BMC Bioinformatics

BioMed Central

Proceedings

Open Access

SPIDer: *Saccharomyces* protein-protein interaction database

Xiaomei Wu†, Lei Zhu†, Jie Guo, Cong Fu, Hongjun Zhou, Dong Dong, Zhenbo Li, Da-Yong Zhang and Kui Lin*

Address: MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing 100875, China

Email: Xiaomei Wu - wuxm@mail.bnu.edu.cn; Lei Zhu - zhulei@mail.bnu.edu.cn; Jie Guo - guojie@mail.bnu.edu.cn; Cong Fu - fucong.cmb@gmail.com; Hongjun Zhou - honglin@mail.bnu.edu.cn; Dong Dong - ddseesea@mail.bnu.edu.cn; Zhenbo Li - limengci@hotmail.com; Da-Yong Zhang - zhangdy@bnu.edu.cn; Kui Lin* - linkui@bnu.edu.cn

* Corresponding author †Equal contributors

from International Conference in Bioinformatics – InCoB2006
New Dehli, India. 18–20 December 2006

Published: 18 December 2006

BMC Bioinformatics 2006, **7**(Suppl 5):S16 doi:10.1186/1471-2105-7-S5-S16

Abstract

Background: Since proteins perform their functions by interacting with one another and with other biomolecules, reconstructing a map of the protein-protein interactions of a cell, experimentally or computationally, is an important first step toward understanding cellular function and machinery of a proteome. Solely derived from the Gene Ontology (GO), we have defined an effective method of reconstructing a yeast protein interaction network by measuring relative specificity similarity (RSS) between two GO terms.

Description: Based on the RSS method, here, we introduce a predicted *Saccharomyces* protein-protein interaction database called SPIDer. It houses a gold standard positive dataset (GSP) with high confidence level that covered 79.2% of the high-quality interaction dataset. Our predicted protein-protein interaction network reconstructed from the GSPs consists of 92 257 interactions among 3600 proteins, and forms 23 connected components. It also provides general links to connect predicted protein-protein interactions with three other databases, DIP, BIND and MIPS. An Internet-based interface provides users with fast and convenient access to protein-protein interactions based on various search features (searching by protein information, GO term information or sequence similarity). In addition, the RSS value of two GO terms in the same ontology, and the inter-member interactions in a list of proteins of interest or in a protein complex could be retrieved. Furthermore, the database presents a user-friendly graphical interface which is created dynamically for visualizing an interaction sub-network. The database is accessible at http://cmb.bnu.edu.cn/SPIDer/index.html.

Conclusion: SPIDer is a public database server for protein-protein interactions based on the yeast genome. It provides a variety of search options and graphical visualization of an interaction network. In particular, it will be very useful for the study of inter-member interactions among a list of proteins, especially the protein complex. In addition, based on the predicted interaction dataset, researchers could analyze the whole interaction network and associate the network topology with gene/protein properties based on a global or local topology view.

Background

Protein-protein interactions play a key role in many biological processes and therefore, are increasingly forming the focus of studies in understanding protein function and cellular behaviour [1,2]. Genome-scale protein interaction networks have now been determined experimentally for several model organisms [3-12]. However, these interaction datasets are often incomplete and noisy [13,14]. Therefore, several *in silico* interaction predictions have been designed as complementary to existing experimental approaches, which are derived from gene context analysis, such as gene neighbourhood, gene fusion events, phylogenetic profiles and correlated mRNA expression patterns. Detailed reviews of these methods can be found elsewhere [15,16]. Currently, the interaction predictions have been improved by integrating different genomic features based on a single probabilistic framework in yeast [17,18] and human [19].

Originating from the Gene Ontology (GO), which organizes biological information for molecular function (MF), biological process (BP) and cellular component (CC) for different model organisms [20], we have defined a new metric of relative specificity similarity (RSS) to predict the functional association of two proteins [21]. Using both CC and BP ontologies and their respective annotations from Organelle DB [22], we derived a gold standard positive (GSP) dataset with the highest confidence of RSS values and then reconstructed a protein-protein interaction network [21]. A schematic explanation of the prediction of interaction network is shown in Figure 1.

Using our computational method, we have now developed a web-accessible database, *Saccharomyces* Protein-protein Interaction Database (SPIDer). This database contains the predicted yeast protein-protein interaction dataset, which is based on a new release of GO and the annotations derived from SGD [23]. Moreover, it provides users with a graphical interface to visualize an interaction sub-network for a list of proteins of interest.

Construction and content

We used the MySQL database system and Perl scripting language to develop and implement our database. The graphical view of network was designed using the Graphviz tool [24]. Our server runs on the Linux operating system. The content of the database is as follows.

GSP dataset

SPIDer houses the data for yeast protein-protein interactions (GSP dataset) solely derived from the GO and yeast annotations by measuring RSS between two GO terms at the genome level (Figure 1 and Ref [21]). In previous work, we used the September 2004 release of GO and yeast annotations from Organelle DB [22] to obtain a GSP dataset, which covers 78% of the integrated high-quality interaction dataset (valid experimental interactions) [21]. Here, we used the March 2006 release of the GO and yeast protein annotations downloaded from SGD submitted on March 31, 2006. The derived GSP dataset is proved to have a high confidence level, as it covers 79.2% of the high-quality interaction dataset which was used in the previous work [21]. The current GSP dataset consists of 92 257 interactions encompassing 3600 proteins, and the reconstructed interaction network forms 23 connected components, out of which the largest contains 3527 proteins and 92 084 interactions.

A detailed help page explaining the contents of the database is available online at the website.

Cross references to other databases

SPIDer provides general links to connect predicted protein-protein interactions with four other related datasets derived from three databases, namely, DIP [25], BIND [26] and MIPS [27]. Both DIP (the release of April 2, 2006 was used here) and BIND (the release of May 21, 2005) are important databases that contain information on protein-protein interactions. MIPS provides all annotated and genome-scale protein interactions (the release of January 18, 2005), and also compiles data on protein complexes (the release of June 20, 2005), which are converted to binary interactions using the matrix model. As a result, there are 15 296 interactions in GSPs that have at least one connection to the four datasets, out of which 2951, 1431, 1135 and 14 452 interactions in our database have cross references to DIP, BIND, MIPS binary interactions and MIPS complexes, respectively.

Utility

Datasets in SPIDer can be accessed through a variety of search features.

(i) Searching by protein information: given an SGD [28] identifier, a group of interactors of the query protein is returned. The interaction table is designed to be clickable and thus directs users to more detailed information, such as cross references to other databases, and hyperlinks to the pages with evidence information retrieved from the datasets we integrated. Before visualizing the network of queried interactions dynamically, two filtering criteria – the RSS score cutoff with default 0.85 and interaction step with default 1 – can be set to navigate the final graphical display of the network. No result to any given query may be due to the absence of the query protein from the GSP dataset. This search option also allows users to specify a keyword of standard gene name, open reading frame (ORF) name or protein description and to retrieve a list of SGD entries that match the keyword. Then, one or multiple proteins of interest can be selected and used in turn as

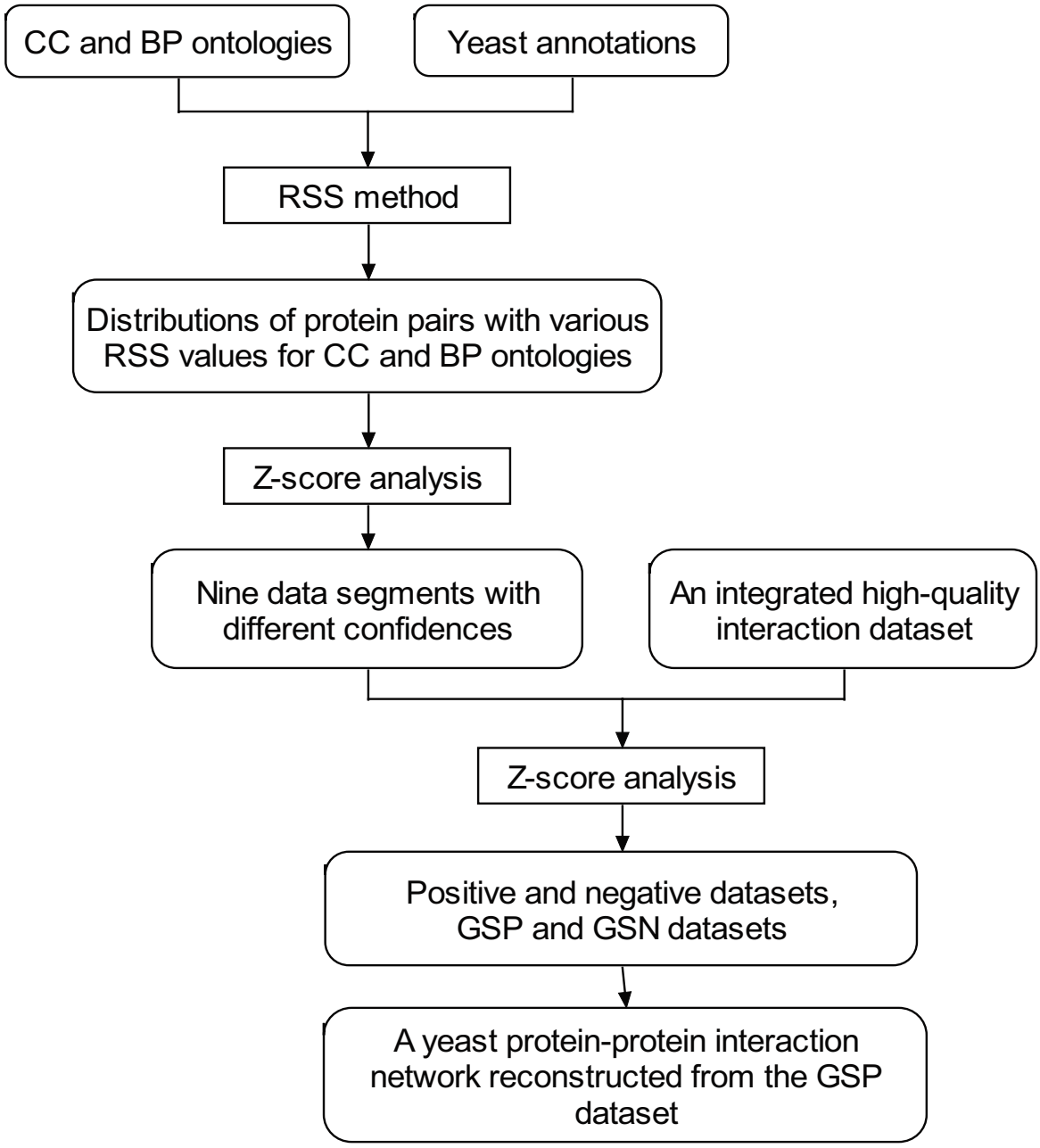

Figure 1
Schematic representation of the GO-based prediction of protein-protein interaction network. It includes four steps: (1) Using both CC and BP ontologies and their respective annotations, the distributions of protein pairs with various RSS values for the two ontologies were obtained. (2) Then a Z-score analysis was applied to draw the statistical significance of the quality scoring system and the nine data segments with different combinations of confidences were obtained. (3) To evaluate the RSS method, an integrated high-quality interaction dataset was applied. Based on a Z-score analysis, a positive dataset and a negative dataset were selected. Moreover, a gold standard positive (GSP) dataset with the highest level of confidence and a gold standard negative (GSN) dataset with the lowest level of confidence were derived. (4) Finally, a yeast protein-protein interaction network was reconstructed from the GSP dataset. The RSS method, a new metric for semantic similarity to score the degree of the functional association between two different proteins by comparing the relative specificity of pairs of GO terms assigned to them in similarity within a GO DAG [21].

A

Name				RSS			Database Cross Reference	
Protein-A		Protein-B		CC	BP	MF	Binary	MIPS_complex
SGT1 (YOR057W, S000005583)	★	CDC4 (YFL009W, S000001885)	★	0.857	0.888	0.933	DIP:2577E	445.10 550.2.358
SKP1 (YDR328C, S000002736)	★	CDC34 (YDR054C, S000002461)	★	0.923	1		BIND:170498 DIP:45139E MIPS:physical	445.10 445.20 445.30
SKP1 (YDR328C, S000002736)	★	CDC4 (YFL009W, S000001885)	★	0.923	1	1	BIND:112279 BIND:130462 BIND:170505 DIP:1808E MIPS:physical	445.10 550.2.358 550.2.44
SKP1 (YDR328C, S000002736)	★	CDC53 (YDL132W, S000002290)	★	0.857	1	1	BIND:355 BIND:209947 BIND:170497 DIP:1423E MIPS:physical	550.2.358 550.2.455 550.2.494 550.2.508 550.2.516 550.2.181 445.30 445.20 445.10 550.2.47 550.2.82

B

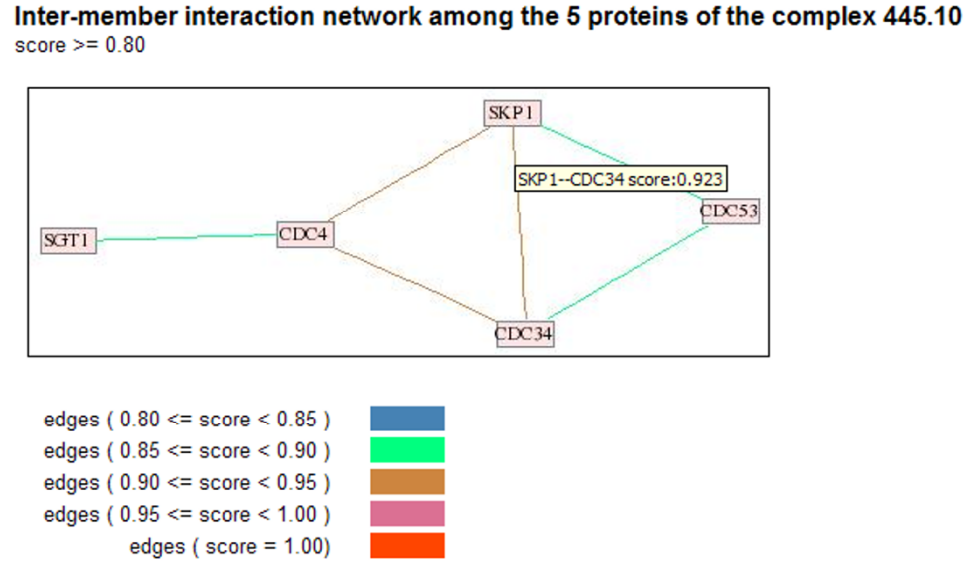

Figure 2

Sample query output and graph representation from SPIDer. (A) Partial results of inter-member interactions from the query for the SCF-CDC4 complex (search by inputting the MIPS identifier '445.10'). The table of cross-member interactions has the following information: the information for two interacting proteins, their RSS values on the three ontologies (CC, BP and MF), the cross references to other binary interaction datasets (DIP, BIND and MIPS physical interactions) and the MIPS complexes, as well as the hyperlinks to the evidence information. This table is designed to be clickable and provides hyperlinks of the proteins to SGD. Both DIP binary interactions and MIPS complexes have links to their home pages. (B) A graphical representation of the inter-member interaction network of the SCF-CDC4 complex using the default score cutoff. Nodes and edges represent proteins and cross-member interactions, respectively. Each node is labeled standard gene name. Each edge is depicted in colour representing the RSS score, which is the minimum of the BP and CC RSS values. In addition, the ORF name and primary SGD identifier are obtained when the cursor is positioned over a node, and score value over an edge (an example of the interaction between CDC34 and SKP1 is shown). Five different scores of edges in the legend are depicted in five colours.

query protein(s) for retrieving their interactors. Moreover, the selection of multiple query proteins provides access to the extraction and visualization of their inter-member interaction relationship.

(ii) Searching by GO term information: by entering a GO term identifier or a keyword of the term description, users can retrieve a list of proteins annotated on the term(s).

(iii) Retrieving the RSS value of two GO terms in the same ontology by entering their identifiers. This option allows users to access the maximum RSS value computed between the two GO terms.

(iv) Browsing inter-member interactions by entering a MIPS complex identifier or a list of proteins: we provide users with a search facility to extract the inter-member interactions of a list of proteins of interest. It allows inputting a MIPS complex identifier, and also a list of gene names, ORF names or SGD identifiers either by simply pasting them into the text box, or by uploading a plain file where proteins are separated by a space or a newline. For example, in order to identify the topology of SCF-CDC4 complex (MIPS identifier 445.10; containing five members), the identifier is entered. Figure 2 shows the search results and a graphical view of the sub-network of the complex.

(v) Searching by sequence similarity: this search feature facilitates users to retrieve protein sequences matching a query sequence or its fragment. The results are returned as a list of proteins ordered by the BLAST e-value.

Discussion

The yeast protein-protein interaction dataset will continue to be updated using the new release of GO and yeast annotations in SGD. Furthermore, in the future, we intend that our database will provide more detailed analysis results of the whole interaction network, which are useful in analyzing and understanding yeast proteome. We also expect to reconstruct the maps of interactions from other completely sequenced genomes with high-quality GO-based annotations for functional genomic research.

Conclusion

SPIDer is an open and public database server for protein-protein interactions which are solely derived from the Gene Ontology (GO), based on the yeast genome. It provides users with convenient access to protein interactions based on various search features. In particular, it allows users to analyze the potential inter-member interactions among a list of proteins of interest, which is especially useful for the analysis of protein complexes. Furthermore, it presents a graphical interface for visualizing an interaction sub-network, facilitating users to associate the network topology with gene/protein properties based on a global or local topology view.

Availability and requirements

The database is freely accessed on the internet [29]. The complete gold standard positive dataset is on request by contacting with the corresponding author.

List of abbreviations used

SGD: *Saccharomyces* Genome Database

DIP: Database of Interacting Proteins

BIND: Biomolecular Interaction Network Database

MIPS: Munich Information Center for Protein Sequences

Authors' contributions

XMW was responsible for the development and construction of the database and drafted the manuscript, while LZ designed and implemented the web interface. JG provided essential suggestions for designing the database and helped in drafting the manuscript. KL designed and coordinated the work. CF, HJZ, DD, ZBL and DYZ made their great efforts in acquiring and analyzing data, and in testing the database and the web interface. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Ministry of Education of China (Grant No. 105011), the National High-Tech Research and Development Program of China (Grant No. 2003AA231030) and Beijing Normal University.

This article has been published as part of *BMC Bioinformatics* Volume 7, Supplement 5, 2006: APBioNet – Fifth International Conference on Bioinformatics (InCoB2006). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/7?issue=S5.

References

1. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era.** *Nature* 2000, **405(6788):**823-826.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl):**C47-52.
3. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al.*: **A map of the interactome network of the metazoan C. elegans.** *Science* 2004, **303(5657):**540-543.
4. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, *et al.*: **A Protein Interaction Map of Drosophila melanogaster.** *Science* 2003, **302(5651):**1727-1736.
5. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, *et al.*: **The protein-protein interaction map of Helicobacter pylori.** *Nature* 2001, **409(6817):**211-215.
6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98(8):**4569-4574.
7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403(6770):**623-627.

8. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868):**141-147.
9. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.** *Nature* 2002, **415(6868):**180-183.
10. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al.*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
11. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
12. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, *et al.*: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13(10):**2363-2371.
13. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417(6887):**399-403.
14. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes.** *Trends Genet* 2002, **18(10):**529-536.
15. Valencia A, Pazos F: **Computational methods for the prediction of protein interactions.** *Curr Opin Struct Biol* 2002, **12(3):**368-373.
16. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M: **Analyzing cellular biochemistry in terms of molecular networks.** *Annu Rev Biochem* 2004, **73:**1051-1087.
17. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**449-453.
18. Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306(5701):**1555-1558.
19. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: **Probabilistic model of the human protein-protein interaction network.** *Nat Biotechnol* 2005, **23(8):**951-959.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.
21. Wu X, Zhu L, Guo J, Zhang DY, Lin K: **Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations.** *Nucleic Acids Res* 2006, **34(7):**2137-2150.
22. Wiwatwattana N, Kumar A: **Organelle DB: a cross-species database of protein localization and function.** *Nucleic Acids Res* 2005, **33(Database):**D598-604.
23. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, *et al.*: **Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).** *Nucleic Acids Res* 2002, **30(1):**69-72.
24. **Graphviz** [http://www.graphviz.org]
25. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32(Database):**D449-451.
26. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, *et al.*: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33(Database):**D418-424.
27. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005.** *Nucleic Acids Res* 2006, **34(Database):**D169-172.
28. Hirschman JE, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hong EL, Livstone MS, Nash R, *et al.*: **Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome.** *Nucleic Acids Res* 2006, **34(Database):**D442-445.
29. ***Saccharomyces* Protein-protein Interaction Database** [http://cmb.bnu.edu.cn/SPIDer/index.html]