Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 353 (2007) 164–169
BBRC
www.elsevier.com/locate/ybbrc

Using amino acid and peptide composition to predict membrane protein types

Xiao-Guang Yang [a,1], Rui-Yan Luo [b,1], Zhi-Ping Feng [c,*]

[a] Department of Biological Engineering, University of Missouri-Columbia, Columbia, MO 65211, USA
[b] Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA
[c] The Walter & Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Vic. 3050, Australia

Received 24 November 2006
Available online 8 December 2006

Abstract

Membrane proteins play an important role in many biological processes and are attractive drug targets. Determination of membrane protein structures or topologies by experimental methods is expensive and time consuming. Effective computational method in predicting the membrane protein types can provide useful information for large amount of protein sequences emerging in the post-genomic era. Although numerous algorithms have addressed this issue, the methods of extracting efficient protein sequence information are very limit. In this study, we provide a method of extracting high order sequence information with the stepwise discriminant analysis. Some important amino acids and peptides that are distinct for different types of the membrane proteins have been identified and their occurrence frequencies in membrane proteins can be used to predict the types of the membrane proteins. Consequently, an accuracy of 86.5% in the cross-validation test, and 99.8% in the resubstitution test has been achieved for a non-redundant dataset, which includes type-I, type-II, multipass transmembrane proteins, lipid chain-anchored and GPI-anchored membrane proteins. The fingerprint features of the identified peptides in each membrane protein type are also discussed.

Keywords: The stepwise discriminant analysis; Type-I membrane protein; Type-II membrane protein; Multipass transmembrane protein; Lipid-chain-anchored membrane protein; GPI-anchored membrane protein

All living cells are enveloped by membranes, which are composed primarily of lipids and proteins. Membrane proteins are embedded in the lipid bilayer, which creates a suitable environment for their actions. Although the physical aspects of protein–lipid interactions is poorly understood, it is believed that most of the specific functions are carried out by the membrane proteins and their associations with the lipid bilayer [1,2]. As a result, membrane proteins are attractive targets for drug design. For example, proteins on the surface of the malaria parasite are attractive candidates for inclusion in a malaria vaccine as they are accessible to plasma antibodies that block parasite invasion of host cells [3]. Membrane apical antigen 1 (AMA1) and a series of merozoite surface proteins (MSP1-10) of human malaria parasite *Plasmodium falciparum* (Pf) belong to such kind of proteins [3–5].

There are mainly five kinds of membrane proteins. The function of each type of membrane proteins is closely related their topologies. For example, PfAMA1 is a type-I membrane protein, which contains one transmembrane domain, N-terminal extracellular domain and C-terminal intracellular domain. The ectodomain appears to be important in orienting the merozoite on the erythrocyte surface prior to invasion, and lead to AMA1 to be an important malaria vaccine candidate. On the other hand, the majority of glycosylation enzymes are type-II membrane proteins (N-term in, C-term out). It is now clear that the targeting of Golgi resident enzymes is intimately associated with the organization of Golgi membranes and the control of protein and lipid traffic in both anterograde and retrograde

* Corresponding author. Fax: +61 3 93470852.
E-mail address: feng@wehi.edu.au (Z.-P. Feng).
[1] These authors should be regarded as joint first authors.

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.12.004

directions [6]. Compared with single transmembrane proteins, E1-E2_ATPase contains six putative transmembrane helices, and has been indicated to transport heavy metal ions, as the function of many multipass transmembrane proteins [7]. In contrast to the above integral membrane proteins, anchored membrane proteins modified by the lipophilic moieties, such as by fatty acids, isoprenoids, and glycosylphosphatidylinositol (GPI) anchor, often regulate protein interaction with membranes or other proteins, and signalling function at only one side of the lipid bilayer [8].

Experimentally identify membrane protein structure or topology type is complicated and time consuming. Many transmembrane topology prediction methods have been proposed, such as TMHMM and HMMTOP [9,10]. They can predict transmembrane region(s) based on the knowledge of signal peptides and/or typical hydrophobic feature of transmembrane helices. There are also several servers to predict GPI modification site (e.g., big-PI [11]) and GPI-SOM (http://gpi.unibe.ch/), myristoylation site (http://mendel.imp.ac.at/myristate/) and palmitoylation site (http://bioinformatics.lcd-ustc.org/css_palm/). Chou et al. have developed many algorithms [e.g., 12–20] to predict five types of the membrane proteins (shown in Fig. 1) using global protein sequence information. However, in all such previous studies, including ourselves [21,22], the protein sequences have been treated uniformly in the algorithms, which is convenient for practical use and objective test, but most high order sequence information has been ignored due to the complexity in calculations.

In this study, we identified some amino acids and peptides that can significantly discriminate five different types of the membrane proteins based on the stepwise discriminant analysis method. The extracted sequence information captures the general features of the amino acid compositions from global sequences, hydropathy from transmembrane helices, signal peptides, and modification sites by lipid. As a result, their occurrence frequencies in protein sequences can be used in predicting membrane protein types. An accuracy of 86.5% in the cross-validation test, and 99.8% in the resubstitution test have been achieved for a non-redundant dataset [21], which includes type-I, type-II, multipass transmembrane, lipid chain-anchored and GPI-anchored membrane proteins. The result indicates that the present method of extracting sequence features from proteins is efficient and it can be an important algorithm in protein bioinformatics.

Materials and methods

Dataset. Five types of membrane proteins (shown in Fig. 1) have been analyzed in this study. (1) Type-I membrane proteins: The protein spans the lipid bilayer only once and its N-terminus is at the extracellular side and C-terminus projects into the intracellular side. (2) Type-II membrane proteins: The protein also spans the lipid bilayer only once, but its C-terminus is outside the cell and N-terminus projects into the cytoplasm. (3) Multipass transmembrane proteins: The protein spans the lipid bilayer many times. (4) Lipid chain-anchored membrane proteins: the anchored membrane protein is associated with the bilayer by means of one or more covalent attached fatty acids or prenyl groups. (5) GPI-anchored membrane proteins: The anchored membrane protein is bound to the membrane by a glycosylphosphatidylinositol (GPI) anchor.

A no redundant dataset has been used in this study. It consists of 879 protein sequences derived from 40 version SWISSPROT databank [22,23] and sequence similarity was reduced to 25% with CD-HIT [24]. During the process of reducing sequence similarity, only 11.3% of the sequences in the original dataset were remained, in which 151 are type-I membrane proteins, 101 type-II membrane proteins, 518 multipass transmembrane proteins, 56 lipid chain-anchored membrane proteins, and 53 GPI-anchored membrane proteins.

Method of extracting sequence features and prediction. The stepwise discriminant analysis [25,26] is performed to extract the sequence features from the five types of membrane proteins in the dataset. In the stepwise discriminant analysis variables are chosen to leave (or enter) the model according to one of two criteria: (1) the significance level of an F-test from an analysis of covariance, where the variables already chosen act as covariations and the variable under consideration is the dependent variable; (2) the squared partial correlation for predicting the variable under consideration from the variable representing the classification of observations, controlling for the effects of the variables already selected for the model.

Let $S(1)$ be the set composed of the 20 amino acids, namely $S(1) = \mathrm{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}$. Let $S(i)$ be such a set that each of its element is a string with i members of $S(1)$ (i is a positive integer greater than 1). In other words, $S(i)$ is the set composed of all possible peptides with length of i, denoted as i-peptide(s) ($i > 1$). For brevity, the elements in $S(i)$ are ordered lexicographically. Therefore,

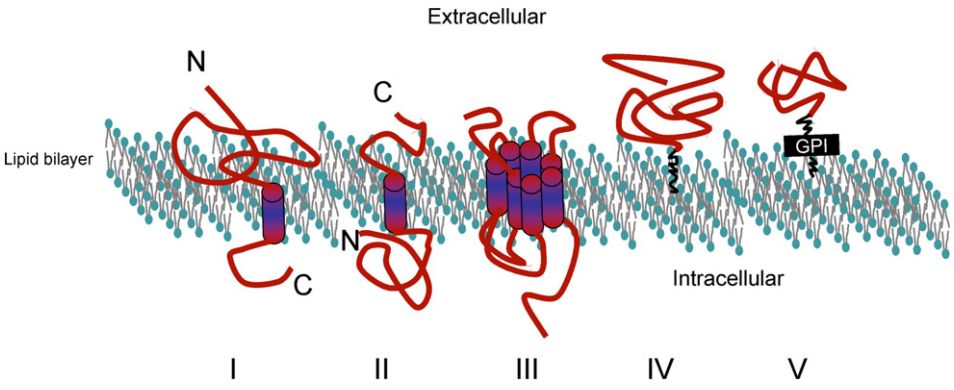

Fig. 1. Schematic show of the five types of membrane proteins: (I) type-I membrane protein, (II) type-II membrane protein, (III) multipass transmembrane protein, (IV) lipid-chain-anchored membrane protein, and (V) GPI-anchored membrane protein. The figure is modified based on [16].

$S(2) = \{AA, AC, AD, \ldots, AY, CA, CC, \ldots, YW, YY\}$,
$S(3) = \{AAA, AAC, AAD, \ldots, AAY, ACA, ACC, \ldots, YYY\}$,
$\ldots \quad \ldots \quad \ldots$

Apparently, the number of elements in $S(i)$ is $20^i (i = 1, 2, \ldots)$, which grows with exponential rate. In the following description, $S(i)$ $(i \geqslant 0)$ is called the set of peptides, but it represents the set of single amino acid residues when $i = 1$ and i-peptides when $i > 1$. In order to predict the membrane protein type of each query membrane sequence, a subset of $S(i)$, denoted as $T(i)$, is formed such that the result of prediction using $T(i)$ is almost the same as using $S(i)$, while the number of elements in $T(i)$ is much smaller than 20^i. To construct $T(i)$, first construct a subset of $T(i-1)$, $T_0(i-1)$ which is called "seeds of $(i-1)$-peptide(s)", then add every amino acid in $S(1)$ in front and at the back of each element in $T_0(i-1)$ to get a set including $T(i)$. Meanwhile, denote the quantitative variable sets that represent the occurrence frequencies of elements in $T(i)$ and $T_0(i)$ as $X(i)$ and $X_0(i)$, respectively. The algorithm (the flowchart shown in Fig. 2) is described as follows:

1. Choose "seeds of the amino acids": let $i = 1$ and $T(1) = S(1)$. Perform stepwise discriminant analysis with elements in $X(1)$ and get a subset of $X(1)$, denoted as $X_0(1)$. Accordingly a subset of $T(1)$ is obtained, denoted as $T_0(1)$. Suppose $T_0(1)$ has n(i) elements which are called "seeds". Denote the result of discrimination with variables in $X_0(1)$ as $R(1)$.
2. "seeds sprout" or construct $T(i+1)$ from $T_0(i)$ $(i=1,2,3,\ldots)$: for all the elements in $T_0(i)$, add each of the 20 elements of $S(1)$ in front and at the back of them and get $n(i) \times 20 \times 2 = 40 \times n(i)$ elements belonging to $S(i+1)$. The repeated elements are deleted to make sure that each variable appears only once. Suppose the number of $(i+1)$-peptides thus obtained is $m(i+1)$. Calculate their frequencies in each membrane protein type and the corresponding means in the subgroups, and choose the variables whose means in one subgroup are above the given threshold. Put these remaining variables together and get a quantitative variable set $X(i+1)$ as well as the corresponding polypeptide set $T(i+1)$.
3. Choose "seeds of $(i+1)$-peptides" or construct $T_0(i+1)$ from $T(i+1)$ $(i = 1, 2, 3, \ldots)$: perform stepwise discrimination with all the elements in $X(i+1)$ to select relatively important variables. Suppose $n(i+1)$ variables are chosen. Put them together and get $X_0(i+1)$, one subset of $X(i+1)$. Meanwhile the corresponding polypeptide set $T_0(i+1)$ consisting of "seeds of $(i+1)$-peptides" is obtained.
4. Check the discriminant results while $(i+1)$-peptides are taken into account: Put together all the variables in $X(1), X(2), \ldots, X(i+1)$ and perform stepwise discriminant analysis. Denote the discriminant result as $R(i+1)$ with the variables chosen. If none of $X(i+1)$ enters, or if $R(i+1)$ is not higher than $R(i)$, then stop the process. The variables obtained from $X(1), X(2), \ldots, X(i)$ by stepwise discrimination are the desired variables in final prediction and $R(i)$ is the highest prediction accuracy. Otherwise go back to step 2, letting $i \leftarrow i+1$, and iterate the following steps to lengthen "seeds of i-peptides" further.

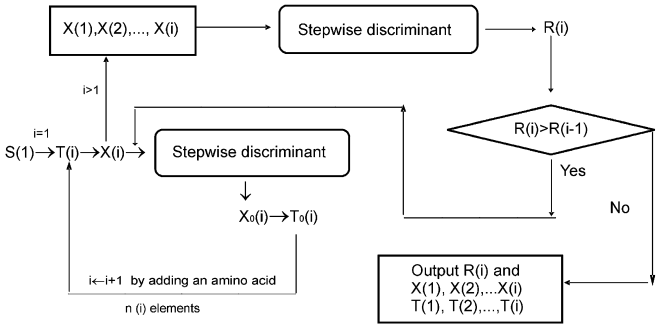

Fig. 2. Flowchart show of the algorithm on extracting protein sequence information based on stepwise discriminant analysis. See detailed description from the text.

Evaluate of the prediction algorithm. In order to assess the accuracy of the prediction algorithms, the overall accuracy, the sensitivity and specificity for each type of the membrane proteins are calculated [27], and denoted as *Overall*, Q^{D}, Q^{M}, respectively. The overall accuracy is the number of totally correctly predicted membrane proteins in relation to the total number of the membrane proteins of all types. The *sensitivity*, Q^{D}, is the true positive rate, which is the percentage of correctly predicted number of membrane proteins in each type in relation to the total number of actual membrane proteins in the type. The specificity, Q^{M}, is the percentage of correctly predicted number of membrane proteins in each type in relation to the total number of membrane proteins that have been predicted to the type. In addition, a prediction is evaluated by the resubstitution and jackknife test. The former reflects the self-consistency and the latter reflects the extrapolating effectiveness of the algorithm studied. Among various cross-validation tests in the literature to evaluate the extrapolating effectiveness of an algorithm, the jackknife test is thought to be more rigorous and reliable [12–22]. In the jackknife test each protein in the training dataset is singled out in turn as an independent test sample, all the frequencies of the chosen peptide(s) are derived from the remaining protein sequences. The final accuracy is obtained by the average results of all test samples.

Results and discussions

The overall prediction results

The results of the overall prediction accuracy in resubstitution test, jackknife test, the number of the variables chosen and the threshold set in each of the iterations of lengthening the peptides are listed in Table 1. We can see from the table that the steady improvement of the overall accuracy with the expanding of the length of the peptides. The overall accuracy in jackknife test becomes more stable between the 4-th and 5-th iterations. If the procedure is

Table 1
Prediction results with peptide(s) composition chosen from different iterations of the stepwise discriminate analysis[a]

Iteration	Threshold	Number of peptide(s)	Overall$^{\mathrm{Resub}}$ (%)	Overall$^{\mathrm{Jack}}$ (%)
1	0.00000	14	63.0	61.0
2	0.00000	96	83.8	72.1
3	0.00050	158	90.4	78.5
4	0.00012	308	99.0	86.0
5	0.00006	356	99.8	86.5

[a] Overall$^{\mathrm{Resub}}$ and Overall$^{\mathrm{Jack}}$ are the overall accuracy in the resubstitution and jackknife tests, respectively. The variables chosen from the ith iteration include the composition of i-peptides and peptides shorter than i-residue. For example, in the second iteration, 96 variables consist of the composition of some dipeptides and some single residues. The peptide(s) with frequencies smaller than the given threshold have been ignored. The peptide(s) finally chosen for the prediction are 356 from single residues to 5-peptides.

stopped at 4-th iteration, 308 variables are chosen, and the final prediction accuracy reaches to 99.0% and 86.0%, respectively, in the resubstitution test and in the jackknife test. If the procedure is stopped at 5-th iteration, 356 variables are chosen, and the final prediction accuracy reaches to 99.8% and 86.5%, respectively, in the resubstitution test and in the jackknife test.

The accuracy for each type of the membrane proteins is summarized in Table 2. From Table 2, we can see that in the resubstitution test all the sensitivities and specificities for the five different types of membrane proteins can be higher than 98% when the pentapeptides are taken into account. In the jackknife test, the sensitivities for multipass transmembrane proteins can reach 92.7% with a specificity

Table 2
Prediction results with the 5-peptides taken into account[a]

	Method	Type-I	Type-II	Multipass	Lipid chain	GPI	Overall
Q^D	Resub	151/151 (100.0)	99/101 (98.0)	518/518 (100.0)	56/56 (100.0)	53/53 (100.0)	877/879 (99.8)
	Jack	133/151 (72.7)	77/101 (76.2)	480/518 (92.7)	38/56 (84.4)	32/53 (60.4)	760/879 (86.5)
Q^M	Resub	151/151 (100.0)	99/99 (100.0)	518/520 (99.6)	56/56 (100.0)	53/53 (100.0)	877/879 (99.8)
	Jack	109/163 (66.9)	44/51 (86.3)	502/590 (85.1)	56/57 (98.2)	11/18 (61.1)	760/879 (86.5)

[a] Q^D and Q^M are sensitivity and specificity defined by Baldi et al. [27]; resubstitution test and the jackknife test are denoted as Resub and Jack, respectively. The prediction is carried out by 356 variables chosen from five iterations of the stepwise discriminate analysis.

Table 3
The peptide(s) chosen for the final prediction[a]

	1	2	3	4	5	6	7	8	9	10
0	F	K	A	DKDGK	M	PDKF	SC	CSSNA	LTACS	AKEA
1	ALKSA	VKPP	YFSKA	ADGVG	ACKPC	LEGNP	DEAIN	TTKI	GGARA	LVAGC
2	DSAA	SLEDL	AAEAL	LAYS	TTTTK	AGARA	GADDT	TVVN	AACSS	ACSSI
3	TITVQ	TVVNS	SDAK	TTKKV	STAS	KLLSQ	LAACS	H	VLEDF	PSPTP
4	STATV	VTVD	TSTEA	TVQ	AADIQ	TAEKA	PAVTA	RPAVT	KLKVV	SKDSS
5	AGCS	YAKP	LLAGC	PDKFA	NLCN	LPSPT	RC	SALK	RNVT	DEANE
6	DKDG	DSG	HTTK	GQD	AKS	LKL	KLLNV	ADATA	TATT	VKADK
7	YDSA	SKKD	TLAGC	YAEK	FTGK	SKAS	TPNTT	Y	YK	NLC
8	PNP	N	GADD	TATE	KSLDE	DSS	SE	TVDE	AEAT	LLAAC
9	AEATP	ACSSK	ISPSE	QARAA	DKD	QH	LCNEI	GH	CH	KAK
10	DADA	DKSA	KDKD	DKDM	DKDK	TNLL	PW	AIG	KVLED	EGLK
11	DADAK	EGLKE	QAKE	PDKA	LGKT	SLAAC	NGN	ACSSN	TACS	YN
12	W	GQDK	MKKIL	ACKQ	ADK	NVTT	NVTTD	NVT	VFVA	GKY
13	AEA	RL	KLLN	KLLS	LTAV	PPA	DKSK	PQ	FTGKG	KLK
14	EKLK	NVTTA	TPSPT	RT	CSST	PR	NVS	PLPP	PLPPL	SPSE
15	CI	EGAP	EGAPA	YEGA	DKDMS	VVACS	IADD	TN	EG	AAK
16	STTTN	LGKE	HE	QNVS	V	KTGK	FLAY	RY	VC	WP
17	EE	SSDAK	SDAKH	KADK	DEAN	TTTK	AANV	HH	TIAD	VTLI
18	YD	TAVS	TSKDK	KPV	ADKS	DV	LEDL	DEA	KDK	DKL
19	TSPAW	NS	EK	SD	TH	YF	SVTLG	DAEK	PM	EQ
20	YM	GARAR	KED	PI	AAAD	AADK	SKDKS	ATSPA	TSPD	KDKSA
21	AKLLS	SQ	STEAP	ADAM	LCL	LDEL	TGKA	VT	SM	KM
22	AFGL	RSD	SPT	SEVT	DLCN	IP	KC	IT	QTAT	LGN
23	AKA	LAY	LEDHG	SEAKP	SKKI	ELGKL	QL	VTL	TLDEL	LEGTL
24	NLCF	TLEGT	VTLK	CV	LKV	VR	HS	KLKM	ARAA	EGT
25	PY	VIGL	EF	AY	ADAA	NG	CD	TC	MLKL	VAGC
26	SEGLA	LPE	LGK	TTK	TVA	KAKE	SATT	APNP	DSA	LN
27	EN	ES	QAR	AVT	FC	KADEA	KEGTV	KEGT	AIGL	HY
28	RW	LAADG	SDG	KGK	AQLE	DG	GE	QE	EAK	EP
29	QT	LFF	ME	QD	CP	YI	LLR	EAKPV	TSP	RAD
30	VE	LA	MV	DEAY	DH	AAE	SKKT	MKK	KLED	RM
31	HN	ALEGT	MR	IQ	PH	NVSS	TPE	AC	PT	DI
32	AARA	LGQD	AR	SGK	YA	VFV	LTAVT	KDSSG	EA	TLK
33	ETT	DE	DKDKY	VEVT	ELK	HF	LEGA	TDEE	DEEC	SKAK
34	AAEK	MT	FSKA	ADGK	ARAD	QQ	Q	IH	SKA	LEGA
35	LAYA	LAK	ACK	ACKP	ITVQ	NLEGN				

[a] The peptides are derived from non-redundant dataset based on five iterations of the stepwise discriminant analysis, and listed in the order of their significant in discrimination of the five types of membrane proteins. Of the 356 peptide(s) chosen, 10 are single amino acids, 85 are dipeptides, 59 are tripeptides, 111 are tetrapeptides, and 91 are pentapeptides. Their frequencies in the membrane protein sequences are used in prediction.

of 85.1% as a result of the increment of the length of the peptides used in prediction. However, for GPI-anchored membrane proteins the improvement is modest and reaches a sensitivity of 60.4% with a specificity of 61.1% when the pentapeptides are taken into account. Although these numbers are not impressive, they are still higher than those by the least Hamming distance algorithm [28], lease Euclidean distance algorithm [29] and ProtLock predictor [30].

The extracted sequence features

In the process of stepwise discriminant analysis, some peptide(s) chosen in the former step are not chosen in the later one. The contribution of the removed peptide(s) is usually replaced by the peptide(s) chosen later, whether they are longer or shorter. The variables finally chosen in prediction are 356 elements consisting of 10 single residues, 85 dipeptides, 59 tripeptides, 111 tetrapeptides, and 91 pentapeptides, which are listed in Table 3 in the order of significance in prediction. We can see that the number of peptide(s) remained in Table 3 is much less than $20 + 400 + 8000 + \cdots + 20^6$, but they contain plentiful information and lead to higher prediction accuracy. For example, 'CSST' is the 145-th element, which corresponds to '14' in the first column and '5' in the first row. Clearly, some longer peptides will provide more information in discriminating a membrane protein type, although their frequencies are generally much smaller than those of shorter ones. The number of tetrapeptides (111) is much higher than that of tripeptides (59), which indicates that considerable sequence information is carried by the 4-residue correlation of the amino acids. This is not surprising given the property of 3.6-residue for a turn of helix. Therefore, an algorithm concerning only amino acid or dipeptide composition will not be able to extract efficient information from a sequence.

Analyzing the protein sequences containing the final chosen 261 peptides longer than 3-residue, we find that many of them have membrane type propensities. Further protein motif fingerprint database (PRINTS) [31] searches indicate that many peptides identified contain information of signal peptide, fatty acid or transmembrane helices. For example, the 27th peptide listed in Table 3, 'GADDT', is likely a signal peptide of saturated fatty acid myristate (N-myristoylation) after the initiating Met has been removed [8]. While the 'aaX', (aliphatic–aliphatic–X) motif in the peptides of the 20th, 36th, 53th, 67th, 90th, 109th, 297th, etc. listed in Table 3 likely contains signal information of prenylation [8]. 'ALKSA' (the 11th peptide), is likely a part of the signature for cation transporting ATPase. 'DKDGK' (the 4th peptide), 'AKEA' (the 10th peptide) and 'SKKT' (307th peptide) are likely involved in the signature of out surface proteins. Therefore, the peptides listed in Table 3 contain important high order protein sequence information in determination of the types of membrane proteins. As a result high accuracy has been achieved by using their occurrence frequencies as input parameters in predicting membrane protein types.

Conclusion

Membrane proteins play an important role in a cell. Effective algorithm to predict their types can expedite the understanding of their functions. A multivariate statistical method of extracting the signal or topology features from protein sequences has been described. Based on the stepwise discriminant analysis and lengthening "seeds peptide(s)" method some peptide(s) that carry important signal or signature information have been identified from five types of membrane proteins, which include single amino acid residues, amino acid pairs or dipeptides, tripeptide, tetrapeptides, and pentapeptides. The occurrence frequencies of these peptide(s) in membrane proteins can be used in predicting the membrane protein types. Consequently, an accuracy of 86.5% in the cross-validation test and 99.8% in the resubstitution test has been achieved, respectively, for a non-redundant dataset. The higher prediction accuracy indicates that significant information is extracted from the different types of the membrane protein sequences. The provided method of extracting protein sequence information can be used in the systematic analysis of the great amount of genome sequences and in prediction of the possible functions for membrane proteins.

Acknowledgment

We thank Prof F. S. Ma for valuable discussions. Z.P. Feng has been supported by an APD award from the Australian Research Council.

References

[1] D.D. Thomas, C. Hidalgo, Rotational motion of the sarcoplasmic reticulum Ca^{2+}-ATPase, Proc. Natl. Acad. Sci. USA 75 (1978) 5488–5492.
[2] H. Lodish, D. Baltimore, A. Berk, S.L. Zipursky, P. Matsudaira, J. Darnell, Molecular Cell Biology, Scientific American Books, New York, 1995, Chapter 3..
[3] Z.P. Feng, D.W. Keizer, R.A. Stevenson, S. Yao, J.J. Babon, V.J. Murphy, R.F. Anders, R.S. Norton, Structure and inter-domain interactions of domain II from the blood-stage malarial protein, apical membrane antigen 1, J. Mol. Biol. 350 (2005) 641–656.
[4] Z.P. Feng, X. Zhang, P. Han, N. Arora, R.F. Anders, R.S. Norton, Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes, Mol. Biochem. Parasitol. 150 (2006) 256–267.
[5] P.R. Sanders, L.M. Kats, D.R. Drew, R.A. O'Donnell, M. O'Neill, A.G. Maier, R.L. Coppel, B.S. Crabb, A set of glycosylphosphatidyl inositol-anchored membrane proteins of *Plasmodium falciparum* is refractory to genetic deletion, Infect. Immun. 74 (2006) 4330–4338.
[6] A.S. Opat, C. van Vliet, P.A. Gleeson, Trafficking and localisation of resident Golgi glycosylation enzymes, Biochimie 83 (2001) 63–773.
[7] P.L. Jorgensen, J.R. Jorgensen, P.A. Pedersen, Role of conserved TGDGVND-loop in Mg^{2+} binding, phosphorylation, and energy transfer in Na,K-ATPase, J. Bioenerg. Biomembr. 33 (2001) 367–377.
[8] M.D. Resh, Trafficking and signaling by fatty-acylated and prenylated proteins, Nat. Chem. Biol. 2 (2006) 584–590.

[9] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[10] G.E. Tusnady, I. Simon, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, J. Mol. Biol. 283 (1998) 489–506.
[11] B. Eisenhaber, P. Bork, F. Eisenhaber, Prediction of potential GPI-modification sites in proprotein sequences, J. Mol. Biol. 292 (1999) 741–758.
[12] S.Q. Wang, J. Yang, K.C. Chou, Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition, J. Theor. Biol. 242 (2006) 941–946.
[13] H. Liu, J. Yang, M. Wang, L. Xue, K.C. Chou, Using fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types, Protein J. 24 (2005) 385–389.
[14] H.B. Shen, J. Yang, K.C. Chou, Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition, J. Theor. Biol. 240 (2006) 9–13.
[15] H. Liu, M. Wang, K.C. Chou, Low-frequency Fourier spectrum for predicting membrane protein types, Biochem. Biophys. Res. Commun. 336 (2005) 737–739.
[16] H. Shen, K.C. Chou, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types, Biochem. Biophys. Res. Commun. 334 (2005) 288–292.
[17] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model. 45 (2005) 407–413.
[18] K.C. Chou, Y.D. Cai, Using GO-PseAA predictor to identify membrane proteins and their types, Biochem. Biophys. Res. Commun. 327 (2005) 845–847.
[19] M. Wang, J. Yang, Z.J. Xu, K.C. Chou, SLLE for predicting membrane protein types, J. Theor. Biol. 232 (2005) 7–15.
[20] M. Wang, J. Yang, G.P. Liu, Z.J. Xu, K.C. Chou, Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition, Protein Eng. Des. Sel. 17 (2004) 509–516.
[21] X.G. Yang, Z.P. Feng, Predicting membrane protein types using residue-pair models based on reduced similarity dataset, J. Biomol. Struct. Dyn. 20 (2002) 163–172.
[22] Z.P. Feng, C.T. Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, J. Protein Chem. 19 (2000) 269–275.
[23] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.
[24] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics 18 (2002) 77–82.
[25] M.C. Costanza, A.A. Afifi, Comparison of stopping rules in forward stepwise discriminant analysis, J. Am. Stat. Asso. 74 (1979) 777–785.
[26] R.Y. Luo, Z.P. Feng, J.K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269 (2002) 4219–4225.
[27] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.
[28] P.Y. Chou, in: G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 549–586.
[29] H. Nakashima, K. Nishikawa, T. Ooi, The folding type of a protein is relevant to the amino acid composition, J. Biochem. 99 (1986) 152–162.
[30] J. Cedano, P. Aloy, J.A. P_erez-Pons, E. Querol, Relation between amino acid composition and cellular location of proteins, J. Mol. Biol. 266 (1997) 594–600.
[31] T.K. Attwood, M.E. Beck, A.J. Bleasby, D.J. Parry-Smith, PRINTS—a database of protein motif fingerprints, Nucleic Acids Res. 22 (1994) 3590–3596.