W84–W89 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 5 May 2010*
doi:10.1093/nar/gkq320

PANDORA: analysis of protein and peptide sets through the hierarchical integration of annotations

Nadav Rappoport[1], Menachem Fromer[1], Regev Schweiger[1] and Michal Linial[2,3,*]

[1]School of Computer Science and Engineering, [2]Department of Biological Chemistry, Institute of Life Sciences and [3]The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Israel

Received January 31, 2010; Revised April 10, 2010; Accepted April 17, 2010

ABSTRACT

Derivation of biological meaning from large sets of proteins or genes is a frequent task in genomic and proteomic studies. Such sets often arise from experimental methods including large-scale gene expression experiments and mass spectrometry (MS) proteomics. Large sets of genes or proteins are also the outcome of computational methods such as BLAST search and homology-based classifications. We have developed the PANDORA web server, which functions as a platform for the advanced biological analysis of sets of genes, proteins, or proteolytic peptides. First, the input set is mapped to a set of corresponding proteins. Then, an analysis of the protein set produces a graph-based hierarchy which highlights intrinsic relations amongst biological subsets, in light of their different annotations from multiple annotation resources. PANDORA integrates a large collection of annotation sources (GO, UniProt Keywords, InterPro, Enzyme, SCOP, CATH, Gene-3D, NCBI taxonomy and more) that comprise ~200 000 different annotation terms associated with ~3.2 million sequences from UniProtKB. Statistical enrichment based on a binomial approximation of the hypergeometric distribution and corrected for multiple hypothesis tests is calculated using several background sets, including major gene-expression DNA-chip platforms. Users can also visualize either standard or user-defined binary and quantitative properties alongside the proteins. PANDORA 4.2 is available at http://www.pandora.cs.huji.ac.il.

INTRODUCTION

Due to advances in biological, experimental and computational methodologies, scientists are able to conduct high-level genomic and proteomic experiments. In most of these, biologists face the need of extracting meaningful biological insights from a large set of proteins or genes (1). A common approach for extracting such insights is to manually examine the set of proteins and attempt to derive biological conclusions. However, this method greatly relies on the expertise of the biologist examining the data and often produces a partial and biased view of the protein set (2). Another approach is using annotation-based computational methods. These methods enable the biologist to reach a global and more objective view of the data (3).

Typically, computational methods use a single annotation source, most commonly the Gene Ontology (GO) (4), and automatically detect annotations that appear at a frequency that is significantly greater than expected (5). However, the strong dependency of such methods on a single source restricts the biological information they can extract. Furthermore, these methods often provide only a limited biological view of the data set and are unable to detect groups that are characterized by sharing multiple biological properties in common. There are some exceptions, however, such as the DAVID (6) and EASE (7) resources, which provide statistical analysis of annotation subsets for the purpose of extracting biological knowledge.

We have developed a web server called PANDORA (Protein ANnotation Diagram ORiented Analysis) whose goal is the biological analysis of protein sets (8). Many protein and gene-annotation systems either explicitly, or implicitly, correspond to some hierarchical structure. For example, being annotated as localizing to the nucleolus necessarily implies localization to the nucleus, though the converse does not hold. Thus, several tools have been developed to address the visualization task for hierarchical annotations (9). We take this concept one step further by dynamically integrating multiple annotation sources into the natural hierarchy deriving from a particular set of user-defined proteins. PANDORA shows the protein set as a graph, which we refer to as the *Concept DAG* (Directed Acyclic Graph). The Concept DAG is a directed graph whose nodes represent protein

*To whom correspondence should be addressed. Tel: +972 2 6585425; Fax: +972 2 6586448; Email: michall@cc.huji.ac.il

subsets that share a unique combination of one or more biological annotations, and whose directed edges represent subset/superset relations between nodes [for further information on the graph construction see (8)]. Importantly, the graph still retains the annotation information for each protein while providing a richer and more accurate view of the data. Furthermore, PANDORA is based on the annotations as extracted from UniProtKB protein entry files. For each file, the annotation provided by UniProtKB and the mapping from external annotation resources encompassing extensive biological aspects is extracted. The rich collection of annotation resources covers biological functions at various levels: participation in biological processes, 3D structural classification, cellular localization, taxonomy, and more (see 'Databases' section). This overcomes the limitation of a single annotation source and permits helpful comparisons between various biological aspects.

We have previously described the underlying logic behind PANDORA and have demonstrated that PANDORA is useful in extracting meaningful and previously overlooked data from protein sets (3,8). PANDORA was valuable in interpretation of large-scale experiments as demonstrated in (10). PANDORA 4.2 is expanded to include most UniProtKB protein sequences and their associated annotations. In this article, we describe new and improved features in PANDORA 4.2 that further extend the power of biological analysis of sets through our system. These features include: (i) User Properties—PANDORA allows incorporation of external user properties, such as differential expression levels or quantitative information from mass spectrometry (MS) proteomics experiments. These custom properties can be included in the PANDORA analysis to further enhance the discovery of biological knowledge; (ii) Statistical evaluation of the input relative to several different background databases; (iii) Using the hit list of protein matches of NCBI-BLAST as an input set and using the BLAST *e*-values as quantitative properties; (iv) Incorporating PANDORA into external biological servers such as ProtoNet, which provides thousands of homology-based clusters for analysis; (v) Expanding PANDORA to handle MS proteomics data—PANDORA now also allows peptides as input for major model organisms. Peptides are mapped to peptide lists representing in-silico cleavage by proteases that are commonly used in MS proteomics research.

OVERVIEW

The user starts using the PANDORA server either by entering a user-defined set of proteins ('User Set' menu), entering a list of proteolytic peptides to be mapped to the proteins from which they are derived ('Peptides'), searching for proteins with a particular annotation ('Keyword'), or considering the proteins detected in a BLAST homology search ('Blast'). Ultimately, these inputs are all transformed into the set of corresponding proteins and the process continues from there. Subsequently, pre-defined quantitative properties can be selected, as desired. Finally, the proteins being analyzed are displayed in their annotation-derived hierarchy, where each node represents a subset of proteins with particular biological properties. In addition, a statistical evaluation of annotation enrichment is provided.

DATABASES AND ANNOTATION RESOURCES

PANDORA 4.2 supports almost 10 times as many proteins than in previous versions (Table 1), covering ~3.2 millions sequences from UniProtKB (11). A sample list of the keywords that are supported is shown in Table 2. On average, each protein is covered by 24 different annotation types (excluding taxonomy). PANDORA is based mainly on annotations extracted from UniProtKB (the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases). The mapping to UniProtKB Keywords, ENZYME, GO annotations, InterPro and Taxonomy is based on the XML file for each protein sequence entry. For structural annotations (CATH, SCOP, GENE3D), a direct mapping was completed from the original resources or through the InterPro compendium. The individual sources underlying InterPro entries are maintained allowing focusing on any of the family and domain based resources (e.g. PROSITE, PRINTS, Pfam, SMART, SUPERFAMILY). All the information is stored locally in ProtoNet database (12). The size of the database supporting PANDORA 4.2 is stored in 88 GB. Several of the databases are structured and hierarchical (such as ENZYME, SCOP, CATH). For these resources, each level of the hierarchy can be selected separately, resulting in ~40 levels of annotations that can be selected for analysis (Figure 1). Note that the coverage of

Table 1. Number of sequences supported by older and new versions of PANDORA, for model organism representatives

Species	PANDORA 2.0	PANDORA 3.0	PANDORA 4.2
Homo sapiens	8507	47 641	106 529
Mus musculus	5678	41 813	61 783
Drosophila melanogaster	2049	22 603	27 942
Arabidopsis thaliana	1680	39 367	46 671
Plasmodium falciparum	153	8434	11 029
Total proteins	114 033	1 072 911	3 188 835

Table 2. Sample of the supported annotations and their coverage in PANDORA database

Annotation resource	Percentage coverage	Number of annotations
ENZYME (10/2006)	8	5010
GENE3D (3.0)	8	410
SMART (5.0)	15	704
CATH (v3.1.0)	19	3301
GO (6/2006)	22	13 603
SCOP (1.71)	25	6039
PFAM (19.0)	73	8534
UniProt (8.1)	78	879
InterPro (12.1)	78	13 147
NCBI Taxonomy	100	283 050

the annotation resources ranges from 8 to 78% (excluding taxonomy) (Table 2) but this level is higher for the main model organisms.

INPUT METHODS AND INTEGRATED BLAST

There are four methods of selecting an initial set for PANDORA to work with:

(i) *User set of genes/proteins:* The user inputs a list of protein accession numbers, either by uploading a file or manually entering it. PANDORA accepts either UniProt protein accession numbers or GenBank gene accessions (which are mapped to proteins in UniProtKB). Users may also include in their file supplementary user properties (see 'User properties' section) in order to access some of the advanced analysis features. Other options include selecting an appropriate background database in order to fine-tune the statistical evaluation and adding to the analysis some intrinsic pre-calculated protein properties (e.g. pI and molecular weight). The default background database covers all proteins in our database.

(ii) *Peptides set:* As MS proteomics is becoming more extensive, we added an MS-based input method. The user can enter a set of peptides and PANDORA will match these peptides to the appropriate proteins. PANDORA supports peptides of >600 daltons that match peptides from MS proteomics data. Currently peptides from the Rat, Mouse, Human, Drosophila and Yeast proteomes are supported. The user must choose the proteolytic enzyme that was used to derive the MS data, from the most commonly used proteases (e.g. Trypsin, Lys-C). Currently, only complete cleavage is supported and post-translational modifications are not taken into account.

(iii) *Keyword:* The user chooses a keyword (annotation), and the set of all proteins in the database that have that particular keyword (along with others) are chosen to be the initial set. This permits the study of all proteins that participate in a certain biological pathway, share a common 3D fold, or share a similar molecular function. These can provide a global view for an evolutionary study. The user may also select the initial protein set to be the union of several keyword-based sets, in order to overcome some inconsistencies in annotation sources (e.g. annotation of 'Voltage-gated potassium channel activity' supported by GO and 'Voltage-dependent potassium channels' supported by InterPro).

(iv) *Integrated BLAST:* This input method integrates the BLAST local alignment search method with the analysis capabilities of PANDORA. The user submits a protein sequence and a routine NCBI-BLAST search is run against the selected database. Instead of the usual result that consists of a long list of matching proteins, the results are sent to PANDORA and are displayed as a Concept DAG, using the BLAST *e*-value as a quantitative property which is added on to the graph (see 'User properties' section). One can easily recognize biological subsets within the results list that have been unified by a specific range of *e*-values. Of course, the results can be viewed from various biological perspectives by integrating the different annotation sources offered by PANDORA.

USER PROPERTIES

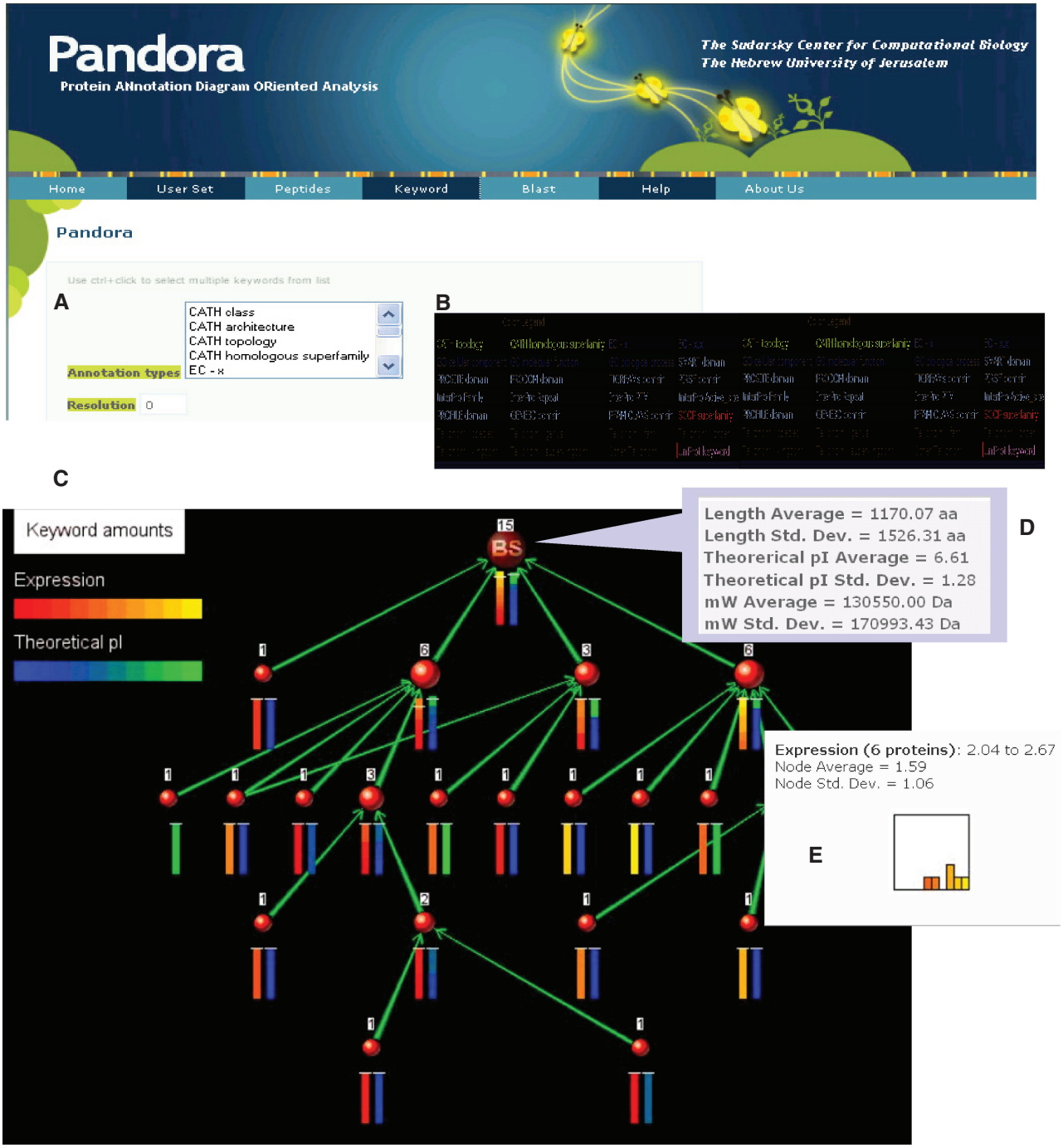
Generally, PANDORA receives a protein set as input, derives all information on the proteins from its integrated database and uses that information to build the Concept DAG (see example in Figure 1). However, in many cases, it would be helpful to let the user introduce external supplementary information about the proteins into the analysis. Examples of such external information are relative change in expression levels (which are typical for microarray experiments), a user-defined division of the protein set into several sets (allowing comparison of the sets from repeated experiments), or even an alignment score such as BLAST *e*-values (see 'Input methods and integrated BLAST' section). To this end, we have developed the 'user properties' option. Generally, properties introduced by the user can be divided into three categories:

(i) *Binary properties:* All annotations used by PANDORA are primarily binary annotations. This means that a protein can either have or not have the annotation (e.g. a protein can either be an enzyme or not). Users can add such external properties to the proteins, which will be considered as genuine annotations when constructing the graph. This allows users to add additional user-specific information that is not included in the annotation sources that PANDORA uses. This can also be used to compare multiple sets of proteins, assigning to each protein an indication of the set it belongs to.

(ii) *Categorical properties:* This is actually a variation on the binary properties: Some properties may be viewed as categorical, meaning that the protein can belong to one category out of a number of different categories. Each category can be considered to be a binary annotation. Examples for such user-based annotations are the level of reliability of the experiment (low, high and intermediate), the source of the data (healthy or cancerous according to different staging scores), identity of the different tissues, or organism strains.

(iii) *Quantitative properties:* Many interesting properties can not be categorized as above. Examples are the differential change in expression in genomic and proteomic experiments.

In order to deal with a quantitative property, PANDORA ignores the property when building the graph, and then examines the distribution of the property on the graph (Figure 1). The PANDORA



F **Statistical Evaluation:**

Keyword type	Keyword	Amount	Sensitivity	Specificity	P-value	Corrected P-value
InterPro Family	Barrier to autointegration factor, BAF	1	0.034	0.067	5.444e-5	0.002
InterPro Family	Ninjurin	1	0.031	0.067	6.007e-5	0.002
InterPro Family	Myelin and lymphocyte (MAL)	1	0.024	0.067	7.696e-5	0.002
InterPro Family	Emopamil-binding	1	0.020	0.067	9.386e-5	0.003
InterPro Domain	Dopey, N-terminal	1	0.018	0.067	1.032e-4	0.003

Figure 1. Result page from PANDORA analysis on a user set. The set of 15 proteins was included in the input set (marked as Basic Set, BS). (**A**) Approximately 40 annotation resources are selected by the user from a menu, multiple selections are encouraged. (**B**) Sample of the keywords annotation source color-coded by their types. (**C**) PANDORA graph for a user set associated with quantitative properties of user-input expression levels (red to yellow) and pre-calculated pI (blue to green). (**D**) Summary and statistics for the quantitative data of the analyzed protein set. (**E**) Distribution histogram of the expression range for a node. (**F**) Table of the statistical significance of the annotations, including a correction for multiple hypothesis test.

graph consists of nodes, where each node represents a subset of proteins that share certain biological properties. Each node, therefore, has a distribution of the quantitative property for its proteins. The distribution of each node is displayed as a histogram below the node. This allows the user to easily recognize nodes with distinct quantitative patterns. For example, if the quantitative property is change in expression level, we could easily identify subsets of proteins that are both related biologically and share similar expression patterns. Of course the user is not limited to any specific kind of quantitative property and could make creative use of this feature. For example, the integrated BLAST feature uses the BLAST *e*-values as a quantitative property (see 'Input methods and integrated BLAST' section) in order to facilitate the detection of biological groups that have statistically significant sequence similarity to an input sequence. For simplicity, the user may display up to three quantitative properties simultaneously, enabling the search for correlation between different orthogonal properties. We added pre-calculated quantitative properties for each protein in the database, including pI, molecular weight (in Dalton) and length (in amino acids). Experimental MS proteomics is a rich source for proteins and peptide sets. We thus added quantitative data that include the number of detectable peptides with various commonly used proteases and the number of validated phosphorylation sites. The later were extracted from UniProtKB XML under 'amino-acid modification'. A further refinement is achieved by partitioning the phosphorylation type to Phosphothreonine, Phosphoserine and Phosphothyrosine.

STATISTICAL EVALUATION

One critical aspect in the evaluation of biological results is their statistical significance. PANDORA deals with this by coloring each node according to the node's sensitivity for that annotation. The node's color represents the highest sensitivity of the node to any of its annotations. A white and red node has a sensitivity of 1 and 0, respectively. For some nodes the sensitivity is not well-defined and these nodes appear as a red–white swirl (undetermined).

In addition PANDORA provides an evaluation table together with each graph. The table gives *P*-values for the appearance of the annotations on the current protein set, estimating the probability that an annotation would randomly appear as frequently as it did. The calculation of *P*-value is based on a binomial approximation of the hypergeometric distribution following Bonferroni correction. An additional correction (multiple hypothesis testing) is added to the table that is based on the FDR adjustment (13). In conjunction with the ability to use several different annotation sources, this evaluation can label statistically significant enrichments (Figure 1). Of course, to properly estimate these *P*-values, it is necessary to know from which background pool of proteins the input proteins were taken and evaluate how frequent each annotation is in that background set. Although PANDORA generally does not assume anything about the origin of the protein set which is analyzed, it allows a selection of background models that fit various experimental models. For microarray experiments, PANDORA offers a variety of background sets, such as the most commonly used Affymetrix microarrays. For proteomic experiments, PANDORA offers background sets of proteomes of several model organisms and proteins according to their partition to SwissProt or TrEMBL. For other purposes, PANDORA simply uses the whole SwissProt+TrEMBL database as its background. Researchers that require background sets that are not currently included in PANDORA are encouraged to contact the authors.

INTERFACE WITH EXTERNAL SERVERS

PANDORA results can be saved at different formats (including FASTA format, accession ID list etc). In addition, PANDORA allows presenting a group of proteins that is unified by an annotation node by a multiple sequence alignment (i.e. CLUSTALW representation). PANDORA can easily interface with other biological servers that deal with protein set analysis. A web server that has recently been linked to PANDORA is ProtoNet (12), which uses PANDORA to gain biological insight into large protein clusters. Web server developers who are interested in interfacing directly with PANDORA may contact the authors.

PANDORA UPDATE

PANDORA is based on an extensive database which integrates several biological databases. An underlying protein database is used as a basis for information on the protein entities, and several annotation sources whose annotations are mapped to the protein databases are used in conjunction.

The underlying protein database initially used by PANDORA (8) has been changed from SwissProt (114 035 proteins) to UniProtKB (3 188 835 proteins), giving a greatly enhanced representation of the proteomes of several model organisms (see examples in Table 1). The annotation sources used by PANDORA have also been updated, and now offer ~200 000 different annotations, spanning several different biological domains. All underlying protein and annotation databases are periodically updated in order to keep up with the most recent biological knowledge available. We are currently planning to add additional annotation sources to PANDORA in order to improve protein set analysis in further biological aspects such as protein–protein interactions.

ACKNOWLEDGEMENTS

The authors would like to thank Solange Karsenty for her support in maintaining and design the Web site. The authors thank Michael Dvorkin for support in managing the immense database and the ProtoNet team.

FUNDING

Prospects consortium (EU framework VII) and the BSF (grant number 2007219); Sudarsky Center for Computational Biology (to N.R., M.F. and R.S.). Funding for open access charge: Prospects consortium (EU framework VII) and the BSF (grant number 2007219).

Conflict of interest statement. None declared.

REFERENCES

1. Loewenstein,Y., Raimondo,D., Redfern,O.C., Watson,J., Frishman,D., Linial,M., Orengo,C., Thornton,J. and Tramontano,A. (2009) Protein function annotation by homology-based inference. *Genome Biol.*, **10**, 207.
2. Artamonova.II, Frishman,G., Gelfand,M.S. and Frishman,D. (2005) Mining sequence annotation databanks for association patterns. *Bioinformatics*, **21(Suppl. 3)**, iii49–iii57.
3. Sasson,O., Kaplan,N. and Linial,M. (2006) Functional annotation prediction: all for one and one for all. *Protein Sci.*, **15**, 1557–1562.
4. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
5. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
6. Huang da,W., Sherman,B.T., Tan,Q., Collins,J.R., Alvord,W.G., Roayaei,J., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183.
7. Hosack,D.A., Dennis,G. Jr, Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
8. Kaplan,N., Vaaknin,A. and Linial,M. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res.*, **31**, 5617–5626.
9. Hu,Z., Hung,J.H., Wang,Y., Chang,Y.C., Huang,C.L., Huyck,M. and DeLisi,C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, **37**, W115–W121.
10. Witzmann,F.A., Arnold,R.J., Bai,F., Hrncirova,P., Kimpel,M.W., Mechref,Y.S., McBride,W.J., Novotny,M.V., Pedrick,N.M., Ringham,H.N. *et al.* (2005) A proteomic survey of rat cerebral cortical synaptosomes. *Proteomics*, **5**, 2177–2201.
11. Stutz,A., Bairoch,A., Estreicher,A. and Grp,S.-P. (2006) UniProtKB/Swiss-Prot: the protein sequence knowledgebase. *FEBS J.*, **273**, 62.
12. Kaplan,N., Sasson,O., Inbar,U., Friedlich,M., Fromer,M., Fleischer,H., Portugaly,E., Linial,N. and Linial,M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res*, **33**, D216–D218.
13. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodol.)*, **57**, 289–300.