D434–D439 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm806

FunSimMat: a comprehensive functional similarity database

Andreas Schlicker* and Mario Albrecht

Max Planck Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

Received August 15, 2007; Revised September 13, 2007; Accepted September 17, 2007

ABSTRACT

Functional similarity based on Gene Ontology (GO) annotation is used in diverse applications like gene clustering, gene expression data analysis, protein interaction prediction and evaluation. However, there exists no comprehensive resource of functional similarity values although such a database would facilitate the use of functional similarity measures in different applications. Here, we describe FunSimMat (Functional Similarity Matrix, http://funsimmat.bioinf.mpi-inf.mpg.de/), a large new database that provides several different semantic similarity measures for GO terms. It offers various precomputed functional similarity values for proteins contained in UniProtKB and for protein families in Pfam and SMART. The web interface allows users to efficiently perform both semantic similarity searches with GO terms and functional similarity searches with proteins or protein families. All results can be downloaded in tab-delimited files for use with other tools. An additional XML–RPC interface gives automatic online access to FunSimMat for programs and remote services.

INTRODUCTION

Sequencing efforts have produced large amounts of genomic data, which are functionally characterized further by experimental techniques and automated methods (1,2). The controlled vocabulary provided by the Gene Ontology (GO) consortium is commonly used for annotating gene products with their function (3). GO consists of three ontologies: biological process, molecular function and cellular component. Each GO term is represented by a node in a directed acyclic graph (DAG). Relationships between different GO terms are established by edges that connect GO term nodes within the DAG.

Different approaches for computing the functional similarity between gene products have been proposed. The simplest, but least sensitive, method is to count the number of GO terms annotated to two functionally related proteins. More advanced approaches are based on the semantic similarity of GO terms and compute a numerical value of the similarity between two GO terms (4–9). These semantic similarity measures rely on an annotation database for defining the importance of every GO term. One popular method based on semantic similarity is the calculation of the average semantic similarity between the GO terms annotated to the two gene products being compared (10). A modification of this method is the definition of the functional similarity between two gene products as the maximal semantic similarity between two annotated GO terms (11–13). Two more sophisticated approaches introduced by Zhang *et al*. (9) and Schlicker *et al*. (7), which are practically identical in most cases, find the most similar GO term of one protein for every GO term annotated to the other protein and then take the average of these best matches. Wang *et al*. (14) developed a new semantic similarity measure that distinguishes different types of relationships. However, regarding the functional similarity measure, they have adapted a methodology closely related to Zhang *et al*. and Schlicker *et al*. All described methods measure the functional similarity according to only one of the three GO ontologies. To our knowledge, the only measure combining different ontologies into a single similarity score is the FunSim approach by Schlicker *et al*. (7).

One important application of GO annotation and functional similarity is the analysis of gene expression data. Many methods exist for the identification of over-represented GO terms in a list of genes (15,16). Khatri and Draghici recently reviewed methods that apply GO to the analysis of microarray data (17). Other approaches use functional similarity for clustering of genes on microarrays because functionally related genes may have similar expression profiles (11,18,19). A different application of GO-based functional similarity is the prediction and validation of molecular interactions. For instance, functional similarity was found to be one of the best predictors for protein–protein interactions (20,21). In other work, it was utilized for the quality assessment of protein–protein or domain–domain interaction data (22–25). Using functional similarity values, it is also possible to

*To whom correspondence should be addressed. Tel: +49 681 9325 321; Fax: +49 681 9325 399; Email: andreas.schlicker@mpi-inf.mpg.de

derive useful confidence thresholds for predicted domain–domain interactions (22). Another application of functional similarity is the prioritization of putative disease genes (26–30). For example, GO-based similarity appears to be a good indicator for disease gene relatedness (27). Further uses of functional similarity include the identification of functional modules in interaction networks (31,32).

Despite the wide applicability of functional similarity measures, no comprehensive resource of similarity values exists. Some tools are available for download that can be used for computing the functional similarity of gene products (10,33–35). However, like the programs GO Graph (10) and DynGO (33), these tools require the user to build a local database or, like FSST (34), to download a large database before the functional similarity can be computed on the own computer. Furthermore, some databases allow functional similarity searches (4,9,36). The Gene Functional Similarity Search Tool (GFSST) supports queries for functionally similar proteins, but restricts the user to either the human or the mouse proteome (9). The FuSSiMeG web service reports the functional similarities between two GO terms annotated to each of them, but the results lack a combined score (4).

To overcome the described limitations of other tools, we implemented the database FunSimMat accessible from a convenient online user front-end (http://funsimmat.bioinf.mpi-inf.mpg.de/) and through a XML–RPC interface. FunSimMat offers the semantic comparison of GO terms and several search options for functionally similar proteins or protein families. FunSimMat contains precomputed functional similarity values for proteins and protein families from UniProtKB (37), Pfam (38) and SMART (39). We implemented three different semantic similarity measures and apply them in the calculation of various functional similarity scores.

MATERIALS AND METHODS

Annotation classes

The current revision of FunSimMat contains more than 4.6 million proteins and protein families. Since functional similarity measures are symmetric, roughly 10 trillion computations of functional similarity values would be required for a complete all-against-all comparison. However, not every protein or protein family is annotated with a unique combination of GO terms. Therefore, we define an annotation class to be a specific, lexically sorted, list of GO terms from one ontology. An annotation class can be identified by a unique accession number.

Each protein or protein family is assigned to three annotation classes that correspond to the annotated GO terms, one class for biological process (BPclass), one for molecular function (MFclass) and one for cellular component (CCclass). For example, the terms 'mitochondrion inheritance' (GO:0000001) and 'actin cortical patch assembly' (GO:0000147) constitute one BPclass. If A and B are two annotation classes, then all pairs of proteins p und q that belong to A and B, respectively, obtain the same functional similarity value. This decreases the amount of required computations by several orders of magnitude. In addition to the definition of BPclasses, MFclasses and CCclasses, we also defined GO annotation classes (GOclasses). Each GOclass consists of one BPclass, one MFclass and one CCclass. Theoretically, more than a trillion different GOclasses could be derived from the available BPclasses, MFclasses and CCclasses. However, only 52 493 GOclasses occur in practice, which reduces the search space considerably when comparing one protein or protein family against the complete database. The definition of annotation classes as well as the mapping of proteins and protein families to annotation classes are available for download on the website.

Data sources

As of September 2007, our MySQL database FunSimMat includes 4 629 251 proteins with GO annotations from UniProtKB release 10.5. Additionally, the database contains 8957 Pfam families (Pfam release 21.0) and 704 SMART families (from InterPro release 15) annotated to one of the proteins. Currently, the proteins and protein families can be assigned to 17 140 biological process annotation classes, 20 649 molecular function classes and 4978 cellular component classes. Since the number of annotation classes is small in contrast to the number of proteins in the database, we anticipate that our approach will scale well with the growing number of proteins and annotations that can be expected in the upcoming years. We intend to update the databases every 3 months, which takes about 1 week of computation time using two CPUs.

Semantic similarity measures

We implemented four different semantic similarity measures. These measures are based on the information content (IC) of a GO term (6). The more specific a GO term is, the smaller is its probability and the higher its IC. The probability of a GO term is defined as its relative frequency in UniProtKB:

$$\mathrm{freq}(t_1) = \mathrm{annot}(t_1) + \sum_{c \in \mathrm{children}(t_1)} freq(c),$$

$$p(t_1) = \frac{\mathrm{freq}(t_1)}{\mathrm{freq}(\mathrm{root})},$$

where $\mathrm{annot}(t_1)$ is the number of proteins annotated in UniProtKB with term t_1, and $\mathrm{children}(t_1)$ is the set of child terms of term t_1 in the GO graph. The IC of a GO term t_1 is then defined as follows:

$$\mathrm{IC}(t_1) = -\log(p(t_1)).$$

Resnik's semantic similarity measure uses the IC of the most informative common ancestor (MIA) to capture the common information shared by two terms t_1 and t_2. It is defined as follows (6):

$$\mathrm{sim}_{\mathrm{Res}}(t_1,t_2) = \mathrm{IC}(\mathrm{MIA}).$$

Lin's measure of semantic similarity additionally takes into account the differences between two terms and is defined as follows (5):

$$\mathrm{sim}_{\mathrm{Lin}}(t_1,t_2) = \frac{2\log p(\mathrm{MIA})}{\log p(t_1)+\log(t_2)},$$

where $p(t_1)$, $p(t_2)$ and $p(\mathrm{MIA})$ are the probabilities of the two terms and their most informative ancestor.

The $\mathrm{sim}_{\mathrm{Rel}}$ measure combines both semantic similarity measures and is defined as follows (7):

$$\mathrm{sim}_{\mathrm{Rel}}(t_1,t_2) = \frac{2\log p(\mathrm{MIA})}{\log p(t_1)+\log p(t_2)}(1-p(\mathrm{MIA})).$$

Jiang and Conrath defined a semantic distance measure based on the IC (40), which can be transformed into a similarity measure. The similarity is defined as follows (41):

$$\mathrm{sim}_{\mathrm{JC}}(t_1,t_2) = \frac{1}{\mathrm{IC}(t_1)+\mathrm{IC}(t_2)-2\mathrm{IC}(\mathrm{MIA})+1}.$$

Functional similarity measures

We implemented several functional similarity measures for proteins and protein families that are based on the DAG structure of GO or on the semantic similarity measures. For two proteins p and q that are annotated with two GO term sets GO^p and GO^q of sizes N and M, respectively, the different functional similarity measures are calculated and are described as follows.

The UI score is calculated from the number of terms in the intersection of the two induced graphs and the number of terms in the union of the two graphs. It is defined as follows (13):

$$\mathrm{UI}(p,q) = \frac{\left|g^p \cap g^q\right|}{\left|g^p \cup g^q\right|},$$

where g^p and g^q are the GO terms in the graphs induced by GO^p and GO^q, respectively.

Pesquita *et al.* defined a functional similarity measure based on the IC of the terms annotated to the two proteins or protein families (42):

$$\mathrm{sim}_{\mathrm{GIC}}(p,q) = \frac{\sum_{t\in g^p\cap g^q}\mathrm{IC}(t)}{\sum_{t\in g^p\cup g^q}\mathrm{IC}(t)}.$$

The functional similarity measures based on the semantic similarity scores for GO terms are calculated as follows. First, the similarity matrix S containing all pair-wise semantic similarity values (s_{ij}) between all terms GO_i^p in GO^p and all terms GO_j^q in GO^q are computed (7):

$$s_{ij} = \mathrm{sim}(\mathrm{GO}_i^p, \mathrm{GO}_j^q), \quad \forall i \in \{1,\ldots,N\}, \forall j \in \{1,\ldots,M\}$$

The row vectors and column vectors of the matrix S represent the two possible directions of comparing protein p to protein q and vice versa. The $\mathrm{GOscore}_{\mathrm{max}}$ is defined as the maximum over all s_{ij} according to Lord *et al.* (10):

$$\mathrm{GOscore}_{\mathrm{max}}(p,q) = \max s_{ij}, \quad \forall i \in \{1,\ldots,N\}, \forall j \in \{1,\ldots,M\}$$

$\mathrm{GOscore}_{\mathrm{avg}}$ is defined as the average over all s_{ij} according to Speer *et al.* (11):

$$\mathrm{GOscore}_{\mathrm{avg}}(p,q) = \frac{1}{N*M}\sum s_{ij}, \quad \forall i \in \{1,\ldots,N\}, \forall j \in \{1,\ldots,M\}$$

The average over the row maxima is defined as the rowScore, and the average over the column maxima is defined as the columnScore. This amounts to finding, for every GO term of one protein, the best-matching GO term annotated to the other protein. $\mathrm{GOscore}_{\mathrm{BM}}$ between two proteins p and q is then computed as follows according to Schlicker *et al.* (7):

$$\mathrm{GOscore}_{\mathrm{BM}}(p,q) = \max\{\mathrm{rowScore}(p,q), \mathrm{columnScore}(p,q)\}$$

$\mathrm{GOscore}_{\mathrm{max}}$, $\mathrm{GOscore}_{\mathrm{avg}}$ and $\mathrm{GOscore}_{\mathrm{BM}}$ can be computed using either of the four semantic similarity measures. The lowest similarity value is 0 for all measures. The maximum similarity is 1 except for Resnik's measure, which has no upper bound. For each protein pair or protein family pair, three different functional measures can be computed: one for biological process (BPscore), one for molecular function (MFscore) and one for cellular component (CCscore).

The GOscores quantify the similarity of two proteins or protein families according to their annotation to one ontology. Full functional similarity can be established by considering all three ontologies and the corresponding three GOscores. Joining different GOscores into one score provides an overall assessment of functional similarity. However, a composite score should be able to distinguish protein or protein family pairs that score average in all ontologies from pairs that score high in one or two ontologies and low in the other ontologies. The funSim and the rfunSim measures combine BPscore and MFscore and are defined as follows (7):

$$\mathrm{funSim}(p,q) = \frac{1}{2}\left[\left(\frac{\mathrm{BPscore}(p,q)}{\mathrm{max}_{\mathrm{BPscore}}}\right)^2 + \left(\frac{\mathrm{MFscore}(p,q)}{\mathrm{max}_{\mathrm{MFscore}}}\right)^2\right],$$

$$\mathrm{rfunSim}(p,q) = \sqrt{\mathrm{funSim}(p,q)}.$$

Here, max(BPscore) and max(MFscore) denote the maximal score for biological process and molecular function, respectively. The funSim score and the rfunSim score are computed using simRel and $\mathrm{GOscore}_{\mathrm{BM}}$. They range from 0 for no functional similarity to 1 for maximum functional similarity.

In addition, we introduce the funSimAll and the rfunSimAll scores that also take the cellular component annotation into account. They are defined as follows:

$$\mathrm{funSimAll}(p,q) = \frac{1}{3}\left[\left(\frac{\mathrm{BPscore}(p,q)}{\mathrm{max}_{\mathrm{BPscore}}}\right)^2 + \left(\frac{\mathrm{MFscore}(p,q)}{\mathrm{max}_{\mathrm{MFscore}}}\right)^2 + \left(\frac{\mathrm{CCscore}(p,q)}{\mathrm{max}_{\mathrm{CCscore}}}\right)^2\right],$$

$$\mathrm{rfunSimAll}(p,q) = \sqrt{\mathrm{funSimAll}(p,q)}.$$

Here, max(BPscore), max(MFscore) and max(CCscore) denote the maximal score for biological process, molecular function and cellular component, respectively. They range from 0 for no functional similarity to 1 for maximum functional similarity.

FunSimMat

Query options

FunSimMat consists of a web front-end and a XML–RPC interface for providing the results returned by a back-end server in a suitable matrix format. The back-end server is responsible for executing the user queries and is implemented in Java 1.5. FunSimMat offers several query options. The first option is the semantic all-against-all comparison of GO terms contained in an input list provided by the user. The GO terms have to be entered using their accession numbers, for example, GO:0000001. The results table contains the computed semantic similarity values.

The second query option is the comparison of one query protein or protein family with a list of proteins or protein families, which can be compiled in different ways. The simplest one is to enter the corresponding accession numbers into the query form of the website. It is also possible to upload a text file containing these accession numbers. Alternatively, users may use all proteins and protein families from a certain taxon by entering the corresponding NCBI Taxonomy identifier. The query form also contains a drop-down box to quickly select from common taxa. It is also possible to compare the query protein or protein family to the whole database. The computation results contain the functional similarity scores between the query protein or protein family with every protein or protein family from the list. If the user selected a taxon or the whole database, the results table contains the annotation classes corresponding to the selected proteins or protein families. By clicking on one annotation class, the user can obtain a list of selected proteins or protein families belonging to that class.

The third query option is the definition of a functional profile. A functional profile consists of a list of GO terms from biological process, molecular function or cellular component. This functional profile is treated as an annotation class and is compared to a list of proteins or protein families. The user can either choose a taxon, as in the case above, or compare the profile to the whole database. This helps finding protein and protein families that are similar to a prototype protein the user is interested in. Similar to the second query option, the results table contains the comparison between the functional profile and the annotation classes. The list of selected proteins or protein families belonging to one class can be accessed by clicking on the class identifier.

Web front-end

The web front-end provides HTML forms for all of the different query options that FunSimMat offers. The results are displayed in a table (Figure 1), and can be downloaded as tab-delimited text file or printed. The results table can be customized to a large extent. It can be sorted according to one column by clicking on the corresponding column header. The score values are colored with a gradient from white for low similarity to blue for high similarity. This gradient can be changed to red or green. Additionally, it is possible to hide and show specific columns or groups of columns, for example all biological process scores at once. However, it is important to note that the features to change the color gradient and to hide columns are only available if JavaScript is enabled. The first two columns of the table contain the GO, UniProtKB, Pfam or SMART accessions linked to the respective source database, or the annotation class accessions. Annotation class accessions are linked to a page listing all proteins and protein families belonging to this annotation class together with their complete GO annotation. Tooltips containing the GO annotation of proteins or protein families are shown when the mouse hovers above an accession. More details can be found in the help pages on the website.

XML–RPC interface

Extensible markup language remote procedure call (XML–RPC) is a protocol for accessing remote services and programs over a network. The XML–RPC interface allows for automatically querying FunSimMat over the Internet and to process the results. This interface is implemented using PHP and is available at http://funsimmat.bioinf.mpi-inf.mpg.de/xmlrpc.php. It provides the same query options as the web front-end. For instance, in order to semantically compare a list of GO terms, the function Semantic.getSemSims() can be called. It accepts the accession numbers of the GO terms as comma-delimited list and returns the rows of the results table in the form of an array. A detailed description can be found in the help pages on the website.

CONCLUSIONS

Functional similarity measures are used in many different applications like gene clustering, protein–protein interaction prediction and validation and disease gene prioritization. However, there is no comprehensive resource of functional similarity values available. Therefore, users have to compute these values themselves using one of the existing tools or their own implementation. Those tools, however, typically require large databases to be downloaded or created. Furthermore, the computation of functional similarity measures is rather time-consuming.

In order to remedy these problems, we have implemented a database of functional similarity values, FunSimMat. The database contains functional similarities between more than 4.6 million proteins and protein families. The web front-end provides several query options for flexible, simple and fast retrieval of these values. All query results are accessible online for 2 days and may be downloaded in tab-delimited files, which facilitates their use in many applications. The additional XML–RPC interface makes it possible to automatically query FunSimMat. This greatly supports the integration of

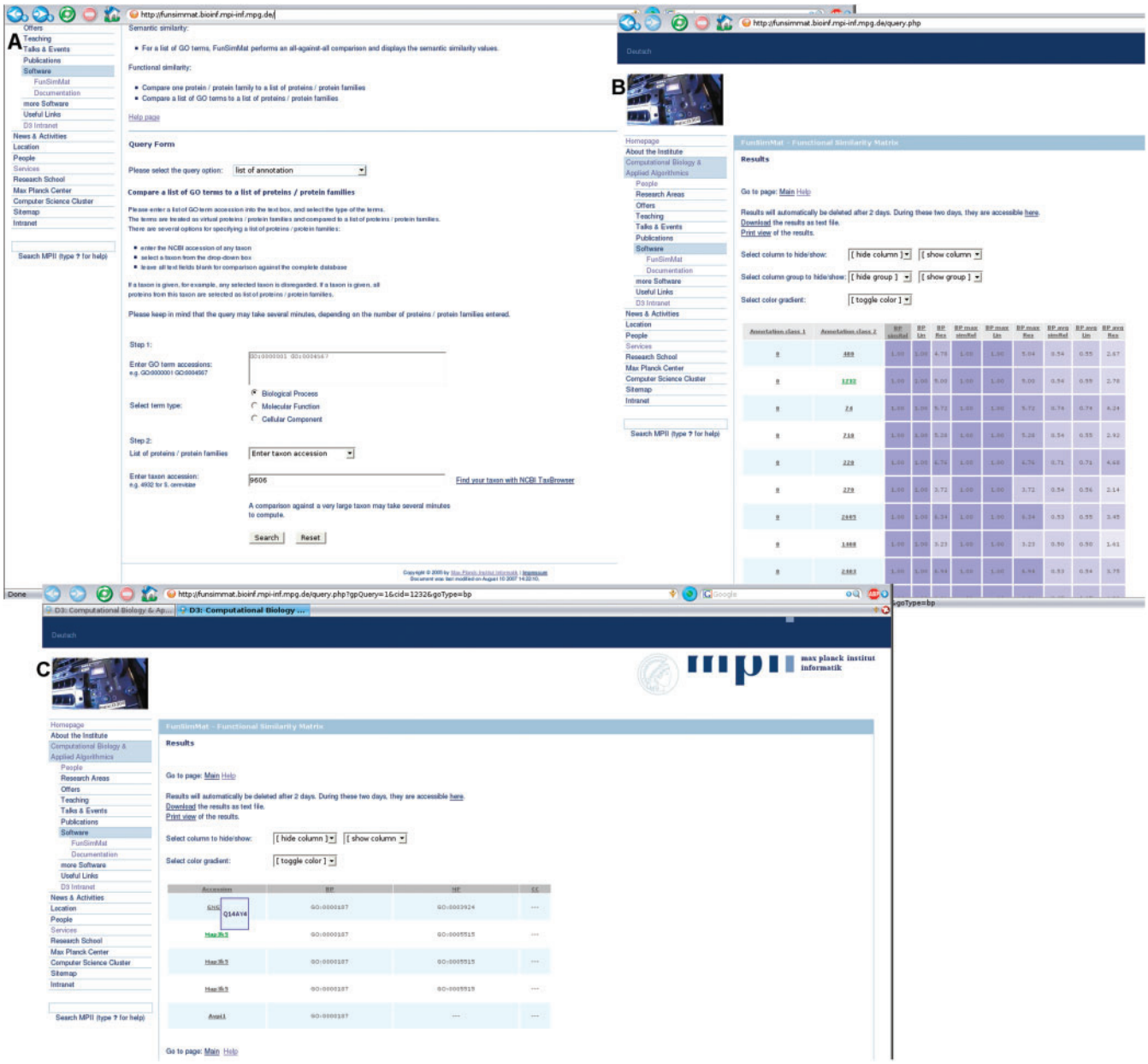

Figure 1. Application example of querying FunSimMat with a functional profile. (**A**) Query form with the user input data. (**B**) Results table listing the similarity scores for the comparison of the functional profile with different BPclasses. (**C**) Web page showing all proteins belonging to BPclass number 1232 and the corresponding GO annotations.

FunSimMat and the use of functional similarity in many existing and new data analysis pipelines and tools. Additionally, FunSimMat provides a novel way of performing rapid functional similarity searches within large protein databases. Future work will include the setup of a Distributed Annotation System (DAS) server (43) that provides functional similarity as annotation of protein–protein and domain–domain interactions.

ACKNOWLEDGEMENTS

We are grateful to Hagen Blankenburg, Dorothea Emig and Fidel Ramírez for useful comments on the manuscript and the design of the website as well as Francisco S. Domingues for helpful discussions on FunSimMat. Part of this study was financially supported by the German National Genome Research Network (NGFN) and by the German Research Foundation (DFG), contract number KFO 129/1-1. The work was conducted in the context of the BioSapiens Network of Excellence funded by the European Commission under grant number LSHG-CT-2003-503265. Funding to pay the Open Access publication charges for this article was provided by Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. *et al.* (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
2. Friedberg,I. (2006) Automated protein function prediction-the genomic challenge. *Brief Bioinform.*, **7**, 225–242.
3. Consortium,G.O. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
4. Couto,F.M., Silva,M.J. and Coutinho,P.M. (2007) Measuring semantic similarity between Gene Ontology terms. *Data Knowledge Eng.*, **61**, 137–152.
5. Lin,D. (1998) An information-theoretic definition of similarity. In Shavlik,J.W. (ed.), *Proceedings of the 15th International Conference on Machine Learning (ICML-98)*, Madison, WI, USA, Morgan Kaufmann, San Francisco, CA, USA, pp. 296–304.
6. Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, (IJCAI-95), Montreal, Canada*. Morgan Kaufmann, San Francisco, CA, USA, pp. 448–453.
7. Schlicker,A., Domingues,F., Rahnenführer,J. and Lengauer,T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.
8. Yu,H., Jansen,R. and Gerstein,M. (2007) Developing a similarity measure in biological function space. *Bioinformatics*, **23**, 2163–2173.
9. Zhang,P., Zhang,J., Sheng,H., Russo,J.J., Osborne,B. and Buetow,K. (2006) Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, **7**, 135.
10. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
11. Speer,N., Spieth,C. and Zell,A. (2004) A memetic clustering algorithm for the functional partition of genes based on the gene ontology. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), La Jolla, CA, USA*. IEEE Press, San Diego, CA, USA, pp. 252–259.
12. Wang,H., Azuaje,F., Bodenreider,O. and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), La Jolla, CA, USA*. IEEE Press, San Diego, CA, USA, pp. 25–31.
13. Guo,X., Liu,R., Shriver,C.D., Hu,H. and Liebman,M.N. (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**, 967–973.
14. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.-F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.
15. Alexa,A., Rahnenführer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
16. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
17. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
18. Brameier,M. and Wiuf,C. (2007) Co-clustering and visualization of gene expression data and gene ontology terms for *Saccharomyces cerevisiae* using self-organizing maps. *J. Biomed. Inform.*, **40**, 160–173.
19. Qu,Y. and Xu,S. (2004) Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, **20**, 1905–1913.
20. Lin,N., Wu,B., Jansen,R., Gerstein,M. and Zhao,H. (2004) Information assessment on predicting protein–protein interactions. *BMC Bioinformatics*, **5**, 154.
21. Lu,L.J., Xia,Y., Paccanaro,A., Yu,H. and Gerstein,M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.
22. Schlicker,A., Huthmacher,C., Ramírez,F., Lengauer,T. and Albrecht,M. (2007) Functional evaluation of domain–domain interactions and human protein interaction networks. *Bioinformatics*, **23**, 859–865.
23. Ramírez,F., Schlicker,A., Assenov,Y., Lengauer,T. and Albrecht,M. (2007) Computational analysis of human protein interaction networks. *Proteomics*, **7**, 2541–2552.
24. Futschik,M.E., Chaurasia,G. and Herzel,H. (2007) Comparison of human protein–protein interaction maps. *Bioinformatics*, **23**, 605–611.
25. Suthram,S., Shlomi,T., Ruppin,E., Sharan,R. and Ideker,T. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360.
26. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
27. Franke,L., van Bakel,H., Fokkens,L., de Jong,E.D., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
28. Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18**(Suppl. 2), S110–S115.
29. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.
30. Rossi,S., Masotti,D., Nardini,C., Bonora,E., Romeo,G., Macii,E., Benini,L. and Volinia,S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res.*, **34**, W285–W292.
31. Pu,S., Vlasblom,J., Emili,A., Greenblatt,J. and Wodak,S.J. (2007) Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. *Proteomics*, **7**, 944–960.
32. Sen,T., Kloczkowski,A. and Jernigan,R. (2006) Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, **7**, 355.
33. Liu,H., Hu,Z.-Z. and Wu,C.H. (2005) DynGO: a tool for visualizing and mining of Gene Ontology and its associations. *BMC Bioinformatics*, **6**, 201.
34. Schlicker,A., Rahnenführer,J., Albrecht,M., Lengauer,T. and Domingues,F.S. (2007) GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol.*, **8**, R33.
35. Fröhlich,H., Speer,N., Poustka,A. and Beissbarth,T. (2007) GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.
36. Cao,S.-L., Qin,L., He,W.-Z., Zhong,Y., Zhu,Y.-Y. and Li,Y.-X. (2004) Semantic search among heterogeneous biological databases based on gene ontology. *Acta Biochim. Biophys. Sin. (Shanghai)*, **36**, 365–370.
37. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
38. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
39. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J.R. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
40. Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In Jiang,J.J. and Conrath,D.W. (eds), *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING X)*. Tapei, Taiwan, pp. 19–33.
41. Couto,F.M., Silva,M.J. and Coutinho,P.M. (2007) Measuring semantic similarity between Gene Ontology terms. *Data Knowledge Eng.*, **61**, 137–152.
42. Pesquita,C., Faria,D., Bastos,H., Falcão,A.O. and Couto,F.M. (2007) In Stevens,R., Lord,P., McEntire,R. and Sansone,S.-A. (eds),Evaluating GO-based semantic similarity measures. Proceedings of the 10th Annual Bio-Ontologies Meeting (Bio-Ontologies 2007). Vienna, Austria, pp. 37–40.
43. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.