PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure, and evolutionary information

Marcin J. Mizianty and Lukasz Kurgan*

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

ABSTRACT

Membrane proteins (MPs) are difficult to identify in genomes and to crystallize, making it hard to determine their tertiary structures. MPs could be categorized into α-helical (AMP) and outer membrane proteins which mostly include beta barrel folds (OMBBs). The AMPs are relatively easy to predict from a protein sequence because they usually include several long membrane-spanning hydrophobic α-helices. The OMBBs play important roles in cell biology, they are targeted by multiple drugs, and they are more challenging to identify as they have shorter membrane-spanning regions which lack a folding patern, that is, as consistent as in the case of the AMPs. Hence, accurate *in silico* methods for prediction of OMBBs from their primary sequences are needed. We present an accurate sequence-based predictor of OMBBs, called OMBBpred, which utilizes a Support Vector Machine classifier and a custom-designed set of 34 novel numerical descriptors derived from predicted secondary structures, hydrophobicity, and evolutionary information. Our method outperforms modern existing OMBB predictors and achieves accuracy of above 98% when tested on two existing benchmark datasets and 96% on a new large dataset. OMBBpred reduces the error rates of the second best method, depending on the dataset used, by between 13 and 65%, and generates predictions with high specificity of above 96%. Our solution is a useful tool for high-throughput discovery of the OMBBs on a genome scale and can be found at http://biomine.ece.ualberta.ca/OMBBpred/OMBBpred.htm.

Proteins 2011; 79:294–303.

Key words: outer membrane proteins; beta barrel; secondary structure segments; evolutionary information; structural genomics; support vector machine.

INTRODUCTION

Although the number of solved proteins structures increases every year, the gap between the number of the known protein sequences and the known protein structures is rapidly growing. Currently, about 59,000 protein structures, which are deposited into the protein data bank (PDB) database,[1] are resolved from among over 9.5 million known nonredundant protein chains. Membrane proteins are among the most challenging targets as they are hard to crystallize and therefore hard to map into 3D structure.[2] At the same time, around 25% of genetic code is assumed to represent membrane proteins.[3] These proteins play crucial role in cell biology, as they are responsible for interaction between interior and exterior of the cell, and they serve as targets for numerous drugs used in human and veterinarian medicine.[4] The membrane proteins could be categorized into the α-helical membrane proteins (AMPs) and the outer membrane proteins (OMP). Almost all known proteins in the latter category, with a handful of exceptions, are outer membrane beta barrel proteins (OMBBs). Recently, two OMPs which fold into an alpha barrel-like structure were found.[5,6] We concentrate on building a predictive model for OMBBs, since the low number of the currently characterized alpha barrel proteins does not allow for building and a reliable validation of a model that would cover all OMPs. AMPs are relatively easy to predict from the protein sequence as they include several long hydrophobic α-helices which span through a cell membrane. OMBBs are found in the outer membranes of gram-negative bacteria, mitochondria, and chloroplasts[7,8] and they perform diverse functional roles, including bacterial adhesion, material transport and they support structural integrity of the cell wall.[7–9] Identification of OMBBs are not easy, mainly because they have shorter membrane spanning region, which assemble into β-barrels, and they lack a clear pattern in their membrane spanning strands.[10]

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: Natural Sciences and Engineering Research Council (NSERC) of Canada, Killam Trusts
*Correspondence to: Lukasz Kurgan, Department of Electrical and Computer Engineering, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, Canada T6G 2V4. E-mail: lkurgan@ece.ualberta.ca.
Received 7 April 2010; Revised 17 August 2010; Accepted 9 September 2010
Published online 22 September 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.22882

Recent years have seen development of several *in silico* methods for the prediction of OMBBs from the protein sequence. These methods usually work in two steps, where first the protein sequence is encoded into a feature vector which is next inputted into a classification model to produce the prediction. Various models have been utilized to date, including Hidden Markov Models,[10–13] simple statistical analysis,[14–19] nearest neighbors,[20–23] quadratic discriminant,[24] neural networks,[25,26] radial basis function networks,[27] and support vector machines (SVMs).[28,29] Most of these methods used a fixed size feature vector to represent the sequence. They utilized various approaches to extract the features, including the usage of a simple amino acid (AA) composition[16,18,20–23,27–29], pseudo AA composition,[24,24] dipeptide composition,[17,23,27,28] β-barrel score,[16] physicochemical AA properties,[14,26,29] and evolutionary information[19–23,27]. Only two methods utilized predicted secondary structure to find OMBBs[14,15] although this information seems crucial since specific arrangements of the secondary structures define the transmembrane region of the OMBBs which differentiates them from the AMPs. Unfortunately these works applied a simple statistical approach to compute the predictions and used the secondary structure in a limited way as they only considered the composition of a few AAs in the predicted strands[15] or the overall content of predicted strands in the protein chains.[14]

Our aim is to build a novel model, referred to as OMBBpred, which improves prediction quality when compared with the modern existing methods. This is accomplished by fusing multiple approaches to extract the features (numerical descriptor of the input sequence) including predicted secondary structure, AA composition, hydrophobicity of residues, and their evolutionary conservation. We also carefully craft novel features based on the predicted secondary structure to not only consider the content of the predicted strands, but also most importantly to quantify the size and arrangements of the predicted secondary structure segments. The latter should be helpful in finding β-barrels that are formed by all-next-neighbor β-sheets.[30] This means that all paired strands that form the β-sheets are adjacent in the sequence, which in turn results in a relatively high number of strand-coil-strand motifs. Our motivation to use these features comes from a similar approach that was recently shown to improve the accuracy of prediction of protein structural classes,[31] which shares similarities with the OMBB prediction problem. Specifically, proteins from given structural classes and OMBBs are characterized by certain arrangements of secondary structures. Finally, instead of relying on simple and less accurate statistical approaches, we use SVM classifier which is considered to be one of the leading methods to perform nonlinear classification[32] and which was previously shown to predict OMBBs with favorable quality when compared with other classifiers.[28,29]

MATERIALS AND METHODS

Datasets

We utilize three benchmark datasets, two of which were commonly used in the recent evaluations of OMBB predictors,[17,19–26,28,29] and a new one that includes OMBB proteins that were characterized after the first two datasets were released. The two older datasets were proposed by Park *et al.*.[28] The first dataset (DS1) consist of 208 OMBBs, 673 globular proteins from various structural classes including 155 all-α, 156 all-β,184 α+β, and 178 α/β proteins, and 206 AMPs. Similarly as in the other studies and since we aim at the prediction of OMBBs, the globular proteins and the AMPs were merged. The similarity between any pair of the sequences from DS1 is below 40%. The second dataset (DS2) is derived from DS1 to include sequences that share pairwise similarity of no more than 25%. The DS2 consists of 112 OMBBs, 673 globular proteins, and 178 AMPs. We use the second dataset to design our predictor using fivefold cross validation tests, as this dataset is more challenging than the DS1. We also introduce the new dataset, named DS3, which was developed using a protocol similar to that in[28] with the 25% maximal similarity between any pair of the included sequences. First, we took well annotated OMBB and AMPs from the PSORT-B database[33] and globular proteins from the ASTRAL 1.75[34] that exclude entries from the membrane proteins SCOP class[35] and that were filtered at 25% sequence identity. Next, the combined set of OMBBs, AMPs, and globular proteins was filtered using blastclust[36] at the 25% maximal pairwise sequence identity; The resulting dataset includes 6988 globular proteins, 1168 AMPs, and 243 OMBBs. Because of the large and likely overrepresented (as a result of the difficulties in characterization of the structure of membrane proteins) number of globular proteins, we removed at random globular and AMP proteins to obtain the OMBB/AMP and OMBB/globular ratios that are similar to the ratios in the DS2 dataset. The final dataset includes 1460 globular proteins, 387 AMPs, and 243 OMBBs, which doubles the size of the DS2 dataset that is based on the same 25% similarity cut-off. The DS3 includes a substantial number of OMBBs that are dissimilar when compared with the chains in the DS1 and DS2 datasets; over half of the OMBB sequences from the DS3 share less than 30% and less than 50% identity with chains from the DS1 and DS2, respectively; see Supporting Information Figure 1.

Encoding of the protein sequences

Prior studies have demonstrated that evolutionary information carries more information than the sequence itself.[22,23,37] We use Position Specific Scoring Matrix (PSSM) computed with PSI-BLAST[38] to quantify the evolutionary information. PSSM gives conservation scores

for each AA at each position in the sequence and is represented by a $N \times 20$ dimensional matrix, where N is the length of the sequence. Conservation scores were normalized to the range between 0 and 1 using a standard logistic function, as in Ref. 39.

Many OMBBs have a well defined tertiary structure, where the transmembrane portion folds into a β-barrel. While prediction of tertiary structure from a sequence is still a challenging task,[40] our approach is to identify β-barrels using the predicted secondary structure that should include high content of long, collocated strand segments. We use PSIPRED[37] to predict secondary structure, since this method was ranked as one of the best in the field according to the EVA server (http://cubic.bioc.columbia.edu/eva/index.html). PSIPRED outputs probability of helix, strand and coil conformation for each input residue.

Hydrophobicity is another important feature of transmembrane proteins.[22] We group AAs into two sets, hydrophilic (D, E, G, H, K, N, Q, R, S, T) and hydrophobic (A, C, F, I, L, M, P, V, W, Y) and we also use a hydrophobicity index to estimate the hydrophobicity of the protein chain. We use the index by Jones[41] as it is characterized by the highest biserial correlation coefficient with the binary annotation of protein chains (OMBBs vs others) among different hydrophobicity indices from the AAIndex database.[42]

For each feature, we use either its raw value or we normalize it to assure that it's maximal and minimal values are in the [0; 1] range. We compare results obtained with the raw (without normalization) and the normalized features to investigate whether normalization leads to improved predictions. We first introduce the features and next we explain how they were normalized. The following 28 features are derived from the primary sequence:

- *seq_length (N)*—sequence length. (1 feature);
- *mol_weight*—molecular weight. (1 feature);
- *composition_AA$_i$*—number of AA$_i$ in a sequence,where $i = 1,2,\ldots,20$ is the AA type. (20 features);
- *content_{hydrophobic/hydrophilic}*—number of hydrophobic/hydrophilic residues in a sequence. (1 features);
- *max_len_{hydrophobic/hydrophilic}*—maximal length of a segment of residues in which each consecutive AA is hydrophobic/hydrophilic. (2 features);
- *avg_len_{hydrophobic/hydrophilic}*—average length of segments of residues in which each consecutive AA is hydrophobic/hydrophilic. (2 features);
- *hydrophobicity*—sum of the hydrophobicity index values for all residue in the sequence. (1 features).

Another 23 features are based on the conservation scores from the PSSM produced by the PSI-BLAST:

- *consScore_AA$_i$*—sum of conservation scores of AA$_i$ for each residue in a sequence. (20 features);
- *consScore_{hydrophobic/hydrophilic}*—sum of conservation scores of hydrophobic/hydrophilic AAs for each residue in a sequence. (2 features);
- *consScore_hydrophobicity*—hydrophobic index values multiplied by the corresponding conservation scores summed for all residues in a sequence. (1 feature).

The following 34 features are based on the predicted secondary structure, linear collocation of secondary structure segments, and some of them also use information concerning hydrophobicity:

- *content_SS$_j$*—number of residues predicted as SS$_j$, where j = {C (coil), H (helix), E (strand)}. (3 features);
- *max_SS$_j$*—maximal number of consecutive residues predicted as SS$_j$ (i.e., length of the longest SS$_j$ segment). (3 features);
- *avg_*SS$_j$—average number of consecutive residues predicted as SS$_j$ (i.e., average size of the SS$_j$ segments). (3 features);
- *probability_*SS$_j$—sum of probabilities of predictions of SS$_j$. (3 features);
- *{hydrophobic/hydrophilic}_SS$_j$*—number of hydrophobic/hydrophilic residues predicted as SS$_j$. (6 features);
- *hydrophobicity_SS$_j$*—sum of hydrophobicity index values of residues predicted as SS$_j$. (3 features);
- {*HH, EE, HE or EH*}—number of helix-coil-helix, strand-coil-strand, and helix(strand)-coil-strand(helix) motifs. (3 features);
- *Seg$_k$_num*—number Seg$_k$ segments in a protein, where k = {HCH, ECE, HCEorECH} represents segments of helices separated only by coils, segments of strands separated by coils, and segments of helices and strands separated by coils where a helix cannot be followed by helix and a strand cannot be followed by another strand. (3 features);
- *Seg$_k$_len*—number of residues in Seg$_k$ segments. (3 features);
- *E_ECE*—number of residues predicted as strands involved in the ECE segments. (1 feature);
- *H_HCH*—number of residues predicted as helices involved in the HCH segments. (1 feature);
- {*H/E*}*_HCE_or_ECH*—number of residues predicted as H/E involved in the HCE or ECH. (2 features).

Most of the features were normalized by dividing their values by N, except the *seq_length* and *mol_weight* which were divided by the corresponding maximal value in the training DS2 dataset (we set the values of these features to 1 in case of longer/heavier proteins in the DS1 or DS3 datasets), *consScore_{hydrophobic/hydrophilic}* that was divided by the product of N and the number of AA in each group (these features are named *avg_consScore_{hydrophobic/hydrophilic}*), *consScore_hydrophobicity* divided by the product of N and 20 (the number of AAs) (these

features are named *avg_consScore_hydrophobicity*), *HH, EE, HE, or EH* which were divided by the sum of the number of the helix and strand segments, and Seg_k_num that was divided by the total number of the predicted secondary structure segments.

The latter set of features that quantify the linear arrangement and size of predicted secondary structure segments as well as the features that are based on the aggregation of the scores from the PSSM matrix and which utilize the hydrophobic/hydrophilic segments are first proposed in this study, and they contribute to the improved predictive quality offered by the proposed here method.

Quality measures

We compute four quality measures including accuracy, Matthews Correlation Coefficient (MCC), sensitivity, and specificity, to evaluate the proposed predictor. Accuracy is defined as the number of correct predictions divided by the total number of test sequences. The MCC values range between −1 and 1, where 0 represents random correlation, and bigger positive (negative) values indicate better (lower) prediction quality. Sensitivity and specificity quantify the percentage of correctly predicted OMBBs and non-OMBBs, respectively, from among all OMBBs/non-OMBBs. We use the MCC to guide our design process, that is, we maximize MCC when performing feature selection and parameterization of the SVM, since this quality index provides better estimates for unbalanced datasets (which is the case in this project) when compared with the other three measures. We also report receiver-operator characteristics (ROC) curves that present a graphical plot of the True Positive (TP) rates (OMBBs predicted as OMBBs) = TP/(all OMBBs) against False Positive (FP) rates (non-OMBBs predicted as OMBBs) = FP/(all non-OMBBs). This is performed by thresholding the probabilities (confidence values) that are generated together with the predicted classes (OMBBs vs. nonOMBBs). These plots are also used to compute the area under the ROC curve (AUC). Higher AUC value indicates better predictive power of the corresponding method.

Proposed prediction model

We use SVM due to its successful use in the prior OMBBs predictors,[28,29] as well as Nearest Neighbour (NN) and RBF Network classifiers that were previously used in this area.[20–23,27] We utilized implementations from the WEKA workbench.[43] The SVM requires parameterization to select the kernel function, its parameters, and the complexity constant *C*. We considered three popular kernels, Radial Basis Function (RBF), polynomial kernel (POLY), and normalized polynomial kernel (NPOLY) and we tuned their gamma (for the RBF kernel) and exponent (for the polynomial kernels) parameters. We also considered each of the three SVMs with and without logistic regression-based estimates of the output probabilities. In case of the NN, we tuned the number of neighbors and the type of the distance function and distance weighting used, whereas for the RBF Network we choose the values of the ridge and minimal standard deviation parameters. We tested total of eight classifiers, including six SVM configurations, NN, and RBF Network. Our feature set includes 85 features and some of them may not be relevant to the prediction of OMBBs. Therefore, we performed a heuristic search to find a well-performing subset of these features and to optimize the classifiers' parameters for this set. First, we parametrized the eight classifiers using the entire feature vector. For each of the classifiers, we performed a grid search for the set of parameters' values that give the highest MCC when tested based on the fivefold cross validation on the DS2. Using the best parameters from the grid search we applied a wrapper-based feature selection with the corresponding classifier as the base method[44] and three different search algorithms including forward best first, backward best first, and ranking-based search. The forward (backward) best first search starts with empty (all features) subset of features and adds (removes) one feature at the time which increases the MCC value the most. In the ranking based feature selection, we first ranked features using an average rank based on the χ^2 and information gain criterions, and next starting with the empty feature set we kept adding the top ranked features if these additions would result in an increase of the MCC value. For each of the selected three feature sets, the corresponding classifiers were parameterized once again. The MCC, χ^2, and information gain values, which were used to tune the classifiers and in the feature selection, were computed based on the fivefold cross validation on DS2 to assure robust (not overtrained) estimates. The same test type was adopted by authors of other predictors in this area. We selected the classifier and its corresponding feature set that provided the highest MCC value based on the fivefold cross validation on the DS2. We performed the above feature selection for both raw (without normalization) and normalized feature sets. The results for the raw features, see the Supporting Information Table I, are in general worse than for the corresponding models computed with the normalized features, see the Supporting Information Table II. Therefore, we concentrate our further efforts on the evaluation and building of predictive models using normalized features.

Next, we investigate whether building of a consensus based predictors would lead to improved predictions. We tried two types of ensembles, a heterogeneous ensemble that combines different types of classifiers and a homogenous ensemble that combines multiple instances of classifiers of the same type. The consensus prediction was implemented based on a majority vote using predictions from individual input predictors; we combine

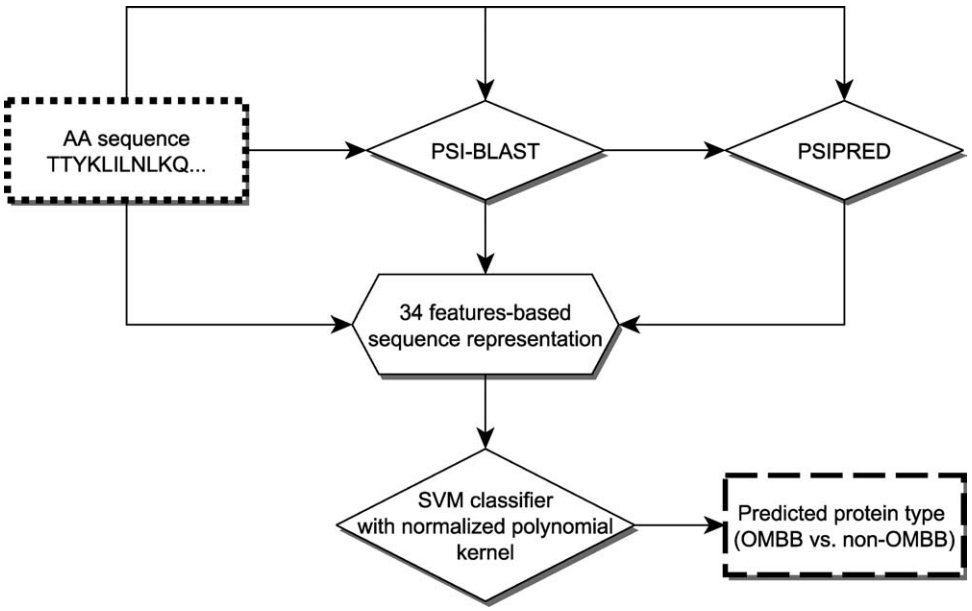

Figure 1

Flowchart of the proposed prediction method. The rectangles with the dotted and dashed border indicate the input and output, respectively.

an odd number of predictors to avoid ties. The heterogeneous ensemble was built by combining the three best-performing, based on the results in the Supporting Information Table I, parameterized predictors that utilize the NN, RBF Network, and SVM classifiers, respectively. The homogenous ensembles combine multiple SVM classifiers since Supporting Information Table I demonstrates that the SVMs provide favorable predictive performance when compared with the NNs and RBF Networks. We tried to combine the best 3, best 5, and best 7 SVM configurations (the bottom 3, 5, and 7 SVMs in the Supporting Information Table I). The results of the considered consensus predictors and the predictions using individual classifiers are compared in the Supporting Information Table I. The heterogeneous ensemble is outperformed by the homogenous ensembles, likely because the NN- and RBF Network-based classifiers provide lower predictive quality than the SVMs. The homogenous ensembles do not improve over the best performing individual SVM-based predictor. This could be explained by the very high quality of this SVM-based classifier and the fact that the third and lower-ranked best SVMs, which are utilized in the homogenous ensembles, introduce additional errors that were not balanced by the new correct predictions due to the voting. Consequently, we selected the best performing, with respect to MCC, SVM classifier to implement the proposed OMBBpred method, see Figure 1. The OMBBpred utilizes the SVM with the normalized polynomial kernel with exponent $= 2$ and $C = 2$ which uses 34 normalized features that are listed in Table I.

RESULTS AND DISCUSSION

The results for all datasets are based either on the five-fold cross validation or based on predictions generated using available web servers in case of methods that could not be re-implemented; the latter methods are marked with asterix in Tables II and III. The results on DS2 and DS3 are based on the same splits into the fivefolds, while the tests on DS1 may include different splits since these

Table I

List of the 34 Features Used by the Proposed Prediction Method

Id	Feature name	Biserial correlation	Id	Feature name	Biserial correlation
1	consScore_N	−0.51	18	composition_G	−0.23
2	consScore_S	−0.5	19	composition_Y	−0.22
3	content_E	−0.36	20	consScore_Q	−0.22
4	consScore_D	−0.35	21	composition_E	0.2
5	avg_len_hydrophobic	0.33	22	composition_C	0.19
6	content_H	0.32	23	consScore_F	−0.17
7	consScore_Y	−0.31	24	composition_H	0.14
8	probability_H	0.31	25	composition_D	−0.13
9	composition_N	−0.29	26	consScore_H	0.13
10	avg_len_hydrophilic	0.29	27	probability_C	−0.12
11	hydrophobic_E	−0.29	28	composition_L	0.1
12	hydrophobicity_E	−0.28	29	composition_K	0.09
13	composition_S	−0.26	30	consScore_K	0.09
14	consScore_A	−0.26	31	consScore_P	−0.08
15	composition_I	0.25	32	composition_A	−0.05
16	content_hydrophilic	−0.24	33	consScore_L	0.03
17	content_hydrophobic	0.24	34	composition_R	0.02

The features are sorted in the descending order based on the magnitude of their biserial correlation coefficients with the binary annotation of protein sequences (OMBBs vs. others).

Table II
Results of the Empirical Comparison Between the Proposed and Competing Methods on the DS1 and DS2

	Prediction method			Prediction quality			
Dataset	Classifier (name of the method, if available)	# Features	Reference	Accuracy	MCC	Sensitivity	Specificity
DS1	Neural Network	20	25	91.0	0.72	79.3	93.8
	Neural network	400	26	94.4	0.81	81.3	97.5
	Quadratic Discriminant	420	24	94.8	0.84	92.8	82.5
	SVM	420	28	95.2	0.84	79.3	99.0
	Nearest Neighbor	20	22	96.1	0.87	91.5	98.2
	Nearest Neighbor (WED with homology)	144	23	97.4	0.89	91.1	98.4
	SVM	546	29	97.8	0.93	91.8	99.2
	SVM (OMBBpred)	34	This paper	**98.1**	**0.94**	**91.3**	**99.7**
DS2	Probabilistic (PROFtmb)*	N/A	19	96.4	0.58	37.0	**100**
	Nearest Neighbor	20	22	95.6	0.77	74.3	98.4
	Probabilistic and Nearest Neighbor (BOMP)*	N/A	16	92.3	0.82	79.8	98.5
	Nearest Neighbor (TMB-Hunt)*	20	20	96.4	0.83	81.5	98.5
	Nearest Neighbor (WED with homology)	144	23	96.8	0.86	**90.7**	97.6
	SVM (OMBBpred)	34	This paper	**98.5**	**0.93**	88.2	99.9

The results were taken from the original publications and they were obtained using fivefold cross validation, except for the methods marked with asterix which were obtained using available web servers. The methods are sorted in the ascending order by their MCC values. The best results for each prediction quality measure and each dataset are shown in bold.

folds were not annotated by the authors of this dataset. Table II compares the quality of prediction generated by the OMBBpred with the predictions of several modern predictors on the benchmark datasets DS1 and DS2, and Table III report the results on the new DS3. In Table II, we report solutions with the highest MCC values for the prior studies that investigated several different OMBB predictors. The OMBBpred achieves above 98% accuracy and 0.93 MCC on both older benchmark datasets. The sensitivity of the proposed method shows that it correctly predicts at least 88% of the OMBBs, while the specificity reveals that only 0.3% or less of the non-OMBBs are mistaken for OMBBs. This means that OMBBpred is geared towards predictions with very low false positive rates. The OMBBpred achieves 96% accuracy, 0.78 MCC, and 99.1% specificity on the new dataset. The lower predictive performance on the DS3 when compared with DS2 that also includes chains with the maximal identity set at 25% is characteristic to all top-performing methods. The MCC, which provides better estimates for unbalanced datasets like DS2 and DS3 when compared with other considered measures, drops from 0.86 to 0.62 for the homology-based WED,[23] from 0.83 to 0.7 for the TMB-Hunt,[20] from 0.82 to 0.71 for the BOMP,[16] and from 0.93 to 0.78 for the proposed OMBBpred. This indicates that the new dataset, which includes recently annotated and larger number of chains, is more challenging. Importantly, our method is shown to consistently outperform the existing solutions; we obtain the highest accuracies and MCC values on all benchmark datasets. The only method that has higher specificity (but only by 0.1%) on the DS2 is PROFtmb,[17] but this result comes as a trade-off for the relatively low sensitivity and MCC values that equal 37% and 0.58, respectively. Similarly, there is only one method, WED,[23] with higher sensitivity on the DS2 but it generates predictions with lower accuracy, MCC and specificity when compared with the OMBBpred. When compared with the second best predictor by Gao *et al.*[29] on the DS1, OMBBpred obtains (98.1–97.8)/(100–97.8) = 14% error rate reduction. However, the error rate reductions for datasets with stricter similarity threshold are 65% for the DS2 when

Table III
Results of the Empirical Comparison Between the Proposed and the Competing Methods on the DS3

Prediction method			Prediction quality				
Classifier (name of the method, if available)	# Features	Reference	Accuracy	MCC	Sensitivity	Specificity	AUC
Nearest Neighbor (WED without homology)	144	23	90.8	0.49	44.4	96.9	N/A
SVM	546	29	92.4	0.60	55.6	97.3	0.930
Nearest Neighbor (WED with homology)	144	23	92.9	0.62	56.8	97.6	N/A
Nearest Neighbor (TMB-Hunt)*	20	20	94.4	0.70	62.1	98.7	0.917
Probabilistic and Nearest Neighbor (BOMP)*	N/A	16	94.4	0.71	63.4	98.5	0.811
SVM (OMBBpred)	34	This paper	**95.8**	**0.78**	**70.4**	**99.1**	**0.961**

The results were obtained using fivefold cross validation using the same folds for all methods, except for the methods marked with asterix which were obtained using available web servers. We used a standalone version of the WED method that was provided by the authors. The methods are sorted in the ascending order by their MCC values. The best results for each prediction quality measures are shown in bold.

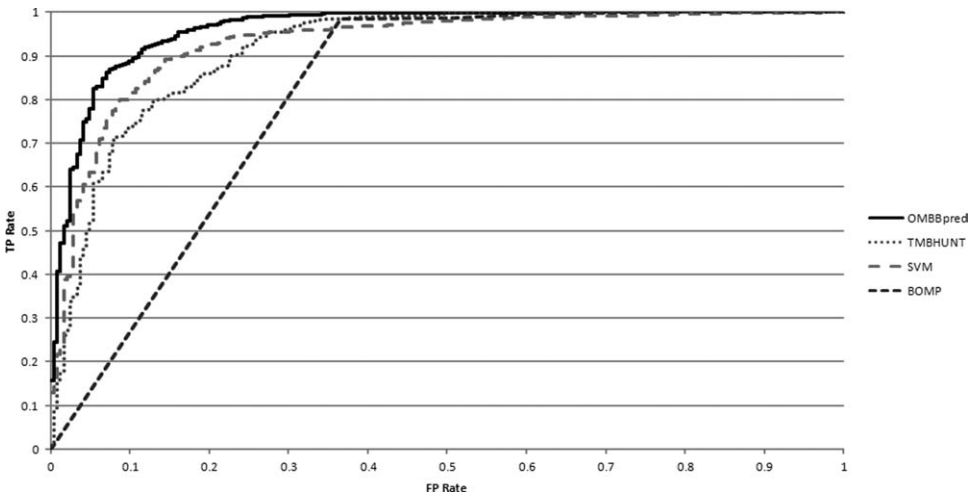

Figure 2
The ROC curves for four predictors, including SVM,[29] TMB-Hunt,[20] BOMP,[16] and OMBBpred for the predictions on the new DS3.

compared with the second best method by Yan *et al.*[23] and 25% for the DS3 when compared with the runner-up BOMP and TMB-Hunt. Both BOMP and TMB-Hunt are web servers which are based on homology detection and they may include some of the OMBBs from the DS3 in their template database. The OMBBpred, which outperformed these solutions on the DS3 for all quality measures, was evaluated using training datasets with chains that share no more than 25% similarity with the test sequences. When compared with the best performing SVM method,[29] which does not utilize homology modeling and that was tested using the same fivefold cross validation on the DS3, the OMBBpred provides 45% error rate reducution. We also report and compare the AUC values, see Table III, and the corresponding ROC curves, see Figure 2, for the predictions on the DS3. Similarly as for the other measures, the proposed predictor outperforms other solutions and these improvements are consistent for the entire range of the False Positive rates. The main reason for these improvements is the fact that the proposed method utilizes carefully crafted input features that combine information from the predicted secondary structure, AA composition, hydrophobicity of residues, and their evolutionary conservation.

We investigated SCOP classes of the false positives predictions (AMP and globular proteins predicted as OMBBs). Since OMBB is characterized by high specificity it rarely misclassifies the non-OMBBs. When considering prediction on the three datasets together, only 14 AMPs, 4 all-β, 1 α/β, and 3 α+β proteins were misclassified as OMBBs by our method. Additionally, we checked whether OMBBpred can successfully discriminate between the OMBBs and the all-β proteins that also have high content of β-strands. We filtered the biggest dataset, DS3, leaving only OMBB and all-β globular proteins and we repeated the 5-fold cross validation test on the resulting dataset using the same architecture of the OMBBpred method (the same features and the parameters of the SVM classifier). These predictions are characterized by high 0.91 MCC and 95.8% accuracy and they demonstrate that OMBBpred can be used to distinguish between the all-β and the OMBB proteins.

Another important aspect is the impact of the sequence identity between the chains used to compute evolutionary information that is inputted to the OMBBpred and the chains in the benchmark datasets. To investigate this, we filtered the nr dataset, which is utilized to generate the PSSM profiles and the predicted secondary structure that uses these profiles as inputs, to exclude any sequence that shares over 25% identity with any chain in the training dataset DS2. We regenerated the profiles and re-predicted the secondary structure utilizing the filtered nr database, which includes 1,324,307 chains, and we used these new inputs to generate the 34 features for the OMBBpred. We evaluated this version of our predictor using fivefold cross validation on the DS2 and we observe that the accuracy and the MCC drop by 2.2% and 0.12, respectively; see Supporting Information Table III. This shows that OMBBpred that uses the filtered nr dataset provides competitive predictions when compared with the top-performing runner-up methods that utilize homology modeling like WED, TMB-Hunt and BOMP, see Table II. To compare, a similar drop in accuracy is observed for WED method on the DS3, see Table III, where addition of the information concerning homologous (i.e., similar) sequences improves the accuracy by 2.1% and the MCC by 0.13.

Until recently the OMPs were believed to be OMBBs. A study in 2006[5] and subsequent works including[6] show examples of OMPs that implement the membrane spanning region as the alpha barrel rather than the beta bar-

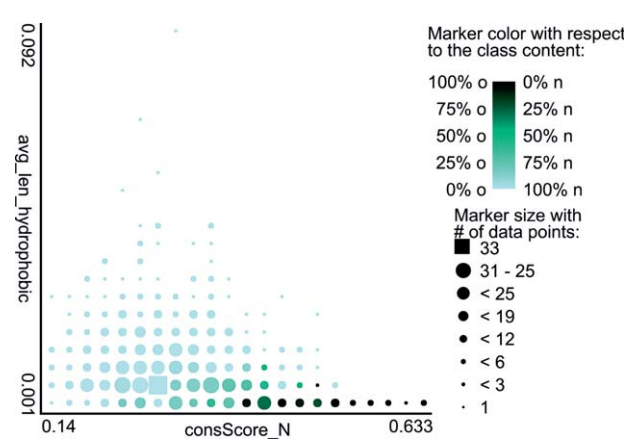

Figure 3

Scatter plot of two representative features (novel features with the highest positive and negative biserial correlations with the OMBB/non-OMBB annotations) used by the proposed predictor. The *x*-axis shows the values of the *consScore_N* (overall sequence conservation of Asp in the input sequences) feature and the *y*-axis the values of the *avg_len_hydrophobic* (average length of hydrophobic segments) features for the proteins from the DS2. The color (shading) of the markers indicates the probability of a given prediction outcome (darker for proteins that are more likely to be OMBBs) for a given combination of the values of the two features. The size of the markers denotes the number of proteins with the given values of the two features.

rel. Although OMBBpred and other related predictors were designed and tested using the OMBB-type OMPs, we investigate how they would predict the two known α-barrel OMPs.[5,6] We found that OMBBpred and TMB-Hunt predict them as OMBBs, BOMP and WED predict only the WZA protein[5] as OMBB, whereas the SVM classifier[29] predicts both of these proteins as not OMBBs. The likely reason why OMBBpred, TMB-Hunt, BOMP, and WED predict the WZA protein as the beta barrel fold is the fact that it includes β-sheets.

Finally, we analyze features that are utilized in the OMBBpred to provide insights into sequence-derived factors that differentiate OMBB proteins from the AMPs and globular proteins. We investigate the two representative novel features which quantify information concerning the evolutionary profiles and hydrophobicity. The *consScore_N* feature is characterized by the highest negative value of the biserial correlation coefficient with the annotation of protein chains (OMBBs vs others) among the 34 features used by the proposed predictor, and the value of *avg_len_hydrophobic* obtains the highest positive correlation, see Table I. The former feature quantifies the overall sequence conservation (computed using PSSMs) of Asp in the input sequence and the latter feature computes the average length of hydrophobic segments. A scatter plot shown in Figure 3 visualizes the predictive value of these two features. The size and color of the markers indicate the number of chains and the probability of these chains to be OMBBs, respectively. The plot reveals that OMBBs are characterized by shorter average hydrophobic segments and have higher average conservation of Asp when compared with the non-OMBBs proteins. The shorter average hydrophobic segments, as well as shorter average hydrophilic segments (the *avg_len_hydrophilic* feature also has positive biserial correlation, see Table I), could be explained by the fact that the beta barrel residues are hydrophobic on the exterior (where they contact membrane lipids) and hydrophilic in the interior, which results in a chain consisting of multiple short hydrophobic/hydrophilic segments. The content of Asp was previously shown to be similar in the OMBB and globular proteins, and substantially higher in the OMBBs when compared with the AMPs.[28] However, our work shows that Asp is more conserved in OMBBs than in the other protein types. In DS2, the average composition of Asp is 3.1% (±1.2%) for AMPs, 4.4% (±2.2%) for globular proteins, and 6% (±1.8%) for OMBBs, whereas the corresponding average conservation scores equal 0.239 (±.057), 0.309 (±.063), and 0.421 (±.078), respectively, see Supporting Information Figure 2. This is also confirmed by the lower magnitude of the biserial correlation of *composition_N* (−0.29) than *consScore_N* (−0.51), see Table I, which means that the composition of Asp has lower discriminative power than the average conservation score of Asp. In our feature set, 6 of the 34 features quantify different aspect of hydrophobicity which confirms that the knowledge of this factor is useful for the discrimination of membrane proteins. Combining the hydrophobicity-derived features with six features which describe the predicted secondary structure helps to discriminate between the OMBBs and the other membrane proteins, such as the α-helical membrane proteins. We also utilize 13 features that quantify AA composition and 11 which describe average AA conservation score, which agrees with previous results that have shown that AA composition is useful for an accurate identification of the OMBBs.[16,18,21–23,27–29] However, analysis of the biserial correlations shown in Table I reveals that the conservation score-based features, which were introduced in this work, are more informative (they have higher magnitutes of the correlations) than the features based on the AA content.

CONCLUSIONS

The novel characteristics of the proposed OMBBpred method, which include effective fusion of multiple information sources including the predicted secondary structure, hydrophobicity, and evolutionary conservation and the careful design that combines feature design and selection, result in its favorable predictive quality when compared with modern solutions on three benchmark datasets that include sequences with low identity. Our predictions are characterized by high specificity, which

reveals that around 90% of OMBBs in the older datasets and above 70% in the new more challenging dataset were correctly identified, and very high specificity that demonstrates that only about 0.9% or less of non-OMBBs are incorrectly predicted as OMBBs. Moreover, OMBBpred achieved the highest accuracy and MCC scores on the three datasets when compared with several modern predictors of the OMBB proteins. The high predictive quality of our method motivates its use for the high-throughput identification of OMBBs on the genomic scale. Our methods could be also used to generate predicitons for methods, such as TMBpro,[45] which predict structural properties of the OMBBs. The prediction model and the datasets can be found at http://biomine.ece.ualberta.ca/OMBBpred/OMBBpred.htm.

ACKNOWLEDGMENTS

The authors thank Drs. Michael Gromiha and Changhui Yan for providing the datasets.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
2. Carpenter EP, Beis K, Cameron AD, Iwata S. Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 2008;18:581–586.
3. Jones D. Do transmembrane protein superfolds exist? FEBS Lett 1998;423:281–285.
4. Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the protein data bank: identification and classification. Bioinformatics 2004;20:2964–2972.
5. Dong C, Beis K, Nesper J, Brunkan-Lamontagne AL, Clarke BR, Whitfield C, Naismith JH. Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. Nature 2006;444:226–229.
6. Chandran V, Fronzes R, Duquerroy S, Cronin N, Navaza J, Waksman G. Structure of the outer membrane complex of a type IV secretion system. Nature 2009;462:1011–1015.
7. Schulz GE. Beta-barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.
8. Wimley WC. The versatile beta-barrel membrane protein. Curr Opin Struct Biol 2003;13:404–411.
9. Koebnik R, Locher KP, Van Gelder P. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol Microbiol 2000;37:239–253.
10. Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ. A hidden Markov model method, capable of predicting and discriminating ß-barrel outer membrane proteins. BMC Bioinformatics 2004;29.
11. Martelli PL, Fariselli P, Krogh A, Casadio R. A sequence-profile-based HMM for predicting and discriminating b-barrel membrane proteins. Bioinformatics 2002;18:S46–S53.
12. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane b-barrels in proteomes. Nucleic Acids Res 2004;32:2566–2577.
13. Deng Y, Liu Q, Li YX. Scoring hidden Markov models to discriminate b-barrel membrane proteins. Comput Biol Chem 2004;28:189–194.
14. Zhai Y, Saier MH, Jr. The b-barrel finder (BBF) program, allowing identification of outer membrane b-barrel proteins encoded within prokaryotic genomes. Protein Sci 2002;11:2196–2207.
15. Liu Q, Zhu Y, Wang B, Li Y. Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. Comp Biol Chem 2003;27:355–361.
16. Berven FS, Flikka K, Jensen HB, Eidhammer I. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 2004;32:W394–W399.
17. Gromiha MM, Ahmad S, Suwa M. Application of residue distribution along the sequence for discriminating outer membrane proteins. Comput Biol Chem 2005;29:135–142.
18. Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 2005;21:961–968.
19. Bigelow HR, Rost B. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res 2006;34:W186–W188.
20. Garrow AG, Agnew A, Westhead DR. TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 2005;33:W188–W192.
21. Garrow AG, Agnew A, Westhead DR. TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. BMC Bioinformatics 2005;56.
22. Yan C, Hu J, Wang Y. Discrimination of outer membrane proteins using a K-nearest neighbor method. Amino Acids 2008; 35:65–73.
23. Yan C, Hu J, Wang Y. Discrimination of outer membrane proteins with improved performance. BMC Bioinformatics 2008;47.
24. Lin H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 2008;252:350–356.
25. Gromiha MM, Suwa M. Discrimination of outer membrane proteins using machine learning algorithms. Proteins 2006;63:1031–1037.
26. Gromiha MM, Suwa M. Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. Biochim Biophys Acta 2006;1764:1493–1497.
27. Ou YY, Gromiha MM, Chen SA, Suwa M. TMBETADISC-RBF: discrimination of b-barrel membrane proteins using RBF networks and PSSM profiles. Comput Biol Chem 2008;32:227–231.
28. Park KJ, Gromiha MM, Horton P, Suwa M. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 2005;21:4223–4229.
29. Gao QB, Ye XF, Jin ZC, He J. Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition. Anal Biochem 2010;398:52–59.
30. Schulz GE. Transmembrane beta-barrel proteins. Adv Protein Chem 2003;63:47–70.
31. Mizianty MJ, Kurgan L. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics 2009;414.
32. Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14:1–37.
33. Gardy JL, Spencer C, Wang K, Ester M, Tusnády GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 2003;31:3613–3617.
34. Brenner S, Koehl P, Levitt M. The ASTRAL cendium for sequence and structure analysis. Nucleic Acids Res 2000;28:254–256.
35. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36: D419–D425.
36. BLASTCLUST, the NCBI software development toolkit. Available at: ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html Accessed on August 9, 2009.

37. Bryson K, McGuffin L, Marsden R, Ward J, Sodhi J, Jones D. Protein structure prediction servers at University College London. Nucleic Acids Res 2005;36:36–38.
38. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
39. Campbell K, Kurgan L. Sequence-only based prediction of b-turn location and type using collocation of amino acid pairs. Open Bioinf J 2008;2:37–49.
40. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction—round VIII. Proteins 2009;77:1–4.
41. Jones D. Amino acid properties and side-chain orientation in proteins: A cross correlation approach. J Theor Biol 1975;50:167–183.
42. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–D205.
43. Hall M, Frank F, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations 2009;11:10–18.
44. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.
45. Randall A, Cheng J, Sweredoski M, Baldi P. TMBpro: secondary structure, beta-contact, and tertiary structure prediction of transmembrane beta-barrel proteins. Bioinformatics 2008;24:513–520.