D540–D544 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 17 November 2009*
doi:10.1093/nar/gkp1026

The Negatome database: a reference set of non-interacting protein pairs

Pawel Smialowski[1,2], Philipp Pagel[1,2], Philip Wong[1], Barbara Brauner[1], Irmtraud Dunger[1], Gisela Fobo[1], Goar Frishman[1], Corinna Montrone[1], Thomas Rattei[2], Dmitrij Frishman[1,2,*] and Andreas Ruepp[1]

[1]Institute for Bioinformatics and Systems Biology/MIPS, HMGU—German Research Center for Environmental Health Ingolstaedter Landstrasse 1, 85764 Neuherberg and [2]Germany Department of Genome Oriented Bioinformatics, Technische Universitaet Muenchen Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

Received October 14, 2009; Revised October 19, 2009; Accepted October 20, 2009

ABSTRACT

The Negatome is a collection of protein and domain pairs that are unlikely to be engaged in direct physical interactions. The database currently contains experimentally supported non-interacting protein pairs derived from two distinct sources: by manual curation of literature and by analyzing protein complexes with known 3D structure. More stringent lists of non-interacting pairs were derived from these two datasets by excluding interactions detected by high-throughput approaches. Additionally, non-interacting protein domains have been derived from the stringent manual and structural data, respectively. The Negatome is much less biased toward functionally dissimilar proteins than the negative data derived by randomly selecting proteins from different cellular locations. It can be used to evaluate protein and domain interactions from new experiments and improve the training of interaction prediction algorithms. The Negatome database is available at http://mips.helmholtz-muenchen.de/proj/ppi/negatome.

INTRODUCTION

Protein–protein interactions are a crucial part of the majority of biological processes. A vast array of high- and low-throughput methods are currently used to expand our knowledge about protein interaction networks (1). Many of these experimental techniques are inherently noisy and suffer from relatively high false positive and negative rates [see, e.g. Huang and Bader (2)].

To some extent, the problem of noisy and contradictory results can be tackled by integrating heterogeneous data within a rigorous machine-learning or statistical framework (3,4). The availability of a high-quality standard of truth is often crucial for the validation of new interaction datasets and machine learning approaches. Positive trusted data describing high-confidence interaction pairs usually stem from careful literature curation efforts. Extensive gold standard datasets are currently available for several model organisms, including yeast (5) and human (6). In contrast, the datasets containing experimentally confirmed non-interacting protein pairs (NIPs) are presently quite sparse (7). The lack of negative training data represents a significant problem because the knowledge about NIPs is as important for developing and evaluating prediction algorithms as the knowledge of true positive pairs (7–9).

A common practice in constructing negative 'gold-standards' is to randomly select pairs of proteins having different cellular localization and/or involved in different biological processes (3,4,10–13). As demonstrated by Ben-Hur *et al.* (14), the latter approach can lead to over-optimistic estimation of method performance. Restricting negative data only to pairs of proteins localized in different cellular compartments allows for the creation of protein sets enriched in non-interacting pairs, but such pairs may introduce substantial functional bias hurting downstream analyses and predictions. The use of such data for building a classifier can result primarily in predictions of protein co-localization. The fact that interacting protein pairs have to be in the same place and time does not imply that all proteins in the same compartment will be interacting with each other. Furthermore, localization to different cellular compartments does not exclude physical binding in all cases: many proteins involved in functional interactions re-locate to different compartments during their life cycle and interactions between

*To whom correspondence should be addressed. Tel: +49 8161 712134; Fax: +49 8161 712186; Email: d.frishman@wzw.tum.de

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

compartments exist, facilitated by organelle membrane proteins, which have the ability to engage in interactions on both sides of the organelle boundary.

The ratio of interacting protein pairs to all possible pairs has been estimated to be below 1%. For example, in *Saccharomyces cerevisiae* ~6000 proteins allow for ~18 million potential pairwise interactions but the true number of interactions has been estimated to be well below 100 000 (15,16). Hence, proteins sharing the same cellular compartment and/or the same biochemical process are not necessarily interacting with each other. Another possible consequence of reducing the negative set to differentially localized proteins is that sequence composition variability between interacting and non-interacting datasets will be artificially high (14). Such bias in sequence composition can lead to over-optimistic performance estimations for sequence-based protein–protein interaction prediction methods. One possible way to address the problem of bias is to use a negative set constructed by random sampling of proteins from a given organism regardless of their localization. Based on the estimated low number of interacting pairs, such dataset will be fairly enriched in negative data but still contain a small background of interacting pairs.

In this publication, we describe two complementary efforts to construct reliable negative interaction datasets. One effort involves the collection of evidence against physical interactions from literature, focusing only on those cases where the lack of interaction between two proteins was experimentally validated by an individual experiment. In parallel, we analyzed complexes consisting of three or more proteins deposited in the PDB (Protein Data Bank) (17) and derived a set of protein pairs that, while being in immediate vicinity in the context of a protein complex, do not interact directly with each other. The resulting database, which we call the Negatome, is freely available from http://mips.helmholtz-muenchen.de/proj/ppi/negatome.

DATABASE CONTENT

The Negatome comprises several datasets based on literature evidence and structural information (Table 1). A representation focusing on non-interacting protein domains is available as well. Experimental evidence was annotated according to the standards established for protein-interaction experiments in the PSI-MI format (18).

Structurally non-interacting protein pairs

From experimental structures of biological units as provided by PDB (17), we derived our non-interacting pairs as follows. First, for each biological unit hosting more than two protein chains, we measured inter-chain distances between all Cβ atoms (Cα for glycine) using the CCP4 software package (19,20). A pair of protein chains was declared to be non-interacting if all inter-chain distances were more than 8 Å (Supplementary Figure S1). Pairs that were nearest neighbors to each other in terms of inter-chain distances and pairs mapping to same UniProt (21,22) accession number were removed. For example, the PDB structure 1U0N (Supplementary Figure S1) contains four chains: A, B, C and D corresponding to von Willebrand factor, botrocetin alpha chain, botrocetin beta chain and platelet glycoprotein Ib. The physically interacting pairs are A–B, A–C, A–D, B–C and C–D. Chains B and D do not interact and are not nearest neighbors, therefore we claim that those two proteins do not interact.

A total of 809 non-interaction pairs were derived for the PDB dataset (Table 1). Because non-interacting pairs derived from these structures may be a consequence of non-observed electron density, truncation or modification of the proteins to allow for crystallization, or other experimental conditions, which do not occur naturally, we performed additional filtering to derive a second dataset, PDB-stringent, by removing interacting protein pairs as described in the IntAct database (23). After IntAct filtering, we were left with 745 NIPs (termed the PDB-stringent dataset) (Table 1). For all protein pairs from PDB-stringent, we list PDB chain ids and UniProt accessions of associated full-length proteins (http://mips.helmholtz-muenchen.de/proj/ppi/negatome).

Manually curated non-interacting pairs

Annotation of the manual dataset was performed analogous to the annotation of protein–protein interactions and protein complexes in previous projects published by our group (24,25). Information about NIPs was extracted from scientific literature using only data from individual experiments but not from high-throughput experiments. Only mammalian proteins were considered. Data from high-throughput experiments were omitted in order to maintain the highest possible standard of reliability. Since negative results are usually of low scientific

Table 1. Overview of the Negatome datasets

Dataset name	Derived from	Description	Number of pairs
PDB	The PDB database	Protein pairs that are members of at least one structural complex but do not interact directly. Organism of origin is not restricted.	809
PDB-stringent	PDB	The PDB dataset filtered against the IntAct dataset.	745
PDB-PFAM	PDB-stringent	Non-interacting PFAM domains found in the same structural complex, filtered as described in 'Methods' section.	458
Manual	Manual literature annotation	Manually annotated literature data describing the lack of protein interaction. High-throughput data are not included. The data is restricted only to mammalian proteins.	1291
Manual-stringent	Manual	The Manual dataset filtered against the IntAct dataset.	1162
Manual-PFAM	Manual-stringent	PFAM domain pairs found in the Manual dataset filtered as described in 'Methods' section.	523

importance for the authors, this kind of data is inherently difficult to find. Full-text searches in journals using terms like 'not interact' reveal large numbers of articles which, upon closer inspection, do not provide explicit experimental evidence supporting the non-interaction status of protein pairs. Such data are mostly generated from investigations of protein–protein interactions and protein complexes.

We focused on the non-interacting pairs selected manually from the scientific articles where they have been used as control experiments or where they appear as a result of testing multiple proteins as potential interactors of target proteins. Often scientists carefully choose negative controls for such experiments to be present in the compartment of interest and/or to be involved in relevant processes.

For example, Snyder *et al.* (26) reported that β-synuclein regulates proteasome activity by interaction with α-synuclein but does not interact with proteasomal subunit S6. In another study, it was shown that Fbxl3 controls clock oscillations by mediating the degradation of the two-cryptochrome proteins Cry1 and Cry2 (27). Immunoprecipitation experiments of Cry proteins with nine further F-box proteins revealed that only Fbxl3 was able to co-immunoprecipitate with Cry1 and Cry2 and the other nine proteins were not interacting with Fbxl3.

A total of 246 articles were used for the generation of the Manual dataset. Due to the relatively large size of the dataset, there is no strong bias toward certain functional systems or cellular locales. In addition to UniProt primary accessions of the non-interacting proteins, experimental method and the PMID (PubMed identification number) of the respective experiment are given in this dataset. We also provide the Manual-stringent dataset obtained by filtering literature derived data against known interaction pairs from the IntAct database and removing pairs involving the same proteins. The Manual and Manual-stringent datasets contain 1291 and 1162 pairs, respectively. Interestingly, the overlap between PDB-stringent and Manual-stringent dataset is only 15 pairs.

Non-interacting domain pairs

We also provide datasets of non-interacting PFAM domains derived from the PDB-stringent and the Manual-stringent dataset, respectively. We mapped proteins to PFAM (28) domains using cross-references from the UniProt database (21). We assume that the PFAM domains residing in the non-interacting PDB amino acid chains do not interact. However, since chains in the PDB do not always contain full-length proteins, interacting domains might be missed. To account for this possibility, we removed all domain–domain pairs found in interacting protein pairs from IntAct. We make a generous assumption that all domains are interacting with each other if they belong to interacting proteins. We also subtract all pairs of known interacting PFAM domains as defined in the 3DID (29) and iPFAM (30) databases. In summary, the number of unique non-interacting PFAM domain pairs provided in the Negatome is 458 and 523 for the PDB-PFAM and Manual-PFAM datasets, respectively. Two domain pairs are common between PDB-PFAM and Manual-PFAM.

DATA ANALYSES

Comparing non-interacting pairs with predictions from STRING

The STRING database (31) aggregates vast amounts of data and predictions of protein–protein associations and interactions including the evidence based on physical binding, genetic and functional context, experimental data and text-mining results. We mapped our NIPs against the STRING using a 100% sequence identity threshold. Only a small fraction (13.8, 9.3, 8.9 and 8.3% for PDB, PDB-stringent, Manual and Manual-stringent, respectively) of our non-interacting pairs is functionally associated by STRING. Most of these associations are by 'text-mining' (Manual: 85%, Manual-stringent: 86.9%, PDB: 75.4%, PDB-stringent: 81.3% of the total number of pairs associated by STRING) and 'experimental' (Manual: 57.5%, Manual-stringent: 52.3%, PDB: 71.4%, PDB-stringent: 56.5%) evidences (Supplementary Table S1). Association by 'text-mining' can be misleading as the names of NIPs derived from manual annotation appear in the same text by definition. Likewise, proteins co-occurring in the same structural complex have a very high chance to be described together in publications. The 'experimental' evidence may contain a significant number of false positive interactions since it is derived from many high-throughput experiments.

Two methods measuring shared evolutionary pressure ('neighborhood' and 'co-occurrence') support association of <27 and 7% of NIPs found in STRING derived from the PDB and Manual datasets, respectively. More frequent association of PDB-derived non-interacting pairs can be explained by tighter evolutionary constraint between proteins sharing the same complex. NIPs from the PDB dataset are additionally associated by 'database' evidence (PDB: 72%, PDB-stringent: 65%) as they are frequently part of the same functional complex and commonly share the same metabolic pathway.

Interestingly, there is higher support for association of PDB non-interacting proteins by the 'co-expression' evidence compared with Manual dataset (~60% versus 4%), which contrasts to the generally poor levels of co-expression of complex members found in yeast (32). More detail analysis (see Supplementary Data).

Will a growing interaction dataset wipe the Negatome?

A possible criticism of our database could be that with more and more complete knowledge of the interactome, an ever-growing fraction of our NIPs will be proven wrong. We attempted to estimate the rate at which our data will be falsified in the near future. For protein interaction data, we used dates provided by IntAct, and for our non-interaction pairs the dates of PDB deposition and PubMed entry creation, respectively. By counting the number of non-interaction pairs known at each given time point from 1995 to 2007 (Supplementary Figure S2), we found that the percentage of PDB non-interacting

pairs contradicted by protein–protein interaction data grew substantially from 0% in 1999 to 7% in 2002 where it stabilized, oscillating between 6.5% and 8%. For the Manual dataset, the percentage of contradicted interactions stabilized after increasing from 0 to 2% in 1999 and from 2 to 6% in 2004. The growth rate of our stringent datasets allows us to believe that an increasing number of non-interacting pairs will be available in the foreseeable future. Finding interactions between a pair of proteins does not necessarily falsify a NIP. Differences in conditions in which experiments are carried out may explain differences in propensities of proteins to interact. Our database provides lists of negative protein interactions which, when compared with newly discovered interactions between the same proteins, may help discern conditions preventing or promoting such interactions.

Functional similarity between non-interacting proteins

Interacting proteins are involved in some common biological function (16). Randomly picked pairs and even more so pairs randomly assembled from proteins found in different compartments are less likely to contain proteins with the same specific function. In order to assess the extent of this common functional bias, we computed a functional similarity score between interacting proteins from IntAct, non-interacting proteins pairs constructed based on non-colocalization and our NIPs. Protein pairs with differential cellular localization were derived using localization data from the DBSubLoc database (33) and filtered against protein interactions from the IntAct database. Functional similarity between proteins was computed using three graph-based similarity measures developed by Resnik (34), GraSM (35) and Jiang-Conrath (36) implemented in the GOSim package (37) version 1.1.5.4 written in R language (version 2.9.1) (38) and using GO.db database version 2.2.11. Computations were carried out with the biological process sub-tree of Gene Ontology (39). For reasons of computational feasibility, similarity for protein pairs from IntAct was computed for a subset of 5000 randomly picked pairs. As expected, on average, interacting protein pairs were found to have a higher functional similarity than protein pairs from different cellular compartments. The Manual subset of NIP showed a score distribution similar to IntAct, while the highest degree of functional similarity was found in the PDB-derived data (Supplementary Figures S3 and S4). More details can be found in Supplementary Data.

Coevolution of interacting and non-interacting protein and domain pairs

We profiled the presence or absence of domains found in known interacting pairs (from iPFAM and 3DID) and those in our non-interaction datasets across 460 genomes (40). We found that domains had significantly greater co-occurrence in these organisms if they were interacting compared to those in non-interacting pairs (Supplementary Table S3). This result did not change even if we profiled against various subsets of these 460 genomes. Detail description of methods used for co-evolution analysis and more detail results are described in Supplementary Data.

Non-interacting proteins are part of the interaction network

Interestingly, proteins that constitute our non-interacting datasets (PDB-stringent, Manual-stringent) have similar average numbers of interacting partners (PDB-stringent 5.37; SD 35.51; Manual-stringent 6.85; SD 19.92) as any other proteins in IntAct (6.88; SD 18.37) indicating that they are not generally biased against interaction. These results show that proteins in our database can engage in interactions with many other partners but interactions have not been observed for certain pairs.

CONCLUSIONS

At the time of writing, we provide a total of 1892 NIPs and 979 predicted non-interacting domain pairs based on the experimental evidence. Phylogenetic profiling of domains showed that pairs of known interacting domains had much tighter domain coevolution compared to our sets of non-interacting domain pairs in terms of coordinated presence or absence of domains across a set of species. This result agrees with the idea that correlated evolution can be helpful for predicting interactions. Further analyses showed that the mean functional similarity between our non-interacting proteins (Manual, PDB datasets) is higher than the similarity between proteins interacting according to IntAct and much higher than the similarity within pairs generated by randomly selecting individual proteins from different cellular locations. The non-interacting pairs derived from PDB show the highest mean functional similarity because each pair belongs to a common protein complex. While the use of randomly generated negative pairs, as well as of those, where proteins are selected from different cellular locations can be helpful to train classifiers for protein–protein interactions, our data should be a valuable contribution as it is not as biased toward functionally dissimilar pairs of proteins as these former types of data. Our data can be useful for assessing the quality of new experimentally extracted protein interaction datasets. The non-interacting PDB pairs, in particular, should be beneficial for predicting protein interactions within the same complex. Our time-course analysis of the Negatome and IntAct databases suggests that in spite of a growing fraction of contradicting pairs between both sets, the absolute number of non-interacting pairs in our gold-standard set is constantly increasing. In summary, the Negatome resource is expected to complement current popular approaches for training predictors of protein–protein interaction.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The Biosapiens Network of Excellence (grant number LHSG-CT-2003-503265). Funding for open access charge: Institute for Bioinformatics and Systems Biology, Neuherberg, Germany.

Conflict of interest statement. None declared.

REFERENCES

1. Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.
2. Huang,H. and Bader,J.S. (2009) Precision and recall estimates for two-hybrid screens. *Bioinformatics*, **25**, 372–378.
3. Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21(Suppl. 1)**, i38–i46.
4. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
5. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
6. Kandasamy,K., Keerthikumar,S., Goel,R., Mathivanan,S., Patankar,N., Shafreen,B., Renuse,S., Pawar,H., Ramachandra,Y.L., Acharya,P.K. *et al.* (2009) Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res.*, **37**, D773–D781.
7. Li,X.L., Tan,S.H. and Ng,S.K. (2006) Improving domain-based protein interaction prediction using biologically significant negative datasets. *Int. J. Data Min. Bioinform.*, **1**, 138–149.
8. Browne,F., Wang,H., Zheng,H. and Azuaje,F. (2009) GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction. *Source Code Biol. Med.*, **4**, 2.
9. Sanchez-Graillet,O. and Poesio,M. (2007) Negation of protein-protein interactions: analysis and extraction. *Bioinformatics*, **23**, i424–i432.
10. Jansen,R. and Gerstein,M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.*, **7**, 535–545.
11. Chen,X.W. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.
12. Shen,J., Zhang,J., Luo,X., Zhu,W., Yu,K., Chen,K., Li,Y. and Jiang,H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.
13. Guo,Y., Yu,L., Wen,Z. and Li,M. (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.
14. Ben-Hur,A. and Noble,W.S. (2006) Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, **7(Suppl. 1)**, S2.
15. Grigoriev,A. (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.*, **31**, 4157–4161.
16. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale datasets of protein-protein interactions. *Nature*, **417**, 399–403.
17. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
18. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
19. Collaborative_Computational_Project. (1994) The CCP4 suite: programs for protein crystallography. *Acta. Crystallogr. D Biol. Crystallogr.*, **50**, 760–763.
20. Winn,M.D. (2003) An overview of the CCP4 project in protein crystallography: an example of a collaborative project. *J. Synchrotron. Radiat.*, **10**, 23–25.
21. UniProt-Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
22. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt knowledgebase. *Methods Mol. Biol.*, **406**, 89–112.
23. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
24. Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stumpflen,V., Mewes,H.W. *et al.* (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.
25. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.
26. Snyder,H., Mensah,K., Hsu,C., Hashimoto,M., Surgucheva,I.G., Festoff,B., Surguchov,A., Masliah,E., Matouschek,A. and Wolozin,B. (2005) beta-Synuclein reduces proteasomal inhibition by alpha-synuclein but not gamma-synuclein. *J. Biol. Chem.*, **280**, 7562–7569.
27. Busino,L., Bassermann,F., Maiolica,A., Lee,C., Nolan,P.M., Godinho,S.I., Draetta,G.F. and Pagano,M. (2007) SCFFbxl3 controls the oscillation of the circadian clock by directing the degradation of cryptochrome proteins. *Science*, **316**, 900–904.
28. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
29. Stein,A., Russell,R.B. and Aloy,P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.
30. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
31. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
32. Liu,C.T., Yuan,S. and Li,K.C. (2009) Patterns of co-expression for protein complexes by size in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **37**, 526–532.
33. Guo,T., Hua,S., Ji,X. and Sun,Z. (2004) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Res.*, **32**, D122–D124.
34. Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *14th International Conference Research on Computational Linguistics*. IJCAI-95, Montreal, Canada, Vol. 1, pp. 448–453.
35. Couto,F.M., Silva,M.J. and Coutinho,P.M. (2005) Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In *14th ACM International Conference on Information and Knowledge Management*. ACM, Bremen, Germany, pp. 343–344.
36. Jiang,J. and Conrath,D. (1998) *International Conference Research on Computational Linguistics*. ROCLING X, Taiwan, Vol. 1.
37. Froehlich,H. (2008) GOSim package (version 1.1.5.4). 1.1.5.4 ed.
38. R_Development_Core_Team. (2009) R language (version 2.9.1). Vienna, Austria.
39. 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.
40. Pagel,P., Wong,P. and Frishman,D. (2004) A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.*, **344**, 1331–1346.