

# TCDB: the Transporter Classification Database for membrane transport protein analyses and information

Milton H. Saier, Jr\*, Can V. Tran and Ravi D. Barabote

Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116, USA

Received April 8, 2005; Revised July 7, 2005; Accepted July 18, 2005

## ABSTRACT

**The Transporter Classification Database (TCDB) is a web accessible, curated, relational database containing sequence, classification, structural, functional and evolutionary information about transport systems from a variety of living organisms. TCDB is a curated repository for factual information compiled from >10 000 references, encompassing ~3000 representative transporters and putative transporters, classified into >400 families. The transporter classification (TC) system is an International Union of Biochemistry and Molecular Biology approved system of nomenclature for transport protein classification. TCDB is freely accessible at <http://www.tcdb.org>. The web interface provides several different methods for accessing the data, including step-by-step access to hierarchical classification, direct search by sequence or TC number and full-text searching. The functional ontology that underlies the database structure facilitates powerful query searches that yield valuable data in a quick and easy way. The TCDB website also offers several tools specifically designed for analyzing the unique characteristics of transport proteins. TCDB not only provides curated information and a tool for classifying newly identified membrane proteins, but also serves as a genome transporter-annotation tool.**

## INTRODUCTION

Transport is an essential function of every living cell. Thousands of researchers worldwide devote their efforts to the study of this basic function. Transmembrane transport protein biology has undergone an explosive growth in scientific discovery in the last several years. The recent high-resolution structural elucidation of many transporters [such as BtuB (1,2), AcrB (3), LacY (4), GlpT (5), EmrE (6) and MsbA

(7,8)] has enabled investigation into the molecular dynamics of fundamental transport processes. As the structures of other unique transport systems are revealed, the power of computational methods in transporter analysis and prediction will grow exponentially.

Transporters play critical roles in the life science industries. Absorption, distribution and excretion of drugs within the human body are regulated by transporters which must be factored into pharmacological studies. Growing numbers of pathogenic microbial strains resistant to many common antibiotics are wreaking havoc on the public health system. Computational prediction of potential inhibitors of multi-drug resistance transporters and of transporters that offer a survival advantage to pathogenic microbes would help in the design of novel anti-microbial drugs. Additionally, drug resistance in cancer cells, caused by drug efflux pumps, is of increasing concern in the field of oncology. The immense importance of studying transport proteins and the enormity of the data available on these proteins have warranted the systematic classification of transport systems in order to promote a comprehensive understanding of one of the basic functions of all living cells (9–11).

The Transporter Classification Database (TCDB) is a freely accessible web resource (<http://www.tcdb.org>) allowing access to the data upon which the transporter classification (TC) system is based. All data in TCDB is a compilation of published information from over 10 000 references. Approximately 3000 distinct proteins from all kinds of known living organisms are organized into >400 transporter families based on the TC system. Data are added on a continuous basis as new functional data are published and new transport systems are identified. Several resources for analyzing transmembrane proteins are provided on the website. Uniting a multitude of resources and biological databases for centralized computational analysis facilitates the ease-of-use that life scientists require when researching transporters. The availability of TCDB has allowed major basic research advances including answering fundamental biological questions and determining the routes of evolution taken for the appearance of these proteins (12,13).

\*To whom correspondence should be addressed. Tel. +1 858 534 4084; Fax: +1 858 534 7108; Email: [msaier@ucsd.edu](mailto:msaier@ucsd.edu)

## THE TRANSPORTER CLASSIFICATION SYSTEM

The TC system consists of a set of representative protein sequences, most of which have been functionally characterized. These transporters are classified with a five-character designation, as follows: D1.L1.D2.D3.D4. D1 (a single digit) corresponds to the transporter class (i.e. channel, carrier, primary active transporter, group translocator or transmembrane electron flow carrier). L1 (a letter) corresponds to the transporter subclass, which, e.g. in the case of primary active transporters, refers to the energy source used to drive transport. D2 (a number) corresponds to the transporter family (sometimes actually a superfamily). D3 (a number) corresponds to the subfamily (or the family of a superfamily) in which a transporter is found. D4 (a number) corresponds to the transporter itself. This refers to a specific transport system with a defined range of substrates, a known polarity of transport, an energy source that drives vectorial movement of the substrate and a mechanism of action. Only in one of the TC classes (class 9) is this information incomplete or absent.

A TC number for proteins in classes 1–5 provides the following information: (i) the type of transporter (D1); (ii) the subtype of transporter; e.g. for primary active transporters, the type of energy source used to drive transport (L1); (iii) the specific family to which the transporter belongs (D2); (iv) the subfamily to which the transporter belongs (D3) and (v) the specific transporter with a given polarity, specificity and mechanism of action (D4). Because phylogeny reflects the mechanism, mode of energy coupling, polarity and substrate specificity of a transporter, a functional/phylogenetic system of classification provides far more information than would be possible with a purely functional one. The basis for the architecture of the TC system as approved by the International Union of Biochemistry and Molecular Biology has been enunciated in detail in (11). The full architectural consideration of the TC system is beyond the scope of this article.

At the heart of TCDB are the protein families. Although there are a few examples of transporters within families that can use more than a single mode of action or can use a mechanism dissimilar from that used by other members of the family, for the most part, family membership implies similar function and mechanism. Any two transport systems in the same subfamily of a transporter family that transports the same substrate(s) are given the same TC number, regardless of whether they are orthologs (e.g. arose in distinct organisms by speciation) or paralogs (e.g. arose within a single organism by gene duplication). However, because different types of information may be available for two proteins of the same specificity (e.g. regulatory data, subcellular localization data, disease association data), two or more such systems may on occasion be included in TCDB. It should be noted that within practical limits, TCDB reflects the current state of our knowledge about the proteins included within it.

If two transporters exhibit weak similarity but operate by the same transport mechanism, two distinct subfamilies will represent the two transporters and their close homologs. Sequenced homologs of unknown function are normally not assigned a TC number unless they represent a unique family/subfamily or are from an underrepresented kingdom. Transporter classes 1–5 are well-defined classes, class 8 is reserved for accessory transport proteins, while class 9 is for

transporters which are incompletely characterized. When sufficient information warrants their transfer to one of the defined classes (1–5), they will acquire a new TC number. Class 9 is therefore in a continual state of flux. TC classes 6 and 7 are currently unused but will be introduced if additional classes of transporters are discovered.

## DATABASE CONTENT AND ACCESS

The TCDB web application is based upon a three-tier architecture. The underlying tier of the system is the open source database MySQL. An Apache-PHP applications server forms the middle tier, which retrieves tuples from the database and returns populated HTML data to the web browser client, the superficial tier. This architecture resides upon a dual processor PowerPC G5 running Mac OS X operating system.

The raw data stored in TCDB originates from multiple sources. Protein sequences are obtained from the Swiss-Prot knowledgebase (14). The 3D macromolecular structures are retrieved from the PDB (15) in mmCIF format. Protein domains from Interpro (16) are integrated with TCDB. Human transporters with nomenclature approved by the Human Genome Nomenclature Committee are presented as reported in GENEW (17). Life science journal citations are integrated, and in the case of human transporters, as well as transporters with structural data, PubMed ID numbers (18) are provided.

Encoded within the TCDB relational schema is the functional and phylogenetic TC system taxonomy. The clickable ‘TC System’ button on the main page (Figure 1) provides access to the data in TCDB. A two-vertical-frame architecture of the web page allows quick browsing through the hierarchical TC system in the left frame as well as access to detailed description or to protein information in the right window. Thus, users can access the classification system through the intuitive interface that allows the user to read descriptions of entries at varying levels of granularity. The user can start at the top of the hierarchy and descend through the taxonomy. At the deepest level, the user can retrieve individual protein information such as Swiss-Prot accession number, the primary sequence, source organism and the protein name, length, molecular weight and probable topology (Figure 2). Several links, such as links to the Swiss-Pfam database, the ExpASY server, the Swiss Institute of Bioinformatics BLAST Network service, and transmembrane segment (TMS) prediction are provided (Figure 2). A link to the FASTA formatted protein sequence as well as a quick link to the hydrophathy and amphipathicity plots for the protein are available. A user can enter the TC family name or TC number to search the database (Figure 1). Additionally, the ‘Search’ link at the top of the page (Figure 1) allows advance searches by keyword, disease name, protein name, etc. Cited literature in TCDB can be searched as well.

## ACCESSORY DATA

Phylogenetic analyses and refined sequence comparisons of many transport systems in our laboratory have revealed distant relationships between many families in TCDB (19). These distant relationships are detailed in a section named ‘TC

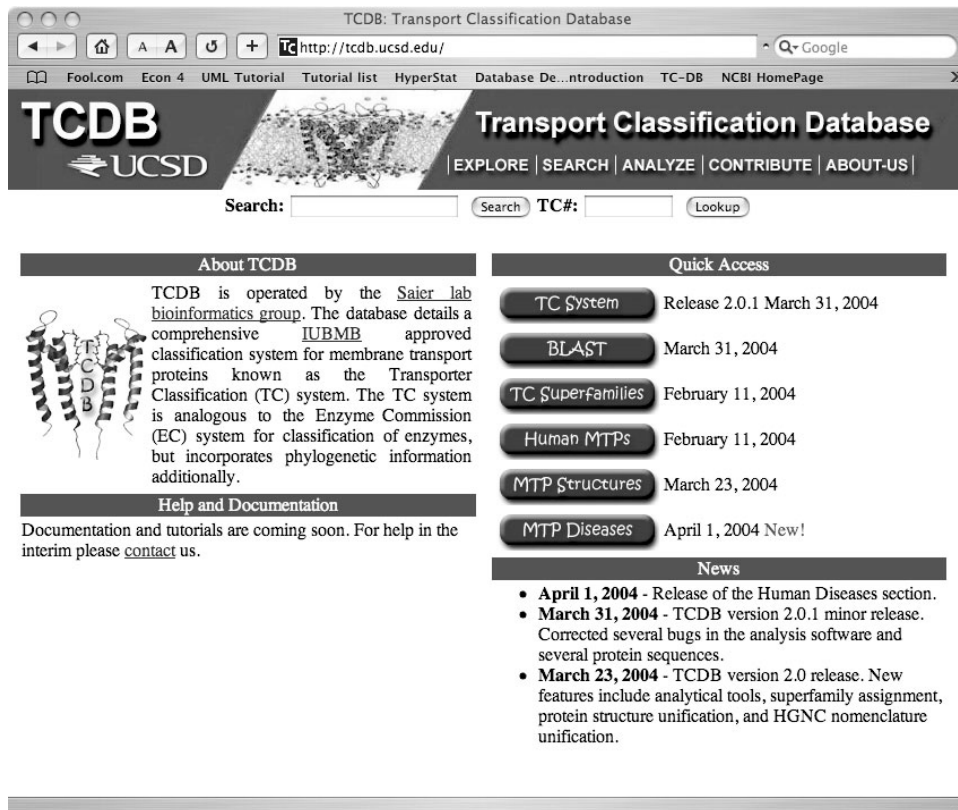


Figure 1. The user-friendly front page of TCDB (<http://www.tcdb.org>) where its features are easily accessible.

### 4.A.1.1.1 Glucose porter (PtsG)

<p><b>Accession Number:</b> P05053  <b>Protein Name:</b> PTGB aka PTSG aka GLCA aka UMG aka B1101 aka C1373 aka Z1740 aka ECS1479  <b>Length:</b> 477  <b>Molecular Weight:</b> 50676.00  <b>Species:</b> <i>Escherichia coli</i>  <b>Transmembrane Segments (TMSs):</b>  <b>Number of Alpha Helical TMSs:</b> 10  <b>Orientation of the first TMS:</b> OUT</p>	<p><b>Links</b></p> <p>Predict  TMSs  Swiss-Pfam  ExPASy  Quick  BLAST</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------

---

Window Size:  Angle:

---

FASTA Formatted Sequence

```

1: MFKNAFANLQ KVGKSLMLPV SVLPIAGILL GVGSANFSWL PAVVSHVMAE AGGSVFANMP
61: LIFAIGVALG FTNNDGVSAL AAVVAYGIMV KTMVVVAPLV LHLPAEEIAS KHLADTGVLG
121: GIISGAIAAY MFNRFYRIKL PEYLGFFAGK RFVPIISGLA AIFTGVVLSF IWPIGSAIQ
181: TFSQWAAAYQN PVVAFGIYGF IERCLVPPGL HHIWNVFPQM QIGEYTNAG QVFHGDIPRY
241: MAGDPTAGKL SGGFLFKMYG LPAAAIWVH SAKPENRAKV GGIMISAALT SFLTGITPEI
301: EFSFMFVAPI LYIIHAILAG LAFFPICLLG MRDGTFSFHG LIDFIVLSGN SKLWLFPIV
361: GIGYAIVYYT IFRVLIKALD LKTPGREDAT EDAKATGTSE MAPALVAAFG KENITNLDA
421: CITRLRVSVV DVSKVDQAGL KKLGAAGVVV AGSGVQAI FG TKSDNLKTEM DEYIRNH
    
```

Figure 2. Protein data for the glucose porter (PtsG) of *Escherichia coli* displayed in the right window of the two-vertical-frame design of the TCDB web application. The six links shown in the upper right hand side of the figure can be accessed directly on screen.

**TCDB**  
UCSD

**Transport Classification Database**  
EXPLORE | SEARCH | ANALYZE | CONTRIBUTE | ABOUT-US

Search:  Search TC#:

**Calculation of Protein Properties**

- [Hydropathy Analysis](#) - Hydropathy analysis using the Kyte-Doolittle scale combined with hydrophobic moment and TMS prediction.
- [EMBOSS Pepwindow](#) - EMBOSS Hydropathy Analysis
- [EMBOSS Hmoment](#) - EMBOSS Hmoment calculates the hydrophobic moment, the hydrophobicity of a peptide measured for a specified angle of rotation per residue.
- [EMBOSS Pepwheel](#) - creates a helical wheel plot that displays a protein sequence looking down the axis of the alpha helix. This view facilitates the identification of amphipathic TMSs.
- [Helixator](#) - creates a helical wheel plot that displays a protein sequence looking down the axis of the alpha helix. This view facilitates the identification of amphipathic TMSs.
- [HMMTOP](#)

**Sequence Similarity Search**

- [BLAST](#) - Search against TC-DB with a protein or nucleotide sequence.
- [SSEARCH](#) - Search against TC-DB using the Smith-Waterman algorithm.

**Pairwise Sequence Alignment**

- [Global Alignment with Displayed TMSs](#)
- [SSEARCH \(Smith-Waterman Local Alignment\)](#)

**Multiple Sequence Alignment Tools and Accessories**

- [Multiple Sequence Alignment with Displayed TMSs](#)
- [cut a multiple sequence alignment in half at specified start and end points for the two halves](#)
- PCMA
- T-Coffee

**Other Sequence Analysis Tools**

- [Biotools](#)

Figure 3. Tools for analyzing transporters.

Superfamilies' on the main website (Figure 1). This information has been integrated with the data in TCDB. Thus, if the user explores the TC hierarchy and inspects a family with known distant relationships to other families, the relationships will be mentioned in the family description.

We have also included a section detailing human transporters that have been approved by the Human Genome Nomenclature Committee. Each of these proteins has been cross-referenced with the TC system. This information can be accessed via the 'Human MTPs' button on the main website (Figure 1). Another new section reports diseases in humans that are associated with human transporters and includes cross-references to the Online Mendelian Inheritance of Man database (OMIM, <http://www.ncbi.nlm.nih.gov/omim/>). The 'MTP Diseases' section (Figure 1) contains these data. The burgeoning number of transporter structures and accessory proteins being sequenced and deposited has led us to catalog known transporter macromolecular structures and cross-reference each structure with its TC number. The 'MTP Structures' link (Figure 1) on the TCDB website provides access to a table listing such information. The presence of 3D structural data for a protein in any given protein family is noted as well in the protein family description in TCDB.

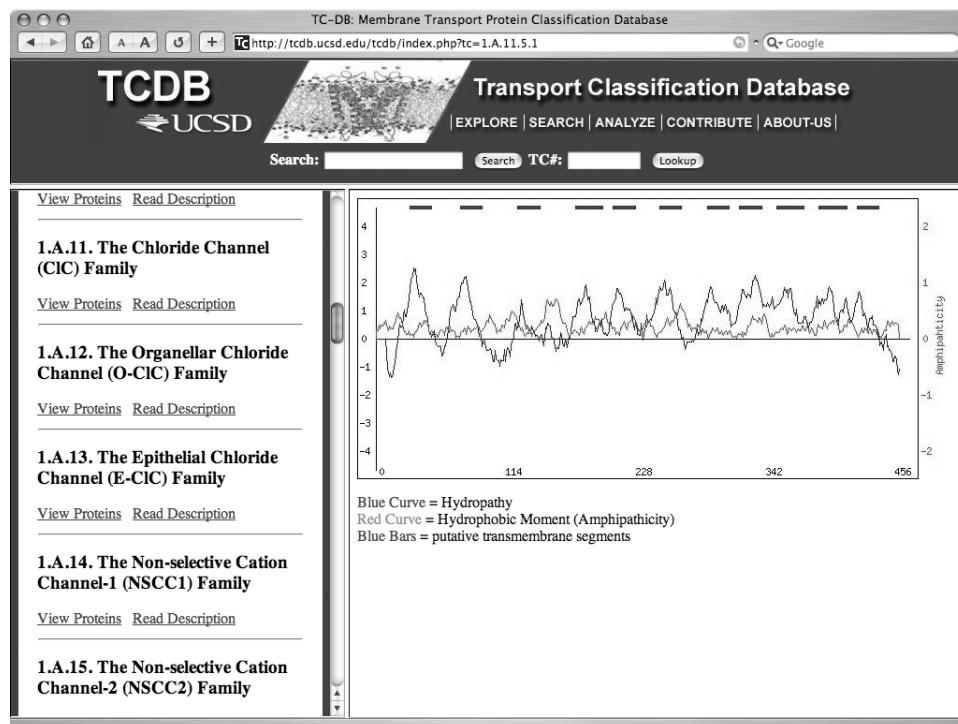
## SEQUENCE ANALYSIS TOOLS

Over the years, we have developed an extensive collection of tools suited to the analysis of transporters. All of these tools

(Figure 3) can be accessed through TCDB by clicking on the 'Analyze' link on the website or by directly visiting <http://www.tcdb.org/analyze.php>. Several tools to analyze transporters are provided such as TMS prediction using HMMTOP 2.0 (20), hydropathy analysis using the Kyte and Doolittle hydropathy scale (21), hydrophobic moment (amphipathicity) analysis (22) using the H moment program from EMBOSS (23) and helical wheel plots using the Pepwheel program from EMBOSS. A protein sequence can be submitted for hydrophobicity and amphipathicity analysis. TMSs predicted by HMMTOP are displayed on the plot (Figure 4). The user may then click on a TMS and view the helical wheel plot of the TMS.

Sequence similarity searches using BLAST (24,25) or SSEARCH (26) are available to search for homologous proteins in TCDB. A quick link to BLAST is provided on the main website as well (Figure 1). A protein or a nucleotide sequence can be submitted to TC-BLAST for a sequence similarity search. The results will specify similar proteins with their TC numbers and numbers of TMSs. The user may then select several sequences and view the multiple sequence alignment and generate a phylogenetic tree by clicking on the 'TC-TREE' button. The user interface displays the multiple sequence alignment with marked predicted TMSs as well as a plot of average hydropathy and conservation. A phylogenetic tree for the sequences can be viewed using ATV (27).

Tools for alignment of two or more sequences are also provided. Two sequences are aligned either locally using SSEARCH or globally using the Needle program from



**Figure 4.** A plot characterizing a transporter with a 12 TMS topology. The plot integrates hydropathy, amphipathicity and TMS prediction. The curves represent hydropathy and amphipathicity of the proteins, and the bars are putative TMSs. Each of the TMSs is a hyperlink that can be traversed to plot a helical wheel representation of the TMS.

EMBOSS. The output of the pairwise global alignment also highlights the TMSs that are predicted using HMMTOP. Multiple sequence alignment with predicted TMSs displayed (28) can also be carried out. A link to additional sequence analysis tools on the Biotools server (<http://saier-144-37.ucsd.edu>) is also provided (Figure 3). Several analytical tools developed in our lab are hosted on this server.

## CONCLUSIONS AND PERSPECTIVES

TCDB is a centralized resource for transporter data and analysis. We are dedicated to bringing data and analytical tools to TCDB users in a timely fashion. Further improvements will include the addition of more analytical sequence tools as well as a bioinformatics process pipeline generator which will enable the user to create work flows for complicated analyses. We will also improve data mining capabilities for analyzing the textual information stored in our database such as PubMed citations and TC family descriptions.

## ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by a grant from The National Institute of Health.

*Conflict of interest statement.* None declared.

## REFERENCES

- Chimento,D.P., Mohanty,A.K., Kadner,R.J. and Wiener,M.C. (2003) Crystallization and initial X-ray diffraction of BtuB, the integral membrane cobalamin transporter of *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 509–511.
- Kurusu,G., Zakharov,S.D., Zhalnina,M.V., Bano,S., Eroukova,V.Y., Rokitskaya,T.I., Antonenko,Y.N., Wiener,M.C. and Cramer,W.A. (2003) The structure of BtuB with bound colicin E3 R-domain implies a translocon. *Nature Struct. Biol.*, **10**, 948–954.
- Yu,E.W., McDermott,G., Zgurskaya,H.I., Nikaido,H. and Koshland,D.E.Jr (2003) Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump. *Science*, **300**, 976–980.
- Abramson,J., Smirnova,I., Kasho,V., Verner,G., Iwata,S. and Kaback,H.R. (2003) The lactose permease of *Escherichia coli*: overall structure, the sugar-binding site and the alternating access model for transport. *FEBS Lett.*, **555**, 96–101.
- Huang,Y., Lemieux,M.J., Song,J., Auer,M. and Wang,D.N. (2003) Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science*, **301**, 616–620.
- Ma,C. and Chang,G. (2004) Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **101**, 2852–2857.
- Chang,G. (2003) Structure of MsbA from *Vibrio cholera*: a multidrug resistance ABC transporter homolog in a closed conformation. *J. Mol. Biol.*, **330**, 419–430.
- Chang,G. and Roth,C.B. (2001) Structure of MsbA from *E.coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science*, **293**, 1793–1800.
- Busch,W. and Saier,M.H.Jr (2002) The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol.*, **37**, 287–337.
- Busch,W. and Saier,M.H.Jr (2003) The IUBMB-endorsed transporter classification system. *Methods Mol. Biol.*, **227**, 21–36.
- Saier,M.H.Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
- Saier,M.H.Jr (2003a) Answering fundamental questions in biology with bioinformatics. *ASM News*, **69**, 175–181.
- Saier,M.H.Jr (2003b) Tracing pathways of transport protein evolution. *Mol. Microbiol.*, **48**, 1145–1156.
- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.*

- (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
15. Bourne, P.E., Address, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W. *et al.* (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
  16. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P. *et al.* (2003) The InterPro Database, 2003, brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
  17. Wain, H.M., Lush, M.J., Ducluzeau, F., Khodiyar, V.K. and Povey, S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.
  18. Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E. *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.
  19. Chang, A.B., Lin, R., Studley, W.K., Tran, C.V. and Saier, M.H.Jr (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Mol. Membr. Biol.*, **21**, 171–181.
  20. Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
  21. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
  22. Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.
  23. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
  24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
  25. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
  26. Pearson, W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
  27. Zmasek, C.M. and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
  28. Zhou, X., Yang, N.M., Tran, C.V., Hvorup, R.N. and Saier, M.H.Jr (2003) Web-based programs for the display and analysis of transmembrane alpha-helices in aligned protein sequences. *J. Mol. Microbiol. Biotechnol.*, **5**, 1–6.