

Databases and ontologies

## Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications

Haiyuan Yu<sup>1,4,5,†</sup>, Ronald Jansen<sup>1,†</sup>, Gustavo Stolovitzky<sup>6</sup> and Mark Gerstein<sup>1,2,3,\*</sup><sup>1</sup>Department of Molecular Biophysics & Biochemistry, <sup>2</sup>Department of Computer Science, <sup>3</sup>Program in Computational Biology and Bioinformatics, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520,<sup>4</sup>Department of Genetics, Harvard University, <sup>5</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, 1 Jimmy Fund Way, Boston, MA 02115 and <sup>6</sup>IBM Computational Biology Center, T.J. Watson Research Center, PO Box 704, Yorktown Heights, NY 10598

Received on December 31, 2006; revised on May 2, 2007; accepted on May 22, 2007

Advance Access publication May 31, 2007

Associate Editor: Alex Bateman

### ABSTRACT

**Motivation:** Many classifications of protein function such as Gene Ontology (GO) are organized in directed acyclic graph (DAG) structures. In these classifications, the proteins are terminal leaf nodes; the categories ‘above’ them are functional annotations at various levels of specialization and the computation of a numerical measure of relatedness between two arbitrary proteins is an important proteomics problem. Moreover, analogous problems are important in other contexts in large-scale information organization—e.g. the Wikipedia online encyclopedia and the Yahoo and DMOZ web page classification schemes.

**Results:** Here we develop a simple probabilistic approach for computing this relatedness quantity, which we call the total ancestry method. Our measure is based on counting the number of leaf nodes that share exactly the same set of ‘higher up’ category nodes in comparison to the total number of classified pairs (i.e. the chance for the same total ancestry). We show such a measure is associated with a power-law distribution, allowing for the quick assessment of the statistical significance of shared functional annotations. We formally compare it with other quantitative functional similarity measures (such as, shortest path within a DAG, lowest common ancestor shared and Azuaje’s information-theoretic similarity) and provide concrete metrics to assess differences. Finally, we provide a practical implementation for our total ancestry measure for GO and the MIPS functional catalog and give two applications of it in specific functional genomics contexts.

**Availability:** The implementations and results are available through our supplementary website at: <http://gersteinlab.org/proj/funcsims>

**Contact:** mark.gerstein@yale.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

### 1 INTRODUCTION

The systematic classification of protein function is of great importance in functional genomics, as it organizes our thinking

about the biological roles of proteins. Various databases collecting functional information from the literature have been developed to meet this need. Among the most important classification schemes are GO, MIPS, GenProtEC and Ecocyc (Harris *et al.*, 2004; Karp *et al.*, 1996; Mewes *et al.*, 2006; Riley and Space, 1996). Whereas GO represents a classification across organisms, MIPS, GenProtEC and Ecocyc focus on well-studied individual organisms (e.g. *Saccharomyces cerevisiae* and *Escherichia coli*).

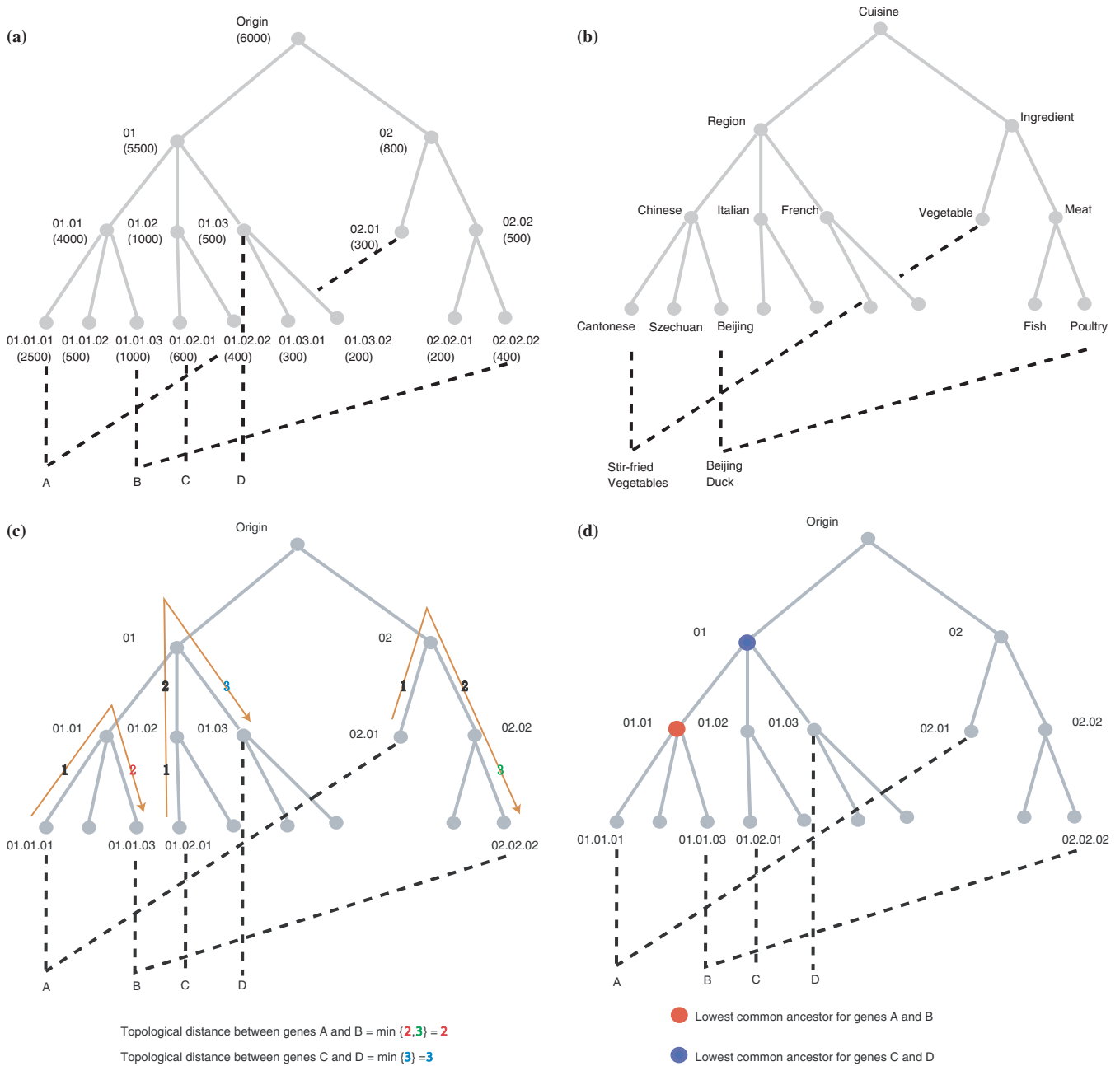
In many contexts, it is important to compare proteins within the framework of a functional classification and to define a proper measure of the degree of functional similarity between them. For instance, in developing sequence and structure comparison methods and annotation transfer schemes, similarity in sequence and structure needs to be correlated with functional similarity (Hegyí and Gerstein, 2001; Orengo *et al.*, 1999; Wilson *et al.*, 2000). Moreover, there are many interesting applications for functional similarity measures in the automatic analysis of high-throughput genomics experiments, e.g. the analysis of mRNA expression profiles, protein–protein interactions or chIP-chip experiments (Horak *et al.*, 2002; Lee *et al.*, 2002).

#### 1.1 The problem of defining functional similarity

It is not obvious how to best compute functional similarity from the existing classification schemes. On a most simple level, researchers often need to ask whether two proteins are members of the same functional class; this can essentially be described as a Boolean functional similarity measure (giving the answer ‘yes’ or ‘no’). There are obviously several problems with this approach. On the one hand, it is desirable to work with systems that contain many different functional classes because detailed information on protein function allows for a high differentiation between proteins. On the other hand, there are limits to the level of detail in many practical applications. This is, e.g. a problem in the automatic classification of protein function from DNA microarray data. If there are too few proteins in a functional class, no reliable statistics can be gathered on it, confounding reasonable statistics (Brown *et al.*, 2000; Mateos *et al.*, 2002; Wu *et al.*, 2002). More importantly,

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



**Fig. 1.** Illustration of different functional similarity measures discussed in the article using a hypothetical tree. (a) Illustration of the example tree structure. This tree structure is adapted from MIPS. Each node represents a functional class, which is named as in the MIPS database. The number below the name in the parentheses shows the size of each functional class. A, B, C and D are different genes belong to distinct classes. Genes A and B each have two different functions. The tree structures in all panels have exactly the same configuration as this one. (b) An example tree structure adapted from the Wikipedia online encyclopedia. Each node represents a cuisine class. Two specific dishes belonging to different cuisine classes are shown at the bottom. Please note that this tree is not complete and might be slightly different from the real one in Wikipedia. (c) Calculation of topological distance. The distances between A and B at two parts of the tree are 2 (red) and 3 (green), respectively. The topological distance of AB is therefore 2. (d) Illustration of finding the lowest common ancestor for protein pairs. (e) Calculation of information-theoretic similarity between genes A and B. The semantic similarity between nodes 01.01.01 and 01.01.03 and that between 02.01 and 02.02 are the information contents of nodes 01.01 (red) and 02 (purple), respectively. (f) Calculation of total ancestry similarity between genes A and B. The genes in the box under each node belong to that functional category. These common edges and ancestors form a specific pattern, which can be represented by only the set of the lowest common ancestors (i.e. in this case, nodes ‘01.01’ and ‘02’, highlighted as red). The idea is to calculate the frequency of a protein pair having this pattern among all possible pairs in the database. This frequency is a natural similarity metric. The lower the frequency, the higher the similarity. There are seven gene pairs sharing the same set of lowest common ancestors as AB, which are connected by thin lines in the graph.

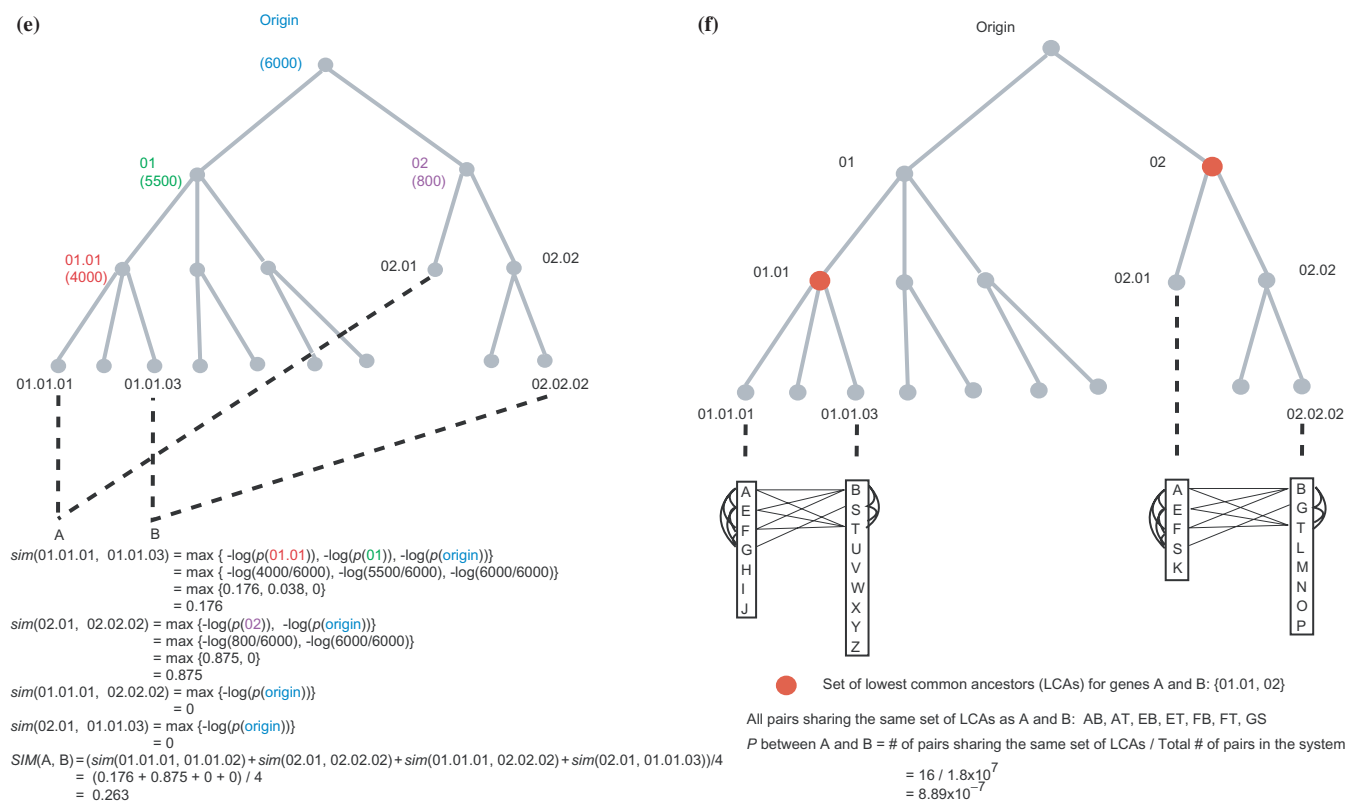


Fig. 1. Continued.

a functional similarity measure is much more desirable to be a continuous quantity that varies in relation to the degree of common functional properties between proteins.

Most classification systems have the structure of a tree (e.g. MIPS; see Fig. 1a) or a directed acyclic graph (DAG) (e.g. GO)—please note that a tree is a special case of a DAG. One could simply cut off the classification graph at some arbitrary level to obtain classes of sufficient size for a particular application. But it is often not clear how to do this in a systematic way. For instance, the DAG structure of GO does not explicitly contain discrete levels of classes. But even for classification systems with tree structures (such as MIPS) it is not clear whether two unrelated classes are at the same level of biological detail just because they are in the same level of the tree hierarchy. The edges between classes may have different weights, but these are not explicitly defined. Furthermore, different parts of a functional classification system may have different degrees of detail, because (1) there are varying amounts of knowledge on specific biological processes; (2) intrinsically, some biological processes have more detail and therefore have more depth in the classification system. In this case, a certain level at one part of a functional classification system clearly reflects different degrees of functional similarity than that at another part.

An additional problem is that many proteins are members of multiple functional classes. This is because (1) some proteins truly have multiple functions; (2) a protein often needs more than one label to describe its function.

An example of the latter is Hac1p, which is a transcription factor that regulates membrane biosynthesis (Cox and Walter, 1996), therefore, Hac1p is associated with two parts (transcription and metabolism) of the MIPS functional classification system. How does one define functional similarity in this context?

In the following, we discuss three non-Boolean measures of similarity between protein pairs (a topological, an information theoretic and a total ancestry measure) that address some of the problems we mentioned. These problems arise in many other contexts with a DAG structure. For example, the Wikipedia online encyclopedia categorizes various terms (see Fig. 1b); Yahoo and DMOZ web page classification schemes categorize web pages; and, the US patent classification system categorizes inventions. We believe that these methods measuring similarities will be applicable to items within these DAG structures, as well. The results can be very helpful for people using these systems: they make it much easier to find relevant terms, web pages and patents, which could significantly expedite the process of searching through these huge systems.

## 2 NON-BOOLEAN MEASURES OF FUNCTIONAL SIMILARITY

### 2.1 Topological similarity

**2.1.1 Topological distance** A more advanced measure of functional similarity than the Boolean measure is to compute the number of edges one has to traverse in going from one

node to another. In case of multiple paths between two nodes (as in a DAG), one can simply take the minimum or average of the different path lengths. The minimum path is appealing as it yields an actual mathematical distance that verifies the three distance axioms in metric spaces. In graph theory, this measure is known as the shortest path length between nodes (see Fig. 1c). (As this measure takes into account the topology of the DAG, we shall call this metric the ‘topological distance’.) There are a number of known algorithms to compute this distance such as the Floyd–Warshall algorithm or the Dijkstra algorithm (Aho *et al.*, 1974; Gusfield, 1997; Swamy and Thulasiraman, 1981). The topological distance is defined for classes rather than proteins. For proteins with multiple classes, one can again simply take the minimum of the set of lengths, yielding an actual distance.

As mentioned above, one might want to weight edges differently. The Dijkstra algorithm also allows the computation of minimum path lengths in case the classification graph has weighted edges. However, given a classification system, it is not clear how to actually go about quantifying the edge weights.

**2.1.2 Lowest common ancestor** The topological distance (as defined above) does not record at what level in a classification hierarchy two proteins share their first common ancestor class although this information is clearly useful in characterizing their similarity. It should be noted that the distance between the root node of a classification system and the lowest common ancestor of two proteins yields another distance metric (see Fig. 1d). In general, this distance correlates with the significance of the similarity between the two proteins. Again, we note that two proteins may be members of multiple functional classes and therefore may have multiple lowest common ancestors. In this case, one can measure the distances between all lowest common ancestors and the root node and simply take the maximal (or minimal) distance.

## 2.2 Azuaje’s information theoretic similarity (ITS)

Semantic and information theoretic approaches provide alternative methods for quantifying similarities between objects in a classification DAG. Such approaches are based on counting the frequency of how many objects each node in the classification graph contains, which, in turn, allows quantifying the probability or information theoretic entropy that an object is member of such a node. Examples of such information theoretic definitions are Resnik’s method for measuring semantic similarity and Lin’s similarity measure (Guo *et al.*, 2006; Lin, 1998; Resnik, 1995; Wang and Azuaje, 2004). The basic principles of these methods are the same. Here, we will use one method to illustrate the use of information theoretic entropy to measure the functional similarity between protein pairs, which we call Azuaje’s ITS method (Wang and Azuaje, 2004):

first, the similarity between two functional terms  $c_i$  and  $c_j$  is calculated by:

$$\text{sim}(c_i, c_j) = \max_{c \in S(c_i, c_j)} [-\log(p(c))]$$

where  $S(c_i, c_j)$  is the set of parent terms shared by both  $c_i$  and  $c_j$ ,  $p(c)$  describes the probability of finding a child of a functional term  $c$  in the whole database. It is calculated as the ratio of the number of objects (i.e. proteins) included in term  $c$  over the total number of objects in the classification system. In information theory,  $-\log(p(c))$  measures the information content of  $c$ . The value of this similarity metric ranges between 0 and infinity.

Next, let us consider a pair of gene products,  $g_i$  and  $g_j$ , which are annotated by a set of terms  $A_i$  and  $A_j$ , respectively.  $A_i$  and  $A_j$  comprise  $m$  and  $n$  terms, correspondingly. The semantic similarity between these two proteins,  $\text{SIM}(g_i, g_j)$ , may be defined as the average inter-set similarity between terms from  $A_i$  and  $A_j$ :

$$\text{SIM}(g_i, g_j) = \frac{1}{m \times n} \times \sum_{c_k \in A_i, C_p \in A_j} \text{sim}(C_k, C_p).$$

Figure 1e uses a specific example to illustrate how the semantic similarity between protein pairs is calculated.

## 2.3 Total ancestry similarity

**2.3.1 Probability formalism** The notion of the lowest common ancestor and the probabilistic nature of the information theoretic measures lead us to another definition of functional similarity that combines several of the qualities of the other measures. One can combine elements of the topological distance and the lowest common ancestor distance within a statistical framework to arrive at a probabilistic level of similarity. Obviously, a pair of proteins should be very similar if they belong to a functional group (i.e. the lowest common ancestor node) with few proteins, whereas the similarity will be less significant if many proteins belong to that group.

Given two proteins that share a specific set of lowest common ancestor nodes in the classification structure, one can count the total number of protein pairs  $n$  that also have the exact same set of lowest common ancestors (see Fig. 1f). This number is expected to be low for proteins that share a very detailed functional description, but very high for proteins that have no function in common. For instance, if a functional class contains only the two proteins of interest, then the count would yield  $n=1$ . On the other hand, if the root node is the lowest common ancestor of two proteins,  $n$  is the number of protein pairs contained in the classification.

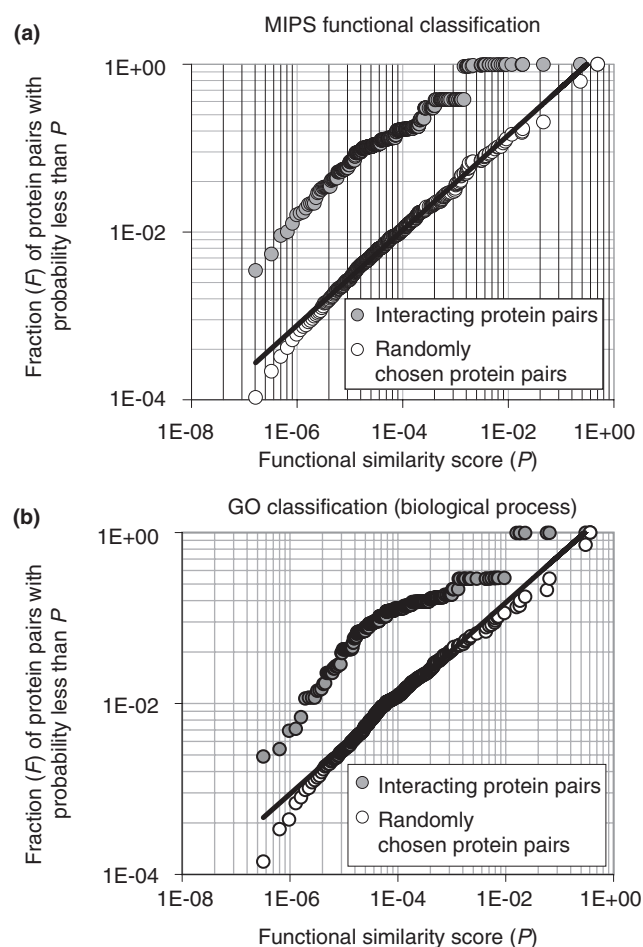
The count  $n$  can be normalized to represent a statistical or probability-like measure  $P$ :

$$P = n/N$$

where  $N$  is the total number of protein pairs in the classification. (For instance, MIPS classifies 4779 of the 6000 *S.cerevisiae* proteins, thus yielding  $N=4779 \times 4778/2$  protein pairs; see Fig. 1f). Please note that we can trivially convert this probability into an information content (to make it more comparable to Azuaje’s ITS method) by simply computing  $-\log P$ .



**2.3.2 P-value calculation** Figure 2 shows that the cumulative density distributions ( $F$ ) of the (normalized) functional similarities ( $P$ ) based on both the GO biological process annotation and the MIPS functional catalog approximately follow power law distributions (see Fig. 2; for MIPS,  $F=1.9438P^{0.5668}$ ; for GO,  $F=1.9945P^{0.5598}$ ; we estimated the parameters of the power law distributions by performing a simple linear regression on the log-log scaled graph). Interestingly, the exponents of the two power laws are virtually the same for both the MIPS and GO annotation. In fact, it can be empirically shown that the cumulative density distribution of any DAG structure follows a power law distribution (see Supplementary Fig. 4). This feature is useful for practically assessing the statistical significance of functional similarity measurements. For a given functional annotation scheme, the parameters of the power law distribution can be relatively easily determined from a random sample of functional similarity values and subsequent regression.



**Fig. 2.** The cumulative density distributions ( $F$ ) of the (normalized) functional similarities ( $P$ ) based on both the GO biological process annotation and the MIPS functional catalog approximately follow power-law distributions. We define ‘interacting’ proteins as those that are members of the same protein complex (according to the MIPS complexes catalog).

Knowledge of the power law parameters then allows the computation of  $P$ -values for observed functional similarity measurements.

### 3 CONSISTENCY OF SIMILARITY MEASURES

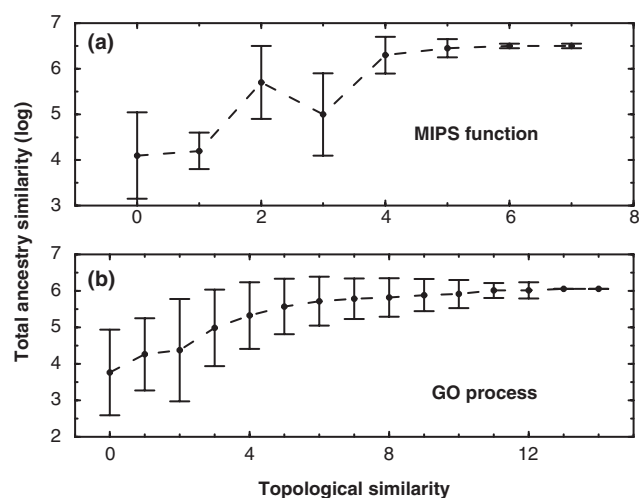
There are several ways to assess the consistency of these functional similarity measures.

#### 3.1 Comparing different measures of functional similarity

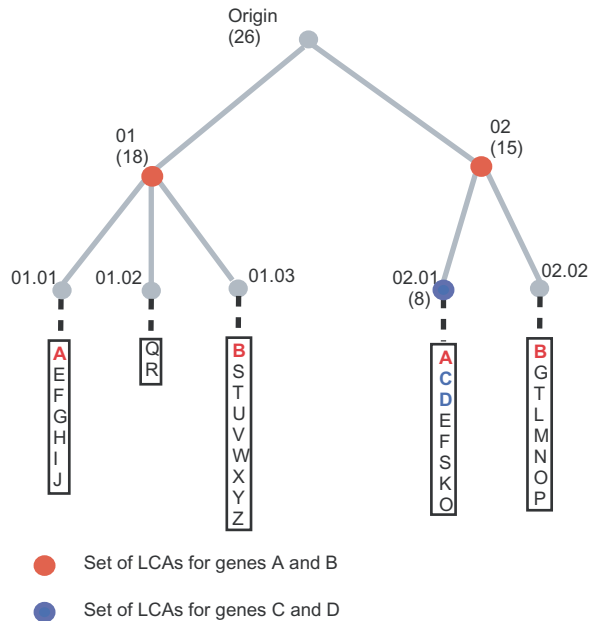
**3.1.1 Topological similarity versus total ancestry similarity** Figure 3 shows that the topological and total ancestry similarity measures are positively correlated, as expected. However, as mentioned above, the topological similarity does not measure the distance of the lowest common ancestor to the root of the classification graph. The total ancestry similarity measures can aid in finding cases where the lowest common ancestor is very near the root node in the classification.

For instance, there are 39 protein pairs in the GO biological process classification that have a topological distance of zero, but for which the total ancestry measure is greater than  $1/3$ . In each of these pairs, the two proteins have one functional class in common, but these tend to be functional classes of a very general type. For instance, 26 of the 39 pairs are members of the general class ‘metabolism’.

**3.1.2 Azuaje’s ITS similarity versus total ancestry similarity** Even though Azuaje’s ITS method provides a semi-probabilistic measure of functional similarity, it is not equivalent to the total ancestry similarity that we proposed (see Fig. 4): it looks for the smallest functional term (i.e. the one



**Fig. 3.** Comparison of the topological and total ancestry similarity measures. (a) The results for MIPS and (b) for GO (biological process). The topological measure is shown on the abscissa, the logarithm of the count  $n$  as a measure of the ancestry similarity on the ordinate (average and SD bars). Note there are only few cases in the MIPS classification for which the topological distance is 1 or 3, which is a consequence of its tree structure.



$$SIM(A, B) = (-\log(18/26) - \log(15/26)) / 4 = (0.160 + 0.239) / 4 = 0.10$$

$$SIM(C, D) = -\log(8/26) = 0.512 \gg SIM(A, B) \rightarrow CD \text{ are more similar}$$

$$P(A, B) = 16 / 325 = 0.05$$

$$P(C, D) = 28 / 325 = 0.09 \gg P(A, B) \rightarrow AB \text{ are more similar}$$

**Fig. 4.** Illustration of the differences between the information-theoretic similarity measure and the total ancestry measure. Genes C and D share the lowest common ancestor 02.01, which includes eight genes. Genes A and B share two lowest common ancestors in nodes 01 and 02, which are both one level higher in the classification system than node 02.01. Therefore, the information-theoretic similarity measure (as well as the topological distance measure and the lowest common ancestor measure) will show that CD are more similar than AB. However, in this schematic example, there are only seven pairs sharing the same set of lowest common ancestors as AB. Therefore, the total ancestry measure will show that AB are actually more similar than CD. A really biological example is proteins Doa4p and Pre1p, two subunits of the 26S proteasome.

contains fewest proteins) that two proteins share. However, a pair of proteins can share only large functional terms; but if they share many such terms, they should still be considered as functionally similar. Because all of the classification systems are far from complete right now, many proteins only have rough functional descriptions, i.e. they are only associated with large terms. Therefore, how to take these large functional terms into consideration is of great importance to a good functional similarity measure. The total ancestry measure solves this problem naturally. It does not look at individual terms, but rather examines the whole set of the shared terms. Even though each individual term may contain a lot of proteins, it is much less likely that many of these proteins will share a number of different terms. For example, Doa4p and Pre1p are two subunits of the 26S proteasome based on the MIPS complex catalog. Clearly, their functions should be extremely similar as they form a stable complex. However, since they only belong

to a number of large functional terms in GO, their functional similarity determined by the information-theoretic method is not significant (less than half of all the protein pairs). On the contrary, the total ancestry method ranks this pair among the top 5% most functionally similar pairs, which makes more biological sense.

Interestingly, there are only 278 protein pairs that are among the top 5% most functionally similar pairs measured by Azuaje's ITS method, but are not significant (less than half of all the protein pairs) according to the total ancestry method. None of the known interacting pairs within the MIPS complexes is included in these 278 pairs. Consequently, we can see that the similarity assignment by Azuaje's ITS method is less optimal, compared with our total ancestry method. Moreover, the distribution of the ITS scores of these pairs is clearly concentrated towards the lower boundary, indicating that these pairs can be easily pushed below the cutoff with the insertion of many more functionally similar pairs in front of them by the total ancestry method (see Supplementary Fig. 1).

Furthermore, all the pairs that are determined as similar by the information-theoretic measures will have a good similarity score by our total ancestry method. The reason is that, if two proteins share a small functional term, the number of proteins sharing the whole set of the terms that these two proteins share will be even smaller. As a result, the scores by the two methods are also positively correlated (Pearson correlation coefficient = 0.54;  $P$ -value  $< 10^{-15}$ ).

### 3.2 Consistency under reorganizations of the classification system

As mentioned above, the classification systems are constantly updated with new information and are sometimes reorganized.

The topological distance measure tends to be very sensitive to any change in the structure of a classification graph. However, ideally, a functional similarity measure would be robust with respect to some of these changes. For instance, as additional classes are introduced into the graph to accommodate newly characterized proteins, the information-theoretic and total ancestry similarity measures tend to be more robust against such changes, and leads to only minor perturbations to the relationships between already classified proteins. For example, the information content of a functional class only takes into consideration the size of this class; it is not related to the organization of its descents. If a new descent class is introduced or the relationships between its descents are re-arranged, its information content is left unchanged, as long as the total number of genes in this class remains the same. Similar reasons also apply to the total ancestry measure.

Overall, an advantage of Azuaje's ITS and the total ancestry similarity measures is that they could practically distinguish between reorganization events that should change functional similarity and those that should leave it unaffected.

## 4 IMPLEMENTATION OF SIMILARITY MEASURES

We have implemented all of the similarity measures discussed above and calculated different functional similarity scores for

both GO and MIPS functional classification systems. The implementations and results are available through our supplementary website at: <http://www.gersteinlab.org/proj/funccsim>.

## 5 APPLICATIONS OF FUNCTIONAL SIMILARITY MEASURES

Well-defined functional similarity measures could facilitate the genome-wide analyses in many applications. Here, we briefly outline two examples.

### 5.1 Analysis of mRNA expression data

The automatic classification of proteins into functional classes given their expression profiles in cDNA microarray using different machine-learning algorithms has had only limited success so far. In this context, it has been observed that machine-learning algorithms often assign proteins as false positives that are actually biologically related to the class of true positives (Mateos *et al.*, 2002).

In Figure 5a we show an example of classifying proteins belonging to the class ‘TCA cycle’ in MIPS, using a neural network machine-learning algorithm [details explained in Mateos *et al.* (Mateos *et al.*, 2002)]. We performed several iterations of machine-learning steps. After each learning step, we determined which proteins were correctly or incorrectly classified by cross-validating the results with the MIPS data. Then we added those proteins that appeared as false positives in the previous round as belonging to the original class (‘TCA cycle’) in the next round. We found that such an iterative procedure converged to a relatively coherent set of proteins, which are functionally related to the original TCA cycle proteins. The functional similarity measures allow us to quantify the relationship of these proteins to the original class (see Fig. 5a), rather than classifying them as false positives in a Boolean way (‘true’ or ‘false’). Please note that we did not develop a new learning algorithm. What we showed here is that the traditional learning algorithms can benefit from the non-Boolean values of the functional similarity that our method produces.

Figure 5b shows the general relationship between distance in the gene expression space and the distance between genes in the MIPS functional catalog. It shows that for small gene expression distances, the proportion of small functional distances is substantially greater than that of large functional distances. [We used the Spellman dataset (Spellman *et al.*, 1998), and a Euclidian distance in gene expression profiles between pairs of genes of the normalized vector of 79 gene expression values.] GO classification yields a similar picture (data not shown). As we increase the gene expression distance, larger functional distances are more likely. One interpretation of this is that clusters that result from Eisen plots (Eisen *et al.*, 1998) are likely to indicate functional similarity of the genes participating in the cluster. The converse is not true. We have observed that the gene expression distance between genes is rather independent of the functional distance, thus indicating that classes with similar function will be ‘far’ in gene expression space. This agrees with our earlier observation that most functional classes are poorly

learned from expression data by machine-learning algorithms (Mateos *et al.*, 2002).

Perhaps surprisingly, the total ancestry measure for MIPS and GO generally yield very different values and are only weakly correlated (Pearson correlation coefficient  $\approx 0.21$ ; see Supplementary Fig. 2). Similar results can be observed for the topological measure. This may be due to the different organization nature of MIPS and GO classification schemes (Ruepp *et al.*, 2004).

### 5.2 Prediction of protein–protein interactions/complexes

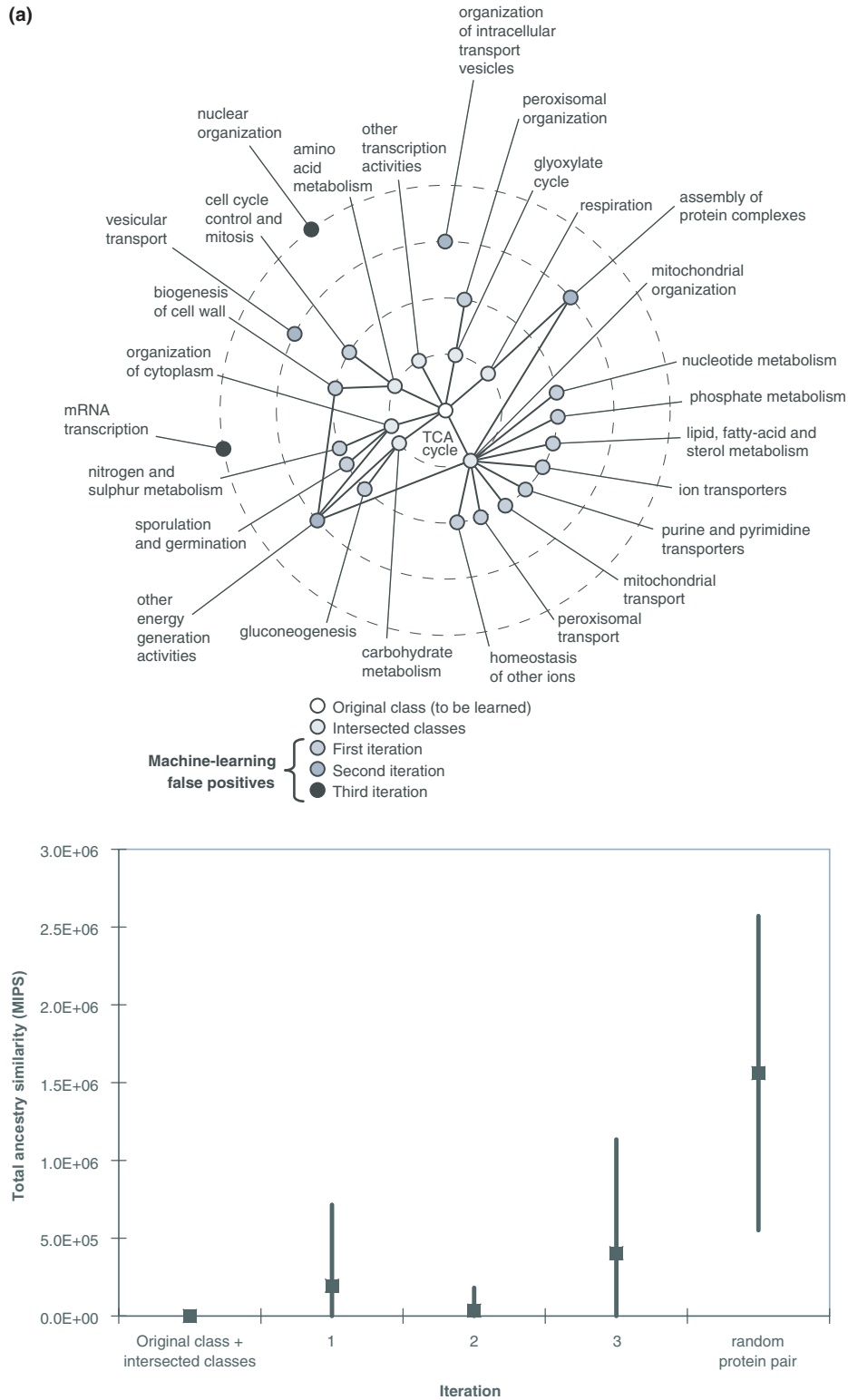
Another potential application of the functional similarity measure is the prediction of protein–protein interactions. This is possible because proteins that participate in similar biological processes are more likely to interact than those that are operating in very different or unrelated processes. It is possible to statistically predict the chance that two proteins are interacting in a complex based on their biological process annotation. Figure 2 also shows that, at all different levels of significance, interacting protein pairs are always more likely to be functionally similar than randomly chosen ones, which, to some degree, confirms the validity of our approach.

A caveat of such a prediction is that many functional annotations were derived from protein–protein interaction experiments in the first place. Thus, using such data blindly to ‘predict’ interactions may amount to circular reasoning.

However, one can explicitly consider only those functional annotations that were derived from experiments other than protein–protein interactions. It is possible to identify these from the evidence codes that indicate which data sources were used to arrive at a functional assignment and that accompany each annotation in the GO database.

Figure 6a shows the statistical distribution of known interacting and non-interacting proteins in a subset of the GO annotation that includes only evidence unrelated to protein–protein interaction experiments: the subset of the GO biological process annotation considered here is solely based on characterizations of genetic phenotypes. Figure 6a shows that there is a strong signal for protein–protein interactions in these data. For instance, protein pairs with a normalized functional similarity value between  $5 \times 10^{-7}$  and  $5 \times 10^{-6}$  are on average 45.55 times more likely to be in the same complex than randomly chosen protein pairs. For a specific example: Ssa1p is an ATPase involved in protein folding and nuclear localization signal (NLS)-directed nuclear transport. It is a member of heat shock protein 70 (HSP70) family (Kim *et al.*, 1998). Ydj1p is also a protein chaperone involved in regulation of the HSP90 and HSP70 functions (Kimura *et al.*, 1995). They have an extremely good functional similarity score of  $1 \times 10^{-7}$  ( $L = 98.7$ ), but they are not included in the MIPS complex catalog. However, previous studies have shown that Ssa1p and Ydj1p form a chaperone complex (Hon *et al.*, 2001).

Finally, we calculated the enrichment of functionally similar protein pairs among the known interacting pairs. Total 98.5% of the interacting pairs are functionally similar as measured by the total ancestry method, whereas the corresponding value is only 56.7% for Azuaje’s ITS method (P-value  $< 10^{-10}$ ).



**Fig. 5.** Relationship between gene expression distance and functional similarity. **(a)** The functional similarity of proteins recruited in successive rounds of automatic classifications of proteins belonging to the class 'TCA cycle' in MIPS, based on gene expression data by Eisen *et al.* (Eisen *et al.*, 1998). Nodes in the concentric circles indicate MIPS functional classes of false positives at different iterations. Solid lines connect classes that share common proteins. The lower graph shows the average and SD of the total ancestry functional similarity between the false positives at different iterations and the proteins in the class 'TCA cycle'. For comparison, we show the expected values of functional similarity between random protein pairs. **(b)** The average values of the topological distance (MIPS) at given distances in expression space (Spellman *et al.*, 1998).



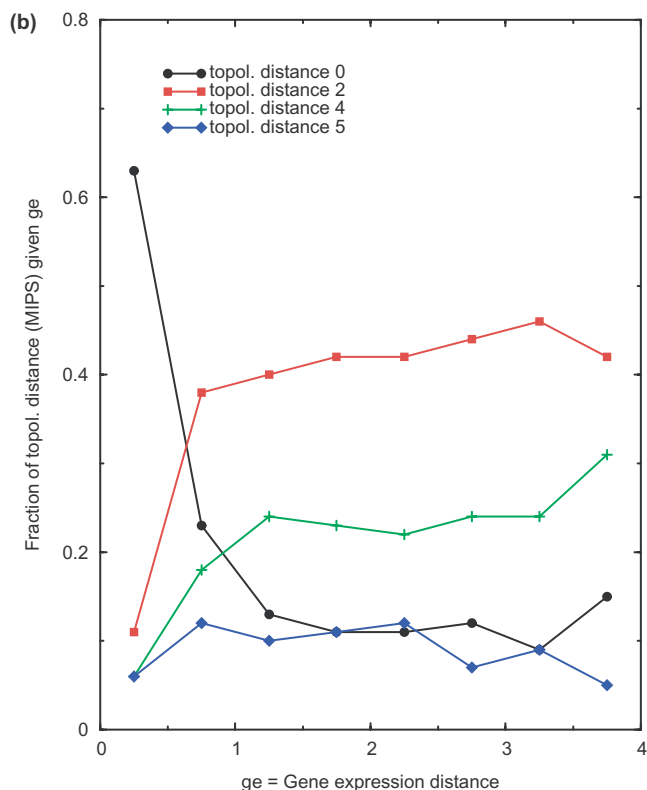


Fig. 5. Continued.

Similar results were also observed for the co-expressed pairs (see Supplementary Fig. 3). Thus, the functional similarity scores calculated by the total ancestry method relate better to protein interaction and gene co-expression, when compared to those calculated by the ITS method.

It should also be noted that similar practices (relating GO or MIPS functional relationships to predicting protein's properties, e.g. interaction, sub-cellular localization) have been carried out by a number of previous studies (Azuaje *et al.*, 2006; Lei and Dai, 2006; Wu *et al.*, 2006).

## 6 DISCUSSION

We have presented a new probabilistic measure, named the total ancestry method, for computing the functional similarity between pairs of proteins and compared it with other measures of functional distance. We explored different properties of these four measures. The total ancestry similarity naturally measures the depth of the lowest common ancestor in a classification graph and changes in response to updates of the classification system in a robust and intuitive way. We discussed potential applications in analyzing genomic data, such as mRNA expression profiles and protein-protein interactions. Furthermore, our measure can be readily applied to other non-biological contexts, e.g. the Wikipedia online encyclopedia and the Yahoo and DMOZ web page classification schemes. In

$y = \text{normalized functional similarity}$	# positives	# negatives	$P(y pos)$	$P(y neg)$	$L$
$< 5 \times 10^{-7}$	22	147	0.0325	0.0003	98.7
$5 \times 10^{-7} - 5 \times 10^{-6}$	81	1,173	0.1198	0.0026	45.5
$5 \times 10^{-6} - 5 \times 10^{-5}$	187	4,396	0.2766	0.0099	28.1
$5 \times 10^{-5} - 5 \times 10^{-4}$	168	13,242	0.2485	0.0297	8.4
$\geq 5 \times 10^{-4}$	218	426,910	0.3225	0.9575	0.3

$y = \text{normalized functional similarity}$	# positives	# negatives	$P(y pos)$	$P(y neg)$	$L$
$< 5 \times 10^{-7}$	-	-	0.0000	0.0000	NA
$5 \times 10^{-7} - 5 \times 10^{-6}$	-	-	0.0000	0.0000	NA
$5 \times 10^{-6} - 5 \times 10^{-5}$	24	61	0.1159	0.0027	42.5
$5 \times 10^{-5} - 5 \times 10^{-4}$	104	382	0.5024	0.0171	29.4
$\geq 5 \times 10^{-4}$	79	21,900	0.3816	0.9802	0.4

**Fig. 6. (a)** The statistical distribution of positive (interacting) and negative (non-interacting) protein pairs for the complete MIPS annotation. As a reference for protein pairs that are members of the same complex ('positives'), we used the MIPS complexes catalog, a list of protein complexes in *S.cerevisiae*. As a reference for protein pairs that are not interacting ('negatives'), we compiled a list of protein pairs that are present in different subcellular compartments (Jansen *et al.*, 2003). These references of positive and negative examples allow us to assess the relationship between functional similarity of proteins and their tendency to be in the same complex. The first column shows different intervals of the normalized functional similarity, and the second and third columns the number of positive and negative protein pairs that were found within these intervals. These numbers allow us to calculate the conditional probabilities in columns four and five. Finally, column six gives the likelihood ratio  $L$ , defined as the ratio of these two conditional probabilities. The posterior odds of finding a protein interaction ('posterior' = after considering the level of functional similarity) is equal to the likelihood ratio  $L$  times the prior odds of finding an interaction ('prior' = before considering functional similarity). Thus, for instance, protein pairs with a normalized functional similarity value  $< 5 \times 10^{-7}$  are on average 98.71 times more likely to be members of the same protein complex than any randomly chosen protein pair. Higher functional similarity is clearly correlated with a higher chance that two proteins are interacting. The signal for protein interactions contained in the functional similarity data is similar to that contained in the (noisy) data gained from high-throughput protein-protein interaction experiments. Although such datasets may only predict protein-protein interactions with a relatively high false positive rate, it has been shown that the combination of multiple datasets of such quality can lead to reliable predictions of protein complexes. **(b)** For comparison, the statistical distribution of positive (interacting) and negative (non-interacting) protein pairs for the complete GO annotation that includes evidence based either solely or partially on protein-protein interaction experiments. As expected, the likelihood ratios are somewhat larger than in part a for the same functional similarity ranges.

these contexts, proteins in our calculation correspond to the leaf nodes in the DAGs of interest.

One major disadvantage of our method is that it is computationally much more expensive than all other methods mentioned in the article because it examines the set of lowest common ancestors of all protein pairs in a given genome for any two proteins of interest. However, in practice, the similarity scores were calculated for the whole *S.cerevisiae*

genome in <30 min using a normal PC [Intel(R) 2.16 GHz CPU with 2 GB RAM].

Another interesting issue is that our total ancestry method focuses on the coinciding categories between two proteins, but ignores the missing ones (in fact, this is true for all current methods). For example, let's consider three proteins A, B and C (A has a function 10.03; B has functions 10.03, 01.01.01, 11.01.02, 12.03.04 and C has a function 10.02; here the functional class names follow the MIPS functional category convention—each number represents one category and a sub-category is separated from its parent category by a dot). Using our total ancestry method, AC and BC will have the same functional similarity (i.e. they all share the same functional category—category 10). One could argue that, since B has many more functions other than 10.03, it is less 'devoted' to the functional category 10; therefore, BC should be functionally less similar than AC. One way to extend our method to accommodate this is to assign a weight to each functional category that a certain protein belongs to so that the sum of all the weights equals to one. Then, instead of counting the exact number of protein pairs within a set of lowest common ancestors, one could use a weighted count to calculate the total ancestry similarity.

However, we still believe that considering only coinciding categories actually makes biological sense for two main reasons. First, as discussed earlier, some proteins need more than one label to describe their function (e.g. see the example of Hac1p described above). In this case, being in one functional category does not make the protein less 'devoted' to another category. Second, one of the main advantages of our method is that it naturally takes into account the fact that current functional classification schemes are far from complete. Thus, many proteins belong to only a few categories perhaps because their other functions have been discovered yet. Under these circumstances, if one were to include the missing ones into the calculation, the incompleteness of the classification scheme becomes a major factor and makes the calculated similarity much less accurate.

There are obvious parallels in other scientific disciplines in taking a 'classification system' and turning it into a distance metric. Networks of computers or telephones have been studied under the aspect of distances. The problem of defining appropriate measures of similarity in these other disciplines, is however not completely resolved. For example, in a recent paper Huffaker *et al.* attempted to define different distance metrics in the internet that best correlates with the lowest latency correlation between hosts (Huffaker, *et al.*, 2002). Two of the distance measures used in this study were related to the shortest path length, and the other two were specific to the problem (actual geographical distance and packet round trip time). It seems that in these fields the issue of defining a measure of distance between pairs of nodes is far from being resolved.

A completely different measure of functional similarity is the dot product measure of functional similarity suggested in Lan *et al.* (2002). This measure relies on representing protein function in a very different way, not in terms of its position in a DAG structure, but in terms of a vector of molecular

interactions. The dot product, or the angle between these two vectors, gives a measure of functional similarity.

Finally, there are also ways of assessing the similarity of the function among groups of more than two proteins. This assessment of functional similarity within such groups of proteins usually makes the assumption that the proteins have been sampled randomly from the genome/proteome without replacement. The standard hypergeometric distribution describes the probability distribution of functional annotations in this sample, given the frequency at which these annotations occur in the genome/proteome as a whole. This is for instance used for finding functional labels that are significantly over-represented in the clusters of genes derived from transcriptional profiling data. In the future, one of the challenges is to look at how to use one of these continuous measures for a group of proteins.

## ACKNOWLEDGEMENTS

This work is supported by a grant from NIH (P50 HG02357-01). Funding to pay the open access charges was provided by a grant from NIH (P50 HG02357-01).

*Conflict of Interest:* none declared.

## REFERENCES

- Aho,A.V. *et al.* (1974) *Design and Analysis of Computer Algorithms*. Addison-Wesley, London.
- Azuaje,F. *et al.* (2006) Predictive integration of Gene Ontology-driven similarity and functional interactions. In *Proceedings of IEEE-ICDM 2006 Workshop on Data Mining in Bioinformatics*.
- Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
- Cox,J.S. and Walter,P. (1996) A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response. *Cell*, **87**, 391–404.
- Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
- Guo,X. *et al.* (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**, 967–973.
- Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York.
- Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
- Hegyri,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
- Hon,T. *et al.* (2001) The Hsp70-Ydj1 molecular chaperone represses the activity of the heme activator protein Hap1 in the absence of heme. *Mol. Cell. Biol.*, **21**, 7923–7932.
- Horak,C.E. *et al.* (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.*, **16**, 3017–3033.
- Huffaker,B. *et al.* (2002) Distance Metrics in the Internet. In *IEEE International Telecommunications Symposium*.
- Jansen,R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
- Karp,P.D. *et al.* (1996) EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res.*, **24**, 32–39.
- Kim,S. *et al.* (1998) Folding in vivo of a newly translated yeast cytosolic enzyme is mediated by the SSA class of cytosolic yeast Hsp70 proteins. *Proc. Natl Acad. Sci. USA*, **95**, 12860–12865.
- Kimura,Y. *et al.* (1995) Role of the protein chaperone YDJ1 in establishing Hsp90-mediated signal transduction pathways. *Science*, **268**, 1362–1365.
- Lan,N. *et al.* (2002) Toward a systematic definition of protein function that scales to the genome level: defining function in terms of interactions. *Proc. IEEE*, **90**, 1848–1858.

- Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
- Lei, Z. and Dai, Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**, 491.
- Lin, D. (1998) An information-theoretic definition of similarity. In *Proceedings of 15th International Conference on Machine Learning*. San Francisco, pp. 296–304.
- Mateos, A. *et al.* (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res.*, **12**, 1703–1715.
- Mewes, H.W. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
- Orengo, C.A. *et al.* (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
- Resnik, R. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal, pp. 448–453.
- Riley, M. and Space, D.B. (1996) Genes and proteins of *Escherichia coli* (GenProtEc). *Nucleic Acids Res.*, **24**, 40.
- Ruepp, A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
- Spellman, P. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Swamy, M.N.S. and Thulasiraman, K. (1981) *Graphs, Networks and Algorithms*. John Wiley and Sons, New York.
- Wang, H. and Azuaje, F. (2004) Gene expression correlation and gen ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of the 2004 IEEE symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 25–31.
- Wilson, C.A. *et al.* (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.
- Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.
- Wu, X. *et al.* (2006) Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res.*, **34**, 2137–2150.