BMC Bioinformatics

BioMed Central

Database

DBMLoc: a Database of proteins with multiple subcellular localizations

Song Zhang†, Xuefeng Xia†, Jincheng Shen, Yun Zhou and Zhirong Sun*

Address: MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China

Email: Song Zhang - zhangsong04@mails.tsinghua.edu.cn; Xuefeng Xia - xxf04@mails.tsinghua.edu.cn; Jincheng Shen - shenjc04@mails.tsinghua.edu.cn; Yun Zhou - zhouyun03@mails.tsinghua.edu.cn; Zhirong Sun* - sunzhr@mail.tsinghua.edu.cn

* Corresponding author †Equal contributors

Published: 28 February 2008

BMC Bioinformatics 2008, **9**:127 doi:10.1186/1471-2105-9-127

Received: 22 July 2007
Accepted: 28 February 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/127

Abstract

Background: Subcellular localization information is one of the key features to protein function research. Locating to a specific subcellular compartment is essential for a protein to function efficiently. Proteins which have multiple localizations will provide more clues. This kind of proteins may take a high proportion, even more than 35%.

Description: We have developed a database of proteins with multiple subcellular localizations, designated DBMLoc. The initial release contains 10470 multiple subcellular localization-annotated entries. Annotations are collected from primary protein databases, specific subcellular localization databases and literature texts. All the protein entries are cross-referenced to GO annotations and SwissProt. Protein-protein interactions are also annotated. They are classified into 12 large subcellular localization categories based on GO hierarchical architecture and original annotations. Download, search and sequence BLAST tools are also available on the website.

Conclusion: DBMLoc is a protein database which collects proteins with more than one subcellular localization annotation. It is freely accessed at http://www.bioinfo.tsinghua.edu.cn/DBMLoc/index.htm.

Background

Knowledge of subcellular localization is crucial to understanding protein function and biological process. During translation or later, proteins will be transported into different compartments such as cytoplasm, membrane system, mitochondrion, etc., or may be secreted out of the cell. Locating to a specific subcellular compartment is essential for a protein to function efficiently. High-throughput experimental approaches like immuno-localization[1], tagged genes and reported fusions[2,3] have made the growth of localization data catch up with the avalanche of protein data. Swiss-Prot is a comprehensive database which includes subcellular localization information. In the recent years, some specific subcellular localization databases are constructed based on experimentation, computational prediction or both. The subcellular localization data of LOCATE[4] are from high-throughput immunofluorescence-based assay and publications. Organelle DB[5] annotates all protein localizations using vocabulary from the Gene Ontology consortium which facilitates data interoperability. DBSubLoc[6] uses a keyword-based system to integrate

Swiss-Prot subcellular localization annotations. LOCtarget[7] and PA-GOSUB[8] implement predictors of subcellular localization based on different methods have been reported. PSORTdb[9] is a database for bacteria that contains both information determined through laboratory experimentation (ePSORTdb) and computational predictions (cPSORTdb). Eukaryotic database, eSLDB[10], collects five species' location data which are experimental-determined, homology-based or predicted. In addition, some bioinformatics methods have been developed to predict the protein subcellular location, which make use of the sorting signals[11], domain information[12], amino acid composition in the sequences [13-15] or other information[16].

However, a lot of proteins have more than one subcellular localization annotations. These proteins may simultaneously locate or move between different cellular compartments, for example, transcription factors and signaling pathway transduction factors. Proteins may play different roles in biological process when they are in different subcellular localizations. For these proteins, single subcellular localization annotation will lose some important information. Usually these proteins have more important biological functions. Their localization annotations will provide more valuable clues to researchers. These proteins are quite common, accounting for about 39% of all organellar proteins in mouse liver[17]. However, there are very few proteins annotated with multiple locations in the available subcellular localization databases. Here we have built the database DBMLoc which collects proteins with multiple subcellular localization annotations. It provides useful information for protein functional research as well as computational prediction. In addition, taxonomy, Swiss-Prot, GO and interaction information are also annotated. If protein has interactions, a subcellular localization quality score is computed on the basis of its interaction proteins' locations.

Construction and content

The DBMLoc database is mainly developed from primary protein databases (Swiss-Prot/TrEMBL[18]), available experimental-determined subcellular localization databases (DBSubloc[6], ePSORTdb[9], MitoProteome[19], Organelle DB[5] and LOCATE[4]) and some literature references. Only full-length and unambiguous proteins are selected from Swiss-Prot, and those whose subcellular localization annotations are marked with "by similarity", "probable", "possible", "potential", "may be" are excluded. At the same time, multiple annotations are collected from subcellular localization databases (DBSubloc, ePSORTdb, MitoProteome, Organelle DB and LOCATE), then they are mapped to the protein set derived from Swiss-Prot. The redundant annotations are filtered. In order to standardize subcellular localization annotation terms, various terms of cellular compartments and complexes are assigned into twelve large organelle categories as follows: extracellular, cell wall, membrane, cytoplasm, mitochondrion, nucleus, ribosome, plastid, endoplasmic reticulum, Golgi apparatus, vacuole and virion. Cell wall, plastid and vacuole are unique in plant cell. Some subcellular localization annotations which can not be classified into the twelve categories are assigned into "others". There are 616 proteins that have "others" annotations. This process is mainly based on the Gene Ontology[20] annotations and original subcellular localization annotations. We annotate the proteins with GO ID from their primary sources or the annotation tools provided by GOA (Gene Ontology Annotation Database)[21]. The proteins are also cross-referenced to the NCBI Taxonomy database[22]. Sub-datasets are derived based on their taxonomy class (i.e. animal, plant, eukaryote, etc.)

Proteins that interact with each other tend to share the same subcellular localizations, so we annotate the protein with interaction data collected from DIP[23], MINT[24] and BIND[25]. To check the subcellular localization annotation quality, if it has interaction proteins, a quality score is computed based on the following formula. The higher the score is, the more reliable the subcellular localization annotations are. All the proteins whose score equals 1 are integrated into a high quality dataset.

$$Score = \frac{N1}{N2}$$

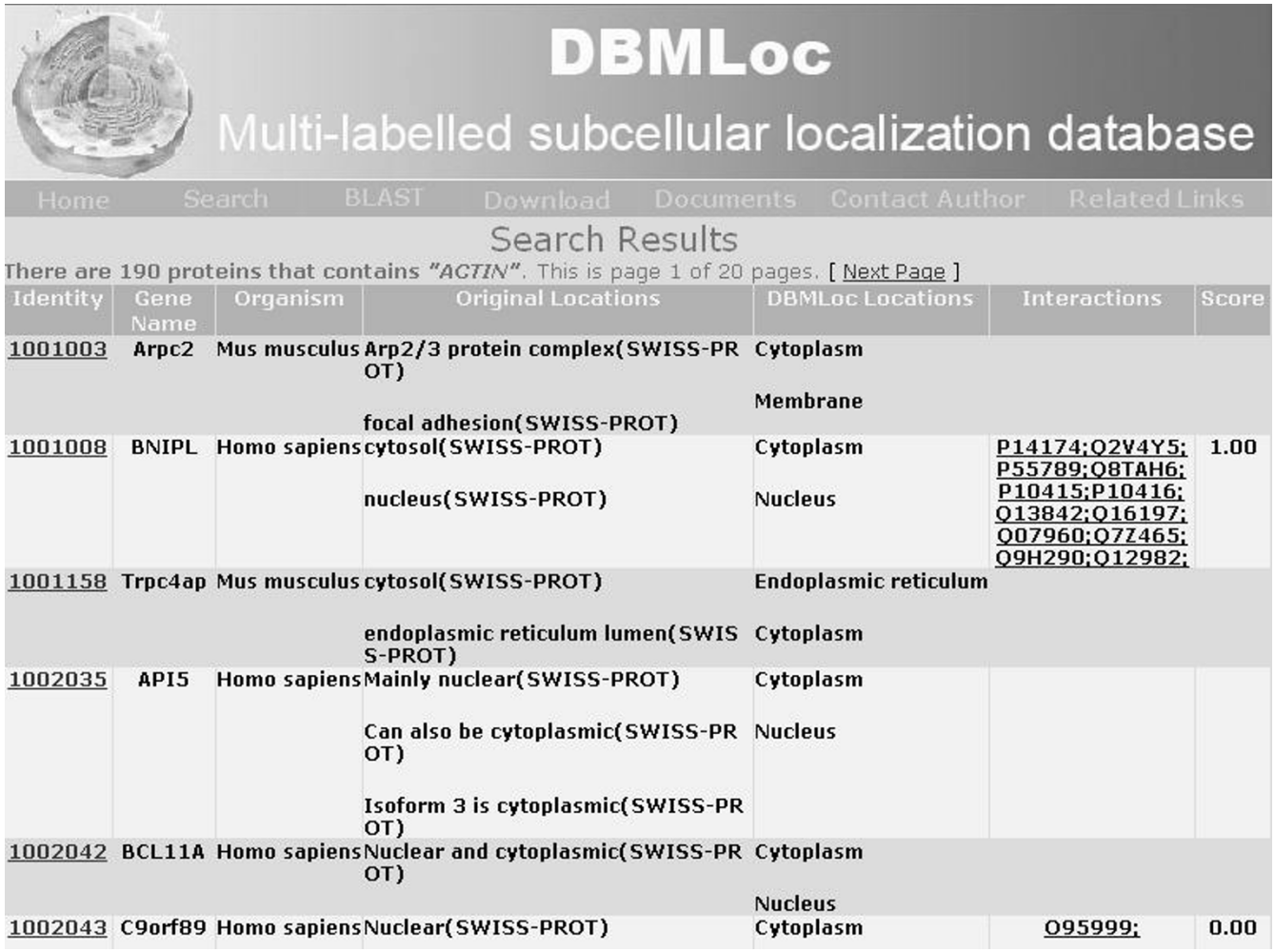
N1: Number of the localizations shared by its interaction proteins' subcellular localizations.

N2: Number of protein's subcellular localizations.

Finally, with some literature annotated proteins added, 10470 protein entries are integrated into DBMLoc database. The downloadable DBMLoc database and non-redundant sub-datasets are released as plain text files. The format is similar to that of Swiss-Prot data file. Each line in the file is one record of an entry in the 'KEY VALUE' format. The cross-reference records begin with a 'CX' key. Each of the value data contains one cross-reference record in the 'Reference Database: Reference ID' format, for example, the 'CX SWISS-PROT: Q85FL3' record means that the protein entry is linked to SWISS-PROT database Q85FL3 entry. More detailed description of the format can be found on the web page.

Utility and discussion

We provide free download of the database, organism specific sub-datasets and taxonomy-categorized files for all the education and research users. Users can search the database with DBMLoc identity, cross-referenced database



DBMLoc

Multi-labelled subcellular localization database

Home Search BLAST Download Documents Contact Author Related Links

Search Results

There are 190 proteins that contains *"ACTIN"*. This is page 1 of 20 pages. [Next Page]

Identity	Gene Name	Organism	Original Locations	DBMLoc Locations	Interactions	Score
1001003	Arpc2	Mus musculus	Arp2/3 protein complex(SWISS-PROT) focal adhesion(SWISS-PROT)	Cytoplasm Membrane		
1001008	BNIPL	Homo sapiens	cytosol(SWISS-PROT) nucleus(SWISS-PROT)	Cytoplasm Nucleus	P14174;Q2V4Y5; P55789;Q8TAH6; P10415;P10416; Q13842;Q16197; Q07960;Q7Z465; Q9H290;Q12982;	1.00
1001158	Trpc4ap	Mus musculus	cytosol(SWISS-PROT) endoplasmic reticulum lumen(SWISS-PROT)	Endoplasmic reticulum Cytoplasm		
1002035	API5	Homo sapiens	Mainly nuclear(SWISS-PROT) Can also be cytoplasmic(SWISS-PROT) Isoform 3 is cytoplasmic(SWISS-PROT)	Cytoplasm Nucleus		
1002042	BCL11A	Homo sapiens	Nuclear and cytoplasmic(SWISS-PROT)	Cytoplasm Nucleus		
1002043	C9orf89	Homo sapiens	Nuclear(SWISS-PROT)	Cytoplasm	O95999;	0.00

Figure 1
Protein name search result with keyword "actin".

identity or protein name. Figures 1 and 2 show the name and identity search results. Protein sequence also can be submitted to search for homologous proteins in the full DBMLoc database or in one of its subsets.

The initial release contains 10470 multiple subcellular localization-annotated protein entries. Non-redundant protein data sets with sequence similarity less than 90% and 25% are also generated by BLAST. Table 1 lists brief statistical information on full and non-redundant data sets. The detailed statistical information is on the web page.

Various databases' annotations integrated together in DBMLoc database might be false annotations or conflicts. So, we will pay more attention to the quality of data in the future development. More experimental data and other available information, like experimental method and post-translation modification, will be integrated to the database. The database will be updated regularly as new version of Swiss-Prot is available. Besides, more web services and analysis tools will be developed.

Conclusion

DBMLoc is a specific database aimed at multiple localization annotated proteins. Proteins are cross-referenced to NCBI taxonomy, Gene Ontology and original database. Proteins that interact with each other tend to share the same subcellular localizations. So, protein-protein interaction information is also integrated into the database. A quality score is derived from protein-protein interactions. These data will be valuable to help experimental and computational biologists understand and analyze biological function.

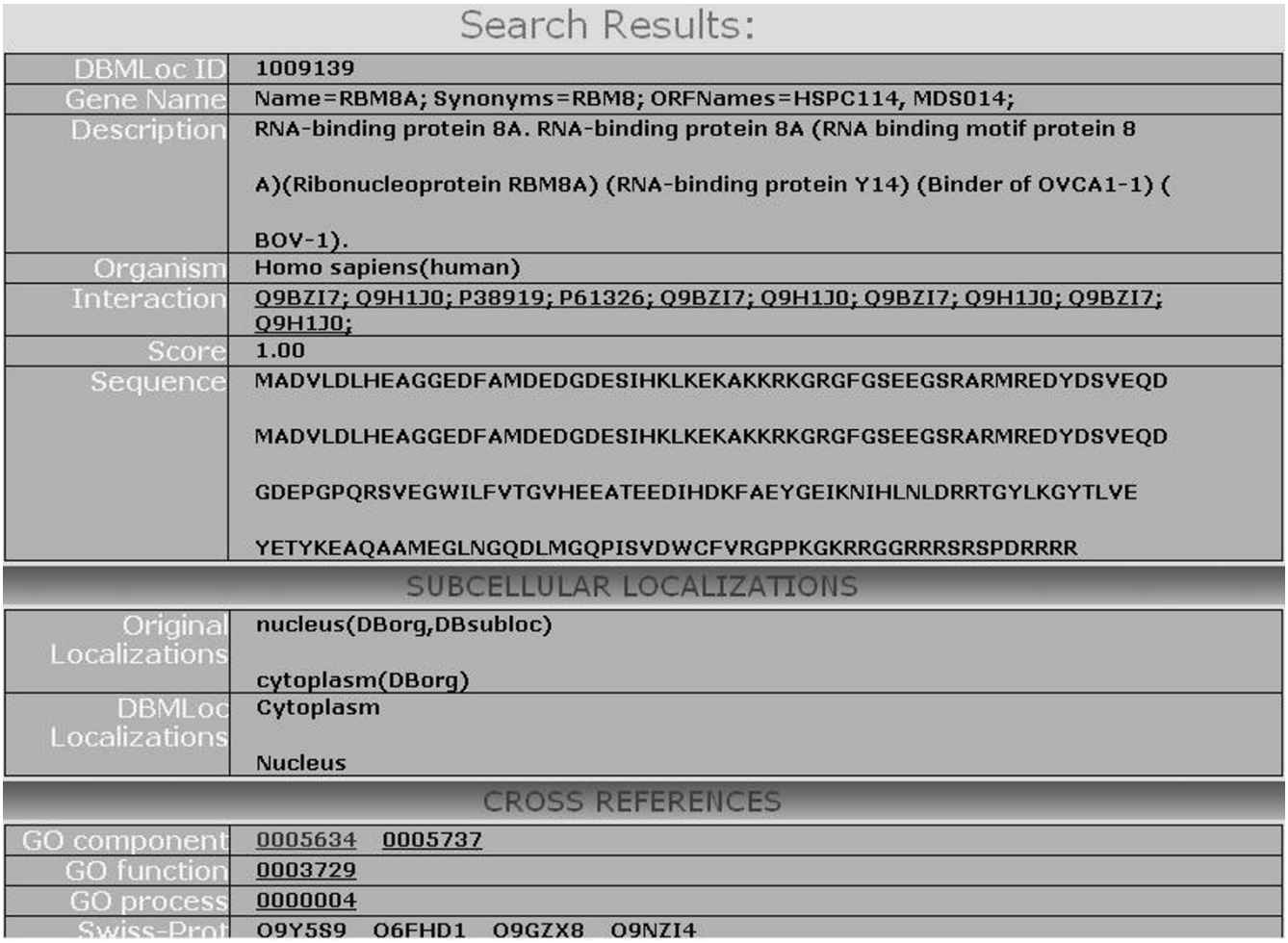


Figure 2
Swiss-Prot identity search result with query "Q9Y5S9".

Availability and requirements

DBMLoc home page: http://www.bioinfo.tsinghua.edu.cn/DBMLoc/index.htm

License: The database is freely available.

List of abbreviations

GO: Gene Ontology.

Authors' contributions

SZ and XX designed and constructed the database. SZ drafted the manuscript. JS and YZ participated in data curation. ZS supervised the project. All authors read and approved the final manuscript.

Table 1: Brief statistics of DBMLoc

	Full data sets	Non-redundant data sets (90%)	Non-redundant data sets (25%)
Two subcellular localizations	8887	6055	2366
Three subcellular localizations	1461	1112	593
Four subcellular localizations	107	100	85
Eukaryote	9954	6727	2549
Animal	6492	4240	1523
Plant	3462	2487	1278

Acknowledgements

This project was supported in part by the National Natural Science Grant in China 863 (no.2006AA020403), 973(no.2003CB715900) and the National Natural Science Grants (no.30770498).

References

1. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425(6959):**686-691.
2. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, Cheung KH, Miller P, Gerstein M, Roeder GS, Snyder M: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16(6):**707-719.
3. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L, Heidtman M, Nelson FK, Iwasaki H, Hager K, Gerstein M, Miller P, Roeder GS, Snyder M: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999, **402(6760):**413-418.
4. Fink JL, Aturaliya RN, Davis MJ, Zhang F, Hanson K, Teasdale MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD: **LOCATE: a mouse protein subcellular localization database.** *Nucleic Acids Res* 2006, **34(Database issue):**D213-7.
5. Wiwatwattana N, Kumar A: **Organelle DB: a cross-species database of protein localization and function.** *Nucleic Acids Res* 2005, **33(Database issue):**D598-604.
6. Guo T, Hua S, Ji X, Sun Z: **DBSubLoc: database of protein subcellular localization.** *Nucleic Acids Res* 2004, **32(Database issue):**D122-4.
7. Nair R, Rost B: **LOCnet and LOCtarget: sub-cellular localization for structural genomics targets.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W517-21.
8. Lu P, Szafron D, Greiner R, Wishart DS, Fyshe A, Pearcy B, Poulin B, Eisner R, Ngo D, Lamb N: **PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization.** *Nucleic Acids Res* 2005, **33(Database issue):**D147-53.
9. Rey S, Acab M, Gardy JL, Laird MR, deFays K, Lambert C, Brinkman FS: **PSORTdb: a protein subcellular localization database for bacteria.** *Nucleic Acids Res* 2005, **33(Database issue):**D164-8.
10. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **eSLDB: eukaryotic subcellular localization database.** *Nucleic Acids Res* 2007, **35(Database issue):**D208-12.
11. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Int J Neural Syst* 1997, **8(5-6):**581-599.
12. Mott R, Schultz J, Bork P, Ponting CP: **Predicting protein cellular localization using a domain projection method.** *Genome Res* 2002, **12(8):**1168-1174.
13. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS: **PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.** *Nucleic Acids Res* 2003, **31(13):**3613-3617.
14. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 2001, **17(8):**721-728.
15. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular location of proteins.** *Nucleic Acids Res* 1998, **26(9):**2230-2236.
16. Sarda D, Chua GH, Li KB, Krishnan A: **pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties.** *BMC Bioinformatics* 2005, **6:**152.
17. Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, Mann M: **A mammalian organelle map by protein correlation profiling.** *Cell* 2006, **125(1):**187-199.
18. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31(1):**365-370.
19. Cotter D, Guda P, Fahy E, Subramaniam S: **MitoProteome: mitochondrial protein sequence database and annotation system.** *Nucleic Acids Res* 2004, **32(Database issue):**D463-7.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.
21. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32(Database issue):**D262-6.
22. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28(1):**10-14.
23. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic Acids Res* 2000, **28(1):**289-291.
24. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513(1):**135-140.
25. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31(1):**248-250.