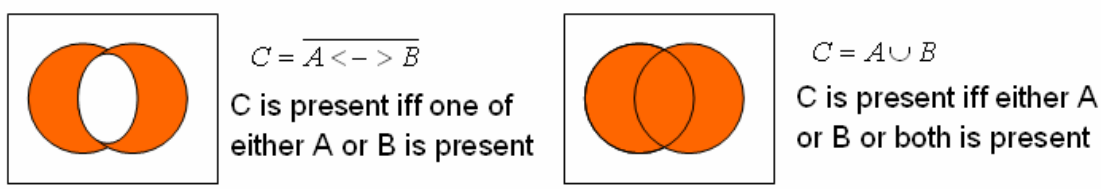

Figure 2: Two examples of logical relationships between 3 profiles.

strong non linear function of group size for exhaustive searches of the kind described here, but will be necessary if we are to fully decipher the logical relationships governing cell behavior.

3 Conclusions

Statistical metrics are developed to identify higher order protein associations in complex cellular networks. Analysis of phylogenetic profile triplets can be readily extended to larger sets of proteins, and to gene expression profile correlations, which in its simplest form is mathematically identical. We envision the general applicability of the method to the analysis of higher order organization of protein networks, and the potential discovery of interesting network motifs that elucidate the biological functions.

References

[1] Aravind, L., Guilt by association: Contextual information in genome analysis, *Genome Res.*, 8:1074–1077, 2000.

[2] Bowers, P.M., Cokus, S.J., Eisenberg, D., and Yeates, T.O., Use of logic relationships to decipher protein network organization, *Science*, 306(5705):2246–2249, 2004.

[3] Date, S.V. and Marcotte, E.M., Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages, *Nat. Biotechnol.*, 21(9):1055–1062, 2003.

[4] Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Bouzoukis, C.A., Protein interaction maps for complete genomes based on gene fusion events, *Nature*, 402:86–90, 2001.

[5] Huynen, M.A. and Bork, P., Measuring genome evolution, *Proc. Natl. Acad. Sci. USA*, 95:5849–5856, 1998.

[6] Huynen, M.A. and Snel, B., Gene and context: Integrative approaches to genome analysis, *Adv. Protein Chem.*, 54:345–379, 2000.

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

[8] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D., Detecting protein function and protein-protein interactions from genome sequences, *Science*, 285:751–753, 1999.

[9] McEliece, R.J., *The Theory of Information and Coding*, Cambridge University Press, 2000.

[10] Mellor, J.C. and DeLisi, C., Toward a high-resolution map of eukaryotic transcription logic, (*submitted*), 2005.

[11] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U., Superfamilies of evolved and designed networks, *Science*, 303(5663):1538–1542, 2004.

[12] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: Simple building blocks of complex networks, *Science*, 298(5594):824–827, 2002.

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

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

[15] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V., The COG database: New developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 29:22–28, 2001.

[16] Wu, J., Kasif, S., and DeLisi, C., Identification of functional links between genes using phylogenetic profiles, *Bioinformatics*, 19(12):1524–1530, 2003.

[17] Wu, J., Hu, Z., and DeLisi, C., Gene annotation and network inferences using phylogenetic profiles, (*submitted*), 2005.

[18] Yanai, I., Derti, A., and DeLisi, C., 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(14):7940–7945, 2001.

[19] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A., Global protein function prediction from protein-protein interaction networks, *Nat. Biotechnol.*, 21(6):697–700, 2003.

[20] The Gene Ontology Consortium, Creating the gene ontology resource: Design and implementation, *Genome Res.*, 11(8):1425–1433, 2001.