Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Proteins. 2016 November ; 84(11): 1706–1716. doi:10.1002/prot.25155.

TMSEG: novel prediction of transmembrane helices

Michael Bernhofer[1,*], **Edda Kloppmann**[1,2], **Jonas Reeb**[1], and **Burkhard Rost**[1,2,3,4]

[1] Department of Informatics & Center for Bioinformatics & Computational Biology – i12, Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching/Munich, Germany

[2] New York Consortium on Membrane Protein Structure, New York Structural Biology Center, 89 Convent Avenue, New York, NY 10027

[3] Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany

[4] Institute for Food and Plant Sciences WZW – Weihenstephan, Alte Akademie 8, Freising, Germany

Abstract

Transmembrane proteins (TMPs) are important drug targets because they are essential for signaling, regulation, and transport. Despite important breakthroughs, experimental structure determination remains challenging for TMPs. Various methods have bridged the gap by predicting transmembrane helices (TMHs), but room for improvement remains. Here, we present TMSEG, a novel method identifying TMPs and accurately predicting their TMHs and their topology. The method combines machine learning with empirical filters. Testing it on a non-redundant dataset of 41 TMPs and 285 soluble proteins, and applying strict performance measures, TMSEG outperformed the state-of-the-art in our hands. TMSEG correctly distinguished helical TMPs from other proteins with a sensitivity of 98±2% and a false positive rate as low as 3±1%. Individual TMHs were predicted with a precision of 87±3% and recall of 84±3%. Furthermore, in 63±6% of helical TMPs the placement of all TMHs and their inside/outside topology was correctly predicted. There are two main features that distinguish TMSEG from other methods. First, the errors in finding all helical TMPs in an organism are significantly reduced. For example, in human this leads to 200 and 1600 fewer misclassifications compared to the 2nd and 3rd best method available, and 4400 fewer mistakes than by a simple hydrophobicity-based method. Second, TMSEG provides an add-on improvement for any existing method to benefit from.

Keywords

membrane protein; protein structure prediction; transmembrane helices; α-helical integral membrane protein; transmembrane protein prediction; transmembrane helix prediction

Introduction

Transmembrane proteins (TMPs) are involved in numerous essential processes within living organisms such as signaling, regulation, and transport[1]. About 20-30% of all proteins within

* Corresponding author: Michael.Bernhofer@mytum.de.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

any organism have been estimated to be TMPs[2,3]. Many TMPs, especially G protein-coupled receptors (GPCRs), are primary drug targets[4] and therefore of high interest.

TMPs cross the membrane bilayer with either transmembrane helices (TMHs) or beta-strands. The latter are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. They make up only about 1-2% of all proteins in Gram-negative bacteria[5]. We concentrated on the more common class of helical TMPs and will refer to these as TMPs in the following. TMPs can cross the membrane only once (single-pass) or multiple times (multi-pass). Due to the apolar and hydrophobic environment in the lipid bilayer, most of the amino acids found in TMHs are hydrophobic, and their orientation in the membrane (called TMP topology), can be discerned through Gunnar von Heijne's positive-inside rule[6,7].

Despite their immense importance, and despite crucial experimental advances[8-11], less than 2% of the structures in the Protein Data Bank[12] (PDB) are TMPs[13-15]. As membrane regions are typically not visible in high-resolution structures, TMHs are assigned to PDB structures by expert resources, most prominently the Orientations of Proteins in Membranes[16] (OPM) database and the Protein Data Bank of Transmembrane Proteins[17] (PDBTM).

Recent advances in experimental structure determination have benefited from advanced computational predictions of TMHs from sequence[8,9]. In the last 25 years, many such tools have been developed, ranging from simple algorithms based solely on hydrophobicity scales (*e.g.* TopPred[18]) to sophisticated uses of hidden Markov models (*e.g.* TMHMM[19], HMMTOP[20], Phobius[21], and PolyPhobius[22]), neural networks (*e.g.* PHDhtm[23,24], and MEMSAT3[25]), and support vector machines (MEMSAT-SVM[26]). Arguably, the most important advance was the incorporation of evolutionary information from sequence profiles or multiple sequence alignments[23,24]. Consequently, almost all methods developed over the last decade are based on evolutionary information. A recent assessment applying strict evaluation measures showed that many methods perform well overall; the best are some recent methods[27]. Here, we show that a few simple ideas improve significantly over the state-of-the-art.

Material and Methods

Dataset TMP166: helical TMPs with known structures

We collected helical TMPs with known structures annotated in OPM[16] and PDBTM[17] (releases 2013_07). Both databases use PDB[12] chain identifiers. We mapped those PDB chains to their UniProtKB[28] protein sequences using SIFTS[29]. We excluded all chimeric PDB chains, model structures, X-ray structures with >8Å, and those for which some TMH residues did not map gapless to UniProtKB sequences. This gave 1087 PDB chains from 455 PDB structures (379 X-ray and 76 NMR structures).

UniqueProt[30] reduced sequence-redundancy at HVAL>0 (the HVAL depends on alignment length and the percentage of pairwise sequence identity[31]). At this threshold no pair of proteins has more than 20% pairwise sequence identity for alignments of more than 250 residues (see Rost 1999[32] for precise definitions). The result of this is our final data set consisting of 166 non-redundant TMPs (called TMP166, Table S1, Supporting Online Material: SOM).

As the TMH annotations in OPM and PDBTM differed for some proteins, we associated TMH annotations from both databases with each sequence. The inside/outside topology of the non-transmembrane regions was assigned based on the ATOM coordinates and topology annotation from OPM (*cf.* Note S1 and Fig. S1, SOM). We considered re-entrant regions[33,34] to be non-transmembrane due to their scarcity in the TMP166 dataset (only 15 proteins with one or two re-entry regions each; Table S1, SOM).

Dataset SP1441: proteins with and without signal peptides

As signal peptides are often confused with TMHs and *vice versa*[27], a second dataset was derived from the SignalP4.1 dataset[35]. This dataset contained UniProtKB sequences of soluble proteins and TMPs with and without signal peptide annotations. Note that these TMPs have no inside/outside topology annotations and many of their TMH annotations are not supported by experimental evidence.

The SignalP4.1 dataset was redundancy reduced twice using UniqueProt. First, all proteins similar to any of those in the TMP166 dataset were removed at HVAL>0. Second, the remaining proteins were redundancy-filtered at HVAL>0. The final dataset contained 1441 proteins sequences (299 TMPs and 1142 soluble proteins, called SP1441; Table S2, SOM). 477 of those had signal peptide annotations (25 TMPs and 452 soluble proteins).

Splitting the datasets

We split the combined TMP166 and SP1441 dataset into four subsets. We partitioned them in a way that all subsets have approximately the same distributions with respect to the number of soluble proteins and TMPs, protein sequences with and without signal peptides, and sequence lengths (Fig. S2, SOM).

We used the first three subsets to develop TMSEG in a three-fold cross-validation approach (*cf.* TMSEG training). The fourth split, the independent test set called BlindTest, was used only for the final performance evaluation, *i.e.* no parameter was optimized on that set. The BlindTest dataset contained 41 TMPs (from TMP166) with known structure and TMH annotations from OPM and PDBTM, and 285 soluble proteins from the SP1441 dataset. The 74 TMPs from the 4th split of SP1441 (Table S2) were not included in the BlindTest dataset, because they lack sufficient experimental annotations. However, we used them for the signal peptide prediction performance analysis as we did not have curated signal peptide annotations for the TMPs from OPM and PDBTM.

Human proteome

We retrieved the human proteome, 20,196 protein sequences, from UniProtKB/Swiss-Prot (release 2015_03). We applied our TMSEG algorithm to the whole proteome to provide a summary of its TMP composition and to estimate run time.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dataset New12

Our original data sets had been based on the PDB release from July 2013, when this work began. Shortly before submission of the work in February 2016, *i.e.* 32 months later, we retrieved all TMPs added to OPM and PDBTM since July 2013. We removed all TMPs similar (HVAL>0) to proteins in data sets used previously (TMP166 and SP1441). Testing the pairwise similarity of the remaining TMPs we found that two pairs were similar (HVAL>0), but we decided to keep them due to their low HVAL. This resulted in 12 new TMPs (New12 dataset, Table S3, SOM) we used for additional testing. Although the statistical power of such a small set is very limited, these 12 constitute the entire addition of completely new structures from 2013/07 to 2016/02. Further, these or structurally related TMPs have most likely not been used to develop any method used for comparison.

Evaluation

As per-protein scores (correct classification as TMP or non-TMP), we compiled the sensitivity (percentage of observed TMPs predicted as TMPs) and the false positive rate (FPR: percentage of soluble proteins predicted as TMPs, Table 1). As per-TMH scores (correct identification and placement of TMHs), we compiled the precision (percentage of predicted TMHs that are correct), recall (percentage of observed TMHs predicted as TMHs), Q_{ok} and $Q_{top.}$ Q_{ok} is the percentage of TMPs for which all TMHs are correctly predicted (Table 1). Q_{top} requires in addition to Q_{ok} correct topology predictions (in/out: Table 1). To resolve conflicts between OPM and PDBTM annotations, we chose whichever fit the prediction best. Note that while sensitivity and recall have the same formula, we used sensitivity in conjunction with TMPs and recall with TMHs to better distinguish between those scores in the text.

Each TMH was considered correctly predicted, if predicted and observed TMH ends were within five residues (Fig. S3, SOM), and if predicted and observed TMH overlapped by at least half of the length of the longer of the two helices. These two criteria are more stringent than those that have commonly been used (typically: overlap >3-5 residues anywhere between observed and predicted TMH[36]) and have recently led to re-evaluating TMH prediction methods[27]. None of our major conclusions changed upon applying values slightly different than five residues for the maximum allowed discrepancy between predicted and observed TMH ends (data not shown).

Error rates for the evaluation measures were estimated by bootstrapping[37], *i.e.* by re-sampling the population of proteins used for the evaluation 1000 times and calculating the sample standard deviation. Each of these sample populations contained 60% of the original proteins (picked randomly without replacement).

State-of-the-art methods

We compared TMSEG to the best methods[27], namely to PolyPhobius[22], MEMSAT3[25], and MEMSAT-SVM[26]. Like TMSEG, these methods also use evolutionary information to predict TMPs: MEMSAT3 and MEMSAT-SVM automatically generate position-specific scoring matrices (PSSMs) with PSI-BLAST, while PolyPhobius generates multiple sequence alignments (MSAs). To ensure equal conditions for all methods we ran them on our local machines and used the UniProt Reference Cluster with 90% sequence identity (UniRef90, release 2015_03) as the homology search database, *i.e.* to generate the MSAs or PSSMs. While we used proteins completely unknown to TMSEG to assess its performance, some of the proteins used in our assessment might have been used to develop PolyPhobius, MEMSAT3, or MEMSAT-SVM. In this sense, our assessment was likely to over-estimate their performance, in particular with respect to TMSEG.

Baseline performance

We also compared all methods to a simple baseline predictor similar to TopPred[18]: for all possible segments of 21 consecutive residues we summed the Eisenberg-hydrophobicity[38] (EisenbergSum, Table S4, SOM). All non-overlapping segments with EisenbergSum≥4 were predicted as TMHs, starting with the segments with the highest sum. The inside/outside topology was predicted based on the difference between arginine and lysine residues on either side of the TMHs, *i.e.* applying Gunnar von Heijne's positive-inside rule[6,7].

TMSEG input/output

TMSEG needs two input files to successfully run a prediction: a FASTA file with the protein sequence and a PSI-BLAST PSSM file for the input protein. The PSSM file is mandatory and used to include homology-based features that greatly increase the prediction accuracy.

Combining evolutionary information (*e.g.* PSSMs and MSAs) with machine learning has been the most important improvement in protein prediction and is commonly used in TMH and secondary structure prediction[24,27,39,40]. TMSEG incorporates evolutionary information through PSI-BLAST profiles[41] generated from UniRef90 (release 2015_03). We used two sets of profiles: a training set with a stringent E-value cutoff of 10^{-5} and five iterations for creating the profile, as well as a test set with a less strict E-value cutoff of 10^{-3} and three iterations. We deactivated PSI-BLAST's low-complexity filter and enabled the option to calculate local optimal Smith-Waterman alignments in order to generate longer and more accurate alignments.

In addition, we used biophysical properties (charge, hydrophobicity, polarity; Table S4, SOM) and the overall amino acid composition. These features were calculated twice for each residue: once for all substitutions with a positive PSSM score and once based on all substitutions with a negative score.

The standard output gives a brief summary of the positions of the TMHs and signal peptide (if any) and the inside/outside topology. In addition, a raw output is available that also contains the unmodified output probabilities of the machine-learning tools.

TMSEG algorithm

TMSEG combines several machine-learning tools and empirical filters. The machine-learning algorithms used are two random forests (RFs) and one neural network (NN), both of which are implementations from the WEKA Java package[42]. The output of these algorithms is further processed with empirically determined filters and thresholds. The TMSEG algorithm executes four separate steps (Fig. 1):

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Step 1: Initial per-residue prediction

An RF detects TMHs from the input sequence. This RF slides a window of 19 consecutive residues through the protein sequence, predicting whether or not the central residue in the window is in a TMH, signal peptide, or non-TM region, *i.e.* the probability of each residue for each state is calculated based on the residue itself and the nine residues left and right of it. For each of the 19 residue positions, we compute the PSSM profile. For the central nine residues in the window, we also compute the average Kyte-Doolittle[43] hydrophobicity, and the percentage of hydrophobic, charged, and polar residues (Table S4, SOM).

In addition to these local features, we compile global features: the distance of the residue to the N- and C-terminus, the length of the protein sequence, and the global amino acid composition. The RF assigns three values to each residue corresponding to the probability to be in a TMH, a signal peptide, or a non-TM region. Runtime is decreased by multiplication of the probabilities by 1000 and transformation into integers.

Step 2: Per-protein filter: TMP or soluble

The per-residue scores are filtered empirically. First to reduce short peaks of one or two residues, all per-residue scores are smoothed by compiling the median score over five consecutive residues and assigning it to the center residue. Next, each residue is assigned to the state with the highest score (TMH, signal peptide, or non-TM). To prevent over-prediction due to the under-sampling of signal peptide residues, we applied a penalty of 185 (*i.e.* 18.5%) to non-TM and 60 *(i.e.* 6%) to TMH residues. These penalties were optimized during cross-training to best balance over- and under-prediction. Finally, TMHs shorter than seven residues are changed into non-TM regions. If a signal peptide of at least four consecutive residues is identified within the first 40 N-terminal residues ending in residue at position i, TMSEG predicts a signal peptide from residue 1 to residue i ($i \leq 40$). Signal peptide predictions outside the first 40 residues (i>40) are changed into non-TM, but do not invalidate signal peptides inside the first 40 residues. Initial predictions with fewer than four consecutive residues are changed into non-TM.

Step 3: Refinement of TMHs

In the third step an NN corrects the predicted TMHs. In contrast to the standard sliding window approach of the RF in Step 1, here we introduced a segment-based solution that used as input the following averages over the predicted TMHs: length of predicted TMH, amino acid composition, average hydrophobicity, as well as the percentages of hydrophobic and charged residues. The output of the NN is the predicted probability for the segment to be a TMH. Based on this probability, the predicted TMHs from Step 2 are adjusted.

First, TMHs≥35 residues are split into two TMHs with at least 17 residues, if these two TMHs increase the overall probability. The minimum length of 35 residues for splitting long TMHs and of 17 residues for the resulting two TMHs were empirically chosen based on the overall performance during cross-training. Second, the start and end positions for each TMH are adjusted by shifting them by up to three residues in either direction. Shifts are accepted if they increase the overall probability. The maximum endpoint adjustment by three residues was empirically chosen based on the overall performance during cross-training. In addition,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the relatively long minimum TMH lengths to allow splitting and the relatively small shift of maximally three residues of the TMH ends allow TMSEG to maintain a short runtime.

Step 4: topology prediction

Another RF predicts the inside/outside topology of the TMP, *i.e.* in which direction the TMHs cross the membrane. During this step the non-transmembrane regions are assigned to inside (*e.g.* cytoplasmic side of the membrane) or outside. This prediction is made for the entire protein. For each TMH, we consider up to 15 residues before and after the TMH, and eight residues at the TMH start and end (for TMHs<16 these residues overlap). As all predicted TMHs are assumed to cross the membrane, the in/out assignment is switched after each TMH. For each side, we compute as input to the RF the amino acid composition, the percentage of positively charged residues (we consider all arginine and lysine residues), and the absolute difference of positively charged residues between the two sides. Based on the RF output, one side is assigned to be inside (*e.g.* cytoplasmic), the other to be outside. Residues immediately after predicted signal peptides are assigned to outside (non-cytoplasmic) and all consecutive segments are assigned accordingly without any further prediction.

TMSEG training

To reduce the risk of over-fitting, we split our combined TMP166 and SP1441 datasets into four even splits (*cf.* Table S1 and S2). Note that the TMPs from the SP1441 dataset were used to train the random forest in the initial prediction (step 1) as they contain signal peptide annotations. They are, however, not used for the neural network (step 3) or the random forest in step 4, since they have no inside/outside topology annotations and many of their TMH annotations are not supported by experimental evidence.

The first of three splits was used to train, the second to cross-train, *i.e.* to optimize all other free parameters (*e.g.* the minimum TMH length), and the last to evaluate performance (test). This procedure was repeated three times, such that each protein had been used exactly once for training, cross-training and testing. The final parameters were frozen according to the overall best performance for all three rotations (on the test set). Given the frozen parameters, we applied the final method to the fourth split, the BlindTest dataset, which had not been used before.

Our careful four-fold split leading to three-fold development (each with training, cross-training, and testing), provided a double protection against overestimating performance. We decided about every detail in the final method before using the BlindTest dataset to evaluate TMSEG as presented here. Many developers use a two-fold split (training/testing), more careful ones the three-fold split (training/cross-training/testing), while the fourth split is occasionally introduced through pre-release data[39] like the New12 dataset that we generated.

Results and Discussion

The novel TMSEG method introduced here distinguishes between proteins with transmembrane helices (TMHs) and soluble proteins. For all helical transmembrane proteins (TMPs), it predicts the placement of the TMHs, and their orientation in the membrane, *i.e.*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

their inside/outside topology. We established sustained performance through cross-validation with two levels of blind testing. We compared our new methods to others, including the best at predicting TMPs[27], namely PolyPhobius[22] and MEMSAT-SVM[26]. Furthermore, we analyzed MEMSAT3[25] because it excels at the inside/outside topology prediction[44], and SignalP4.1 as the leading method for signal peptide identification[35]. In addition, we compared to a simple hydrophobicity-based prediction similar to TopPred[18].

Outstanding per-protein distinction between TMPs and other proteins

TMSEG correctly identified 40 of the 41 TMPs in the BlindTest dataset (98±2% sensitivity) and incorrectly predicted 8 of 285 soluble proteins as TMPs (3±1% false positive rate: FPR). TMSEG performed similar to PolyPhobius (100% sensitivity and 5±1% FPR) and significantly better than MEMSAT3 and MEMSAT-SVM (Table 2).

Although signal peptides can be confused with TMHs due to the similarity of their signal, only one of the 8 mistakes of predicting soluble proteins as TMPs originated from incorrectly predicting a signal peptide as a TMH. This shows that training on a dataset containing signal peptides helped significantly to reduce false positive predictions. PolyPhobius, which also includes a sophisticated signal peptide prediction, did not confuse any signal peptides with TMHs. However, MEMSAT-SVM, MEMSAT3, and the Baseline predictor had 13, 41, and 69 predicted TMHs, respectively, that overlapped by at least half their length with annotated signal peptides. Overall, TMSEG was able to reliably detect signal peptides and to not predict them as TMHs (Table S5, SOM).

We used the 74 TMPs from the 4th subset of the SP1441 dataset (*cf.* Table S2, SOM) to further test the prediction of signal peptides and TMHs. For these proteins, TMSEG and PolyPhobius incorrectly predicted several single-pass TMPs as soluble proteins, because they confused their TMHs near the N-terminus with signal peptides (Table S5, SOM). This trend did not occur with the TMPs from the TMP166 dataset (evident by their high sensitivity values; Table 2). An explanation might be that TMPs with TMHs within the first 40 residues are more prevalent in the SP1441 dataset, which makes this misclassification more likely to happen. Although these misclassification rates would lower our previous sensitivity estimates for TMSEG and PolyPhobius (at least for single-pass TMPs with their TMH near the N-terminus), we hesitate to generalize the results to everyday applicability since the SP1441 dataset is biased (it was generated to develop the signal peptide predictor SignalP4.1) and contains many TMPs with a TMH near the N-terminus. Further, only 2 of the 9 TMHs that were incorrectly predicted as SPs had experimental evidence.

While all methods reached high sensitivity, they differed vastly in their false positive rates, *i.e.* soluble proteins incorrectly considered to contain TMHs (Table 2). By translating the error rates, the number of proteins that would be misclassified in the entire human proteome can be estimated using two reasonable assumptions: (i) the error estimates for all methods based on the 326 non-redundant proteins (41 TMPs and 285 soluble proteins) in the BlindTest dataset hold true for the (redundant) human proteome, (ii) the human proteome has 20,196 proteins and 4791 of those are TMPs (*cf.* Section below *"Application to the human proteome"*). Under these assumptions, TMSEG achieves 97% per-protein accuracy and misclassifies only about 558 human proteins. The 2nd best method, PolyPhobius, makes

770 mistakes (212 more than TMSEG) and MEMSAT-SVM as the 3rd best method already misclassifies 2253 proteins (1695 more than TMSEG, Table 2). In fact, TMSEG is almost 8.8-times superior to the Baseline predictor, PolyPhobius over 6.5-times better, and MEMSAT-SVM 2.2-times better than the Baseline predictor (Table S6, SOM).

Best overall per-TMH prediction

Overall, TMSEG achieved a sustained level of precision (87±3%) and recall (84±3%) for the TMHs, *i.e.* 87±3% of all predicted TMHs were at the correct position and 84±3% of all observed TMHs had been accurately predicted (Fig. S4A and S4B, SOM). These values were second to no other method, however, only slightly above the 2nd best method MEMSAT-SVM (85±3% precision at 83±3% recall). All other methods had scores below 80%. For 66±6% of all TMPs, TMSEG predicted all observed TMHs at their correct positions, *i.e.* Q_{ok}=66±6% (Fig. 2). MEMSAT-SVM followed as second best with Q_{ok}=61±7% (Fig. 2). Nevertheless, given the small data sets, the top performance of TMSEG remained within one standard deviation of all compared methods, except the baseline hydrophobicity prediction (Fig. 2: error bars).

When comparing the performance on TMP subsets based on the number of TMHs, the performance got worse the more TMHs a protein had (Fig. S4C and S4D, SOM). This might be misunderstood to imply that prediction methods perform better in placing the TMHs in single-pass TMPs than in, *e.g.* GPCRs (with 7 TMHs). However, this simple numerical comparison ignores the difference in the difficulty of the task: The Baseline predictor reached a high value in Q_{ok} for single-pass TMPs, but failed to predict all TMHs correctly for any TMP with more than 5 TMHs (Fig. S4C, SOM). In fact, when we simply compiled performance for the subset of proteins for which the Baseline predictor failed, we found similar values for proteins with one TMH, those with 2-5, and those with more than 5 TMHs (Fig. S5, SOM).

In contrast, it surprised us that even for the trivial cases, *i.e.* those for which the Baseline predictor had all TMHs correct, the more advanced methods failed for some of them. This suggests that the large number of different features used by the more advanced methods sometimes interfere with and obscure a strong hydrophobicity signal. Indeed, only 11 of the 19 trivial TMPs were correctly predicted by all four other methods. However, TMSEG still performed best with Q_{ok}=89±6%, followed by MEMSAT3 and MEMSAT-SVM with Q_{ok}=84±7% (data not shown).

Best inside/outside topology prediction

TMSEG and MEMSAT3 correctly placed the N-terminus as inside (*e.g.* cytoplasmic) or outside (*e.g.* extracellular), *i.e.* correctly predicted the topology, for 93±4% of all TMPs (Table 2). When taking into account the global topology and correct TMH placement (*i.e.* Q_{top}), TMSEG performed better than all other methods reaching Q_{top}=63±6% (Fig. 2). This is five percentage points higher than the 2nd best method, MEMSAT-SVM (albeit still within one standard deviation). Most advanced methods predicted the topology correctly for almost all proteins for which they correctly predicted all TMHs (Q_{top} almost identical to Q_{ok} for all methods, except for the Baseline predictor in Fig. 2).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Application to the human proteome

We applied TMSEG to predict all helical TMPs in the human proteome (20,196 proteins from UniProtKB/Swiss-Prot). TMSEG predicted a total of 5157 TMPs, almost half of these (2300 = 45%) were predicted with one TMH. Given the sensitivity and false positive rate of TMSEG (98±2% and 3±1%, respectively; Table 2), we estimate that 462 TMPs were incorrectly predicted (over-predicted) and 96 were missed (under-predicted). In total, we thus misclassified 558 proteins, and our corrected estimate was that humans have about 4791 TMPs, *i.e.* about 24% of all proteins cross the membrane. While TMSEG misclassified about 558 human proteins, the mistake in the estimate of this percentage appeared to be less than a per-mille, i.e. ±0.01%. However, our error estimate might be too simplistic due to the high number of single-pass TMPs for which the error rates are much higher than for proteins with more TMPs.

Confirming previous observations[2,3], we also observed two peaks of predicted TMPs for proteins with 7 TMHs (819 proteins) and 12 TMHs (189 proteins). These likely represent G protein-coupled receptors (GPCRs) and transporter proteins. Applying UniqueProt to the 5157 predicted TMPs we found around 500 non-redundant TMPs of which 320 are single-pass TMPs.

Latest experimental structures confirmed our estimates

The 12 new TMPs (New12 dataset) that have recently been added to the PDB constituted the only data set with truly identical conditions for all methods assessed. The New12 dataset allowed us to confirm the outstanding performance of our new method TMSEG. TMSEG and PolyPhobius correctly identified 10 of the 12 TMPs (83±10% sensitivity), while MEMSAT3, MEMSAT-SVM and the Baseline predictor identified 11 (92±7% sensitivity). However, TMSEG correctly predicted every TMH of those 10 TMPs, resulting in a Q_{ok}=83±10%, compared to Q_{ok}=58±13% for PolyPhobius, MEMSAT3, and MEMSAT-SVM (Baseline predictor Q_{ok}=50±13%). TMSEG also performed best taking into account the topology prediction and reached Q_{top}=66±12%, compared to a Q_{top}=58±13% for MEMSAT3 and MEMSAT-SVM, and Q_{top}=50±13% for PolyPhobius and the Baseline predictor.

Comparisons complicated by small data sets

The two small datasets available for evaluation (BlindTest with 41 TMPs and New12 with 12 TMPs) implied high standard errors for many performance estimates. Especially standard errors for the TMH-segment based scores are so high (up to 16 percentage points, Fig. S4, SOM) that comparisons between methods hardly provide statistically significant differences on the TMH-segment level. Nevertheless, TMSEG seemed to perform on par with any existing method. Note that the differences in the distinction between helical TMPs and other proteins in the BlindTest dataset were statistically significant even in considering TMSEG as slightly better than the 2nd best PolyPhobius (Table 2).

Further, we could not use a single gold standard, because OPM and PDBTM differed in their TMH annotations: comparing the OPM annotations to the PDBTM annotations (*i.e.* ‘predicting’ one with the other) yielded Q_{ok}=56±7%. In other words, if we considered one of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

those experiment-based annotations as the prediction of the other, the average performance would be similar to that of TMSEG and the other methods. When using only OPM or PDBTM annotations to evaluate the prediction performance, TMSEG still performed excellently (Fig. S6, SOM). However, this was also the only comparison in which one other method reached a numerically higher value for a data set than TMSEG, namely MEMSAT-SVM on the PDBTM annotations. Overall, all predictions agreed more with OPM than with PDBTM annotations (Fig. S6, SOM).

Performance best with diverse alignments

TMSEG strongly depends on the evolutionary information taken from PSI-BLAST PSSMs. We recommend using a sufficiently large search database (*e.g.* UniRef90) to generate the PSSMs. Additionally, redundancy reduction might help (*e.g.* at 90% pairwise sequence identity as in UniRef90).

Alignments built from smaller search-databases (e.g. UniRef50 and Swiss-Prot) only slightly lowered the per-protein performance: the sensitivity never dropped below 90±4%, while the false positive rate remained at or below 3±1%. However, the TMH-based precision and recall values dropped substantially (Fig. S7, SOM). Thus, for sequences that produce no PSI-BLAST hits, we recommend using a larger search database or – in the rare case that the protein is a true singleton – a method that is independent of evolutionary information, *e.g.* Phobius[21,27].

Re-entrant membrane helices not predicted correctly

Our dataset contained only few re-entrant helices, insufficient to learn their prediction (Table S1, SOM). Therefore, we considered re-entrant helices as non-TM during training to avoid later interference with the inside/outside topology prediction. Due to the lack of data we could not reliably assess how well TMSEG distinguishes TMHs from re-entrant membrane helices: The BlindTest dataset included only seven re-entrant regions (OPM and PDBTM annotations combined). TMSEG incorrectly predicted 5 of 7 as TMHs; 2 of these 5 were predicted as two separate TMHs, thus the overall inside/outside topology was not influenced. MEMSAT-SVM, the only tested method that predicts re-entrant helices, identified 5 of the 7 as re-entrant, predicted one as a TMH, and missed the last. When considering re-entrant regions as TMHs, Q_{ok} remained the same for TMSEG and PolyPhobius and dropped by 2-5 percentage points for MEMSAT-SVM, MEMSAT3, and the Baseline predictor.

TMSEG easily combined with other methods

Due to the modularity of TMSEG (*i.e.* its four separate steps, Fig. 1), it can be used to refine other methods. This includes the adjustment of the TMHs as well as the inside/outside topology prediction. We used the TMH predictions of the reference methods, and applied steps 3 and 4 of TMSEG to their prediction (Fig. 2). Applying TMSEG as refinement improved the performance for most methods (Fig. 3; Fig. S8, SOM). While the improvement was small for the TMH placement (Q_{ok}), TMSEG improved most methods by over eight percentage points in Q_{top} (correct TMHs and topology).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Runtime estimation

We estimated the runtime by applying TMSEG to the human proteome (20,196 proteins). As the time to run PSI-BLAST differs depending on the database size, we decided to use pre-computed PSSMs to measure only the time needed by TMSEG. Given those PSI-BLAST profiles, the prediction for the entire human proteome took about 90 minutes (Intel Core i7-3632QM 2.2GHz, 8GB RAM; no multithreading), which corresponds to three to four protein sequences per second.

Conclusion

In our hands, our new method TMSEG almost always outperformed existing state-of-the-art prediction methods (Table 2, Fig. 2). However, due to the small data sets, many improvements on the per-TMH level remained too small for the large margin of statistical significance (standard errors up to 16 percentage points, Fig. S4, SOM). Most importantly, TMSEG achieved the significantly best per-protein classification in the distinction between helical TMPs and all other proteins. For instance, for the prediction of all human proteins, this implied about 558 incorrectly predicted proteins. This number might appear high, however, no method tested reached such a low level, *e.g.* PolyPhobius misclassified about 200 more proteins than TMSEG and MEMSAT-SVM fared about four times worse (corresponding to over 2000 incorrect predictions).

The highest per-protein performance resulted from a combined prediction of TMHs, non-TM regions, and signal peptides. In order to predict re-entrant helices, another state would have to be introduced; as is, TMSEG predicted 5 of 7 re-entrant helices in our data set as TMHs. The sustained high levels of per-segment predictions resulted from our new segment-focused algorithm. Another major advantage of our new concept is that it can be used to improve the predictions of most other TMH prediction methods.

Availability and speed

Other than its top performance, using TMSEG may also be recommended due to its speed and because it might help to improve over the method that you run locally. The method is easily and freely available: online through the PredictProtein[45] webserver (www.predictprotein.org), and as standalone Debian package from the Rostlab Debian repository (www.rostlab.org/owiki) and GitHub (www.github.com/Rostlab/TMSEG). A tutorial on how to use PSI-BLAST and TMSEG can be found in the Rostlab Wiki (www.rostlab.org/owiki/index.php/TMSEG).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Thanks to Tim Karl for technical and to Inga Weise (both TUM) for administrative assistance. This work was supported in part by a grant from the Alexander von Humboldt foundation through the German Federal Ministry for Education and Research (BMBF) and by the grant U54 GM095315 to the New York Consortium on Membrane Protein Structure (NYCOMPS) from the Protein Structure Initiative (PSI) of the National Institutes of Health (NIH). Thanks to all authors who made their methods openly available and provided us with versions to run on our own machines. Last but not least, thanks to all who practice open science and deposit their data into public databases and those who maintain these excellent databases.

Abbreviations used

3D	three-dimensional
GPCR	G protein-coupled receptor
NN	(artificial) neural network
OPM	Orientations of Proteins in Membranes
PDB	Protein Data Bank
PDBTM	Protein Data Bank of Transmembrane Proteins
RF	random forest
TMH	transmembrane alpha-helix
TMP	transmembrane protein

References

1. von Heijne G. The membrane protein universe: what's out there and why bother? J Intern Med. 2007; 261(6):543–557. [PubMed: 17547710]
2. Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci. 2001; 10(10):1970–1979. [PubMed: 11567088]
3. Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L. Prediction of the human membrane proteome. Proteomics. 2010; 10(6):1141–1149. [PubMed: 20175080]
4. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006; 5(12):993–996. [PubMed: 17139284]
5. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res. 2004; 32(8):2566–2577. [PubMed: 15141026]
6. Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J. 1986; 5(11):3021–3027. [PubMed: 16453726]
7. von Heijne G, Gavel Y. Topogenic signals in integral membrane proteins. Eur J Biochem. 1988; 174(4):671–678. [PubMed: 3134198]
8. Punta M, Love J, Handelman S, Hunt JF, Shapiro L, Hendrickson WA, Rost B. Structural genomics target selection for the New York consortium on membrane protein structure. J Struct Funct Genomics. 2009; 10(4):255–268. [PubMed: 19859826]
9. Love J, Mancia F, Shapiro L, Punta M, Rost B, Girvin M, Wang DN, Zhou M, Hunt JF, Szyperski T, Gouaux E, MacKinnon R, McDermott A, Honig B, Inouye M, Montelione G, Hendrickson WA. The New York Consortium on Membrane Protein Structure (NYCOMPS): a high-throughput platform for structural genomics of integral membrane proteins. J Struct Funct Genomics. 2010; 11(3):191–199. [PubMed: 20690043]
10. Caffrey M. A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. Acta Crystallogr F Struct Biol Commun. 2015; 71(Pt 1):3–18. [PubMed: 25615961]
11. Moraes I, Evans G, Sanchez-Weatherby J, Newstead S, Stewart PD. Membrane protein structure determination - the next generation. Biochim Biophys Acta. 2014; 1838(1 Pt A):78–87. [PubMed: 23860256]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
13. Kloppmann E, Punta M, Rost B. Structural genomics plucks high-hanging membrane proteins. Curr Opin Struct Biol. 2012; 22(3):326–332. [PubMed: 22622032]
14. White SH. Biophysical dissection of membrane proteins. Nature. 2009; 459(7245):344–346. [PubMed: 19458709]
15. White SH. The progress of membrane protein structure determination. Protein Sci. 2004; 13(7): 1948–1949. [PubMed: 15215534]
16. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: orientations of proteins in membranes database. Bioinformatics. 2006; 22(5):623–625. [PubMed: 16397007]
17. Kozma D, Simon I, Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res. 2013; 41(Database issue):D524–529. [PubMed: 23203988]
18. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol. 1992; 225(2):487–494. [PubMed: 1593632]
19. 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. 2001; 305(3):567–580. [PubMed: 11152613]
20. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001; 17(9):849–850. [PubMed: 11590105]
21. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004; 338(5):1027–1036. [PubMed: 15111065]
22. Käll L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics. 2005; 21(Suppl 1):i251–257. [PubMed: 15961464]
23. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci. 1995; 4(3):521–533. [PubMed: 7795533]
24. Rost B, Casadio R, Fariselli P. Refining neural network predictions for helical transmembrane proteins by dynamic programming. Proc Int Conf Intell Syst Mol Biol. 1996; 4:192–200. [PubMed: 8877519]
25. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics. 2007; 23(5):538–544. [PubMed: 17237066]
26. Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics. 2009; 10:159. [PubMed: 19470175]
27. Reeb J, Kloppmann E, Bernhofer M, Rost B. Evaluation of transmembrane helix predictions in 2014. Proteins. 2015; 83(3):473–484. [PubMed: 25546441]
28. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43(Database issue):D204–212. [PubMed: 25348405]
29. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 2013; 41(Database issue):D483–489. [PubMed: 23203869]
30. Mika S, Rost B. UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res. 2003; 31(13):3789–3791. [PubMed: 12824419]
31. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991; 9(1):56–68. [PubMed: 2017436]
32. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12(2):85–94. [PubMed: 10195279]
33. Granseth E, Viklund H, Elofsson A. ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. Bioinformatics. 2006; 22(14):e191–196. [PubMed: 16873471]
34. Papaloukas C, Granseth E, Viklund H, Elofsson A. Estimating the length of transmembrane helices using Z-coordinate predictions. Protein Sci. 2008; 17(2):271–278. [PubMed: 18096645]
35. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011; 8(10):785–786. [PubMed: 21959131]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

36. Chen CP, Kernytsky A, Rost B. Transmembrane helix predictions revisited. Protein Sci. 2002; 11(12):2774–2791. [PubMed: 12441377]

37. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall; New York: 1993.

38. Eisenberg D. Three-dimensional structure of membrane and surface proteins. Annu Rev Biochem. 1984; 53:595–623. [PubMed: 6383201]

39. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology. 1993; 232:584–599. [PubMed: 8345525]

40. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cellular and Molecular Life Sciences. 2003; 60(12):2637–2650. [PubMed: 14685688]

41. 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. 1997; 25(17):3389–3402. [PubMed: 9254694]

42. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009; 11(1):10–18.

43. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157(1):105–132. [PubMed: 7108955]

44. Rath EM, Tessier D, Campbell AA, Lee HC, Werner T, Salam NK, Lee LK, Church WB. A benchmark server using high resolution protein structure data, and benchmark results for membrane helix predictions. BMC Bioinformatics. 2013; 14:111. [PubMed: 23530628]

45. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Honigschmid P, Schafferhans A, Roos M, Bernhofer M, Richter L, Ashkenazy H, Punta M, Schlessinger A, Bromberg Y, Schneider R, Vriend G, Sander C, Ben-Tal N, Rost B. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res. 2014; 42(Web Server issue):W337–343. [PubMed: 24799431]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

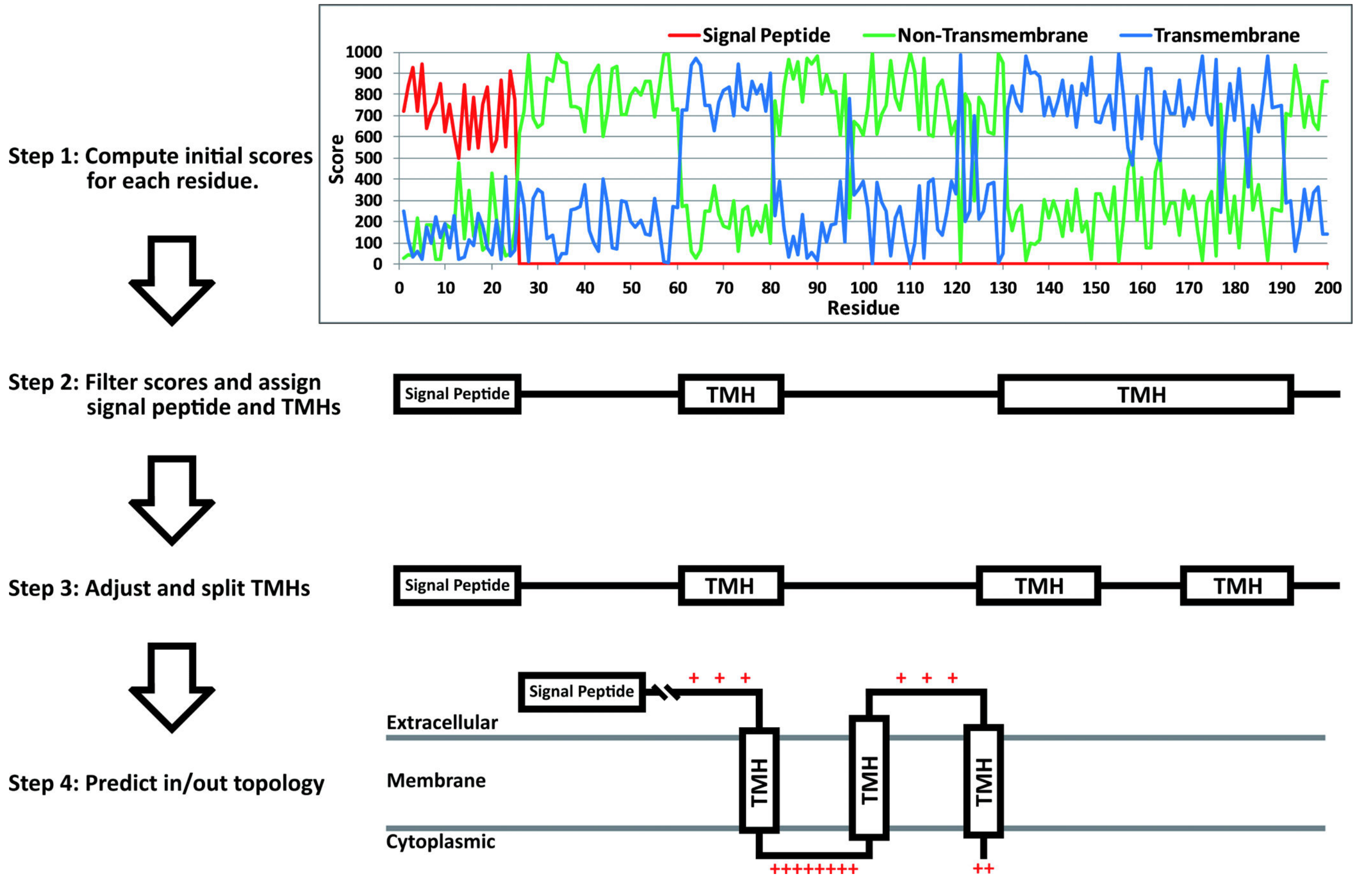

Fig. 1. TMSEG algorithm

The new method TMSEG has four steps of machine learning and optimization. **Step 1:** A random forest (RF) assigns a score to each residue for the three states transmembrane helix (TMH), signal peptide, and non-TM region. **Step 2:** The previous scores are smoothed (median over 5 residues), all residues are assigned to the state with the highest score, and short segments are removed. **Step 3:** A segment-based neural network (NN) adjusts the exact position of predicted TMHs, and their length, sometimes splitting TMHs, sometimes shifting, extending, or compressing them. **Step 4:** The inside/outside topology is predicted by another RF.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

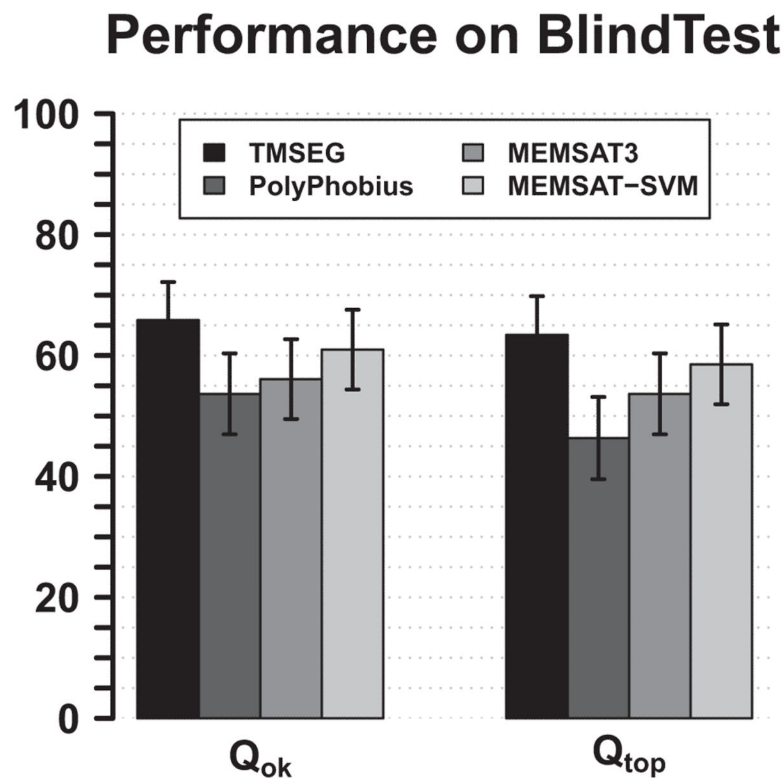

Fig. 2. TMSEG compared favorably to state-of-the-art

Results are provided for all 41 TMPs in the BlindTest dataset. Error bars are the sample standard deviation based on bootstrapping (*cf.* Methods). Shown is on the left the percentage of proteins for which all TMHs were predicted correctly (Q_{ok}, Table 1) and on the right the percentage of proteins with correctly predicted TMHs and inside/outside topology (Q_{top}, Table 1; note that $Q_{ok} \geq Q_{top}$ by definition).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

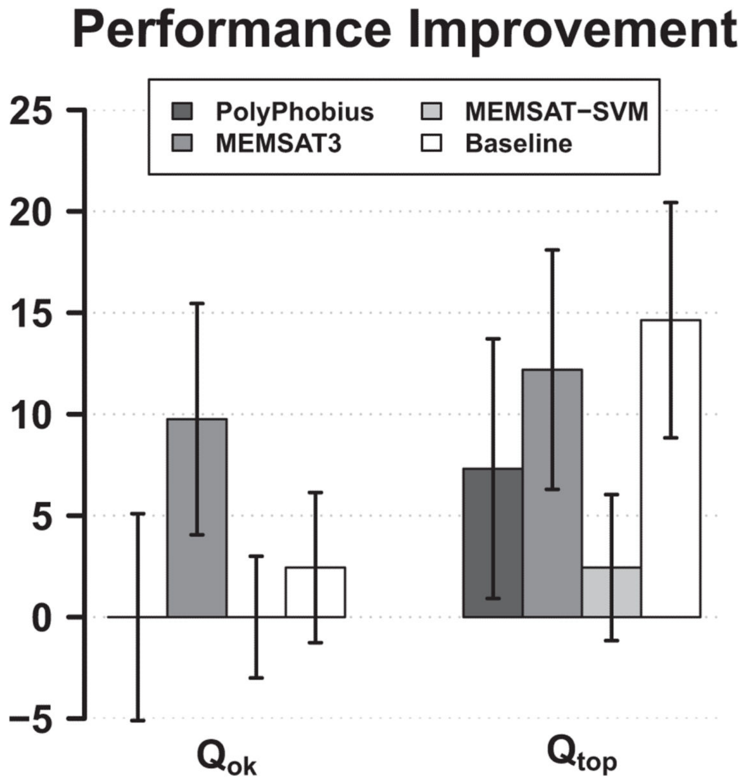

Fig. 3. TMSEG applied to refine other methods

The TMSEG algorithm iteratively refines performance through four consecutive steps. Here, we applied steps 3 and 4 as post-filters to other methods (data set and error bars as in Fig. 2). Given is the improvement of Q_{ok} and Q_{top} (*cf.* Table 1 for definitions) of the prediction method by applying TMSEG, *i.e.* Q(method+TMSEG) – Q(method). Note that PolyPhobius (1st bar on the left) and MEMSAT-SVM (3rd bar on the left) showed, on average, no improvement in Q_{ok}.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Evaluation measures

Listed are the evaluations measures used and how they were calculated. Precision and recall for the performance evaluation of the TMH prediction were computed by combining all TMHs within the dataset (*i.e.* not averaged over each protein). Q_{ok} and Q_{top} were calculated based on all TMPs, where N was the number of TMPs in the dataset, p_i and r_i were the TMH precision and recall for protein i within the dataset, and $t_i = 100\%$ indicated a correctly predicted N-terminal inside/outside topology for protein i.

Measurement	Formula	Description
Precision (%)	$100 * \frac{\#\ of\ correctly\ predicted\ TMHs}{\#\ of\ predicted\ TMHs}$	Precision of TMH prediction
Recall (%)	$100 * \frac{\#\ of\ correctly\ predicted\ TMHs}{\#\ of\ observed\ TMHs}$	Recall of TMH prediction
Q_{ok} (%)	$\frac{100}{N} * \sum_{i=1}^{N} x_i; x_i = \begin{cases} 1, if\ p_i = r_i = 100\% \\ 0, else \end{cases}$	Percentage of TMPs with correct TMH placement
Q_{top} (%)	$\frac{100}{N} * \sum_{i=1}^{N} y_i; y_i = \begin{cases} 1, if\ p_i = r_i = t_i = 100\% \\ 0, else \end{cases}$	Percentage of TMPs with correct TMH placement and inside/ outside topology
FPR (%)	$100 * \frac{\#\ of\ incorrectly\ predicted\ TMPs}{\#\ of\ soluble\ proteins}$	False positive rate of TMP prediction
Sensitivity (%)	$100 * \frac{\#\ of\ correctly\ predicted\ TMPs}{\#\ of\ observed\ TMPs}$	Sensitivity of TMP prediction

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2
Per-protein distinction between helical TMPs and other proteins

Results are provided for all 41 TMPs and 285 soluble proteins in the BlindTest dataset. Error rates are the sample standard deviation based on bootstrapping (*cf.* Methods). Listed are the *TMP sensitivity* (percentage of correctly predicted helical TMPs), the *TMP FPR* (percentage of non-TMP proteins incorrectly predicted as TMP), *Topology correct* (percentage of proteins for which the topology (inside/outside) was correctly predicted; this differs from Q_{top} which requires topology and all TMHs to be predicted correctly), *Misclassified in human* (estimates the number of proteins misclassfied for the entire human proteome), and *More mistakes than TMSEG in human* (estimates the number of proteins misclassfied more by the method than by TMSEG). The estimates for the human proteome are based on two assumptions: (i) the error estimates on the BlindTest dataset hold true for the human proteome, (ii) the human proteome has 20,196 proteins, 4791 of which are TMPs (*cf.* Results section *"Application to the human proteome"*).

Method	TMP sensitivity	TMP FPR	Topology correct	Misclassified in human	More mistakes than TMSEG in human
TMSEG	98 ± 2	3 ± 1	93 ± 4	558	-
PolyPhobius[22]	100 ± 0	5 ± 1	78 ± 7	770	212
MEMSAT3[25]	100 ± 0	28 ± 2	93 ± 4	4,313	3,755
MEMSAT-SVM[26]	98 ± 2	14 ± 2	88 ± 5	2,253	1,695
Baseline	95 ± 3	31 ± 2	75 ± 7	5,015	4,457