Wang et al. BMC Bioinformatics 2010, 11:290
http://www.biomedcentral.com/1471-2105/11/290

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Revealing and avoiding bias in semantic similarity scores for protein pairs

Jing Wang[1], Xianxiao Zhou[1], Jing Zhu[1], Chenggui Zhou[1] and Zheng Guo*[1,2]

Abstract

Background: Semantic similarity scores for protein pairs are widely applied in functional genomic researches for finding functional clusters of proteins, predicting protein functions and protein-protein interactions, and for identifying putative disease genes. However, because some proteins, such as those related to diseases, tend to be studied more intensively, annotations are likely to be biased, which may affect applications based on semantic similarity measures. Thus, it is necessary to evaluate the effects of the bias on semantic similarity scores between proteins and then find a method to avoid them.

Results: First, we evaluated 14 commonly used semantic similarity scores for protein pairs and demonstrated that they significantly correlated with the numbers of annotation terms for the proteins (also known as the protein annotation length). These results suggested that current applications of the semantic similarity scores between proteins might be unreliable. Then, to reduce this annotation bias effect, we proposed normalizing the semantic similarity scores between proteins using the power transformation of the scores. We provide evidence that this improves performance in some applications.

Conclusions: Current semantic similarity measures for protein pairs are highly dependent on protein annotation lengths, which are subject to biological research bias. This affects applications that are based on these semantic similarity scores, especially in clustering studies that rely on score magnitudes. The normalized scores proposed in this paper can reduce the effects of this bias to some extent.

Background

Many scores for measuring semantic similarity (also termed functional similarity) between proteins have been proposed, based on the Gene Ontology (GO) terms [1] used to annotate the proteins. Some semantic similarity scores for a protein pair [2,3] are calculated by combining the similarity scores for the term pairs [4-7] describing the two proteins. Other scores between proteins that do not use pairwise similarity scores between terms have also been proposed [8,7,9-12]. Similarity scores for protein pairs have been widely applied in functional genomic research [13]. These scores are commonly used to analyze the correlation between functional similarity and similarities on other aspects, such as amino acid sequence similarity [2,8,14-16], or expression similarity [17-19]. Another type of applications is finding functional clusters of proteins [7,20-22], or functional modules in physical or genetic protein-protein interaction networks [23-28]. Similarity scores are also used to predict protein functions [29-35], protein-protein interactions [36-41] and putative disease genes [42-45].

GO protein annotations are known to be incomplete [46], and suffer from a large research bias, because certain proteins, such as those related to diseases, tend to be studied more intensively [43,47,48]. Such an annotation bias may affect protein semantic similarity scores. In this paper, we evaluated 14 common semantic similarity scores for protein pairs, and demonstrated that the scores significantly correlated with the numbers of annotation terms for the proteins (i.e., the annotation length). Thus, we proposed normalizing the scores based on their power transformation to reduce annotation bias effects, and we provide evidence that this improves performance in some applications.

* Correspondence: guoz@ems.hrbmu.edu.cn
[1] Bioinformatics Centre, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China
Full list of author information is available at the end of the article

BioMed Central

© 2010 Wang et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Methods

Gene Ontology (GO)

The GO annotation data for human was downloaded from the UniProt database http://www.ebi.ac.uk/GOA/index.html, including versions in November (Nov) 2001, Nov 2002, Nov 2003, Nov 2004, Nov 2005, Nov 2006, Nov 2007 and August 2008. The GO vocabulary data were downloaded from the GO website http://www.geneontology.org in August 2008. Here, we considered only IS-A links in GO [5,6], and we mainly present the results based on the "Biological Process" (BP) sub-ontology. We also observed that all the semantic similarity scores for pairs of term groups are dependent on the annotation lengths when using "Molecular Function" and "Cellular Component" (data not shown).

Online Mendelian Inheritance in Man (OMIM) database and disease classification

The data for 1996 genes associated with 2192 diseases were downloaded from the OMIM database [49]ftp://ftp.ncbi.nih.gov/repository/OMIM in August 2008, of which 1752 genes were annotated to GO BP terms. According to Goh et al. [50], the 2192 diseases were classified into 20 primary disorder classes based on the affected physiological systems. Diseases with multiple clinical features were assigned to the "multiple" class, and disease assigned to "Unclassified" class were not analyzed.

Similarity scores for term pairs

Many semantic similarity scores for two proteins are based on combinations of the similarity scores for term pairs between two groups of protein annotation terms. We evaluated four semantic similarity scores for term pairs based on the information contents: Resnik score [6], Lin score [5], Relevance score (RS) [4] and Jiang score [7]. The information content of a term c was defined as $IC(c) = -\log(p(c))$, where $p(c)$ is the number of gene products annotated to the term and its descendants, divided by the number of all gene products annotated to the GO BP ontology. Let $P(m, n)$ represent the set of common ancestor terms of m and n, then the four scores between terms m and n were calculated as:

$$sim_{\mathrm{Re}\,snik}(m,n) = \max_{c\in P(m,n)}[IC(c)]$$

$$sim_{Lin}(m,n) = \max_{c\in P(m,n)}\left(\frac{2\times IC(c)}{IC(m)+IC(n)}\right)$$

$$sim_{RS}(m,n) = \max_{c\in P(m,n)}\left(\frac{2\times IC(c)\times(1-p(c))}{IC(m)+IC(n)}\right)$$

$$sim_{Jiang}(m,n) = \frac{1}{1+IC(m)+IC(n)-2\times\max\limits_{c\in P(m,n)}[IC(c)]}$$

Similarity scores for protein pairs based on pairwise similarity scores between term groups

In some methods, the similarity scores for term pairs describing two proteins are combined to calculate the semantic similarity scores of the two proteins. Here, two combination methods were evaluated: the arithmetic average (AVG) of the pairwise semantic similarity scores between two groups of GO terms describing the two proteins [2] and the best-match average (BMA) approach [3].

A_1 and A_2 were the groups of annotation terms for proteins P_1 and P_2, and $\#P_1$ and $\#P_2$ were the number of terms included in A_1 and A_2. Then the two combined scores between the two proteins were defined as:

$$AVG(P_1,P_2) = \frac{1}{\#P_1\times\#P_2}\sum_{m\in A_1, n\in A_2} sim(m,n)$$

$$BMA(P_1,P_2) = \frac{S(P_1,P_2)+S(P_2,P_1)}{\#P_1+\#P_2}$$

where $S(P_1,P_2) = \sum_{m\in A_1}\max_{n\in A_2}(sim(m,n))$,

$S(P_2,P_1) = \sum_{m\in A_2}\max_{n\in A_1}(sim(m,n))$.

In total, eight semantic similarity measures for protein pairs were evaluated, using the four semantic similarity scores for term pairs (Resnik, Lin, RS and Jiang) combined with the AVG and BMA methods (see Table 1).

Similarity scores for protein pairs based on groupwise similarity scores between term groups

We also evaluated six protein semantic similarity scores that do not use pairwise similarity scores between two term groups. These similarity scores are briefly described as below (please see details in the original papers).

(1) The TO (Term Overlap) score [9] simply counts the number of overlapped terms for two proteins P_1 and P_2 as follows:

$$TO(P_1,P_2) = \#(GA_1\cap GA_2)$$

where GA_1 and GA_2 include the terms directly annotated with proteins P_1 and P_2 and all their ancestral terms, respectively. $\#(GA_1\cap GA_2)$ is the number of the overlapped terms between GA_1 and GA_2.

(2) The NTO (Normalized Term Overlap) score [9] is defined as:

$$NTO(P_1,P_2) = \frac{\#(GA_1\cap GA_2)}{\min(\#GA_1,\#GA_2)}$$

Table 1: Summary of 14 semantic similarity scores for protein pairs.

Measure	Description	Range
	Similarity scores for term pairs	
Resnik [6]	Information content of the most informative common ancestor of two terms	≥ 0
Lin [5]	Normalized Resnik similarity score by assessing how close two terms are to their most informative common ancestor	[0, 1)
RS [4]	Weighted Lin similarity score by using the probability of annotations of the most informative common ancestor	[0,1)
Jiang [7]	Based on the difference between two terms and their most informative common ancestor in information content	(0,1]
	Similarity scores for protein pairs based on pairwise similarity scores between term groups	
AVG [2]	The average of the similarity scores for all pairs of terms between two groups of protein annotations	Same with those for the corresponding similarity scores for term pairs
BMA [3]	The score of the best-matching pairs between two groups of protein annotations	
	Similarity scores for protein pairs based on groupwise similarity scores between term groups	
TO [9]	The number of terms shared by the annotations for two proteins	≥ 1
NTO [9]	Dividing TO by the minimum of the annotation lengths of two proteins	(0,1]
Dice [12]	Dividing TO by the average of annotation lengths of two proteins	(0,1]
Kappa [11]	A chance-corrected measure of co-occurrence between two groups of protein annotations	[0, 1]
GIC [8]	Jaccard index weighted by the information content of each GO term	[0, 1]
VSM [10]	Cosine similarity weighted by the information content of each GO term	[0, 1]

(3) The Dice score [12] is defined as:

$$Dice(P_1, P_2) = \frac{2\times\#(GA_1 \cap GA_2)}{\#\,GA_1 + \#\,GA_2}$$

(4) The Kappa score [11] is defined as:

$$Kappa(P_1, P_2) = \frac{O_{P_1,P_2} - A_{P_1,P_2}}{1 - A_{P_1,P_2}}$$

where O_{P_1,P_2} and A_{P_1,P_2} represent the observed and random co-occurrence of GO annotation terms for the two proteins, respectively.

(5) The Graph Information Content (GIC) score [8] is defined as:

$$GIC(P_1, P_2) = \frac{\Sigma_{c\in GA_1\cap GA_2} IC(c)}{\Sigma_{c\in GA_1\cup GA_2} IC(c)}$$

(6) The Vector Space Model (VSM) score [10] is defined as follow:

$$VSM(P_1, P_2) = \frac{\sum_{k=1}^{n} \omega_{1_k} \times \omega_{2_k}}{\sqrt{\sum_{k=1}^{n} \omega_{1_k}^2 \times \sum_{k=1}^{n} \omega_{2_k}^2}}$$

where n is the number of all the GO BP terms and $\omega_{1_k}(\omega_{2_k})$ is the information content of term k if it is annotated with protein P_1 (P_2) while $\omega_{1_k}(\omega_{2_k})$ is 0 if the term k is not an annotation of the protein P_1 (P_2).

In total, we evaluated 14 semantic similarity scores for protein pairs (see Table 1). We note that some other semantic similarity scores for protein pairs [13,51] were not evaluated in this paper. For example, the score proposed by Wang et al. [22], which weights the IS-A and

PART-OF links of GO, was not analyzed, because we considered only IS-A links in this study.

Random experiments

Using randomly selected pairs of term groups, we evaluated the increase in protein semantic similarity score that resulted from only the increased annotation length, regardless of other biological factors. First, we randomly selected 10,000 pairs of term groups with the same sizes (corresponding to the annotation lengths of proteins) ranging from 1 to 10, since only 1.5% of proteins had more than 10 annotations in GO BP ontology. Then, using each of the 14 semantic similarity scores described above, we calculated the semantic similarity scores for random term group pairs, and analyzed whether these scores increased as the group size increased using the Spearman rank correlation coefficient [52].

Normalization based on power transformation

As demonstrated in the *Results* section, a similarity score for two groups of terms is dependent on the lengths of the term groups. To reduce the effect of the lengths on the scores, we took two steps to make the scores for pairs of term groups with given length combinations follow the standard normal distribution.

Firstly, we applied the commonly used power transformation approach to transform the scores to achieve normality [53,54]. Suppose $SS(TG_1, TG_2)$ is the score for term groups TG_1 with length L_1 and TG_2 with length L_2, we power-transformed it to $TSS(TG_1, TG_2) = SS(TG_1, TG_2)^{\lambda_{L_1,L_2}}$. Here, the power λ_{L_1,L_2} was estimated as follow [53,54]:

$$\lambda_{L_1,L_2} = 1 - median\left(\frac{[(T_{1-q}+T_q)/2]-M_{SS}}{[(T_{1-q}-M_{SS})^2+(M_{SS}-T_q)^2]/4M_{SS}}\right)$$

where M_{SS} is the median of the random $SS(TG_1, TG_2)$ distribution which was estimated by the similarity scores for 10,000 pairs of random term groups (with lengths L_1 and L_2). T_q and T_{1-q} are the lower and upper qth quantiles of this distribution ($\left(q = \frac{1}{4}, \frac{1}{8}, \cdots, \frac{1}{64}\right)$). By the one-sample Kolmogorov-Smirnov test for distribution goodness-of-fit [55], at the significance level of 0.1, we tested whether the power-transformed scores for pairs of term groups with given length combinations fit normal distributions.

Secondly, we normalized the above power-transformed scores to make them fit the standard normal distribution as follow:

$$NSS(TG_1, TG_2) = \frac{NSS(TG_1,TG_2)-M_{TSS}}{STD_{TSS}}$$

In the above normalization formula, M_{TSS} and STD_{TSS} are the median and standard deviation of the power-transformed scores respectively. Here, we used the median rather than the mean in the normalization formula because it might be more appropriate for measuring the location parameter of a distribution when the distribution might be skewed [56-58]. As shown in the *Results* section, most of the normalized scores for pairs of term groups with given length combinations follow normal distributions. In this situation, the means and the medians are equal.

Sequence similarity scores for protein pairs

Amino acid sequence data for human proteins was downloaded from UniProt ftp://ftp.uniprot.org in August 2008. The sequence similarity between two proteins was measured by the ln(bit score), and calculated by the NCBI "blastall" program [2]. Sequence similarity scores were obtained for a total of 499,878 protein pairs with GO BP annotations.

Clustering algorithm and enrichment analysis

Suppose the original and normalized similarity scores for two proteins (P_1 and P_2) annotated with two groups of terms are $SS(P_1, P_2)$ and $NSS(P_1, P_2)$ respectively, we firstly transformed both $SS(P_1, P_2)$ and $NSS(P_1, P_2)$ to the range [0, 1] by the Min-Max normalization method [59,60] as follows

$$MM(P_1, P_2) = \frac{SS(P_1,P_2)-Min_{SS}}{Max_{SS}-Min_{SS}}$$

$$NM(P_1, P_2) = \frac{NSS(P_1,P_2)-Min_{SS}}{Max_{SS}-Min_{SS}}$$

where Max_{SS} and Min_{SS} are the maximum and minimum values of the original similarity scores for all protein pairs from a protein set (e.g., a set of proteins encoded by a set of disease genes). Max_{NSS} and Min_{NSS} are the maximum and minimum values of the normalized similarity scores for all these protein pairs.

Then, we calculated the distance between two proteins as $D(P_1, P_2) = 1\text{-}MM(P_1, P_2)$ based on the original score. Similarly, based on the normalized score, the distance was calculated as $ND(P_1, P_2) = 1\text{-}NM(P_1, P_2)$. Both $D(P_1,$

P_2) and $ND(P_1, P_2)$ take values ranging from 0 to 1 and satisfy three main properties of distance metrics [61]: (i) symmetry, $D(P_1, P_2) = D(P_2, P_1)$ ($ND(P_1, P_2) = ND(P_2, P_1)$); (ii) non-negative, $D(P_1, P_2) \geq 0$ ($ND(P_1, P_2) \geq 0$); (iii) triangle inequality, $D(P_1, P_3) \leq D(P_1, P_2)+D(P_2, P_3)$ ($ND(P_1, P_3) \leq ND(P_1, P_2)+ND(P_2, P_3)$). Using $D(P_1, P_2)$ and $ND(P_1, P_2)$ respectively, we clustered disease genes by the complete linkage clustering algorithm [62].

To evaluate the clustering results, using the hypergeometric distribution model [63,64], we calculated the probability p of detecting at least the observed number of genes related to a disease category proposed by Goh et al. [50] in a cluster of disease genes by random chance. The p values were corrected by the false discovery rate (FDR) by the Benjamini-Hochberg (BH) procedure [65]. With FDR of 1%, we found the disease categories enriched in a cluster of disease genes found by the clustering algorithm.

Results

The dependence of the semantic similarity scores on annotation lengths

From 2001 to 2008, the average number of GO BP terms annotated with disease genes increased from 2.6 to 5.1, as shown in Figure 1(A). In contrast, the average annotation length of "non-disease" genes increased slightly from 1.7 to 2.1 (Figure 1(B)). These results indicated that disease genes tend to be studied more extensively, and are biased to have more annotations.

As shown in Figure 2, for each of the 14 protein semantic similarity scores analyzed, the median score for 10,000 random pairs of term groups increased significantly as the annotation lengths (the group sizes) increased (Spearman $r \geq 0.99$, $p <$ 1E-07). Based on the Resnik [6], Lin [5], RS [4] and Jiang [7] semantic similarity scores for term pairs, all four AVG combined scores for protein pairs were relatively stable when the annotation length was greater than three, especially for the Jiang(AVG). In contrast, the combined scores based on Resnik(BMA), Lin(BMA) and RS(BMA) increased rapidly with the annotation length. We evaluated six other semantic similarity scores for protein pairs, which do not use pairwise similarity scores between terms. As shown in Figure 2, the TO score was linearly dependent on the annotation length, while other scores increased slightly but significantly as the annotation length increased (Spearman r = 1, $p <$ 1E-09). Notably, as shown in Figure 2, as the annotation lengths increased, the standard deviation of TO scores increased but it decreased for other similarity scores, which could be explained statistically. For example, because TO scores follow the hypergeometric probability model [63,64], we can derive that its standard deviation increases with the annotation lengths.

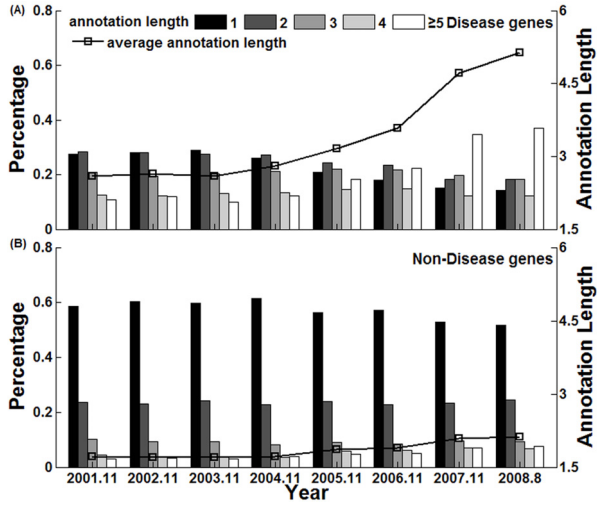

Figure 1 Changes of the gene annotation lengths in GO BP ontology. The bar plots represent the percentages of disease and non-disease genes with different annotation lengths, while the solid-square lines represent the average annotation lengths in various versions of GO.

Applications of the normalized scores

As shown in Table 2, based on each of the 14 similarity measures for term groups, most of the original scores ($SS(TG_1, TG_2)$) for pairs of term groups with given length combinations did not fit normal distributions ($p \geq 0.1$, one-sample Kolmogorov-Smirnov test). For nine similarity measures, namely the Resnik(AVG), Resnik(BMA), Lin(AVG), Lin(BMA), RS(AVG), RS(BMA), Jiang(AVG), Dice and Kappa scores, over 80% of the power-transformed scores for pairs of term groups with given length combinations followed normal distributions. Then, these power-transformed scores were normalized to the standard normal distribution. Thus, for these nine similarity measures, the normalization method based on the power transformation is largely suitable for comparing scores for pairs of term groups with different length combinations. However, based on each of the other five similarity measures, less than 60% of the power-transformed scores fitted normal distributions. We also analyzed another five simple transformation methods and the results (see Table 2) showed that all these simple transformation methods performed worse than the power-transformation method using the estimated $\lambda_{L_1' L_2}$.

Then, for two types of applications, by comparing the results based on the original scores and their corresponding normalized scores, we showed that the bias affects certain analysis more than others. One type of applications based on semantic similarity scores for protein pairs study the correlation between functional similarity and similarities on other aspects [2,8,14-19]. Based on the normalized RS(BMA) score (the corresponding $\lambda_{L_1' L_2}$

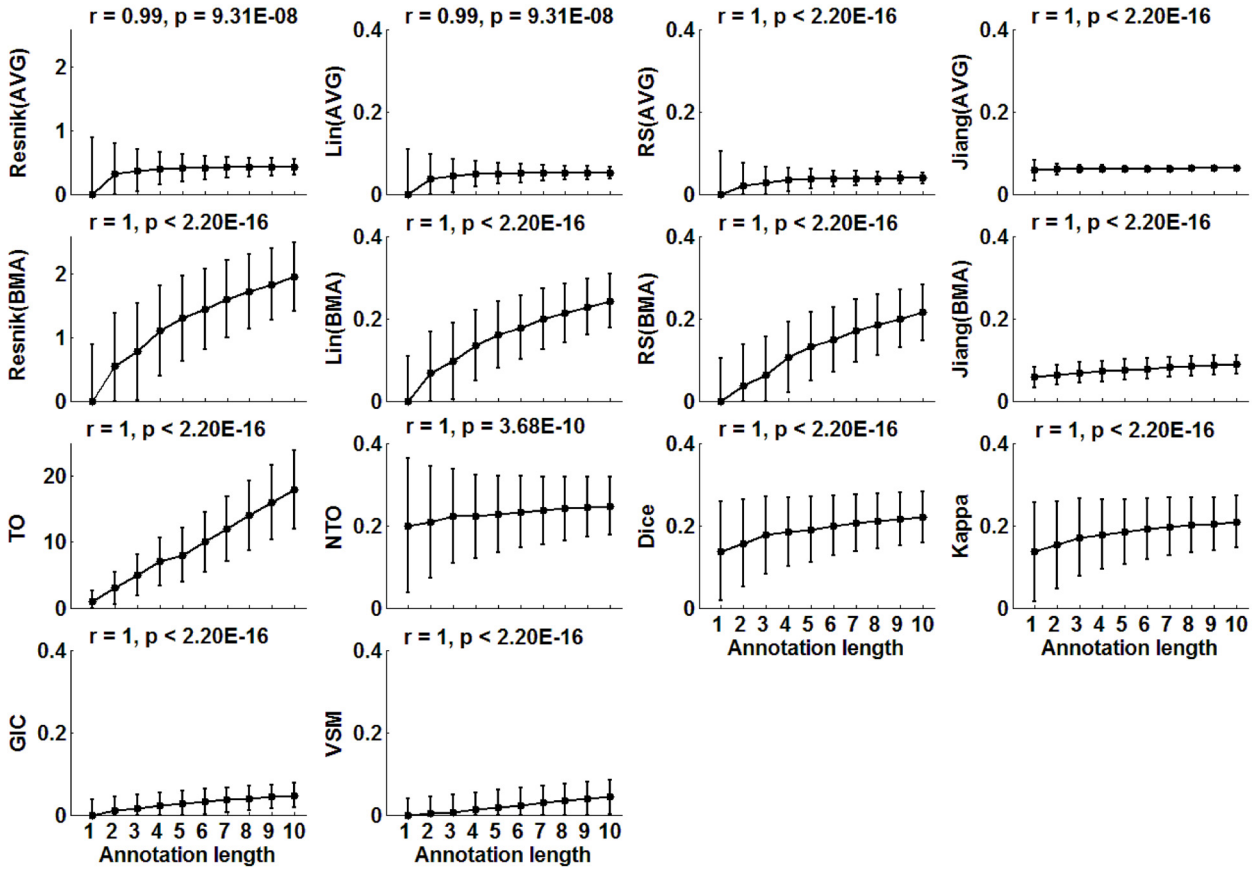

Figure 2 Semantic similarity scores for pairs of term groups increase as the annotation lengths increase. For each of the 14 semantic similarity measures, the solid circles represent the median score of 10,000 random pairs of term groups annotated with the same number of GO BP terms. The error bars represent the standard variation of the scores (because the similarity scores were positive, we reduced the negative part of error bars). *r* and *p* are the Spearman correlation coefficient and the corresponding significance between the median of the scores and the annotation lengths. To make the 14 plots more informative, according to the ranges of the scores (see Table 1), we classify the plots into three groups, in each of which the plots have the same y-axis scale.

distribution for this measure was shown in Figure 3), we analyzed the correlation between protein semantic similarity scores and their amino acid sequence similarity scores (ln(bit score)). As shown in Figure 4, the correlation was significant (p < 2.20E-16), supporting the model that proteins with similar sequences tend to be functionally similar [2,14]. Based on the RS(BMA) scores, similar results were observed, because of significant correlation between the ranks of the RS(BMA) scores and the normalized RS(BMA) scores for protein pairs (Spearman r = 0.88, p < 2.20E-16).

Another type of applications is clustering of functionally similar proteins [7,20-22] or finding functional modules in physical or genetic protein-protein interaction networks [23-28]. Using the RS(BMA) and the normalized RS(BMA) distance, we applied a hierarchy clustering algorithm to cluster the disease genes into 21 categories, and compared the results with the categories determined by Goh et al. [50]. As evaluated by the hypergenomic distribution test, using FDR of 1%, 16 clusters based on the normalized distance were enriched with disease genes with the same or similar phenotypes while only 6 clusters were enriched based on the original distance. To analyze more clearly the effect of annotation length on the cluster results, we clustered only the genes determined to the "Hematological" and "Immunological" categories. As shown in Figure 5(A), based on the normalized RS(BMA) distance, 73.5% of "Hematological" genes (red) were clustered into one group (p = 7.3E-13), while 78.4% of "Immunological" genes (blue) were in another (p = 3.96E-13). In contrast, as shown in Figure 5(B), when clustering these two classes of disease genes into two groups based on the RS(BMA) distance, no group was significantly enriched with a class of disease genes (p > 0.10). As shown in Figure 6, after normalization, the ranks of some disease gene pairs with different annotation lengths changed, improving the clustering results based on the normalized RS(BMA) distance. In general, based on the normalized

Table 2: The performance of different data transformation methods*.

Measure	Estimated λ**	λ = 1	Inverse (λ = -1)	Cube-root (λ = 1/3)	Square-root (λ = 1/2)	Square (λ = 2)	Log
Resnik(AVG)	0.878	0	-***	0.645	0.370	0	-
Lin(AVG)	0.890	0	-	0.659	0.474	0	-
RS(AVG)	0.925	0	-	0.632	0.355	0	-
Jiang(AVG)	0.812	0	0.081	0	0	0	0.002
Resnik(BMA)	0.938	0.661	-	0.025	0.248	0	-
Lin(BMA)	0.940	0.706	-	0.012	0.156	0.002	-
RS(BMA)	0.927	0.650	-	0.004	0.042	0.001	-
Jiang(BMA)	0.010	0	0	0	0	0	0
TO	0	0	0	0	0	0	0
NTO	0.555	0.001	0	0.366	0.478	0	0.009
Dice	0.926	0.014	0	0.384	0.890	0	0.001
Kappa	0.896	0.010	-	0.518	0.866	0	-
GIC	0.552	0	-	0.096	0	0	-
VSM	0.291	0	-	0.006	0	0	-

* The numbers in the table represent the percentages of the scores that fitted normal distributions after data transformation, among all group pairs with different length combinations.
** λ was estimated by the method described in the *Methods* section.
*** "-" indicates the transformation method was not suitable for the similarity measure.

RS(BMA) scores, our results suggested that disease genes related to the same or similar diseases tend to work together in the same disease-related functional gene modules [50,66-78].

Discussion

In this paper, we found that most semantic similarity scores for protein pairs increased as protein annotation lengths increased. Because protein annotations are likely to be subject to biological research bias, most applications based on current semantic similarity scores for protein pairs will be biased. Without the annotation bias, one could argue that over-annotated proteins might be more likely to be similar than under-annotated proteins, when considering only shared functions, and disregarding differences. However, currently, most semantic similarity scores for protein pairs evaluate the overall functional

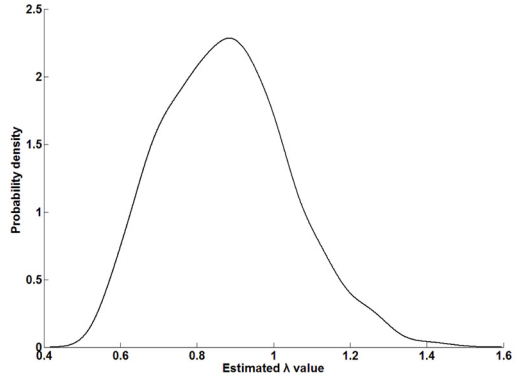

Figure 3 Probability density of λ_{L_1,L_2} estimated for pairs of term groups with different annotation length combinations based on the RS(BMA) measure.

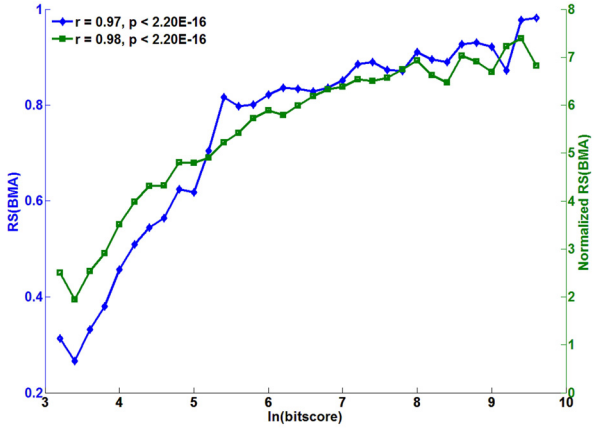

Figure 4 Relationship between sequence similarity and semantic similarity of protein pairs. The semantic similarity scores of protein pairs were calculated by RS(BMA) (blue) and normalized RS(BMA) (green) measure. The results showed that the original and normalized RS(BMA) scores had similar correlation with sequence similarity scores, and confirmed that proteins with similar sequences tend to have higher "functional similarity" [2,14].

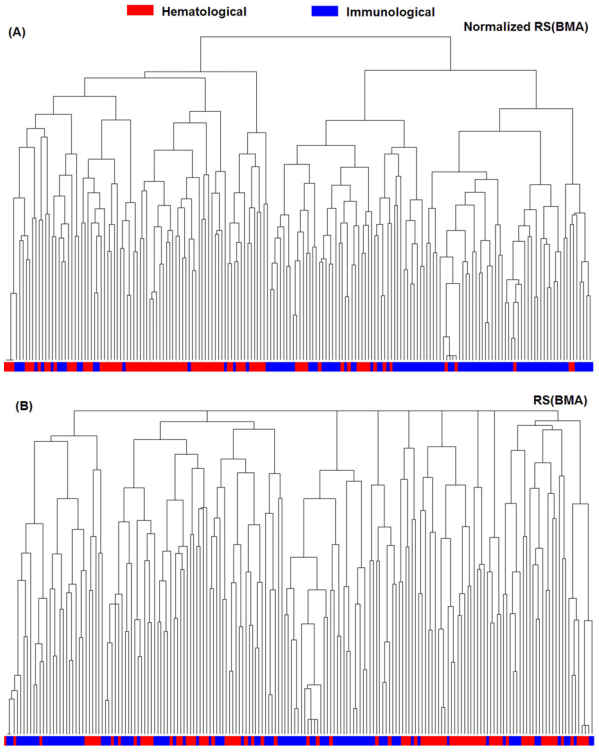

Figure 5 Cluster analysis of genes in "Hematological" and "Immunological" disease categories. Two types of disease genes were clustered by the normalized (A) and original (B) RS(BMA) methods. Horizontal axis represents all disease genes of "Hematological" (red) and "Immunological" (blue) categories.

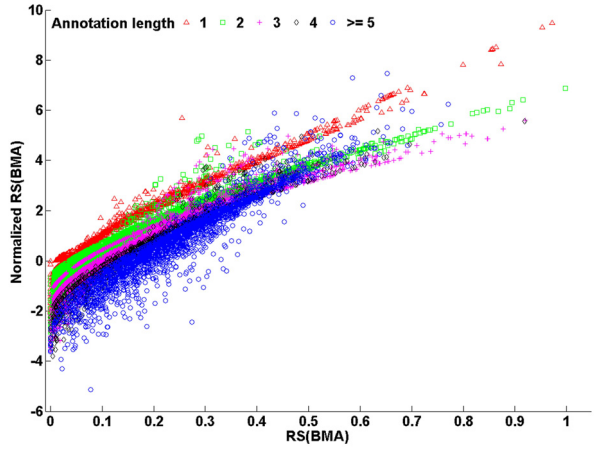

Figure 6 Scatter plot of RS(BMA) vs. normalized RS(BMA) scores for pairs of genes in "Hematological" and "Immunological" disease categories. Disease gene pairs were classified into five groups, according to the minimum of annotation lengths of the two genes: one (red triangles), two (green squares), three (magenta plus signs), four (black diamonds) and more than four (blue circles).

similarity between proteins. Depending on the available knowledge about domains, and the final aim of the application, different criteria could be used to define similarity between proteins. We note that protein annotations in GO for most model organisms (e.g., *Saccharomyces cerevisiae*) are also incomplete and suffer from the research bias because important genes such as the homologues of human disease genes tend to be studied more intensively [79,80]. By analyzing the *Saccharomyces cerevisiae* data, we also found that the similarity scores between two groups of terms increased significantly with the annotation lengths (data not shown). Thus, our conclusion on the bias of semantic similarity scores for proteins would be applicable to other organisms.

A protein is usually annotated to a group of GO terms. Often, the semantic similarity scores between two proteins are calculated using some combination methods [2,3] based on the semantic similarity scores for pairs of terms annotated with the two proteins. Many semantic similarity scores for term pairs such as the Resnik [6], Lin [5], Relevance (RS) [4] and Jiang [7] are based on the information content (related to the annotation specificity) of the terms. Based on these similarity scores for term pairs, the similarity scores for two proteins might not always increase, if the proteins have many annotations with low-specificity. However, as shown here, all the AVG and BMA scores for protein pairs based on the Resnik, Lin, RS and Jiang scores for term pairs still significantly correlated with the protein annotation lengths.

To reduce the effects of protein annotation bias, we normalized the scores based on the power transformation by estimating power λ_{L_1,L_2}. The normalization method based on the power transformation can transform most scores based on nine of the similarity measures to fit normal distributions but it performs poorly for the other five similarity measures. Thus, future works are needed to further improve the data transformation and normalization method.

The feasibility of the normalized scores was analyzed for two types of applications and the results showed that the normalized scores were useful in these applications. Analysis of the correlation between functional similarities and similarities on other aspects [2,8,14-19] might be less affected by the annotation bias, because the ranks of semantic similarity scores for protein pairs and their corresponding normalized scores were highly correlated. Our results also showed that clustering analysis [7,20-22] using the magnitude of the semantic similarity scores might be more seriously affected by biased protein annotations, and the results could be improved by using the normalized scores.

A third type of applications that uses protein semantic similarity scores is predicting protein functions [29-35],

protein-protein interactions [36-41] and disease genes [42-45]. However, because many other factors, such as the selection of algorithms and the definition of positive and negative sets [81] can affect the prediction results, we did not evaluate the effect of the annotation bias on these uses. Nevertheless, because this type of applications also uses the similarity score magnitudes, we recommend also using normalized scores in prediction studies, to reduce the effects of the annotation bias.

To avoid the influence of annotation bias, other approaches may be attempted. For example, the statistical *p*-value of a semantic similarity score for a protein pair could be evaluated by comparing this score with the scores of random protein pairs with the same annotation lengths. If the semantic similarity score of the two proteins was significantly larger than the score expected by random chance, at a given significance level (*p*-value), we could determine that the two proteins are functionally similar [82]. Functional modules could be found by linking functionally related proteins. To analyze the functional relationships of proteins more comprehensively, the semantic similarity scores should be combined with other functional data, such as protein-protein interaction, co-expression and co-conservation of proteins [83-85].

Conclusions

Current protein semantic similarity scores are highly dependent on protein annotation lengths, which are subject to biological research bias. This bias may affect many current applications based on these scores. The proposed normalization method might solve this problem to some extend.

Authors' contributions

Conceived and designed the experiments: ZG, JW. Performed the experiments: JW, XZ and CZ. Analyzed the data: ZG, JW and JZ. Wrote the paper: JW, JZ and ZG. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant nos. 30571034, 30970668, 30670539).

Author Details

[1]Bioinformatics Centre, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China and [2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China

Received: 9 February 2010 Accepted: 28 May 2010
Published: 28 May 2010

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.
2. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19(10):**1275-1283.
3. Marino-Ramirez L, Bodenreider O, Kantz N, Jordan IK: **Co-evolutionary Rates of Functionally Related Yeast Genes.** *Evol Bioinform Online* 2006, **2:**295-300.
4. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7:**302.
5. Lin D: **An information-theoretic definition of similarity.** *Proc 15th International Conf on Machine Learning: 1998* 1998:296-304.
6. Resnik P: **Using information content to evaluate semantic similarity in a taxonomy.** *Proceedings of the 14th International Joint Conference on Artificial Intelligence: 1995* 1995:448-453.
7. Ovaska K, Laakso M, Hautaniemi S: **Fast Gene Ontology based clustering for microarray experiments.** *BioData Min* 2008, **1(1):**11.
8. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, Couto FM: **Metrics for GO based protein semantic similarity: a systematic evaluation.** *BMC Bioinformatics* 2008, **9(Suppl 5):**S4.
9. Mistry M, Pavlidis P: **Gene Ontology term overlap as a measure of gene functional similarity.** *BMC Bioinformatics* 2008, **9:**327.
10. Chabalier J, Mosser J, Burgun A: **A transversal approach to predict gene product networks from ontology-based similarity.** *BMC Bioinformatics* 2007, **8:**235.
11. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biol* 2007, **8(9):**R183.
12. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **GOToolBox: functional analysis of gene datasets based on Gene Ontology.** *Genome Biol* 2004, **5(12):**R101.
13. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM: **Semantic similarity in biomedical ontologies.** *PLoS Comput Biol* 2009, **5(7):**e1000443.
14. Joshi T, Xu D: **Quantitative assessment of relationship between sequence similarity and function similarity.** *BMC Genomics* 2007, **8:**222.
15. Yang L, Yu J: **A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes.** *BMC Evol Biol* 2009, **9:**55.
16. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol* 2009, **5(1):**e1000262.
17. Elo LL, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T: **Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process.** *Bioinformatics* 2007, **23(16):**2096-2103.
18. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14(6):**1085-1094.
19. Wang H, Azuaje F, Bodenreider O, Dopazo J: **Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships.** *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB'04: 2004* 2004:25-31.
20. Chen JL, Liu Y, Sam LT, Li J, Lussier YA: **Evaluation of high-throughput functional categorization of human disease genes.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S7.
21. Du Z, Li L, Chen CF, Yu PS, Wang JZ: **G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery.** *Nucleic Acids Res* 2009:W345-349.
22. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF: **A new method to measure the semantic similarity of GO terms.** *Bioinformatics* 2007, **23(10):**1274-1281.
23. Ulitsky I, Shlomi T, Kupiec M, Shamir R: **From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions.** *Mol Syst Biol* 2008, **4:**209.
24. Bandyopadhyay S, Kelley R, Krogan NJ, Ideker T: **Functional maps of protein complexes from quantitative genetic interaction data.** *PLoS Comput Biol* 2008, **4(4):**e1000065.
25. Sen TZ, Kloczkowski A, Jernigan RL: **Functional clustering of yeast proteins from the protein-protein interaction network.** *BMC Bioinformatics* 2006, **7:**355.
26. Lubovac Z, Gamalielsson J, Olsson B: **Combining functional and topological properties to identify core modules in protein interaction networks.** *Proteins* 2006, **64(4):**948-959.

27. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, *et al.*: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440(7084):**637-643.
28. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome.** *Nat Biotechnol* 2009, **27(2):**199-204.
29. Chen XW, Liu M, Ward R: **Protein function assignment through mining cross-species protein-protein interactions.** *PLoS ONE* 2008, **3(2):**e1562.
30. Zhu M, Gao L, Guo Z, Li Y, Wang D, Wang J, Wang C: **Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities.** *Gene* 2007, **391(1-2):**113-119.
31. Tao Y, Sam L, Li J, Friedman C, Lussier YA: **Information theory applied to the sparse gene ontology annotation network to predict novel gene function.** *Bioinformatics* 2007, **23(13):**i529-538.
32. Tu K, Yu H, Guo Z, Li X: **Learnability-based further prediction of gene functions in Gene Ontology.** *Genomics* 2004, **84(6):**922-928.
33. Cakmak A, Ozsoyoglu G: **Discovering gene annotations in biomedical text databases.** *BMC Bioinformatics* 2008, **9:**143.
34. Cho YR, Shi L, Ramanathan M, Zhang A: **A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge.** *BMC Bioinformatics* 2008, **9:**382.
35. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S: **Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology.** *PLoS ONE* 2009, **4(2):**e4619.
36. Futschik ME, Chaurasia G, Herzel H: **Comparison of human protein-protein interaction maps.** *Bioinformatics* 2007, **23(5):**605-611.
37. Wu X, Zhu L, Guo J, Zhang DY, Lin K: **Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations.** *Nucleic Acids Res* 2006, **34(7):**2137-2150.
38. Ofran Y, Yachdav G, Mozes E, Soong TT, Nair R, Rost B: **Create and assess protein networks through molecular characteristics of individual proteins.** *Bioinformatics* 2006, **22(14):**e402-407.
39. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW: **An in vivo map of the yeast protein interactome.** *Science* 2008, **320(5882):**1465-1470.
40. Xu T, Du L, Zhou Y: **Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data.** *BMC Bioinformatics* 2008, **9:**472.
41. Soong TT, Wrzeszczynski KO, Rost B: **Physical protein-protein interactions predicted from microarrays.** *Bioinformatics* 2008, **24(22):**2608-2614.
42. Gaulton KJ, Mohlke KL, Vision TJ: **A computational system to select candidate genes for complex human traits.** *Bioinformatics* 2007, **23(9):**1132-1140.
43. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics* 2006, **22(6):**773-774.
44. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009:W305-311.
45. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18(Suppl 2):**S110-115.
46. Rhee SY, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations.** *Nat Rev Genet* 2008, **9(7):**509-515.
47. Verver O, Ridder Jd, Reinders MJT, Wessels LFA: **Prioritization of Candidate Disease Genes using Microarray Data and Functional Relations.** *Bioinformatics* 2007, **00:**1-12.
48. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6:**55.
49. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005:D514-517.
50. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci USA* 2007, **104(21):**8685-8690.
51. Chagoyen M, Carazo J, Pascual-Montano A: **Pairwise similarity scores using functional annotations: review and comparison.** *8th Spanish Symposium on Bioinformatics and Computational Biology: 2008* 2008.
52. Spearman C: **The Proof and Measurement of Association Between Two Things.** *American Journal of Psychology* 1904, **15:**72-101.
53. Tan W, Gan F, Chang T: **Using normal quantile plot to select an appropriate transformation to achieve normality.** *Computational Statistics & Data Analysis* 2004, **45(3):**609-619.
54. Emerson J, Stoto M: **Exploratory Methods for Choosing Power Transformations.** *Journal of the American Statistical Association* 1982, **77:**103-108.
55. Massey J, Frank J: **The Kolmogorov-Smirnov test for goodness of fit.** *Journal of the American Statistical Association* 1951, **46:**68-78.
56. Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, Flegal KM, Guo SS, Wei R, Mei Z, Curtin LR, Roche AF, Johnson CL: **CDC growth charts: United States.** *Adv Data* 2000, **314:**1-27.
57. Kuk A, Mak T: **Median estimation in the presence of auxiliary information.** *Journal of the Royal Statistical Society Series B Methodological* 1989, **51:**261-269.
58. Waterlow JC, Buzina R, Keller W, Lane JM, Nichaman MZ, Tanner JM: **The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years.** *Bull World Health Organ* 1977, **55(4):**489-498.
59. Indovina M, Uludag U, Snelick R, Mink A, Jain A: **Multimodal biometric authentication methods: a COTS approach.** *Proceedings of the MMUA: 2003* 2003:99-106.
60. Sorace JM, Zhan M: **A data review and re-assessment of ovarian cancer serum proteomic profiling.** *BMC Bioinformatics* 2003, **4:**24.
61. Wang J, Wang X, Lin K, Shasha D, Shapiro B, Zhang K: **Evaluating A Class of Distance-Mapping Algorithms for Data Mining and Clustering.** *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 1999* 1999:307-311.
62. Yamada T, Kanehisa M, Goto S: **Extraction of phylogenetic network modules from the metabolic network.** *BMC Bioinformatics* 2006, **7:**130.
63. Fury W, Batliwalla F, Gregersen PK, Li W: **Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion.** *Conf Proc IEEE Eng Med Biol Soc* 2006, **1:**5531-5534.
64. Gonin H: **The use of factorial moments in the treatment of the hypergeometric distribution and in tests for regression.** *Philosophical Magazine Series 7* 1936, **21(139):**215-226.
65. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B Methodological* 1995, **57:**289-330.
66. Oti M, Brunner HG: **The modular nature of genetic diseases.** *Clin Genet* 2007, **71(1):**1-11.
67. Brunner HG, van Driel MA: **From syndrome families to functional genomics.** *Nat Rev Genet* 2004, **5(7):**545-551.
68. Oti M, Huynen MA, Brunner HG: **Phenome connections.** *Trends Genet* 2008, **24(3):**103-106.
69. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, *et al.*: **A human phenome-interactome network of protein complexes implicated in genetic disorders.** *Nat Biotechnol* 2007, **25(3):**309-316.
70. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S: **A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes.** *Proc Natl Acad Sci USA* 2008, **105(52):**20870-20875.
71. Girirajan S, Truong HT, Blanchard CL, Elsea SH: **A functional network module for Smith-Magenis syndrome.** *Clin Genet* 2009, **75(4):**364-374.
72. Moran LB, Graeber MB: **Towards a pathway definition of Parkinson's disease: a complex disorder with links to cancer, diabetes and inflammation.** *Neurogenetics* 2008, **9(1):**1-13.
73. Li Y, Agarwal P: **A pathway-based view of human diseases and disease relationships.** *PLoS One* 2009, **4(2):**e4346.
74. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, *et al.*: **A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration.** *Cell* 2006, **125(4):**801-814.
75. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, *et al.*: **A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease.** *Mol Cell* 2004, **15(6):**853-865.
76. Bergholdt R, Storling ZM, Lage K, Karlberg EO, Olason PI, Aalund M, Nerup J, Brunak S, Workman CT, Pociot F: **Integrative analysis for finding genes**

and networks involved in diabetes and other complex diseases. *Genome Biol* 2007, **8(11):**R253.
77. Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ: **A network biology approach to prostate cancer.** *Mol Syst Biol* 2007, **3:**82.
78. Jiang X, Liu B, Jiang J, Zhao H, Fan M, Zhang J, Fan Z, Jiang T: **Modularity in the genetic disease-phenotype network.** *FEBS Lett* 2008, **582(17):**2549-2554.
79. Kaletta T, Hengartner MO: **Finding function in novel targets: C. elegans as a model organism.** *Nat Rev Drug Discov* 2006, **5(5):**387-398.
80. Langenau DM, Jette C, Berghmans S, Palomero T, Kanki JP, Kutok JL, Look AT: **Suppression of apoptosis by bcl-2 overexpression in lymphoid cells of transgenic zebrafish.** *Blood* 2005, **105(8):**3278-3285.
81. Yu S, Van Vooren S, Tranchevent LC, De Moor B, Moreau Y: **Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining.** *Bioinformatics* 2008, **24(16):**i119-125.
82. Yang D, Li Y, Xiao H, Liu Q, Zhang M, Zhu J, Ma W, Yao C, Wang J, Wang D, *et al.*: **Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories.** *Bioinformatics* 2008, **24(2):**265-271.
83. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306(5701):**1555-1558.
84. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *Am J Hum Genet* 2006, **78(6):**1011-1025.
85. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36(10):**1090-1098.

doi: 10.1186/1471-2105-11-290
Cite this article as: Wang *et al.*, Revealing and avoiding bias in semantic similarity scores for protein pairs *BMC Bioinformatics* 2010, **11**:290

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central