Published online 1 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W491–W495***
doi:10.1093/nar/gkn241

MassNet: a functional annotation service for protein mass spectrometry data

Daeui Park[1], Byoung-Chul Kim[1], Seong-Woong Cho[1], Seong-Jin Park[1], Jong-Soon Choi[2], Seung Il Kim[2], Jong Bhak[1] and Sunghoon Lee[1,*]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-806 and [2]Proteomics Research Team, Korea Basic Science Institute, Daejeon 305-333, Korea

Received February 23, 2008; Revised April 4, 2008; Accepted April 15, 2008

ABSTRACT

Although mass spectrometry has been frequently used to identify proteins, there are no web servers that provide comprehensive functional annotation of those identified proteins. It is necessary to provide such web service due to a rapid increase in the data. We, therefore, introduce MassNet, which provides (i) physico-chemical analysis information, (ii) KEGG pathway assignment (iii) Gene Ontology mapping and (iv) protein–protein interaction (PPI) prediction for the data from MASCOT, Prospector and Profound. MassNet provides the prediction information for PPIs using both 3D structural interaction and experimental interaction deposited in PSIMAP, BIND, DIP, HPRD, IntAct, MINT, CYGD and BioGrid. The web service is freely available at http://massnet.kr or http://sequenceome.kobic.re.kr/MassNet/.

INTRODUCTION

Mass spectrometry (MS) is the key method for proteomics (1). MS is widely used to study complex cellular proteomes and low abundance proteins (1–3). With it we can rapidly identify proteins and obtain information for protein complexes and posttranslational modification (3). MS data are used to produce genome-scale data (4). Presently, the functional annotation of MS data often requires researchers to navigate numerous web-accessible primary data servers. In order to analyze large-scale data, one approach is to provide access to an integrated web server that contains rich bio-information with graphic interfaces (5). Several MS data processing systems have been developed to handle these challenges. They are MASCOT (http://www.matrixscience.com) (2), Prospector (http://prospector.ucsf.edu) and Profound (http://prowl.rockefeller.edu) (6). These systems provide protein identification data using public databases such as SwissProt (http://www.ebi.ac.uk/swissprot) and NCBInr (http://www.ncbi.nlm.nih.gov). These web services do not include the functional annotation of MS data and do not supply the latest version of the analysis tools. To provide an easy and automated pipeline for functional annotation of given MS results, we constructed a web-based server, MassNet. The use of MassNet does not require any application installation and it is easy to use.

METHODS

To analyze MS data, various protein annotation resources are required. Therefore, we integrated major protein sequence databases, protein–protein interaction (PPI) databases, Gene Ontology (GO) (http://www.geneontology.org) (7), KEGG pathway (http://www.genome.jp/kegg) (8) and bioinformatics analysis tools such as SignalP. This system has four major parts: (i) a nonredundant protein database, (ii) a physico-chemical property analysis module, (iii) a function annotation module and (iv) a PPI prediction module. A schematic workflow of MassNet is shown in Figure 1.

Construction of the nonredundant protein database

In order to identify proteins from MS data, researchers use various protein sequence databases such as NCBInr, SwissProt and trEMBL. However, there can be confusion among protein identifiers. Because of this problem, all protein identifiers were relationally linked. We integrated the protein sequence databases (Swiss-Prot, trEMBL, NCBInr, RefSeq, Ensembl and IPI) using only perfect-matching sequences. The database unifies protein IDs of the same sequence, summarizes annotations and descriptions of proteins from a range of organisms representing all three major kingdoms of life: eukaryotes,

*To whom correspondence should be addressed. Tel: +82 42 879 8530; Fax: +82 42 879 8519; Email: ishoon@kribb.re.kr
Correspondence may also be addressed to Jong Bhak. Tel: +82 42 879 8500; Fax: +82 42 879 8519; Email: jong@kribb.re.kr

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

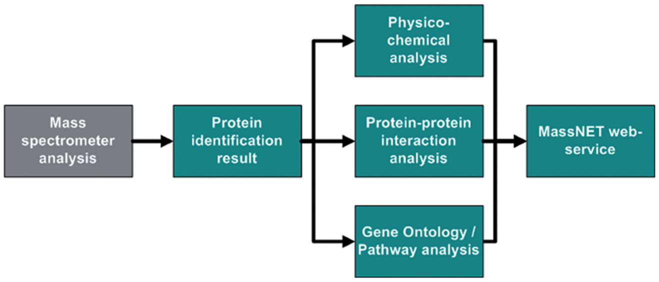

Mass spectrometer analysis
Protein identification result
Physico-chemical analysis
Protein-protein interaction analysis
Gene Ontology / Pathway analysis
MassNET web-service

Figure 1. The schematic workflow of MassNet.

prokaryotes and viruses. Therefore, the root identifier (Sequenceome_ID) can contain several protein identifiers from all available databases. The Sequenceome_ID database is a nonredundant sequence database of 6 856 434 proteins (April 2008).

Analysis of physico-chemical properties of proteins

The physico-chemical properties of MS data are important to understand biological functions. Especially, the prediction of hydropathy and subcelular localization of MS data is closely related to find membrane proteins which are involved in cellular processes and protein classes as drug targets (9). We used modules from Biopython (http://biopython.org) (10) to calculate hydropathy profile, GRAVY score (the average hydropathy score for all the amino acids), protein length, molecular weight, amino acid distribution, isoelectric point and protein instability index (11). For the subcellular localization prediction, we predicted transmembrane helices and signal peptides using Phobius (http://phobius.sbc.su.se) (12) and SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP) (13) programs. In order to provide physico-chemical information without any time delay, we provide precalculated physico-chemical properties for all nonredundant protein sequences. Whole proteins' physico-chemical properties are also provided as summary tables or figures. If a set of proteins was input, the user can acquire information on the protein set's physico-chemical distribution against whole-protein distribution of the organism. If the identified protein set was from the membrane fraction of an organism, the user compares the relative transmembrane protein abundances between the organism's whole-protein set and the identified protein set. Therefore, this summary information can be used to evaluate the input data quality.

Integration of annotation information

MassNet provides biological function information by using KEGG pathways and GO. The KEGG pathway database and GO represent an attempt to assign known proteins into known biological pathways and are updated regularly (8). MassNet assigns proteins to KEGG pathways thorough ID mapping and shows color-coded proteins in the context of biochemical pathway maps using KEGG API. In order to find significant associations of GO terms with queried proteins, we assigned proteins into GO categories and GO-slim (14) through ID mapping. In order to gain more accurate statistical test results of KEGG and GO assignment, we added Fisher's exact test algorithm (*P*-value).

Prediction of PPI

The prediction of PPI is based on PSIMAP (protein structural interactome MAP) (http://psimap.com, http://psibase.kobic.re.kr) (15,16) and PEIMAP (protein experimental interactome MAP) (17). The basic algorithm of PSIMAP infers interactions among proteins by using their homologs. Interactions among domains or proteins for known PDB (Protein Data Bank) (http://www.rcsb.org/pdb) structures are the basis of the predictions. If an unknown protein has a homolog to a domain, PSIMAP assumes that the query tends to interact with its homolog's partners. Its concept is called 'homologous interaction' (18–20). The original interaction between two proteins or domains is based on the Euclidean distance. Therefore, PSIMAP gives a structure-based interaction prediction (15). On the other hand, PEIMAP is a well-established method that uses public resources of experimentally confirmed protein interaction information such as BIND (http://bond.unleashedinformatics.com) (21), DIP (http://dip.doe-mbi.ucla.edu) (22), IntAct (http://www.ebi.ac.uk/intact) (23), MINT (http://mint.bio.uniroma2.it/mint) (24), HPRD (http://www.hprd.org) (25), CYGD (http://mips.gsf.de/genre/proj/yeast) (26) and BioGrid (http://www.thebiogrid.org) (27). We constructed a nonredundant PPI database from the source databases. We carried out a redundancy check to remove identical protein sequences from the source interaction databases using PERL (http://www.perl.org). Now, it contains 116 773 proteins and 229 799 interactions. The accuracy of PEIMAP is dependent on the confidence of each resource. In order to reduce the false positive rate of PEIMAP, we computed the final 'combined score' for each pair of proteins which were predicted by PEIMAP and PSIMAP algorithms. This scoring methodology has been proposed by published articles including the STRING server (http://string.embl.de) (28). Users can easily predict PPI for queried proteins in a list and can examine PPIs with a network viewer.

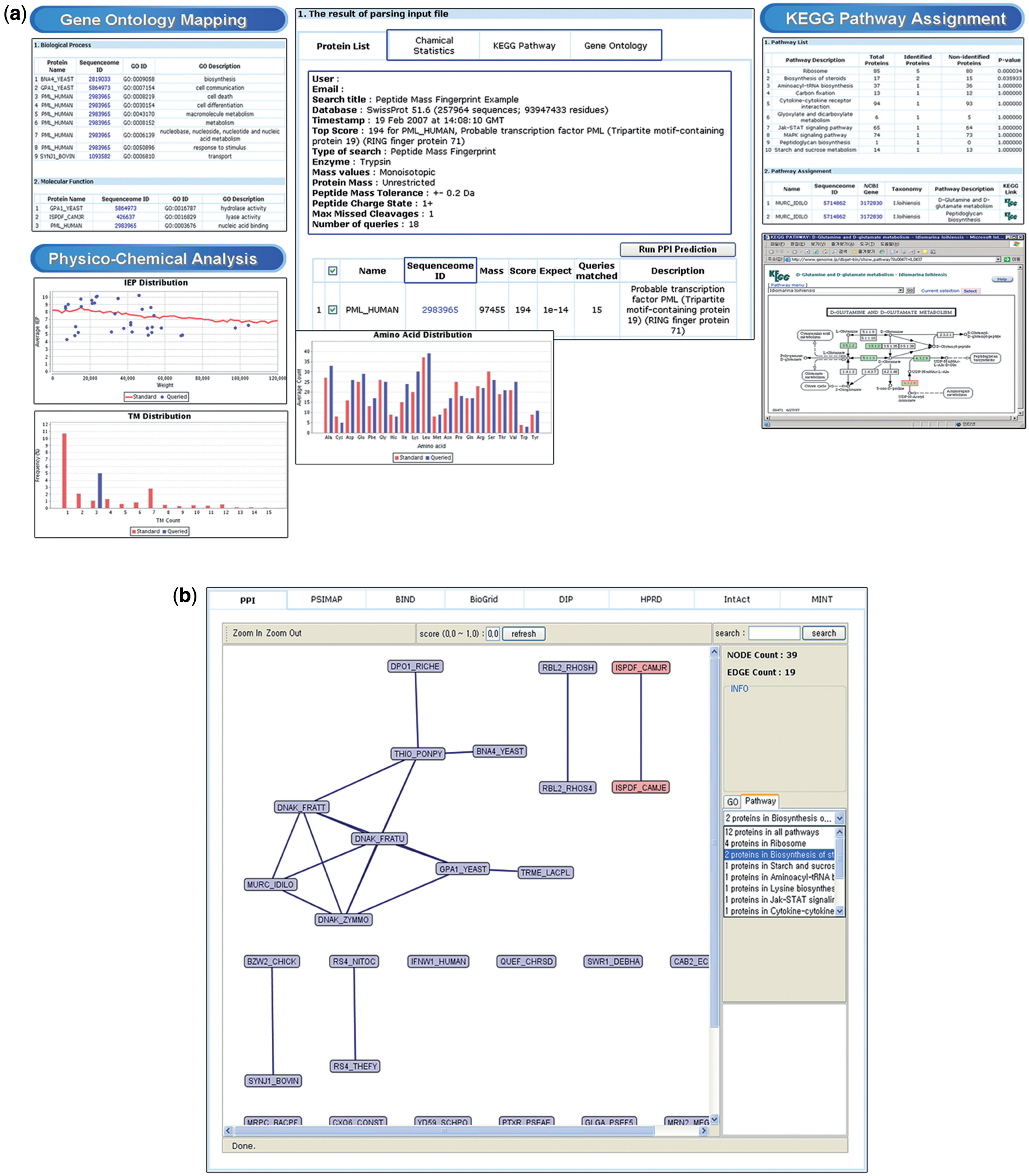
USER INTERFACE

Input

The query interface allows the user to submit an HTML file from the mass spectrometry or a TAB-delimited text file. The tab-delimited file must contain protein names in the first column. Detailed information about the TAB-delimited file format is described on the 'HOW TO USE' page. MassNet can use four types of MS data formats, i.e. MASCOT, Prospector, Profound and TAB-delimited file.

Output

After uploading the query file, users can obtain the annotation information as in Figure 2a. The annotation results consist of five parts: (i) a protein list page, (ii) the physico-chemical property of each protein, (iii) a PPI



1. Biological Process

	Protein Name	Sequenceome ID	GO ID	GO Description
1	BNA4_YEAST	2819033	GO:0009058	biosynthesis
2	GPA1_YEAST	5864973	GO:0007154	cell communication
3	PML_HUMAN	2983965	GO:0008219	cell death
4	PML_HUMAN	2983965	GO:0030154	cell differentiation
5	PML_HUMAN	2983965	GO:0043170	macromolecule metabolism
6	PML_HUMAN	2983965	GO:0008152	metabolism
7	PML_HUMAN	2983965	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
8	PML_HUMAN	2983965	GO:0050896	response to stimulus
9	SYNJ1_BOVIN	1093582	GO:0006810	transport

2. Molecular Function

	Protein Name	Sequenceome ID	GO ID	GO Description
1	GPA1_YEAST	5864973	GO:0016787	hydrolase activity
2	ISPDF_CAMJR	426637	GO:0016829	lyase activity
3	PML_HUMAN	2983965	GO:0003676	nucleic acid binding

1. The result of parsing input file

Protein List | Chamical Statistics | KEGG Pathway | Gene Ontology

User :
Email :
Search title : Peptide Mass Fingerprint Example
Database : SwissProt 51.6 (257964 sequences; 93947433 residues)
Timestamp : 19 Feb 2007 at 14:08:10 GMT
Top Score : 194 for PML_HUMAN, Probable transcription factor PML (Tripartite motif-containing protein 19) (RING finger protein 71)
Type of search : Peptide Mass Fingerprint
Enzyme : Trypsin
Mass values : Monoisotopic
Protein Mass : Unrestricted
Peptide Mass Tolerance : +- 0.2 Da
Peptide Charge State : 1+
Max Missed Cleavages : 1
Number of queries : 18

Run PPI Prediction

	☑	Name	Sequenceome ID	Mass	Score	Expect	Queries matched	Description
1	☑	PML_HUMAN	2983965	97455	194	1e-14	15	Probable transcription factor PML (Tripartite motif-containing protein 19) (RING finger protein 71)

1. Pathway List

	Pathway Description	Total Proteins	Identified Proteins	Non-identified Proteins	P-value
1	Ribosome	85	5	80	0.000034
2	Biosynthesis of steroids	17	2	15	0.035933
3	Aminoacyl-tRNA biosynthesis	37	1	36	1.000000
4	Carbon fixation	13	1	12	1.000000
5	Cytokine-cytokine receptor interaction	94	1	93	1.000000
6	Glyoxylate and dicarboxylate metabolism	6	1	5	1.000000
7	Jak-STAT signaling pathway	65	1	64	1.000000
8	MAPK signaling pathway	74	1	73	1.000000
9	Peptidoglycan biosynthesis	1	1	0	1.000000
10	Starch and sucrose metabolism	14	1	13	1.000000

2. Pathway Assignment

	Name	Sequenceome ID	NCBI Gene	Taxonomy	Pathway Description	KEGG Link
1	MURC_IDILO	5714862	3172830	I.loihiensis	D-Glutamine and D-glutamate metabolism	KEGG
2	MURC_IDILO	5714862	3172830	I.loihiensis	Peptidoglycan biosynthesis	KEGG

NODE Count : 39
EDGE Count : 19

2 proteins in Biosynthesis o...

12 proteins in all pathways
4 proteins in Ribosome
2 proteins in Biosynthesis of st
1 proteins in Starch and sucros
1 proteins in Aminoacyl-tRNA b
1 proteins in Lysine biosynthes
1 proteins in Jak-STAT signalir
1 proteins in Cytokine-cytokine

Figure 2. Screenshots of MassNet annotation results. (**a**) Panel in the middle is the protein list table. KEGG Pathway tab shows KEGG pathway assignment and metabolic pathway graph (right panels). Gene Ontology tab shows proteins assigned to GO categories (left top panel). Chemical Statistics tab shows the input protein set's physico-chemical distribution against whole protein distribution of the organism (left bottom panels). (**b**) Protein-protein interactions of user-selected proteins are visualized by a network viewer. Rectangular shapes are protein nodes. The black connecting lines indicate interactions among the nodes. The two red rectangular nodes are proteins that are selected by the users through the right hand side panel. When users select the right pull down menus in the right panel, the left drawing canvas shows highlighted protein nodes.

prediction page, (iv) a KEGG pathway page and (v) a GO page.

The protein list page shows a table describing protein names and scores, which are parsed from the query file. The KEGG pathway and the GO pages show the number of proteins, which belong to the categories of KEGG pathways and GO. By clicking the 'Run PPI Prediction' button at the top of the protein list table, the user can acquire the PPI information for selected proteins. The PPI page shows PEIMAP and PSIMAP (see Methods section) data at two separated tables.

By clicking Sequenceome_IDs at all pages, users can access two pages, i.e. a Same IDs page and a Chemical Property page. The same IDs page shows the identical sequences at various protein sequence databases and provides the hyperlinks to original database web pages. In order to provide clear information, MassNet provides a viewer for PPI networks as in Figure 2b.

IMPLEMENTATION

The MassNet web server runs on a Linux server. It combines a MySQL (http://www.mysql.com) database with a dynamic web interface using Java Server Pages (http://java.sun.com/products/jsp). Data preprocessing is implemented in Perl and Python, and the network viewer for PPI was constructed using Java.

CONCLUSION

The functional analysis and interpretation of the large-scale MS data are still a challenging task. An automatic approach is necessary for tens of thousands of MS data collected throughout the world. MassNet is the first web server that provides various kinds of functional information, such as physico-chemical properties, biological pathways, gene ontology and PPI, for MS data. MassNet is easy to use and provides information through an automatic annotation for queried proteins.

ACKNOWLEDGEMENTS

We thank Maryana Bhak for editing the article and Suan Cho for making the figures. This work was supported by a grant from the KRIBB Research Initiative Program of Korea, by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (No. M10508040002-07N0804-00210) and supported by a Korea Basic Science Institute K-Mep grant (T28021) to J-.S.C. Funding to pay the Open Access publication charges for this article was provided by MOST (No. M10508040002-07N0804-00210).

Conflict of interest statement. None declared.

REFERENCES

1. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
2. Perkins,D.N., Pappin,D.J., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
3. Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.
4. Kemmeren,P., van Berkum,N.L., Vilo,J., Bijma,T., Donders,R., Brazma,A. and Holstege,F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, **9**, 1133–1143.
5. Dennis,G. Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, **4**, P3.
6. Eng,J.K., McCormack,A.L. and Yates,J.R. III. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.
7. 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.
8. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
9. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
10. Chapman,B. and Chang,J. (2000) Biopython: python tools for computational biology. *ACM SIGBIO Newsletter*, **20**, 15–19.
11. Guruprasad,K., Reddy,B.V.B. and Pandit,M.W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, **4**, 155–161.
12. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
13. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protocols*, **2**, 953–971.
14. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
15. Park,J., Lappe,M. and Teichmann,S.A. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, **307**, 929–938.
16. Gong,S., Yoon,G., Jang,I., Bolser,D., Dafas,P., Schroeder,M., Choi,H., Cho,Y., Han,K., Lee,S. *et al.* (2005) PSIbase: a database of protein structural interactome map (PSIMAP). *Bioinformatics*, **21**, 2541–2543.
17. Kim,J.G., Park,D., Kim,B.C., Cho,S.W., Kim,Y.T., Park,Y.J., Cho,H.J., Park,H., Kim,K.B., Yoon,K.O. *et al.* (2008) Predicting the interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. *BMC Bioinform.*, **9**, 41.
18. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
19. Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.
20. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.
21. Bader,G.D. and Hogue,C.W. (2000) BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, **16**, 465–477.
22. Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.
23. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P.,

Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.

24. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
25. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32**, D497–D501.
26. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al*. (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res*., **33**, D364–D368.
27. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535–D539.
28. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.