SUPPLEMENTARY MATERIALS

"Enhanced Recognition of Transmembrane Protein Domains with Prediction-based Structural Profiles"

Baoqiang Cao, Aleksey Porollo, Rafal Adamczak, Mark Jarrell and Jaroslaw Meller
Contact: jmeller@chmcc.org; Server available at http://minnou.cchmc.org

SP1. Multistage protocol for transmembrane helix prediction in MINNOU

In addition to a simple NN-based classifier developed and assessed in the cross-validation study (see the Systems and Methods section of the main body of the paper), we also developed a multistage protocol for enhanced prediction of transmembrane (TM) helices. For the final predictor we do not consider the MA-based representation, which is shown using cross-validation to yield a lower accuracy compared with the proposed compact prediction-based "structural profiles". We also excluded hydropathy scales from the representation of an amino acid residue, as it was observed to lead to a higher level of confusion with globular proteins and signal peptides. Moreover, statistical propensities of amino acids (e.g. to lipid – aqueous phase interfaces) that could be used, in principle, to improve the prediction of TM segments, are not taken into account at this stage. Consequently, each residue is initially represented by five numbers: the predicted real valued RSA, confidence of RSA prediction and probabilities of each of the three secondary structures (as predicted by SABLE, http://sable.cchmc.org). SABLE predictions are derived from the multiple alignment, hydropathy scales and other attributes, which are commonly used by other state-of-the-art methods. Therefore, it is expected that other accurate methods for RSA and SS prediction will be useful in that regard as well. This is a subject of a future investigation.

Following in the footsteps of other studies (Rost et al., 1995), we use a two-stage prediction system, with the second layer (structure-to-structure) NNs allowing one to "average" and smooth over the initial classification obtained using the first (sequence-to-structure) layer predictor. The architecture of the first and second layer NNs is similar to that used for the cross-validation study. Namely, a simple feed-forward topology with one hidden layer, fully interconnected with the input and output layers, is employed. The choice of the sliding window size, the number of nodes in the hidden layer, training protocols and other characteristics of these NNs are discussed below.

In order to reduce the danger of overfitting and achieve regularization as well as improvement in accuracy, a consensus of twenty different networks was used to generate predictions at each stage. These different networks were trained on different subsets of the training set that were also used for the cross-validation study. Multiple NNs were trained on each subset of the data, with different number of nodes in the hidden layer and with different size of the sliding window. The number of nodes in the hidden layer was again varied between 8 and 18 and the size of the sliding window was varied between 11 and 31. The stopping criteria were defined in terms of improvement (or lack thereof) on the corresponding validation set. From multiple networks trained on each subset of the data, the one providing the highest per residue accuracy on the corresponding validation set was then selected for the final consensus predictor. In that sense, the whole training set of 73 protein chains is used here to optimize the parameters of each of the networks. However, using different subsets of the data provides a better sampling of local minima.

The first ten networks were trained on different subsets of the data using the representation with five attributes per amino acid residue, as described above. The other ten networks included in the consensus were trained with only four attributes per residue. Namely, the two nodes related to relative solvent accessibility prediction are replaced by one node, which represents a discretized RSA assignment. In other words, such a binary node indicates if a given residue is predicted to be buried or exposed (with a threshold of 25% RSA used to project the real valued SABLE prediction into two classes). Adding these additional networks to the consensus was found to improve somewhat the per segment classification accuracy. The consensus predictor is based on a simple majority voting. The source code for the MINNOU package, with the definition of all the NNs used for the consensus prediction (including a detailed description of their topology and parameters) can be downloaded from ftp://ftp.chmcc.org/pdi/jmeller/minnou/.

To further improve the performance of the first layer classifier, we introduced a second layer prediction system. The output of the first layer is used at this stage as input. In addition, the SABLE predictions used in 1st layer are included in the input as well. After some experimentation we found that the best results are obtained when the output of the first layer system is combined as follows to be included in the input of the 2nd layer. For each residue, two additional nodes are used to represent averaged (over the

networks included in the consensus) excitation of the two output nodes corresponding to two different classes. Another pair of nodes is used to represent the maximum (again over the different networks) excitation for each class. Thus, the number of nodes per residue in the 2nd layer is either nine or eight, depending on the number of attributes used in the first layer. The choice of optimal topology and other settings for the neural networks and the training procedures follows that for the 1st layers (see also the MINNOU package for further details).

While the second layer NNs lead to significant smoothing of the prediction and improves the overall accuracy in terms of both: sensitivity and specificity, some long or short helices are still occasionally predicted. We estimated the probability density distribution for the length of TM helices and used it as a guideline in the design of a filter, applied to the second layer prediction in order to avoid such unphysical predictions. Similar filters have been used before by other groups (Rost et al., 1995). Basically, the final filter is applied to either split predicted long TM helices or delete too short ones, which are observed with very low frequency in the known sample of helical TM domains. The presence of relatively short (e.g., horizontally oriented) membrane embedded helices as well as relatively long ("skewed") helices that occur in some ion channels, for instance, influences the choice of the length thresholds applied here.

Specifically, if only a single membrane segment is predicted in a protein, then it is deleted if its length is shorter than 14 residues; if more than one membrane segment is predicted, then it is deleted if shorter than 8 residues. On the other hand, if there is a continuous membrane segment of length greater than 44, it is split into two segments in the middle (by introducing an artificial loop of length one); and if the predicted segment is longer than 66 residues (the latter happened seven times in the set of 2247 segments predicted for the TMH Benchmarks test), then it is split into three segments. Even though this final post-processing step does not affect significantly the per residue accuracy, it does help to "smooth" the predictions and to improve the per segment accuracy, Q_{OK}, as defined in (Chen et al., 2002; Kernytsky and Rost, 2003). It also helps reducing further the observed level of confusion with signal peptides and globular proteins by filtering out unphysical short TM helices (see Table 6 and also the Results section of the main body of the paper for further discussion). However, this final prediction stage is likely to be improved further, for example, by considering explicitly the overall topology of membrane proteins and characteristics of loops connecting TM segments. This is a subject of a future work.

SP2. Tables and figures

Table 5 Per residue classification accuracy (Q_3) of secondary structure predictions obtained using SABLE (Adamczak et al., 2005) and PSIPRED servers (Jones, 1999). The results on non-redundant sets of 73 alpha-helical and 15 beta-barrel membrane proteins, respectively, are shown.

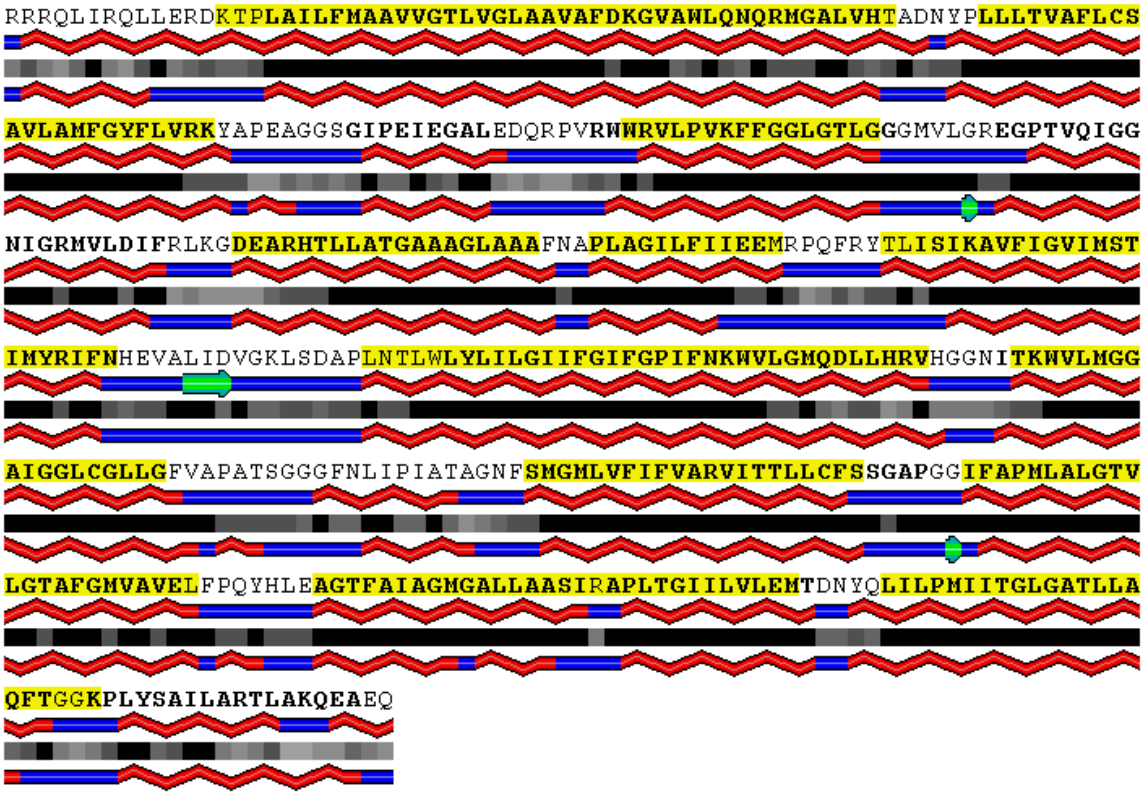
Test set	SABLE	PSIPRED
Alpha-helical: all	78.2%	76.6%
Alpha-helical: TM segments only	83.3%	80.6%
Beta-barrel: all	68.7%	75.7%
Beta-barrel: TM segments only	64.9%	77.6%

Table 6 Improvements due to 1st layer, 2nd layer and filter (F) based predictors according to TMH Benchmark evaluation. Per segment (Q_{OK}) and per residue (Q_2) classification accuracies and confusion levels with signal peptides (cSP) and globular proteins (cGP) are given in the units of per cent. Five different prediction systems are compared. The first two involve the first layer predictions: (a) training with the standard training set containing membrane proteins only; and (b) training with the augmented training set that includes signal peptides and some false positives from the first iteration. The next two include the corresponding results for the second layer predictors that use the results of the first layer classifiers as part of their input. The last row contains results of the second layer prediction with filtering of short, unphysical segments, which improves the prediction in some respects.

	High resolution		Low resolution		cSP	cGP
	Q_{OK}	Q_2	Q_{OK}	Q_2		
1st layer (a)	66	88	32	83	78	27
1st layer (b)	61	88	40	85	54	12
2nd layer (a)	66	89	24	82	64	23
2nd layer (b)	66	88	39	84	27	8
2nd layer (F)	80	89	55	85	8	1

Figure 1 Examples of MINNOU transmembrane helix predictions for an ion channel, PDB code 1OTS, (panel A), and a glutamate transporter protein, PDB code 1XFH (panel B). The amino acid sequence is included in the first row, with the actual and predicted membrane segments highlighted using bold and yellow boxes, respectively. The secondary structures and relative solvent accessibilities, predicted using the SABLE server, are shown in the second and third rows, respectively. The alpha-helices are represented by red ribbons, the beta-strands by green arrows and coils by blue lines, respectively. The level of predicted aqueous solvent exposure is represented by shaded boxes, with fully "buried" and fully exposed residues represented by black and white boxes, respectively. The secondary structures observed in the experimentally resolved structures are shown in the last row for comparison. See text for further details.

A.


RRRQLIRQLLERDKTPLAILFMAAVVGTLVGLAAVAFDKGVAWLQNQRMGALVHTADNYPLLLTVAFLCS
AVLAMFGYFLVRKYAPEAGGSGIPEIEGALEDQRPVRWWRVLPVKFFGGLGTLGGGMVLGREGPTVQIGG
NIGRMVLDIFRLKGDEARHTLLATGAAAGLAAAFNAPLAGILFIIEEMRPQFRYTLISIKAVFIGVIMST
IMYRIFNHEVALIDVGKLSDAPLNTLWLYLILGIIFGIFGPIFNKWVLGMQDLLHRVHGGNITKWVLMGG
AIGGLCGLLGFVAPATSGGGFNLIPIATAGNFSMGMLVFIFVARVITTLLCFSSGAPGGIFAPMLALGTV
LGTAFGMVAVELFPQYHLEAGTFAIAGMGALLAASIRAPLTGIILVLEMTDNYQLILPMIITGLGATLLA
QFTGGKPLYSAILARTLAKQEAEQ

B.

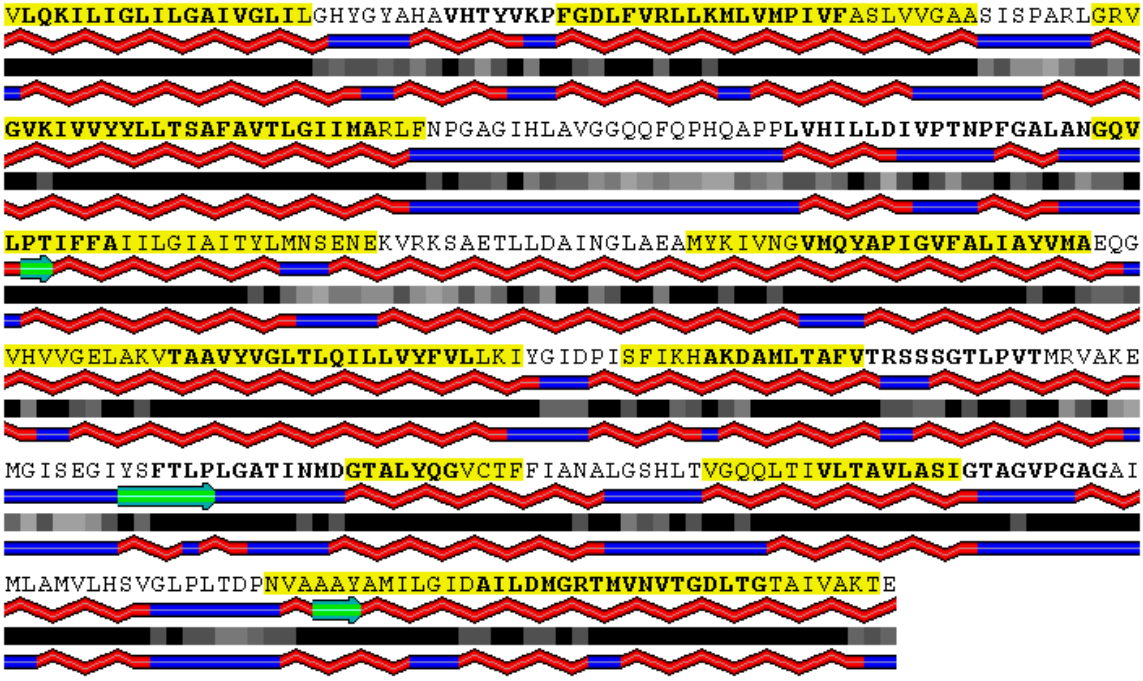

VLQKILIGLILGAIVGLILGHYGYAHAVHTYVKPFGDLFVRLLKMLVMPIVFASLVVGAASISPARLGRV
GVKIVVYYLLTSAFAVTLGIIMARLFNPGAGIHLAVGGQQFQPHQAPPLVHILLDIVPTNPFGALANGQV
LPTIFFAIILGIAITYLMNSENEKVRKSAETLLDAINGLAEAMYKIVNGVMQYAPIGVFALIAYVMAEQG
VHVVGELAKVTAAVYVGLTLQILLVYFVLLKIYGIDPISFIKHAKDAMLTAFVTRSSSGTLPVTMRVAKE
MGISEGIYSFTLPLGATINMDGTALYQGVCTFFIANALGSHLTVGQQLTIVLTAVLASIGTAGVPGAGAI
MLAMVLHSVGLPLTDPNVAAAYAMILGIDAILDMGRTMVNVTGDLTGTAIVAKTE

Figure 2 The distribution of lengths for TM segments included in the MPtopo (Jayasinghe et al., 2001) database of membrane proteins. 3D_helix and 1D_helix refer to helical membrane proteins with or without resolved 3D structure, respectively (see text for details). Thus, in case of 1D_helix set only low resolution information derived from various experimental studies is available. Consequently, the delineation of TM segments is laden with additional uncertainty. Moreover, it was suggested that prediction methods may have been used to derive the actual boundaries of TM segments for some of the 1D_helix proteins (Chen et al., 2002). As can be seen from the figure, the distribution of lengths for low resolution structures is indeed qualitatively different, compared to high resolution (3D_helix) set. In particular, a sharp peak is observed around the canonical length of an alpha-helical TM segment (20-22 residues) for low resolution structures. This is in stark contrast to a much wider distribution for high resolution structures. Because of these differences and higher uncertainty in the annotation of low resolution structures, we used only 3D_helix set for training of MINNOU. The 3D_other set includes beta–barrel membrane proteins with known 3D structures and is characterized by much shorter membrane spanning segments.

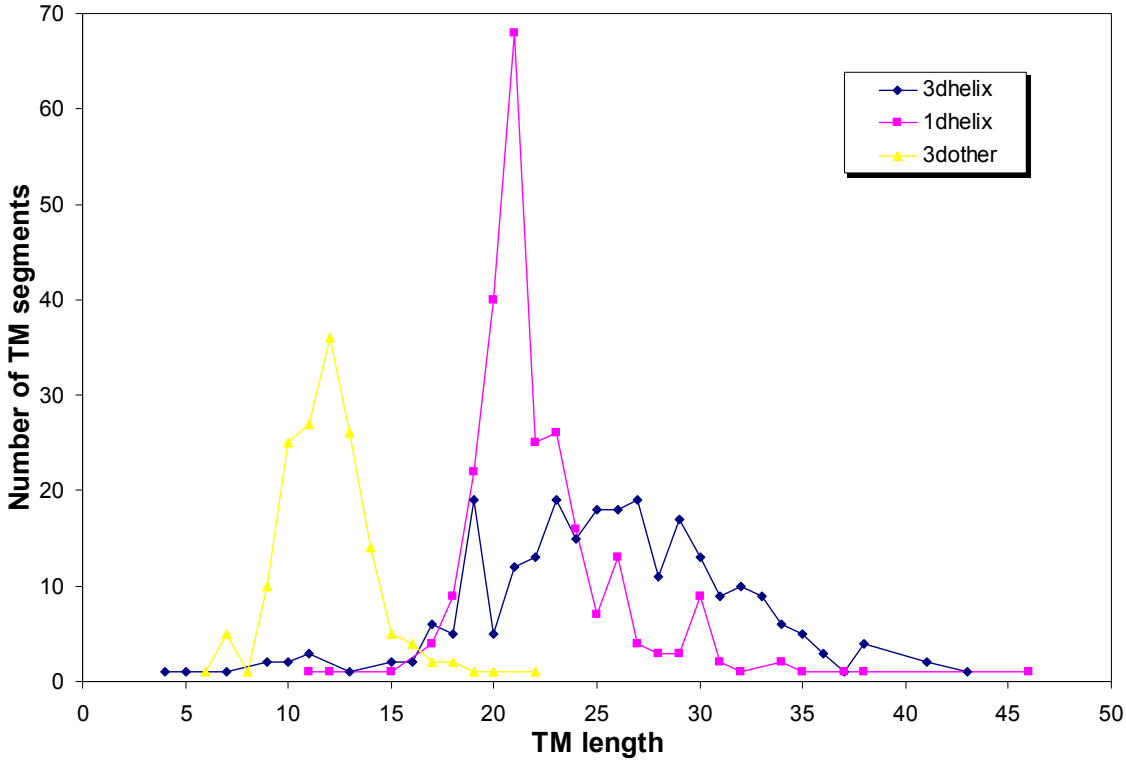


Figure 3 Dependence of prediction accuracy on the size of the sliding window, using 10-fold cross-validation on the non-redundant training set of 73 alpha-helical membrane protein chains (see text for details). The results of a simple NN, with one hidden layer consisting of 10 nodes, trained for 1000 cycles with the default learning parameter and the Rprop delta parameter set to 0.1, are compared in terms of Matthews Correlation Coefficients (MCC). Fixed topology and meta-parameters of the networks allow one to compare the results directly and assess the extent of overfitting due to increased windows size. Each amino acid residue in the sliding window is represented here by five numbers, using the prediction-based structural profiles described in the text. Thus, increasing the size of the sliding window from 11 to 21 increases the number of weights to be optimized from 550 to 1050 (neglecting the weights for edges between the hidden layer and the two nodes in the output layer as well as the parameters for the activation functions of the nodes in the hidden and output layers). As can be seen from the figure, the cross-validation estimates of the accuracy quickly increase initially in order to reach a plateau and then gradually decrease with the growing size of the sliding window, pointing out problems with overfitting for more complex models. It should be noted that changing the topology of the network we were able to obtain somewhat better results (at the level of MCC=0.74, as reported in Tables 1 and 2 for different sliding windows). We would also like to comment that while very short sliding windows seem to be capable of capturing relatively well the essential information about an amino acid residue and its environment (since not too many residues predicted to be fully buried and in alpha helices are found in the loop domains of TM proteins included in the training set) the overall quality of predictions based on such short windows is low due high level of confusion with globular proteins and low per segment accuracy.