Bioinformatics Advance Access published June 4, 2004

Transmembrane proteins in Protein Data Bank: identification and classification

Gábor E. Tusnády, Zsuzsanna Dosztányi and István Simon†

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

Running head: PDB_TM: transmembrane proteins in PDB

†Correspondence : Institute of Enzymology, BRC, Hungarian Academy of Sciences, H-1518 Budapest, P.O. Box 7., HUNGARY; Tel: (36-1) 466-9276, Fax: (36-1) 466-5465, e-mail: simon@enzim.hu

Abstract

Motivation: Integral membrane proteins play important roles in living cells. Although these proteins are estimated to constitute around 25% of proteins at a genomic scale, the Protein Data Bank (PDB) contains only a few hundred membrane proteins due to the difficulties with experimental techniques. The presence of transmembrane proteins in the structure data bank, however, is quite invisible, as the annotation of these entries is rather poor. Even if a protein is identified as a transmembrane one, the possible location of the lipid bilayer is not indicated in the PDB because these proteins are crystallized without their natural lipid bilayer, and currently no method is publicly available to detect the possible membrane plane using the atomic coordinates of membrane proteins.

Results: Here we present a new geometrical approach to distinguish between transmembrane and globular proteins using structural information only and to locate the most likely position of the lipid bilayer. An automated algorithm (TMDET) is given to determine the membrane planes relative to the position of atomic coordinates, together with a discrimination function which is able to separate transmembrane and globular proteins even in cases of low resolution or incomplete structures such as fragments or parts of large multi chain complexes. This method can be used for the proper annotation of protein structures containing transmembrane segments and paves the way to an up-to-date database containing the structure of all known transmembrane proteins and fragments (PDB_TM) which can be automatically updated. The algorithm is equally important for the purpose of constructing databases purely of globular proteins.

Availability: The PDB_TM database is available for academic users on {{http://www.enzim.hu/PDB_TM}}.

Contact: tusi@enzim.hu, zsuzsa@enzim.hu, simon@enzim.hu

Supplementary Information: Data files used in this article can be found under the PDB_TM homepage ({{http://www.enzim.hu/PDB_TM/index.php?method=docs}}).

Introduction

Integral membrane proteins form about 20-30% of all protein sequences (Jones, 1998; Wallin and von Heijne, 1998; Krogh *et al.*, 2001). They are of vital importance for living cells, playing role in communication and in transport between the cells and the outside world. Beside the obvious academic research interest, they are also targets of many pharmaceutical developments, as the overwhelming majority of the drugs used in human and veterinarian medicine act on this kind of proteins. Despite their great importance, transmembrane proteins (TMPs)[†] are highly underrepresented in the protein structure database, due to difficulties in crystallizing them in an aqueous environment. TMPs are usually larger than globular[§] proteins, making their structure determination quite difficult by NMR technique as well (Arora and Tamm, 2001). This gives an explanation for their relatively low occurrence among the more than 20000 structures deposited into the Proteins Data Bank (PDB) so far (Berman *et al.*, 2000). Currently, more than 300 membrane protein structure files can be found in PDB, representing around 30-40 different folds. The size of this subset is approaching the level, where an automatic procedure is required to construct and maintain a database specific for TMPs.

Although solving the structure of a membrane protein is still regarded as a major achievement, one vital component, the membrane itself is missing from these structures. For structure determination they are taken out from the lipid bilayer, and crystallized by masking their exposed hydrophobic parts by amphiphilic detergents, so that the protein-detergent complex can be treated similarly to soluble proteins (Ostermeier and Michel, 1997). The detergent molecules are highly unstructured and are usually not visible in the X-ray picture. With the exception of a few tightly bound lipid or detergent molecules, the deposited experimental data have no direct indication that the protein is immersed into the membrane under native conditions, and do not contain information about the exact location of the lipid bilayer (Lee, 2003). The topology of transmembrane proteins can be predicted form the sequence alone with relatively high accuracy, however, transmembrane prediction methods are not adequate for distuingishing TMPs

[†]Abbreviations used: TMP: transmembrane protein; PDB: Protein Data Bank; PDB_TM: Protein Data Bank of Transmembrane Proteins

[§]Note: we use globular and water soluble as a synonym of non-transmembrane through the article.

form globular, water soluble proteins, as most of them identify at least one false „transmembrane segment" in more than a quarter of the globular proteins (Tompa *et al.*, 2001). When the 3D structures of TMPs are available, these provide the most reliable source of information for benchmarking membrane topology prediction methods and form the basis of comparing and analyzing different TMP structures.

Currently, there is no exact algorithm available to identify transmembrane proteins and to determine the membrane location using the protein 3D structure as an input (Tusnády and Simon, 2001). The hydrophobic membrane surrounding TMPs and the aqueous environment of globular proteins are drastically different, thus distinguishing TMPs from globular ones should be straightforward. There are several reasons that the separation between the two groups is not unequivocal. The surface of globular proteins is usually not entirely hydrophilic as apolar atoms of polar residues may be exposed and larger hydrophobic patches involved in ligand binding can also be found on the surface. Analogously, the membrane embedded parts of a TMP may contain polar and charged residues playing role in enzymatic activity or ion transport. While the surface of transmembrane and globular proteins are adopted to their different environment, the inside of the two groups is commensurable in their hydrophobicity (Rees *et al.*, 1989). This can make distinguishing short fragments located inside an intact globular or transmembrane protein quite difficult. Similar problems can occur in the case of large multi-chain complexes. The interface of globular oligomers is often hydrophobic, while in the case of transmembrane chains it is more likely to be polar. Thus, when considering individual chains without the valid quaternary structure, the difference between the surface compositions can also diminish.

Another factor influencing the discrimination of transmembrane and globular proteins is related to the quality of the structure. The crystal structure of several membrane proteins is of low resolution, often reflected in distorted secondary structures with incomplete hydrogen bond network or a structure with C_α atoms only. Determination of the structure by NMR in a detergent solvent instead of the lipid bilayer can result in a highly flexible structure with the structural boundaries of the membrane regions melted. As a result of all these factors, the objective function aiming to distinguish between transmembrane and globular proteins should

not only account for the physical difference in their environments, but it should also incorporate practical limitations associated with the structure determination.

PDB, the data bank which originally contained only soluble proteins, does not clearly distinguish TMPs form globular ones. Entries which correspond to soluble fragments of a transmembrane or membrane-anchored protein, like MHC, are often annotated as membrane proteins. Also, structure files of true TMPs can completely lack the specific annotation regarding its membrane character. Although the description of these proteins usually implies that they are TMPs, this information is difficult to extract automatically. The annotation of some bacterial toxins, which are able to immerse into the lipid bilayer under appropriate conditions, can also be misleading if their structures correspond to the water soluble form. As a result, the HEADER, TITLE, COMPND or REMARK records in a PDB file do not faithfully describe whether the given file contains transmembrane or globular proteins. Instead, TMPs can be identified only by analyzing the structure coordinates.

Some collections of TMPs are already available. The aim of these databases was to collect all experimental information regarding TMPs (Möller *et al.*, 2000; Jayasinghe *et al.*, 2001; Ikeda *et al.*, 2003), hence they are not specific to proteins with known structures. Common to all of these databases is that they are constructed manually, relying on an expert to continuously follow the release of novel transmembrane structures.

These difficulties necessitate the development of a new automated algorithm to distinguish transmembrane and globular proteins by their atomic coordinates as well as to identify the transmembrane segments of TMPs using only their atomic coordinates. To this end, we developed an algorithm called TMDET and collected the newly determined transmembrane segments in a database called Protein Data Bank of Transmembrane Proteins (PDB_TM). The automated TMDET program ensures that PDB_TM will be automatically updated. The PDB_TM database is public for academic usage at the web site: `http://www.enzim.hu/PDB_TM`. Upon scanning the entire structure data base several structure discrepancies have been discovered, which are discussed as well.

Methods

The TMDET algorithm processes the input coordinate data file as follows: First, it investigates the protein and chain types and omits virus and pilus proteins as well as entries containing nucleotides (RNA or DNA sequences). Entries with less than 15 standard amino acid residues are also ignored. Low-resolution structures containing chains with C_α or backbone atoms only are handled separately (see below). The next step is the construction of the possible biological oligomer structure. This step is divided into two parts, first the algorithm builds up the biomolecule, if the BIOMOLECULE record is given in the input PDB file, secondly it investigates the oligomer structure to eliminate chains which form non-biological contacts as a result of the crystallographic process. Next, the membrane-exposed "water accessible area" is calculated. The core of the algorithm is the search for the most probable position of membrane planes relative to the given coordinates, by measuring the fitness of membrane localization via an objective function. The protein classification is made according to the best value of the objective function (called Q-value). The Q-value is composed of a measure of hydrophobicity and a structural part. These steps are described in details in the following sections. The TMDET algorithm is written in standard C language. A typical run including the calculation of water accessible surface as well as the search for the most probable membrane plane, takes a few seconds on a Pentium 4, 2.4 GHz personal computer.

Construction of biological molecule

The biological molecule, i.e. the macromolecule that has been shown, or is believed, to be functional, is constructed by using the matrix operations described in the BIOMOLECULE records. In a few cases, however, we found non-biological contacts as a results of crystallization artifacts. The superfluous chains are detected and omitted from further analysis in the following cases: i) if there are no interactions between clusters of identical chains, ii) if they occupy the same position (i.e. superposition of two molecules, or the symmetry operation is not correct), or iii) if chains with the same sequence are not related by a simple rotation but by an additional translation.

The advantage of analyzing the internal symmetry between different subunits is that it can directly lead to finding the membrane axis, since the rotational axis has to be parallel with the membrane normal in the case of the transmembrane chains. Therefore, if the Q-value of the objective function (see below) by using the given rotational axis as the membrane normal exceeds a certain threshold, the algorithm yielded the rotational axis as the membrane normal. In the opposite case, a non-redundant cluster of non-identical chains is tested again for the most likely position of the membrane.

Calculating the membrane-exposed "water accessible surface area"

"Water accessible surface area" is calculated according to the algorithm of Lee and Richards (Lee and Richards, 1971) (We use the "water accessible surface area" idiom, even if we know, that these protein surfaces are in fact membrane-exposed). In order to improve the membrane detection algorithm, the "water accessible surface" is considered only for those atoms which could potentially interact with the lipid bilayer. These membrane-exposed atoms are selected by the following approximate filtering procedure: the protein is cut into 1 Å wide slices along a predefined axis, and around each slice of atoms, test points are placed on a rectangle which embeds all the atoms within that slice. Those atoms lying closest to any of the test points are defined to be on the outside (i.e. possible membrane exposed) of the surface. For all other atoms the "water accessible surface area" is set to zero. As seen, the filtering algorithm depends on the given axis, therefore, if a likely membrane axis had been found, this procedure was iterated until convergence.

Definition of the objective function

The heart of the TMDET algorithm is the objective function, which measures the fitness of a given membrane position to the protein. The objective function depends only on the membrane position and direction, which is defined by the membrane's normal vector. The protein is cut into 1 Å wide slices along this normal vector. In each slice the membrane-exposed "water accessible surface area" of hydrophobic and hydrophilic residues are summed separately. We

use the simplest "hydrophobicity scale" of residues, by dividing residues into hydrophobic (F, G, I, L, M, V, W and Y) and hydrophilic residues (A, C, D, E, H, K, N, P, Q, R, S and T), because the use of various hydrophobicity scales did not improve the procedure. Alanine is handled as a polar residue, as it has a weak but distinct preference for the interface region (Nilsson *et al.*, 2003), and it occurs frequently in soluble polypeptides, for example in the so called anti-freeze proteins (Davies and Hew, 1990). The hydrophobic factor of the objective function is defined as the relative hydrophobic membrane-exposed surface area (hydrophobic area divided by all surface area).

A structure factor is also incorporated into the objective function. It is defined as a product of three factors, the straightness, turn and end-chain factor. The definition of these factors are as follows: for the straightness factor, the i^{th} residue in a protein chain is part of a straight triplet if the projection of the C_α atoms of the previous third (i-3) residue, itself (i) and the next third (i+3) residue onto a predefined vector (the normal vector of membrane planes, see below) are in a monotone decreasing or increasing order. The straightness factor is defined as the relative frequency of "straight" residues in a given protein slice.

The turn factor is defined as one minus the relative frequency of "turn" residues in a given slice. Turn triplets have a similar definition as the straight triplets: the i^{th} residue in a protein chain is the center of a turn if the projection of the C_α atoms of the previous third (i-3) residue, itself (i) and the next third (i+3) residue onto the predefined vector are not in a monotone decreasing or increasing order.

The end-chain factor is one minus the relative frequency of chain end residues in a given slice.

The return value of the objective function (called *Q*-value) is the average of the products of the hydrophobic factor and the structure factor in each slice over a predefined width (i.e. a predefined number of slices). For proteins containing chains with only C_α or backbone atoms, the relative hydrophobic surface area can not be calculated, and this part of the objective function is replaced by measuring the relative hydrophobicity of residues using the hydrophobicity scale of von Heijne (von Heijne, 1992).

Searching the best membrane plane

The simplest scenario can occur when the analysis of internal symmetry between different chains yielded a rotational axis. If there was a 15 Å slice along this axis for which the *Q*-value exceeded the predefined threshold, then it was accepted as the normal vector of the membrane plane. In all other cases an exhaustive search was needed to find the most probable orientation of the membrane. Possible membrane normals were sampled as unit vectors pointing to test points placed equidistantly on the surface of a ball and the best *Q*-value was searched in each direction by calculating the *Q*-value of 15 Å wide slices moving along the given axis. The vector with highest *Q*-value obtained during the rotation gives the normal vector of the best membrane planes. Then the width of the membrane is broadened as much as possible in such a way that the number of crossing segments do not change. For proteins containing C_α or backbone atoms only, 22 Å wide slices were used in the calculation.

Classification of the proteins

The classification of a given protein was based on the calculated best *Q*-value. If the best *Q*-value was below a predefined lower selection limit, it was classified as globular protein. If the best *Q*-value was below the lower selection limit and the corresponding Swissprot (Boeckmann *et al.*, 2003) entry contained "TRANSMEM" FT line(s), then the protein was classified as the globular fragment of a transmembrane protein. If the best *Q*-value was above of a predefined upper selection limit, it was classified as transmembrane protein. Transmembrane proteins were further classified as transmembrane alpha, beta and coil proteins based on the dominant secondary structure of their membrane spanning segments determined by the DSSP algorithm (Kabsch and Sander, 1983). Between the lower and upper selection limit the decision was made manually. We have to emphasize, that only a very small percent of proteins (cca 2%) falls between the lower and upper selection limit (see Figure 1).

Databases used

The databases used were the following: i, the MPtopo dataset (Jayasinghe *et al.*, 2001) contains 46 polytopic membrane proteins of known 3D structure; ii, the White dataset, a large collection found on the web (White and Wimley, 1999), which comprises 95 structures; iii, the Möller dataset (Möller *et al.*, 2000) containing 55 transmembrane proteins of known 3D structure; iv, a selection of TMPDB database (Ikeda *et al.*, 2003) contains 208 entries; v, the "Membrane and cell surface proteins and peptides" class of the SCOP database (Murzin *et al.*, 1995) comprises 171 proteins. These five datasets were collected on July, 2003 and altogether contained 254 unique PDB entries. The list of these entries as well as the filtering of these databases are described in the PDB_TM home page (`http://www.enzim.hu/PDB_TM/docs/database_used.html`).

A further dataset comprising a globular selection of the PDB database were also used in the TMDET validation. This set were generated from the PDB select database (Hobohm and Sander, 1994) by eliminating short, pilus, viral and transmembrane (!) proteins, as well as nucleotide containing proteins. After removing these entries 489 proteins remained, which were used as the globular test set in the validation process.

Results

Validating the TMDET algorithm

The TMDET algorithm described in the Methods section in details is the first algorithm in the literature, which takes the atomic co-ordinates of a membrane protein and calculates the most probable localization of the lipid bilayer. The algorithm can be used for making a selection between transmembrane and globular proteins as well. As a test, it was applied on the selected 489 globular and 254 transmembrane proteins and the calculated Q-values vs frequencies are shown on Figure 1. These distributions can be used to assign how likely is a protein to be TMP or globular at a given Q-value, by taking the partial integrals of the two curves. The Q-values are clearly separated for transmembrane and globular proteins. Generally, there are

only very few entries in the overlapping region, 9 (1.8%) and 3 entries have Q-values larger than 40.0 (lower limit) and 46.0 (upper limit), respectively, in the globular set and only one TMP has a Q-value lower than 40.0. The accuracy of the selection of TMDET algorithm on these datasets is 98.7% overall. The largest Q-value in the globular protein set is 62.5 for the protein apolipoprotein A-II (1l6k), which is a lipid binding protein (Kumar *et al.*, 2002). A voltage-dependent potassium channel, complexed with an Fab (1orq) has the lowest Q-value (33.8) among TMPs, while the second lowest Q-value is 41.3, showing that 1orq has a unique structure (see Discussion too) (Jiang *et al.*, 2003). These data show the high selectivity power of TMDET algorithm making it possible to use TMDET to find all the TMPs in the PDB database.

Scanning the PDB database

After validating the TMDET algorithm, the entire PDB database was scanned by the algorithm to find all the TMPs. To ensure that all TMPs will be found, even at the expense of collecting some more globular ones, the lower Q-value limit was lowered (from 40.0 to 38.0). After scanning the 22178 PDB entries, the program collected 472 proteins having Q-value greater than 38.0. All of these 472 proteins were visually checked by using PyMOL molecular visualization program (DeLano, 2003), as well as the PDB headers and the corresponding Swissprot entries were carefully investigated. If it was necessary, we checked the literature as well. Finally, 148 proteins were eliminated and 324 proteins were classified as TMPs (226 alpha, 73 beta, 9 unstructured and 16 low-resolution proteins). These 324 TMPs cover 1673 protein chains, 1021 chains contain real transmembrane segments and 652 chains are globular parts of transmembrane protein complexes. The transmembrane segment distribution of the transmembrane polypeptide chains can be seen on Table 1. The numbers of β-barrel proteins containing two and four transmembrane segments are somewhat artifacts, as these chains build up a 14 (7*2) and a 12 (3*4) stranded β-barrel. The high value of the 372 for one helical transmembrane segment containing protein chains is also an artifact, because 89 of them are fragments of polytopic TMPs.

Comparing transmembrane databases

We collected transmembrane proteins of known 3D structure from five transmembrane protein databases published earlier for the sake of validating the TMDET algorithm and to make some comparisons. Some of these databases contain TMPs, whose topology is confirmed by experiments, and therefore these contain sequential data rather than structural ones. Thus, these databases had to be filtered for proteins, which have atomic coordinates in the PDB database. The whole filtering process is described in the PDB_TM home page in details and briefly in the Methods section. There are only 9 proteins, which can be found all the five transmembrane protein databases published earlier. The MPtopo and Möller databases do not contain any protein which has not been found in other databases, while the SCOP, White and TMPDB databases cover 11, 20, and 28 unique proteins, respectively. There are 70 proteins among the newly collected TMPs, which can not be found in any of the five databases. Most of them were deposited to the PDB database after the creation of the five databases, but the White dataset and SCOP database should cover these entries, as these databases have been recently updated.

The PDB_TM database

The results for all proteins in the PDB (transmembrane and non transmembrane) were collected in a database, called the PDB_TM database. The database contains the classification of each proteins as well as the localization of membrane planes in the molecular coordinate system and the localization of transmembrane segments in the sequence, both determined by the TMDET algorithm. Because the extracellular and intracellular side of proteins can not be determined from its coordinates, we use side-one and side-two notation to distinguish between the two sides of proteins. We made the database search-able and public for academic users (`http://www.enzim.hu/PDB_TM`).

The transmembrane proteins are grouped into structural families based on pairwise alignment. More remote homologous were detected by PSI-BLAST (Altschul *et al.*, 1997). Single transmembrane helices, as well as proteins without proper sequence assignment were not classified. Alpha helical structures can be grouped into 29 distinct structural families, and there are

10 kind of beta-barrel structures (Table 2).

Discussion

Although experimental studies of membrane proteins usually reveal many important structural properties, they do not give the exact location of the membrane, because the atomic coordinates of the lipid molecules are not determined. In most cases, transmembrane regions have significantly different characteristics from soluble proteins, like a band with high hydrophobicity, ordered secondary structures arranged in a roughly parallel fashion or folds specific to membrane proteins. These features usually give a clear indication of the approximate position of the membrane. Some structures contain a few lipid molecules remaining in the structure, which also helps locating the membrane plane. There are, however, more problematic cases, when the classification into membrane and soluble proteins as well as finding which part of the protein resides inside the membrane bilayer needs a more careful consideration.

Non-biological contacts between chains in the PDB structure can mislead the transmembrane detection algorithm. This makes it necessary to analyze the quaternary structure of the protein as given in the experimental structure file. Unfortunately, the information on the oligomeric form of molecules is among the least reliable records in PDB files, and often it is completely missing even if it is known by the authors. Non-biological contacts can cause a failure to recognize the membrane spanning regions and can lead to misclassification. From this viewpoint, it is more crucial to recognize non-biological complexes than to predict the full oligomeric state of the protein.

Membrane proteins are taken out form their native environment and this can have a significant impact on the structure. The presence of weak detergents, for example, can induce the formation of non-biological oligomer structures. In the PDB structure 1f88 of bovine rhodopsin, two identical chains are arranged in parallel, but in a head-to-tail orientation (Figure 2, Panel A) (Palczewski *et al.*, 2000). This arrangement allows strong contacts between the transmembrane helices as well as the soluble parts of the molecule, but because of the differences of

the extracellular and intracellular environment, this orientation cannot be the native complex. Nevertheless, the significant number of interactions makes it difficult to distinguish these cases from real oligomeric structures. The server for the quaternary structure of proteins originally developed for globular proteins (PQS)(Henrick and Thornton, 1998), failed to recognize these artifacts, and we found this wrong oligomeric structure in the Protein Data Bank biounit classification as well. However, the TMDET algorithm recognized the wrong oligomeric form and was able to correctly locate the biological molecule with the correct membrane position.

Another interesting example can result from not knowing the composition of the biological complex in advance. The complete structure of calcium-gated potassium channel (1lnq) is made up of 4 copies of the full sequence composed of a water soluble and a transmembrane domain and 4 additional copies of the water soluble domains generated by alternative splicing (Jiang *et al.*, 2002*a*; Jiang *et al.*, 2002*b*). The solved structure, however, contains four additional transmembrane domains, forming a second transmembrane region below the water soluble domains, which can not exist in the native structure (Figure 2, Panel B).

Proteins embedded into the membrane bilayer can reduce the movement of the lipid side chains, while the lipid molecules surrounding the protein impose constraints on the polypeptide chain. The elimination of these restrictions can result in a more flexible structure. This effect can be clearly seen in the case of the NMR structure of an outer membrane enzyme, pagP (1mm4), where the strands in the β-barrel are less ordered with their ends melted (Hwang *et al.*, 2002). Low-resolution structures can also look more disordered, with regular secondary structure elements becoming unrecognizable. For this reason, the proposed algorithm does not rely on secondary structure assignment, but on a more general description of straight or turn regions.

The collection of known folds of membrane proteins is much less diverse than globular folds. There are two basic structural motifs, bundles of alpha helices and beta-barrels, but the latter ones have only been found in the outer membrane of bacteria. The secondary structure elements in these folds have the characteristics length of the membrane width, but longer and shorter elements can also occur. Some of the more recent alpha-helical membrane structures

deviate from the classic themes even more. The structure of the ClC chloride channel has a complex topology, where the length of alpha helices varies widely and they are remarkably tilted (Dutzler *et al.*, 2002). The first structure of a voltage-gated ion channel provided another surprising example for the possible arrangement of transmembrane helices (Jiang *et al.*, 2003). In this structure (1orq), one of the helices lies almost parallel to the membrane plane, with four arginines exposed towards the surface. Is is not surprising that this protein gives the most unfavorable score in our algorithm. This helix is subject to large movements during gate opening, serving as some kind of a paddle, which allows the transport of potassium ion across the membrane. The Fab used for crystallyzing this protein may also alter the native conformation of the protein. Whether these unusual structures are exceptions, or will become more general as the number of known transmembrane folds grows, remains to be seen. The question also arises, whether the currently known transmembrane folds cover all the existing folds, or new structure-determination methods would yield new classes of TMPs.

Most of the TMPs deposited into the PDB are only fragments. The globular and transmembrane domains often form independent structural units which can be studied separately. Many structural studies take it even further, and analyze the structural properties of smaller fragments of TMPs such as single transmembrane helices. Actually, these single transmembrane helices dominate in the collection of TMPs (see Table 1). Some of these fragments of around 20 residues adopt their native conformation, while others remain highly unstructured in solution. The experimental conditions can also significantly influence these structures. The two structures of a beta amyloid fragment solved under two different conditions differed so significantly that one structure was classified as a transmembrane while the other one was not (1ba4, 1ba6) (Coles *et al.*, 1998; Watson *et al.*, 1998). In this case, the ambiguity of the structure is likely to contribute to its amyloid forming capability. Nonetheless, it is not clear, what is the size of the smallest part of protein which can independently form the structure similar to the one adopted under its native context; thus, the structures of small fragments should be treated carefully.

During the development of the algorithm, our main goal was to make the procedure as automatic as possible. The objective function was created in a way that it assigns lower scores

to more problematic cases, hence, it also indicates the reliability of the classification. The algorithm is able to detect and handle some discrepancies of the deposited structure files. On this basis, the number of cases requiring manual checking is automatically reduced to a small subset. There is a special group of proteins, however, which cannot be strictly assigned into soluble or transmembrane classes, as both can be correct. Some hemolytic toxins, virus coat proteins and pilus proteins are "amphibious" proteins, which can be transiently or permanently immersed into the membrane or share the life of soluble proteins depending on their structural or oligomeric forms. In the PDB_TM database, these proteins are treated as a separate group from transmembrane proteins and assigned on the basis of keyword search in the PDB file and the corresponding Swissprot entry.

One clear application of the PDB_TM database is to help the validation of transmembrane topology prediction algorithms and the structural analysis of transmembrane proteins. However, this classification is also important when globular proteins are in focus. Current selections of representative set of all PDB structures, like PDB_select (Hobohm and Sander, 1994), contain transmembrane proteins as well, although they are usually used to analyze the properties of globular proteins only. From this viewpoint, transmembrane proteins are simply contaminations, which introduce a bias because of their higher hydrophobicity and increased portion of regular secondary structure elements. The classification given in the PDB_TM database can help to create databases specific for globular proteins as well.

Acknowledgments

We wish to thank Dr. Peter Tompa (Institute of Enzymology) for his helpful comments on the manuscript. This work has been sponsored by grants BIO-0005/2001, OTKA T34131, D42207 and F043609. Zs.D. and G.E.T. were supported by the Bolyai Janos Scholarship.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Arora, A. and Tamm, L. K. (2001). Biophysical approaches to membrane protein structure determination. *Curr. Opin. Struct. Biol.* **11**, 540–547.

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.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003). The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.

Coles, M., Bicknell, W., Watson, A. A., Fairlie, D. P. and Craik, D. J. (1998). Solution structure of amyloid beta-peptide(1-40) in a water-micelle environment. Is the membrane-spanning domain where we think it is? *Biochemistry,* **37**, 11064–11077.

Davies, P. L. and Hew, C. L. (1990). Biochemistry of fish antifreeze proteins. *FASEB J.* **4**, 2460–2468.

DeLano, W. L. (1998-2003). *The PyMOL Molecular Graphich System*. DeLano Scientific LLC San Carlos, California, USA. http://www.pymol.org.

Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. and MacKinnon, R. (2002). X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature,* **415**, 287–294.

Henrick, K. and Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.

Hwang, P. M., Choy, W. Y., Lo, E. I., Chen, L., Forman-Kay, J. D., Raetz, C. R., Prive, G. G., Bishop, R. E. and Kay, L. E. (2002). Solution structure and dynamics of the outer membrane enzyme PagP by NMR. *Proc. Natl. Acad. Sci. USA,* **99**, 13560–13565.

Ikeda, M., Arai, M., Okuno, T. and Shimizu, T. (2003). TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.* **31**, 406–409.

Jayasinghe, S., Hristova, K. and White, S. H. (2001). MPtopo: A database of membrane protein topology. *Protein Sci.* **10**, 455–458.

Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. T. and MacKinnon, R. (2002*a*). The open pore conformation of potassium channels. *Nature,* **417**, 523–526.

Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. T. and MacKinnon, R. (2002*b*). Crystal structure and mechanism of a calcium-gated potassium channel. *Nature,* **417**, 515–522.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T. and MacKinnon, R. (2003). X-ray structure of a voltage-dependent K+ channel. *Nature,* **423**, 33–41.

Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Letters,* **423**, 281–285.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

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.

Kumar, M. S., Carson, M., Hussain, M. M. and Murthy, H. M. (2002). Structures of apolipoprotein A-II and a lipid-surrogate complex provide insights into apolipoprotein-lipid interactions. *Biochemistry,* **41**, 11681–11691.

Lee, A. G. (2003). Lipid-protein interactions in biological membranes: A structural perspective. *Biochim. Biophys. Acta,* **1612**, 1–40.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Möller, S., Kriventseva, E. V. and Apweiler, R. (2000). A collection of well characterised integral membrane proteins. *Bioinformatics,* **16**, 1159–1160.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540. http://scop.mrc-lmb.cam.ac.uk/scop/.

Nilsson, I., Johnson, A. E. and von Heijne, G. (2003). How hydrophobic is alanine? *J. Biol. Chem.* **278**, 29389–29393.

Ostermeier, C. and Michel, H. (1997). Crystallization of membrane proteins. *Curr. Opin. Struct. Biol.* **7**, 697–701.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science,* **289**, 739–745.

Rees, D. C., DeAntonio, L. and Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science,* **245**, 510–513.

Tompa, P., Tusnády, G. E., Cserző, M. and Simon, I. (2001). Prion protein: evolution caught en route. *Proc. Natl. Acad. Sci. USA,* **98**, 4431–4436.

Tusnády, G. E. and Simon, I. (2001). Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.* **41**, 364–368.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494.

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.

Watson, A. A., Fairlie, D. P. and Craik, D. J. (1998). Solution structure of methionine-oxidized amyloid beta-peptide (1-40). Does oxidation affect conformational switching? *Biochemistry,* **37**, 12700–12706.

White, S. H. and Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu Rev. Biophys. Biomol. Struct.* **28**, 319–365. http://blanco.biomol.uci.edu/Membrane_proteins_xtal.html.

Distribution of transmembrane chains with different topologies

NS	1	2	3	4	5	6	7	8	9	10	11
α	372	103	39	15	99	46	131	19	0	18	14
β	0	7	0	3	0	0	0	9	0	0	0

NS	12	13	14	15	16	17	18	19	20	21	22
α	38	1	0	0	0	0	0	0	0	0	0
β	9	0	0	0	51	0	31	0	0	0	16

Table 1: NS is the number of transmembrane segments per protein chain for α-helical and β-barrel proteins.

Representative protein families in the PDB_TM database

Beta-barrel transmembrane proteins			
Family name	**NC**	**NTM**	**PDB code**
Alpha-hemolysin	7	2	7ahlA
Outer membrane protein, tolC	3	4	1ek9A
Outer membrane enzyme, pagP	2	8	1mm4A
Outer membrane protein A, ompA	7	8	1g90A
Outer membrane protease, ompT	1	10	1i78A
Outer membrane adhesin/invasin, opcA	1	10	1k24A
Outer membrane phospholipase A	9	12	1qd5A
Porin	51	16	1prn
Maltoporin	31	18	1oh2P
Outer membrane transporter, fecA	16	22	1kmoA

Table 2. to be continued

Representative protein families in the PDB_TM database (cont)

Alpha-helical transmembrane proteins			
Family name	**NC**	**NTM**	**PDB code**
Photosystem I, subunit psaK	1	2	1jb0K
Mechanosensitive channel, mscL	5	2	1mslA
Sensory rhodopsin II transducer	2	2	1h2sB
ATP synthase, subunit C	40	2	1ijpA
Band 3 anion transport protein	2	2	1bzkA
Cytochrome c oxidase polypeptide II	16	2	1occB
Voltage-gated potassium channel	45	2	1bl8A
Photosystem I, subunit psaL	1	3	1jb0L
Mechanosensitive channel, mscS	7	3	1mxmA
Fumarate reductase, 13kDa	16	3	1l0vD
Acetylcholine receptor	12	4	1oedA
ATP synthase A chain	1	4	1c17M
Thromboxane A2 receptor	1	4	1lbnA
FDN cytochrome b556 subunit	2	4	1kqfC
Fumarate reductase, cytochrome B	6	5	1qlbC
Photosynthetic reaction centers	84	5	1aigL
Aquaporins	29	6	1ih5A
Photosystem II, subunit psbC	4	6	1izlC
PTH/PTHR receptor	2	7	1et2S
Cytochrome c oxidase polypeptide III	14	7	1occC
Bacteriorhodopsins	107	7	1qhjA
Rhodopsins	14	7	1f88A
Cytochrome bc1 complex, cytochrome b	17	8	1bccC
ClC-type chloride channel, clcA	10	10	1kplA
Plasma membrane ATPase	6	10	1mhsA
Photosystem I, subunit psaA	2	11	1jb0A
Glucose transporter, glut1	1	12	1ja5A
Cytochrome c oxidase polypeptide I	16	12	1occA
ABC transporters	22	12	1iwgA

Table 2: NC: number of protein chains in the family, NTM: number of transmembrane segments in the given chain, the four letter PDB code with the protein chain indicator shows one representative protein from the family. Protein families with single transmembrane helices or without valid sequences are not listed. A more detailed list of protein classification is available at the PDB_TM homepage (http://www.enzim.hu/PDB_TM/families.html).

Figure legends

Figure 1: Frequencies of Q-values. Distribution of values of the objective function (Q-values) for transmembrane (solid line) and globular (dashed line) proteins. The region between the lower and upper selection limits (see text) is highlighted in light gray.

Figure 2: Two examples of structural discrepancies of membrane proteins. Panel A: The structure of bovine rhodopsin (1f88) from outside (top), and from the membrane plane (bottom). The two identical chains are associated in a head-to-tail orientation. The color scheme is the following: the membrane spanning region is colored green, the side containing the N-terminal is red, and the side containing the C-terminal is blue for both chains. Panel B: The structure of calcium-gated potassium channel (1lnq) with an extra transmembrane region. The eight chains (A-H) are shown in different colors in side-one (A-D: blue-magenta, E-H: green-cyan), while in side-two each is red. The membrane-spanning regions are colored in yellow and secondary structure is shown only in these regions.

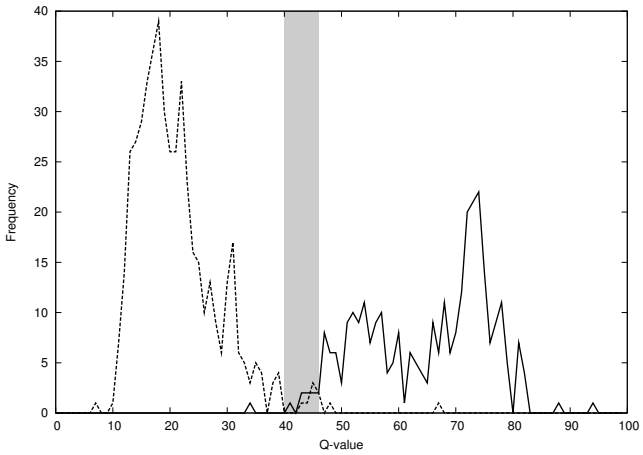

Figure 1:

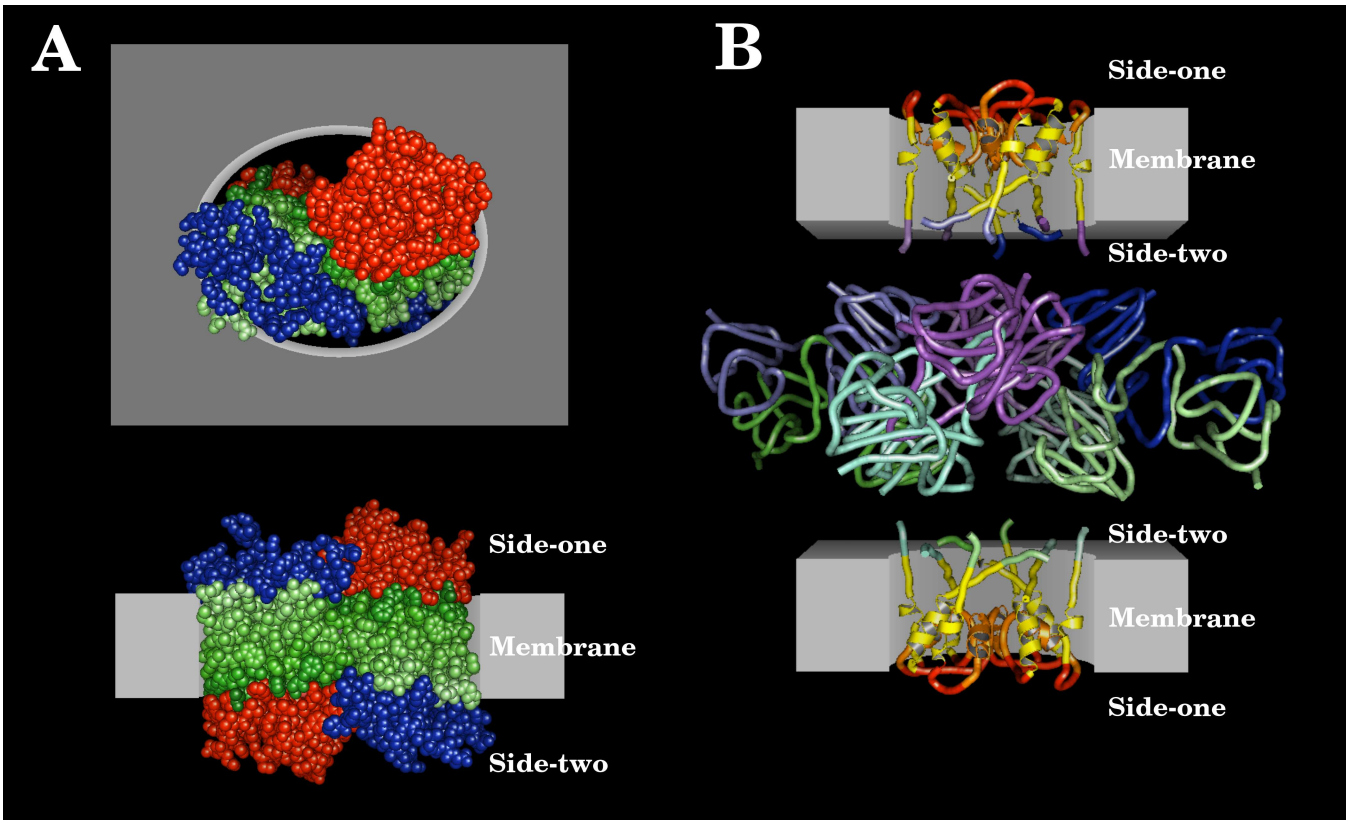

Figure 2: