ELSEVIER

Biophysical Chemistry 117 (2005) 65 – 71

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Motifs in outer membrane protein sequences: Applications for discrimination

M. Michael Gromiha*

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received 24 February 2005; received in revised form 1 April 2005; accepted 1 April 2005
Available online 17 May 2005

Abstract

Discriminating outer membrane proteins (OMPs) from other folding types of globular and membrane proteins is an important problem for predicting their secondary and tertiary structures and detecting outer membrane proteins from genomic sequences as well. In this work, we have systematically analyzed the distribution of amino acid residues in the sequences of globular and outer membrane proteins with several motifs, such as A*B, A**B, etc. We observed that the motifs E*L, A*K and L*E occur frequently in globular proteins while S*S, N*S and R*D predominantly occur in OMPs. We have devised a statistical method based on frequently occurring motifs in globular and OMPs and obtained an accuracy of 96% and 82% for correctly identifying OMPs and excluding globular proteins, respectively. Further, we noticed that the motifs of transmembrane helical (TMH) proteins are different from that of OMPs. While I*A, I*L and L*I prefer in TMH proteins S*S, N*S and N*N predominantly occur in OMPs. The information about the occurrence of A*B motifs in TMH and OMPs could discriminate them with an accuracy of 80% for excluding OMPs and 100% for identifying OMPs. The influence of protein size and structural class for discrimination is discussed.

Keywords: Discrimination; Motif; Outer membrane protein

1. Introduction

Outer membrane proteins (OMPs) perform a variety of functions, such as mediating non-specific, passive transport of ions and small molecules, selectively passing the molecules like maltose and sucrose [1,2] and are involved in voltage dependent anion channels [3]. These proteins contain β-strands as their membrane spanning segments and are found in the outer membranes of bacteria, mitochondria and chloroplast [4]. A comparative analysis of the distribution of amino acid residues in α-helical and β-barrel membrane proteins shows that the membrane part of OMP is more complex than transmembrane helical (TMH) proteins due to the intervention of many charged and polar residues in the membrane [5,6]. Consequently, the success rate of discriminating OMPs from other proteins is considerably lower than that of TMH proteins [7,8].

For the past two decades the parameter, “amino acid composition” and the properties related with composition have been used for predicting the structural classes of globular proteins and discriminating TMH proteins [9–14]. Further, methods based on dipeptide composition have been used for predicting protein structural class [15,16]. These methods perform well in distinguishing the structural class of globular proteins into all-α, all-β, α + β or α/β. On the other hand, several methods have been proposed for discriminating outer membrane proteins from amino acid sequences. These methods include structure based sequence alignment [17], hydrophobicity [18], amino acid composition in the membrane spanning regions [19], Hidden Markov Model [20,21] and machine learning techniques [22]. However, the accuracy of discriminating OMPs is rather modest.

In our earlier work, we have proposed a method based on the amino acid composition of globular and outer membrane

Abbreviation: OMP, outer membrane protein.

* Tel.: +81 3 3599 8046; fax: +81 3 3599 8081.
E-mail address: michael-gromiha@aist.go.jp.

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2005.04.005

proteins for discriminating outer membrane proteins [23]. In this work, we have systematically analyzed the residue distribution along the sequence by means of several motifs, and based on this information, we have developed a statistical method for discriminating OMPs. We have examined our approach with several sets of globular proteins belonging to four different structural classes, transmembrane helical proteins, and OMPs obtained from both well annotated sequences and known three dimensional structures. Our predicted results showed an accuracy of 96% for correctly picking up the OMPs from well-known annotated sequences and the present method is able to exclude up to 82% of globular proteins correctly. The comparison of A*B motifs between 268 TMH proteins and 377 OMPs showed that these classes of proteins are discriminated with the accuracy of 80% and 100%, respectively. These accuracy levels are comparable to or better than other methods in the literature.

2. Materials and methods

2.1. Data sets

The following datasets have been used for calculating the composition of A*B motifs, identifying OMPs and excluding globular and α-helical transmembrane proteins: (i) 377 well annotated OMPs from the PSORT-B database [24], (ii) 674 globular protein chains were extracted from the PDB40D_1.37 database of SCOP with the sequence identity of less than 40% [25–27]. This dataset includes 155 all-α proteins, 156 all-β proteins, 184 α+β proteins and 179 α/β proteins, (iii) non-redundant dataset of 19 OMPs available in Protein Data Bank with sequence identity of less than 25% [26] and (iv) a dataset of 268 well-annotated TMH proteins [24]. Further, we have used subsets of 208 OMPs and 206 TMH proteins with less than 40% sequence identity obtained using the program CD-HIT [28].

2.2. Training and cross-validation tests

We have calculated the training accuracy using the datasets of 377 OMP and 674 globular protein sequences. All these sequences have been used for deriving the amino acid composition and predicting the type of the protein.

Further, we have performed a repeated, class-balanced, 2-fold cross-validation test for estimating the accuracy of discrimination: a set of N proteins is split into equally balanced subsets; parameters are developed on M proteins and then tested on the remaining $N-M$ proteins [29,30]; the procedure is repeated for all subsets of data to obtain the average accuracy. We have used 189 outer membrane and 337 globular proteins (*Set A*) to derive the amino acid composition and the result obtained with *Set A* was used to discriminate the remaining proteins in the dataset (188 outer membrane and 337 globular proteins, *Set B*). The same procedure has been repeated by keeping *Set B* as the training and *Set A* as test set. Further, we have shuffled the sequences in the whole dataset of globular and OMPs, separately, and divided the proteins into training (*Set A1*) and test sets (*Set B1*) as described above. Using Sets A1 and B1, we have repeated the calculation. This procedure was repeated several times to validate the performance of our method. This type of test is known as cross-validation test.

We have also used the datasets of 268 TMH and 377 OMPs for discriminating proteins belonging to these groups.

2.3. Computation of dipeptide motif composition

The composition of all the 400 dipeptides motifs (A*B, where A and B are specific amino acids of 20 types and * is any amino acid) based on the distribution of amino acid residues along the sequences of globular and OMPs has been computed using the following expression:

$$\mathrm{Dipep}(i,j) = \sum N_{ij} * 100.0 / \left(\sum N_i + \sum N_j\right) \quad (1)$$

where i,j stands for the distribution of 20 amino acid residues at positions i and $i+2$. $N_{i,j}$ is the number of residues of type i followed by the residue j. $\sum N_i$ and $\sum N_j$ are the total number of residues of type i and j, respectively. The same procedure was repeated for the globular proteins for obtaining their amino acid composition. The total number of dipeptides in the datasets of globular and OMPs, are 122,569 and 203,814, respectively. We have also computed the dipeptide compositions of AB, A**B and A***B.

2.4. Discrimination of OMPs

We have followed the following steps to discriminate OMPs: (i) calculated the dipeptide composition for both globular ($\mathrm{Dipep_{glob}}$) and OMPs ($\mathrm{Dipep_{OMP}}$) and the difference between them ($\sigma_{\mathrm{OMP-glob}}$); (ii) for a new protein, X, we have calculated the dipeptide composition using Eq. (1) and given weights to the dipeptide composition with $\sigma_{\mathrm{OMP-glob}}$; (iii) calculated the sum of weighted dipeptide composition and (iv) the protein X is predicted to be an OMP if the total weighted dipeptide composition is positive and globular protein otherwise.

We have performed several tests including the compositions of AB, A**B, A***B and observed that the performance of A*B is better than others. Hence, we report the results based on A*B only.

3. Results and discussion

3.1. Dipeptide motif composition in globular and OMPs

The dipeptide motif composition for all possible 400 pairs in globular and OMPs have been computed using Eq. (1) and the difference between them are presented in Table 1.

Table 1
Difference of dipeptide composition of A*B motif in globular and OMPs

Residue	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.92	−0.27	−0.76	−0.71	0.06	−0.11	−0.89	−0.05	**−1.24**	0.71	−0.73	0.13	0.73	0.23	−0.15	0.81	0.57	0.47	−0.01	0.65
Asp	−0.06	0.26	−0.64	−0.52	−0.54	0.53	−0.12	−0.86	0.24	−0.54	0.08	0.37	−0.23	0.74	0.84	0.37	0.13	−0.50	−0.05	−0.17
Cys	−0.60	−1.00	0.75	−0.82	−0.51	**−1.03**	−0.84	−0.47	−0.33	−0.74	−0.34	−0.75	−0.42	−0.78	−0.84	−0.78	−0.43	−0.69	−0.51	−0.61
Glu	−0.66	−0.16	−0.51	−0.03	−0.82	−0.36	−0.46	−0.95	−0.13	**−1.56**	−0.86	0.28	−0.43	0.45	0.16	0.55	0.12	**−1.06**	−0.69	−0.19
Phe	0.14	−0.50	−0.70	−0.75	0.25	0.44	−0.78	−0.07	−0.57	0.49	−0.49	−0.42	−0.24	−0.31	−0.25	−0.20	0.07	0.06	0.24	0.21
Gly	0.31	−0.39	**−1.02**	0.17	0.29	0.49	−0.64	−0.29	−0.05	0.47	−0.24	0.90	**−1.03**	0.23	0.32	**1.03**	0.32	−0.30	0.08	0.38
His	−0.80	−0.55	−0.60	−0.36	0.01	−0.99	**−1.04**	−0.70	−0.40	−0.58	−0.78	−0.23	−0.80	0.08	−0.43	−0.26	−0.62	−0.94	−0.21	−0.80
Ile	−0.44	0.10	−0.58	−0.94	−0.23	−0.26	−0.81	−0.13	−0.73	−0.01	−0.42	−0.45	0.40	0.27	−0.75	−0.03	−0.30	0.06	0.04	0.16
Lys	−0.20	0.26	−0.51	−0.32	−0.85	−0.40	−0.18	−0.88	0.27	−0.91	−0.58	0.46	−0.86	0.23	0.38	0.52	0.49	−0.70	−0.18	−0.12
Leu	0.84	−0.11	−0.69	**−1.23**	0.12	0.67	−0.73	0.13	**−1.15**	0.93	−0.40	−0.20	−0.15	0.02	0.06	0.35	−0.22	0.94	−0.01	0.79
Met	−0.41	−0.44	−0.50	−0.81	−0.25	−0.30	−0.59	−0.48	−0.14	−0.22	−0.03	−0.29	−0.38	−0.27	−0.27	−0.45	−0.17	−0.44	−0.16	−0.12
Asn	0.47	0.19	−0.77	0.38	0.23	0.72	−0.35	−0.38	0.08	0.44	−0.29	0.86	−0.26	0.68	0.36	**1.40**	0.50	0.31	−0.15	0.14
Pro	−0.43	−0.58	−0.61	−0.54	−0.15	−0.70	−0.52	0.55	0.01	−0.05	−0.34	0.15	−0.28	−0.42	−0.02	−0.01	−0.53	−0.20	−0.35	−0.67
Gln	0.04	0.69	−0.61	0.27	0.15	0.33	−0.29	0.05	0.49	0.08	−0.54	**1.11**	−0.05	**1.14**	0.21	0.62	0.80	0.14	−0.42	0.01
Arg	0.14	**1.21**	−0.85	0.58	−0.27	0.51	−0.65	−0.61	−0.03	−0.79	−0.50	**1.05**	−0.35	0.40	0.38	0.68	0.28	−0.44	−0.60	0.14
Ser	0.55	0.67	−0.68	−0.08	−0.10	0.95	−0.51	−0.20	0.14	0.44	−0.39	**1.06**	0.21	0.90	0.66	**1.65**	0.57	0.34	−0.20	0.30
Thr	0.18	0.12	−0.70	−0.54	−0.02	0.58	−0.78	0.00	0.06	0.45	−0.18	**1.04**	−0.70	0.30	0.32	0.40	0.69	0.32	−0.19	0.43
Val	0.34	−0.38	−0.85	−0.63	−0.26	0.10	−0.71	0.09	−0.52	0.73	−0.69	−0.28	0.14	0.26	−0.54	−0.15	0.05	0.48	0.24	0.78
Trp	0.05	−0.14	−0.03	−0.48	0.20	−0.06	−0.09	−0.06	−0.47	0.20	−0.28	−0.27	−0.34	−0.13	−0.04	−0.25	−0.65	0.06	−0.26	−0.06
Tyr	−0.02	0.22	−0.43	−0.07	**1.05**	0.03	0.17	0.24	0.48	0.29	−0.40	0.32	−0.83	0.59	0.13	0.21	0.54	0.15	0.03	0.40

The dipeptides that have high (>1.0) difference between OMP and globular proteins are highlighted in bold.

Table 2
Discrimination of globular and outer membrane proteins in a typical set of 50 sample proteins

No.	PDB code/name	Dipep* ($\sigma_{OMP-glob}$)	Assignment
All-α proteins			
1	1AB3	−58.48	GP
2	1ABV	−30.70	GP
3	1ACA	−42.47	GP
4	1ACP	−60.43	GP
5	1ADT	−59.32	GP
6	1AEP	−58.45	GP
7	1AGR	−83.41	GP
8	1AK4	−88.23	GP
9	1ALO	−111.74	GP
10	1AOF	−48.81	GP
All-β proteins			
11	1BBT	−34.98	GP
12	1BCO	−0.16	GP
13	1BDO	−123.75	GP
14	1BIA	−50.53	GP
15	1BMF	−67.01	GP
16	1BNC	−101.78	GP
17	1BPL	−15.95	GP
18	1BTY	−0.56	GP
19	1BW3	−1.83	GP
20	1CD1	−7.52	GP
α+ β proteins			
21	1AFI	−2.12	GP
22	1AG2	−14.13	GP
23	1AH6	−7.08	GP
24	1AHQ	−34.28	GP
25	1AIH	−67.63	GP
26	1AKO	−48.10	GP
27	1ADR	−25.34	GP
28	1APS	−1.43	GP
29	1APY	−80.20	GP
30	1AST	−47.55	GP
α/β proteins			
31	1AD3	−86.53	GP
32	1ADD	−99.84	GP
33	1AK5	−24.04	GP
34	1AMP	−21.02	GP
35	1ART	−32.37	GP
36	1ASU	−39.30	GP
37	1AYL	−9.81	GP
38	1BAM	−28.28	GP
39	1BCO	−88.10	GP
40	1BKS	−60.25	GP
	Total number of proteins: 674		
	Number of proteins correctly excluded: 554		
	Accuracy: 82.2%		
Outer membrane proteins			
1	outD (479227)	98.41	OMP
2	mrpC (485956)	65.64	OMP
3	OpcM (1061410)	59.30	OMP
4	OMP1 (1262291)	62.53	OMP
5	OmpK17 (1279830)	175.84	OMP
6	HrcC (1336093)	47.49	OMP
7	Oma87 (1401350)	80.17	OMP
8	SpiA (1498307)	55.18	OMP
9	HecB (1772622)	65.58	OMP
10	PscC (1781385)	60.84	OMP
	Total number of proteins: 377		
	Number of proteins correctly assigned: 361		
	Accuracy: 95.8%		

The NCBI gi numbers for the outer membrane proteins are given in parenthesis. GP: globular protein; OMP: outer membrane protein.

In this table, positive values indicate the higher occurrence in OMPs than globular proteins and negative values show the preference of dipeptide motifs in globular proteins. We observed that the occurrence of dipeptide motifs, S*S, V*S, R*D, Q*Q, Q*N, S*N, R*N, Y*F, T*N and G*S are significantly higher in OMPs than in globular proteins. Interestingly, most of the dipeptide motifs involve the residues Ser, Asn and Gln, which are among the most favored residues in OMPs [23]. Further, it has been reported that these residues play an important role to the structure and stability of OMPs [31–33]. On the other hand, the dipeptide motifs, E*L, A*K, L*E, L*K, H*H, C*G and G*C have higher occurrence in globular proteins than OMPs. It is noteworthy that most of these dipeptide motifs involve the charge residues, Lys, Glu and His, which have significantly higher occurrence in globular proteins compared with OMPs [23]. Further, the occurrence of two adjacent polar or hydrophobic residues in OMPs indicates the amphipathic nature of amino acid residues in β-strand segments of OMPs [11].

3.2. Discrimination of globular and OMPs

We have calculated the dipeptide motif composition for each of the 674 globular and 377 OMPs using Eq. (1) and the difference between them (Table 1). For each protein, we have calculated the weighted sum of dipeptide motif compositions and the results for a sample set of 50 proteins are shown in Table 2. For 1AB3, the weighted sum of dipeptide composition is −97.30 and hence this protein is predicted as a non-OMP, as known from its structural information [23,34]. On the other hand, for OutD protein, the weighted sum of dipeptide composition is 71.35 and hence it is identified as an OMP, showing the agreement with experimental observations. From the set of 377 OMPs and 674 globular proteins, we have correctly identified 361 OMPs (95.8%) and excluded 554 globular proteins (82.2%). The cross-validation test described earlier yielded an average accuracy of 93% and 77% respectively, for correctly identifying OMPs and excluding globular proteins. Further, the present method correctly identified 95.7% of OMPs and excluded 81.9% of globular proteins in a dataset of 208 OMPs and 674 globular proteins with the sequence identity of less than 40%.

3.3. Analysis on different structural classes and proteins of different size

We have analyzed the prediction results based on different structural classes, all-α, all-β, α+β and α/β. We observed that the prediction accuracies for these four classes of proteins are, respectively, 90%, 64%, 84% and 90%, indicating the better performance of all-α, α+β and α/β proteins. We have also estimated the accuracy of discrimination for proteins belonging to each structural class when proteins of such specific class alone are used to compute the dipeptide motif composition. We observed that the discriminative accuracy significantly improved in all structural classes of proteins and are, respectively, 93%, 80%, 91% and 96% in all-α, all-β, α+β and α/β class. In all these classes, the OMPs are correctly identified with the accuracy in the range of 96–98%.

Furthermore, we have divided the proteins based on their size and we observed that the proteins with less than 300 residues are correctly excluded from OMPs with an accuracy of about 82%, and the proteins of large size are excluded at an accuracy of 86%. Proteins that have 301–400 residues are excluded with 84% accuracy.

In OMPs, the large size proteins (more than 800 residues) are correctly identified with an accuracy of 100% and the proteins with more than 300 residues are picked up with about 98% accuracy. Proteins with 300 residues or less are predicted with an accuracy of 86%.

Furthermore, we have set up a representative set of 19 non-redundant OMPs. The weighted sum of the dipeptide motif composition for 18 OMPs is positive and hence the prediction accuracy of known structures is 95%.

3.4. Dipeptide motif composition in TMH and OMPs

The difference between the dipeptide motif composition for all possible 400 pairs in TMH and OMPs are presented in Table 3. We observed that the occurrence of dipeptide motifs, S*S, N*S, N*H, S*N, N*G and Q*N is significantly higher in OMPs than in TMH proteins. This result shows the importance of Ser and Asn in the folding pattern of OMPs and the formation of β-strands in the membrane compared with TMH proteins. On the other hand, the dipeptide motifs with hydrophobic residues, I*A, L*I and I*L have higher occurrence in TMH proteins than OMPs. It is due to the presence of long stretches of hydrophobic residues in the membrane part of TMH proteins.

3.5. Discrimination of TMH and OMPs

We have calculated the dipeptide motif composition for each of the 268 TMH proteins and 377 OMPs using Eq. (1) and the difference between them (Table 3). For each protein, we have calculated the weighted sum of dipeptide motif compositions. From the set of 377 OMPs and 268 TMH proteins, our method correctly identified all the 377 OMPs (100%) and excluded 213 TMH proteins (80%). The dataset with 208 OMPs and 206 TMH proteins with less than 40% sequence identity also showed similar results. The cross-validation test described earlier yielded an average accuracy of 99.9% and 78.4% respectively, for correctly identifying OMPs and excluding TMH proteins. The high accuracy of discriminating TMH and OMPs is due to the significant difference in the dipeptide motif composition between them.

3.6. Comparison with other methods

Gnanasekaran et al. [17] devised a method based on sequence alignment profiles of porins to identify the β-stranded OMPs and reported an accuracy of about 80%. Liu et al. [20] proposed a method based on the amino acid composition of residues in transmembrane β-strand segments of 12 proteins to discriminate β-barrel membrane proteins and claimed an accuracy of 84% in a set of 241 OMPs. This method missed to identify OMPs with fewer membrane spanning β-strand segments, such as, 7AHL (α-hemolysin), which has two membrane spanning β-strands. Martelli et al. [19] devised a neural network method using 12 OMPs and tested the method in 145 OMPs, which has yielded the accuracy of 84%. Bagos et al. [21] used a HMM for discriminating β-barrel OMPs and reported an accuracy of 88% for a set of 133 OMPs. The method based on amino acid compositions showed an accuracy of 89% in a dataset of 377 OMPs [23]. In this work, we have used a set of 377 OMPs and discriminated them with an accuracy of 96%. Further, in a pool of TMH and OMPs the present method correctly excluded 80% of the transmembrane α-helical proteins and identified the OMPs with 100%. The high accuracy achieved by the present method is due to the superiority of the method as well as the information gained from the large dataset of globular, TMH and OMPs. As the sequence information alone is sufficient to derive the dipeptide motif composition, one can refine the parameters frequently, as the number of OMP sequences is growing rapidly, which may improve the accuracy.

3.7. Applications for detecting OMPs in genomic sequences

The present method has the following applications to detect OMPs in genomic sequences. First, one can eliminate the TMH proteins using the dipeptide motif composition of TMH and OMPs. This step will eliminate 80% of the TMH proteins. On the other hand, it will include all the OMPs in genomic sequences and none of the OMP will be excluded. The result obtained with this discrimination can be used to exclude the globular protein sequences using the dipeptide motif composition of globular and OMPs. The combination of two methods could be able to dissect the OMPs in genomic sequences successfully.

Table 3
Difference of dipeptide composition of A*B motif in α-helical membrane and OMPs

Residue	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.29	0.74	−0.44	−0.16	−1.10	−0.34	−0.18	−1.26	0.18	−0.79	−1.49	0.40	0.11	0.52	0.25	0.32	0.54	−0.61	−0.49	0.96
Asp	0.91	0.90	0.15	0.11	0.18	1.13	0.06	0.35	0.81	0.98	0.10	1.46	−0.13	0.90	1.20	1.23	1.23	0.54	0.06	0.63
Cys	−0.19	−0.13	0.91	−0.22	−0.21	−0.33	−0.17	−0.47	−0.07	−0.52	−0.13	0.01	0.08	0.00	−0.23	−0.26	−0.04	−0.42	−0.28	−0.63
Glu	−0.18	0.35	0.15	0.34	0.14	0.61	−0.28	0.17	0.12	0.51	−0.57	0.56	0.01	0.36	0.24	0.80	0.40	−0.19	−0.13	0.75
Phe	−1.16	0.57	−0.32	0.31	−0.88	−0.61	−0.30	−1.20	0.32	−0.86	−1.09	0.18	−0.67	−0.32	−0.56	−0.71	−0.57	−0.96	−0.23	−0.33
Gly	−0.43	0.84	−0.39	0.74	−0.73	0.25	−0.26	−1.36	0.88	−0.94	−1.12	1.77	−0.42	0.41	0.82	0.66	0.94	−0.70	−0.60	0.81
His	−0.47	−0.55	0.02	−0.31	0.34	−0.27	−0.30	−0.20	0.03	−0.17	−0.27	−0.13	−0.41	0.22	−0.60	−0.05	−0.14	−0.30	−0.60	−0.09
Ile	−**1.79**	1.15	−0.49	0.53	−1.37	−1.22	−0.47	−1.15	0.26	−**1.53**	−0.95	0.24	0.15	0.36	−0.42	−0.53	−0.84	−0.94	−0.27	0.03
Lys	0.74	0.70	0.16	0.21	0.09	0.64	0.25	0.18	1.39	0.71	−0.53	0.68	−0.14	0.90	0.23	0.81	0.85	0.88	−0.12	0.86
Leu	−0.64	1.21	−0.52	0.05	−1.12	−0.66	−0.36	−**1.62**	0.65	−1.46	−1.30	0.29	−0.13	0.51	0.00	−0.20	−0.84	−1.04	−0.67	0.79
Met	−0.95	−0.69	−0.37	−1.01	−1.07	−1.12	−0.33	−1.20	0.07	−1.21	−0.55	−0.72	−0.67	−0.61	−0.54	−1.19	−1.03	−1.13	−0.63	−0.66
Asn	0.66	0.87	−0.08	0.65	0.09	**1.53**	−0.46	0.37	1.22	0.72	−0.27	**1.69**	−0.04	1.00	1.06	**1.99**	1.27	0.67	−0.21	0.96
Pro	−0.24	0.13	−0.13	−0.16	−0.28	−0.23	−0.45	0.01	0.42	−0.35	−1.01	0.39	0.47	−0.29	−0.17	0.40	0.15	−0.12	−0.99	−0.35
Gln	0.26	0.95	0.00	0.66	−0.25	0.86	−0.11	0.32	0.63	0.34	−0.39	**1.51**	−0.32	0.50	0.15	0.54	0.81	0.57	−0.31	0.70
Arg	0.60	1.33	0.00	0.62	0.04	0.21	−0.15	−0.07	0.04	−0.05	−1.05	1.10	−0.57	0.17	−0.83	0.82	0.17	0.03	−0.69	0.66
Ser	0.42	1.25	−0.25	0.24	−0.85	0.98	−0.32	−0.42	0.22	−0.16	−1.15	**1.61**	0.54	0.66	0.65	**2.05**	0.92	0.03	−0.52	0.87
Thr	0.12	0.96	−0.05	0.56	−0.63	0.55	−0.10	−0.72	0.86	−0.57	−0.63	1.45	−0.01	0.57	0.80	0.95	1.29	0.05	−0.52	0.94
Val	−0.47	0.69	−0.49	0.49	−1.05	−0.99	−0.42	−0.72	0.38	−1.05	−1.14	0.45	0.19	0.57	−0.09	−0.35	0.15	0.30	−0.09	0.55
Trp	−0.39	0.00	0.05	−0.27	−0.52	−0.34	−0.18	−0.65	−0.31	−0.69	−0.59	0.06	−0.57	−0.13	−0.11	−0.60	−0.48	−0.50	−1.36	−0.72
Tyr	0.29	1.01	−0.41	0.64	1.22	0.66	0.26	0.12	1.09	0.21	−0.77	1.13	−0.69	1.09	0.47	0.56	0.74	0.35	−0.21	0.90

The dipeptides that have high (>1.5) difference between OMP and TMH proteins are highlighted in bold.

4. Conclusions

We have systematically analyzed the amino acid sequences of globular and OMPs and developed the dipeptide motif composition for these classes of proteins. The preferences of dipeptide motif composition between globular and OMPs and that between TMH proteins and OMPs have been explored. Based on these results, we have devised a statistical method based on the weighted sum of dipeptide motif composition to identify the OMPs and to exclude globular proteins. Our method correctly identified 96% of the OMPs and excluded up to 82% of the globular proteins in a data set of 377 OMPs and 674 globular proteins. Further, from the pool of 268 TMH proteins and 377 OMPs, the present method correctly excluded 80% of the α-helical membrane proteins and identified 100% of OMPs. These accuracy levels are comparable to or better than other methods in the literature. We suggest that this simple method could be effectively used to discriminate OMPs from globular and membrane proteins.

References

[1] T. Schirmer, T.A. Keller, Y.F. Wang, J.P. Rosenbusch, Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution, Science 267 (1995) 512–514.

[2] D. Forst, W. Welte, T. Wacker, K. Diederichs, Structure of the sucrose-specific porin ScrY from Salmonella typhimurium and its complex with sucrose, Nat. Struct. Biol. 5 (1998) 37–46.

[3] C.A. Mannella, Conformational changes in the mitochondrial channel protein, VDAC, and their functional implications, J. Struct. Biol. 121 (1998) 207–218.

[4] G.E. Schulz, The structure of bacterial outer membrane proteins, Biochim. Biophys. Acta 1565 (2002) 308–317.

[5] M.M. Gromiha, R. Majumdar, P.K. Ponnuswamy, Identification of membrane spanning beta strands in bacterial porins, Protein Eng. 10 (1997) 497–500.

[6] M.M. Gromiha, A simple method for predicting transmembrane alpha helices with better accuracy, Protein Eng. 12 (1999) 557–561.

[7] T. Hirokawa, S. Boon-Chieng, S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, Bioinformatics 14 (1998) 378–379.

[8] C.P. Chen, B. Rost, State-of-the-art in membrane protein prediction, Appl. Bioinformatics 1 (2002) 21–35.

[9] P. Klein, Prediction of protein structural class by discriminant analysis, Biochim. Biophys. Acta 874 (1986) 205–215.

[10] K.C. Chou, C.T. Zhang, A correlation-coefficient method to predicting protein-structural classes from amino acid compositions, Eur. J. Biochem. 207 (1992) 429–433.

[11] M.M. Gromiha, P.K. Ponnuswamy, Prediction of protein secondary structures from their hydrophobic characteristics, Int. J. Pept. Protein Res. 45 (1995) 225–240.

[12] W.S. Bu, Z.P. Feng, Z. Zhang, C.T. Zhang, Prediction of protein (domain) structural classes based on amino-acid index, Eur. J. Biochem. 266 (1999) 1043–1049.

[13] K.C. Chou, G.M. Maggiora, Domain structural class prediction, Protein Eng. 11 (1998) 523–538.

[14] S. Mitaku, T. Hirokawa, T. Tsuji, Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces, Bioinformatics 18 (2002) 608–616.

[15] T.S. Kumarevel, M.M. Gromiha, M.N. Ponnuswamy, Structural class prediction: an application of residue distribution along the sequence, Biophys. Chemist. 88 (2000) 81–101.

[16] K.C. Chou, Prediction of protein structural classes and subcellular locations, Curr. Protein Pept. Sci. 1 (2000) 171–208.

[17] T.V. Gnanasekaran, S. Peri, A. Arockiasamy, S. Krishnaswamy, Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins, Bioinformatics 16 (2000) 839–842.

[18] W.C. Wimley, Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures, Protein Sci. 11 (2002) 301–312.

[19] Q. Liu, Y. Zhu, B. Wang, Y. Li, Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure, Comput. Biol. Chem. 27 (2003) 355–361.

[20] P.L. Martelli, P. Fariselli, A. Krogh, R. Casadio, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins, Bioinformatics 18 (2002) S46–S53.

[21] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, S.J. Hamodrakas, A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins, BMC Bioinformatics 5 (2004) 29.

[22] N.K. Natt, H. Kaur, G.P. Raghava, Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods, Proteins 56 (2004) 11–18.

[23] M.M. Gromiha, M. Suwa, A simple statistical method for discriminating outer membrane proteins with better accuracy, Bioinformatics 21 (2005) 961–968.

[24] J.L. Gardy, C. Spencer, K. Wang, M. Ester, G.E. Tusnady, I. Simon, S. Hua, K. de Fays, C. Lambert, K. Nakai, F.S. Brinkman, PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria, Nucleic Acids Res. 31 (2003) 3613–3617.

[25] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[26] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[27] Z.X. Wang, Z. Yuan, How good is prediction of protein structural class by the component-coupled method? Proteins 38 (2000) 165–175.

[28] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282–283.

[29] J.A. Cuff, G.J. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, Proteins 34 (1999) 508–519.

[30] S. Ahmad, M.M. Gromiha, NETASA: neural network based prediction of solvent accessibility, Bioinformatics 18 (2002) 819–824.

[31] A. Pautsch, G.E. Schulz, High-resolution structure of the OmpA membrane domain, J. Mol. Biol. 298 (2000) 273–282.

[32] K. Zeth, K. Diederichs, W. Welte, H. Engelhardt, Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with a periplasmic peptide at 2.1 A resolution, Structure 8 (2000) 981–992.

[33] L. Vandeputte-Rutten, R.A. Kramer, J. Kroon, N. Dekker, M.R. Egmond, P. Gros, Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site, EMBO J. 20 (2001) 5033–5039.

[34] H. Berglund, A. Rak, A. Serganov, M. Garber, T. Hard, Solution structure of the ribosomal RNA binding protein S15 from *Thermus thermophilus*, Nat. Struct. Biol. 4 (1997) 20–23.