Published online 10 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D737–D741***
doi:10.1093/nar/gkl876

SynDB: a Synapse protein DataBase based on synapse ontology

Wuxue Zhang, Yong Zhang, Hui Zheng[1], Chen Zhang[2], Wei Xiong[1], John G. Olyarchuk, Michael Walker[3], Weifeng Xu[4], Min Zhao, Shuqi Zhao, Zhuan Zhou[1,*] and Liping Wei*

Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P.R. China, [1]Institute of Molecular Medicine, Peking University, Beijing 100871, P.R. China, [2]Center for Basic Neuroscience, UT Southwestern Medical Center, Dallas, TX 75235, USA, [3]Department of Medicine and [4]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA 94305, USA

Received August 15, 2006; Revised and Accepted October 6, 2006

ABSTRACT

A synapse is the junction across which a nerve impulse passes from an axon terminal to a neuron, muscle cell or gland cell. The functions and building molecules of the synapse are essential to almost all neurobiological processes. To describe synaptic structures and functions, we have developed Synapse Ontology (SynO), a hierarchical representation that includes 177 terms with hundreds of synonyms and branches up to eight levels deep. associated 125 additional protein keywords and 109 InterPro domains with these SynO terms. Using a combination of automated keyword searches, domain searches and manual curation, we collected 14 000 non-redundant synapse-related proteins, including 3000 in human. We extensively annotated the proteins with information about sequence, structure, function, expression, pathways, interactions and disease associations and with hyperlinks to external databases. The data are stored and presented in the Synapse protein DataBase (SynDB, http://syndb.cbi.pku.edu.cn). SynDB can be interactively browsed by SynO, Gene Ontology (GO), domain families, species, chromosomal locations or Tribe-MCL clusters. It can also be searched by text (including Boolean operators) or by sequence similarity. SynDB is the most comprehensive database to date for synaptic proteins.

INTRODUCTION

Recent developments in genomics, proteomics and systems biology have significantly impacted fields such as oncology and immunology (1–5) and are beginning to be applied to neuroscience research, generating an exponentially increasing amount of data (6–11) and calling for efficient databases. However, neuroinformatics databases at the molecular level are currently limited. For instance, databases listed in the Society for Neuroscience Database Gateway (NDG, http://ndg.sfn.org/eavObList.aspx?cl=81) principally contain imaging, anatomic or clinical data, while few focus on the gene or protein level and their functions.

The synapse is a specialized intercellular junction between neurons or between neurons and other excitable cells such as muscle. The synapse plays a key role in information processing in the nervous system that underlies many neurobiological processes, including neurotransmission, learning and memory. Defects in synaptic activity are associated with many neurological disorders, including Alzheimer's disease (12). The synapse has also been proposed as an excellent candidate for large-scale systems biology studies (7,8,13). There is a critical need for a focused yet comprehensive database resource for the synapse 'proteome'. Creating such a database is non-trivial, because the proteins involved in synaptic activities are numerous and diverse and information is scattered in multiple heterogeneous sources. No simple keyword search and no small number of domains can retrieve all the proteins. These complexities may explain why such a database has not been reported thus far.

*To whom correspondence should be addressed: Tel: +86 10 6276 4970; Fax: +86 10 6275 2438; Email: weilp@mail.cbi.pku.edu.cn
*Correspondence may also be addressed to Zhuan Zhou. Tel: +86 10 6276 4986; Fax: +86 10 6275 3212; Email: zzhou@pku.edu.cn

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

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

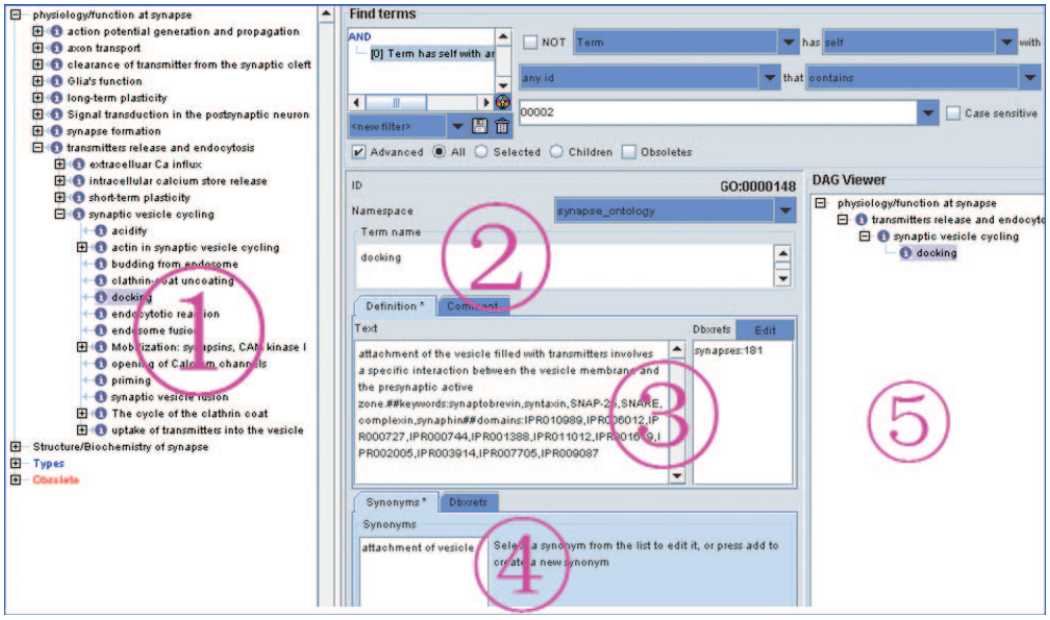

Figure 1. DAG-Edit view of Synapse Ontology (SynO): SynO is stored and managed in DAG-Edit. ① Hierarchical display of names and relationships of SynO terms; ② the term of interest; ③ description of the term and list of keywords and domains associated with the term and sources from which term was derived; ④ synonyms of the term; ⑤ the path from root to the term.

Here, we present the Synapse protein DataBase (SynDB, http://syndb.cbi.pku.edu.cn) as an information hub for synapse-related proteins.

CONSTRUCTION OF SYNAPSE ONTOLOGY

Ontology is defined as the 'specification of a conceptualization' (14). It describes a domain using a collection of concepts or terms and includes the hierarchical relationships between the terms. In order to formally describe synaptic functions and structures, we extensively reviewed three sources of information: (i) three classic text books, *Synapses* (15), *Principles of Neural Science* (16) and *Ion Channels of Excitable Membranes* (17); (ii) 115 recent (2000–2006) review papers published in *Nature Reviews Neuroscience* and *Annual Review of Neuroscience*; and (iii) relevant terms in two general ontologies, *Gene Ontology* (GO) (18) and *Medical Subject Headings* (MeSH) (19).

By reviewing these resources and iteratively organizing the information, we constructed the first synapse ontology (SynO), a hierarchical description of synaptic structures and functions. SynO has two top-level categories: structure and function. Structure is divided into categories such as presynaptic compartment, postsynaptic compartment and glia; and function is divided into categories such as transmitter release and endocytosis, synapse formation and signal transduction in the postsynaptic neuron. In total, SynO contains 177 terms with hundreds of synonyms and up to eight levels deep. SynO is constructed, as is GO, as a directed acyclic graph (DAG). If the terms are represented by vertices and the relationships between terms are represented by edges, the terms in a DAG can be connected via a directed graph without cycles.

We used DAG-edit (20) to input, manage and update SynO (Figure 1). We annotated each term with name, synonyms, definition and source references, as well as the 'part-of' or 'is-a' relationship to other terms. In the definition field, we recorded additional protein keywords associated with the term as well as InterPro (21) domains related to the term (see details below in 'Association of Proteins'). SynO is available for download in the Open Biomedical Ontologies flat file format (20) at http://syndb.cbi.pku.edu.cn/download/SynO.obo.

We developed a Perl script to generate a list of search keywords based on SynO, including and expanding from SynO terms and synonyms. If a SynO term consists of more than one word, the Perl script specified which word can be expanded and whether the order of the words can be flexible. All possible combinations were automatically generated. The expanded list of search keywords was used in the next step.

ASSOCIATION OF PROTEINS

We searched the InterPro database using the search keywords and retrieved 400 protein domains. Through careful manual screening we identified 109 domains as being involved in synaptic activities and assigned them to the most appropriate SynO terms. We retrieved over 5000 proteins using the mapping between InterPro and UniProt (22) and associated these proteins with SynO terms.

We then searched UniProt to retrieve additional protein entries that contain the search keywords. While domain-based searches tend to have a high false-negative rate (as not all domains can be modeled), keyword-based searches tend to have a high false-positive rate, requiring that we impose both automated and manual quality control. For example, entries containing 'immune' or 'immunological' were removed, because 'immunological synapse' is a term defining a process in the immunological system that occurs

Table 1. SynDB protein annotations and cross-referenced molecular databases

Protein annotations	Cross-referenced databases
Gene name	NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)
Species	NCBI Taxonomy DB (ftp://ftp.ncbi.nih.gov/pub/taxonomy/)
Sequences	GenBank (http://www.ncbi.nlm.nih.gov/)
Chromosomal location	GoldenPath (http://hgdownload.cse.ucsc.edu/goldenPath/)
GO functional category	Gene Ontology (GO) (http://www.ebi.ac.uk/GOA/)
Protein domain	InterPro (http://www.ebi.ac.uk/interpro/)
Structure	Protein Data Bank (PDB) (http://www.rcsb.org/pdb/)
	ModBase (http://salilab.org/modbase-cgi)
Gene expression	GEO (http://www.ncbi.nlm.nih.gov/geo/)
	BodyMap-Xs (http://bodymap.jp)
Antisense transcripts	NATsDB (http://natsdb.cbi.pku.edu.cn/)
Post-transcriptional modification	dbPTM (http://dbptm.mbc.nctu.edu.tw/)
Protein family	Ensembl Family (http://www.ensembl.org/Homo_sapiens/familyview)
Pathway	KEGG (http://www.genome.jp/kegg/)
Protein–protein interaction	PPID (http://www.anc.ed.ac.uk/mscs/PPID/)
	DIP (http://dip.doe-mbi.ucla.edu/dip/Download.cgi)
Disease association	OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM)
References	PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi)

in hundreds of protein entries. In another example, thousands of false-positive entries were removed because they were annotated as being submitted by a company named Synapse. After manual review of thousands of entries, we retrieved over 10 000 proteins and assigned them with SynO.

We combined the two sets of proteins and removed redundant entries following the strategy of International Protein Index (23). We considered two UniProt proteins in a species redundant if they were ⩾ 95% identical over ⩾ 95% of the length of the shorter sequence, based on pair-wise BLASTP of all sequences in the species. Among redundant proteins we selected SwissProt sequences over Trembl sequences. For those sequences from the same data source, we selected longer sequences over shorter ones. The resulting SynDB contains 14 000 non-redundant proteins, including 3000 in human and is the most comprehensive collection of synapse-related proteins to date.

ANNOTATIONS AND WEB INTERFACE DESIGN

To enhance SynDB's utility as an information resource, we developed parsers in Perl to retrieve extensive information on protein sequences, expression, protein–protein and protein-small molecule interactions, disease associations and literature references. Known 3D structures or potential structure templates were retrieved by pair-wise BLASTP comparison between SynDB proteins and non-redundant proteins with known structures from PDB_SELECT_25 (24). In addition, cross-references to ModBase (25) are also provided. Potential metabolic pathways involved were identified by running the KOBAS system against the KEGG database (26,27). Table 1 shows the protein features and related external databases. The information for each protein is integrated and presented in a single graphical web page. For example, the SynDB entry page for Huntingtin Interacting Protein 1 (HIP1) (Figure 2 and http://syndb.cbi.pku.edu.cn/sdb_pro.php?id=HIP1_HUMAN) shows that HIP1 is located on chromosome 7, highly expressed in brain and involved in the Huntington disease pathway. It is included in the Online Mendelian Inheritance in Men database (28) as providing an important molecular link between huntingtin and the neuronal cytoskeleton and has sequence available for download.

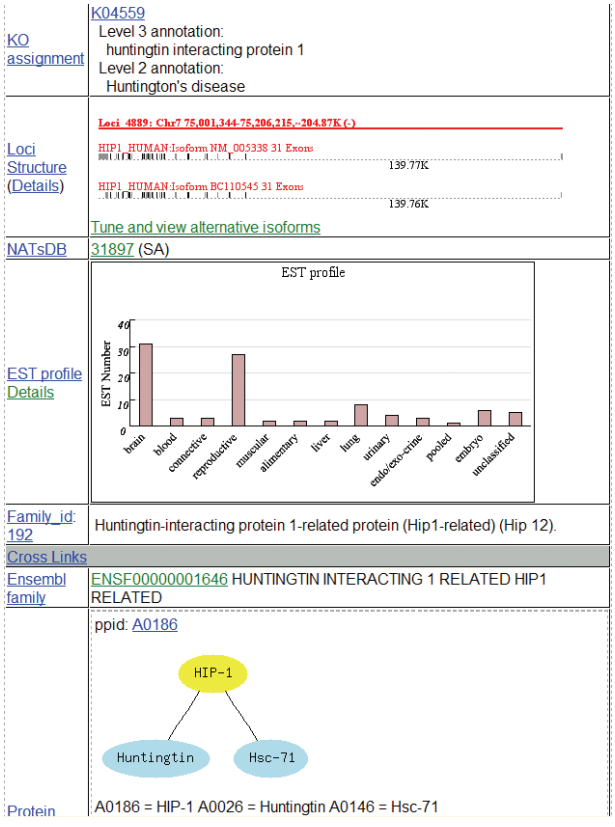

Figure 2. A part of protein entry page for human Huntingtin Interacting Protein 1 (HIP1). See 'Annotations and Web Interface Design' for a brief description and http://syndb.cbi.pku.edu.cn/help/sdb_help.php for detail.

We implemented six interactive browsing options in SynDB. Users can browse synapse proteins by 'SynO' or 'GO', displayed as hierarchical trees. They can zoom in on a particular branch of the ontology by clicking on the '+' sign to expand the branch. For example, a user interested in 'transmitters release and endocytosis' may expand this

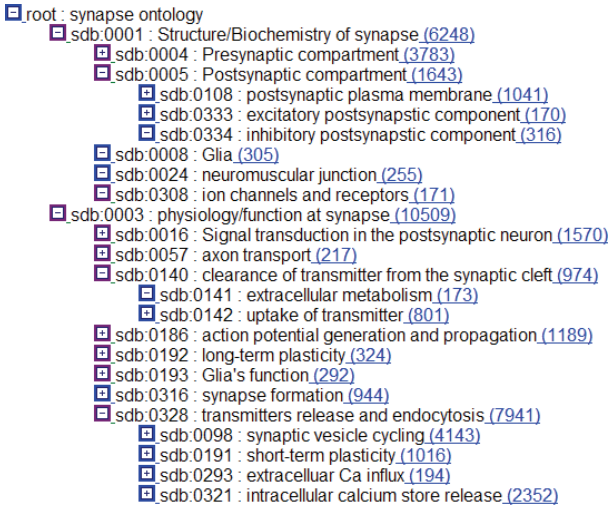

Figure 3. The Synapase Ontology browser. '+' indicates this term could be expanded to list it's child terms.

category and focus on 'synaptic vesicle cycling' (Figure 3). 'Protein Domains' were grouped into InterPro domain family groups which could be expanded by clicking on the group name. For each domain, the numbers of total, human, mouse and rat proteins are shown. To facilitate study of the evolution of a domain, an 'Expand' link shows all species that contain the domain and its prevalence in each. To further facilitate study of the evolution of the synapse proteome across different species, we clustered all SynDB proteins by sequence similarity (BLAST *E*-value cutoff e^{-10}) using Tribe-MCL (29) and made the clusters available in the 'MCL Cluster' browser. A separate 'Species' browser lists all species represented in SynDB, in order of decreasing number of proteins. Finally, the 'Chromosomal Location' browser (Figure 4), available for human, mouse and rat, allows users to cursor over or click on chromosomal locations to see gene details. Because a number of neural gene families, such as olfactory receptors, have been known to form gene clusters along chromosomes (30), we implemented a 'Locus number' field in the 'Chromosomal Location' browser that allows the user to enter a cutoff number and view gene clusters of at least that size along a chromosome. Two loci are considered to belong to a cluster if their intergenic distance is less than 500 kb (30,31).

SynDB supports searching by text with Boolean operators. It also supports searching by amino acid or nucleotide sequence similarity with BLAST. Information in SynDB is stored in a MySQL relational database comprised of over 100 tables. Sequences can be downloaded directly from the web and the complete database is available from the authors. We will keep SynO up-to-date by regular review of the latest literature as well as users' and collaborators' comments. We used the Perl scripts which will automatically update the sequences in SynDB followed by manual review.

DISCUSSION

The brain is a complex and subtle network of neurons that communicate with each other via synapses. Chemical synapses are asymmetric contact and play key roles in

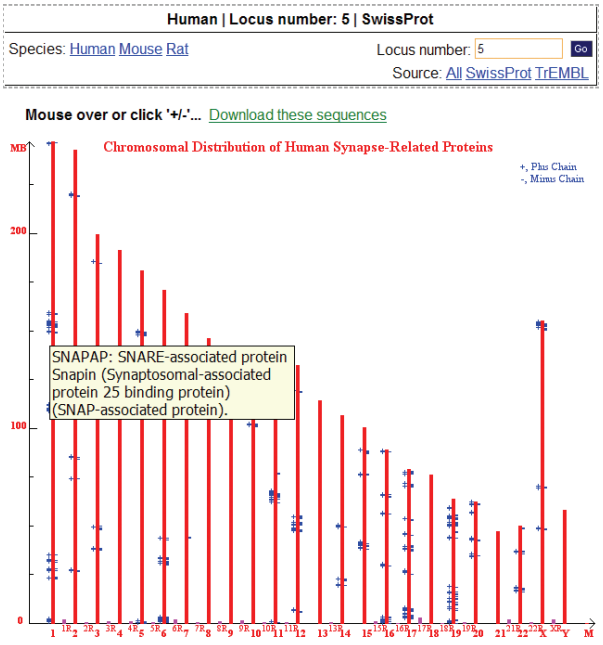

Figure 4. Chromosomal browser of SynDB: The Chromosomal browser is available for human, mouse and rat. The *x*-axis shows the different human chromosomes. Users can mouse over or click on a '+' to view a protein translated from the gene locating in that loci of the chromosome. Users can also input a number in 'Locus number' to view gene clusters with as few as that members. From this figure, user can get the information which chromosome and which region of the chromosome derives more synapse-related genes.

information processing and storage, behavior and disease. In order to better organize the wealth of synapse-related information and facilitate understanding of synapses, we developed SynDB, an online database for the synapse proteome. SynDB aims to enable systematic studies of the synaptic functions and structures at proteomic level. A focused ontology is essential for the development of such a database because of the numerous and diverse proteins involved. Beyond general-purposed ontologies such as MeSH and GO (18,19), focused ontologies such as SynO are important because they can provide more specific, complete and resolved information to scientists, such as neuroscientists interested in synaptic function. In fact, of 177 SynO terms, only 24 were derived from MeSH and GO.

In its first year online, SynDB has had over 600 000 external hits (excluding search engine crawlers). SynDB's objective is to serve as a repository for current knowledge and a potential starting point for experimental design or *in silico* data mining.

ACKNOWLEDGEMENTS

This work was supported by grants from the China National High-tech 863 Program to L.W. and 973 Program (2006CB500800) and NSFC to Z.Z. We thank Drs Peace Cheng, Tim Qing-Rong Liu and John Reiland for valuable suggestions. Funding to pay the Open Access publication charges for this article was provided by China Ministry of Education 'Program of Introducing Talents of Discipline to Universities' (B06001).

Conflict of interest statement. None declared.

REFERENCES

1. Buetow,K.H. (2004) The NCI Center for Bioinformatics (NCICB): building a foundation for *in silico* biomedical research. *Cancer Invest*., **22**, 117–122.
2. Coleman,W.B. (2004) Cancer bioinformatics: addressing the challenges of integrated postgenomic cancer research. *Cancer Invest*., **22**, 161–163.
3. Nakagawara,A. and Ohira,M. (2004) Comprehensive genomics linking between neural development and cancer: neuroblastoma as a model. *Cancer Lett*., **204**, 213–224.
4. Lefranc,M.P. (2004) IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics. *Mol. Immunol*., **40**, 647–660.
5. Rammensee,H.G. (2003) Immunoinformatics: bioinformatic strategies for better understanding of immune function. Introduction. *Novartis Found. Symp*., **254**, 1–2.
6. Boguski,M.S. and Jones,A.R. (2004) Neurogenomics: at the intersection of neurobiology and genome sciences. *Nature Neurosci*., **7**, 429–433.
7. Choudhary,J. and Grant,S.G. (2004) Proteomics in postgenomic neuroscience: the end of the beginning. *Nature Neurosci*., **7**, 440–445.
8. Grant,S.G. (2003) Systems biology in neuroscience: bridging genes to cognition. *Curr. Opin. Neurobiol*., **13**, 577–582.
9. Skuse,D. (2004) Genetics and genomics of neurobehavioral disorders. *J. Child Psychol. Psychiat*., **45**, 1180–1181.
10. Hamacher,M., Klose,J., Rossier,J., Marcus,K. and Meyer,H.E. (2004) 'Does understanding the brain need proteomics and does understanding proteomics need brains?'—second HUPO HBPP workshop hosted in Paris. *Proteomics*, **4**, 1932–1934.
11. Insel,T.R., Volkow,N.D., Landis,S.C., Li,T.K., Battey,J.F. and Sieving,P. (2004) Limits to growth: why neuroscience needs large-scale science. *Nature Neurosci*., **7**, 426–427.
12. Nelson,P.G. (2005) Activity-dependent synapse modulation and the pathogenesis of Alzheimer disease. *Curr. Alzheimer Res*., **2**, 497–506.
13. Husi,H. and Grant,S.G. (2002) Construction of a protein–protein interaction database (PPID) for synaptic biology. In Kotter,R. (ed.), *Neuroscience Databases: A Practical Guide*. Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 1–62.
14. Gruber,T.R. (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**, 199–220.
15. Cowan,W.M., Sūdhof,T.C., Stevens,C.F. and Davies,K. (2000) *Synapses*. The Johns Hopkins University Press, Baltimore and London.
16. Kandel,E.R., Schwartz,J.M. and Jessell,T.M. (2000) *Principles of Neural Science*, 4th edn. McGraw-Hill Companies, New York, NY.
17. Hille,B. (2001) *Ion Channels of Excitable Membranes*, 3rd edn. Sinauer Associates, Inc., Sunderland, Massachusetts.
18. 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. *Nature Genet*., **25**, 25–29.
19. Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc*., **88**, 265–266.
20. Smith,B., Ceusters,W., Klagges,B., Kohler,J., Kumar,A., Lomax,J., Mungall,C., Neuhaus,F., Rector,A.L. and Rosse,C. (2005) Relations in biomedical ontologies. *Genome Biol*., **6**, R46.
21. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
22. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
23. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
24. Boberg,J., Salakoski,T. and Vihinen,M. (1992) Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins*, **14**, 265–276.
25. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*., **32**, D217–D222.
26. Mao,X., Cai,T., Olyarchuk,J.G. and Wei,L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**, 3787–3793.
27. Wu,J., Mao,X., Cai,T., Luo,J. and Wei,L. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res*., **34**, W720–W724.
28. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **30**, 52–55.
29. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.
30. Niimura,Y. and Nei,M. (2003) Evolution of olfactory receptor genes in the human genome. *Proc. Natl Acad. Sci. USA*, **100**, 12235–12240.
31. Young,J.M., Kambere,M., Trask,B.J. and Lane,R.P. (2005) Divergent V1R repertoires in five species: amplification in rodents, decimation in primates, and a surprisingly small repertoire in dogs. *Genome Res*., **15**, 231–240.