D234–D239 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 5 October 2007*
doi:10.1093/nar/gkm751

TOPDB: topology data bank of transmembrane proteins

Gábor E. Tusnády*, Lajos Kalmár and István Simon

Institute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary

Received August 8, 2007; Revised September 3, 2007; Accepted September 10, 2007

ABSTRACT

The Topology Data Bank of Transmembrane Proteins (TOPDB) is the most complete and comprehensive collection of transmembrane protein datasets containing experimentally derived topology information currently available. It contains information gathered from the literature and from public databases available on the internet for more than a thousand transmembrane proteins. TOPDB collects details of various experiments that were carried out to learn about the topology of particular transmembrane proteins. In addition to experimental data from the literature, an extensive collection of structural data was also compiled from PDB and from PDBTM. Because topology information is often incomplete, for each protein in the database the most probable topology that is consistent with the collected experimental constraints was also calculated using the HMMTOP transmembrane topology prediction algorithm. Each record in TOPDB also contains information on the given protein sequence, name, organism and cross references to various other databases. The web interface of TOPDB includes tools for searching, relational querying and data browsing as well as for visualization. TOPDB is designed to bridge the gap between the number of transmembrane proteins available in sequence databases and the publicly accessible topology information of experimentally or computationally studied transmembrane proteins. TOPDB is available at http://topdb.enzim.hu.

INTRODUCTION

Integral membrane proteins play crucial roles in living cells. They are involved in almost all cellular processes of living organisms such as communication with the outside world, transport of molecules and ions across membranes and energy generation processes. Their vital importance is reflected in their high frequencies (~20–30% of total number of proteins) in various genomes (1–3). Despite these facts, only a few hundreds of transmembrane protein structures have been determined to date. The structure determination of this type of proteins by X-ray crystallography and by NMR techniques is hampered by the difficulties in crystallizing them in an aqueous environment and by their relatively high molecular weight (4).

Because of the bottlenecks in conventional structure determination, numerous biochemical and molecular biology techniques have been developed to investigate the localization of sequence segments relative to the membrane. The full topology of a transmembrane protein defines the membrane spanning and extra/intra cellular segments of a given protein. There are various techniques that enable us to get information about the topology of transmembrane proteins (5), including immunolocalization, molecular biology modifications of proteins, such as inserting/deleting glycosylation sites, making fusion proteins, etc.

Although there are hundreds of articles dealing with experimental determination of topologies according to PubMed, these data have not been collected in a database yet. A well-characterized dataset of integral membrane proteins—containing 320 records—was collected by Moller *et al.* (6), however, a large part of the collected data (about one-third) is based on the analysis of hydrophathy plots and not on experiments. The authors underlined that the interpretation of individual experiments was sometimes difficult and the transmembrane annotation was provided by human experts, considering the results of the hydrophathy plot analysis and experiments. Another collection is the TMPDB dataset (7), containing 302 transmembrane proteins, with experimentally established topology. This dataset includes topology data of 17 beta barrel transmembrane proteins as well. Although the references to PubMed are given for each entry, the experimental details and the method of data processing are not included in the database nor they are described in the article. While the authors of both datasets planed maintenance and further updates, the Moller dataset was not updated, while TMPDB was updated only once, in 2003, but without adding any new entries.

*To whom correspondence should be addressed. Tel: +36 1 466 9276; Fax: +36 1 466 5465; Email: tusi@enzim.hu

Previously, we have developed an algorithm (8,9), called TMDET, to find the most likely orientation of a transmembrane protein in the lipid membranes and to distinguish between transmembrane and non-transmembrane proteins or protein segments using their coordinates only. By scanning all entries in the PDB database with TMDET a new database, PDBTM, has been established (8,10), which collects structures of all transmembrane proteins deposited in the PDB and describes the most likely orientation of proteins relative to the membrane. This orientation can be projected onto the protein sequences giving the most likely position of transmembrane segments in the sequence. It should be noted that the complete topology is not given in PDBTM, namely the location of transmembrane segments without the sidedness information. Using solely the information given in the 3D structures, one cannot assign which part of a structure is inside or outside of the cell or cell compartment. A similar database—OPM (Orientation of Proteins in Membranes)—utilizing a more elaborated biophysical computation method has been created by Lomize *et al.* (11,12). While PDBTM is updated on a weekly basis, the homepage of OPM does not contain information about updates.

In the past three decades, numerous topology predictions have been developed. HMMTOP was the first among these methods that is able to take into account experimental constrains in the prediction (13,14). In many cases the various experiments contradict each other, leading to uncertainties about the topology of a particular protein. These uncertainties can be resolved by collecting topology data and drawing the most likely topology suggested by HMMTOP. This may also verify reliability of various experimental methods.

Here, we present the details of the construction and World Wide Web interface of TOPDB database, which is designed to be the most comprehensive resource of experimental data related to topology of transmembrane proteins as well as of topologies themselves.

DATABASE CREATION

TOPDB was created through three main steps. First, raw experimental data were collected. In the second step, an overall topology was 'predicted' by the HMMTOP method (13,14) using the collected experimental results as constrains. In the last step, a reliability score was calculated that reflects how the experimental data correspond to the final topology and how much of the entire protein sequence is covered by the experiments.

TOPDB integrates data from three sources. One part of the data has been collected directly from the literature. The second part has been derived from PDBTM database, the third part of the data has been obtained from the comparison of PDB (15), PDBTM (8,10) and UniProt (16,17) database.

From the literature, articles containing experiments related to topology of transmembrane proteins have been collected by searching PubMed. The various experimental techniques used in these articles have been categorized into five main classes: (i) fusion experiments, (ii) topology determination by post-translational modifications (e.g. glycosylation), (iii) experiments using proteases, (iv) various techniques using immunolocalization and (v) experiments utilizing chemical modification techniques. The distribution of the occurrences of these experiment types among TOPDB entries is shown in Table 1. Sequential positions used in articles describing topology data were corrected as necessary by using the UniProt/TrEMBL sequences and numbering. Experimental details, like the activity of the reporter enzyme fused to the investigated protein, are also collected and appended to the entry. The lists of various techniques and their description can be found on the manual page of TOPDB.

Table 1. Distribution of experiment types over the TOPDB entries and the total topology data

Experimental type	Entry counts	Topology data counts
Fusion	647	3859
Post-translational modification	31	134
Protease	63	259
Immunolocalization	66	253
Chemical modification	21	167
Structure	820	18 405
Other	22	88
Total	1497	23 162

A detailed description of the various experiment types may be found under the documents section of the TOPDB website.

Topology data derived from PDBTM database (http://pdbtm.enzim.hu) were also incorporated into TOPDB. PDBTM contains the structure of transmembrane proteins with the most likely orientation to the membrane. This results in two membrane planes, which cut the structures into three parts: 'side one', membrane embedded and 'side two' parts. These spatial localizations are projected onto the sequences. Because the 3D structures of transmembrane proteins do not contain topological information for sidedness, we checked the original article of each PDBTM entry for the proper localization of 'side one' and 'side two' parts. This evidence is given in <SideDefinition> section of the entries. Most PDBTM entries cannot be directly mapped to TOPDB entries, because PDBTM entries often contain oligomer structures, while a TOPDB entry stores data for one polypeptide chain only. Moreover, in many cases the same polypeptide chain can be found in various PDB files (and therefore in various PDBTM files), which are collected in a single TOPDB entry.

A large number of structures in the PDB correspond to the soluble fragment of a transmembrane protein. These cases also contain information about the topology in an indirect way. We have collected all these structures from the PDB, by comparing each sequence of PDB entries defined as not transmembrane protein in PDBTM, with the corresponding UniProt entry. If a protein with a globular fragment in the PDB was annotated as transmembrane protein in the UniProt database, we

entered it into TOPDB. Evidences of the exact localization of these structures have also been extracted from the articles, where the structure was published. These data were regarded as additional experimental data, specifying the inside/outside localization of a given part of a transmembrane protein

From the 1899 articles processed until now, 188 entries have been collected, which contain only experimental topology data, 346 entries have been derived from PDBTM database and we found 474 proteins, where a non-transmembrane fragment of a transmembrane protein have been crystallized. The current holdings and various additional statistics can be found under the Statistics menu on the TOPDB home page.

It should be noted that, while topology data derived from PDBTM and from the comparison of PDB and UniProt entries can be obtained semi-automatically, literature searching requires careful interpretation of the experimental data by human experts. This explains why the number of topology data derived from the literature is less than the semi-automatically generated data (see Table 1).

TOPOLOGY GENERATION

After collecting raw topological data, we applied a fully automatic procedure to generate the most probable topology of each protein in the database. First, all similar sequences have been identified in TOPDB by BLAST (18) using an e-value $1E{-}10$. Topological data of similar entries have been projected on each entry according to the BLAST alignments. Then, HMMTOP transmembrane topology prediction method (13,14) was applied using the collected topology data as constrains. Following the logic of the Baum–Welch algorithm the experimental constrains may be naturally incorporated into the calculation. That is, they were applied by the product of forward and backward probabilities for a state and conditional probabilities of the experimental results on the condition that position is inside, membrane or outside, as described previously (14). The conditional probabilities are close to one if the experiment agrees with the state of the hidden Markov Model and it is a small and constant error otherwise. Two new architectures created for HMMTOP (Tusnady and Simon, in preparation) enables us to predict topologies of alpha-helical transmembrane proteins with re-entrant loop as well as the topologies of beta-barrel proteins. We have calculated a reliability index of each predicted topology over the sequence, as the sum of the number of experiments which agree minus those which do not agree with the predicted topology, normalized by the length of the sequence. The lower limit for reliability index is 20%, while the upper one is 90% if all experiments correspond to one side of the given protein only.

DATABASE ORGANIZATION, FILE FORMATS AND ANNOTATIONS

Each TOPDB entry has a unique identifier composed of two characters and five digits. The two characters give a

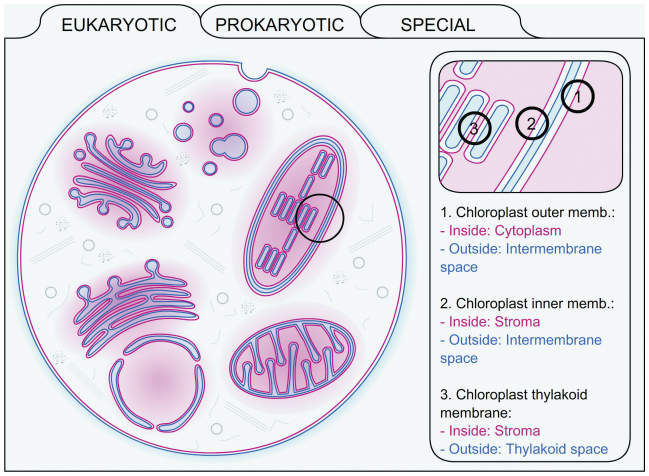

Figure 1. A representation of the interactive flash movie found under the documents section of the TOPDB website. It serves to depict eukaryotic, prokaryotic and special membrane types. The animation assists the determination of interior and exterior membrane faces. In the given image the chloroplast membranes can be seen.

crude characterization of entries: 'AB' for alpha helical bitopic proteins, 'AP' for alpha helical polytopic proteins and 'BP for' beta barrel transmembrane proteins. The primary file format of TOPDB is XML (Extensible Markup Language), validated by the corresponding Schema definition, which can be found on the TOPDB homepage (http://topdb.enzim.hu/TOPDB.xsd). The main nodes in the XML files are the Name, Organism, CrossRef, Sequence, Membrane, Topology, Experiments and References nodes. Name and Organism sections describe the protein name and source organism. This information is obtained from the UniProt database (16,17), if the given sequence was found in the SwissProt or TrEMBL database. In the CrossRef section, we collected all cross references of entries related to UniProt and PDB databases. When PDB structures are referenced the corresponding structure determination method, resolution, chain identifier and the PDB sequence transformation are also given (<Fragment Begin = 'from' End = 'to'>). Sequence node contains the amino acid sequence of the given entry itself with an md5 checksum, and information about the amino acid sequence of the matured protein (<Cleavable Begin = 'from' End = 'to' Type = 'Signal | Transit | Propep'>). These data are also derived from UniProt. The next node, Membrane, defines the membrane type where the given protein resides, as well as the appropriate localization of the inside and outside spaces. We created an interactive flash movie about the membranes used in TOPDB entries (and UniProt) to help understanding the localization and facing of various membranes in the eukaryotic and prokaryotic cells (Figure 1). Topology node contains the final predicted topology by the HMMTOP prediction algorithm, using all experimental data as constrains (see previous section). The next section is the list of all collected and derived topology data, describing the source of the data together with the experiment types and the details of the experiments (e.g. the normalized activity of alkaline-phoshphatase of

a fusion product). The full reference of experiment types can be found on the manual page in the TOPDB homepage. Finally, the References node lists the PubMed identifiers with the publication source related to the experiment data stored in the database entry.

Entries in TOPDB database are stored in a Subversion repository, allowing the tracing of modifications as well as the downloading of updates. The most up-to-date version of the database is also accessible under the download menu of the homepage in XML and traditional formats. A history tab for each entry is also provided (see next section), where the versions of an entry can be compared directly.

DATA ACCESS AND VISUALIZATION

TOPDB database is available at http://topdb.enzim.hu. There are various ways to access information related to a protein in TOPDB database. The simplest way is to use the quick search form on the header part on TOPDB web pages. TOPDB, PDB and UniProt identifier or keyword can be submitted to the quick search. If a user would like to restrict the search to a specific experiment, organism or structural type, a detailed search form is also available. If the search results in several entries, the server sends back a brief, selectable list of the matched entries, while all data are visualized instantly in the case of a single hit.

Although XML format is useful for computer programming, it is hard to understand the raw data. Therefore, numerous possibilities are provided to help the interpretation of data. Various visualization modes can be selected by clicking the appropriate tab on the Result page. Users can choose between graphical representations (flash movie or image) or text representations (HTML or XML). The flash movie shows the topology model of proteins with clickable sequence parts. By selecting a given region, the experiments related to the specified region appear in a new pop-up window. In the graphical view, experiments are grouped by references, and are mapped on the protein sequence, which is represented by a horizontal line. A colour code helps to grasp the image at a glance (Figure 2). Moving the mouse over a coloured box on the image opens a small window containing the experimental details and links to related references. In HTML view, data are shown in a tabular form, while choosing XML tab, the raw XML file is shown. On an additional tab, the history of the given entry can be traced back, while by clicking the download tab, the selected entry can be downloaded separately (in XML format).

COMPARISON WITH OTHER DATASETS AND DATABASES

We have compared a reliable subset of our topology data in TOPDB of α-helical transmembrane proteins, with reliability score >80%, to data in four other databases and datasets including UniProt (16,17), the training set of TMHMM algorithm (2), the Moller dataset (6) and TMPDB database (7). The results of the comparison are summarized in Table 2.

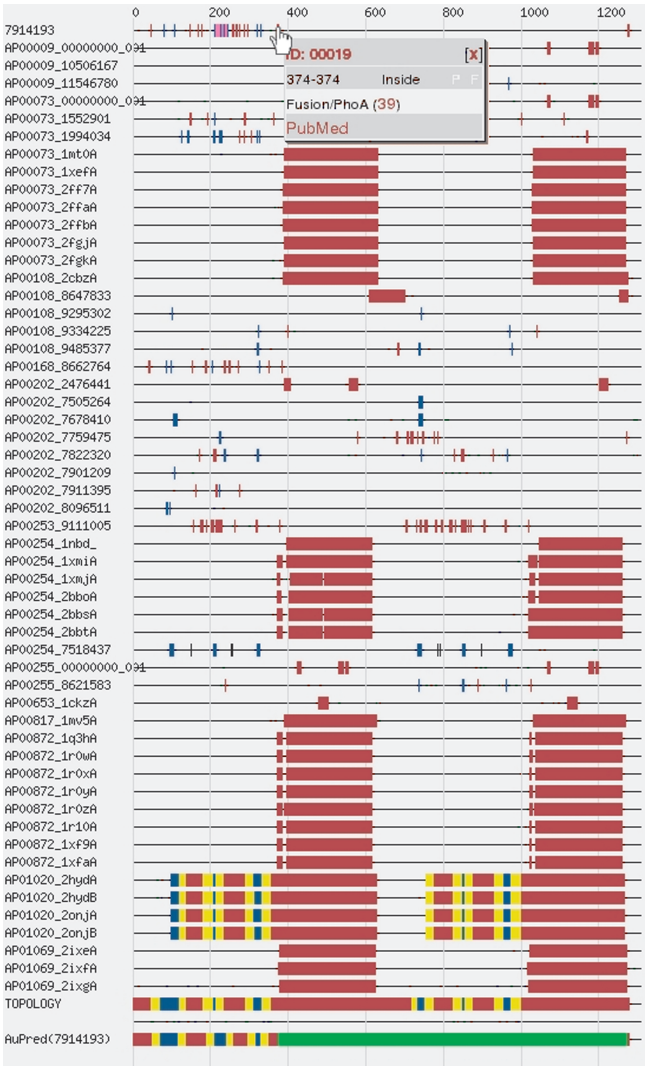

Figure 2. The graph shows the topology of TOPDB entry of mouse multidrug resistance protein 1 (AP00199). The distinct experiments have unique identifiers (e.g. PDB ID, PUBMED ID, etc.), and if the topology data are derived from homologous proteins the identifier is extended with the TOPDB ID. The graph is colour coded to show protein segment localization: membrane interior (red), membrane exterior (blue) and transmembrane (yellow). In the interactive view, additional information can be obtained by rolling the mouse over the individual bars, e.g. experiment type, residue position and cross references.

Total 869 entries of UniProt database are also present in TOPDB. Out of the 869 database records, 57 (6%) contain neither information about the position of transmembrane helices, nor the localization of extramembranous parts of the proteins in the UniProt database. One hundred and seventy-two records (20%) out of the 869 do not contain the orientation of the proteins in UniProt, just the position and the numbers of the transmembrane helices. For the remaining 697 entries, 83 ones do not have the same topology in TOPDB and in UniProt. By carefully investigating these individual entries, we have found that in some cases the available topology data do not enable us to decide which topology definition is correct, while in other cases

Table 2. Comparison of TOPDB topology data with other datasets

Name	N_{entry}	TMH_{TOPDB}	TMH_X	TMH_{same}	$AllTMH_{same}$	TOP_{same}
UniProt	696	2655	2669 (97%)	2579 (97%)	619 (89%)	614 (88%)
TMHMM	63	285	284 (99%)	282 (99%)	60 (95%)	60 (95%)
Moller	85	366	371 (98%)	360 (98%)	80 (94%)	79 (93%)
TMPDB	154	796	809 (97%)	779 (98%)	154 (90%)	125 (81%)

Abbreviations: N_{entry}: number of entries with larger than 80% reliability in TOPDB and were also found in the dataset/database compared; TMH_{TOPDB}: number of transmembrane helices of the common entries in TOPDB; TMH_x: number of transmembrane helices in the dataset/database compared; TMH_{same}: number of transmembrane helices that have the same sequential position in TOPDB and the database/dataset compared; $AllTMH_{same}$: number of entries where all transmembrane helices positions are the same in; and TOP_{same}: number of entries whose topologies are the same in TOPDB and in the dataset/database compared.

uncleavable/cleavable signals, signal/transit or transmembrane/membrane loops are swapped. However, for the largest part of the discrepancies the topology definitions in the UniProt contradicted experimental data.

There are only 63 entries common in TMHMM training set and TOPDB. The low percentage of the identity might be explained by two reasons. First, we chose only such TOPDB entries, which have reliability index above 80%, on the other hand TOPDB contains entries that have their own topology data. Comparing these 63 entries we found different topology definitions in three cases only (AP00183, AP00201 and AB00728). In all these cases, the TMHMM definitions are inconsistent with the available experimental data.

Comparison of Moller dataset and TOPDB gave similar results as the comparison of TMHMM's training set and TOPDB. There are only 85 entries common between the two dataset, and 6 entries among them have different topologies. In one out of the six cases, we cannot decide which definition is correct, the available topology data allow both definitions. In the remaining five cases, the topology data in the Moller dataset do not comply with the available experimental data.

The largest differences can be seen between TMPDB and TOPDB (Table 2). Apart from 22 mitochondrial inner membrane proteins in TMPDB database with inverted In and Out definitions, there are 29 entries out of the 154 shared entries, which have different topology definitions in the two databases. In six cases there are not enough data to decide which definition is correct, but in twenty-three cases the topology data in TMPDB are in conflict with the available experimental data.

COMPARISON OF EXPERIMENT TYPES

Using the 3D structure as the most reliable source of the topology data, we performed a comparison between the different topologies with respect to various experiment types. In general, the results of various experiments are in good agreement with the 3D structures, but we have found two common pitfalls. One of them emerges in the case of utilizing reporter enzymes as fusion proteins. In cases, when the N-terminus is outside, but there was no transmembrane helices prior to the fusion point, the reporter enzyme could not be transferred outside, and remained inactive. Moreover, if the fusion points were after the first transmembrane helix, the topologies were frequently inverted. A typical example of this experimental error can be found in the case of the iron–sulfur subunit of formate dehydrogenase-O (19) (AB00057). The other common error is the misinterpretation of the lack of immunoglobulin binding to extra inserted or endogen epitopes localized in the cytosol. In these cases, the binding can be seen only after membrane permeabilization by using special detergents. However, if the extra-cytosolic epitope is shaded, i.e. the antibody cannot bind due to structural reasons, the epitope will be accessible after the use of detergent, which in turn results in binding.

FUTURE DIRECTION

Topology data based on structures deposited into PDB database can be updated semi-automatically following the weekly update of PDB and PDBTM databases. However, the update of other topology data collected from literature is difficult to carry out automatically. This would require the work of numerous annotators, similarly to the maintenance of SwissProt database. Alternatively, the help of the community of transmembrane protein scientists could be utilized in a way similar to wikipedia systems. We plan to change the web interface of TOPDB to enable scientists to add and/or maintain their own experimental data by turning it into a Web2-like platform.

A further way to expand the database is to transfer topology annotation for sequences showing sequence similarity to an existing entry in TOPDB, but lacking direct experimental data. We also plan to give a classification system of transmembrane proteins in the database. Naturally, we are open for any users' advice as well.

ACKNOWLEDGEMENTS

Thanks to Zs. Dosztányi and A. Fiser for their suggestions and editing the manuscript. This work was supported by grants from Hungarian research and development funds: OTKA T049073, K61684, NI68950; NKFP MediChem2 1/A/005/2004; GVOP-3.1.1-2004-05-0143/3.0 and GVOP-3.1.1-2004-05-0195/3.0. The Öveges fellowship for I.S. and the Bolyai János Scholarship for G.E.T. are also gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by OTKA.

Conflict of interest statement. None declared.

REFERENCES

1. Jones,D.T. (1998) Do transmembrane protein superfolds exist? *FEBS Lett*., **423**, 281–285.
2. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
3. Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*., **7**, 1029–1038.
4. Arora,A. and Tamm,L.K. (2001) Biophysical approaches to membrane protein structure determination. *Curr. Opin. Struct. Biol*., **11**, 540–547.
5. van Geest,M. and Lolkema,J.S. (2000) Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiol. Mol. Biol. Rev*., **64**, 13–33.
6. Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
7. Ikeda,M., Arai,M., Okuno,T. and Shimizu,T. (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res*., **31**, 406–409.
8. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, **20**, 2964–2972.
9. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.
10. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*., **33**, D275–D278.
11. Lomize,A.L., Pogozheva,I.D., Lomize,M.A. and Mosberg,H.I. (2006) Positioning of proteins in membranes: a computational approach. *Protein Sci*., **15**, 1318–1333.
12. Lomize,M.A., Lomize,A.L., Pogozheva,I.D. and Mosberg,H.I. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.
13. Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol*., **283**, 489–506.
14. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
16. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
17. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
18. 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.
19. Benoit,S., Abaibou,H. and Mandrand-Berthelot,M.A. (1998) Topological analysis of the aerobic membrane-bound formate dehydrogenase of Escherichia coli. *J. Bacteriol*., **180**, 6625–6634.