BIOINFORMATICS

Vol. ? no. ? 200?
Pages 1–1

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

Baoqiang Cao[2], Aleksey Porollo[1], Rafal Adamczak[1], Mark Jarrell[2] and Jaroslaw Meller[1,3]

[1]*Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, OH 45229*

[2]*Department of Physics, University of Cincinnati, Cincinnati, OH 45221*

[3]*Department of Informatics, Nicholas Copernicus University, 87-100 Torun, Poland*

ABSTRACT

Motivation: Membrane domain prediction has recently been re-evaluated by several groups, suggesting that the accuracy of existing methods is still rather limited. In this work, we revisit this problem and propose novel methods for prediction of alpha-helical as well as beta-sheet transmembrane (TM) domains. The new approach is based on a compact representation of an amino acid residue and its environment, which consists of predicted solvent accessibility and secondary structure of each amino acid. A recently introduced and trained on a set of soluble proteins method for solvent accessibility prediction is used here to indicate segments of residues that are predicted not to be accessible to water and, therefore, may be "buried" in the membrane. While evolutionary profiles in the form of a multiple alignment are used to derive these simple "structural profiles", they are not used explicitly for the membrane domain prediction and the overall number of parameters in the model is significantly reduced. This offers the possibility of a more reliable estimation of the free parameters in the model with a limited number of experimentally resolved membrane protein structures.

Results: Using cross-validated training on available sets of structurally resolved and non-redundant alpha and beta membrane proteins, we demonstrate that membrane domain prediction methods based on such a compact representation outperform approaches that utilize explicitly evolutionary profiles and multiple alignments. Moreover, using an external evaluation by the TMH Benchmark server we show that our final prediction protocol for the TM helix prediction is competitive with the state-of-the-art methods, achieving per-residue accuracy of about 89% and per-segment accuracy of about 80% on the set of high resolution structures used by the TMH Benchmark server. At the same time the observed rates of confusion with signal peptides and globular proteins are the lowest among the tested methods. The new method is available on-line at http://minnou.cchmc.org.

Contact: jmeller@chmcc.org

1 INTRODUCTION

Due to difficulties in applying experimental techniques, such as X-ray crystallography or NMR, the number of high resolution structures of membrane proteins that have been solved to date is still very limited, especially when compared with the number of resolved soluble proteins. As of December 2004, there were only 86 unique integral membrane proteins with known 3D structures in the Protein Data Bank, of which 71 were alpha-helical and 15 were beta-barrel proteins (http://blanco.biomol.uci.edu/mpex/), respectively. On the other hand, it is estimated that integral membrane proteins constitute about 20 to 30% of all proteins in the sequenced genomes (Wallin and von Heijne, 1998). The computational prediction of membrane proteins has therefore become an important alternative and complementary tool for genomic annotations and membrane protein studies.

The reader is referred to (Chen and Rost, 2002) for a comprehensive review of the state of the art in membrane protein prediction and concepts underlying different prediction protocols. Many methods for transmembrane helix prediction rely primarily on hydropathy scales, observed preferences of amino acids for membrane proteins and other physico-chemical properties of amino acids in order to identify sufficiently long stretches of mostly hydrophobic residues that may coincide with TM helices. Successful examples of such methods include SOSUI (Hirokawa et al., 1998), TopPred (von Heijne, 1992) and TMpred (Hofman and Stoffel, 1993).

Another very successful approach to membrane protein prediction is based on Hidden Markov Models (HMM) that allow one to model the global topology of membrane domains, rather than simply identifying local membrane segments. Highly successful examples of such methods include TMHMM (Krogh et al., 2001), Phobius (Kall et al., 2004) and HMMTOP (Tusnady and Simon, 1998, 2001). Either individual sequence or evolutionary profiles of protein families, as encoded by their multiple alignment (MA), may be used in this approach (see, e.g., Viklund and Elofsson, 2004). Alternative group of methods uses advanced machine learning techniques, such as Neural Networks (NN), in order to map the single sequence-based or evolutionary information about an amino acid residue and

its context into the classification (e.g. membrane vs. non-membrane). The NN-based PHDhtm method (Rost, 1996), which incorporates evolutionary profiles, is a particularly successful example of such an approach.

An independent assessment of state-of-the-art TM helix prediction methods, referred to as the TMH Benchmark, was recently performed in the Rost Lab (Kernytsky and Rost, 2003). The top performing methods (both in terms of sensitivity and specificity), such as PHDhtm or TMHMM, achieved the per-residue accuracy of up to 80% and per-segment accuracy of up to 84%, with the estimated 1% and 20-30% rate of false positive matches for globular proteins and signal peptides, respectively. Other recent studies also pointed out the overall limited accuracy of TM domain prediction (see, e.g. Chen and Rost, 2002; Chen et al., 2002; Moller et al., 2001). At the same time, published estimates and benchmark results (Viklund and Elofsson, 2004; Kernytsky and Rost, 2003; Chen and Rost, 2002) suggest that methods that employ evolutionary information appear to be more accurate than methods based on information derived from a single sequence.

While similar concepts apply to prediction of beta-barrel membrane proteins as well, this problem appears to be more difficult due to a weakly hydrophobic nature of membrane spanning beta-strands. Another fundamental limitation comes from the very limited number of known prototypes available for training. Nevertheless, a number of methods for prediction of beta-membrane proteins, both based on single sequences and evolutionary profiles, have been proposed and they are being used to annotate newly sequenced genomes (see, e.g., Wimley, 2002; Casadio et al., 2003; Bigelow et al., 2004).

The prediction methods considered here represent an amino acid residue and its environment using a sliding window of certain length. Typically, in MA-based prediction methods each residue in a sliding window is represented by a column of 20 substitution scores from a position-specific scoring matrix (PSSM) obtained using Psi-BLAST (Altschul et al., 1997). Such generated PSSMs effectively represent the frequencies of different amino acid substitutions at a given position in a protein family. Consequently, a MA-based representation may imply several hundred attributes to describe a residue (e.g. 420 numerical values when using a sliding window of length 21). In conjunction with certain machine learning techniques, such as NNs, this may lead to thousands of free parameters to be optimized. Higher complexity of the model may, in turn, hinder our ability to train a successful prediction protocol with good generalization properties, especially if only a limited number of training examples is available. The latter is particularly relevant in case of membrane protein prediction.

In this paper, we consider an alternative strategy to membrane protein prediction, which is based on a compact representation of an amino acid and its environment and may be applied to improve recognition of both: helical and beta transmembrane domains. Instead of using directly evolutionary profiles we propose to use prediction-based "structural profiles" consisting of predicted relative solvent accessibility (RSA) and secondary structure (SS) of each residue. This initial prediction step may be viewed as an effective projection of the information encoded by MA into a reduced representation defined by the predicted RSA/SS profiles.

We used here our recently introduced, accurate real valued RSA prediction method, which is available through the SABLE server (Adamczak et al., 2004). The SABLE server, which was rigorously evaluated by the EVA meta-server for evaluation of SS prediction servers (Eyrich et al., 2001) is also used to generate state-of-the-art SS predictions (Adamczak et al., 2005). Each amino acid residue is thus represented by up to five numbers: the predicted real valued RSA, confidence of RSA prediction and predicted probabilities for each of the three secondary structure classes (i.e. helix, beta-strand and coil or other).

This compact representation stemmed from the observation that a method that was trained on soluble proteins only (Adamczak et al., 2004) in order to effectively identify residues buried in the hydrophobic core of globular proteins, may, in fact, be also used to indicate residues that are "buried" in the membrane (see also Figure 1 in the Supplementary Materials). In addition, secondary structure prediction methods trained on soluble proteins achieve, perhaps surprisingly, a relatively high (comparable to that for soluble proteins) accuracy on membrane proteins (see Table 5 in the Supplementary Materials). The latter may indicate interesting hints as to how membrane proteins fold and stability of their secondary structures also in an aqueous environment.

The accuracy of the new approach is estimated using both: cross-validated training and the TMH Benchmark evaluation server (Kernytsky and Rost, 2003). Alternative representations and classification models are assessed using several different machine learning techniques, including Linear Discriminant Analysis (LDA) and Neural Networks. In particular, the MA-based representation is directly compared with the reduced representation proposed here.

2 SYSTEM AND METHODS

2.1 Training sets

In order to derive a non-redundant and as much representative as possible set of membrane proteins of known structures, we explored the MPtopo membrane protein database, which provides a comprehensive and up-to-date collection of membrane proteins with experimentally verified topologies (Jayasinghe et al., 2001). Using MPtopo, version as of June 2004, we initially obtained 167 protein chains, including 101 3D_helix chains, 38 1D_helix chains and 28 3D_other chains.

In MPtopo terminology, 3D_helix and 1D_helix refer to helical membrane proteins with or without resolved 3D structure, respectively (in the latter case only a low resolution information derived from various experimental studies is known). The 3D_other set includes, in turn, beta–barrel and monotopic membrane proteins with known 3D structures (Jayasinghe et al., 2001). This initial data set was subsequently reduced as follows. First, we excluded the low resolution structures from the 1D_helix set. It has been recently shown that the location of TM helices in this set is likely to be identified with the help of prediction methods (Chen et al., 2002), making them rather unsuitable for training. The above choice reduces the number of examples available for training significantly. However, at the same time, our compact representation of the amino acid environment and limited number of parameters to optimize make this problem somewhat less severe than in case of alignment-based methods, while at the same time offering a possibility of reducing the level of noise in the class assignment. It should be noted, however, that even for the structurally resolved membrane proteins the identification of the membrane segment boundaries is laden with some uncertainty (Tusnady et al., 2004).

For the remaining protein chains we used sequence alignment to identify homologous chains and remove redundant entries from the training set. Specifically, the BLASTP program (Altschul et al., 1997) with the BLOSUM62 substitution matrix was used to generate pairwise alignments. Sequences resulting in matches with E-values smaller than 10^{-10} were removed. However, four relatively distant homologs, resulting in matches with E-values between 10^{-3} and 10^{-10}, were left in the training set. We also removed the monotopic membrane proteins from the 3D_other set in order to obtain a suitable training set for beta–barrel membrane proteins. As a result, a set of 73 alpha-helical membrane protein chains (with 15,598 residues) and 15 beta-barrel membrane proteins (with 4,623 residues) were obtained, respectively.

In addition to results obtained using such reduced non-redundant set of membrane proteins, we also augmented the training set by adding examples of signal peptides and falsely classified (after the first iteration of the training and validation procedure) globular proteins. This augmented set is used to train the final prediction system and to further reduce the confusion with signal peptides and globular proteins. For that purpose, we randomly selected 15 signal peptides: 5 from eukaryotic proteins, 5 from gram positive, and 5 from gram negative bacteria, included in the PrediSi database of signal peptides (Hiller et. al., 2004). For each of the proteins with a signal peptide attached we included in the training 100 residue-long fragments that comprised the signal peptides and the adjacent fragment of the protein, next to the cleavage site.

Furthermore, we used our intermediate prediction system, which was trained using the initial, non-augmented training set, to identify relatively long segments of globular proteins falsely predicted to be in the membrane. We used a control set of globular proteins, which was independent of the training set used for the optimization of our RSA and SS prediction methods, and found a set of 13 soluble proteins with (in general multiple) false membrane segments. These fragments, together with flanking sequencing of length 15 residues on both sides, were then added to the augmented training set. Therefore, our augmented training set used to retrain the final TM helix prediction system consisted of membrane protein chains, globular protein fragments, and signal peptides, comprising together 18,637 residues. The globular proteins and signal peptides used to identify false positives for augmented training were subsequently excluded from our validation sets.

2.2 Architecture and training of NNs

In order to compare the results of linear and non-linear classifiers we used in this study both: LDA and NN approaches. The LDA-based classifiers were trained using the Tooldiag package (Rauber et al., 1993). The results of LDA are briefly discussed in the next section, as they are used primarily as a reference to assess the non-linear models. Here, we focus on the setup of NNs that were used to derive our final prediction system and for cross-validation study, assessing relative contributions of the multiple alignment, hydropathy scales, and the novel prediction-based "structural profiles".

The architecture of all NNs used here is similar. Namely, a feed forward topology with three layers: the input layer, one hidden layer, and the output layer, is used. The adjacent layers are fully connected and the logistic activation function for the nodes in the hidden and output layers is used. The number of features used to represent each amino acid residue (and thus the size of the input layer) varies between one and six in our tests for models that do not use evolutionary profiles and 20 for MA-based methods (see also Tables 2 and 3). For example, when five features per amino acid residue are used, the input consists of up to 155 numbers, which represent amino acid residues in a sliding window of length up to 31 (the longest sliding window tested here). The two output nodes represent the class of "membrane residues" and "non-membrane residues", respectively. Each residue (input vector) is assigned to a class with a larger excitation of its output node.

All the networks were trained using the Rprop (Riedmiller and Braun, 1993) algorithm, as implemented in the SNNS package (Zell et al.). The order of training examples was random and the number of training iterations (epochs) was set to 500 since no significant improvement in terms of the sum of squares error function was observed in a longer training for the networks considered here. For each

of the representations and sliding windows, we trained a number of networks with a different number of nodes in the hidden layer that was varied between 8 and 18 (with an increment of 2). In the cross-validation study, which aims at assessing relative merits of different representations discussed here, multiple networks with a different number of nodes in the hidden layer were trained and evaluated. Alternative representations imply different size of the input vectors and may require a different number of nodes in the hidden layer in order to achieve the best performance. Therefore, for a fair comparison between different representations, the results for networks with a number of nodes in the hidden layer that yielded the best generalization in each case are included in Table 1.

For helical membrane proteins, the cross-validation involves splitting the training set into ten subsets, each consisting of several protein chains with approximately equal number of residues in each subset. Similarly to standard 10-fold cross-validation, the training is then performed 10 times using different unions of nine subsets and the remaining subset as a validation set. The results on controls sets are averaged and are referred to as 10-fold cross-validation results (even though it is not strictly correct). Protein (rather than residue)-based definition of training and control sets makes it more likely to observe specific patterns resulting from distinct membrane domains in a control set only. Thus, this approach is expected to yield a more realistic assessment of generalization for novel membrane proteins that have not been resolved structurally as yet. For beta-barrel membrane proteins we used leave-one-out (protein) procedure to estimate the accuracy of alternative representations.

In addition to a simple NN-based classifier developed and assessed in the cross-validation study, we also developed a multistage protocol for enhanced prediction of transmembrane helices (a similar system for beta-barrel membrane proteins is a subject of a future study). 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 our compact prediction-based "structural profiles". Because inclusion of hydropathy scales led in our tests to a higher level of confusion with globular proteins and signal peptides, we also excluded hydropathy from the representation of an amino acid residue. 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.

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 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 in detail in the Supplementary Materials.

While the second layer NN leads 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 prediction. Basically, the final filter is applied to either split predicted long TM helices or delete too short ones (the details are discussed in the Supplementary Materials). Similar filters have been used before by other groups (see, e.g., Rost et al., 1995).

3 RESULTS AND DISCUSSION

3.1 Cross-validation study

We used non-redundant training sets of alpha-helical and beta-barrel membrane proteins, as described in the previous section and 10-fold (or leave-one-out) cross-validation in order to train and evaluate both LDA-based and NN-based classification systems. In Table 1 we compare results obtained using NNs and different sliding windows for two alternative representations considered here. The novel compact representation is estimated to achieve the per-residue classification accuracy of 88% and 78% and correlation coefficients of 0.73 and 0.53 for prediction of transmembrane helices and beta membrane proteins, respectively. For comparison, the MSA-based prediction achieves in cross-validation per-residue classification accuracy of up to 87% and correlation coefficient of 0.7 for alpha-helical and 74% and 0.42 for beta proteins, respectively.

Table 1 Accuracy of membrane protein prediction using alternative representations, consisting of predicted RSA/SS profiles and MA-based evolutionary profiles, are compared using cross-validation on non-redundant sets of alpha-helical and beta-barrel proteins for three different sizes of the sliding window (11, 21 and 31 residues, respectively). Averaged two class per-residue classification accuracy (Q_2), Matthews correlations coefficients (MCC) (Matthews, 1975) and standard deviations are included.

	Alpha-helical		Beta-barrel	
Features	Q_2 %	MCC	Q_2 %	MCC
RSA+SS (11)	87.9±0.8	0.74±0.02	77.9±3.3	0.50±0.08
RSA+SS (21)	88.0±0.6	0.73±0.02	78.7±3.3	0.53±0.08
RSA+SS (31)	87.4±0.7	0.73±0.02	77.9±3.6	0.53±0.08
MA (11)	85.0±1.3	0.67±0.03	71.6±2.9	0.37±0.07
MA (21)	86.0±1.4	0.70±0.03	73.3±3.4	0.41±0.08
MA (31)	86.5±1.4	0.70±0.03	73.6±3.6	0.42±0.09

It is interesting to note that the differences in accuracy between the two alternative representations as well as error

bars (as measured by standard deviations from cross-validated training) are significantly higher for beta-barrel proteins, reflecting a very limited number of training examples in this case and highlighting the problems with reliable parameter estimation. While the difference in accuracy is not as large for alpha-helical proteins, for which the number of training examples is significantly higher, the prediction-based structural profiles still yield improved accuracies and lower error bars in cross-validation, despite (or perhaps thanks to) much simpler representation.

The error bars observed in cross-validation and the drop in accuracy between training and validation sets may also be used to assess the level of overfitting for both representations. In general, a higher variability on the control sets (e.g. 1.4% for MA-based vs. 0.7% for RSA-based representation in case of alpha-helical proteins) and higher accuracy in the training with respect to control sets indicate a higher level of overfitting. In that regard, the classification accuracy in the training (as measured by the average accuracy on ten training sets used in cross-validation) is only about 3% higher than on control sets in case of the novel RSA-based representation, as opposed to about 5% difference in case of the more complex MA-based models. Given that some of the proteins included in our limited training and control sets are likely to share common characteristics (despite the lack of sequence homology), the above estimates of the level of overfitting are expected to be overly optimistic. Nevertheless, the relative differences between the two representations are clear and reinforce our proposition that the novel compact representation enables more reliable parameter estimation for prediction of transmembrane domains.

Note also, that the differences in accuracy between shorter and longer sliding windows considered in Table 1 are not statistically significant, even though the error bars tend to be somewhat higher for longer sliding windows. In principle, longer sliding windows imply more parameters to be estimated from the limited data and are more prone to overfitting. However, for fair comparison between the two alternative representations, we report here the best results achieved for each size of the sliding window by selecting the optimal number of the nodes in the hidden layer (see Section 2.2). In contrast, if the topology of the network is fixed, implying a monotone increase of the number of parameters with the size of the sliding window, then a decrease in accuracy is observed in cross-validation for windows longer than 20 residues. A significant drop in accuracy is also observed for very short windows (for details see Supplementary Materials). In that context, it is also important to realize that the distribution of lengths of TM segments is quite wide, with many "non-canonical" TM helices providing difficult to classify prototypes. Therefore, we decided to incorporate into the final consensus-based predictor networks that use different sliding windows for further smoothing of the results (see Supplementary Materials).

Further results are summarized in Table 2. Several compact representations are compared, including the simple hydropathy scale based methods (with one attribute per amino acid in the sliding window), the RSA and SS predictions alone (with two and three attributes per residue, respectively) and the combined profiles. A sliding window of length 25 is used for this comparison since it was found to be optimal for the compact representations introduced here. For both alpha-helical and beta-barrel membrane proteins the RSA and SS based structural profiles perform significantly better than a simple hydropathy-based method (with the KD scale (Kyte and Doolittle, 1982) working somewhat better than the "biological" scale proposed recently (White and Wimley, 1999)). Combining RSA and SS predictions with hydropathy profiles does not result in further increase in accuracy. On the other hand, the RSA and SS predictions alone are sufficient to achieve performance close to that of the best combination of features.

Table 2 Accuracy of membrane protein prediction using structural profiles including: KD and WW hydropathy profiles (Kyte and Doolittle, 1982; White and Wimley, 1999), predicted RSA and SS, estimated using cross-validation on non-redundant sets of alpha-helical and beta-barrel proteins.

	Alpha-helical		**Beta-barrel**	
Features	Q_2 %	MCC	Q_2 %	MCC
H_{KD}	84.2±1.5	0.67±0.02	67.8±2.0	0.32±0.04
H_{WW}	82.6±1.5	0.63±0.02	67.3±2.6	0.31±0.04
SS	81.2±5.0	0.66±0.06	77.0±3.9	0.52±0.08
RSA	86.8±0.8	0.71±0.02	76.0±2.2	0.42±0.08
SS+RSA	88.2±0.6	0.74±0.02	77.4±3.7	0.53±0.08
SS+RSA+H_{KD}	87.8±0.7	0.74±0.02	77.7±3.1	0.53±0.07
SS+RSA+H_{WW}	88.0±0.8	0.74±0.01	77.4±3.3	0.53±0.07

Even though we do not present a detailed analysis of the LDA results here, it is interesting to note that the same trends are observed in that case as well (with the overall level of classification accuracy lower by about 4%). The fact that cross-validated accuracy of the LDA based classification is significantly lower, suggests that the more general, non-linear characteristics of NN-based classifiers play an important role in this case. On the other hand, however, the risk of overestimating the accuracy increases when using more complex models. From that point of view, the accuracy of the simple LDA model with its 125 free parameters (when using five attributes per residue and a sliding window of length 25), provides an additional support for the hypothesis that the compact representation proposed here, consisting of prediction-based structural profiles, is likely to contribute to a further increase in accuracy of membrane domain prediction methods. We also hypothesize that the new representation may prove advantageous over explicit use of evolutionary profiles not only in the context of machine learning-based methods, as directly tested here, but also in the context of grammar-based methods. This is the subject of a future work.

3.2 TMH Benchmark assessment

In this section we present the evaluation of our final two-stage NN-based prediction system for prediction of transmembrane helices. The new method will be referred to as "Membrane protein IdeNtificatioN withOUt explicit use of hydropathy profiles and alignments" (MINNOU). Using the TMH Benchmark server, we assess both sensitivity and specificity of the new method (especially in terms of confusion with globular proteins and signal peptides that may be incorrectly predicted as having TM segments). At the same time, the performance of MINNOU is compared with that of other state-of-the-art prediction methods. Table 3 summarizes the results on a set of high resolution structures, with well defined boundaries of transmembrane helices. The levels of confusion with globular proteins and signal peptides are also shown in Table 3, whereas Table 6, included in the Supplementary Materials, illustrates contributions due to different steps in our multistage protocol.

As can be seen from Table 3, MINNOU achieves the highest per residue accuracy (89%) among the methods included in the TMH Benchmark evaluation. At the same time per segment accuracy (80%) is worse than that of PHDhtm and HMMTOP2. It should be noted, however, that per segment accuracy is very sensitive to falsely predicted short TM helices and short non-membrane segments in true TM helices. This can be also seen from the effects of including a filter that removes some of these incorrectly predicted short segments, without affecting significantly per residue accuracy (see Table 6). Therefore, further improvement in that regard is likely to be achieved by optimizing this step.

We would also like to comment that several methods (including the two mentioned above) achieve a higher per residue accuracy (89-90%, as opposed to 85% for MINNOU) on the set of low resolution structures included in the TMH Benchmark. These structures were not included in our original training set because of the uncertain assignment of their membrane segments. However, for comparison we performed a cross-validated training using both high and low resolution structures and observed a decreased accuracy on such joint set (by about 2% in terms of classification accuracy and 0.03 in terms of correlation coefficient). Thus, the two sets of TM segments appear to have distinct characteristics. This is further highlighted by much narrower (with respect to high resolution structures) distribution of lengths of the TM segments derived from the low resolution structures (see Figure 2 in Supplementary Materials). The fact that some of the prediction methods actually achieve higher accuracy on low resolution than on high resolution structures could indicate that prediction methods may have played a role in delineating TM segments in these low resolution structures, as suggested before by (Chen et al., 2002).

It should be also noted that due to the very small number of structurally resolved membrane proteins, any benchmark is likely to use for evaluation proteins homologous to those used by most of the evaluated methods (including MINNOU in case of high resolution structures) for training. Therefore, the observed levels of accuracy are, in fact, based on variations of the training set and are unlikely to hold in the future. Nevertheless, TMH Benchmark is a very useful resource for independent (static) evaluation of the results and comparison between different methods. Moreover, the TMH Benchmark evaluation revealed significant levels of confusion with globular proteins and even higher levels of confusion with signal peptides. It is encouraging that the new method appears to be significantly better in that regard than any other method evaluated in (Kernytsky and Rost, 2003), which is partly to be attributed to the use of an augmented training set. In the latter aspect, MINNOU is similar to another recently published Phobius method (Kall et al., 2004).

Table 3 Assessment of TM helix prediction methods using the TMH Benchmark server. Per segment (Q_{OK}) and per residue (Q_2) classification accuracies (Chen et al., 2002) and confusion levels with signal peptides (cSP) and globular proteins (cGP) are given in the units of per cent.

Method	Q_2	Q_{OK}	cGP	cSP
MINNOU	**89**	**80**	**1**	**8**
PHDhtm	80	84	2	23
HMMTOP2	80	83	6	48
TMHMM1	80	71	1	34
DAS	72	79	16	97
TopPred2	77	75	10	82
SOSUI	75	71	1	61

We would like to point out that we also performed an independent analysis of the specificity of our predictions. Using non-redundant sets of signal peptides and globular proteins we found good agreement with the low estimates of confusion observed in Table 3. For example, on a set of 314 non-redundant soluble proteins that had no homology to proteins used in the training of our SABLE prediction method (a subset of proteins used before in (Adamczak et al., 2005) in order to evaluate SABLE), only three proteins were falsely predicted to have membrane segments.

Another interesting observation is a relatively high accuracy of MINNOU predictions for ion channel proteins, which are characterized by occurance of very long and relatively short membrane helices. For seven ion channels included in the set of high resolution structures used for training, MINNOU achieved the average per residue accuracy of 92.0% and correlation coefficients of 0.81, as opposed to 86.4% and 0.68 for HMMTOP2 or 85.4% and 0.68 for DAS, for instance. For a newly solved ion channel structure, 1ots, which is however homologous to one of the proteins included in the training set, MINNOU also achieved significantly higher accuracy than any other method (correlation coefficient of 0.67 as opposed to the

second best of 0.44). While these differences are statistically not significant and are clouded by the lack of truly indepndent test sets, we find these trends and the ability of MINNOU to predict largely correctly membrane segments in ion channels, without losing the ability to make relatively accurate predictions for other types of helical TM proteins, rather encouraging.

Table 4 Accuracy of different methods as measured by per-residue accuracy and Matthews correlation coefficients (second line in each row) on a set of five helical membrane proteins not included in the training set (including two, 1u7c and 1xfh, that are not homologous to proteins included in the training).

Method	1tn0_A	1vfp_A	1umx_L	1u7c_A	1xfh_A
HMMTOP2	75.6	88.3	81.9	79.0	70.0
	0.51	0.62	0.64	0.59	0.38
SOSUI	82.8	91.8	88.3	78.0	66.7
	0.64	0.74	0.76	0.56	0.34
TopPred2	80.0	88.5	91.8	85.2	63.1
	0.59	0.64	0.83	0.71	0.27
DAS	80.4	91.0	84.0	82.1	68.5
	0.62	0.70	0.66	0.66	0.33
MINNOU	**84.8**	**91.4**	**77.6**	**75.6**	**65.5**
	0.68	**0.75**	**0.63**	**0.52**	**0.34**

Finally, in order to illustrate the performance of several top ranking methods on individual proteins we used a set of five recently solved membrane proteins. The results are shown in Table 4. Three out of these five proteins (including a bacteriorhodopsin structure, 1tn0, and a photosynthetic reaction center protein, 1umx) exhibit homology to those included in the training and are merely used to show the variation in accuracy observed for all the methods on different proteins. It is interesting to note that none of these methods appears to be clearly better and they all have failed quite badly for one of the non-redundant new proteins, a glutamate transporter 1xfh, for which the highest correlation coefficient is only 0.38 (the MINNOU prediction for 1xfh, which is at level of other methods, is included in Figure 1 in the Supplementary Materials). One should note, however, that the assignment of TM segments for these newly solved proteins is based on a theoretical analysis of the structures using the method by (Tusnady et al., 2004), and it is, thus, laden with additional uncertainty. Nevertheless, we believe that the limited accuracy of the top ranking methods included in Table 4 further underscores the need for continued development of improved methods for membrane domain prediction.

4 CONCLUSION

We proposed a novel representation of an amino acid residue and its environment for membrane protein prediction. The new approach does not use explicitly evolutionary profiles or hydropathy scales. Instead, the new method relies on prediction-based structural profiles, consisting of predicted relative solvent accessibility and secondary structures of amino acid residues. In particular, the predicted level of aqueous solvent exposure, which is indicated by an accurate RSA prediction method trained on soluble proteins only (Adamczak et al., 2004), is used to identify segments of residues that are "buried" in the membrane in order to "avoid" contact with water.

In cross-validation with a simple, one layer NN-based classifier, the new representation is estimated to yield an accuracy of 0.74 for TM helix and 0.53 for beta membrane prediction, as measured by the correlation coefficient between the predicted and observed classes. For comparison, the MA-based prediction is estimated to achieve a lower accuracy, with correlation coefficients of 0.70 for alpha-helical and 0.42 for beta proteins, respectively. The final prediction protocol for TM helix prediction, based on two-step NN-based classifier, is estimated by the TMH Benchmark server to achieve per-residue accuracy of 89% (significantly higher than any of the methods evaluated in (Kernytsky and Rost, 2003)) and per-segment accuracy of 80%, with the lowest rates of confusion with globular proteins and signal peptides among the methods tested (and similar to the recently published Phobius method (Kall et al., 2004)).

Thus, using the new representation we were able to achieve accuracy competitive with that of other state-of-the-art methods for alpha helical TM domains, as assessed by the TMH Benchmark server. High sensitivity of TM domain prediction is achieved with very low levels of confusion with globular proteins and signal peptides. Moreover, in our internal cross-validation tests, the new representation outperformed multiple alignment-based approaches for both: alpha helical and beta-barrel membrane proteins. Therefore, we conclude that applying predicted RSA and SS is likely to further contribute to the development of accurate methods for prediction of protein membrane domains.

ACKNOWLEDGMENTS

This work has been supported by NIH grants AI055338 and 5R01GM067823-02.

REFERENCES

Adamczak, R., Porollo, A. and Meller, J. , (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753-67.

Adamczak, R., Porollo, A., and Meller, J. (2005) Combining Prediction of Secondary Structures and Solvent Accessibility in Proteins, *Proteins,* **59**:467-75.

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

Bigelow H.R., Petrey, D.S., Liu, J., Przybylski, D., and Rost, B., (2004), Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res.*, **32**, DOI: 10.1093/nar/gkh580

Casadio, R., Fariselli, P., Finocchiaro, G., and Martelli P. L. (2003) Fishing new proteins in the twilight zone of genomes: The test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria. *Protein Sci.*, **12**(6): 1158 - 1168.

Chen, C.P. and Rost, B., (2002) State-of-the-art in membrane protein prediction. *Applied Bioinformatics* **1**, 21-35.

Chen, C.P., Kernytsky, A., and Rost, B., (2002) Transmembrane helix predictions revisited, *Protein Sci.*, **11**, 2774-91.

Eyrich, V., Marti-Renom, M., Madhusudhan, M., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B., (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics,* **17**: 1242-1243.

Hiller, K., Grote, A., Scheer, M., Münch, R., and Jahn, D., (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res., 32, W375-W379

Hirokawa, T., Seah, B.C. and Mitaku. S., (1998) SOSUI: Classification and Secondary Structure Prediction System for Membrane Proteins. *Bioinformatics*, **14**, 378-9.

Jayasinghe, S., Hristova, K., and White, S. H., (2001) MPtopo: A database of membrane protein topology. *Protein Sci.*, **10**, 455-8.

Jones, D., (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.,* **292**: 195-202.

Kall L., Krogh, A., and Sonnhammer, E.L.L., (2004) A combined transmembrane topology and signal peptide prediction method, *J. Mol. Biol*., **338**, 1027-36.

Kernytsky, A. and Rost, B., (2003) Static benchmarking of membrane helix predictions. *Nucleic Acids Res*., **31**, 3642-4.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567-580.

Kyte, J. and Doolittle, R. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*. **157**, 105-132.

Matthews, B.W., (1975) Comparison of predicted and observed secondary structure of T4 ohage lysozyme. *Biochim Biophys Acta* **405**, 442-451.

McGuffin, L.J., Bryson, K., Jones, D.T., (2000) The PSIPRED protein structure prediction server. *Bioinformatics*. **16**, 404-405

Moller, S., Croning, M.D.R. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Rauber, T.W., Barata, M.M., and Steiger-Garcao, A.S., (1993) A Toolbox for Analysis and Visualization of Sensor Data in Supervision. Proceedings of the International Conference on Fault Diagnosis, Toulouse, France.

Riedmiller M, Braun H., (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *Proc IEEE Int Conf Neural Networks,* pp. 123–134.

Rost, B., (1996) PHD: predicting one dimensional protein structure by profile based neural networks. *Meth Enzymol*, **266**:525-539.

Rost B., Fariselli, P. and Casadio, R., (1996) Topology prediction for helical TM proteins at 86% accuracy *Protein Sci*, **5**,1704-18.

Rost, B., Casadio, R., Fariselli, P., and Sander, C., (1995) Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* **4**, 521-533.

Tusnády, G. E., and Simon, I., (2001) The HMMTOP transmembrane topology prediction server" *Bioinformatics*, **17**, 849-850.

Tusnády, G. E., Dosztányi,Z. S. and Simon I. (2004), Transmembrane proteins in protein data bank: identification and classification. *Bioinformatics*, **20**, 2964-2972.

Tusnády, G.E. and Simon I., (1998) Principles governing amino acid composition of integral membrane ptoeins: application to topology prediction. *J. Mol. Biol.*, **283**, 489-506.

Viklund H, and Elofsson A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using HMMs and evolutionary information. *Protein Sci.* **13**, 1908-17.

Wallin E. and von Heijne G., (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029-38.

White S.H., and Wimley W.C., (1999) Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct.* **28**, 319-65.

Wimely, W.C., (2002) Toward genomic identification of ß-barrel membrane proteins: Composition and architecture of known structures. *Protein Sci.*, **11**:301-312.

Zell A., Mamier G., Vogt, M., et al., The SNNS users manual version 4.1 Available online at http://www-ra.informatik.uni-tuebingen.de/snns.