D244–D248 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 18 November 2009*
doi:10.1093/nar/gkp979

FunSimMat update: new features for exploring functional similarity

Andreas Schlicker* and Mario Albrecht

Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany

Received August 17, 2009; Revised October 14, 2009; Accepted October 15, 2009

ABSTRACT

Quantifying the functional similarity of genes and their products based on Gene Ontology annotation is an important tool for diverse applications like the analysis of gene expression data, the prediction and validation of protein functions and interactions, and the prioritization of disease genes. The Functional Similarity Matrix (FunSimMat, http://www.funsimmat.de) is a comprehensive database providing various precomputed functional similarity values for proteins in UniProtKB and for protein families in Pfam and SMART. With this update, we significantly increase the coverage of FunSimMat by adding data from the Gene Ontology Annotation project as well as new functional similarity measures. The applicability of the database is greatly extended by the implementation of a new Gene Ontology-based method for disease gene prioritization. Two new visualization tools allow an interactive analysis of the functional relationships between proteins or protein families. This is enhanced further by the introduction of an automatically derived hierarchy of annotation classes. Additional changes include a revised user front-end and a new RESTlike interface for improving the user-friendliness and online accessibility of FunSimMat.

INTRODUCTION

Annotations with terms from the Gene Ontology (GO) provide important information on the functions of genes and gene products (1). GO consists of three hierarchically structured vocabularies for biological process, molecular function and cellular component. Nodes in these ontologies represent terms and edges the relationships between different terms. GO annotation can be leveraged for performing functional comparisons between gene products (2–7). Simple approaches measure the functional similarity by counting the number of terms shared between different gene products (4), while more sophisticated methods utilize the semantic similarity between GO terms (3,5–7). Semantic similarity methods commonly rely on the GO structure and an annotation database for quantifying the similarity between two GO terms (5,8–10).

Many diverse applications make use of semantic and functional similarity. A number of methods were developed for analyzing gene expression data considering functional similarity (11–16). In the field of interactomics, functional similarity measures were found to be particularly useful for predicting and validating protein and domain interactions (17–19). Lately, functional similarity was incorporated into methods for prioritizing disease gene candidates (2,20–22). The GO4genome method that was recently introduced by Merkl and Wiezer applies functional similarity in the comparison of genomes for deriving a phylogeny of prokaryotic organisms (23). The Functional Similarity Matrix (FunSimMat) was utilized by Xie and colleagues for assessing the functional similarity between the cholesteryl ester transfer protein (CETP) and other proteins that are targeted by CETP inhibitors (24). Faria *et al.* (25) investigated the protein function space as described by GO using the concept of annotation classes introduced by FunSimMat.

FunSimMat (http://www.funsimmat.de) is the only publicly available comprehensive database of precalculated semantic and functional similarity values (26) for all proteins in UniProtKB (27) and protein families in Pfam (28) and SMART (29). Since its first publication, it has received over 1.4-million user queries. With the current FunSimMat release 3.1, we considerably increase the number of available GO annotations by adding data from the Gene Ontology Annotation (GOA) project (30). The introduction of a new hierarchy of annotation classes and of two visualization tools (Figure 1) affords innovative approaches for the analysis of functional similarity data by the user. More functional similarity measures, a RESTlike (31) web interface, and further performance optimizations were implemented for enhancing the usability of FunSimMat. Furthermore, we provide a new method for prioritizing disease gene candidates using FunSimMat and included information from OMIM (32) about proteins known to be involved in diseases.

*To whom correspondence should be addressed. Tel: +49 681 9325 329; Fax: +49 681 9325 399; Email: andreas.schlicker@mpi-inf.mpg.de

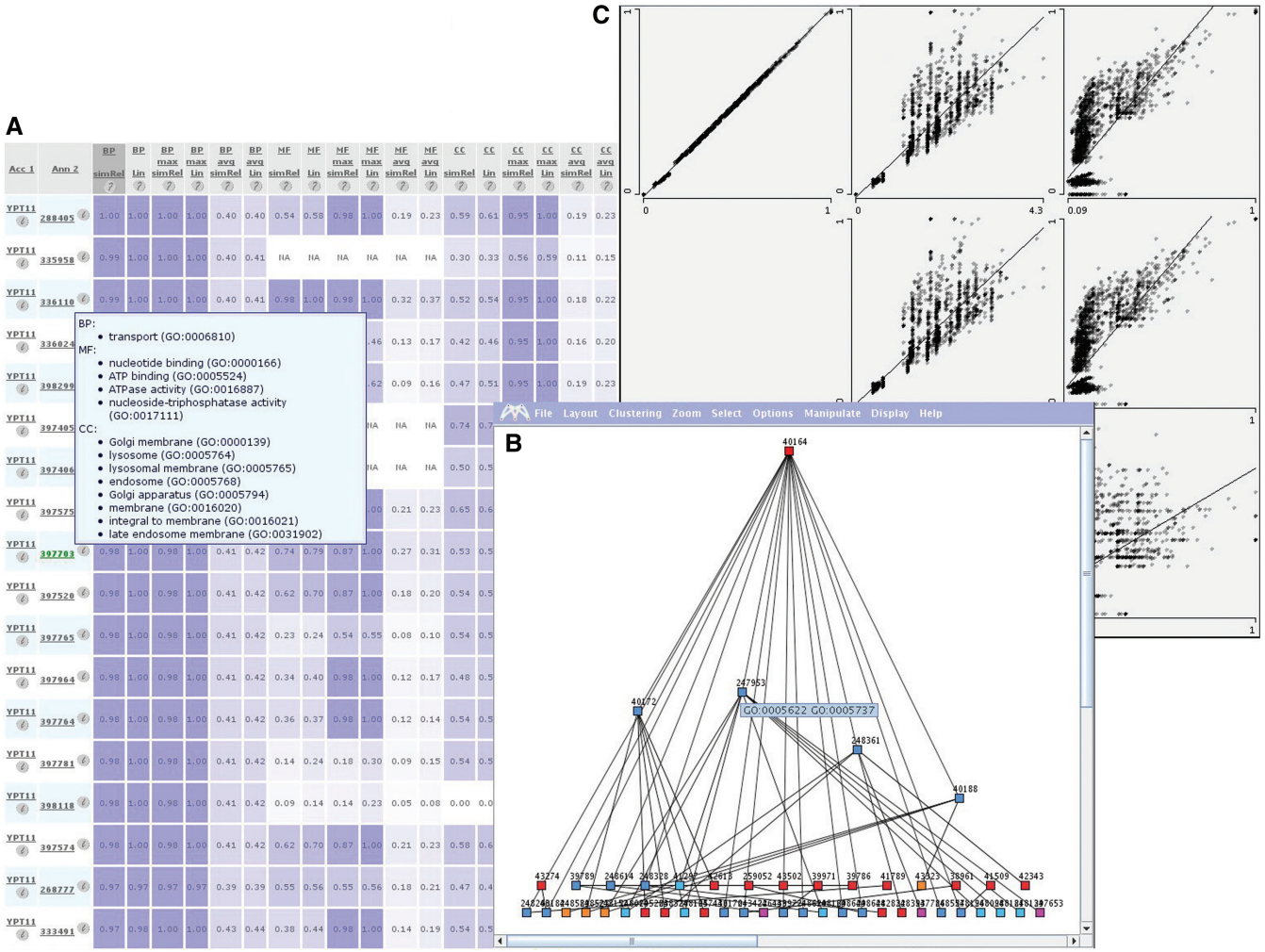

Figure 1. Different visualization options for a result set provided by FunSimMat. The figure shows some of the results obtained by the functional comparison of GTP-binding protein YPT11 (UniProtKB P48559) with GO annotation superclasses of human proteins. (**A**) The results table lists all functional similarity scores of the query protein with different GOclasses. Each table cell is colored by a gradient; white color represents no similarity and blue color high similarity. The popup box gives all GO terms for the GOclass 397703. (**B**) Medusa visualization of some CCclasses contained in the results. The classes were clustered using the k-means algorithm with k set to 20 and placed by applying a hierarchical layout. The nodes are colored according to cluster membership. (**C**) Mondrian scatter plots that compare biological process similarities obtained by different semantic similarity measures. The three plots in the first row show, for example, that the results obtained with *simRel* (5) are strongly correlated with Lin's similarity (8) (left), less correlated with Resnik's similarity (10) (center), and only weakly correlated with scores computed using Jiang & Conrath's similarity (9) (right). The straight lines in the scatter plots are least-squares regression calculated by Mondrian.

This greatly expands the applicability of FunSimMat in biomedical research.

DATA SETS

The current FunSimMat release 3.1 contains almost 8.4-million proteins from UniProtKB (release 15.3) and ~26.9-million GO annotations of proteins extracted from UniProtKB and from GOA (release of May 2009). Additionally, FunSimMat includes over 10 000 Pfam families (release 23) and 720 SMART families (from InterPro release 20). The annotations of protein families with GO terms were derived from the pfam2go and smart2go mapping files (both from April 2009). The database also contains 19 481 entries from OMIM (downloaded on 10 June 2009). In total, release 3.1 of the FunSimMat database is 326 GB in size, which is almost four times the size of the previous release.

EXTENDING ANNOTATION CLASSES

FunSimMat eliminates data redundancy and improves computational efficiency by introducing annotation classes, which subsume all proteins and protein families that are annotated with the same set of GO terms. An annotation class is defined as a unique, lexically sorted list of GO terms from a single ontology and can be identified by a unique accession number, which is stable between database releases. There are three types of annotation classes: BPclass (biological process), MFclass (molecular function) and CCclass (cellular component).

Each protein and protein family is assigned to the annotation classes that exactly correspond to its annotated GO terms. Ancestors of annotated GO terms are not included in the annotation classes because the various functional similarity scores account for the GO structure. A GOclass represents a combination of one BPclass, one MFclass and one CCclass, and each protein and protein family is associated with the GOclass that corresponds to its BPclass, MFclass and CCclass. The all-against-all comparison of proteins and protein families is performed by computing the functional similarity values between all possible pairs of annotation classes.

Previous releases of FunSimMat were built using protein GO annotations from UniProtKB only. The increased availability of GO annotations and the inclusion of data from GOA almost doubled the number of available annotations between proteins and GO terms. This provides a significantly larger coverage as well as an improved functional characterization of proteins and protein families sharing similar functions. This is signified by the number of annotation classes in the current release, which is four times higher than in the previous release: 47 538 BPclasses, 59 814 MFclasses, 18 753 CCclasses and 151 151 GOclasses. Many of these classes differ by a single term only, which results in a very high functional similarity between them.

In order to exploit this relatedness, we introduce hierarchically structured networks of annotation classes for biological process, molecular function and cellular component. In these networks, nodes represent annotation classes and two classes, c_1 and c_2, are connected by an edge if the following two conditions are satisfied: (i) all terms from c_1 are contained in c_2, and (ii) c_2 contains exactly one additional term. The second condition restricts the number of edges in the network and prevents it from becoming too complex. Annotation classes consisting of solely one term constitute the source nodes in the network. The most specific classes that are not contained in any other class are defined as annotation superclasses.

The newly established hierarchy of annotation classes enables refining comparisons of a specific protein or protein family with a list of proteins or families. The user can restrict the query to superclasses and thus concentrate on the largest functional differences. By including all annotation classes in a subsequent query, it is possible to obtain a comprehensive overview for identifying smaller differences in functional similarity.

TOOLS FOR VISUALIZING RESULT SETS

FunSimMat provides two basic query options: (i) semantic all-against-all comparison of GO terms and (ii) functional comparison of a query protein or protein family with a list of proteins or protein families. The result sets from both query types are summarized in a table (Figure 1), which provides special means for easily investigating the similarity between a pair of GO terms, proteins, or protein families in detail. However, if the query result set is large, a visual analysis may be advantageous for quickly obtaining an overview. Therefore, we offer two new tools for displaying and analyzing FunSimMat results (Figure 1). The first tool Mondrian allows a comprehensive statistical analysis of the result set (33). It has the particular functionality of drawing different types of plots, for instance, scatter plots, bar charts, box plots, and histograms. Various plots can be opened simultaneously and compared directly, which can be used to investigate the correlation between different functional similarity scores in a specific result set. Data points selected in one plot are highlighted instantly in all other plots, which aids in studying an interesting subset of results from various perspectives. The second tool Medusa visualizes the hierarchical relationships between the annotation classes contained in the result set from functional comparisons (34). Users can apply different layout and cluster algorithms for discovering relationships between annotation classes in the result set. Furthermore, it is possible to search for all classes that contain selected GO terms. The original implementations of both tools were modified to enable their deployment using Java Web Start. Both are started by clicking on the corresponding link on the results page, and the result set is then loaded. Plots generated by both tools can be saved in various bitmap and vector image formats.

NEW FUNCTIONAL SIMILARITY MEASURES

Previously, most functional similarity scores were based on semantic similarity between GO terms. In this update, we included two recently published scores that are based on the number of overlapping terms, the term overlap (TO) and the normalized term overlap (NTO). For two proteins p and q that are annotated with the GO term sets GO^p and GO^q, respectively, the term overlap score is defined as follows (4):

$$\mathrm{sim}_{\mathrm{TO}} = |g^p \cap g^q|,$$

where g^p and g^q are the sets of GO terms in the ontology subgraphs induced by GO^p and GO^q, respectively, excluding the root terms. The NTO score is defined as term overlap divided by the size of the smaller one of the two GO term sets (4):

$$\mathrm{sim}_{\mathrm{NTO}} = \frac{|g^p \cap g^q|}{\min(|g^p|, |g^q|)},$$

where g^p and g^q are defined as in the case of the TO score. Both scores range from 0, for no similarity to positive infinity, and larger scores indicate higher similarity.

DISEASE GENE PRIORITIZATION

Recently, we developed a new method for prioritizing disease gene candidates based on functional similarity (Schlicker *et al.*, submitted). Our MedSim approach exploits GO annotation of genes or proteins known to be involved in a disease of interest and uses functional similarity for ranking candidate genes or proteins. Briefly, MedSim prioritizes candidates in two steps. First, GO terms are transferred automatically from UniProtKB proteins cross-referenced to OMIM diseases

to the corresponding OMIM entry. Second, the list of candidates is ranked by functional similarity between the candidate proteins and the disease of interest. Candidates with higher functional similarity are more likely to be involved in the disease of interest. In order to implement our prioritization method in FunSimMat, each disease was mapped to the annotation classes matching the transferred GO terms, and all functional similarity values between human proteins and the diseases were precomputed. This allows the use of FunSimMat for the fast prioritization of a list of candidates by entering the OMIM accession number of the disease of interest and the list of UniProtKB accessions of the candidate proteins.

FURTHER IMPROVEMENTS

RESTlike interface

Two different interfaces have been available for accessing FunSimMat, the web front-end for manual queries and the XML-RPC interface for automatically accessing FunSimMat. In addition, we now provide a RESTlike interface, which supports the same query options as the other two front-ends, but all query parameters are specified inside an URL. In this way, web links for querying FunSimMat can be added easily to other web sites and services. A detailed description of the available URL parameters is given in the online documentation of FunSimMat.

More technical optimizations

A functional similarity query in FunSimMat compares a query protein to a list of proteins. This list can be defined in several ways, for instance, by entering the corresponding accession numbers or by selecting a specific taxon. Additionally, it is now possible to compare the query protein to all proteins associated with an OMIM entry by entering the accession number of the disease. To focus on certain results, users can choose to receive a specified number of results with the highest similarity.

Furthermore, we added a link to the results page for modifying a previous query. After clicking on the link, the query form is loaded with all the information that was previously entered for performing the query. This also enables sharing the query link with colleagues or bookmarking specific queries and re-running them, for instance, after a database update. Further improvements of the FunSimMat web site concern the use of the results table and the online documentation.

Internal programmatic optimizations accelerate considerably building and accessing the FunSimMat database. Thus the response time to large user queries was reduced from several minutes to seconds. Although the database size almost quadrupled to currently over 300 GB, the computation time for updating FunSimMat was decreased from about one week to only two days. This will allow frequent database updates in the future even if the number of available GO annotations continues to rise.

CONCLUSIONS

The expanding availability and accumulation of GO annotation will provide increasingly detailed functional information on genes and gene products. The described inclusion of the GOA project as a new source of GO annotation in FunSimMat increases significantly its coverage of functional annotation. Notably, the achieved performance improvements in database design and access allow FunSimMat to efficiently cope with the expected future increase in functional annotation. The additional implementation of a new method for disease gene prioritization and of functional similarity measures also broadens the scope and applicability of FunSimMat considerably. Furthermore, the introduction of a hierarchy of annotation classes and of visual analysis tools affords innovative ways of analyzing large sets of functional similarity results, while the new RESTlike interface now supports accessing FunSimMat simply by parameterized query URLs.

FUNDING

German National Genome Research Network (NGFN) (contract number 01GR0453, partial); German Research Foundation (DFG) (contract number KFO 129/1-2, partial). The work was conducted in the context of the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction and the BioSapiens Network of Excellence funded by the European Commission under grant number LSHG-CT-2003-503265. Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
2. 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.
3. 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.
4. Mistry,M. and Pavlidis,P. (2008) Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, **9**, 327.
5. 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.
6. 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.
7. Pesquita,C., Faria,D., Bastos,H., Ferreira,A.E.N., Falcão,A.O. and Couto,F.M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9(Suppl 5)**, S4.
8. Lin,D. (1998) An information-theoretic definition of similarity, In Proceedings of the 15th International Conference on Machine Learning (ICML-98), Madison, WI, USA. Morgan Kaufmann, San Francisco, CA, USA, pp. 296–304.

9. Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING X). Tapei, Taiwan, pp. 19–33.
10. 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.
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. 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.
13. Qu,Y. and Xu,S. (2004) Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, **20**, 1905–1913.
14. Yang,D., Li,Y., Xiao,H., Liu,Q., Zhang,M., Zhu,J., Ma,W., Yao,C., Wang,J., Wang,D. *et al.* (2008) Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265–271.
15. Cho,Y.-R., Zhang,A. and Xu,X. (2009) Semantic similarity based feature extraction from microarray expression data. *Int. J. Data Min. Bioinform.*, **3**, 333–345.
16. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
17. Ramírez,F., Schlicker,A., Assenov,Y., Lengauer,T. and Albrecht,M. (2007) Computational analysis of human protein interaction networks. *Proteomics*, **7**, 2541–2552.
18. 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.
19. 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.
20. Chen,J., Aronow,B.J. and Jegga,A.G. (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 73.
21. Ortutay,C. and Vihinen,M. (2009) Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.*, **37**, 622–628.
22. Yilmaz,S., Jonveaux,P., Bicep,C., Pierron,L., Smaïl-Tabbone,M. and Devignes,M.D. (2009) Gene-disease relationship discovery based on model-driven data integration and database view definition. *Bioinformatics*, **25**, 230–236.
23. Merkl,R. and Wiezer,A. (2009) GO4genome: a prokaryotic phylogeny based on genome organization. *J. Mol. Evol.*, **68**, 550–562.
24. Xie,L., Li,J., Xie,L. and Bourne,P.E. (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.*, **5**, e1000387.
25. Faria,D., Pesquita,C., Couto,F.M. and Falcão,A.O. (2009) GOclasses: molecular function as viewed by proteins. In Lord,P., Shah,N., Sansone,S.-A., Stephens,S. and Soldatova,L. (eds), *The 12th Annual Bio-Ontologies Meeting*, http://bio-ontologies.org.uk/download/Bio-Ontologies2009.pdf pp. 29–32.
26. Schlicker,A. and Albrecht,M. (2008) FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res.*, **36**, D434–D439.
27. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
28. Sammut,S.J., Finn,R.D. and Bateman,A. (2008) Pfam 10 years on: 10,000 families and still growing. *Brief. Bioinform.*, **9**, 210–219.
29. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
30. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
31. Fielding,R.T. and Taylor,R.N. (2002) Principled design of the modern Web architecture. *ACM Trans. Internet Technol.*, **2**, 115–150.
32. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
33. Theus,M. (2002) Interactive Data Visualization using Mondrian. *J. Statist. Software*, **7**, 1–9.
34. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.