Improved prediction of trans-membrane spans in proteins using an Artificial Neural Network

Julia Koehler, Ralf Mueller, Jens Meiler

***Abstract*—Tools for the identification of trans-membrane spans from the protein sequence are widely used in the experimental community. Computational structural biology seeks to increase the prediction accuracy of such methods since they represent a first step towards membrane protein tertiary structure prediction from the amino acid sequence. We introduce a predictor that is able to identify trans-membrane spans from the sequence of a protein. The novelty of the approach presented here is the simultaneous prediction of trans-membrane spanning α-helices and β-strands within a single tool. An artificial neural network was trained on databases of 102 membrane proteins and 3499 soluble proteins. Prediction accuracies of up to 92% for soluble residues, 75% for residues in the interface, and 73% for TM residues are achieved. On average the algorithm predicts 79% of the residues correctly which is a substantial improvement from a previously published implementation which achieved 57% accuracy (Koehler *et al.*, *Proteins: Structure, Function, and Bioinformatics*, 2008). The algorithm was applied to four membrane proteins to illustrate the applicability to both α-helical bundles and β-barrels.**

I. Introduction

MEMBRANE proteins (MPs) account for about 30% of the proteins in the human genome and are involved in many essential functions in the cell. For instance, they act as transporters, participate in signaling pathways and function as ion-channels. Even though almost 50,000 protein structures are deposited in the ProteinDataBank (PDB), only about 900 belong to the class of MPs. This discrepancy reflects the difficulty of crystallizing MPs and they often exceed the size limitation for NMR spectroscopy. In contrast, the structures of MPs are arguably easier to predict computationally because of the constraints the membrane imposes on their fold [1].

First attempts to identify membrane spanning regions along the sequence utilize hydrophobicity scales. A free energy value of transfer from a polar medium (the cytosol) to an apolar medium (the membrane) is assigned to each of the 20 amino acids. Depending on the preference of an amino acid for a specific environment the sign of this transfer free energy value changes. For the prediction of trans-membrane (TM) spans the transfer free energy values are added over a sequence window.

There is a wealth of hydrophobicity scales available that were derived using experimental (for example Wimley & White [2, 3] or GES [4]), knowledge-based (UHS [5]), and consensus approaches (Kyte & Doolittle [6]). The scales are mostly derived considering two phases: solution (SOL) and membrane (TM). Only two scales [5, 7] include a third interface or transition region (TR). Potentials considering the depth of the residue in the membrane bilayer have also been reported for α-helical MPs [8, 9].

The differences between various hydrophobicity scales can be explained by the different experimental setups used during their derivation. Wimley & White for instance examine unfolded peptides in solution and membrane bilayer [2, 7] whereas Hessa et al. consider folded proteins [8, 10].

Hydrophobicity scales that include an interface region between solution and membrane are rare even though three-state scales have a higher information content than two-state scales. In addition, three-state scales are able to provide information about the location of the polar headgroups of the membrane lipids, which are distinctly different than the soluble phase. Further, three-state scales can identify amphipathic helices located in the interface region. Wimley & White experimentally derived a hydrophobicity scale using penta-peptides that were unfolded in all three phases [2, 7]. For this reason the unsaturated hydrogen-bonds in the membrane bilayer lead to a bias of this scale towards solution. On average ~50% of the residues are correctly predicted in this three-state scenario. We derived a knowledge-based hydrophobicity scale from a database of known MP structures containing both α-helical and β-barrel proteins [5]. This Unified Hydrophobicity Scale (UHS) yields accuracies of ~57% in the three-state prediction scenario. Both scales were tested on a database containing both α-helical bundles and β-barrel proteins. A Mammalian Hydrophobicity Scale (MHS) was derived from 16 α-helical bundles and yields accuracies of ~61% tested on an only α-helical database [5].

Subsequent specialized prediction tools for TM spans use machine learning techniques such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs), or Support Vector Machines (SVMs). According to Cuthbertson et al.

This work was supported in part by grant R01-GM080403 from the National Institute of General Medical Sciences.

Abbreviations: TM: trans-membrane; TR: transition; SOL: solution/ soluble; ANN: Artificial Neural Network; HMM: Hidden Markov Model; SVM: Support Vector Machine; PDB: ProteinDataBank; PDBTM: ProteinDataBank for trans-membrane proteins; MP: Membrane Protein.

Corresponding author: Jens Meiler, Center for Structural Biology, Vanderbilt University, Nashville, TN, 37232, USA; (phone: 615-936-5662; fax: 615-936-2211; e-mail: jens.meiler@vanderbilt.edu; html: www.meilerlab.org).

J. Koehler and R. Mueller are both in the Center for Structural Biology at Vanderbilt University, Nashville, TN, 37232, USA.

978-1-4244-2756-7/09/$25.00 ©2009 IEEE

[11] Split4 [12], TMHMM2 [13], and HMMTOP2 [14, 15] are the most successful TM α-helix prediction tools available. Split4 [12] uses basic charge clusters and amino acid attributes to define the correct topology of the helices. TMHMM2 [13] is an HMM trained on a dataset of 160 both single- and multi-spanning proteins and has according to their developers 97% accuracy. HMMTOP [14] utilizes the evolutionary information of multiple-sequence alignments and is based on the notion that topology is governed by the difference of the amino acid distributions in different parts of the protein rather than the amino acid composition itself. The successor HMMTOP2 [15] incorporates experimental information into the topology prediction. Other methods include PhDhtm [16] (which uses two consecutive ANNs and multiple-sequence alignments), TMMOD [17] (which is based on TMHMM, but differs in training procedure and loop models), and TopCons [18] (a consensus prediction server combining five different predictors).

The most successful methods for β-barrel proteins are according to Bagos et al. [19] HMM-B2TMR [20] and PROFtmb [21, 22], both HMM-based methods. HMM-B2TMR is sequence-profile based and therefore uses multiple-sequence alignments. A dynamic programming algorithm is employed for optimization of the location of TM segments. PROFtmb is also profile-based and is trained on eight non-redundant β-barrels. Their developers state a four-state accuracy of 86%. Bagos and co-workers tested the performance of various combinations of β-barrel predictors and implemented the best-performing consensus predictor as ConBBPRED.

Objective of this work is to establish the first integrated tool that identifies both α-helical and β-strand TM spans in a single three-state prediction for the residue being either in TM, TR, or SOL region. Advantage of this method is that sequences can be screened for TM spans with a single tool. Furthermore, synergistic effects during the ANN training lead to an increased prediction accuracy.

II. METHODS

A. Creation of the databases of non-redundant protein structures

For the MP database all TM chains from the PDBTM [23] were culled using the PISCES server [24, 25] with the following parameters: sequence identity <= 25%, resolution 3Å, R-factor 0.3, sequence length 40-10,000 residues, non-X-ray entries as well as Cα-only entries were included, and the PDB was culled by chains. Thereafter, structures derived from electron-microscopy data were excluded due to low resolution resulting in a database of 102 proteins with 136 polypeptide chains. The PDB files were downloaded from the PDBTM.

For the definition of the TM, TR, and SOL regions a fixed membrane thickness of 20Å (TM region) followed by a 10Å TR region was used. Furthermore, a 2.5Å gap region between the TM/TR regions as well as TR/SOL regions was introduced to more cleanly distinguish between the different environments (see ref. [5]). This procedure was implemented rather than using the membrane thickness given by the PDBTM (determined by the TMDET algorithm [26]) in order to avoid a recurrent influence of this predicted membrane thickness onto our method. The resulting database contained 28,379 residues in total, 9,510 residues being in the TM region, 9,079 classified as TR, and 9,790 classified as SOL. A total of 3,882 residues residing in the gap region were excluded from the training process to minimize noise due to incorrect assignment to regions.

Even though the MP database contained a large fraction of soluble residues a soluble protein database was established to account for different properties of soluble proteins that are not equally represented by the soluble parts of the MPs (like solvent-accessible surface area, compactness, length of secondary structure elements).

For the soluble protein database the entire PDB was culled with the PISCES server [24] using the same parameters as above with two exceptions. Due to the much larger size of the database a resolution limit of 2Å was used. Moreover, we excluded non-X-ray and Cα-only entries. The resulting database contained 3,499 proteins with a total of 3,623 polypeptide chains and 820,485 residues.

Both the MP as well as the soluble protein database were used as a basis for the input to the ANN.

B. Knowledge-based free energies for secondary structure type and membrane location were used as input

The MP database served as a basis for the derivation of knowledge-based free energies. The procedure is the same as described in [5] but updated databases allowed for more data to be included. Briefly, three-state free energies for the regions TM, TR, and SOL were derived by normalizing the amino acid frequencies in each region to 20. The propensities P [27] were then calculated by

$$P = \frac{number(region, AA) / number(region)}{number(AA) / number(total)} \qquad (1)$$

and the free energies ΔG were computed using

$$\Delta G = -RT \ln P \qquad (2)$$

with R being the gas constant, and T=293K.

The same procedure was applied to obtain the three-state free energies for the secondary structures helix, strand, and coil. The nine-state free energies for each combination of region and secondary structure type were calculated as in the three-state scenario but normalizing the amino acid occurrences to nine instead of three.

We chose to include the free energies for the secondary structure types for the prediction of the TM region since the two phenomena are interrelated: when a nascent polypeptide chain in solution reaches the membrane interface the influence of the altered dielectric environment (as described by the free energies) leads to an increased formation of backbone hydrogen bonds and therefore to the formation of secondary structure.

The obtained free energies for these different scenarios were taken as input parameters for the ANN. Furthermore, several amino acid properties such as the steric parameter, polarizability, volume, iso-electric point, the solvent-accessible surface area [28], and the position-specific scoring matrices obtained from PSIBLAST [29] were used as input parameters as they increased prediction accuracy in previous experiments [28]. PSIBLAST was run with three iterations and an E-value cutoff of 0.001.

C. Training procedure

For each dataset (i.e. for each residue) the above mentioned input parameters were employed over a sequence window of 31 residues. Therefore (20 property descriptors + 20 numbers in the PSIBLAST profile) x 31 residues = 1240 inputs were used for each dataset. The MP database (28,379 residues) served as a basis for the TM and TR region datasets, whereas the soluble protein database (820,485 residues) together with the MP database were used for the SOL region datasets. To construct the input files the residues were randomly chosen from the databases. In addition, the residues were chosen as to equally represent TM, TR, and SOL residues using an over-sampling procedure. Three dataset sizes of 9,000, 90,000, and 450,000 datasets (i.e. residues) were used for training where the training was started on the smallest dataset and consecutively increased to larger dataset sizes.

This balancing procedure was chosen to avoid an intrinsic bias of the method to predict one region over the other. It also maximizes the entropy in the training data and therefore the information content added by the ANN prediction.

For the training procedure the datasets were shuffled and then split into three subsets: 80% were used for training, 10% for monitoring the training progress, and 10% as an independent test set. Two ANNs were trained with 32 and 64 nodes in the hidden layer, respectively. The ANN with 64 nodes performed best in this case and the results are shown for this network.

The ANN is a feed-forward network with bias neurons trained with back-propagation of errors. Other network architectures have not been tested. In initial training phases the resilient propagation algorithm [30] displayed accelerated training behavior, faster convergence and higher robustness with respect to the initial training parameters than simple propagation. Therefore, the ANN was trained using the resilient propagation algorithm whereas simple propagation was used for final optimization of the weights.

D. Four examples illustrate the performance of the prediction tool

The ANN prediction was applied to four MPs not included in the training phase: two α-helical bundles and two β-barrel proteins. The crystal-structures of the potassium channel KcsA (PDB ID 1k4c) elucidated by Rod McKinnon at a resolution of 2Å was chosen as first helical example protein. Furthermore, we chose lens aquaporin-0 (PDB ID 2b6p) in the open state that was determined by Walz and co-workers at 2.4Å. Unusual structural features in both proteins are half-helices with their adjacent loops returning to the extra-membrane region. As β-barrel proteins the Outer Membrane Protein W (OmpW – PDB ID 2f1t) crystallized by Tamm and van den Berg at 3Å and the NMR structure of OmpA (PDB ID 2ge4) determined by Tamm and Bushweller were selected.

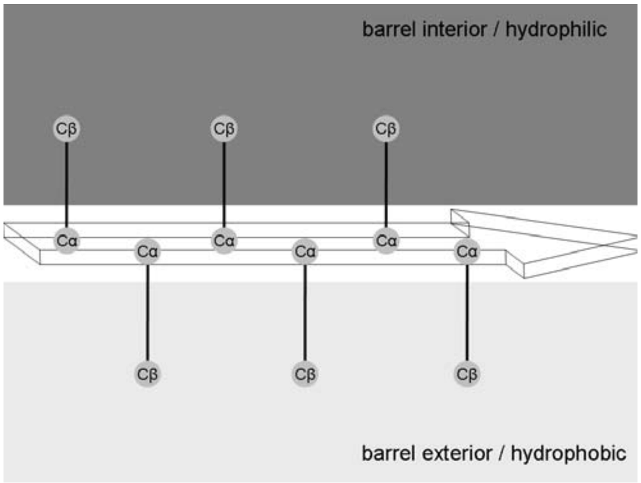

Fig. 1. This figure illustrates the reason for the failure of a simple averaging procedure of hydrophobicity values for the prediction of β-barrel MPs. The arrow indicates a β-strand with its consecutive side-chains (Cα and Cβ atoms indicated as circles) pointing in opposite directions. Averaging over these positive and negative contributions yields a negligible transfer free energy value resulting in a very small probability of predicting this stretch as a TM span.

III. RESULTS AND DISCUSSION

Most of the TM prediction methods are specialized methods for α-helical proteins. β-strand TM spans, on the other hand, are much more difficult to predict because a simple averaging procedure is less effective when consecutive side-chains alternate in facing the polar interior and the apolar exterior of the barrel (see Fig. 1). This obstacle can be overcome using machine learning techniques such as ANNs, HMMs, or SVMs that are capable of recognizing such alternating patterns while distinguishing between α-helices and β-strands at the same time. In addition, α-helices require ~19 residues to cross the lipid bilayer while β-strands require only ~9 residues. This difference results in a different optimal sequence window size for simple linear averaging strategies. However, non-linear functions like ANNs can be optimized on a single larger window (here 31 residues) to work equally well for both scenarios.

A. *Resilient propagation accelerates training*

The ANN is implemented within the Bio-Chemical-Library developed in the Meiler laboratory (www.meilerlab.org) and written in the C++ programming language. It serves as a framework for a wide variety of biomedical applications, such as *de novo* protein tertiary structure prediction [31, 32] and virtual high-throughput screening. The training was started with a small dataset (9,000 datasets). Subsequently the number of datasets was increased to 90,000 and 450,000 datasets. The ANN was trained on each dataset using the resilient propagation algorithm until the error of the monitoring dataset was minimized (see Fig. 2). Afterwards the ANN was trained in simple propagation mode for several 100 iterations to reach the RMSD minimum. This procedure became necessary as resilient propagation is known to display unstable minimization behavior close to minima in the target function [30].

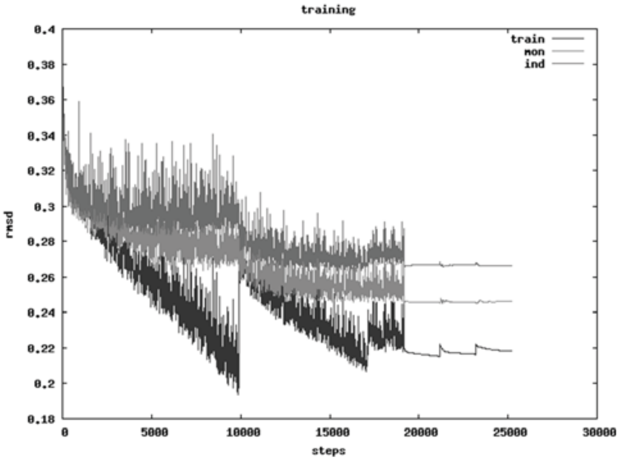

Fig. 2. The RMSD is plotted over the number of steps during the training procedure. Black indicates the RMSD of the training dataset, light gray for the monitoring dataset, and dark gray for the independent test set. The 'jumps' at 9,890 and 17,110 steps indicate a switch to a larger dataset (started with 9,000, then continued with 90,000, and 450,000 datasets). The flat line represents training using the simple propagation algorithm.

B. *Trans-membrane free energies are important for training*

Fig. 3a) shows the sum of the input sensitivities plotted over the 31 residues in the sequence window used for input. The input sensitivity is defined as a partial derivative of an output value with respect to an input variable. The values are determined numerically after ANN training is completed. As expected, the center of the sequence window has the highest impact as reflected in the increased input sensitivities. This represents the importance of the pattern immediately adjacent to the residue of interest within an α-helix or β-strand. The sensitivities converge to a smaller constant value towards the edges of the window which reflects the significance of long-range interactions within the protein. Such interactions are attributed to backbone hydrogen-bonds that stabilize β-barrel proteins as well as helix-helix contacts in α-helical bundles. The large window size facilitates capturing part of this effect. The optimal window size was determined by testing window sizes of 15, 23, 31, 39, and 47 residues with 31 residues performing best.

Fig. 3b) shows the sum of the input sensitivities for the individual input properties. The highest sensitivity is observed for the PsiBlast position-specific scoring matrices with a sensitivity of 2.0. The profile reflects evolutionary information of the protein sequence which is important for the distinction between α-helical bundles and β-barrel proteins. Furthermore, it is essential for the identification of TM spans because the likelihood for mutations contained in this profile provides information about the exposure to the polar solvent, membrane bilayer, or protein core.

Considerable influence have the free energies for the TM region, both in the three-state scenario (sensitivities TM = 1.2, TR = 0.6, SOL = 0.8) and in the nine-state scenario in conjunction with secondary structure types (see below). When considering secondary structure types the free energy for helices (sensitivity = 0.8) contains more information than for strands (0.7). Both have a higher weight than the free energy for coil residues (0.5). Similarly, if the free energies for the secondary structure types are summed over TM, TR, and SOL regions, strands contain with 3.0 more information than helices with 2.8.

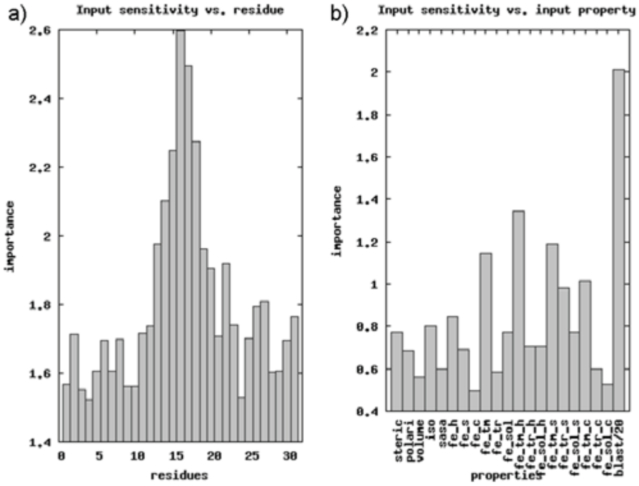

Fig. 3. In a) the sum of the weights are plotted over the residues in the sequence window. As expected, the weights for the center of the window are the largest, therefore having the most impact on the prediction. Residues at the edges of the window have less importance, although they might be involved in long-range hydrogen bonds for the prediction in β-barrels. Figure b) shows the sum of the weights versus the amino acid properties used as input for the ANN: steric = steric parameter; polari = polarizability; iso = isoelectric point; sasa = solvent accessible surface area; fe = free energy for the following secondary structure types and regions: h = helix; s = strand; c = coil; tm = trans-membrane; tr = transition region; sol = solution; blast = blast profile: the sum of the weights is normalized by 20 to represent the weight for a single amino acid.

The sensitivities for the free energies of the TM region in the 9-state scenario sum up to 3.6, whereas for the TR and SOL these sums are smaller (2.3 and 2.0, respectively). The sum of the six amino acid properties (excluding the PSIBLAST matrices) is 3.4 reflecting a smaller per property influence when compared to the free energy values. It is known, that the environment of residues plays a critical role in the formation of secondary structure. We therefore speculate that the ANN uses the free energy patterns efficiently for the identification of TM spans.

TABLE I
PREDICTION ACCURACY

		prediction		
		sol	***tr***	***tm***
observed independent	***sol (SOL proteins)***	**92.2**	5.5	2.3
	sol (MPs)	**74.9**	17.7	7.4
	tr	10.4	**74.7**	14.9
	tm	5.4	22.1	**72.6**
observed training	***sol (SOL proteins)***	**91.4**	6.2	2.4
	sol (MPs)	**80.5**	14.0	5.5
	tr	15.5	**76.8**	7.8
	tm	4.2	19.4	**76.5**

Accuracies of the prediction method on the independent and training datasets with the percentage of predicted residues in these regions. The percentage of correctly predicted residues is 79.6% for the independent and 81.3% for the training dataset. sol = solution, tr = transition region, tm = trans-membrane.

C. *Per-residue accuracy is highest for soluble region*

We have shown previously [5] that the per-residue accuracy of the Wimley-White hydrophobicity scale is ~50% for the three-state prediction scenario using a simple averaging strategy. The UHS correctly classifies up to 57% of the residues. However, it was also shown, that this averaging procedure is much less effective when identifying TM β-strands in β-barrel proteins due to the alternating hydrophobicities of consecutive amino acids. Furthermore, such a simple scheme is not able to incorporate different window lengths for helices and strands, as discussed above.

Table I shows the percentage of per-residue predictions for the three regions TM, TR, and SOL using the ANN method. The data is shown for both the independent and the training dataset. The diagonal matrix elements indicate correct predictions whereas off-diagonal elements represent false classifications. The agreement for the SOL is broken down into the accuracy for soluble proteins and MPs. It can be seen that the highest agreement is achieved in SOL for soluble proteins where 92% of the residues in the independent dataset and 91% of the residues in the training dataset are correctly predicted. For MPs the percentage agreement is lower with 75% for the independent and 81% for the training dataset. The interface region has an agreement of 75% and 77% correct predictions, respectively. This is expected since the interface region has two adjacent regions that detract correct predictions. In addition, the usage of a fixed membrane thickness will reduce prediction accuracy in this region [1]. The TM region has an agreement of 73%. Therefore, the prediction accuracies for MPs are similar for all of the three regions. The smaller agreement in the SOL for soluble parts of MPs than for soluble proteins has been observed earlier [5] and can be attributed to the

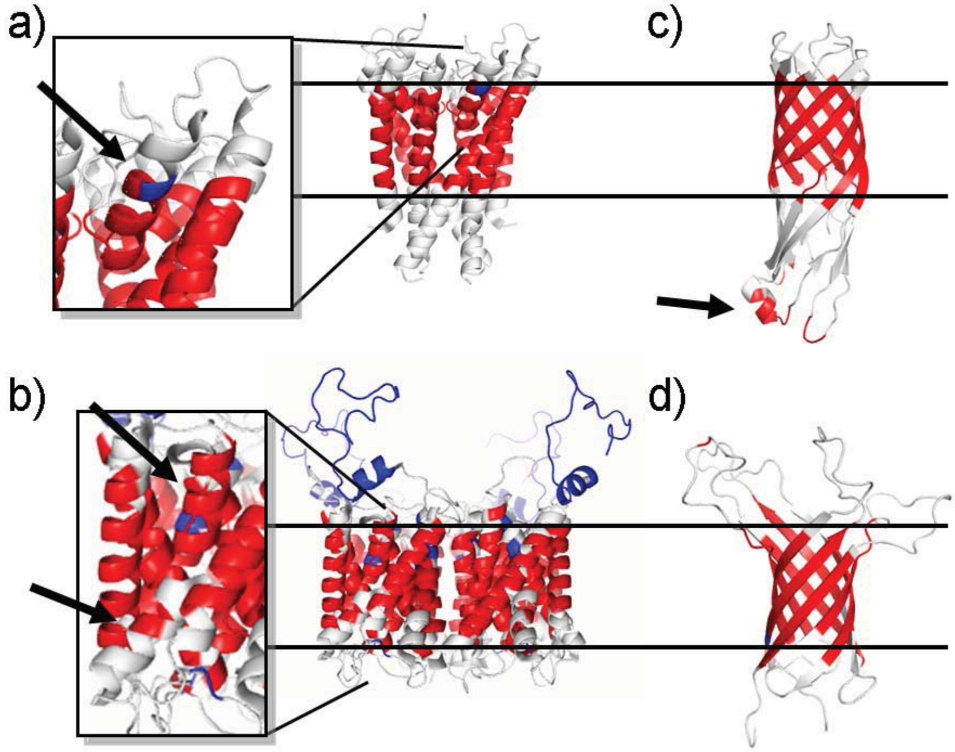

Fig. 4. The algorithm was applied to the sequence of four proteins and mapped onto the known protein structures. a) KcsA potassium channel (PDBID 1k4c); b) lens aquaporin-0 (PDBID 2b6p); c) Outer membrane protein W (PDBID 2f1t); d) Outer membrane protein A (PDBID 2ge4). Red indicates a prediction for being in TM, white represents a prediction for TR, and blue indicates a prediction for SOL. The membrane location is indicated by the black lines. The arrow in the close-up of panel a) points to the pore helix of the tetrameric channel which is a half-helix with the adjacent loop (representing the selectivity filter) returning to the extra-cellular side.

difficulty of accurately pinpointing the exact beginnings and ends of the TM spans. In other words, the residues on the membrane surface are more often predicted as TM although they belong to the SOL region. Such residues are absent in soluble proteins resulting in a better performance.

D. *Four examples illustrate a successful prediction*

The algorithm was tested on four examples: two α-helical proteins and two β-barrel proteins. Only the sequence of the proteins was used as input and the prediction was mapped onto the known protein structures as shown in Fig. 4.

Panel a) shows the crystal structure of the potassium channel KcsA (PDB ID 1k4c). The figure shows the correct prediction of the membrane location. The structure contains a half-helix (selectivity filter) with the adjacent loop returning to the extra-cellular side of the channel (see close-up). Since the correct prediction of such half-helices represents a particular challenge to the algorithm this indicates the ANNs ability to identify the correct location of these pore helices. For this example the ANN predicts 83% of the residues correctly. 95% of the TR residues and 90% of the TM residues are correctly identified. The unified hydrophobicity scale in conjunction with the simple window function implemented earlier [5] identifies 68% of the residues correctly with an accuracy of 21% for SOL, 55% for TR, and 90% for TM.

The prediction for the crystal structure of lens aquaporin-0 in the open state (PDB ID 2b6p) is shown in panel b). Again, all of the three regions are correctly identified. Overall, 75% of the residues are correctly classified. The accuracy is 93% for SOL, 81% for TR, and 68% for TM. The lower agreement in TM is due to the fact that there are isolated residues in the membrane that are predicted to be in SOL. One of the two half-helices is correctly predicted to be in the membrane (as seen by the upper arrow in the inset). The half-helices dip into the membrane and the adjacent loops return to the extra-membrane region. This represents a particular challenge for prediction algorithms since TM helices are usually much longer (~19 residues) and can be confused with hydrophobic regions in soluble proteins. This difficulty might be addressed by feeding the output of this prediction algorithm into a second ANN to obtain the final output. Such a procedure was applied in PsiPred, one of the best secondary structure prediction algorithms to date [33].

Panel c) shows the structure of the Outer Membrane Protein W (OmpW – PDB ID 2f1t). The algorithm is able to correctly identify the location of TM strands. Overall, 73% of the residues are correctly identified with an accuracy of 100% for the TR, and 86% for the TM. The soluble region is not predicted as such since 71% of these residues are predicted to be in TR and 29% in the TM. This is indicated by the small helix at the bottom (see arrow) which is predicted to be in TM although it resides in SOL. For comparison, the unified hydrophobicity scale in conjunction with the simple window function implemented earlier [5] identifies 43% of the residues correctly with an accuracy of 29% for SOL, 75% for TR, and 27% for TM.

Panel d) shows the Outer Membrane Protein A (OmpA – PDB ID 2ge4). Also this example suggests that the algorithm is able to distinguish the different regions for β-barrel proteins. In this protein the overall prediction accuracy averages to 81%. 97% of the TR residues are correctly identified and 77% of the TM residues are correctly predicted. The algorithm identifies all of the 12 soluble residues as being in TR. However, they constitute only ~7% of the total residues in this small β-barrel.

IV. Conclusion

An artificial neural network was trained to predict the location of trans-membrane spans from the protein sequence. In contrast to earlier prediction tools which are specialized for either α-helical or β-barrel proteins, the method represents the first tool that predicts trans-membrane spans for both classes of proteins.

The artificial neural network was trained on a membrane protein and soluble protein database. As input served several amino acid properties and the position-specific scoring matrices from PsiBlast. Furthermore, we used the free energies for (1) the three-state scenario of the residue being in helix, strand, and coil, (2) the three-state scenario of the residue being in trans-membrane, transition, and soluble region, and (3) the nine-state scenario with pair-wise combinations of the former. We found that the position-specific scoring matrices and the free energies for the trans-membrane region (both for individual secondary structure types as well as combined) had the highest impact on the prediction. In contrast, other amino acid properties were less important for the prediction.

Soluble residues were correctly predicted in 92% of the cases, for interface residues the accuracy was 75%, and for trans-membrane residues 73%. Therefore, in the three-state scenario, on average 79% of the residues are correctly predicted, which is a remarkable improvement compared to the prediction using simple hydrophobicity scales.

The algorithm was applied to four membrane proteins, two of α-helical nature and two β-barrel proteins. In these examples the prediction tool is able to classify 78% of the residues correctly. Even though half-helices are intrinsically difficult to predict, the predictor correctly identified two of three half-helices as trans-membrane spans. Since the tested proteins lack large soluble domains, the network has difficulties to identify short soluble loops and correctly classifies them only for one of the four examples.

Acknowledgment

The authors would like to thank Nils Woetzel and Mert Karakas for their help in establishing the databases and Ricardo Lugo and Danilo Selic for their work on the ANN.

REFERENCES

[1] Bowie JU. Solving the membrane protein folding problem. Nature 2005;438(7068):581-589.

[2] Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry 1996;35(16):5109-5124.

[3] White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct 1999;28:319-365.

[4] Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 1986;15:321-353.

[5] Koehler J, Woetzel N, Staritzbichler R, Sanders CR, Meiler J. A Unified Hydrophobicity Scale for Multi-Span Membrane Proteins. Proteins, Structure, Function, and Bioinformatics; in press.

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

[7] Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 1996;3(10):842-848.

[8] Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, von Heijne G. Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature 2007;450(7172):1026-U1022.

[9] Senes A, Chadi DC, Law PB, Walters RF, Nanda V, Degrado WF. E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. J Mol Biol 2007;366(2):436-448.

[10] Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 2005;433(7024):377-381.

[11] Cuthbertson JM, Doyle DA, Sansom MS. Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 2005;18(6):295-308.

[12] Juretic D, Zoranic L, Zucic D. Basic charge clusters and predictions of membrane protein topology. J Chem Inf Comput Sci 2002;42(3):620-632.

[13] 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.

[14] Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 1998;283(2):489-506.

[15] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17(9):849-850.

[16] Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4(3):521-533.

[17] Kahsay RY, Gao G, Liao L. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics 2005;21(9):1853-1858.

[18] http://topcons.net/.

[19] Bagos PG, Liakopoulos TD, Hamodrakas SJ. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics 2005;6:7.

[20] Martelli PL, Fariselli P, Krogh A, Casadio R. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 2002;18 Suppl 1:S46-53.

[21] Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 2004;32(8):2566-2577.

[22] Bigelow H, Rost B. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res 2006;34(Web Server issue):W186-188.

[23] http://pdbtm.enzim.hu/.

[24] Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589-1591.

[25] Wang GL, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Research 2005;33:W94-W98.

[26] Tusnady GE, Dosztanyi Z, Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 2005;21(7):1276-1277.

[27] Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11(1):18-26.

[28] Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Journal of Molecular Modeling 2001;7(9):360-369.

[29] 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.

[30] Anastasladis AD, Magoulas GD, Vrahatis MN. New globally convergent training scheme based on the resilient propagation algorithm. Neurocomputing 2005;64:253-270.

[31] Alexander N, Bortolus M, Al-Mestarihi A, McHaourab H, Meiler J. De novo high-resolution protein structure determination from sparse spin-labeling EPR data. Structure 2008;16(2):181-195.

[32] Dong E, Smith J, Heinze S, Alexander N, Meiler J. BCL::Align-sequence alignment and fold recognition with a custom scoring function online. Gene 2008;422(1-2):41-46.

[33] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2):195-202.