PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PRIMSIPLR: Prediction of inner-membrane situated pore-lining residues for alpha-helical transmembrane proteins

Duy Nguyen, Volkhard Helms, and Po-Hsien Lee*

Center for Bioinformatics, Saarland University, D-66041 Saarbrücken, Germany

ABSTRACT

Transmembrane proteins such as transporters and channels mediate the passage of inorganic and organic substances across biological membranes through their central pore. Pore-lining residues (PLRs) that make direct contacts to the substrates have a crucial impact on the function of the protein and, hence, their identification is a key step in mechanistic studies. Here, we established a nonredundant data set containing the three-dimensional (3D) structures of 90 α-helical transmembrane proteins and annotated the PLRs of these proteins by a pore identification software. A support vector machine was then trained to distinguish PLRs from other residues based on the protein sequence alone. Using sixfold cross-validation, our best performing predictor gave a Matthews's correlation coefficient of 0.41 with an accuracy of 0.86, sensitivity of 0.61, and specificity of 0.89, respectively. We provide a novel software tool that will aid biomedical scientists working on transmembrane proteins with unknown 3D structures. Both standalone version and web service are freely available from the URL http://service.bioinformatik.uni-saarland.de/PRIMSIPLR/.

Proteins 2014; 00:000–000.

Key words: transmembrane protein; pore identification; support vector machine; amino acid composition; evolutionary conservation; imbalanced data.

INTRODUCTION

The function of a biological membrane is to separate cellular and subcellular compartments from their surroundings. With the assistance of elaborate membrane protein machineries that mediate material exchange or signal transduction across the membrane, organelles, or even entire cells can maintain specific ion and metabolite concentrations to perform their intrinsic functions. These machineries comprise a variety of transmembrane proteins such as receptors, transporters and channels. For example, G-protein-coupled receptors trigger signal transduction pathways inside the cell after ligands bind on the extracellular side of the membrane.[1] Adenosine triphosphate (ATP)-binding cassette (ABC) transporters utilize the energy from ATP hydrolysis to facilitate the translocation of a wide range of substrates from ions to oligopeptides and lipids.[2] The translocon complexes Sec61 and SecYEG integrate newly synthesized proteins into the membrane or translocate them across the membrane.[3] The tetrameric K^+ channels regulate the electrical potential or maintain the balance of electrolytes across the membrane by rapid and highly selective permeation of potassium ions.[4] The water-specific membrane channel protein, aquaporin, regulates water homeostasis in different kinds of cells.[5] The trimeric AcrB proteins function as multidrug efflux pumps in gram-negative bacteria and are associated with the resistance against various antibiotics.[6] All mentioned proteins contain pores (cavities, pockets, and channels) in their three-dimensional (3D) structures, either roughly in the center of a protein monomer or in the space between several monomers, to accommodate and transport their substrates. The malfunction of substrate binding or of the translocation mechanisms of these proteins has been related to diseases and thus they may become the targets of novel therapies and drug discovery.[7,8]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Deutsche Forschungsgemeinschaft through the Graduiertenkolleg 1276.

*Correspondence to: Po-Hsien Lee, Center for Bioinformatics, Building E2 1, R. 301, P.O. Box 15 11 50, D-66041 Saarbrücken, Germany.
E-mail: p.lee@mx.uni-saarland.de

Received 18 October 2013; Revised 8 January 2014; Accepted 16 January 2014
Published online 27 January 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24520

Since X-ray crystallographic data on such proteins is rare, dozens of structural prediction methods addressing various features of transmembrane proteins have been developed.[9] For example, OCTOPUS[10] and MEMSAT-SVM[11] predict the topology of α-helical transmembrane proteins including transmembrane helices, re-entrant helices and signal peptides. Interestingly, Nugent and Jones[12] also developed an extended version of MEMSAT-SVM that predicts pore-lining helices and residues. Tools such as TMX[13] and BTMX[14] predict the burial status, that is, whether a residue is exposed to the lipid bilayer or not, of each residue in α-helical and β-barrel transmembrane proteins, respectively. TMHcon[15] and MEMPACK[16] predict the helix–helix contact map of α-helical transmembrane proteins. All of these methods adopted a common strategy to predict structural features from the protein primary sequence. First, they consider evolutionary information of the protein that is typically represented by the position-specific scoring matrix (PSSM) derived from multiple alignments of homologous sequences. Second, all methods use machine learning algorithms such as support vector machines (SVMs), artificial neural networks, or hidden Markov models to build a model according to the features extracted from a set of membrane proteins with known 3D structures. Furthermore, methods such as TMDET[17] and PPM[18] predict the spatial arrangement inside the membrane (including position and orientation) for proteins with known 3D structures. For example, TMDET determines the position of a protein in the membrane by maximizing the relative hydrophobic membrane exposed surface area. PPM optimizes three structural parameters to obtain the minimal transfer energy of the protein from water to the lipid bilayer, namely the thickness of the hydrophobic slab, the tilt angle between the protein and the membrane normal, and the position of the protein along the membrane normal.

Pore-lining residues (PLRs) are crucial for the function of membrane transporters and channels because they directly contact the substrates or the water shell surrounding them. They are thus involved in recognition, desolvation, binding, and transportation processes of protein-substrate interactions. To our knowledge, there exists so far only the prediction method mentioned above for identifying PLRs developed by Nugent and Jones.[12] Their method identifies PLRs located in transmembrane helices of proteins and has been integrated into MEMSAT-SVM as an extension. MEMSAT-SVM first predicts the topology of TM helices and then identifies PLRs with respect to the predicted helices. Although most PLRs are indeed located in transmembrane helices, a certain portion of PLRs are located in loop regions or other locations. In this article, we present a single-step method to predict PLRs of α-helical transmembrane proteins from primary amino acid sequences. This method is based on a comprehensive data set termed PH90 and a new SVM classifier (see Materials and Methods section), and has been tested by stringent cross-validation. With improved prediction accuracy, this method should be of great use in the annotation of genomic sequence data and also provides clues to experimental biologists working on transmembrane proteins.

MATERIALS AND METHODS

Preparation of the data set

To collect a comprehensive data set of pore-containing α-helical transmembrane proteins, we accessed the regularly updated database PDBTM[17] (October 19, 2012 version) to obtain the list of PDB IDs with chain indices of all α-helical transmembrane proteins deposited in the RCSB Protein Data Bank.[19] We retrieved PDB files according to this list and filtered out structures that were not determined by experimental techniques (i.e., theoretical models). In PDB files, the primary sequence is recorded in the "SEQRES" section. However, to facilitate purification or crystallization, the protein sequence of the determined structure sometimes differs from that of the wild-type protein due to, for example, mutations, non-native amino acids, insertions, deletions, tags, fusion proteins, and so on. These modifications may be misleading when applying sequence identity-based clustering. Hence, we adopted wild-type protein sequences taken from the database Uniprot[20] according to their Uniprot IDs recorded in PDB files instead of the sequences in the "SEQRES" section.

Protein sequences of 1329 membrane protein structures were then clustered by the program BLASTClust requiring length coverage above 90% and sequence identity below 25%. Among all members of one cluster, we manually picked the protein with the visually most obvious and largest pores inside its 3D structure as the representative (exemplar) of the respective cluster. Some structures containing only one transmembrane helix or without any detectable pores inside the protein were discarded during this step.

We then used the program PoreID of our package PROPORES[21] to identify pores and to annotate PLRs of the representative structures. PoreID considers cylindrical volumes flanked by atom pairs and defines the space crossed by two approximately perpendicular cylinders as pore volume. This approach avoids the orientation dependency of the results obtained with some grid-based methods. However, a pore identification software such as PoreID usually provides tens of pores for an input structure. To confirm that the identified pores are biologically functional, we reviewed the original articles where the determined structures were reported and annotated only the pores mentioned in these articles. Figure S1 shows an example for the glycerol facilitator protein. After the mentioned data processing steps, 92 chains from 90

protein structures were retained. This data set with the obtained structure based annotations was named "PH90" set. The PDB IDs and protein names are listed in Table S1 of the Supporting Information.

From the OPM database and PPM server,[18] clear boundaries of the water/lipid interface were estimated for each protein. With the annotation of PLRs and the water/lipid boundary of the protein structure, we then classified each residue of a protein sequence into four types according to its location. The type "P" is used for PLRs in the transmembrane region, "M" denotes all other residues in the transmembrane region, "O" is assigned to residues outside of the membrane, and "N" are those residues whose coordinates are not contained in the PDB file. Figure S2 shows an example of this topological classification for the primary sequence of the bacterial glycerol facilitator (PDB id: 1FX8). The full data set is available at http://service.bioinformatik.uni-saarland.de/PRIMSIPLR/.

Machine learning and prediction

For each single residue of the protein sequences of the training set as well as for input sequences uploaded to our webserver, we generate 24 features as input for the machine learning algorithm. The first 20 features are position-specific scores generated by PSI-BLAST.[22] This profile captures the conservation pattern of the protein from a multiple sequence alignment of homologous protein sequences. Three iterations of PSI-BLAST were run for an input sequence against the nonredundant (nr) database and using commonly used parameters (i.e., the word size was set to 3, the penalty of gap opening to 11, the penalty of gap extension to 1, and threshold to 0.001). The next three features characterize physical properties of the 20 native amino acids, namely hydropathy, polarity, and flexibility. As hydropathy scale, we used a modified version of the Kyte–Doolittle hydropathy scale proposed by Juretić *et al.*[23] In their study, the hydropathy of tryptophan was increased and that of alanine was decreased with respect to the Kyte–Doolittle scale to obtain a better prediction of transmembrane helices. The polarity indices for amino acids were taken from the study of Zimmerman *et al.*[24] They approximated the polarity index of an amino acid as the relative electric potential generated by the dipole and the charged group of its side chain. Flexibility scales typically account for the typical degree of conformational flexibility of amino acid side chains. The flexibility scale of amino acids adopted here was proposed by Vihinen *et al.*[25] to predict continuous epitopes on proteins. It was derived by averaging the B-factors of each amino acid type in the PDB files of 92 protein structures. The last feature is the evolutionary conservation score computed by the program Rate4Site.[26] Rate4Site implements a Bayesian approach to estimate the position-specific evolutionary rate of a protein sequence.

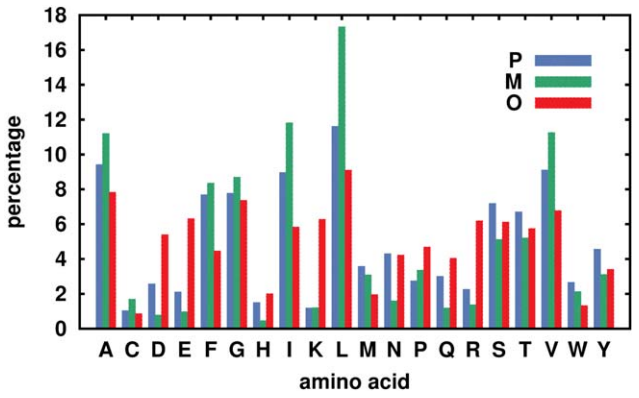

Figure 1

Amino acid composition of pore-lining residues (type P) and nonpore-lining residues (types M and O). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The machine learning algorithm applied in this study is support vector classification implemented in the software package LIBSVM.[27] The SVM was implemented to classify and identify PLRs in the transmembrane region (type P of our annotation) against all other residues of the protein structure (types M, N, and O). We tested several combinations of features to determine how these features affect the performance of the predictions (see Results section). The performance of each SVM classifier was evaluated by the common measures accuracy, sensitivity, specificity and Matthews's correlation coefficient (MCC).

RESULTS AND DISCUSSION

Amino acid composition

Our nonredundant PH90 dataset consists of 92 structures of alpha-helical transmembrane proteins with central pores. 3460 of the 36,885 residues were assigned by PoreID as PLR or type "P." The remaining 33,425 residues were classified to any of the other types M, N, or O. The ratio between P and M, N, and O residues is 1:9.66. N-type residues were included in the group of non-PLR residues together with M- and O-type residues. The reason for this is that most of the unresolved parts of X-ray structures of TM proteins (residues of type N) are located in N-terminal, C-terminal, or flexible loop regions of transmembrane proteins. Hence, they are very unlikely to belong to the group of PLR residues.

The amino acid composition for residues of type P, M, and O is shown in Figure 1. As expected, nonpolar amino acids such as A, I, L, and V are relatively abundant (>10%) in M-type residues because they are either buried inside the protein structure or exposed to the

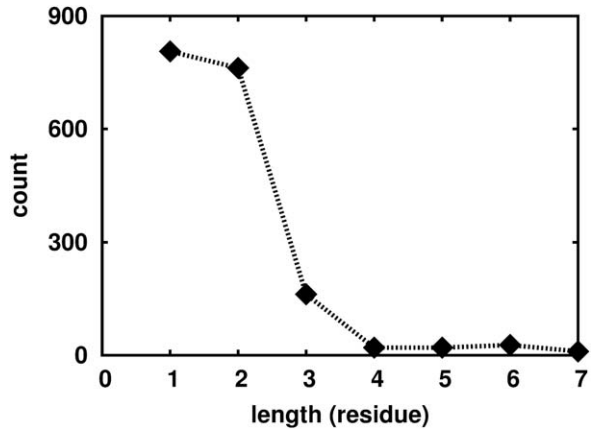

Figure 2

Length of pore-lining fragments when mapped on the primary sequences.

hydrophobic lipid bilayer. Unexpectedly, A, I, L, and V are also frequently found as P-type residues (>8%). This means that nonpolar residues are important components of pores. Even if the substrate is hydrophilic, nonpolar residues form hydrophobic patches and facilitate substrate transportation as in aquaporin.[28] PLR positions ("P" positions) contain larger fractions of charged (D, E, and R) and polar residues (H, N, and Q) than M-type residues because they may be involved in contacts with buried solvent molecules in the pore or with the charged or polar transported substrates. Amino acids W and Y with aromatic side chains occur slightly more often in P-type residues than in the other two cases. These amino acids often function as gating residues of channel proteins or transporters as mentioned in our previous study.[21] In contrast to membrane positions, O-type residues are substantially enriched in charged amino acids such as D, E, K, and R.

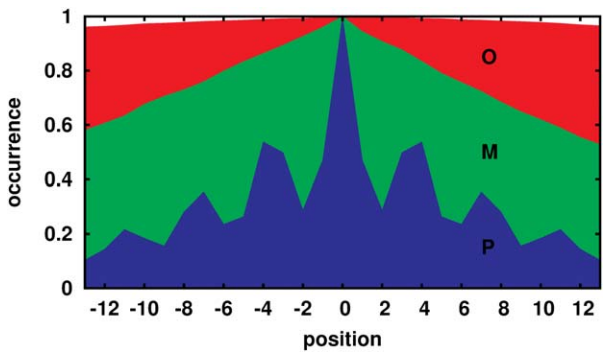

Figure 3

Distribution of neighboring residues around a central pore-lining residue in the PH90 set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

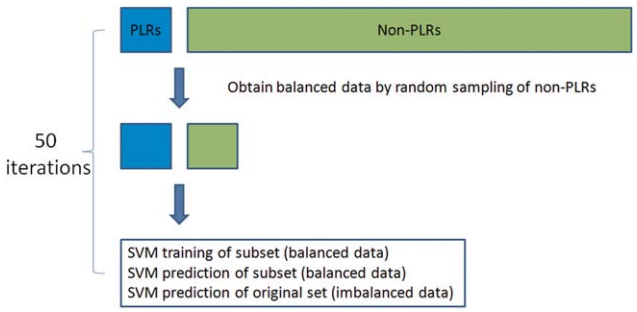

Figure 4

Multiple independent random sampling and training scheme for imbalanced data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

When mapping structurally identified PLRs onto the protein sequences, 94.4% of the pore-lining fragments are shorter than three residues. This is due to the fact that generally only one side of pore forming α-helices faces the pore and perfect α-helices contain 3.6 residues per turn (Fig. 2). One of the exceptions in the PH90 set is the ABC transporter with PDB ID: 3QF4 chain B. Its longest pore-lining fragment is 17 residues long. This protein is heterodimeric and contains a large channel formed between two protomers. This pore-lining fragment is part of the transmembrane helix that is not aligned with the other helices and protrudes into the central cavity, see Figure S3.

To characterize the local environment of PLRs in the primary sequence, we also analyzed the distribution of neighboring residues of PLRs annotated by PoreID (see Fig. 3). The pattern of neighboring PLR peaks is consistent with the periodicity of 3.6 residues per turn of an ideal α-helix. The second and third highest frequency of PLRs was observed at the third and fourth positions downstream (+3, +4) and upstream (−3, −4) from the central PLR. This means that a PLR located in a TM helix frequently has neighboring PLR residues one helix turn away in both directions. Although further peaks of PLR were observed at the 7th and 11th residues, the fraction of non-PLR (types M and O) in these positions is considerably larger.

Training scheme for imbalanced data

Our PH90 set is an imbalanced data set that comprises 9.38% of positive data (3460 PLRs) and 90.62% of negative entries (33,425 non-PLRs). In this study, we used the radial basis function kernel for all SVM trainings. Training an SVM on the entire PH90 set by either equal costs or weighted costs for two classes resulted in high overall prediction accuracy above 90%, a nearly perfect specificity (0.99), but a low sensitivity (about 0.3). It showed an obvious bias toward non-PLRs and is only

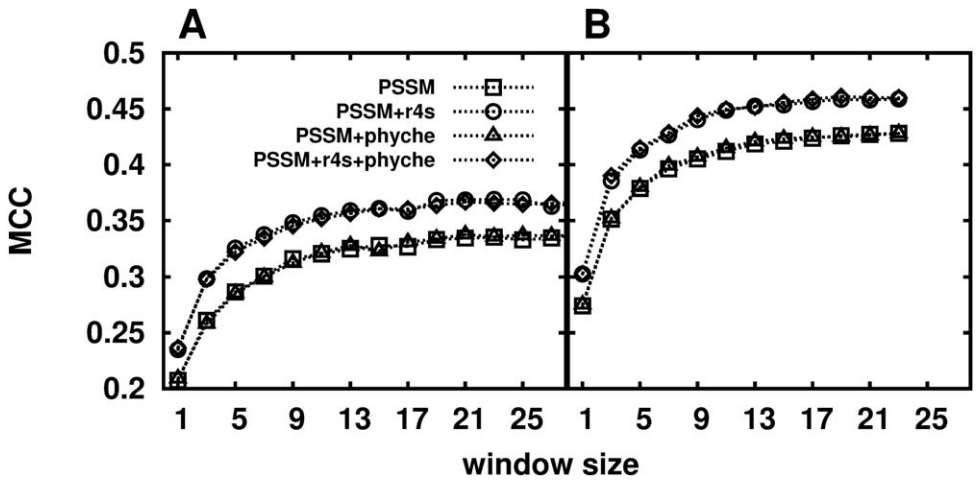

Figure 5

Performance of (**A**) the PH90 set and (**B**) the PH90ext set under different combinations of features and for different window sizes. "r4s" stands for conservation scores generated by Rate4Site. "phyche" is used as abbreviation for physicochemical properties, which comprises of hydropathy, polarity, and flexibility in this study.

slightly better than the naive prediction that assigns all data points to the non-PLR class. Such a classification clearly defeats the purpose of this study which puts a higher emphasis on identifying PLR residues than non-PLR residues. To deal with this issue, we adopted a multiple independent random sampling and training scheme shown in Figure 4. An SVM was trained on all PLRs and a randomly sampled equal number of non-PLRs. Then the predictive power of the trained model was evaluated on the entire data set. Training of the random sampling data was performed 50 times and the model with highest MCC value was then taken as the model of the configuration (features and parameters) for the subsequent prediction process. The idea of this approach is to find a balanced and most representative subset of non-PLRs by multiple random sampling, training and evaluation. With this training approach, the MCC and sensitivity of classification of the imbalanced data was substantially improved whereas some sacrifice of specificity is unavoidable.

Feature and window size selection

The PSSM obtained from PSI-BLAST is a frequently used feature in protein structure prediction (see Introduction). To test whether additional features such as physicochemical properties and conservation score improve the prediction of PLRs, we performed SVM model training for different windows sizes varied over a large range and different combinations of additional features and PSSM. Default values of two parameters in LIBSVM (gamma and cost) were used in these test runs. Since the parameters used here may not be the most optimal ones, we would rather like to compare the overall performance of a profile than a single window size to determine the best combination of features. Besides the PLRs annotated by PoreID, we also tested the case that includes directly adjacent residues. This strategy was suggested by Nugent and Jones[12] to balance the amount of positive and negative data and to indirectly account for conformational dynamics that cannot be captured in crystal structures. We termed this PLR-annotation extended set as "PH90ext" to be distinguished from the original PH90 set. The PH90ext set contains 6413 PLRs and 30,472 non-PLRs. In Figure 5, PH90ext has higher MCC values than those of PH90 under all combinations of features and window sizes. However, sensitivity and specificity are similar in both cases. This means that the performances are likewise similar for both cases. Because the number of positive data (PLRs) in PH90ext is almost twice as much as in PH90, the higher MCC in PH90ext is mainly due to the relatively higher ratio between true-positive predictions and all residues. In both datasets, MCC values increased quickly for longer windows up to a window size of 13 where they started to gradually converge. Figure 5 shows that adding the conservation score computed by Rate4Site substantially improves the performance of the SVM predictor for the PH90 and PH90ext sets. This is likely the case because Rate4Site considers the relationship between the homologous sequences, which is not captured by PSSMs. Rate4Site builds a phylogenetic tree for homologous sequences and estimates the evolutionary rates of each position from the topology and branch lengths of the phylogenetic tree. On the other hand, the three physicochemical properties that were used at the same time to train the SVM models showed no clear effect for PH90 [Fig. 5(A)] and only a slight improvement for PH90ext. This suggests that the

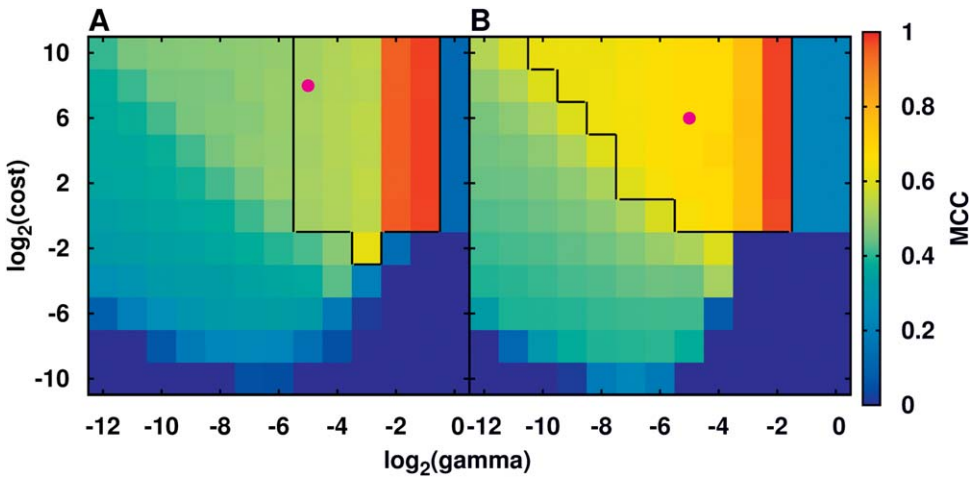

Figure 6

Grid search of gamma and cost parameters used in SVM training for (**A**) PH90 set and (**B**) PH90ext set. MCC values indicate the performance of an SVM that was trained on the entire data set. For the (gamma, cost) pairs in the framed areas, performance was also evaluated by sixfold cross-validation (see text). Those MCC values after sixfold cross-validations are not shown here except for the magenta points that show the optimized parameters with highest averaged MCCs after cross-validation.

PSSMs already contain most information about these three properties in an implicit way. The configurations with best MCC were then taken to be further optimized for the gamma and cost parameters. For the PH90 set, the best configuration is a window size of 23 with the combinations of PSSM and conservation score. For the PH90ext case, a smaller window size of 19 with all features including PSSM, conservation scores and physicochemical properties was used as the start for the optimization of SVM parameters.

Grid search and cross-validation of optimal SVM training parameters

After determining the best window size and features of the training data, we optimized the two parameters of SVMs, gamma and cost. Gamma is an adjustable parameter of the radial basis function kernel (RBF kernel) that was used for mapping the feature space in the SVM. The radial basis function is defined as $\exp(-\text{gamma}^*|u - v|^2$ where u and v are two feature vectors. Gamma can be set as a positive value no greater than 1. The cost is the penalty of misclassification for each data point in the SVM. For both the PH90 and the PH90ext sets, we performed a wide ranged grid search for both parameters. The results of the training evaluations (MCC values) are shown in Figure 6. The results for the two sets follow a similar trend. Larger gamma with higher cost leads to better performances. The least successful separation of training data happened when gamma was set between 2^{-4} and 1 with cost between 2^{-10} and 2^{-2}. In contrast, the red areas show nearly perfect classification when gamma is 2^{-2} and cost is larger than 1. Although Figure 6 shows the training accuracies of the entire data set, we then estimated the true performance of the predictor by sixfold cross-validation. The framed areas in Figure 6 mark the (gamma, cost) pairs with MCC larger than 0.5 for the PH90 set and larger than 0.6 for the PH90ext set. These pairs were further evaluated by sixfold cross-validation. (Note that the MCC values after sixfold cross-validations are not shown in Fig. 6.) For both data sets, we found that the averaged MCCs for the previously perfect classification area of entire data set (red areas in Fig. 6) are only around 0.15 after sixfold cross-validation. This means that the model over-fitted the features of the training data and this model could not accurately predict cases whose features deviate from those of the training data. For the PH90 set, the best prediction emerged when (gamma, cost) were set to (2^{-5}, 2^{8}). The averaged per residue level MCC is 0.39, accuracy is 0.78, sensitivity is 0.82, and specificity is 0.77. For the PH90ext set, the best averaged MCC of sixfold cross-validation was obtained for (2^{-5}, 2^{6}) with averaged per residue level MCC, accuracy, sensitivity and specificity as 0.49, 0.78, 0.83, and 0.77, respectively. As mentioned in the previous section, the prediction performances of both data sets are similar because of close accuracy, sensitivity, and specificity. The higher MCC of the PH90ext set is due to its higher true-positive ratio. Besides, we then checked the classification of the entire data sets by a SVM with optimized parameters. We found that the percentage of the false positives that are due to type M residues being classed as type P is about 74% in both PH90 and PH90ext (about 24% due to O-type residues and 2% from N). Finding more specific and representative features to distinguish these two types of residues (P and M) remains a tough challenge for future work.

For imbalanced data as was used here, the specificity of 0.77 obtained by cross-validation contains a

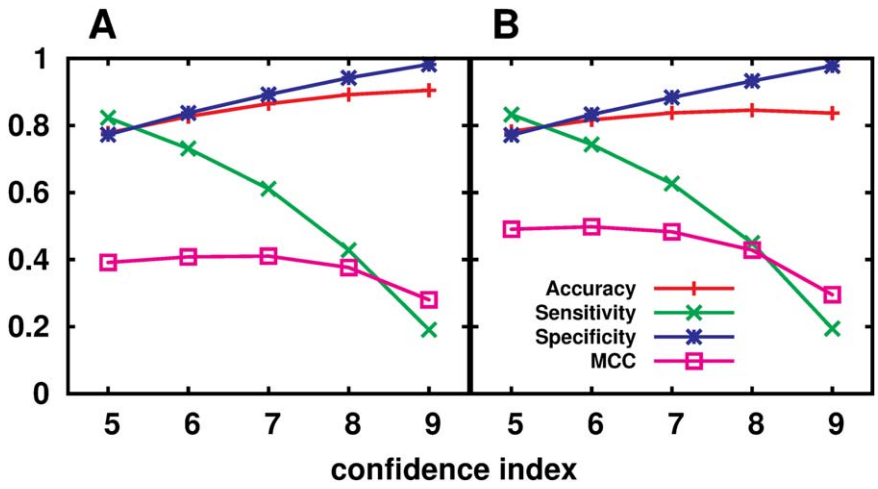

Figure 7

The relationship between confidence index and averaged performance of prediction in sixfold cross-validation. (**A**) PH90 set and (**B**) PH90ext set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

considerable fraction of false positives compared to the true positives. Thus, we introduced a confidence index for each positive prediction by converting the probabilistic output provided by LIBSVM into a discrete scale from 5 to 9. This mapping of an SVM output into a probability by a sigmoid function was proposed by Platt.[29] Lin *et al.*[30] then improved the algorithm and implemented it in LIBSVM. As shown in Figure 7, when the threshold of positive prediction was set to tighter values according to the confidence index, accuracy and specificity gradually increased whereas the sensitivity sharply decreased. For the PH90 set with a threshold of 7, MCC was slightly raised to 0.41 with sensitivity as 0.61 and specificity as 0.89. For the same threshold and the PH90ext set, MCC, sensitivity and specificity were 0.48, 0.63, and 0.88, respectively. Generally, a larger confidence index implies a higher ratio of true positives to false-positive predictions and provides more reliable predictions for experimental biologists whose studies are related to PLRs.

Evaluation on novel protein structures

To evaluate the performance of our predictor on a set of novel protein structures that were not used during the training of the method, we collected a test set including only structures that were submitted to the PDB databank after composing the PH90 set. We compared two lists of α-helical transmembrane proteins obtained from PDBTM[17] either on October 19, 2012, or on July 26, 2013, and processed the novel protein sequences and structures by the same procedure mentioned in "Materials and Methods" section. This resulted in 23 structures that are only contained in the latter list and that share less than 25% sequence identity with any member of the training set. They contain 10,251 residues with 693 assigned as PLR and 9558 as non-PLR (the ratio is 1:13.8). We named this test set Test23 (details are shown in Supporting Information Table S3) and evaluated the predictive power of our PRIMSIPLR method on this set. When the threshold of confidence score was set to 7, we obtained a similar performance to the cross-validation results reported before with higher sensitivity and slightly lower specificity and MCC value (Table I).

To have a fair comparison with MEMSAT-SVM, we derived the modified Test23ext set from Test23 by extending the PLR labels to directly adjacent residues, as was done for the PH90ext set. The same threshold of the confidence score was applied when testing our model on the Test23ext set. According to Tables I and II, the optimized models of PRIMSIPLR showed equivalent performance for both Test23 and Test23ext. When considering the set of residues predicted as PLR (TPs + FPs), about one in four predicted PLR residues is correct for Test23. When

Table I
Performance of PRIMSIPLR Evaluated on the Test23 Set

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	MCC
PRIMSIPLR	469	7862	1221	216	0.85	0.68	0.87	0.37

MCC, Matthews's correlation coefficient.
TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative.

Table II
Performance of PRIMSIPLR and MEMSAT-SVM Evaluated on the Test23ext Set

	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	MCC
PRIMSIPLR	962	7358	1079	461	0.84	0.68	0.87	0.48
MEMSAT-SVM	791	7187	1250	632	0.81	0.56	0.85	0.35

MCC, Matthews's correlation coefficient.
TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative.

including next-neighbors of PLRs as is done in Test23ext, about one in two predicted PLRs is correct (TP). In addition, our method outperformed MEMSAT-SVM and yielded more than 10% improved sensitivity and MCC. When considering the performance of each individual protein, PRIMSIPLR has better MCC values than MEMSAT-SVM for 19 of the 23 proteins (Supporting Information Table S4). One of the cases where MEMSAT-SVM made better predictions than PRIMSIPLR is the protein Magnesium transporter (PDB ID: 4EV6) that was not included in their training set. Remarkably, MEMSAT-SVM gave a nearly perfect prediction with a MCC value of 0.94 compared to 0.65 with PRIMSIPLR. We also evaluated the performance of PLR prediction obtained by PRIMSIPLR when PLR annotations for the members of the Text23 set were made by PoreWalker as was done in the training phase of MEMSAT-SVM. The results, see Supporting Information Table S5, are highly similar to those in Table II. This suggests that the pore identification programs used here are not the crucial factor for the performance of PLR prediction.

Some structures of transporters such as Leucine transporter,[31,32] glutamate/GABA antiporter[33] were determined in one of several alternative ligand transportation states. For the Leucine transporter, the complete view of PLRs can be captured by including multiple structures resolved in different states (i.e., 3F3A[31] is in outward-open conformation and 3TT3[32] is in inward-open conformation). However, the glutamate/GABA antiporter[33] was crystallized either in a blocked or in an inward-open state so that the identification of PLRs of these structures cannot present all possible residues that form contacts with the substrate. This may have caused some errors in the prediction.

CONCLUSIONS

Here, we presented the new method PRIMSIPLR for predicting PLRs from the sequences of α-helical transmembrane proteins. This method was developed on a comprehensive data set containing 90 protein structures. The amino acid composition of the data set reflects the expected characteristics of residues in different environments. Residues outside the membrane prefer to be charged and polar, whereas pore-lining amino acids have a higher content of charged and polar side chains compared to the other residues embedded in the membrane. Amino acids with aromatic side chains are crucial for pore lining residues by virtue of their function as gate. The typical length of a PLR stretch and the periodic pattern of PLRs are related to the structural trait of the α-helix.

We trained an SVM with a multiple independent random sampling scheme to account for the imbalanced nature of our data set. The evolutionary conservation score calculated by Rate4Site[26] gave substantial improvements in the prediction results whereas the three physicochemical properties had no apparent effects. The best MCC is 0.39 with an accuracy of 0.78. Furthermore, we provide a confidence index for each positive prediction. A higher confidence index implies a more reliable prediction. Our predictor outperforms MEMSAT-SVM on 19 of 23 nonredundant novel protein structures. The most challenging issue is to distinguish "P" and "M" residue types inside the membrane. More characteristic and representative features are necessary and the key issues to be overcome in future work.

We provide a tool for experimental biologists who work on α-helical transmembrane proteins. The predictions may be useful for designing mutations or for the design of inhibitors. Both the standalone version and the web service are freely available from the URL http://service.bioinformatik.uni-saarland.de/PRIMSIPLR/. The standalone software was developed in the PERL programming language and was tested on a Linux system.

ACKNOWLEDGMENTS

The authors thank Dr. Peter Hildebrand and Dr. Michael Hutter for helpful comments on the article.

REFERENCES

1. Marinissen MJ, Gutkind JS. G-protein-coupled receptors and signaling networks: emerging paradigms. Trends Pharmacol Sci 2001;22:368–376.
2. Rees DC, Johnson E, Lewinson O. ABC transporters: the power to change. Nat Rev Mol Cell Biol 2009;10:218–227.
3. Zimmermann R, Eyrisch S, Ahmad M, Helms V. Protein translocation across the ER membrane. Biochim Biophys Acta 2011;1808:912–924.
4. Miller C. An overview of the potassium channel family. Genome Biol 2000;1(4):reviews0004.0001-0004.0005.
5. King LS, Kozono D, Agre P. From structure to disease: the evolving tale of aquaporin biology. Nat Rev Mol Cell Biol 2004;5:687–698.

6. Yu EW, Aires JR, Nikaido H. AcrB multidrug efflux pump of Escherichia coli: composite substrate-binding cavity of exceptional flexibility generates its extremely wide substrate specificity. J Bacteriol 2003;185:5657–5664.
7. King LS, Yasui M, Agre P. Aquaporins in health and disease. Mol Med Today 2000;6:60–65.
8. Shieh CC, Coghlan M, Sullivan JP, Gopalakrishnan M. Potassium channels: molecular defects, diseases, and therapeutic opportunities. Pharmacol Rev 2000;52:557–593.
9. Tsirigos KD, Hennerdal A, Kall L, Elofsson A. A guideline to proteome-wide alpha-helical membrane protein topology predictions. Proteomics 2012;12:2282–2294.
10. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 2008;24:1662–1668.
11. Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 2009;10:159.
12. Nugent T, Jones DT. Detecting pore-lining regions in transmembrane protein sequences. BMC Bioinformatics 2012;13.
13. Park Y, Hayat S, Helms V. Prediction of the burial status of transmembrane residues of helical membrane proteins. BMC Bioinformatics 2007;8:302.
14. Hayat S, Walter P, Park Y, Helms V. Prediction of the exposure status of transmembrane beta barrel residues from protein sequence. J Bioinform Comput Biol 2011;9:43–65.
15. Fuchs A, Kirschner A, Frishman D. Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. Proteins 2009;74:857–871.
16. Nugent T, Jones DT. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol 2010;6.
17. Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 2004;20:2964–2972.
18. Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI. Positioning of proteins in membranes: a computational approach. Protein Sci 2006;15:1318–1333.
19. Rose PW, Bi CX, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM, Bourne PE. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 2013;41:D475–D482.
20. Apweiler R, Bairoch A, Wu CH. Protein sequence databases. Curr Opin Chem Biol 2004;8:76–80.
21. Lee PH, Helms V. Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. Proteins 2012;80:421–432.
22. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
23. Juretić D, Lučić B, Zucić D, Trinajstić N. Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. In: Cyril P, editor. Theoretical and computational chemistry, Vol. 5. The Netherlands: Elsevier; 1998. pp 405–445.
24. Zimmerman J, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170–201.
25. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
26. Mayrose I, Graur D, Ben-Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol 2004;21:1781–1791.
27. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intel Syst Tec 2011;2.
28. Murata K, Mitsuoka K, Hirai T, Walz T, Agre P, Heymann JB, Engel A, Fujiyoshi Y. Structural determinants of water permeation through aquaporin-1. Nature 2000;407:599–605.
29. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 1999;10:61–74.
30. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Mach Learn 2007;68:267–276.
31. Singh SK, Piscitelli CL, Yamashita A, Gouaux E. a competitive inhibitor traps LeuT in an open-to-out conformation. Science 2008;322:1655–1661.
32. Krishnamurthy H, Gouaux E. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. Nature 2012;481:469–474.
33. Ma D, Lu PL, Yan CY, Fan C, Yin P, Wang JW, Shi YG. Structure and mechanism of a glutamate-GABA antiporter. Nature 2012;483:632–636.