SPiD: An SVM-Based Protein Discriminator for Outer Membrane Proteins

Babak Alipanahi*

David R. Cheriton School of Computer Science, University of Waterloo
200 University Ave W., Waterloo, ON N2L 3G1, Canada

ABSTRACT

Discrimination of Outer Membrane Proteins (OMP) from other types of membrane and globular proteins is an important step in their secondary and tertiary structure prediction. Moreover, a reliable discrimination method can be used for whole genome analysis and hence discovery of new OMPs. In this paper, we propose an SVM-based protein discriminator for OMPs (SPiD) from other types of proteins, i.e., globular and inner membrane proteins. This approach uses amino acid and amino acid pair composition values, the length of protein sequence, and a newly defined feature called β-barrel score. When applied to a dataset consisting of 1,087 proteins, SPiD achieves an overall accuracy of 96%; to the best of authors' knowledge, this is higher than the accuracy of other previous studies. When SPiD is trained to pick up only outer membrane β-barrels, it reaches an overall accuracy of 99%.

Keywords: SVM, protein classification, membrane proteins

1 INTRODUCTION

The most remarkable fact about Gram-negative bacteria is their cell envelopes. It consists of two layers: Inner Membrane (IM) and Outer Membrane (OM), which are separated by periplasm. IM is in direct contact with cytoplasm and periplasm while OM is in contact with extracellular environment and periplasm [1]. Integral IM proteins span the membrane by α-helices while integral OM proteins span the membrane by β-strands and form a β-barrel. OM Proteins (OMPs) have diverse functions and are divided into several families: selective active and passive transporters of molecules, enzymes, defense proteins, structural proteins, and toxins.

Several methods for discrimination of OMPs have been proposed in the recent years. These methods can be divided into three major groups: methods using sequence alignment information and/or HMM [2, 3, 4, 5, 6], methods based on amino acid composition values [7, 8, 9], and methods using amino acid sequence properties like estimated folding pseudo-energy or average hydrophobicity [10, 11, 12]. In this paper, we propose an SVM-based protein discriminator for OMPs (SPiD), using amino acid and amino acid pair composition values, sequence length, and a feature especially tailored for β-barrels.

2 Materials and methods

2.1 Datasets

The dataset used here is the same as in [9], since it is one of the most challenging and comprehensive available datasets. Moreover, the authors report the best results to date and it facilitates the fair comparison of our results with those of previous studies. This dataset primarily consists of 377 OMPs and 268 α-helical membrane proteins extracted from PSORT-B database [13], and 674 globular proteins from the PDB40D_1.37 database of SCOP [14, 15]. It is filtered for sequence identity of less than 40% using CD-HIT algorithm [16]. The resulting dataset has 208 OMPs, 206 α-helical membrane proteins, and 673 globular proteins consisting of all-α, all-β, $\alpha+\beta$, and α/β proteins. Herein after, we will define the following notation for the dataset used for discrimination: the set of outer membrane proteins (OMP), the set of α-helical membrane proteins (TMH), the set of globular proteins (GLB), and the set of non-OMPs (NOM).

In order to verify the capability of the proposed approach in discrimination of transmembrane β-barrels from α-helical membrane and other non-β-barrel proteins, we also used `TMPDB_alpha_non_redundant` and `TMPDB_beta_non_redundant` datasets [17]. These datasets consist of proteins with experimentally known structures that are filtered

*Corresponding author e-mail: balipana@cs.uwaterloo.ca

for sequence similarity of less than 30%, using CLUSTALW version 1.81 [18]. `TMPDB_alpha_non_redundant` contains 231 and `TMPDB_beta_non_redundant` contains 15 proteins.

2.2 Features

In a protein sequence of length N, for amino acids a and b, the peptide (amino acid) and dipeptide (amino acid pair) composition values are defined by

$$C_a = \frac{n_a}{N}, \quad D_{ab} = \frac{p_{ab}}{N-1}, \tag{1}$$

where n_a and p_{ab} are the number of occurrences of amino acid a and amino acid pair ab in the sequence, respectively. OMPs usually have longer amino acid sequences than some of globular proteins, since for example forming a β-barrel structure requires a minimum number of amino acids, so it was added too. We denote peptide composition value features by C, dipeptide features by D, and sequence length feature by L. There are 20 "C" and 400 "D" features.

We define a new feature called β-barrel score whose original idea is taken from [4]. Every amino acid in the membrane spanning section of the protein sequence can be either Lipid Exposed (LE) or barrel Interior Exposed (IE). The β-strand score for every position i, B_i, is defined as [4]:

$$B_i^1 = \sum_{j\in\mathcal{E}} L(a_{i+j}|a_{i+j} \in \text{IE}) + \sum_{j\in O} L(a_{i+j}|a_{i+j} \in \text{LE}), \tag{2}$$

$$B_i^2 = \sum_{j\in\mathcal{E}} L(a_{i+j}|a_{i+j} \in \text{LE}) + \sum_{j\in O} L(a_{i+j}|a_{i+j} \in \text{IE}), \tag{3}$$

and $B_i = \max(B_i^1, B_i^2)$; where in equations (2) and (3), $L(a_{i+j} \mid a_{i+j} \in \text{IE}) = \log(Pr\{a_{i+j}|a_{i+j} \in \text{IE}\})$, which is estimated from the real data; the same is true for the LE case. Moreover, the set of even and odd shifts are defined as $\mathcal{E} = \{0,2,4,6,8\}$ and $O = \{1,3,5,7,9\}$. When summing $L(a_{i+j}|a_{i+j})$ values, we actually multiply the corresponding probabilities. If we assume independence between consequent residues, B_i^k, $k = 1,2$ values are the logarithm values of the probability that a β-strand starts at residue i with different assumptions, whether it is IE or LE. It turns out that a window size of ten is optimum. After calculating B_i for all residues 1 to $N-9$; we form the new feature called β-barrel score (B) which is defined as $B = \frac{1}{N-9}\sum_{i=1}^{N-9} B_i^2$.

In Table 1, mean values of 20 peptide composition values, sequence length, and β barrel score features are listed for GLB, TMH, NOM, and OMP datasets. To perform feature selection, we used backward elimination-forward selection (BE-FS) method. Since running backward elimination on 400 dipeptide features is not feasible, we only ran forward selection on them.

3 Results and Discussion

3.1 Implementation

We have used the LIBSVM software package[1], which is both very fast and reliable. In each scenario, we repeated each experiment for 100 times and each time randomly changed the permutation of proteins. The standard deviation of the calculated results is approximately 0.1%. In order to optimize the RBF kernel parameters, we have used grid optimization technique in two coarse and fine steps. It turned out that optimal values for penalty parameter and kernel width were 10 and 2, respectively.

3.2 Performance evaluation

In order to participate all data points in the performance evaluation process, we use 5-fold cross validation. In order to compare SPiD's results with previous studies, we use measures that have been widely used before. Suppose TP, FP, TN, and FN denote true positive, false positive, true negative and false negative assignments, respectively. By *positive* we mean a correctly classified protein, and by *negative* we mean an incorrectly classified protein. We use the following widely-used performance measures:

$$\text{SEN} = \frac{\text{TP}}{\text{TP+FN}}, \quad \text{SPC} = \frac{\text{TN}}{\text{TN+FP}}, \quad \text{ACC} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}, \tag{4}$$

where SEN, SPC and ACC stand for sensitivity (ability to discover OMPs, a small sensitivity value indicates that many OMPs will not be discovered), specificity (ability to correctly sift OMPs from non-OMPs), and overall accuracy which is an indicator of overall performance. Moreover, we use the Matthew's correlation coefficient (MCC) which is a better performance measure defined as follows [19]:

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{5}$$

MCC value is zero for a completely random assignment and one for a perfect discrimination.

[1] Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Table 1. Mean values of features

Feature	GLB	TMH	NOM	OMP
L	183	424	240	552
B	0.82	1.06	0.88	1.32
Ala	8.42	10.27	8.86	9.37
Arg	5.05	4.45	4.91	5.23
Asn	4.40	3.06	4.09	5.44
Asp	5.84	3.32	5.25	5.88
Cys	1.47	0.85	1.32	0.41
Gln	3.94	3.21	3.77	4.71
Glu	6.72	3.74	6.02	4.86
Gly	7.65	8.33	7.81	8.69
His	2.20	1.68	2.08	1.25
Ile	5.77	7.50	6.17	4.72
Leu	8.50	12.72	9.49	8.94
Lys	6.17	3.38	5.52	4.89
Met	2.19	3.58	2.52	1.66
Phe	3.77	5.49	4.17	3.75
Pro	4.47	4.29	4.43	3.74
Ser	5.77	5.88	5.79	8.04
Thr	5.71	5.20	5.59	6.31
Trp	1.34	2.05	1.51	1.24
Tyr	3.42	2.82	3.28	4.13
Val	7.20	8.19	7.43	6.75

All composition values are in percentile.

Table 2. OMP-GLB discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	89.3	98.4	96.1	0.896
L+9C+B	89.0	98.9	96.5	0.904
L+9C+2D+B	**89.9**	**98.8**	**96.6**	**0.908**

Table 3. OMP-TMH discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	95.9	93.2	94.6	0.892
L+10C	96.3	94.7	95.5	0.910
L+10C+3D	**98.2**	**96.3**	**97.3**	**0.945**

Table 4. OMP-NOM discrimination results

Features	SEN	SPC	ACC	MCC
L+20C+B	86.2	97.7	95.4	0.855
L+13C+B	85.0	98.3	95.6	0.861
L+13C+3D+B	**87.0**	**98.2**	**96.0**	**0.872**

SEN, SPC, and ACC are in percentile.
The best results are shown in bold.
First line, shows the performance measures without dipeptide features; second line after backward elimination; and the third line after forward selection ran only on dipeptide composition features.

3.3 Discrimination of OMPs

We experimented discrimination of OMPs in three scenarios: from globular proteins (OMP-GLB), from α-helical membrane proteins (OMP-TMH), and from non-OMPs (OMP-NOM).

3.3.1 OMP-GLB and OMP-TMH discrimination

In Table 2, performance results of OMP-GLB discrimination are listed. Initial ACC and MCC values are better than previous studies' results; while backward elimination and forward selection even more improved these results. After backward elimination, remaining amino acids were aromatic (Trp and Tyr) and polar residues (Cys, Gln, Pro, His and Thr). The added dipeptide composition features were Asp-Phe and Tyr-Asn, both a combination of a polar and an aromatic residue, which are abundant in OMPs.

In OMP-TMH scenario (results listed in Table 3), backward elimination improved ACC and MCC by 0.9% and 0.018, respectively; and removed β-barrel score and half of the peptide composition values. The remaining residues were good α-helix formers (Ala and Met), bad α-helix formers (Gly, Pro, Ser, Tyr), and Asp, His, Ile and Lys. Forward selection added Gln-Ala (polar-aliphatic), Asp-Ala (charged-aliphatic), and Glu-Phe (charged-aromatic), and boosted ACC and MCC values by 1.8% and 0.035, respectively.

3.3.2 OMP-NOM discrimination

OMP-NOM discrimination is more important than other scenarios, therefore, we elaborate more on it. Discrimination accuracy using all 20 amino acid composition values, length of protein sequence, and β-barrel score was quite acceptable and better than all previous studies. Backward elimination, did not improve the performance so much but reduced the number of features from 22 to 15; most of the omitted features had close mean values. Forward selection, added 3 dipeptide features: Asp-Ala (charged-aliphatic), Glu-Phe (charged-aromatic) and Asn-Lys (polar-charged); These added features boosted ACC and MCC values from 95.6% and 0.861 to 96% and 0.872, respectively. It is important that sensitivity was improved from 85% to 87%. Detailed performance measures are listed in Table 4. The relation between MCC and SEN values for different values of SPC for OMP-NOM discrimination is depicted in Figure 1. It can be seen that MCC value nearly grows linearly with the growth of SEN value.

In order to better analyze the performance, for each protein, the probability of being classified incorrectly is estimated by running the discrimination experiment 500 times and counting the number of times that any protein is classified incorrectly, whenever it is in the validation dataset. The estimated probabilities are depicted in Figure

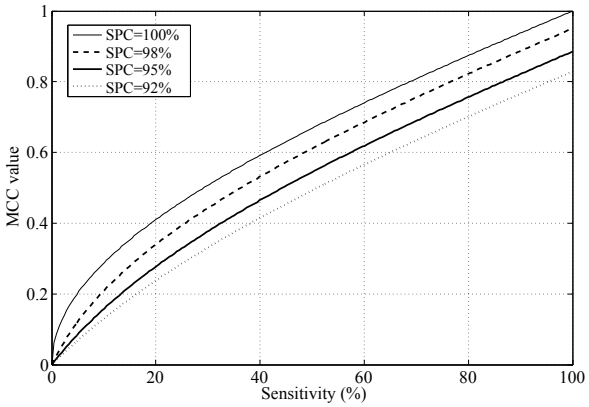

Figure 1. MCC values vs. sensitivity for different values of SPC for OMP-NOM discrimination.

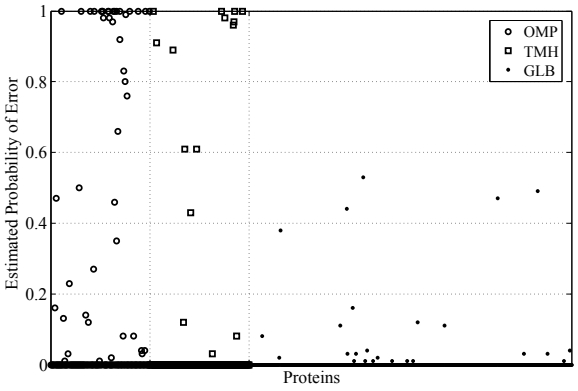

Figure 2. Estimated probability of error for all proteins.

2. It is interesting that some proteins (mostly OMPs) are always misclassified, i.e., their estimated probability of error is one and some proteins (mostly globular proteins) are never misclassified. The detailed results of estimated misclassification error probabilities are listed in Table 7.

3.4 Discrimination of β-barrels

In a more specific discrimination scenario, we apply our approach to discrimination of transmembrane β-barrels from other types of proteins. To do so, we have used the `TMPDB_alpha_non_redundant`, `TMPDB_beta_non_redundant` and GLB datasets. It is interesting that the mean β-barrel score feature of β-barrels (1.79) is nearly twice as other datasets (0.81 for GLB, 0.98 for TMH, and 0.85 for α-helical membrane and globular proteins dataset), and has a very large mean difference in all scenarios. It is mainly because this feature is especially tailored for β-barrels. We use the same set of features that are found after performing backward elimination-forward selection in OMP-NOM discrimination. Performance results are listed in Table 6. In all cases discrimination accuracy is very high.

Table 5. Comparison with other studies.

Method	Scenario	SEN	SPC	ACC
SPiD	OMP-GLB	89.9	98.8	96.6
(SVM)	OMP-TMH	98.2	96.3	97.3
	OMP-NOM	87.0	98.2	96.0
[7]	OMP-GLB	85.5	92.5	92.1
(Linear Classifier)				
[9]	OMP-GLB	88.0	90.4	94.4
(SVM)	OMP-TMH	99.0	92.7	95.9
	OMP-NOM	90.9	94.7	93.9
[12]	OMP-GLB	83.7	97.6	94.3
(Neural Network)	OMP-TMH	91.8	91.7	91.8
	OMP-NOM	81.3	97.5	94.4

Table 6. Results of β-barrel discrimination.

Scenario	SEN	SPC	ACC	MCC
BB-GLB	75.3	99.6	98.9	0.782
BB-AA	95.8	100	99.6	0.974
BB-NBB	80.0	99.8	99.3	0.829

BB: `TMPDB_beta_non_redundant`
AA: `TMPDB_alpha_non_redundant`
NBB: AA + GLB

Table 7. Estimated probability of error analysis.

Dataset	size	$P_e = 0$	$0 < P_e < 1$	$P_e = 1$
GLB	673	0.96	0.04	0.00
TMH	206	0.93	0.05	0.02
OMP	208	0.78	0.13	0.09
Total	1087	0.92	0.06	0.02

4 Discussion and Conclusion

In this study, we poposed SPiD, a new SVM-based approach for discrimination of OMPs and in a more specific case, transmembrane β-barrels. By adding two features to the peptide and dipeptide composition values and performing feature selection, SPiD was proved to be very accurate. Park *et al.* proposed a method based on amino acid and amino acid pair composition values and reported an accuracy of 93.9% in a set of 208 OMPs that was calculated using 5-fold cross validation [9]. Gromiha and Suwa developed a method based on amino acid properties and reported a prediction rate of 94.4% [12]. In comparison to the aforementioned studies, SPiD achieved OMP-GLB, and OMP-TMH, OMP-NOM discrimination accuracies of 96.6%, 97.3%, and 96.0%, respectively. When trained to pick only β-barrels, SPiD was able to reject 913 out of 919 non-β-barrels and cover 12 out of 15 β-barrel families. Moreover,

by performing error probability analysis, it turned out that some proteins were always misclassified because they were more similar to non-OMPs.

References

[1] N. Ruiz, D. Kahne, and T. J. Silhavy, "Advances in understanding bacterial outer-membrane biogenesis.," *Nature reviews. Microbiology* **4**, pp. 57–66, January 2006.

[2] T. V. Gnanasekaran, S. Peri, A. Arockiasamy, and S. Krishnaswamy, "Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins," *Bioinformatics* **16**, pp. 839–842, September 2000.

[3] P. L. L. Martelli, P. Fariselli, A. Krogh, and R. Casadio, "A sequence-profile-based hmm for predicting and discriminating beta barrel membrane proteins.," *Bioinformatics (Oxford, England)* **18 Suppl 1**, 2002.

[4] W. C. Wimley, "Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures.," *Protein science : a publication of the Protein Society* **11**, pp. 301–312, February 2002.

[5] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, "A hidden markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins.," *BMC bioinformatics* **5**, March 2004.

[6] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost, "Predicting transmembrane beta-barrels in proteomes," *Nucl. Acids Res.* **32**, pp. 2566–2577, May 2004.

[7] Q. Liu, "Identification of β-barrel membrane proteins based on amino acid composition properties and predicted secondary structure, url = http://dx.doi.org/10.1016/S1476-9271(02)00085-3, volume = 27, year = 2003," *Computational Biology and Chemistry* , pp. 355–361, July.

[8] A. G. Garrow, A. Agnew, and D. R. Westhead, "Tmb-hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins.," *BMC bioinformatics* **6**, 2005.

[9] K.-J. J. Park, M. M. Gromiha, P. Horton, and M. Suwa, "Discrimination of outer membrane proteins using support vector machines.," *Bioinformatics (Oxford, England)* **21**, pp. 4223–4229, December 2005.

[10] Y. Zhai and M. H. Saier, "The beta-barrel finder (bbf) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes.," *Protein science : a publication of the Protein Society* **11**, pp. 2196–2207, September 2002.

[11] J. Waldispühl, B. Berger, P. Clote, and J.-M. M. Steyaert, "Predicting transmembrane beta-barrels and interstrand residue interactions from sequence.," *Proteins* **65**, pp. 61–74, October 2006.

[12] M. M. Gromiha and M. Suwa, "Influence of amino acid properties for discriminating outer membrane proteins at better accuracy.," *Biochimica et biophysica acta* **1764**, pp. 1493–1497, September 2006.

[13] J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. L. Brinkman, "Psort-b: improving protein subcellular localization prediction for gram-negative bacteria," *Nucl. Acids Res.* **31**, pp. 3613–3617, July 2003.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures.," *Journal of molecular biology* **247**, pp. 536–540, April 1995.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank.," *Nucleic acids research* **28**, pp. 235–242, January 2000.

[16] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics* **17**, pp. 282–283, March 2001.

[17] M. Ikeda, M. Arai, T. Okuno, and T. Shimizu, "Tmpdb: a database of experimentally-characterized transmembrane topologies," *Nucl. Acids Res.* **31**, pp. 406–409, January 2003.

[18] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, "Multiple sequence alignment with the clustal series of programs.," *Nucleic acids research* **31**, pp. 3497–3500, July 2003.

[19] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme.," *Biochimica et biophysica acta* **405**, pp. 442–451, October 1975.