Nucleic Acids Research, 2006, Vol. 34, Database issue ***D213–D217***
doi:10.1093/nar/gkj069

LOCATE: a mouse protein subcellular localization database

J. Lynn Fink[1,2,*], Rajith N. Aturaliya[1,2], Melissa J. Davis[2], Fasheng Zhang[2], Kelly Hanson[1,2], Melvena S. Teasdale[2], Chikatoshi Kai[3], Jun Kawai[3,4], Piero Carninci[3,4], Yoshihide Hayashizaki[3,4] and Rohan D. Teasdale[1,2]

[1]ARC Centre in Bioinformatics and [2]Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia, [3]Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan and [4]Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Received August 15, 2005; Revised and Accepted October 8, 2005

ABSTRACT

We present here LOCATE, a curated, web-accessible database that houses data describing the membrane organization and subcellular localization of proteins from the FANTOM3 Isoform Protein Sequence set. Membrane organization is predicted by the high-throughput, computational pipeline MemO. The subcellular locations of selected proteins from this set were determined by a high-throughput, immunofluorescence-based assay and by manually reviewing >1700 peer-reviewed publications. LOCATE represents the first effort to catalogue the experimentally verified subcellular location and membrane organization of mammalian proteins using a high-throughput approach and provides localization data for ∼40% of the mouse proteome. It is available at http://locate.imb.uq.edu.au.

INTRODUCTION

Determination of the membrane organization and the subcellular location of a protein are essential to understanding its biochemical function. A cell is divided into different cellular compartments and each compartment is associated with a different range of biochemical processes; by localizing a protein to a specific compartment, or set of compartments, the cellular role of the protein can be inferred. This information can provide insight into the functions of hypothetical or novel proteins and can provide a more specific organellar context in which to investigate a particular protein. Historically, these data have been difficult to produce on a large scale for higher eukaryotic organisms. However, recent advances in membrane organization prediction methods and high-throughput subcellular localization assays have made it possible to generate these datasets. We used high-throughput methods to predict the membrane organization for the entire mouse proteome and to determine the subcellular localization of a subset of the proteome. We then developed a database, LOCATE, to organize and warehouse these data.

DATABASE CONTENT

Dataset

The mouse proteome dataset we used was the FANTOM3 Isoform Protein Sequence set (IPS7) generated by the RIKEN FANTOM Consortium (1). This dataset is comprised of protein sequences based on transcript sequences generated from direct sequencing of full-length transcripts. The sequenced transcripts were clustered into transcriptional units (TUs) where a TU is a grouping of transcripts that arise from a single genomic locus and share at least one nucleotide having the same genomic location and orientation. The IPS7 dataset contains 33 451 protein sequences encoded by 19 853 TUs.

Membrane organization

Protein orientation with respect to the membrane was predicted by MemO, a high-throughput, automated pipeline, which combines publicly available feature predictors with empirically determined annotation rules (1,2) (M. J. Davis, F. Clark, J. L. Fink, Z. Yuan, F. Zhang, T. Kasukawa, Y. Hayashizaki, P. Carnici and R. D. Teasdale, manuscript in preparation). The pipeline is described briefly here.

*To whom correspondence should be addressed. Tel: +61 7 3346 2633; Fax: +61 7 3346 2101; Email: l.fink@imb.uq.edu.au

Table 1. Distribution of membrane organization classes and high-quality localization data in LOCATE

Membrane organization class	MemO data	Subcellular localization data		
	IPS proteins in class (TUs/isoforms)	Isoforms with experimental data	TUs with literature-mined data	Total represented (TUs/isoforms)
Soluble, intracellular protein	13 105/22 265	0	302	302/353
Soluble, secreted protein	2190/2948	0	340	340/469
Type I membrane protein	1038/1548	0	377	377/653
Type II membrane protein	2149/2869	207	408	549/766
Multi-pass membrane protein	2538/3821	210	325	460/652
Total proteins analyzed	19 538/33 451	417	1752	2028/2893

The MemO Data columns show the absolute numbers of proteins classified by MemO into each membrane organization class. The 'Subcellular localization data' columns show the number of protein isoforms that have an experimentally determined subcellular location and the number of transcriptional units (TUs) that have a literature-mined subcellular location as well as the total numbers of TUs and isoforms that have subcellular localization data. Localization data mined from other databases is not included here.

Prediction of signal peptides was performed by a local implementation of SignalP v2.0 (3) and by the Australian National Genomic Information Service (ANGIS, http://biomanager.angis.org.au) version of SPScan. A protein was predicted to contain a signal peptide if the averaged and normalized raw output scores from both methods exceeded a threshold identified to maximize the proportions of true positives and true negatives on a training set.

α-Helical transmembrane domain prediction was performed by a consensus method consisting of five currently available predictors: HMMTOP (4), TMHMM v2.0 (5), SVMTM v3.0 (6), MEMSAT (7) and DAS (8). A protein was said to contain a transmembrane domain if at least 7, but no more than 42, consecutive residues in the protein (ignoring a gap of <4 residues) were predicted to participate in a transmembrane domain by at least three of the five predictors.

The prediction of the absence or presence of the signal peptide and transmembrane domain provided a classification into one of five categories of membrane organization:

- soluble intracellular proteins (no transmembrane domains or signal peptide);
- soluble secreted proteins (signal peptide, no transmembrane domains);
- type I membrane proteins (one transmembrane domain, signal peptide) (9);
- type II membrane proteins (one transmembrane domain, no signal peptide) (9);
- multi-pass membrane protein (multiple transmembrane domains) (9).

We applied this pipeline to the 33 451 protein sequences in the IPS7 dataset and identified 5116 (∼15%) proteins containing signal peptides, and 8238 (∼25%) proteins containing transmembrane domains. These proteins were then allocated to the five membrane organization categories based on combinations of those features. The class breakdown of proteins is shown in Table 1.

Subcellular localization

Proteins were selected for experimentation based on clone availability and the extent of previous characterization of their subcellular localization. When selecting multipass membrane proteins, only those without a predicted ER signal peptide were chosen. N-terminally tagged myc-gene of interest expression constructs were generated using a modified overlapping PCR methodology originally reported by Suzuki *et al.* (10). The expressed protein, within fixed transfected HeLa cells, was detected by indirect immunofluorescence and representative images were collected and analyzed to determine the protein's subcellular localization. To date, experimental subcellular localization data have been generated for 417 of these selected proteins and localization data based on primary literature review have been gathered for 1752 TUs.

Both the experimental and literature-mined localization data were manually examined and evaluated for sufficient quality prior to addition to the database. When evaluating literature-mined localization data, only papers describing the localization of full-length proteins in individual mammalian cells in which the protein is detected directly were included in our analysis. These peer-reviewed observations were not reinterpreted. However, some observations were excluded when considered not to be of a sufficient quality.

Because it was not always possible to determine to which protein isoform the literature data referred, we assigned the literature-mined location to all protein isoforms encoded by the corresponding TU. Table 1 summarizes the subcellular localization statistics by membrane organization class.

To provide as complete a location description as possible for any given protein, we also include localization data mined from other online databases including LIFEdb (11), Mouse Genome Informatics (12), UniProt (13), RefSeq (14) and others. A total of 7410 TUs and 11 353 protein isoforms are annotated with these data. In total, we have localization data for 8017 TUs and 12 598 protein isoforms representing 41 and 37% of the IPS7 set, respectively.

Data presentation

General information. Information in LOCATE is displayed as a web page which describes a particular protein entry in detail. The page is divided into sections which summarize several types of data. The top of the page contains a summary of the MemO classification and the subcellular localization of the protein as well as associated metadata provided by FANTOM3 annotations such as the protein identifier, a functional description, protein name synonyms, the source organism and links to other databases which also contain this protein.

Transmembrane topology and predicted domains. Knowing what functional domains and motifs exist in a protein is

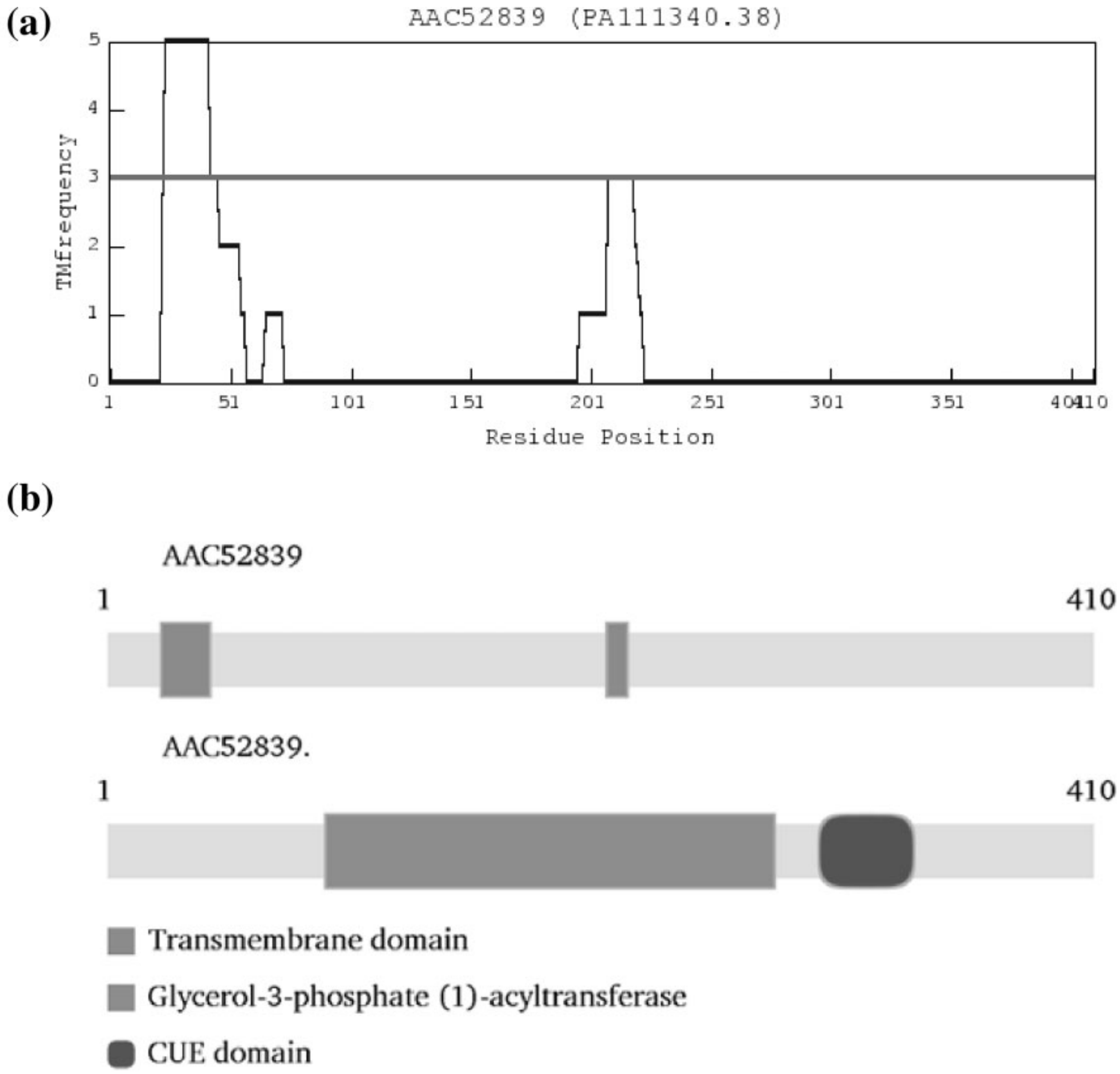

Figure 1. Visualization of MemO- and Pfam- and SCOP-predicted motif data. (**a**) Plots the number of computational methods (from 0 to 5) that predict whether a residue in the protein sequence participates in a helical transmembrane domain. Five independent methods are used in the TMD prediction; we assign a residue to a TMD if at least three of the five methods have a positive prediction at that position in the sequence and the range of the predicted TMD fulfils a set of rules defined in the MemO pipeline (M. J. Davis, F. Clark, J. L. Fink, Z. Yuan, F. Zhang, T. Kasukawa, Y. Hayashizaki, P. Carnici and R. D. Teasdale, manuscript in preparation). (**b**) A schematic diagram of a protein sequence with predicted domains mapped onto it. In this particular diagram, the transmembrane domains predicted by MemO are shown at the top of the figure and the domains predicted by Pfam or SCOP are shown in the bottom of the figure. The schematics are vertically aligned to show the positional relationships of the predicted TMDs and other domains.

extremely useful when attempting to decipher the cellular role of the protein. We have generated predictions of Pfam and SCOP domains for all proteins in the database and have displayed the predicted domains on a graphical protein schematic diagram alongside the membrane organization data (Figure 1). The presence and position of certain domains in relation to predicted transmembrane domains can provide insights into the validity of the functional annotation of the protein (if one exists) as well as the validity or range of the transmembrane domain prediction.

Subcellular location data. If a protein entry has high-throughput subcellular localization data, we display the subcellular location(s) in which that particular protein isoform was observed and a high-resolution fluorescent-image which best illustrates the observed localization. Information about the experimental conditions such as the cell type and epitope used in the localization assays is also displayed. If a protein entry has subcellular localization data mined from literature, we display the determined subcellular location(s), the PubMed ID, and a full citation of the data source.

Controlled vocabulary. Consistent naming of subcellular locations is critical to the integrity and extensibility of the LOCATE data. Therefore, we have constructed a controlled vocabulary which describes both experimentally determined and literature-mined subcellular locations. In the case of high-throughput experimental subcellular localization assays, it is not always possible to determine the exact cellular compartment to which the protein is observed to localize. To address

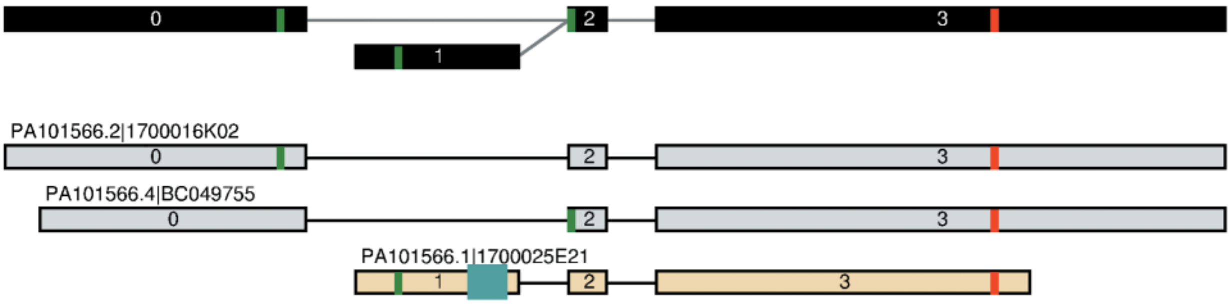

Figure 2. Splicing graph. This graph shows the observed exons and splice junctions for the transcriptional unit 101566 and the splice isoforms of the transcripts that arise from this transcriptional unit. The light gray color represents soluble, cytoplasmic proteins (PA101566.2 and PA101566.4); light orange represents a Type II membrane protein (PA101566.1); black represents all observed exons. The green and red bars represent the observed start and stop codons, respectively. The teal rectangle represents the position and range of the MemO-predicted transmembrane domain; note that the transmembrane domain occurs in the exon that only appears in the Type II membrane protein and not in the soluble, cytoplasmic proteins. This is a clear example of how alternate splicing of these transcripts may change the proteins' membrane organization.

this problem, our controlled vocabulary contains a hierarchical set of terms that allows the call to be only as specific as the data allow. This system also reflects the confidence of the localization call; use of a very specific term implies higher confidence. Some proteins have been observed to localize to more than one subcellular compartment; in these cases, we allow the use of multiple terms to describe the observed locations. When mining subcellular localization data from the literature, we use terms that allow for different levels of location resolution and for cellular components that are specific to cells with a lineage or morphology that differs from the model cells used in our experiments. In both vocabularies, we use Gene Ontology (15) terms to describe subcellular locations whenever possible (see the LOCATE website for more details).

Observed spliced isoforms. For each protein in the database, we display a list of all proteins that belong to the same TU to allow comparisons between each of the observed protein isoforms. Specifically, we display the membrane organization and length of each isoform on a splicing graph which illustrates the observed exons and the various alternate splice forms for that particular TU (Figure 2). These graphs enable analysis of the pattern of membrane organization variation within the observed protein isoforms and examination of the possible effects of alternative splicing on membrane organization. The graphs were generated by a customized version of the Splicing Graph Module (16).

Data accessibility

This database does not seek to duplicate information contained in other databases unless it is particularly useful when viewed in juxtaposition with the subcellular localization or membrane organization data. However, we understand the value of convenient data accessibility and provide links to offsite resources such as SymAtlas (17), GenBank (18), RIKEN (1), MGI (19), READ (20), Pfam (21), SCOP (22), UniProt (13), OMIM (23), Entrez Gene (24), BIND (25), the GeneNetwork (26) and the Mouse Retrovirus Tagged Cancer Gene Database (RTCGD) (20) where applicable.

Because the major aim of this database effort is to present protein subcellular location data and the predicted membrane organization of the protein, these two features are the primary search mechanisms; proteins can be retrieved by protein class, subcellular localization or both. Alternatively, individual protein entries can be retrieved by searching the database with a protein ID (RIKEN clone/IPS ID, GenBank accession number, Entrez Gene ID), by protein name, by Pfam or SCOP accession number, or by functional description. BLAST searches against the database, and subsets of the database, are also available. The BLAST results are enhanced to display the membrane organization of the hits. We also offer a number of batch data retrieval options. The proteins in any given search can be retrieved as FASTA-formatted protein or transcript sequences, subcellular localization data, membrane organization data or protein schematics. XML-marked-up documents containing these data can also be obtained.

CONCLUSIONS

LOCATE represents a significant contribution to the biological research community by organizing and presenting membrane organization and subcellular localization data for the mouse proteome. The LOCATE search interface allows users to retrieve data and sets of data using several different approaches. The interface to individual proteins was designed to maximize ease of interpretation by providing summaries or visualizations that contain the most relevant points of data; links are provided to the raw data or other details that are necessary for a careful evaluation of the experimental results. LOCATE data can be retrieved as individual entries or downloaded as HTML, plain text or XML files.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Nicholas Hamilton for implementing DomainDraw, the domain drawing program; Robert Luetterforst for assistance with the literature mining; and Emma Redhead for designing the LOCATE XML schema and XML document generator. The work was supported by funds from the Australian Research Council (ARC) and by the Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H., and the Research Grant for the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. R.D.T. is supported by a National Health and Medical Research Council of Australia R. Douglas Wright Career Development Award. R.N.A. is supported by a Postgraduate Research Scholarship from the IMB, University of Queensland. M.J.D. is supported by the National Institute for Diabetes, Digestion and Kidney Disease, National Institutes of Health (DK63400) as part of the Stem Cell Genome Anatomy Project (http://www.scgap.org/). Funding to pay the Open Access publication charges for this article was provided by University of Queensland and Australian Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
2. Kanapin,A., Batalov,S., Davis,M.J., Gough,J., Grimmond,S.M., Kawaji,H., Magrane,M., Matsuda,H., Schonbach,C., Teasdale,R.D. *et al.* (2003) Mouse proteome analysis. *Genome Res.*, **13**, 1335–1344.
3. Nielsen,H. and Krogh,A. (1998) In Glasgow,J. (ed.), *Sixth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Vol. 1, pp. 122–130.
4. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
5. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
6. Yuan,Z., Mattick,J.S. and Teasdale,R.D. (2004) SVMtm: support vector machines to predict transmembrane segments. *J. Computat. Chem.*, **25**, 632–636.
7. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.
8. Cserzo,M., Wallin,E., Simon,I., vonHeijne,G. and Elofsson,A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.*, **10**, 673–676.
9. Goder,V. and Spiess,M. (2001) Topogenesis of membrane proteins: determinants and dynamics. *FEBS Lett.*, **504**, 87–93.
10. Suzuki,H., Fukunishi,Y., Kagawa,I., Saito,R., Oda,H., Endo,T., Kondo,S., Bono,H., Okazaki,Y. and Hayashizaki,Y. (2001) Protein–protein interaction panel using mouse full-length cDNAs. *Genome Res.*, **11**, 1758–1765.
11. Bannasch,D., Mehrle,A., Glatting,K.H., Pepperkok,R., Poustka,A. and Wiemann,S. (2004) LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. *Nucleic Acids Res.*, **32**, D505–D508.
12. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
13. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
14. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
15. 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.
16. Lee,B.T., Tan,T.W. and Ranganathan,S. (2004) DEDB: a database of *Drosophila melanogaster* exons in splicing graph form. *BMC Bioinformatics*, **5**, 189.
17. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
18. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
19. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
20. Akagi,K., Suzuki,T., Stephens,R.M., Jenkins,N.A. and Copeland,N.G. (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.*, **32**, D523–D527.
21. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
22. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
23. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
24. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
25. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
26. Wu,C.C., Huang,H.C., Juan,H.F. and Chen,S.T. (2004) GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. *Bioinformatics*, **20**, 3691–3693.