DOI 10.1002/prot.24258

Simultaneous prediction of protein secondary structure and trans-membrane spans

Secondary structure prediction in membranes

Julia Koehler Leman[1,2], *Ralf Mueller*[1,2], *Mert Karakas*[1,2], *Nils Woetzel*[1,2], *Jens Meiler*[1,2]*

[1]Department of Chemistry, and [2]Center for Structural Biology
Vanderbilt University, Nashville, Tennessee, United States

* *To whom correspondence should be addressed:*

Jens Meiler
Associate Professor
Vanderbilt University
Departments of Chemistry and Pharmacology
Center for Structural Biology
465 21st Ave South
BioSci/MRB III, Room 5144B
Nashville, Tennessee 37232-8725
USA

phone: (615) 936 5662
fax: (615) 936 2211
jens.meiler@vanderbilt.edu
http://www.meilerlab.org/

Keywords:
secondary structure prediction, trans-membrane span prediction, membrane proteins, protein structure prediction, ProteinDataBank

Abbreviations:

PDB	ProteinDataBank
PDBTM	ProteinDataBank for Trans-Membrane proteins
RMSD	Root Mean Square Deviation
MC	Membrane Core
TR	Transition Region
SO	Solution
SS	secondary structure
TM	trans-membrane
EPR	Electron Paramagnetic Resonance

Received: May 29, 2012; Revised: Jan 03, 2013; Accepted: Jan 09, 2013

MTSL	Methanethiosulfonate spin label
MP	Membrane Protein
ANN	Artificial Neural Network

Abstract

Prediction of trans-membrane spans and secondary structure from the protein sequence is generally the first step in the structural characterization of (membrane) proteins. Preference of a stretch of amino acids in a protein to form secondary structure and being placed in the membrane are correlated. Nevertheless, current methods predict either secondary structure or individual trans-membrane states. We introduce a method that simultaneously predicts the secondary structure and trans-membrane spans from the protein sequence. This approach not only eliminates the necessity to create a consensus prediction from possibly contradicting outputs of several predictors but bears the potential to predict conformational switches, i.e. sequence regions that have a high probability to change for example from a coil conformation in solution to an α-helical trans-membrane state. An Artificial Neural Network was trained on databases of 177 membrane proteins and 6048 soluble proteins. The output is a 3x3-dimensional probability matrix for each residue in the sequence that combines three secondary structure types (helix, strand, coil) and three environment types (membrane core, interface, solution). The prediction accuracies are 70.3% for nine possible states, 73.2% for three-state secondary structure prediction and 94.8% for three-state trans-membrane span prediction. These accuracies are comparable to state-of-the art predictors of secondary structure (e.g. Psipred) or trans-membrane placement (e.g. OCTOPUS). The method is available as web-server and for download at www.meilerlab.org.

Introduction

The prediction of secondary structure (SS) and trans-membrane (TM) segments from sequence is the first step towards structural characterization of proteins. It is typically applied before more laborious experimental methods are employed: CD spectroscopy only yields an overall SS composition of the protein – no amino acid specific values. The chemical shift index (CSI) derived from NMR experiments requires signal assignment of the protein backbone which is a time-consuming task. Moreover, identification of SS and TM spans is the first step of computational modeling of (membrane) proteins. The output of SS and TM prediction tools are therefore a basic requirement for algorithms performing sequence alignment, fold recognition, and *de novo* protein structure prediction. Furthermore, it facilitates the design of EPR experiments to find an optimal position for MTSL spin labels [1] or to select detergents to screen for membrane protein NMR experiments based on the thickness of the hydrophobic region of the membrane protein.

The identification of SS and TM spans is typically accomplished using a variety of SS and TM prediction methods in parallel (see below). However, the formation of SS and TM spans is interrelated because the occurrence of SS is greatly increased in the TM region. Peptides or proteins can exist in a disordered state in a polar solution because backbone carbonyls and amide protons form hydrogen bonds with the surrounding water molecules. When these peptides are inserted into the membrane the hydrophobic environment drives the same polar groups to form intra-molecular hydrogen bonds – SS is formed. BCL::Jufo9D leverages this interrelation by simultaneously predicting SS and TM segments, i.e. predicting SS propensity in polar and apolar environments. It thereby

enables the prediction of conformational switches, i.e. sequence regions that are stable in two different conformations, for example, as coil in solution or an α-helix in the membrane. This is an important achievement as isolated prediction of secondary structure might recognize a high helix and coil probability, and isolated prediction of trans-membrane spans might recognize the ability to exist in solution or as a TM span, however, the correlation between these probabilities is missing.

Machine learning techniques are widely used for prediction of SS and TM placement

Most recent methods for SS prediction use machine learning techniques (see [2]) such as Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), or Support Vector Machines (SVMs). These algorithms are pattern recognition techniques that associate a given input (e.g. the sequence information of a protein) to an output (e.g. the structural information such as SS or TM spans). For supervised learning the output is provided during the training process using structural information of proteins with known structure. When training is complete, the algorithms predict i.e. the SS for a target sequence. The use of machine learning approaches in SS prediction has been pioneered by Rost and co-workers through the development of their PhD program [3-4].

For soluble proteins SS prediction tools usually provide a three-state probability for each residue being either in helix, strand, or coil. Accuracy is often reported as a Q_3 value which is the percentage of correctly predicted SS if the state with the highest predicted probability is compared to the experimentally determined SS. Accuracies of up to 80% are achieved [5] with Psipred [6-7] being one of the most accurate SS prediction tools

available [5]. Psipred is a two-stage feed-forward ANN that was trained on a sequence database of soluble proteins with position-specific scoring matrices (PSSM) from PSIBLAST [8] as an input. JUFO [9-10] is an ANN that uses dimension-reduced amino acid representations to predict the SS of soluble proteins. It is trained on a database of 430 soluble peptides from the FSSP database [11] using an input window of 31 residues. JUFO was applied to the simultaneous prediction of secondary and tertiary structures probing their interrelation [9]. The SS prediction tool PROFPHD [4,12-13] as part of the PredictProtein server is also based on ANNs. It is a three-layer feed-forward ANN trained on sequence-to-structure and structure-to-structure context that uses a multiple sequence alignment and global amino acid composition as inputs. The developers state a three-state accuracy of 76%.

TM span prediction methods are specialized to either α-helical proteins or β-barrels

Early attempts to predict the location of TM spans in membrane proteins (MPs) involve averaging hydrophobicity values over a sequence window. Many different hydrophobicity scales have been developed using a variety of experimental [14-20], theoretical [21-25], and consensus approaches [26-28], some of them are reviewed in [25]. Most of the scales consider the two states membrane bilayer and solution. The scales of Wimley & White [16,29] as well as a recently developed knowledge-based unified hydrophobicity scale (UHS) [25] take a third interface region into account. Considering an interface region is important since the dielectric environment characterized by the polar lipid head-groups is distinctly different from the aqueous solution as well as from the membrane core

region. Aromatic residues like Tyr or Trp as well as amphipathic α-helices usually reside there [16,25]. Predicting the location of TM-spans using averaging schemes for hydrophobicity values achieves accuracies up to 73% in the two-state scenario (membrane bilayer and solution) and up to 60% in the three-state scenario (with interface region) [25].

Considerable improvement is achieved by the application of machine learning approaches; however, these methods are specialized to either TM α-helical bundles or β-barrels: For identification of TM spans in α-helical MPs OCTOPUS [30] is one of the best methods available. It uses four separately trained ANNs to identify one of the four states (membrane, interface, loop, globular) at the residue level and combines the predictions globally using a Hidden Markov Model (HMM). It is designed as a topology predictor and is able to model reentrant/membrane dipping regions and TM hairpins. The prediction accuracies on an independent benchmark dataset were reported to be as high as 94% for identification of the correct topology. Other available methods use HMMs (such as TMHMM [31] and TMMOD [32]), SVMs (such as MEMSAT-SVM [33]) or a consensus of multiple SS prediction servers, such as ConPredII [34-35].

For identification of TM β-barrels, TMbeta-Net [36] is one of few methods available. It consists of an ANN that was trained on 13 outer membrane proteins with a jack-knife approach for cross-validation. Other methods, mostly HMMs, include ProfTMB [37-38] as part of the PredictProtein server [39] and TMBHMM [40].

The method presented here seeks to simultaneously predict SS and TM regions leveraging their interrelation. It alleviates the necessity to combine multiple contradicting outputs into a single prediction. Moreover, it overcomes the specialization of TM span

prediction tools for either α-helical bundles or β-barrels and the specialization of SS methods to soluble proteins. BCL::Jufo9D is a set of ANNs that predicts a nine-state probability distribution combining three SS states (helix, strand, coil) and three protein environment states (membrane, interface, solution) for each residue in the protein sequence. The ANNs were trained on databases of 226 MP chains in 177 MPs and 6223 soluble protein chains in 6048 soluble proteins. The approach achieves per residue accuracies of 70.3% in a nine state prediction scenario for the independent dataset compared to an accuracy of a random prediction of 11.1%.

Methods

Establishing the membrane protein database

A list of all membrane protein chains, for which a structure has been determined, was downloaded from the PDBTM [41-42] website (Nov. 2011). Similar sequences were excluded by culling this list with the PISCES server [43-44]. The parameters included a percent sequence identity ≤ 30%, resolution 0 – 3 Å, R-factor 0.25, sequence length 40 – 10,000 residues, non X-ray entries as well as CA-only chains were included. BCL::PDBConvert (Woetzel, N. submitted) was used to convert non-natural amino acids into their natural counterparts and to transform the protein into the membrane coordinate frame using the membrane definitions provided by the PDBTM website. The membrane normal aligns with the z-coordinate in the PDB file with the membrane center being at z = 0. We assume a constant thickness of 20 Å for the membrane core and 10 Å for the transition region on either side of the membrane (Figure 1A). Residues in the 2.5 Å gap regions between membrane core and transition region or transition region and solution

were disregarded to obtain more distinct regions for the ANN to identify. DSSP [45] (version of 2011) was used for all PDB structures to obtain a consistent SS identification. Helices with less than five residues and strands with less than three residues were disregarded to focus the prediction on long SS elements. This procedure resulted in a list of 226 chains in 177 membrane proteins.

Establishing the database of soluble proteins

A pre-compiled list of PDB chains was downloaded from the PISCES protein sequence culling server (date 12/02/2011) [43-44]. The list contained sequences with a percentage sequence identity ≤ 30%, resolution 0 – 2 Å, R-factor 0.25, sequence length 40 – 10,000 residues, non X-ray entries as well as CA-only chains were excluded. Membrane proteins were excluded from this list. BCL::PDBConvert (Woetzel, N. submitted) was used to convert non-natural amino acids into their natural counterparts and DSSP [45] was used to identify SS elements. Helices shorter than five residues and strands shorter than three residues were disregarded. The result was a list of 6,223 chains in 6,048 soluble proteins.

The residue counts for all regions both for MPs as well as soluble proteins are shown in Supplementary Figure 2. The counts for soluble proteins are much higher and there are almost twice as many soluble helix residues present in the database (~113,000) than soluble strand residues (~66,000). The counts range from 1852 for coil residues in the membrane core to ~137,000 for coil residues in solution.

Experimental design allows for cross-validation

The databases were split into five subsets for cross-validation. For the membrane proteins α-helical bundles as well as β-barrels were distributed as equally as possible. The soluble proteins were distributed randomly.

To train a single ANN, three of the five subsets were used for training (see Figure 1B) and one subset was used for monitoring the training process to avoid overtraining. The fifth subset was used as an independent test set for computing the prediction accuracies. 20 networks were trained such that the independent as well as the monitoring permuted through the five datasets (Figure 1B).

Evolutionary and property profiles are used as ANN input

Figure 2 shows the input parameters used (see Supplementary Table S1): (a) five amino acid properties including steric parameter, volume, polarizability, iso-electric point, solvent-accessible surface area [10]; (b) six free energies for SS type (helix, strand, coil), residue environment (membrane bilayer, interface, solution) [25] and the nine combinations of both; (c) the position-specific scoring matrices (PSSM) from PSIBLAST [8] after six iterations (see [46]). For each residue all of these parameters were collected over a sequence window of 31 residues. The optimal size of the input window was determined by testing all odd window sizes between 15 and 39 residues.

In addition, "global" parameters were considered for each protein: (a) the number of residues in the protein chain; (b) the oligomeric state (monomer vs. oligomer); (c) the average of all amino acid specific parameters over the entire protein chain including their properties, free energies, and the PSSM values. This resulted in (31 residues x (20

numbers from PSSM + 20 amino acid properties)) + (2 parameters: oligomeric state, length) + (40 averages) = 1282 input parameters that represent the residue at the center of the window.

Balanced training avoids prediction bias towards over-represented states

The datasets (the term "dataset" corresponds to the input and output parameters for each residue in a protein sequence) were randomized and balanced for each protein subset independently. For balancing, an over-sampling procedure was used to represent each of the nine states equally often and avoid a bias in the predictions towards the more abundant states. This approach also increases the entropy in the input data and maximizes the information gain the ANN can achieve.

The ANNs were three-layer feed-forward networks with a sigmoidal activation function and trained through back-propagation of errors. The hidden layer contained 32 neurons – a number that was optimized by testing 4, 8, 16, 32, 64, and 128 neurons. The three subsets used for training contained a total of 270,000 instances, 90,000 instances were in the monitoring dataset, and 90,000 instances in the independent dataset. The training protocol consisted of three consecutive steps using a simple propagation algorithm: (1) 50 steps with weight update after each step with momentum $\alpha = 0.0$ and the learning rate $\eta = 10^{-3}$; (2) 10 steps with batch update with momentum $\alpha = 0.5$ and the learning rate $\eta = 5 \bullet 10^{-6}$; (3) 100 steps with weight update after each step with momentum $\alpha = 1.0$ and the learning rate $\eta = 5 \bullet 10^{-6}$. As a post-processing step the outputs of the four ANNs were averaged that used the same independent subset.

Prediction accuracies are calculated on a per-residue basis on independent datasets as an average over four ANNs used for cross-validation

To report the prediction accuracies as well as the confidence measure, an average of four network outputs was computed only considering ANNs belonging to a single set that share the same independent dataset (Figure 1B). This setup (a) ensures that the reported accuracies originate from ANNs that were not trained on the test set, and (b) prediction accuracies can be reported for each protein in the dataset as always four ANNs exist that were trained with this particular protein in the independent dataset. The final output from the web-server is the average over all 20 ANNs and computes a confidence measure that constitutes the difference between the highest and second highest output probability for each residue.

To calculate the per-residue prediction accuracies, the outputs of the four ANNs in a single set were averaged. The outputs per set were compared to the actual state on a per-residue basis: if the predicted state was a TM helix and the actual state was a TR helix, the counts in this particular 9x9 matrix element (see Figure 3) was increased by one. After obtaining all counts for the 9x9 matrix over a single set, the counts were divided by the number of residues in this region (sum over each row) to arrive at the percentage of predicted residues in each matrix element. The percentages of predicted residues were then averaged over the five sets of ANNs. This cross-validation and averaging procedure circumvents that a "bad choice" of proteins in an independent dataset biases the prediction accuracies.

The counts for the three-state SS prediction, three-state TM span prediction, or two-state TM helix/TM strand prediction were calculated as described in Supplementary

Figure S1. The counts were divided by the total number of counts per row to arrive at the percentages of predicted residues and these percentages were later averaged over the five sets of ANNs.

Consolidating per-residue predictions into two-state prediction of complete TM spans increases accuracy

To directly compare the nine-state output of BCL::Jufo9D to the two-state output of, for instance OCTOPUS, we summed the non-TM helix probabilities to arrive at a two-state prediction (Supplementary Figure S1). To remove the resulting bias towards non-TM states from adding background probabilities of 11.1%, the result needs to be corrected by adding or subtracting ½•(8•11.1%-11.1%)=38.9% from the two states, respectively. This procedure ensures that the total of all prediction probabilities remains 100%. The identical correction was applied to the TM strand prediction.

Furthermore, a post-processing step has been applied for noise reduction. Lengths of SS elements were calculated where kinks of one or two residues were regarded as TM helix residues but were retained in the final prediction. Since the topology prediction output considers only long SS elements that can span the membrane, helices shorter than 11 residues (including kinks) and strands shorter than 5 residues were removed. Including this post-processing step resulted in an increase in prediction accuracy of ~3% over all residues in the dataset.

Results and Discussion

BCL::Jufo9D achieves nine-state per-residue accuracies of 70.3%

Figure 3 shows the percentage of predicted residues for all nine states whereas the percentages are averages over all independent datasets. The rows correspond to the "true" state as represented in the structure and the columns correspond to the predicted state. Ideally, highest percentages should be seen in the matrix diagonal.

Overall, in the nine-state scenario BCL::Jufo9D predicts the correct state for 70.3% of the residues in the independent dataset.

As seen from Figure 3, "true" soluble states are distinguished most accurately from the membrane core or the interface region because their characteristics are distinctly different. Furthermore, helices and strands in solution and the membrane core have highest prediction accuracies (up to 74%), whereas the states in the transition region have lower accuracies ranging up to 60%. We attribute the reduced prediction accuracy in the transition region to two causes: First, the transition region borders to soluble and membrane core regions allowing for two types of errors – prediction as membrane core residues or prediction in the soluble region. In contrast, membrane core and soluble region border only to the transition region eliminating one source of error for these regions. Secondly, membrane proteins cover a range of thicknesses of the membrane core region. Choosing a constant membrane thickness of 20 Å for training BCL::Jufo9D introduces some error in classifying amino acids as membrane core, transition region, or solution. This effect is partly offset through the introduction of the 2.5 Å gaps between the regions. Excluding the gap regions for the predictions results in on average 0.3% improved prediction accuracies where largest improvements up to 4% are seen for the

membrane core and the transition region (Supplementary Figures S3 and S4). Nevertheless, the transition region continues to contain most misclassified residues.

Variable membrane thickness does not improve prediction accuracy

Membrane thicknesses can be computed from experimental MP structures using specific algorithms. We tested the TMDET algorithm provided by the PDBTM [42,47] to compute membrane thickness (data not shown). The performance was overall comparable to usage of a constant membrane thickness. A constant membrane thickness was chosen to circumvent a potentially circular influence of the TMDET algorithm onto BCL::Jufo9D.

Common mistakes include swapping of coil regions with helix or strand and membrane core with transition regions

Whereas helices and strands in solution and the membrane core have highest prediction accuracies, the prediction accuracies of coil states are lower, irrespective of their environment (Figure 3). This is expected, since the coil regions are more diverse in sequence lacking some of the characteristic properties than enable the identification of patterns. Coil states in TM spans are under-represented complicating their reliable identification. Additionally, helix and strand states were rarely mixed irrespective of their environment. This is expected because the properties characteristic for helices with a periodicity of 3.6 are distinctly different than for strands with a periodicity of 2. The trends for swapping predictions between membrane core and transition region (but not solution) and helix/coil and strand/coil is observed most readily when considering the three-state SS and TM prediction as seen in Figure 3.

Three-state secondary structure predictions or TM span predictions achieve accuracies of 73.2% and 94.8%

The ANN output can be analyzed by summing the three probabilities for each of the SS states helix, strand, and coil. The resulting three-state SS prediction accuracies are shown in Figure 4A. On average, in the three-state scenario the SS is correctly identified for 73.2% of the residues. Similar accuracies are obtained for helix and strand states for each of the different environments (Figure 3), however the accuracies in the transition region are lower than for membrane core or solution for reasons discussed above.

Figure 4A also shows the prediction accuracies for the other SS prediction methods Psipred, ProfPhD, and JUFO. Even though accuracies of BCL::Jufo9D are marginally lower than for Psipred or JUFO, accuracies of competing methods are likely somewhat inflated as the testing dataset is not independent from their training set. As discussed below, Psipred has very high accuracies at the termini of SS elements which is presumably one reason for the differences in overall prediction accuracies. However, we believe that BCL::Jufo9D's ability to predict the protein environment in addition to the SS more than compensates for these minimal differences in prediction accuracy.

The ANN output can also be analyzed by summing the three probabilities for each of the TM states membrane core, transition region, solution to arrive at a three state TM span prediction. The accuracies are given in Figure 3. Overall, the environment of 94.8% of the residues in the independent datasets is correctly identified, a number that reflects the bias towards soluble proteins in the datasets. For training, the oversampling procedure guarantees that this bias does not impact the weights in the ANNs.

Nine output states contain more information than both SS prediction and TM span prediction combined

As stated above, BCL::Jufo9D overall classifies 70.3% of the residues correctly into the nine possible states. This compares to an expected accuracy of 11.1% for a random predictor. We wanted to explore how the 70.3% in nine possible states compare to the typical ~73% prediction accuracy of a three-state SS prediction and whether it contains more information. As a direct measure, we computed the information gain for the three-state SS prediction, which is 0.173 ± 0.003, and for the three-state TM prediction, which is 0.294 ± 0.004. Therefore, the sum of the information gain for both SS and TM prediction is with 0.467 ± 0.005 lower than the information gain of the nine-state prediction which is 0.527 ± 0.004. This supports the hypothesis that the nine-state prediction generally contains more information than both of the three-state predictions combined, possibly because the influence of residue environment onto the formation of hydrogen bonds and therefore SS.

Two-state TM span identification yields accuracies of up to 98.0%

Available TM span prediction tools predict their output in two states: OCTOPUS, for instance, identifies whether a residue is located in a TM helix or not. To directly compare BCL::Jufo9D to OCTOPUS we summed the non-TM helix probabilities to arrive at a two-state prediction and applied a post-processing step as described in the Methods section. Using the described consolidation of per-residue predictions, BCL::Jufo9D correctly predicts 97.9% of the residues in the independent datasets (Figure 4B).

OCTOPUS correctly predicts the states of 97.3% of the residues in our dataset. Note that the accuracies of alternative methods tend to be somewhat inflated as these might have been trained on membrane proteins from our independent dataset.

For the TM strand prediction BCL::Jufo9D correctly predicts 94.6% of the residues, whereas TMBetaNet correctly identifies 50.9% of the residues. This number is rather low due to the high over-prediction rate that this method achieves (see Figure 4B).

We want to point out that our method is not set up to directly distinguish TM β-barrels from other proteins. However, we do believe that the high accuracy in TM strand prediction is useful to identify proteins that could be TM β-barrels solely from sequence information. A difference of BCL::Jufo9D to other TM strand prediction methods is that most other methods are trained solely on TM β-barrels and do not include barrels that are formed by multiple chains in the protein. For example, TMBetaNet extensively over-predicts TM β-barrels. BCL::Jufo9D, on the other hand, is trained on these proteins and higher prediction accuracies may be expected.

The ultimate goal would be the establishment of a topology prediction method that distinguishes different protein orientations in the membrane. Currently, the challenges with establishing such a method are the small number of β-barrel MPs resulting in small residue counts in the membrane and interface regions. Training of BCL::Jufo9D splits these few counts into nine output states multiplied by five datasets for cross-validation; further separation would ultimately result in lower prediction accuracies and more noise in the predictions.

Over- and under prediction

In addition to the two-state predictions Figure 4B also shows the over- and under-predictions of TM spans for complete datasets. For TM helix prediction methods, the percentage of over-predicted residues is 2.0% for BCL::Jufo9D and 2.6% for OCTOPUS, for under-prediction 9.1% for BCL::Jufo9D, and 9.8% for OCTOPUS. Similar trends are seen for TM strand prediction methods, where the percentage of over-predicted residues is 5.4% for BCL::Jufo9D and 49.2% for TMBetaNet, for under-prediction 9.3% for BCL::Jufo9D and 24.6% for TMBetaNET. The tendencies to over-/or under-predict certain states are a result of the training procedure and post-processing steps and represent advantages/disadvantages of certain methods for certain applications.

The BCL::Jufo9D server available at www.meilerlab.org provides the two-state outputs in addition to nine-state and three-state outputs. If, for instance, it is known that a particular protein is an α-helical MP, the two-state output more accurately defines the membrane boundaries compared to a nine-state output which, on the other hand, is more useful to describe the overall architecture of the protein without *a priori* knowledge.

Examples demonstrate high prediction accuracies

Figure 5 shows some example cases where the protein sequence was used to predict the SS and TM regions with BCL::Jufo9D. These predictions were mapped onto the known protein structures. The examples are the outer membrane protein OmpX (PDB: 1qj8), the TolC receptor (PDB: 1yc9), the photosynthetic reaction center of cyanobacteria (PDB: 1jb0), and the *E.coli* quinol fumarate reductase (PDB: 1kf6). The prediction accuracies are reported in the figure. For these examples, the SS prediction

accuracy ranges from about 70-90% correctly predicted residues and the TM span prediction accuracy ranges from 68-90%.

Challenges and limitations

Panel B in Figure 5 shows challenges of our method where some of the residues are incorrectly identified. The first example is the human mitochondrial ABC transporter (PDB: 4ayt) with 76.8% of the residues correctly identified in terms of SS, and 54.8% of the residues for TM span prediction. Whereas the TM region in 4ayt is accurately identified, a number of residues in solution are predicted to be in the transition region or even in the membrane. Interestingly, the SS prediction does not suffer from the inaccurate identification of TM regions.

For the main porin of *mycobacteria smegmatis* (PDB: 1uun) the SS is correctly predicted for 76.9% and the TM spans are correctly identified for 44.6% of the residues. In this example, stretches of residues in the membrane are predicted to be soluble. In addition, a large number of residues in solution are identified as transition region or membrane states. Again, the SS prediction does not suffer from this incorrect identification. We point out that overall the percentage of incorrectly classified amino acids remains low and the examples presented here are extreme outliers of a generally very accurate method.

Inevitably, prediction of SS and TM spans from the sequence only is affiliated with some error margin as formation of secondary and tertiary structure is coupled [9]. Specific mistakes made by BCL::Jufo9D, especially for β-barrel MPs do not uniformly correlate with β-barrel diameter, number of charged residues in the TM region, or

orientation of an amino acid side chain towards a polar interior cavity within a membrane protein (data not shown). We tested an ANN architecture with additional output states for residues that point towards polar interior cavities in the membrane region and observed no improved prediction accuracy. We further tested if the limited space of MP sequences could be supplemented with sequence information from homologous MPs, however, no improvement in prediction accuracy was observed.

Secondary structure prediction accuracies are higher for soluble proteins than for soluble parts of membrane proteins

Supplementary Figure 6 shows the prediction accuracies for soluble parts of MPs. It can be seen that in panel A) that the prediction accuracies for soluble states (columns - predicted) in solution (rows - actual) seem with up to 40% very low. The major cause for this is the difficulty of the method to distinguish residues in solution from the interface region. Conversely, SS prediction for soluble proteins is very accurate (panel B), i.e. BCL::Jufo9D recognizes soluble proteins well. The lower prediction accuracies for soluble parts of MPs are primarily explained by a few MP examples with large soluble domains which bias these prediction accuracies, as shown in the examples in Figure 5B. We found that training solely on MPs would alleviate these errors and increase the prediction accuracies for the soluble domains of MPs. Currently, this procedure also decreases the accuracies for some membrane and transition states (up to 5%) as well as for the SS prediction of soluble proteins (up to 9%) at the same time. This result repeatedly demonstrates the interrelation between SS and protein environment, and

suggests that the establishment of a "perfect" prediction method remains challenging since optimizing one aspect is only achieved at the expense of another.

At the current stage one possible reason for the reduced prediction accuracy might be the conditions under which MP structures were determined. Artificial membrane-mimicking environments used in crystallography and NMR spectroscopy perturb the structure from its native state to an unknown degree. Domains outside the membrane might be pushed into a non-native location by the artificial conditions imposed by a three-dimensional crystal lattice or by detergent environments, such as micelles, typically used for NMR spectroscopy. This leads to misclassification of residues in particular in the non-membrane regions of membrane proteins when training the method, ultimately reducing its prediction accuracy.

One example is the cholera cytolysin heptamer (PDB: 3o44) whose cytolysin domain contains many aromatic residues which are expected to reside in the transition region. In the crystal structure, which was determined in detergent, the cytolysin domain is most likely placed on the micelle surface [48]. However, if considered in a membrane bilayer, this domain incorrectly protrudes deep into the membrane. Another example is the recently determined crystal structure of the β2 adrenergic receptor-Gs protein complex (PDB: 3sn6) where a helical domain that resides in the soluble region is incorrectly placed in the transition region [49, 50]. It is currently difficult to account for such structural perturbations and since a flat membrane bilayer is defined for training BCL::Jufo9D few examples of incorrect predictions in these regions may be the result.

To obtain a three-state SS prediction accuracy, the accuracies from all predicted regions need to be added together (Supplementary Figure 6C). This is a feature of our

nine-state prediction and does not apply to other SS prediction methods. Therefore, panel C as the sum of three regions should be compared to panel D which displays the prediction accuracies of the soluble parts of MPs for the old JUFO (version from 2003), Psipred, and ProfPhD.

For BCL::Jufo9D the accuracy for helix and strand states (panel C) are comparable to the old JUFO which is more accurate for helices and coil states in solution than in the membrane. Psipred has much higher accuracies in the membrane (up to 16% higher - data not shown). ProfPhD has similar accuracies in both the membrane as well as solution, except the accuracy for coil regions which is higher for soluble states (data not shown).

Prediction percentages level off after five residues from the termini

The percentages of per-residue predictions at the beginnings and ends of SS elements and TM spans have been investigated (Figure 6) and compared to Psipred and OCTOPUS. As described in the figure legend, the N- and C-termini were not distinguished. Generally, the prediction percentages for both SS and TM span prediction increase at the termini and level off after the fifth residue where further increase is only marginal. In the figure, an ideal prediction is denoted by the dotted line with the black dot as the inversion point. Psipred has about 70% prediction accuracy for the first residue and remains at higher accuracies for the SS prediction than BCL::Jufo9D.

For the TM span prediction the percentages for BCL::Jufo9D were determined using the two-state 'topology prediction' discussed above. This is necessary to not distort the prediction percentages by comparing the three SS states from BCL::Jufo9D to the

two-state output of OCTOPUS. Furthermore, a distinction between shorter and longer TM helices is needed since OCTOPUS does not predict short TM helices, i.e. is unable to predict half-helices or re-entrant helices; it only predicts TM helix lengths of 15, 21, and 31 residues. When averaging over TM helices of length 8 – 19 residues OCTOPUS has accuracies up to 40% lower than BCL::Jufo9D. When considering helices of length 14 – 19 residues only, OCTOPUS' accuracies are between 3 – 8% lower than for BCL::Jufo9D.

Both BCL::Jufo9D and OCTOPUS do not match the 50% inversion point very well. This is due to the fact that for the true states only residues in the membrane core were considered whereas residues in the transition region were counted towards "solution". This means that the membrane core by itself is too thin to represent the full membrane and the two-state prediction methods predict the membrane longer than just the core. This should not be considered as an error since the transition region is not predicted by two-state TM span prediction methods. In contrast, if the transition region was considered as belonging to the membrane, the prediction percentages for both BCL::Jufo9D as well as OCTOPUS would be substantially lower (data not shown) since the termini of the predicted spans are located closer to the center of the transition region.

BCL::Jufo9D correctly predicts more than one third of kinks in TM helices

We defined a kink as one or two coil residues in TM helices longer than 11 residues. We considered the kink as accurately identified if it was predicted within five residues in either direction (N- or C-terminal of the actual kink). Out of 115 kinks in the database, 41 (36%) were correctly predicted by BCL::Jufo9D, whereas Psipred correctly predicted 19 (17%). Though Psipred is extremely good at predicting exact lengths of TM

spans as discussed above, the prediction of kinks is lacking behind BCL::Jufo9D. It is possible that the second layer ANN from Psipred reduces noise and smoothes out these features. We omit this step but use a simple averaging procedure as a post-processing step which does not remove features such as kinks. OCTOPUS correctly predicted 4 kinks corresponding to a prediction of 3% of the total number of kinks. However, this result is expected since OCTOPUS is designed to predict long TM spanning helices, not kinks.

TMkink, a method recently developed in the Bowie lab specifically designed to predict kinks in TM helices, predicted 59 out of 115 kinks which corresponds to an accuracy of 51%. A direct comparison of the results of BCL::Jufo9D with TMkink remains difficult for several reasons: Both methods were trained on MPs so the datasets are overlapping. Many of the proteins that were used for training TMkink were in our database, others had a high sequence similarity. To report accuracies for BCL::Jufo9D the proteins were in the independent dataset. For TMkink these proteins were in the training set which inflates its prediction accuracy. Furthermore, the definition of a kink is completely different in our work compared to Bowie's work. Whereas we use a very simplified method of considering one or two coil residues in TM helices which is not necessarily an indication of an actual kink, Bowie et al. defines a kink by the bend angle and uses a much more sophisticated definition. Our definition also results in rather low prediction accuracies for all tested methods, even for TMkink. In contrast, the bend angle definition results in the identification of several "kinks" in a single helix which is frequently observed in the TMkink output. This is likely to be the observation of a bent helix rather than a single kink.

Reentrant helices are correctly predicted for three examples

Reentrant helices were identified by considering helices that dipped into the membrane to a distance of 5-7 Å away from the membrane center but not entering the opposite leaflet of the membrane. This definition systematically excluded amphipathic helices that reside in the transition region. Using this definition three examples of reentrant helices were found in our database. The examples are MHP1, a nucleobase-cation transport protein (PDB: 2jln), the photosynthetic reaction center of cyanobacteria (1jb0), and the potassium channel MthK (3ldc). The SS and TM span predictions are presented in Figure 7. For MHP1 the two half-helices are clearly identified as sitting in the membrane. For the photosynthetic reaction center the reentrant helix is partially identified as in the membrane core, the other half is predicted to be in the transition region. For MthK, the part dipping deepest into the membrane is predicted to be in the membrane core, the other half "sticking out" is predicted to be in the transition region. Based on these examples we consider the prediction of reentrant helices successful, however, better statistics are needed to fully support this conclusion.

BCL::Jufo9D has the potential to predict protein conformational switches

Since BCL::Jufo9D is designed to produce a high probability in one of the nine output states representing the most likely combination of SS and TM state it maps the correlation between both states and can therefore potentially identify protein conformational switches. We expect that regions of the protein chain that can adopt two different states in terms of SS or TM placement will have a predicted high probability for these two states out of nine states. We investigated on four examples whether

BCL::Jufo9D identifies such switches: (a) the 40 residue form of the amyloid β peptide (PDB: 1iyt) which forms either a TM helix or can exist as a two-stranded β-sheet fibril in solution (PDB: 2beg); (b) the pore-forming toxin perfringolysin where a soluble helix unwinds and inserts into the membrane as a β-sheet conformation [51-52]; (c) the pore-forming toxin α-hemolysin where a soluble β-sheet detaches from the rest of the protein to insert into the membrane – the β-sheet remains intact during that process [53-54]; (d) elongation factor thermo unstable (EF-TU) which contains two switch regions switching from a soluble helix to soluble strand (switch I) or to soluble coil (switch II) [55-56].

Table I details the specific switch regions with states and summarizes the prediction of four different prediction methods for these examples. The BCL::Jufo9D outputs are shown in Supplementary Figure S8. For the Aβ peptide (1iyt) BCL::Jufo9D unambiguously identifies the correct switch regions and states. OCTOPUS identifies a single TM helix which is correct for one state, whereas Psipred predicts two soluble strands, which represents the correct identification of the second conformational state. TMBetaNet incorrectly predicts two TM strands.

For perfringolysin (1pfo) BCL::Jufo9D identifies the switch regions and states, although the probabilities are somewhat reduced. OCTOPUS identifies a signal peptide and 'outside' topology whereas Psipred predicts a single, unambiguous state for the first switch region, and two helices for the second switch region representing one conformation. TMBetaNet predicts 20 TM spans over the whole protein the correct switch regions are included, representing the second conformation.

BCL::Jufo9D also identifies the switch region and states for α-hemolysin (7ahl) where OCTOPUS predicts a globular protein. This is expected since OCTOPUS is not able to

identify TM strands. Psipred identifies three strands, even though only two are truly observed. TMBetaNet predicts 12 TM spans, again distributed over the whole protein but also including the correct strand locations.

Even though these results seem encouraging, BCL::Jufo9D does not always correctly identify conformational switches. The elongation factor thermo unstable EF-TU (1eft) contains two switch regions which are both incorrectly identified by BCL::Jufo9D. OCTOPUS correctly predicts this protein to be globular and Psipred does not recognize the helix-strand conversion either. TMBetaNet incorrectly identifies 22 TM spans over the entire protein which is a soluble protein.

In summary, specific examples show that BCL::Jufo9D is potentially able to predict protein conformational switches. However, there are examples where BCL::Jufo9D does not identify the switch region and/or switch states. This is expected since BCL::Jufo9D was not optimized to predict switch regions. We expect an increase in prediction accuracy for conformational switches once a sufficient number of conformational switches is represented in the PDB with both states so that the method can be optimized for recognizing conformational switches.

Conclusions

BCL::Jufo9D integrates the prediction of SS with the identification of TM spans. An Artificial Neural Network was trained on a database containing soluble proteins and membrane proteins. The output is a combination of the three SS states (helix, strand, coil) with the three environment states (membrane core, transition region, solution) into a nine-state probability vector for each residue in the sequence. It was shown that the per-

residue accuracy in nine states is 70.3%. When combined into a three-state prediction, BCL::Jufo9D achieves accuracies for SS prediction of 73.2% and TM span prediction of 94.8%. These results are comparable to or higher than current SS and TM span prediction tools and BCL::Jufo9D integrates both at the same time. We demonstrated that our method has higher accuracies than other SS prediction methods to predict kinks in helical TM spans and that it has the capability to predict re-entrant TM helices. We have shown that a potential advancement of our method would be the prediction of conformational switches where preliminary results on a few examples seem encouraging.

Acknowledgements

Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192) and NSF (Career 0742762).

References

1. Kazmier K, Alexander NS, Meiler J, McHaourab HS. Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination. J Struct Biol 2011;173(3):549-557.
2. Punta M, Forrest LR, Bigelow H, Kernytsky A, Liu J, Rost B. Membrane protein prediction methods. Methods 2007;41(4):460-474.
3. Rost B, Sander C, Schneider R. Phd - an Automatic Mail Server for Protein Secondary Structure Prediction. Computer Applications in the Biosciences 1994;10(1):53-60.
4. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266:525-539.
5. Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 2005;21(2):152-159.
6. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2):195-202.
7. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404-405.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-410.

9. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci U S A 2003;100(21):12105-12110.
10. 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.
11. Holm L, Sander C. Mapping the protein universe. Science 1996;273(5275):595-603.
12. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A 1993;90(16):7558-7562.
13. Rost B, Sander C, Schneider R. PHD--an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10(1):53-60.
14. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 1981;78(6):3824-3828.
15. 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.
16. 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.
17. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Annual review of biophysics and biomolecular structure 1999;28:319-365.
18. 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.
19. 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):1026-1030.
20. Moon CP, Fleming KG. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. Proc Natl Acad Sci U S A 2011;108(25):10174-10177.
21. Janin J. Surface and inside volumes in globular proteins. Nature 1979;277(5696):491-492.
22. Punta M, Maritan A. A knowledge-based scale for amino acid membrane propensity. Proteins 2003;50(1):114-121.
23. Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics 2004;20(12):1822-1835.
24. 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. Journal of molecular biology 2007;366(2):436-448.
25. Koehler J, Woetzel N, Staritzbichler R, Sanders CR, Meiler J. A unified hydrophobicity scale for multispan membrane proteins. Proteins 2008.
26. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157(1):105-132.
27. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. Hydrophobic Moments and Protein-Structure. Faraday Symposia of the Chemical Society 1982(17):109-120.

28. Guy HR. Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophys J 1985;47(1):61-70.
29. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 1996;3(10):842-848.
30. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 2008;24(15):1662-1668.
31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology 2001;305(3):567-580.
32. 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.
33. Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 2009;10.
34. Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, Shimizu T. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. Nucleic acids research 2004;32(Web Server issue):W390-393.
35. Xia JX, Ikeda M, Shimizu T. ConPred_elite: a highly reliable approach to transmembrane topology predication. Comput Biol Chem 2004;28(1):51-60.
36. Gromiha MM, Ahmad S, Suwa M. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. J Comput Chem 2004;25(5):762-767.
37. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels in proteomes. Nucleic acids research 2004;32(8):2566-2577.
38. Bigelow H, Rost B. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic acids research 2006;34(Web Server issue):W186-188.
39. Rost B, Liu J. The PredictProtein server. Nucleic Acids Res 2003;31(13):3300-3304.
40. Singh NK, Goodman A, Walter P, Helms V, Hayat S. TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. Biochim Biophys Acta 2011;1814(5):664-670.
41. Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 2004;20(17):2964-2972.
42. Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33(Database issue):D275-278.
43. Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589-1591.
44. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic acids research 2005;33(Web Server issue):W94-98.
45. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.
46. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic acids research 2003;31(13):3804-3807.
47. http://pdbtm.enzim.hu/.

48. De S, Olson R. Crystal structure of the Vibrio cholerae cytolysin heptamer reveals common features among disparate pore-forming toxins. Proc Natl Acad Sci U S A 2011;108(18):7385-7390.
49. Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK. Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 2011;477(7366):549-555.
50. Van Eps N, Preininger AM, Alexander N, Kaya AI, Meier S, Meiler J, Hamm HE, Hubbell WL. Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. Proc Natl Acad Sci U S A 2011;108(23):9420-9424.
51. Shatursky O, Heuck AP, Shepard LA, Rossjohn J, Parker MW, Johnson AE, Tweten RK. The mechanism of membrane insertion for a cholesterol-dependent cytolysin: a novel paradigm for pore-forming toxins. Cell 1999;99(3):293-299.
52. Tilley SJ, Orlova EV, Gilbert RJ, Andrew PW, Saibil HR. Structural basis of pore formation by the bacterial toxin pneumolysin. Cell 2005;121(2):247-256.
53. Song L, Hobaugh MR, Shustak C, Cheley S, Bayley H, Gouaux JE. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science 1996;274(5294):1859-1866.
54. Olson R, Nariya H, Yokota K, Kamio Y, Gouaux E. Crystal structure of staphylococcal LukF delineates conformational changes accompanying formation of a transmembrane channel. Nat Struct Biol 1999;6(2):134-140.
55. Abel K, Yoder MD, Hilgenfeld R, Jurnak F. An alpha to beta conformational switch in EF-Tu. Structure 1996;4(10):1153-1159.
56. Polekhina G, Thirup S, Kjeldgaard M, Nissen P, Lippmann C, Nyborg J. Helix unwinding in the effector region of elongation factor EF-Tu-GDP. Structure 1996;4(10):1141-1151.

Figure 1

(A) Definition of membrane thicknesses in our MP database. For training, residues in the gap region of 2.5 A between membrane core/transition region and transition region/solution were disregarded. To report prediction accuracies the gap region was removed, i.e. all residues were taken into account, and the thicknesses of the regions were adjusted as shown. (B) Each ANN is trained on three subsets for training, one for monitoring the training process, and one as an independent test set. To avoid a bias in neither the independent test set, nor the monitoring set, both of these sets are permuted through all five subsets. This results in 20 ANNs that were trained. To report prediction accuracies, the outputs of the four ANNs in one set were summed, the prediction accuracies calculated, and then averages over all five sets were computed. This procedure boosts prediction accuracies in the three state outputs by 1-2% compared to the individual networks. It was tested whether post-processing the output with a second ANN further reduces the noise, but no significant improvements were obtained (data not shown).

Figure 2

Setup for a single ANN. The residue at the center of the window is described by 1282 inputs. For this residue, a normalized nine-state prediction vector is the output. In this example, the predicted state for this residue is a helix in the membrane core (MC).

Figure 3

Averages of percent predicted residues over all independent datasets. The rows represent the "true" state, the columns represent the predicted state. Desired are large percentages in the matrix diagonals and low percentages in the off-diagonal elements. The overall

nine-state accuracy is 70.3%, for SS prediction 73.2%, and for TM span identification 94.8%. The nine-state accuracies are summed to yield three-state SS predictions and three-state TM span predictions shown at the bottom.

Figure 4

(A) Three-state secondary structure prediction comparing to methods trained on soluble proteins. (B) Performance of other two-state TM span prediction methods compared to BCL::Jufo9D that outperforms both of them. Since all other methods are limited to predicting either TM helices or TM strands, a separate comparison is required.

Figure 5

The sequences of these examples are used to predict the SS and TM state for each residue. These predictions are mapped onto the known structure. On the right panels the membrane core and transition regions on either side of the membrane are indicated by gray planes. H = prediction for helix, E = strand, C = coil, MC = membrane core, TR = transition region, SO = solution.

Figure 6

Percent of predicted residues vs. residue position of actual SS elements or TM spans. The residue position denotes the position from either side (N-terminal or C-terminal) of the SS element/TM span where position -1 is outside the SS element or TM span and position one is the first residue within. The dotted line denotes a perfect prediction with the black dot at the inversion point. The percent predictions for each position are averages over SS elements/TM spans between (2*residue position – 1) residues up to 19

residues corresponding to position 10. As an example, the TM state accuracy at position 4 is the average percentage at that position over TM spans of length 7 to 19. For the TM span percentages the 8-19 denotes the length of TM spans considered: 8 to 19 residues. Similarly, 14-19 only considers TM spans between 14 and 19 residues. This distinction was necessary since OCTOPUS only predicts TM helices with the length of 15, 21, or 31 residues.

Figure 7

Prediction of reentrant helices into the membrane. The reentrant helices are highlighted with the rest of the protein shown transparent.

Table I:

Protein conformational switches with residues of known switch regions and predicted switches.

PDBID/ Protein	Switch region	BCL::Jufo9D	Octopus MC-H / else	PsiPred H / E / C	TMbetanet MC-E / else
1iyt Abeta 40	7-25 MC-H/SO-E 27-40 MC-H/SO-E	**switch seen**	***1 TM helix: 21-41***	***2 strands: 11-20 E 31-41 E***	2 TM strands: 10-22 29-41
1pfo Perfringolysin	190-217 SO-H/MC-E 288-311 SO-H/MC-E	**switch seen**	signal peptide/ outside	***3 helices: 184-195 H 287-298 H 305-314 H***	20 TM strands in whole protein; ***182-192 197-206 294-301***
7ahl α-hemolysin	108-149 SO-E/MC-E	**switch seen**	globular	3 strands: ***110-120 E*** 123-126 E ***136-148 E***	12 TM strands in whole protein; ***110-119 131-137 139-150***
1eft EF-TU	40-62 SO-H/SO-E 80-100 SO-H/SO-C	incorrect switch: SO-C/SO-H ambiguous	globular	1 strand, 4 helices: 44-46 E 47-50 H 55-58 H ***85-93 H 96-98 H***	22 TM strands in whole protein; 89-96

Bold font represents an identified switch (BCL::Jufo9D) where predicted outputs are shown in Supplementary Figure 8. Bold italic font indicates that one of the two conformations is identified (methods other than BCL::Jufo9D). Horizontal lines in the predicted cells separate different switches, if two are known. If at all possible, BCL::Jufo9D is the only method that is able to identify switches since it has the ability to output probabilities for secondary structure coupled with membrane environment. For helix-strand switches, Psipred may theoretically be able to detect them if the probabilities of both states are similar. However, in these examples Psipred unambiguously identified a single state with prediction probabilities above 0.75.

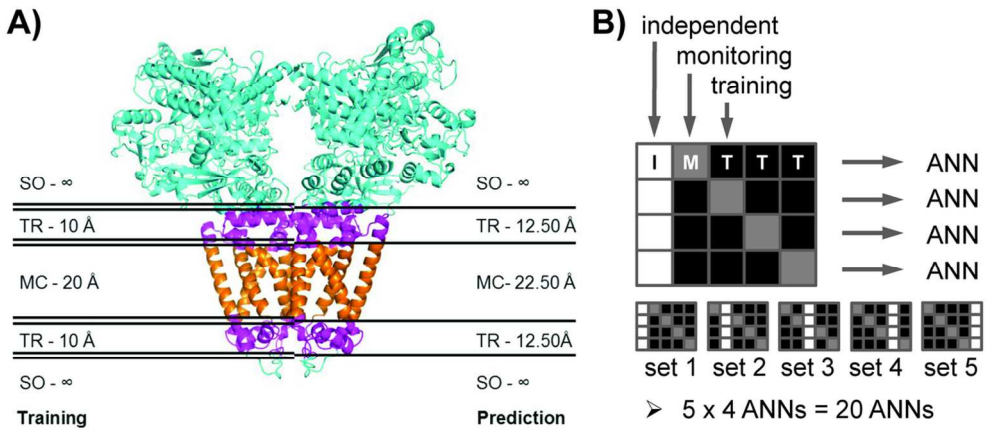

Figure 1
(A) Definition of membrane thicknesses in our MP database. For training, residues in the gap region of 2.5 A between membrane core/transition region and transition region/solution were disregarded. To report prediction accuracies the gap region was removed, i.e. all residues were taken into account, and the thicknesses of the regions were adjusted as shown. (B) Each ANN is trained on three subsets for training, one for monitoring the training process, and one as an independent test set. To avoid a bias in neither the independent test set, nor the monitoring set, both of these sets are permuted through all five subsets. This results in 20 ANNs that were trained. To report prediction accuracies, the outputs of the four ANNs in one set were summed, the prediction accuracies calculated, and then averages over all five sets were computed. This procedure boosts prediction accuracies in the three state outputs by 1-2% compared to the individual networks. It was tested whether post-processing the output with a second ANN further reduces the noise, but no significant improvements were obtained (data not shown).

114x51mm (300 x 300 DPI)

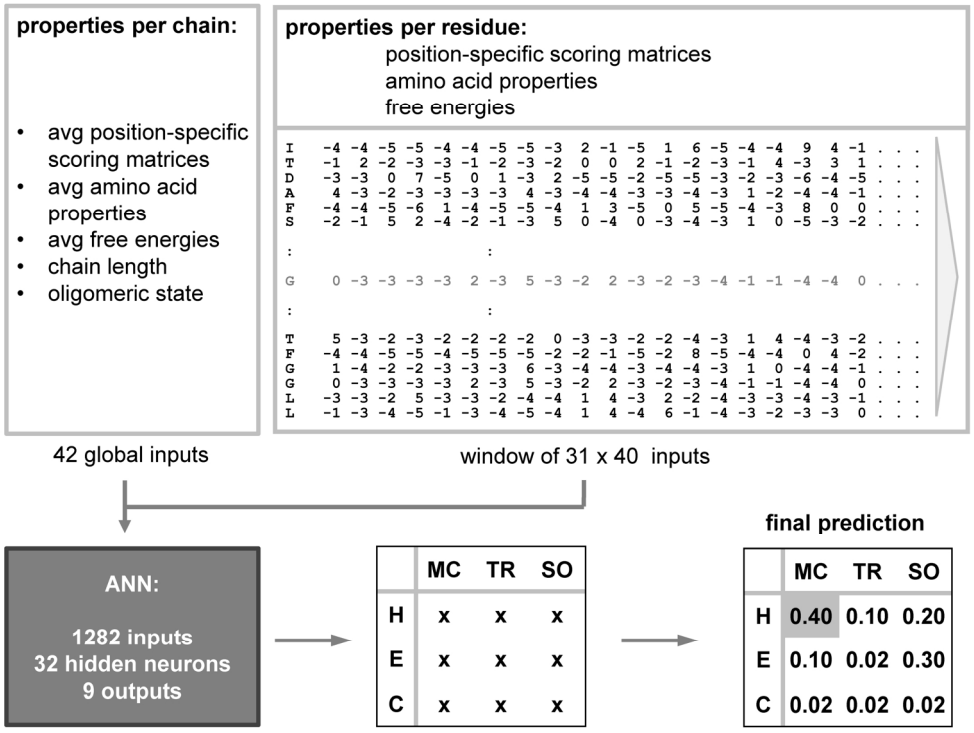

Figure 2
Setup for a single ANN. The residue at the center of the window is described by 1282 inputs. For this residue, a normalized nine-state prediction vector is the output. In this example, the predicted state for this residue is a helix in the membrane core (MC).

190x142mm (300 x 300 DPI)

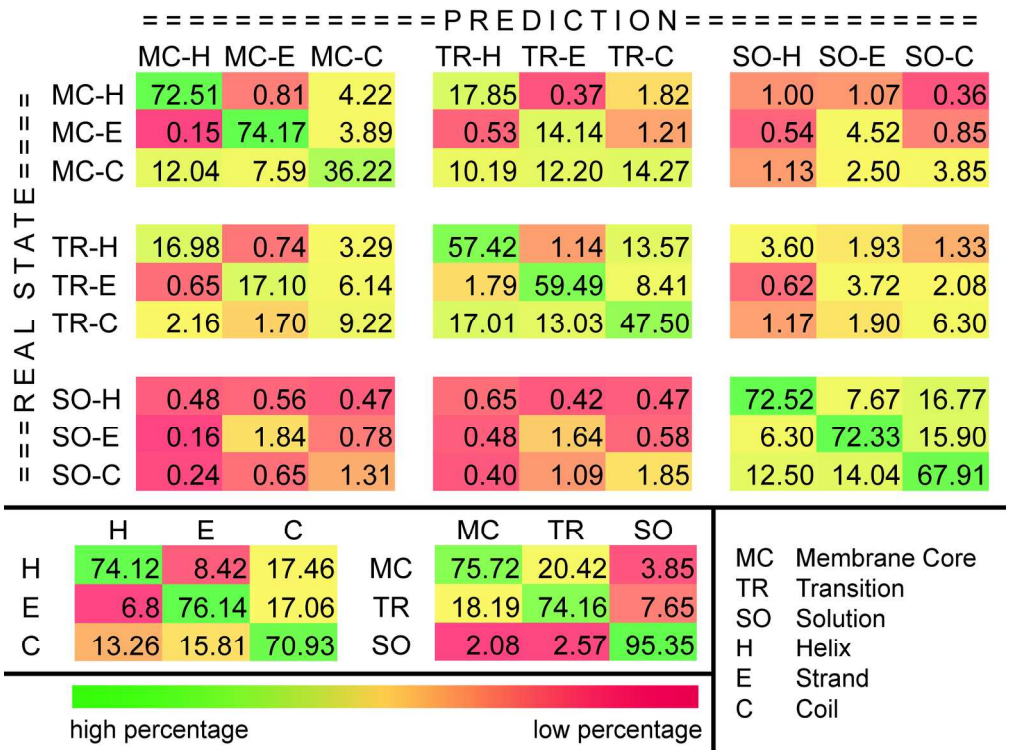

Figure 3
Averages of percent predicted residues over all independent datasets. The rows represent the "true" state, the columns represent the predicted state. Desired are large percentages in the matrix diagonals and low percentages in the off-diagonal elements. The overall nine-state accuracy is 70.3%, for SS prediction 73.2%, and for TM span identification 94.8%. The nine-state accuracies are summed to yield three-state SS predictions and three-state TM span predictions shown at the bottom.

190x142mm (300 x 300 DPI)

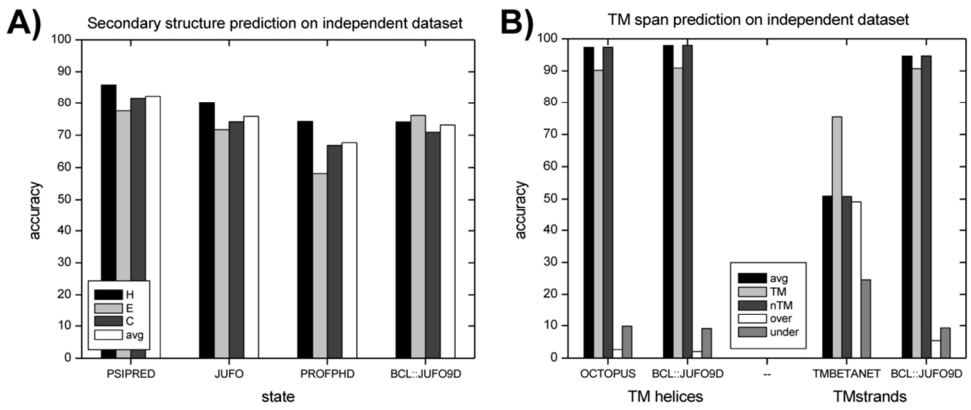

Figure 4
(A) Three-state secondary structure prediction comparing to methods trained on soluble proteins. (B) Performance of other two-state TM span prediction methods compared to BCL::Jufo9D that outperforms both of them. Since all other methods are limited to predicting either TM helices or TM strands, a separate comparison is required.

107x45mm (300 x 300 DPI)

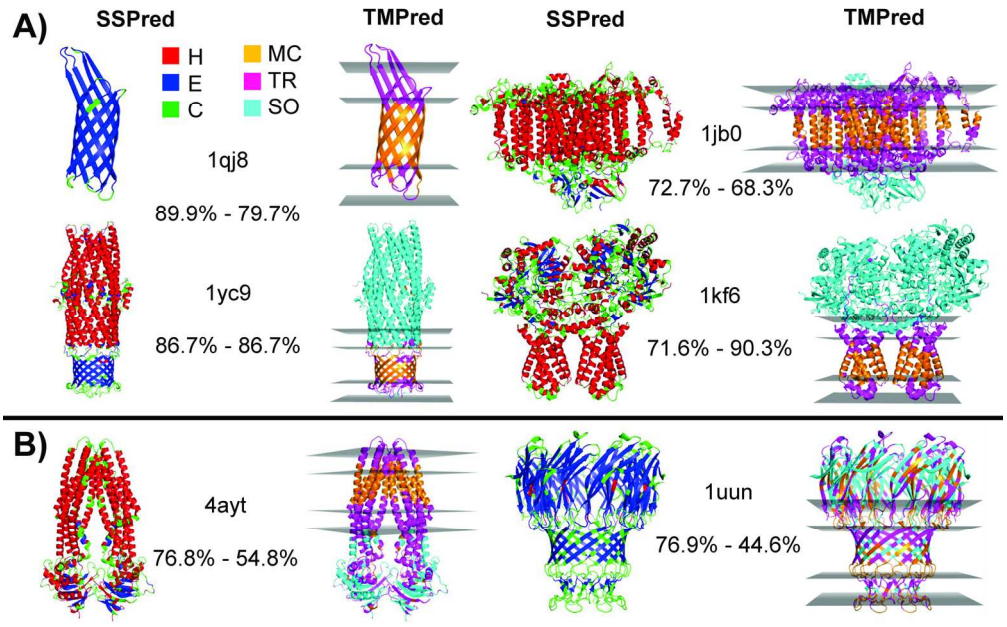

Figure 5
The sequences of these examples are used to predict the SS and TM state for each residue. These predictions are mapped onto the known structure. On the right panels the membrane core and transition regions on either side of the membrane are indicated by gray planes. H = prediction for helix, E = strand, C = coil, MC = membrane core, TR = transition region, SO = solution.
161x103mm (300 x 300 DPI)

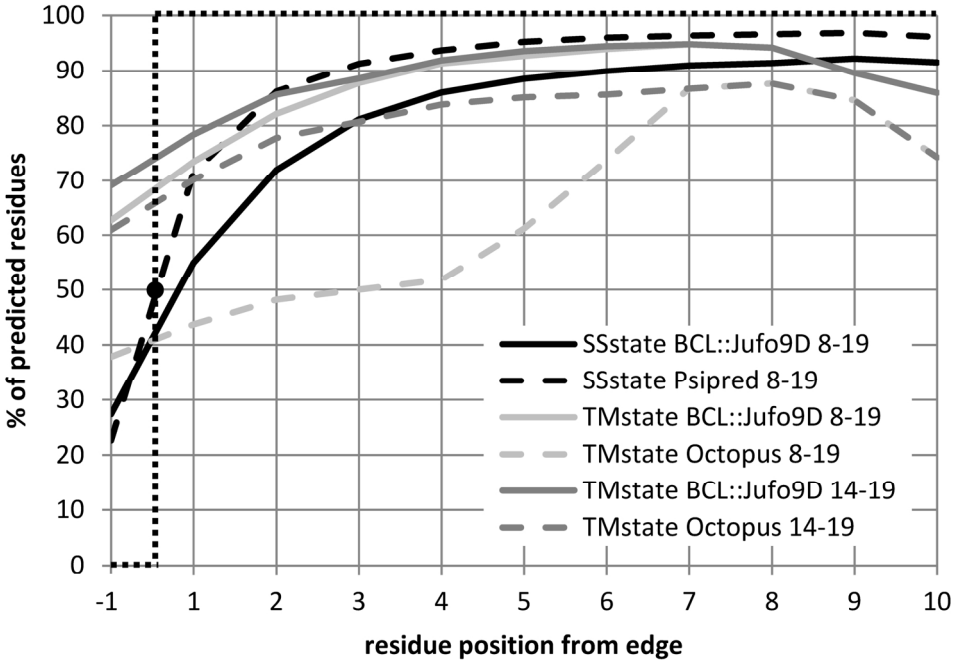

Figure 6
Percent of predicted residues vs. residue position of actual SS elements or TM spans. The residue position denotes the position from either side (N-terminal or C-terminal) of the SS element/TM span where position -1 is outside the SS element or TM span and position one is the first residue within. The dotted line denotes a perfect prediction with the black dot at the inversion point. The percent predictions for each position are averages over SS elements/TM spans between (2*residue position – 1) residues up to 19 residues corresponding to position 10. As an example, the TM state accuracy at position 4 is the average percentage at that position over TM spans of length 7 to 19. For the TM span percentages the 8-19 denotes the length of TM spans considered: 8 to 19 residues. Similarly, 14-19 only considers TM spans between 14 and 19 residues. This distinction was necessary since OCTOPUS only predicts TM helices with the length of 15, 21, or 31 residues.

88x63mm (600 x 600 DPI)

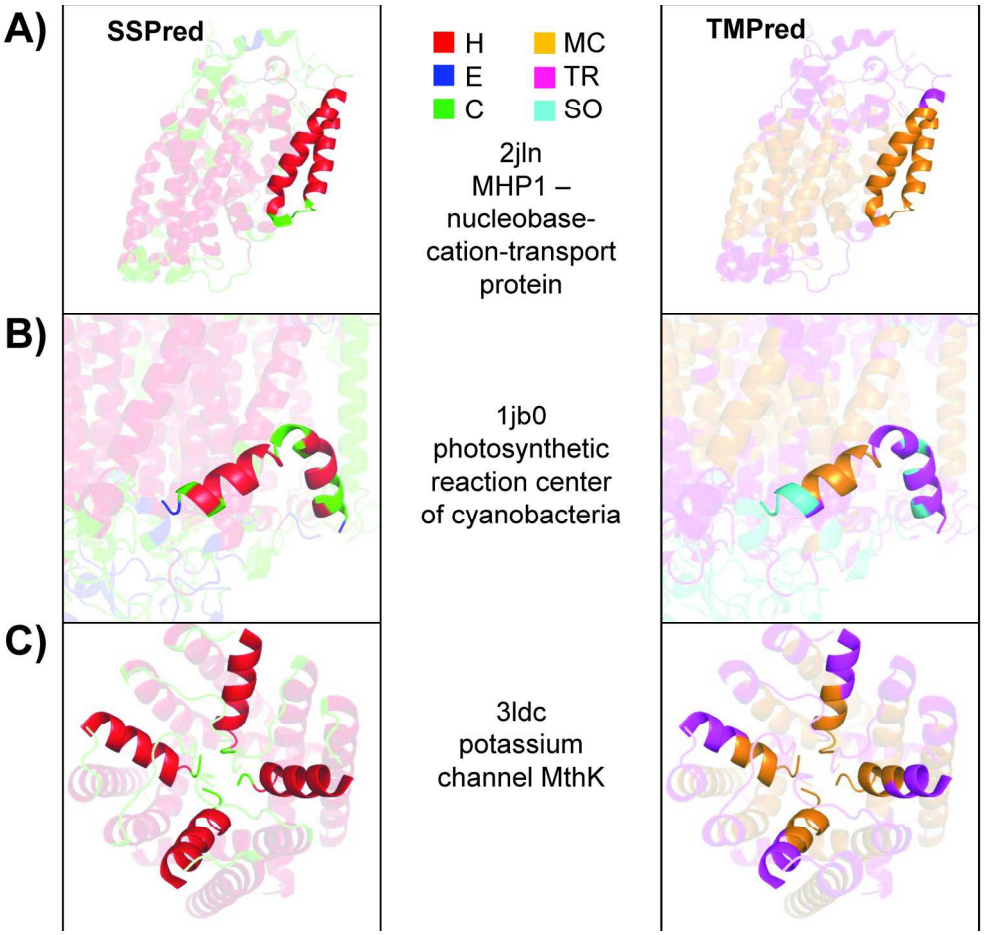

Figure 7
Prediction of reentrant helices into the membrane. The reentrant helices are highlighted with the rest of the protein shown transparent.

190x175mm (300 x 300 DPI)