Journal of Bioinformatics and Computational Biology
Vol. 4, No. 5 (2006) 1033–1056
© Imperial College Press

Imperial College Press
www.icpress.co.uk

RECOGNITION OF TRANSMEMBRANE SEGMENTS IN PROTEINS: REVIEW AND CONSISTENCY-BASED BENCHMARKING OF INTERNET SERVERS

NATALIYA S. SADOVSKAYA

Institute for Information Transmission Problems, Russian Academy of Science
Bolshoi Karetny per. 19, Moscow 127994, Russia
and
State Scientific Center "GosNIIGenetika"
1st Dorozhny proezd 1, Moscow 113545, Russia
nathalie@ippi.ru

ROMAN A. SUTORMIN

State Scientific Center "GosNIIGenetika"
1st Dorozhny proezd 1, Moscow 113545, Russia
and
Department of Bioengineering and Bioinformatics
Moscow State University, Vorobievy Gory 1-73, Moscow, Russia
sutor_ra@mail.ru

MIKHAIL S. GELFAND*

Institute for Information Transmission Problems, Russian Academy of Science
Bolshoi Karetny per. 19, Moscow 127994, Russia
and
Department of Bioengineering and Bioinformatics
Moscow State University, Vorobievy Gory 1-73, Moscow, Russia
gelfand@iitp.ru

Received 28 February 2006
Revised 21 June 2006
Accepted 22 June 2006

Membrane proteins perform a number of crucial functions as transporters, receptors, and components of enzyme complexes. Identification of membrane proteins and prediction of their topology is thus an important part of genome annotation. We present here an overview of transmembrane segments in protein sequences, summarize data from large-scale genome studies, and report results of benchmarking of several popular internet servers.

Keywords: Transmembrane proteins; transmembrane helix prediction; benchmarking.

*Corresponding author.

1. Introduction

Membrane proteins constitute 15–30% of a typical organism's proteome[1–3] and include important functional classes such as transporters,[4] ion channels,[5–7] receptors,[8] components of respiration chains,[9] etc. They can be roughly divided into three classes: globular proteins anchored to the membrane via a single hydrophobic alpha-helix, proteins located in the membrane and containing several membrane-spanning alpha-helices, and beta-structural membrane proteins. The latter are relatively rare, although they also include numerous important representatives, e.g. outer membrane porins.[4]

Much less is known about the structure of transmembrane (TM) domains than that of globular domains. The MPtopo database (last updated May 17, 2005) contains a total of 167 proteins with 1028 TM segments.[10] The entire TM section of PDB, PDB_TM, contains 482 proteins (as of April 2005), which is less than 2% of all PDB entries.[11–13]

This is mainly due to technical difficulties: membrane proteins are usually present at low levels in biological membranes, and most membrane proteins cannot be readily obtained in sufficient amounts from their native environments. Furthermore, they are difficult to overexpress[14] and to crystallize.[15–17]

Properties of membrane proteins are quite different from those of globular proteins,[17] and generic algorithms for prediction of protein secondary structure do not work well when applied to membrane proteins.[18] Although recently a major international project has been launched that aims at large-scale determination of TM-protein structures (http://www.utoronto.ca/AlEdwardsLab/membrane_proteomics_index.html), it still has not produced sufficient amount of data to cover all structural classes of membrane proteins. Moreover, in several cases, there is discrepancy in experimental evidence about the secondary structure of TM-protein, in particular, the number of TM-segments.[19–23]

2. Statistical Properties of TM-Segments

The need to identify and characterize TM-proteins yielded numerous algorithms for identification of alpha-helical TM-segments. They differ in features of TM-segments taken into account and in algorithmic techniques. The most important distinguishing feature of TM-helices is the amino acid composition. As these segments span the lipid bilayer, they predominantly contain hydrophobic residues. Most TM-helices have length between 12 and 35 residues,[3] with the average being 21 residues.[24] However, these numbers should be considered with due caution.[25] Indeed, boundaries of alpha-helices are difficult to define rigorously,[26] and no experimental methods can exactly map the boundary between membrane-embedded and external residues in a TM-helix. The ends of TM-helices, the so-called caps, can be recognized by specific amino acid composition.[27,28] Due to the special mechanism of membrane translocation of TM-proteins, the amino acid composition of inner and outer loops is different: short non translocated (cytoplasmic side) loops contain many positively charged residues (Arg, His, Lys, Glu, and Asp), whereas translocated loops contain few such residues.[28–36] Inside and outside caps of TM-helices also differ in the amino acid composition;[27,28] specifically, these regions are abundant in polar and aromatic residues (Lys, Arg, Trp, His, Glu, Gln, and Tyr).[37,38] Globular regions and loops between TM-helices are usually shorter than 60 residues.[1] Finally, topological constraints put natural restrictions on the order of TM-helices and inside and outside loops that can also be taken into account.[39,40] There are also differences in the amino acid composition of inside/outside domains and TM-caps between single-TM-helix and multiple-TM-helix proteins.[27,28]

An additional complication in the analysis of single-TM-helix membrane-anchored proteins is to distinguish them from secreted proteins containing an *N*-terminal signal peptide. There are specific programs for recognition of signal peptides that take into account not only amino acid composition of the signal peptide, but also the positional residue frequencies around the cleavage sites.[41,42] These programs are, in particular:

SignalP[41,43] (http://www.cbs.dtu.dk/services/SignalP/),
PrediSi[44] (http://www.predisi.de/),
SIGSEQ[45] (ftp://ftp.ebi.ac.uk/pub/software/unix/sigseq.tar.Z),
SPEPlip[46] (http://gpcr.biocomp.unibo.it/predictors/).

On the other hand, there exist specific methods for prediction of single TM-helices.[47]

A special case is that of TM-proteins that have the antiparallel beta-barrel fold, such as bacterial porins.[48] The amino acid composition of TM-beta-strands is different from both TM-alpha-helices[35,49] and beta-strands of globular proteins.[50,51] Although there exist several methods for prediction of transmembrane beta-strands,[52–68] only 12 of them are available over the internet (Table 1). We do not consider them here.

Table 1. Servers for the prediction of transmembrane beta strands.

Server	References	URL
Omp_topo_predict	61	http://turn18.biologie.uni-konstanz.de/test/om_topo_predict2b.html
B2TMPRED	66	http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi
TRAMPLE	59	http://gpcr.biocomp.unibo.it/biodec/
HMM-B2TMR, B2TMR	66, 69	http://gpcr.biocomp.unibo.it/predictors/
PRED-TMBB	64	http://bioinformatics.biol.uoa.gr/PRED-TMBB
ConBBPRED	68	http://bioinformatics.biol.uoa.gr/ConBBPRED/
PROFtmb	65	http://www.rostlab.org/services/PROFtmb/
TBBpred	62	http://www.imtech.res.in/raghava/tbbpred
TMBETA-NET	60	http://psfs.cbrc.jp/tmbeta-net/
TMBETA-SVM	67	http://tmbeta-svm.cbrc.jp

3. Algorithmic Approaches to the Identification of TM-Segments

The simplest technique for identification of TM-segments is the computation of a hydrophobicity score in a sliding window. Many different hydrophobicity scales have been suggested based on physical characteristics of amino acid residues.[38,70–78] Alternatively, one can use statistical approaches, comparing residue frequencies in TM-segments, inner and outer loops, and calculating residue propensity toward these regions.[27,79,80] For that, one score may be computed in a sliding window;[70] or a combined score may be computed by merging several physiochemical indices characterizing TM-segments, caps of TM-segments, and loops.[37,79,81] A similar technique is to predict conformation/location of each residue with subsequent averaging at adjacent positions.[80]

Instead of simple statistical procedures, windows can be classified into TM-segments and nonmembrane ones using more complicated pattern recognition techniques such as logical rules,[36] *k*-nearest neighbor analysis,[82] classification,[83] and neural networks.[84,85] However, isolated prediction of TM-segments and their orientation is less reliable than an analysis of the entire protein. One approach to that problem is to use greedy algorithms,[86] dynamic programming,[27] or other algorithms[87] to reconcile individually characterized TM-segments with additional constraints. A more consistent approach is to use Hidden Markov Models (HMMs) with states corresponding to TM-segments, inner and outer caps, inner and outer loops, and globular domains.[88,89] Residue frequencies in different regions and segment lengths (symbol probabilities and duration of states, respectively, in the HMM terminology) are learned from training samples, and then each new sequence is decomposed into the most probable sequence of hidden states, satisfying natural topology constraints.

Thus, although some algorithms predict only TM-segments,[25,37,38,79,84,90] most current methods predict the entire protein topology[30,40,61,88,89,91–93] (Table 2). Some methods allow the user to constrain the prediction by available experimental or comparative data, so that particular regions of a protein are set into a user-defined state.[94]

In addition to prediction of topology of TM-proteins with multiple TM-helices, especially transporters, it is important to perform rotational positioning of these helices. Indeed, substrates may interact with sites within a TM-domain; and these sites may define the specificity and kinetics of transport.[108] This analysis is based on the assumption that the side of a TM-helix that contacts the membrane is hydrophobic, whereas the side facing the channel may contain hydrophilic residues. Indeed, TM-segments in transporters demonstrate periodicity in positions of hydrophilic residues.[109] In densely packed proteins, there are differences in the composition of lipid-exposed and buried helix sides.[110–112] However, this statistical trend is too weak to analyze relatively short TM-segments; and such an analysis requires additional data.[80] One possible complication obscuring the simple channel side *versus* membrane side picture above comes from helix–helix interactions. Specific indices were developed for the analysis of such interactions.[113–115]

Table 2. Servers for prediction of TM-segments and topology of TM-proteins. Boldface: ten servers used in benchmarking.

Server	References	What is Predicted	Algorithm	Parameters	URL
SOSUI	37, 38	TM-segments	Slidng window	Hydrophobicity, amphiphilicity at caps, charge, length	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
TopPred II	95	TM-segments, orientation	Multiple alignment, sliding window, topology	Hydrophobicity, length, charge in inner/outer loops, amino acid composition of long loops	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
PRED-TMR	79	TM-segments	Sliding window	Hydrophobicity, length, TM and cap propensities	http://o2.biol.uoa.gr/PRED-TMR/
OrienTM	40	Orientation	Topology	Inner/outer loop propensities	http://o2.biol. uoa.gr/orienTM/
SPLIT	80, 93	TM-segments, orientation	Residue analysis with averaging	Hydrophobicity and other scales, hydrophobic moment, conformational propensity, charge motifs	http://split.pmfst.hr/split/ http://split.pmfst.hr/split/4/
PHDhtm	84, 86	TM-segments, orientation	Multiple alignment, sliding window, neural network, dynamic programming	TM and inner loop propensitiy, correlations between adjacent residues, conservation	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA /npsa_htm.html
DAS	25	TM-segments	Averaging of dot-plots	TM propensities (indirectly)	http://www.sbc.su.se/~miklos/DAS/
PSORT	56, 82, 96–99	TM-segments, orientation	k-nearest neighbour method	Hydrophobicity, charge difference, signal motifs	http://www.psort.org/

Table 2. (*Continued*)

Server	References	What is Predicted	Algorithm	Parameters	URL
TMHMM 2.0	89, 100	TM-segments, orientation	HMM	Amino acid composition and lengths of TM-segments, caps and loops	http://www.cbs.dtu.dk/services/TMHMM-2.0/
HMMTOP 2.0	88, 94	TM-segments, orientation	HMM	Amino acid compositions and lengths of TM-segments, caps and loops	http://www.enzim.hu/hmmtop/index.html
MEMSAT 2	101	TM-segments, orientation	Dynamic programming	PSI-BLAST profiles for topological models	http://www.sacs.ucsf.edu/secure/cgi-bin/memsat.pl
CoPreTHi	102	TM-segments	Sequence profile	3 out of DAS, ISREC-SAPS, PHD, SOSUI, TMpred, TopPred II, PRED-TMR	http://o2.biol.uoa.gr/CoPreTHi/
ConPred II	103	TM-segments, orientation	Sliding window	Average of KKD, TMpred, TopPred II, DAS, TMAP, MEMSAT 1.8, SOSUI, TMHMM 2.0 and HMMTOP 2.0.	http://bioinfo.si.hirosaki-u.ac.jp/∼ConPred2/
BPROMPT	104	TM-segments, orientation	Bayesian Belief Network	Combination of HMMTop 2, DAS, SOSUI, TMpred, TopPred II	http://www.jenner.ac.uk/BPROMPT
TMpred	105	TM-segments, orientation	Not described	Amino acid composition, length	http://www.ch.embnet.org/software/TMPRED_form.html
SAPS	83	TM-segments	Statistics	Hydrophobicity	http://www.isrec.isb-sib.ch/software/SAPS_form.html
TMAP	106, 107	TM-segments, orientation	Multiple alignment, sliding window, topology	Amino acid composition, length	http://www.mbb.ki.se/tmap/single.html

An important feature of many methods is that they do not make isolated predictions, but place them in context. There are three different approaches to that. One approach is to predict TM-segments in aligned homologous proteins. The basic assumption here is that the number and positions of TM-segments in homologous proteins are the same; and taking that into account decreases the noise. This technique was used in Refs. 24, 84, 86, 91, 95, and 106. However, alignment of TM-proteins is not trivial, as they consist of segments with clearly different statistical properties, and different amino acid substitution matrices may be needed in these segments, while boundaries are unknown.[116] MEMSAT uses host-slave-type multiple alignments generated by PSI-BLAST,[117] and averages the predictions across the alignment.[101] The use of homology increased the fraction of correctly predicted topologies from 78% to 93%. An alignment algorithm based on averaging of dot-plots is more robust as regards the level of sequence identity;[77] and the DAS algorithm for prediction of TM-segments is based on matching of the query protein to all TM-proteins in the database.[25] This is an indirect way of computing residue propensities to TM-helical regions.

Another approach is to consider all possible structural assignments for a protein at once. This is especially important for proteins with one TM-helix and for the analysis of N-terminal TM-helices. Indeed, one of the problems is to distinguish between TM-segments and predicted signal peptides;[89] and some methods even require prior prediction and removal of signal peptides.[93] To avoid this, PSORT predicts TM-segments and various signals for cellular localization of proteins.[82] Finally, several servers map TM-segments by averaging predictions of other servers. For example, a simple voting can be done separately for each residue[102] or for sliding windows,[118] or a Bayesian decision procedure can be implemented.[104]

Overall success rate for the predictions of TM-protein topology was claimed to be 95% for bacterial and 83% for eukaryotic proteins by TopPred II,[95] 86% by PHDhtm,[86] 85% by HMMTOP,[88] 93% by MEMSAT 2,[101] 87% to 97% for prokaryotic and 94% to 97% for eukaryotic proteins by OrienTM,[40] 78% by TMHMM,[89] and 75% by SPLIT.[93]

It was been shown that the use of taxonomy-specific parameters may improve the performance on bacterial,[34] eukaryotic,[30,33] mitochondrial,[119] and chloroplast[120] proteins. Some structural families may present specific problems, e.g. G-coupled receptors containing seven TM-helices.[24]

4. Neural Networks and Hidden Markov Models for Identification of TM-Segments

We describe in this section two of the most successful algorithmic techniques implemented in three programs.

The most successful application of the neural network methodology to the TM-segment prediction is PHDhtm. Initially created as an algorithm for the prediction of the protein secondary structure,[18] it was then specifically re-designed for the identification and analysis of TM-proteins.[84,86]

PHDthm consists of two successive neural networks and a number of pre-processing and post-processing filters. The first, a sequence-to-structure network, processes overlapping sliding windows of length 13 and estimates the probability of TM localization for the residue in the middle of the window taking into account the identity of all 13 residues and also global parameters such as the protein length, the window distance to the N- and C-termini, and the protein amino acid composition. The network consists of an input layer, a hidden layer with three units, and an output layer that produces measures of TM-propensity and non-TM-propensity, that is, the preferences of the central residue to be in a TM helix or in a non-TM loop. The feed-forward network with a complete set of connections and a sigmoid trigger function is trained using the gradient descent procedure. The technical details are described in Ref. 18.

Then one more neural network is applied. This second network, a structure-to-structure network, takes into account the context of predictions, e.g. to filter out TM-segments that are too short. Its input is the output of the first network in 21-residue windows supplemented with global information, and its hidden layer contains 15 units.

At the next step, a decision is made on whether the considered protein is a TM-protein. This is done by computing the total TM-propensity of sliding windows of length 18: if no such window passes a threshold, the protein is predicted to be a non-TM one.

For a predicted TM-protein, a dynamic programming-based procedure is used to reconcile conflicting predictions and to determine the optimal TM-segment boundaries.[86] A set of possible TM-segments is compiled by considering all windows of length 18 through 25. The average TM-propensity is calculated for each window. An optimal path through all windows is constructed using a dynamic programming procedure, with a restriction that the TM-segments should be separated by loops of the minimal length 4. This path represents the optimal decomposition of the protein into TM-segments and non-TM-loops.

Finally, for a given set of TM-segments, the topology (that is, the protein orientation relative the membrane) is predicted.[86] This is done by computing the overall charge of each non-TM-segment, and then taking the difference between the total charge of odd and even non-TM-regions. The set with the more positive charge should reside on the cytoplasmic side, based on the so-called "positive-inside rule".

Although quite powerful, to perform well, any neural network-based approach needs to involve a considerable number of *ad hoc* filters and procedures (*cf.* the evolution of post-processing filtering/reconciliation steps in the three papers describing PHDthm[18,84,86]), and thus lacks conceptual clarity. A uniform language for the description of symbol sequences containing segments of different statistical properties is the HMMs.[121]

A HMM is defined by a set of transition probabilities $Q(S \to T)$ between states (here, the probability to move into state T if the current state is S) and a set of emission probabilities $R(A|S)$ for each state (here, the probability to observe an outcome, that is, residue A if the current state is S). Each realization $(A_1, \ldots, A_n; S_1, \ldots, S_n)$ of a HMM is a set of observable outcomes and hidden states, and its probability is defined as

$$P(A_1, \ldots, A_n; S_1, \ldots, S_n) = q(S_1) \cdot r(A_1|S_1) \cdot \Pi_{k=2,\ldots,n} Q(S_{k-1} \to S_k) \mathrm{R}(A_k|S_k)$$

where $q(S)$ and $r(A|S)$ are the vectors of initial probabilities to be in the state S and generate in this state the outcome A, respectively.

This simplest definition can be easily generalized for Markov chains of higher order for states. It is also simple to require that an outcome depends not only on the state, but also on the previous outcome, thus introducing the Markovian dependence to the sequence of outcomes.

These general definitions are applied to the recognition of TM-segments as follows. The protein is assumed to represent a sequence of outcomes, and the aim is to reconstruct the most likely sequence of hidden states, so that $P(A_1, \ldots, A_n; S_1, \ldots, S_n)$ is maximized over all state sequences $S_1, \ldots, S_n$.

The hidden states may be TM-segment, external loop, internal loop, external cap (region of transition between the membrane and non membrane surroundings), cytoplasmic cap, N-terminus, C-terminus, external globular domain, internal globular domain etc., dependent on the architecture of the HMM. The natural restrictions (e.g. an external loop should follow an external cap, etc.) are easily taken into account by allowing only certain transitions (Fig. 1(a)); to take into account the length distribution in TM-segments, loops and caps, these states are multiplied (Fig. 1(b)) and the transition probabilities are set according to the distribution. The transition and emission probabilities are learned at the training step.

The reconstruction of the most likely sequence of hidden states is done via a dynamic programming-like technique called the Baum-Welch (or forward-backward) algorithm.

This approach was implemented, in particular, in HMMTOP[88,94] and TMHMM.[89,100] These two algorithms differ in minor details of their HMM architecture, *cf.* Fig. 3 of Ref. 88 and Fig. 1 of Ref. 89.

Finally, it should be noted that both neural network and HMM algorithms can be easily modified to accept as input not single sequences, but multiple alignments.[84,121]

5. Applications to Genomic Analysis

Large-scale analyses of ORFs in complete genomes by various algorithms, in particular JTT2,[122] TopPred,[1] SOSUI,[2] MEMSAT,[28] PHDhtm,[3,84,86] TMHMM,[89] produced a consistent estimate of about 25% and seemed to show no difference in the TM-protein fraction in bacterial, archaeal, and eukaryotic genomes, although an increase of the TM-proteins fraction with the genome complexity was reported.[1]

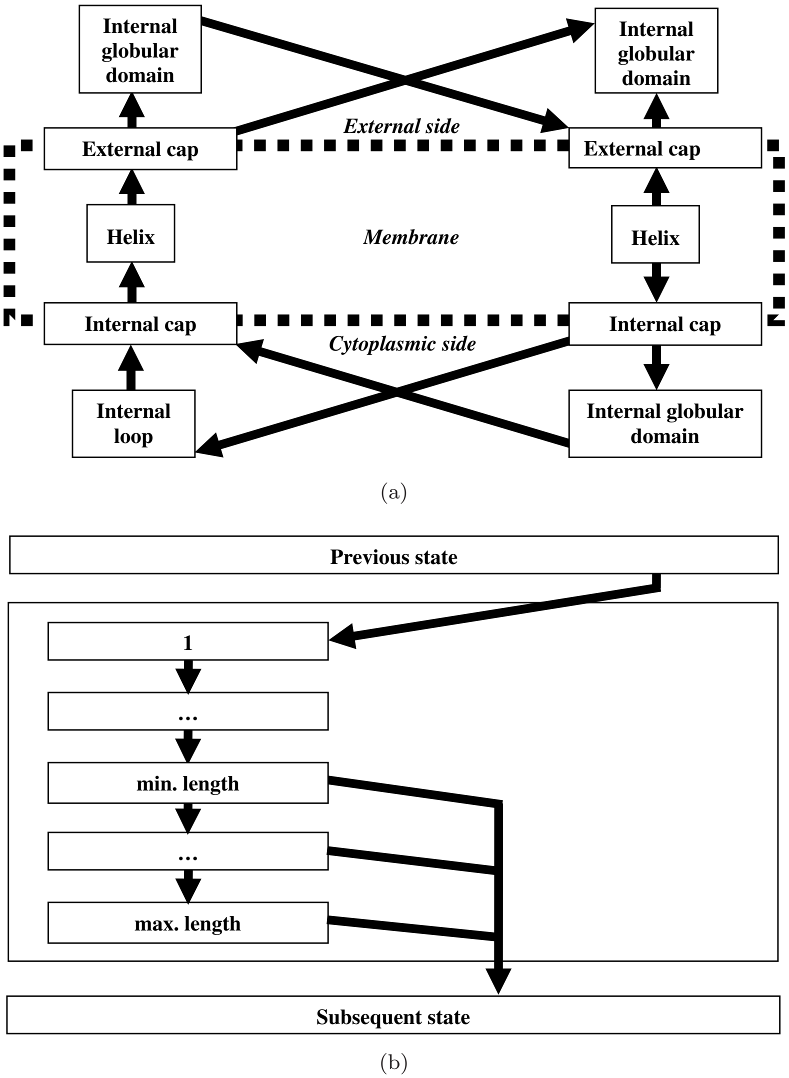

Fig. 1. (a) Architecture of a Hidden Markov Model for identification of TM-segments. (b) Module with multiplied steps, allowing for exact modeling of length distribution.

In individual genomes, this fraction ranges from 19% through 30% (Table 3). The difference between individual predictors was only a few percent points (Table 3).

However, there exists a controversy about the existence of preferred TM-protein topologies. Several studies observed simple monotonic decrease in the fraction of TM-proteins as the number of TM-segments increased,[123–125] whereas over-representation of proteins with seven TM-segments was observed in multicellular eukaryotes, and over-representation of proteins with 6 and 12 TM-segments, in prokaryotes.[3] Membrane proteins seem to come in two basic varieties: those with many TM-segments and short connecting loops, and those with few TM-segments

Table 3. Estimated fraction of TM-proteins in selected complete genomes.

Algorithm	TopPred[a]	TopPred[b]	TopPred[c]	SOSUI	MEMSAT	PHDhtm	TMHMM
References	1	1	1	2	28	3, 84, 86	89
H. sapiens	43%	37%	26%	n/a	27%	14–18%	n/a
D. melanogaster	n/a	n/a	n/a	n/a	n/a	18%	20%
C. elegans	46%	41%	30%	n/a	25%	30%	30%
S. cerevisiae	40%	34%	23%	25%	18%	22–25%	21%
E. coli	40%	34%	24%	26%	26%	25%	21%
H. influenzae	30%	26%	18%	23%	23%	19%	19%
H. pylori	32%	27%	19%	25%	n/a	21%	19%
C. jejuni	n/a	n/a	n/a	n/a	n/a	24%	21%
R. prowazekii	n/a	n/a	n/a	n/a	n/a	28%	26%
M. tuberculosis	n/a	n/a	n/a	n/a	n/a	19%	18%
B. subtilis	33%	29%	23%	30%	n/a	19%	24%
M. genitalium	29%	25%	18%	25%	n/a	19%	20%
M. pneumoniae	29%	23%	16%	26%	n/a	18%	18%
T. pallidum	n/a	n/a	n/a	23%	n/a	21%	24%
B. burgdorferi	n/a	n/a	n/a	29%	n/a	27%	29%
C. pneumoniae	n/a	n/a	n/a	n/a	n/a	27%	28%
C. trachomatis	n/a	n/a	n/a	26%	n/a	22%	25%
A. aeolicus	n/a	n/a	n/a	25%	n/a	18%	21%
Synechocystis sp.	41%	35%	24%	27%	n/a	n/a	26%
D. radiodurans	n/a	n/a	n/a	n/a	n/a	17%	19%
T. maritima	n/a	n/a	n/a	n/a	n/a	20%	24%
M. thermoautotrophicum	n/a	n/a	n/a	n/a	n/a	18%	22%
A. fulgidus	30%	26%	19%	28%	n/a	18%	20%
P. abyssi	n/a	n/a	n/a	n/a	n/a	21%	23%
P. horikoshii	n/a	n/a	n/a	n/a	n/a	23%	26%
M. jannaschii	23%	20%	14%	21%	n/a	18%	n/a

[a] TopPred: set 1: at least two "putative" or "certain" TM;
[b] TopPred: set 2: at least one "certain" and one "putative" or "certain" TM;
[c] TopPred: set 3: at least two "certain" TM.

and large extra-membrane domains, whereas proteins with multiple TM-segments and large extra-membrane domains are relatively rare.[1] Further, if the orientation relative to the membrane is taken into account, all organisms excluding *C. elegans* (but including some other multicellular eukaryotes) prefer the topology with an even number of TM-segments and the *N*-terminus in the cytoplasmic side, whereas *C. elegans* has more proteins with an odd number of TM-segments, external *N*-terminus and internal *C*-terminus.[89] A more detailed analysis in Ref. 28 showed that bacterial TM-proteins with an even number of TM-segments (4, 6, 10, 12, 14) tend to have the topology with the *N*-terminus in the internal side of the membrane (nontranslocated), whereas the proteins with 5 TM-segments tend to have external *N*-termini. In yeast, the preferred topologies are almost the same (even with *N*-termini inside), whereas in multicellular eukaryotes the preferred topologies are 4-TM and 12-TM with the internal *N*-terminus, and 7-TM with the external *N*-terminus. The latter are likely due to the large family of G protein-coupled receptors,[126] whereas 12-TM proteins are mainly permeases and TM-components of ABC-transporters.[4]

6. Benchmarking

Benchmarking of algorithms for prediction of TM-segments is complicated by a relatively small number of resolved TM-protein structures and the low rate of newly arising structures, leading to difficulty in finding examples new for each particular method. The same datasets have been used over and over again, e.g. the 64 TM-protein dataset in Refs. 27, 79, 84, and 90. Several hydrophobicity scales were compared in Ref. 75, and benchmarking of several popular methods (TMHMM 1.0, 2.0[89,100] and a retrained version of 2.0,[127] MEMSAT,[27] Eisenberg,[128] Kyte/Doolittle,[70] TMAP,[107] DAS,[25] HMMTOP,[88] SOSUI,[37] PHD,[129] TMpred,[105] KKD,[29] ALOM2[36] and Topred II[95]) was done in Ref. 127 using a newly compiled database of about 188 TM-protein topologies obtained by low-resolution experiments such as *C*-terminal fusions with indicator proteins and antibody binding.[130] A commonly used database of TM-protein structures is MP-topo.[10] A combined database of structures and topologies supplemented by proteins whose structural similarity could be inferred from sequence comparison[131] was used for comprehensive benchmarking in Refs. 132 and 133. This database also serves as an automated benchmark server for evaluation of new algorithms (http://cubic.bioc.columbia.edu/services/tmh_benchmark/). However, despite specific attempts to perform benchmarking on new data samples or to apply bootstrapping approaches, it has been repeatedly noted that it is difficult to compile a clearly independent sample. Thus, the performance of all methods tends to be over-estimated.[132,134] The results of these benchmarks are discussed below.

Here we tried to benchmark TM-segment prediction algorithms using a comparative approach. It is based on the assumption that homologous proteins have the same number of TM-helices, and these TM-helices can be aligned. Thus instead of matching predictions to experimental data, we measure the server's reliability using the consistency of its predictions on a group of aligned homologs. This approach is complementary to the experimental data-based benchmarking.

It is clear that this approach has limitations. Firstly, a consistently wrong algorithm would get a high score. However, we feel that such possibility is unlikely, and the results presented below seem to confirm this assumption. Secondly, algorithms using multiple alignments as input cannot be considered, as they use the same type of data and thus the consistency approach would be based on a circular argument. Thirdly, protein families having an intermediate level of similarity should be analyzed. Indeed, the proteins should be sufficiently similar to each other to satisfy the assumption of the conserved topology and, moreover, to allow for reliable alignment. On the other hand, it is useless to consider proteins that are too similar, as they would provide no independent data.

7. Data and Methods

Initially, all representatives of bacterial transporters from class TC.2A were considered, downloaded from TCDB databases http://www-biology.ucsd.edu/∼msaier /transport/[4,135] and TransportDB http://www.membranetransport.org/.[136] This sample consisted of 1312 proteins from 101 families. The majority of sequences were from four families from the MFS superfamily (438 proteins), five families from the APC family (103 sequences), and four families from the RND family (114 proteins). The CPA3 and NFE families, consisting of subunits of multicomponent systems, were not considered.

Then each sequence from the initial sample was used to scan the ERGO database (http://ergo.integratedgenomics.com/ERGO/)[137] using BLAST.[138] Only relatively complete genomes were considered, such that at least 500 genes were sequenced as not more than 10 contigs. It produced 860 additional proteins homologous to proteins from the initial sample.

The obtained sample was divided into clusters using the single linkage procedure, with the BLAST identity value serving as the similarity measure. Clusters with identity 40% to 49% and 50% to 59% were considered. Clusters larger than 50 proteins were further subdivided by increasing the identity threshold. Each cluster was aligned using CLUSTALW.[139] The comparisons were done for all pairs of proteins belonging to one cluster with pairwise alignment induced by the multiple alignment. Overall, 2356 pairs of proteins from the first group (40% to 49%) and 909 pairs of proteins from the second group (50% to 59%) were considered.

For each pair, TM-segments were predicted independently, and the consistency indices between the predictions by one method were computed as follows. Overlapping or immediately adjacent (with zero size loop) TM-segments were merged and considered as one TM-segment. The overlap index Q compared predictions at the residue level. For each pair it was equal to the size of the intersection divided by the size of the union of the segment projections. More formally, let S be the number of aligned residue pairs where both members were predicted to belong to TM-segments, and let U be the number of aligned residue pairs where at least one member is predicted to belong to a TM-segment. Then

$$Q = S/U. \tag{1}$$

Another measure, the segment consistency index C, measured the fraction of common TM-segments in two proteins. Let n_1 and n_2 be the numbers of TM-segments in the first and second protein, respectively. Let $i = 1, \ldots, n_1, j = 1, \ldots, n_2$ be TM-segments in the first and second proteins, respectively, and consider all pairs ij of TM-segments whose projections intersect by at least one residue pair. Define the indicator value, the local overlap V_{ij}, for segment i with respect to segment j. By definition,$V_{ij} = 1$, if at least half of segment i intersects with segment j, and $V_{ij} = 0$ otherwise (more exactly, intersection of projections is considered). Formally, let L_i be the size of segment i, let M_j be the size of segment j, and let K_{ij} be the size of intersection of (projections of) segments i and j. Then,

$$\begin{aligned} V_{ij} &= 1, \quad \text{if } K_{ij}/L_i \geq 0.5, \\ V_{ij} &= 0, \quad \text{if } K_{ij}/L_j < 0.5, \end{aligned} \tag{2}$$

and, symmetrically, the local overlap of segment j with respect to segment i is

$$\begin{aligned} W_{ji} &= 1, \quad \text{if } K_{ij}/M_j \geq 0.5, \\ W_{ji} &= 0, \quad \text{if } K_{ij}/M_j < 0.5. \end{aligned} \tag{3}$$

The total segment consistency C is the sum of local overlaps for all pairs of TM-segments whose projections intersect:

$$C = \Sigma_{ij}(V_{ij} + W_{ji})/(n_1 + n_2). \tag{4}$$

For similar predictions, both Q and C are close to 1. If the same numbers of TM-segments are predicted in two aligned proteins and these segments roughly correspond to each other (have more than half of each segment in common), but positions of the TM-segment termini differ, then $Q < 1$, but still $C = 1$. On the other hand, if a single TM-segment in one protein often corresponds to two segments in the other protein (broken segments), $Q \approx 1$ and $C < 1$. Thus, these two indices capture two different aspects of the prediction consistency.

The following ten servers were considered: DAS, HMMTOP, PHDhtm, PRED-TMR, PSORT, SOSUI, TMAP, TMHMM 2.0, TMpred, and TopPred II. This choice depended on server's availability and popularity, possibility to make multiple queries, and possibility to make predictions for single proteins (not only multiple alignments).

8. Results and Discussion

Here we have analyzed ten advanced methods for the analysis of single sequences. The URLs of the considered servers are listed in Table 2. The average values of the consistency indices Q and C and standard deviations (σ) are listed in Table 4.

Table 4. The Q and C consistency indices and standard deviation (σ) in different ID intervals. Proteins are divided into several groups according to their identity level.

ID	40–49%	40–49%	50–59%	50–59%
	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$
PHDhtm	0.88 ± 0.12	0.94 ± 0.11	0.89 ± 0.12	0.94 ± 0.11
HMMTOP 2.0	0.73 ± 0.12	0.93 ± 0.12	0.76 ± 0.13	0.94 ± 0.13
TMHMM 2.0	0.72 ± 0.13	0.92 ± 0.12	0.76 ± 0.11	0.93 ± 0.12
TMpred	0.70 ± 0.10	0.91 ±0.08	0.74 ± 0.10	0.93 ± 0.08
TopPred II	0.71 ± 0.12	0.89 ± 0.10	0.76 ± 0.13	0.91 ± 0.09
PRED-TMR	0.69 ± 0.13	0.89 ± 0.12	0.73 ±0.13	0.90 ± 012
SOSUI	0.69 ± 0.11	0.88 ± 0.13	0.72 ± 0.11	0.89 ± 0.13
TMAP	0.64 ± 0.10	0.85 ± 0.11	0.67 ±0.11	0.87 ± 0.10
DAS	0.64 ± 0.11	0.83 ± 0.10	0.69 ± 0.10	0.87 ± 0.09
PSORT	0.63 ± 0.14	0.84 ± 0.14	0.69 ± 0.14	0.86 ± 0.14

The most consistent predictions were made by PHDhtm, closely followed by HMMTOP and TMHMM. It is instructive to compare these results with benchmarks that used test sets. Among six hydrophobicity scales and eight advanced methods considered in Ref. 127, the best results were demonstrated by TMHMM; the predictions by PHDhtm and HMMTOP were considerably worse.

Eight internet servers and 19 hydrophobicity scales were compared in Ref. 132. The results depend on the details of the procedure used to evaluate the predictions. HMMTOP and PHDhtm usually were among the leaders, whereas TMHMM produced average results.

TopPred and TMpred consistently occupy the middle of the list. This was observed in our study as well as Ref. 127. The last group of servers is PSORT, DAS, and TMAP. DAS demonstrated average performance in Refs. 127 and 132; TMAP was in the middle in Ref. 127; the remaining servers were not considered in the cited papers. The four methods shown to be consistently performing well in Ref. 140 are SPLIT4, TMHMM, HMMTOP, and TMAP.

Thus, we can see that the results obtained in several benchmark studies are largely similar, despite the use of different criteria. Most successful methods use the HMM technique. That seems to be the best technique for capturing various statistical properties of membrane-spanning proteins. An alternative, the use of the sliding windows approach, suffers from various problems ranging from insufficient specificity[127,134] to inexact definition of TM-segment boundaries and spurious merging of adjacent segments, requiring introduction of *ad hoc* fine-tuning procedures.[84]

One specific problem, not addressed here, is the initial identification of membrane-spanning proteins. Indeed, no single method but DAS could identify all such protein in Ref. 132; but it was rather non specific and predicted TM-segments in globular proteins.

Two more technical but important points, whose detailed analysis is beyond the scope of this study, are the fact that advantages and disadvantages of different methods balance each other, and thus averaging over methods improves the reliability of predictions.[134] Indeed, this observation has been exploited by a number of successful programs (Table 2).[102,104,118,141] However, these programs require reliable alignments, and the next generation of methods probably will combine multiple alignment and TM-segment identification using the HMM language common to these two problems.[142]

Finally, as signal peptides are often taken for TM-segments by the TM-recognition software, identification of TM-segments should be combined with the analysis of signal peptides either manually or algorithmically.[134,143]

Overall, if one single server is to be recommended, it is probably TMHMM followed by PHDhtm and HMMTOP. However, in important cases, it is clearly advisable to submit a query to several servers; and also to analyze not one protein, but an entire family of homologs.

Acknowledgments

We are grateful to Alexandra Rakhmaninova, Vsevolod Makeev, and Andrei Mironov for numerous useful discussions.

This study was partially supported by grants from the Howard Hughes Medical Institute (55005610), the Russian Academy of Sciences (programs "Molecular and Cellular biology"), The Russian Fund of Basic Research (04-04-49361), INTAS.

References

1. Wallin E, von Heijne G, Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Sci* **7**(4):1029–1038, 1998.
2. Mitaku S, Ono M, Hirokawa T, Boon-Chieng S, Sonoyama M, Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system, *Biophys Chem* **82**(2–3):165–171, 1999.
3. Liu J, Rost B, Comparing function and structure between entire proteomes, *Protein Sci* **10**(10):1970–1979, 2001.
4. Saier MH, Jr, A functional-phylogenetic system for the classification of transport proteins, *J Cell Biochem* **Suppl 32–33**:84–94, 1999.
5. Miller C, 1990: annus mirabilis of potassium channels, *Science* **252**(5010):1092–1096, 1991.
6. Barnard EA, Receptor classes and the transmitter-gated ion channels, *Trends Biochem Sci* **17**(10):368–374, 1992.
7. Stephenson FA, Ion channels, *Curr Opin Struct Biol* **1**:569–574, 1991.
8. Savarese TM, Fraser CM, In vitro mutagenesis and the search for structure-function relationships among G protein-coupled receptors, *Biochem J* **283 (Pt 1)**:1–19, 1992.
9. Capaldi RA, Structural features of the mitochondrial electron-transfer chain, *Curr Opin Struct Biol* **2**:511–518, 1991.
10. Jayasinghe S, Hristova K, White SH, MPtopo: A database of membrane protein topology, *Protein Sci* **10**(2):455–458, 2001.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank, *Nucleic Acids Res* **28**(1):235–242, 2000.
12. Tusnady GE, Dosztanyi Z, Simon I, Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics* **20**(17):2964–2972, 2004.

13. Tusnady GE, Dosztanyi Z, Simon I, PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank, *Nucleic Acids Res* **33** (Database issue):D275–D278, 2005.
14. Grisshammer R, Tate CG, Overexpression of integral membrane proteins for structural studies, *Q Rev Biophys* **28**(3):315–422, 1995.
15. Qutub Y, Reviakine I, Maxwell C, Navarro J, Landau EM, Vekilov PG, Crystallization of transmembrane proteins in cubo: mechanisms of crystal growth and defect formation, *J Mol Biol* **343**(5):1243–1254, 2004.
16. Seddon AM, Curnow P, Booth PJ, Membrane proteins, lipids and detergents: not just a soap opera, *Biochim Biophys Acta* **1666**(1–2):105–117, 2004.
17. von Heijne G, Principles of membrane protein assembly and structure, *Prog Biophys Mol Biol* **66**(2):113–139, 1996.
18. Rost B, Sander C, Prediction of protein secondary structure at better than 70% accuracy, *J Mol Biol* **232**(2):584–599, 1993.
19. Lewis MJ, Chang JA, Simoni RD, A topological analysis of subunit alpha from Escherichia coli F1F0-ATP synthase predicts eight transmembrane segments, *J Biol Chem* **265**(18):10541–10550, 1990.
20. Bjorbaek C, Foersom V, Michelsen O, The transmembrane topology of the a [corrected] subunit from the ATPase in Escherichia coli analyzed by PhoA protein fusions, *FEBS Lett* **260**(1): 31–34, 1990.
21. Esposti MD, De Vries S, Crimi M, Ghelli A, Patarnello T, Meyer A, Mitochondrial cytochrome b: evolution and structure of the protein, *Biochim Biophys Acta* **1143**(3):243–271, 1993.
22. Hucho F, Gorne-Tschelnokow U, Strecker A, Beta-structure in the membrane-spanning part of the nicotinic acetylcholine receptor (or how helical are transmembrane helices?), *Trends Biochem Sci* **19**(9):383–387, 1994.
23. Stokes DL, Taylor WR, Green NM, Structure, transmembrane topology and helix packing of P-type ion pumps, *FEBS Lett* **346**(1):32–38, 1994.
24. Persson B, Argos P, Prediction of transmembrane segments in proteins utilising multiple sequence alignments, *J Mol Biol* **237**(2):182–192, 1994.
25. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A, Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method, *Protein Eng* **10**(6):673–676, 1997.
26. Grigor'ev IV, Mironov AA, Rakhmaninova AB, [Refinement of helix boundaries in alpha-helical globular proteins], *Mol Biol (Mosk)* **33**(2):242–251, 1999.
27. Jones DT, Taylor WR, Thornton JM, A model recognition approach to the prediction of all-helical membrane protein structure and topology, *Biochemistry* **33**(10):3038–3049, 1994.
28. Jones DT, Do transmembrane protein superfolds exist?, *FEBS Lett* **423**(3):281–285, 1998.
29. Klein P, Kanehisa M, DeLisi C, The detection and classification of membrane-spanning proteins, *Biochim Biophys Acta* **815**(3):468–476, 1985.
30. Sipos L, von Heijne G, Predicting the topology of eukaryotic membrane proteins, *Eur J Biochem* **213**(3):1333–1340, 1993.
31. Boyd D, Manoil C, Beckwith J, Determinants of membrane protein topology, *Proc Natl Acad Sci USA* **84**(23):8525–8529, 1987.
32. von Heijne G, Gavel Y, Topogenic signals in integral membrane proteins, *Eur J Biochem* **174**(4):671–678, 1988.
33. Hartmann E, Rapoport TA, Lodish HF, Predicting the orientation of eukaryotic membrane-spanning proteins, *Proc Natl Acad Sci USA* **86**(15):5786–5790, 1989.

34. von Heijne G, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, *J Mol Biol* **225**(2):487–494, 1992.
35. Nakashima H, Nishikawa K, The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins, *FEBS Lett* **303**(2–3):141–146, 1992.
36. Nakai K, Kanehisa M, A knowledge base for predicting protein localization sites in eukaryotic cells, *Genomics* **14**(4):897–911, 1992.
37. Hirokawa T, Boon-Chieng S, Mitaku S, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics* **14**(4):378–379, 1998.
38. Mitaku S, Hirokawa T, Tsuji T, Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces, *Bioinformatics* **18**(4):608–616, 2002.
39. Gafvelin G, Sakaguchi M, Andersson H, von Heijne G, Topological rules for membrane protein assembly in eukaryotic cells, *J Biol Chem* **272**(10):6119–6127, 1997.
40. Liakopoulos TD, Pasquier C, Hamodrakas SJ, A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm, *Protein Eng* **14**(6):387–390, 2001.
41. Nielsen H, Engelbrecht J, Brunak S, von Heijne G, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Eng* **10**(1): 1–6, 1997.
42. Pascarella S, Bossa F, CLEAVAGE: a microcomputer program for predicting signal sequence cleavage sites, *Comput Appl Biosci* **5**(1):53–54, 1989.
43. Bendtsen JD, Nielsen H, von Heijne G, Brunak S, Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol* **340**(4):783–795, 2004.
44. Hiller K, Grote A, Scheer M, Munch R, Jahn D, PrediSi: prediction of signal peptides and their cleavage positions, *Nucleic Acids Res* **32** (Web Server issue):W375–W379, 2004.
45. Popowicz AM, Dash PF, SIGSEQ: a computer program for predicting signal sequence cleavage sites, *Comput Appl Biosci* **4**(3):405–406, 1988.
46. Fariselli P, Finocchiaro G, Casadio R, SPEPlip: the detection of signal peptide and lipoprotein cleavage sites, *Bioinformatics* **19**(18):2498–2499, 2003.
47. Landolt-Marticorena C, Williams KA, Deber CM, Reithmeier RA, Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins, *J Mol Biol* **229**(3):602–608, 1993.
48. Cowan SW, Rosenbusch JP, Folding pattern diversity of integral membrane proteins, *Science* **264**(5161):914–916, 1994.
49. Gromiha MM, Ponnuswamy PK, Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins, *Int J Pept Protein Res* **48**(5):452–460, 1996.
50. Fasman GD, Protein conformational prediction, *Trends Biochem Sci* **14**(7):295–299, 1989.
51. Gromiha MM, Ponnuswamy PK, Prediction of protein secondary structures from their hydrophobic characteristics, *Int J Pept Protein Res* **45**(3):225–240, 1995.
52. Paul C, Rosenbusch JP, Folding patterns of porin and bacteriorhodopsin, *Embo J* **4**(6):1593–1597, 1985.
53. Vogel H, Jahnig F, Models for the structure of outer-membrane proteins of Escherichia coli derived from raman spectroscopy and prediction methods, *J Mol Biol* **190**(2):191–199, 1986.
54. Jahnig F, Structure predictions of membrane proteins are not that bad, *Trends Biochem Sci* **15**(3):93–95, 1990.

55. Welte W, Weiss MS, Nestel U, Weckesser J, Schiltz E, Schulz GE, Prediction of the general structure of OmpF and PhoE from the sequence and structure of porin from Rhodobacter capsulatus. Orientation of porin in the membrane, *Biochim Biophys Acta* **1080**(3):271–274, 1991.
56. Nakai K, Kanehisa M, Expert system for predicting protein localization sites in gram-negative bacteria, *Proteins* **11**(2):95–110, 1991.
57. Schirmer T, Cowan SW, Prediction of membrane-spanning beta-strands and its application to maltoporin, *Protein Sci* **2**(8):1361–1363, 1993.
58. Wimley WC, Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures, *Protein Sci* **11**(2):301–312, 2002.
59. Fariselli P, Finelli M, Rossi I, Amico M, Zauli A, Martelli PL, Casadio R, TRAMPLE: the transmembrane protein labelling environment, *Nucleic Acids Res* **33** (Web Server issue):W198–W201, 2005.
60. Gromiha MM, Ahmad S, Suwa M, Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins, *J Comput Chem* **25**(5):762–767, 2004.
61. Diederichs K, Freigang, J, Umhau S, Zeth K, Breed J, Prediction by a neural network of outer membrane beta-strand protein topology, *Protein Sci* **7**(11):2413–2420, 1998.
62. Natt NK, Kaur H, Raghava GP, Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods, *Proteins* **56**(1):11–18, 2004.
63. Kaur H, Raghava GP, Prediction of beta-turns in proteins from multiple alignment using neural network, *Protein Sci* **12**(3):627–634, 2003.
64. Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ, PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins, *Nucleic Acids Res* **32** (Web Server issue):W400–W404, 2004.
65. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B, Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res* **32**(8):2566–2577, 2004.
66. Jacoboni I, Martelli PL, Fariselli P, De Pinto V, Casadio R, Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor, *Protein Sci* **10**(4):779–787, 2001.
67. Park KJ, Gromiha MM, Horton P, Suwa M, Discrimination of outer membrane proteins using support vector machines, *Bioinformatics* **21**(23):4223–4229, 2005.
68. Beagos PG, Liakopoulos TD, Hamodrakas SJ, Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, *BMC Bioinformatics* **6**:7, 2005.
69. Martelli PL, Fariselli P, Krogh A, Casadio R, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins, *Bioinformatics* **18 Suppl 1**:S46–S53, 2002.
70. Kyte J, Doolittle RF, A simple method for displaying the hydropathic character of a protein, *J Mol Biol* **157**(1):105–132, 1982.
71. Eisenberg D, Schwarz E, Komaromy M, Wall R, Analysis of membrane and surface protein sequences with the hydrophobic moment plot, *J Mol Biol* **179**(1):125–142, 1984.
72. Eisenberg D, McLachlan AD, Solvation energy in protein folding and binding, *Nature* **319**(6050):199–203, 1986.
73. Engelman DM, Steitz TA, Goldman A, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Biophys Chem* **15**:321–353, 1986.

74. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, *J Mol Biol* **195**(3):659–685, 1987.
75. Degli Esposti M, Crimi M, Venturoli G, A critical evaluation of the hydropathy profile of membrane proteins, *Eur J Biochem* **190**(1):207–219, 1990.
76. Ponnuswamy PK, Gromiha MM, Prediction of transmembrane helices from hydrophobic characteristics of proteins, *Int J Pept Protein Res* **42**(4):326–341, 1993.
77. Cserzo M, Bernassau JM, Simon I, Maigret B, New alignment strategy for transmembrane proteins, *J Mol Biol* **243**(3):388–396, 1994.
78. Samatey FA, Xu C, Popot JL, On the distribution of amino acid residues in transmembrane alpha-helix bundles, *Proc Natl Acad Sci USA* **92**(10):4577–4581, 1995.
79. Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ, A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm, *Protein Eng* **12**(5):381–385, 1999.
80. Juretic D, Jeroncic A, D, Z, Sequence analysis of membrane proteins with the web server SPLIT, *Croatica Chemica Acta* **72**(4):975–997, 1999.
81. Argos P, Rao JK, Prediction of protein structure, *Methods Enzymol* **130**:185–207, 1986.
82. Nakai K, Horton P, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem Sci* **24**(1):34–36, 1999.
83. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S, Methods and algorithms for statistical analysis of protein sequences, *Proc Natl Acad Sci USA* **89**(6):2002–2006, 1992.
84. Rost B, Casadio R, Fariselli P, Sander C, Transmembrane helices predicted at 95% accuracy, *Protein Sci* **4**(3):521–533, 1995.
85. Lohmann R, Schneider G, Behrens D, Wrede P, A neural network model for the prediction of membrane-spanning amino acid sequences, *Protein Sci* **3**(9):1597–1601, 1994.
86. Rost B, Fariselli P, Casadio R, Topology prediction for helical transmembrane proteins at 86% accuracy, *Protein Sci* **5**(8):1704–1718, 1996.
87. Fariselli P, Finelli M, Marchignoli D, Martelli PL, Rossi I, Casadio R, MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments, *Bioinformatics* **19**(4):500–505, 2003.
88. Tusnady GE, Simon I, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, *J Mol Biol* **283**(2):489–506, 1998.
89. Krogh A, Larsson B, von Heijne G, Sonnhammer EL, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol* **305**(3):567–580, 2001.
90. Aloy P, Cedano J, Oliva B, Aviles FX, Querol E, 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins, *Comput Appl Biosci* **13**(3):231–234, 1997.
91. Persson B, Argos P, Topology prediction of membrane proteins, *Protein Sci* **5**(2):363–371, 1996.
92. Fariselli P, Casadio R, HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins, *Comput Appl Biosci* **12**(1):41–48, 1996.
93. Juretic D, Zoranic L, Zucic D, Basic charge clusters and predictions of membrane protein topology, *J Chem Inf Comput Sci* **42**(3):620–632, 2002.

94. Tusnady GE, Simon I, The HMMTOP transmembrane topology prediction server, *Bioinformatics* **17**(9):849–850, 2001.
95. Claros MG, von Heijne G, TopPred II: an improved software for membrane protein structure predictions, *Comput Appl Biosci* **10**(6):685–686, 1994.
96. Horton P, Nakai K, Better prediction of protein cellular localization sites with the k nearest neighbors classifier, *Proc Int Conf Intell Syst Mol Biol* **5**:147–152, 1997.
97. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S, Extensive feature detection of N-terminal protein sorting signals, *Bioinformatics* **18**(2):298–305, 2002.
98. Gardy JL, Spencer C, Wang K, Ester M, Tusndy GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS, PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria, *Nucleic Acids Res* **31**(13):3613–3617, 2003.
99. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS, PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis, *Bioinformatics* **21**(5):617–623, 2005.
100. Sonnhammer EL, von Heijne G, Krogh A, A hidden Markov model for predicting transmembrane helices in protein sequences, *Proc Int Conf Intell Syst Mol Biol* **6**:175–182, 1998.
101. McGuffin LJ, Bryson K, Jones DT, The PSIPRED protein structure prediction server, *Bioinformatics* **16**(4):404–405, 2000.
102. Promponas VJ, Palaios GA, Pasquier CM, Hamodrakas JS, Hamodrakas SJ, CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods, *In Silico Biol* **1**(3):159–162, 1999.
103. Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, Shimizu T, ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability, *Nucleic Acids Res* **32** (Web Server issue):W390–W393, 2004.
104. Taylor PD, Attwood TK, Flower DR, BPROMPT: A consensus server for membrane protein prediction, *Nucleic Acids Res* **31**(13):3698–3700, 2003.
105. Hofmann K, Stoffel W, TMBASE — a database of membrane spanning protein segments, *Biol Chem Hoppe-Seyler* **374**:166, 1993.
106. Milpetz F, Argos P, Persson B, TMAP: a new email and WWW service for membrane-protein structural predictions, *Trends Biochem Sci* **20**(5):204–205, 1995.
107. Persson B, Argos P, Prediction of membrane protein topology utilizing multiple sequence alignments, *J Protein Chem* **16**(5):453–457: 1997.
108. Holland IB, Blight MA, ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans, *J Mol Biol* **293**(2):381–399, 1999.
109. Kalinina OV, Makeev VJ, Sutormin RA, Gelfand MS, Rakhmaninova AB, The channel in transporters is formed by residues that are rare in transmembrane helices, *In Silico Biol* **3**(1–2):197–204, 2003.
110. Baldwin JM, The probable arrangement of the helices in G protein-coupled receptors, *Embo J* **12**(4):1693–1703, 1993.
111. Suwa M, Hirokawa T, Mitaku S, A continuum theory for the prediction of lateral and rotational positioning of alpha-helices in membrane proteins: bacteriorhodopsin, *Proteins* **22**(4):363–377, 1995.
112. Efremov RG, Vergoten G, Hydrophobic organization of alpha-helix membrane bundle in bacteriorhodopsin, *J Protein Chem* **15**(1):63–76, 1996.
113. Lemmon MA, Engelman DM, Specificity and promiscuity in membrane helix interactions, *FEBS Lett* **346**(1):17–20, 1994.

114. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM, Sequence specificity in the dimerization of transmembrane alpha-helices, *Biochemistry* **31**(51): 12719–12725, 1992.
115. Lemmon MA, Treutlein HR, Adams PD, Brunger AT, Engelman DM, A dimerization motif for transmembrane alpha-helices, *Nat Struct Biol* **1**(3):157–163, 1994.
116. Sutormin RA, Rakhmaninova AB, Gelfand MS, BATMAS30: amino acid substitution matrix for alignment of bacterial transporters, *Proteins* **51**(1):85–95, 2003.
117. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**(17):3389–3402, 1997.
118. Ikeda M, Arai M, Lao DM, Shimizu T, Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, *In Silico Biol* **2**(1):19–33, 2002.
119. Gavel Y, von Heijne G, The distribution of charged amino acids in mitochondrial inner-membrane proteins suggests different modes of membrane integration for nuclearly and mitochondrially encoded proteins, *Eur J Biochem* **205**(3):1207–1215, 1992.
120. Gavel Y, Steppuhn J, Herrmann R, von Heijne G, The 'positive-inside rule' applies to thylakoid membrane proteins, *FEBS Lett* **282**(1):41–46, 1991.
121. Durbin R, Eddy SR, Krogh A, G, M, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, *Cambridge University Press*, 1999.
122. Boyd D, Schierle C, Beckwith J, How many membrane proteins are there?, *Protein Sci* **7**(1):201–205, 1998.
123. Frishman D, Mewes HW, Protein structural classes in five complete genomes, *Nat Struct Biol* **4**(8):626–628, 1997.
124. Arkin IT, Brunger AT, Engelman DM, Are there dominant membrane protein families with a given number of helices?, *Proteins* **28**(4):465–466, 1997.
125. Gerstein M, Hegyi H, Comparing genomes in terms of protein structure: surveys of a finite parts list, *FEMS Microbiol Rev* **22**(4):277–304, 1998.
126. Iyengar R, Teaching resources. Structure of G-protein-coupled receptors and G proteins, *Sci STKE* **2005**(276): tr10 (2005).
127. Moller S, Croning MD, Apweiler R, Evaluation of methods for the prediction of membrane spanning regions, *Bioinformatics* **17**(7):646–653, 2001.
128. Eisenberg D, Weiss RM, Terwilliger TC, The helical hydrophobic moment: a measure of the amphiphilicity of a helix, *Nature* **299**(5881):371–374, 1982.
129. Rost B, Casadio R, Fariselli P, Refining neural network predictions for helical transmembrane proteins by dynamic programming, *Proc Int Conf Intell Syst Mol Biol* **4**:192–200, 1996.
130. Moller S, Kriventseva EV, Apweiler R, A collection of well characterised integral membrane proteins, *Bioinformatics* **16**(12):1159–1160, 2000.
131. Kernytsky A, Rost B, Static benchmarking of membrane helix predictions, *Nucleic Acids Res* **31**(13):3642–3644, 2003.
132. Chen CP, Kernytsky A, Rost B, Transmembrane helix predictions revisited, *Protein Sci* **11**(12):2774–2791, 2002.
133. Chen CP, Rost B, Long membrane helices and short loops predicted less accurately, *Protein Sci* **11**(12):2766–2773, 2002.
134. Chen CP, Rost B, State-of-the-art in membrane protein prediction, *Appl Bioinformatics* **1**(1):21–35, 2002.
135. Busch W, Saier MH, Jr, The IUBMB-endorsed transporter classification system, *Mol Biotechnol* **27**(3):253–262, 2004.

136. Ren Q, Kang KH, Paulsen IT, TransportDB: a relational database of cellular membrane transport systems, *Nucleic Acids Res* **32** (Database issue): D284–D288, 2004.
137. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr, Kyrpides N, Fonstein M, Maltsev N, Selkov, E, WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction, *Nucleic Acids Res* **28**(1):123–125, 2000.
138. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**(3):403–410, 1990.
139. Thompson JD, Higgins DG, Gibson TJ, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**(22):4673–4680, 1994.
140. Cuthbertson JM, Doyle DA, Sansom MS, Transmembrane helix prediction: a comparative evaluation and analysis, *Protein Eng Des Sel* **18**(6): 295–308, 2005.
141. Nilsson J, Persson B, von Heijne G, Consensus predictions of membrane protein topology, *FEBS Lett* **486**(3):267–269, 2000.
142. Sutormin RA, Mironov AA, [Membrane probability profile construction based on amino acids sequences multiple alignment], *Mol Biol (Mosk)* **40**(3):1–5, 2006.
143. Lao DM, Okuno T, Shimizu T, Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction, *In Silico Biol* **2**(4):485–494, 2002.

Natalya S. Sadovskaya is a junior fellow researcher in the Laboratory of Bioinformatics, State Scientific Center GosNIIGenetika, Moscow, Russia. She obtained her M.Sc. degree from the Department of Molecular and Biology Physics, Moscow Institute of Physics and Technology (State University) in 2000. Her research interests are the evolution of metabolic pathways and secondary structure of membrane proteins.

Roman A. Sutormin graduated as M.Sc. from the Department of Mechanics and Mathematics, Moscow State University, Russia. He is currently a Ph.D. student at the State Scientific Center of Genetics and Selection and programming teacher at the Department of Bioengineering and Bioinformatics, Moscow State University. His research interests include statistical properties of amino acid sequences of membrane proteins and their evolution, prediction of the secondary structure of TM-proteins both from single sequence and using multiple alignment information, development of the alignment techniques in case of membrane proteins.

Mikhail S. Gelfand is the Head of the Research and Training Center in Bioinformatics of the Institute for Information Transmission Problems, RAS in Moscow, Russia and a Professor at the Department of Bioengineering and Bioinformatics of the Moscow State University. He graduated from the Department of Mathematics of the Moscow State University, received his Ph.D. (Math.) degree from the Institute of Theoretical and Experimental Biophysics, RAS (Pushchino), and the Doctor of Sciences degree from the State Research Institute for Genetics and Selection of Industrial Microorganisms (Moscow). He is a member of editorial boards of several journals, in particular, “PLoS Biology”, “Bioinformatics”, “BMC Bioinformatics”, “Journal of Bioinformatics and Computational Biology,” and “Journal of Computational Biology”. He received the Baev prize (1999) from the Russian State “Human Genome” Council, and Best Scientist of the Russian Academy of Sciences” award (2004). His research interests include comparative genomics, metabolic reconstruction and modeling, evolution of metabolic pathways and regulatory systems, function and evolution alternative splicing, functional annotation of genes and regulatory signals.