生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2008, April 25; 24(4): 651-658
Chinese Journal of Biotechnology ISSN 1000-3061

研究报告

使用多特征联合变量的支持向量机方法预测外膜蛋白

邹凌云[1], 王正志[1], 王勇献[2]

1 国防科技大学机电工程与自动化学院, 长沙 410073
2 国防科技大学计算机学院, 长沙 410073

摘 要: 外膜蛋白(Outer Membrane Proteins, OMPs)是一类具有重要生物功能的蛋白质, 通过生物信息学方法来预测OMPs能够为预测OMPs的二级和三级结构以及在基因组发现新的OMPs提供帮助。文中提出计算蛋白质序列的氨基酸含量特征、二肽含量特征和加权多阶氨基酸残基指数相关系数特征, 将三类特征组合, 采用支持向量机(Support Vector Machine, SVM)算法来识别OMPs。计算了包括四种残基指数的多种组合特征的识别结果, 并且讨论了相关系数的阶次和权值对预测性能的影响。在数据集上的十倍交叉验证测试和独立性测试结果显示, 组合特征识别方法对OMPs和非OMPs的识别精度最高分别达到96.96%和97.33%, 优于现有的多种方法。在五种细菌基因组内识别OMPs的结果显示, 组合特征方法具有很高的特异性, 并且对PDB数据库中已知结构的OMPs识别准确度超过99%。表明该方法能够作为基因组内筛选OMPs的有效工具。

关键词: 外膜蛋白, 联合特征, 氨基酸指数, 相关系数, 支持向量机

Prediction of Outer Membrane Proteins Using Support Vector Machine with Combined Features

Lingyun Zou[1], Zhengzhi Wang[1], and Yongxian Wang[2]

1 *School of Mechatronics and Automatization, National University of Defence Technology, Changsha* 410073, *China*
2 *School of Computer, National University of Defence Technology, Changsha* 410073, *China*

Abstract: Outer membrane proteins (OMPs) are embedded in the outer membrane of Gram-negative bacteria, mitochondria, and chloroplasts. The cellular location and functional diversity of OMPs makes them an important protein class. Researches on prediction of OMPs by bioinformatics methods can bring helpful methodologies for identifying OMPs from genomic sequences and for the successful prediction of their secondary and tertiary structures. In this paper, three feature classes were calculated from protein sequences: amino acid compositions, dipeptide compositions and weighted amino acid index correlation coefficients. Then, three feature classes were combined and inputted into a support vector machine (SVM) based predictor to identify OMPs from other folding types of proteins. The results of discrimination using several combined features including four amino acid index categories were calculated, and the influence on discrimination accuracy using different correlation coefficients with different orders and weights was discussed. In cross-validated tests and independent tests for identifying OMPs from a dataset of 1087 proteins belonging to all different types of globular and membrane proteins, the method using combined features obtains an overall accuracy of 96.96% and 97.33% respectively. And these results outperform that of other methods in the literature. Using this method, high specificities are shown from the results of identifying OMPs in five bacterial genomes, and over 99% OMPs with known three-dimensional

Received: August 15, 2007; **Accepted:** September 15, 2007
Supported by: the National Natural Science Foundation of China (No. 60603054).
Corresponding author: Lingyun Zou. Tel: +86-731-4574991, E-mail: lyzou@nudt.edu.cn
国家自然科学基金(No. 60603054) 资助。

structures in the PDB database are correctly discriminated. These results indicate that the method is a powerful tool for OMPs discrimination in genomes.

Keywords: outer membrane protein, combined features, amino acid index, correlation coefficient, support vector machine

外膜蛋白(Outer membrane proteins, OMPs)具有重要的生物功能，如非特异性调控、组成运输离子和小分子的通道、控制分子(如麦芽糖、蔗糖分子)通过外膜、参与构成电位调控型阴离子通道等等[1–3]。OMPs通常包含一些由β折叠构成的β链，其中大多数是由8~22 条β链通过反平行排列构成类似于桶状的跨膜结构，称为β桶跨膜蛋白，还有少数则形成一些不完全规则的跨膜结构。OMPs 发现于细菌(革兰氏阴性菌)、线粒体和叶绿体的外膜[4]，其周围的脂质环境使得其结构特征不同于那些具有全β结构的球蛋白(Globular proteins, GPs); 和α螺旋跨膜蛋白(Transmambrane α-helical proteins, TMHs)相比，它们也具有不同的结构模体。对 OMPs 识别方法的研究具有两个重要的应用方向: 预测 OMPs 的跨膜拓扑结构和基因组内识别 OMPs。

近年来，一些基于蛋白质序列特征的 OMPs 识别方法先后被提出来。Gnanasekaran 等人通过比对保守的结构模体来识别 OMPs，准确率为 80%[5]。Wimly 通过统计 OMPs 的疏水残基分布来识别基因组内的 OMPs，识别精度约为 75%[6]。Martelli 等人使用 12 个已知结构的 OMPs 数据作为训练数据，通过隐马尔可夫模型(Hidden markov model, HMM)预测 OMPs，在具有 145 个 OMPs 数据的数据集上的预测精度为 84%[7]。Bagos 等人设计了一个基于 HMM 的算法，在包含 133 个 OMPs 的数据集上获得了89%的识别精度[8]。Natt 等人利用 16 个 OMPs 数据训练一个神经网络预测器，在随机挑 GPs 和 OMPs 混合数据集上的平均识别精度为 90%[9]。Garrow 提出了一个改进的 K 最近邻分类器，利用加权的氨基酸组成和进化信息作为分类特征，获得了 92.5%的识别精度[10]。Park 等人利用氨基酸和二肽含量特征，通过机器学习方法对 OMPs 和非 OMPs 的整体识别精度达到了 94%[11]。Gromiha 等人在氨基酸组成特征上加入氨基酸物理化学性质特征来识别 OMPs，预测精度达到了 94.4%[12]。除了在挑选的数据集上测试以外，人们还提出了在基因组序列内筛选 OMPs 的方法。Zhai 和 Saier 研制了一个基于二级结构、疏水性和两性分子参数的搜索程序，在大肠杆菌基因组内搜索 OMPs，并成功识别出 10 个家族的 OMPs [13]。Bigelow 等人提出了一个基于剖面的 HMM 方法来识别 OMPs，并在 72 个革兰氏阴性菌基因组内搜索到了一些可能的 OMPs[14]。Gromiha 等人在 24 个古生菌、254 个细菌和 149 个真核生物的基因组内进行了大规模的 OMPs 识别，得到的结果有很高的特异性，并对识别结果建立了相应的数据库[15]。

基于机器学习技术来识别 OMPs，其识别精度取决于两个重要方面: 特征选择和分类算法。采用氨基酸组成和二肽组成这两个常用特征，一般就能够获得较好的识别精度。但是为了进一步提高预测精度，人们还提出了利用包括氨基酸物理化学性质、二级结构信息、保守模体在内的更多特征作为分类器的输入。但是，并不是特征越多，分类精度就越高。有些特征组合在一起，反而会降低识别率。如何选择有效的特征组合以获得最好的识别精度，是我们所关心的问题。本文分析了氨基酸组成、二肽组成、多种氨基酸残基指数的相关系数三类特征及其组合对 OMPs 识别精度的影响，提出了具有较高识别性能的特征组合，并采用支持向量机(support vector machine, SVM)算法作为分类算法，在测试数据集上对 OMPs 和非 OMPs 数据进行的交叉验证测试和独立性测试中，分别取得了 96.96%和 97.33%的识别精度，并在 5 种细菌基因组内的分析中，获得了高度特异性的预测结果。

1 材料与方法

1.1 数据集

我们采用了由 Gromiha 和 Suwa 收集的一个包含 1318 条蛋白质数据的数据集[16]，包括 OMPs(377)、TMHs(267)和 GPs(674)三大类蛋白质数据，其中的膜蛋白数据从 PSORT-B 数据库中筛选而来，GPs 数据从 PDB40D_1.37 数据库中筛选而来。Park 等人采用 CD-HIT 程序(http://bioinformatics.org/cd-hit/)对数据集中序列相似度进行了分析，去除了序列相似度

大于 40%的冗余序列[11]，最后得到的数据集组成如下：OMPs(208)，TMHs(206)，GPs(673)，我们称为 MCP1087 数据集。该非冗余数据集可以从下列网址下载: http://www.cbrc.jp/~gromiha/omp/dataset2.html。

1.2 特征选择

我们从蛋白质序列中提取的特征矢量由三部分组成，一是蛋白质的氨基酸含量特征；二是二肽含量特征；三是多阶氨基酸残基指数相关系数特征。

蛋白质序列由 20 种氨基酸残基组成，计算每一种氨基酸残基在蛋白质序列中的含量 f_i，得到 20 个特征[f_1,f_2,…,f_{20}]，蛋白质序列的二肽(氨基酸对)含量组成通过下列公式计算:

$$d_{xy}=\frac{\sum_{i=1}^{N-1}x_i y_{i+1}}{N-1} \tag{1}$$

这里，d_{xy}表示第 x 种氨基酸残基后面联接第 y 种氨基酸的二肽类型的含量，i 表示序列中的第 i 个氨基酸残基，N 表示序列残基总数。这样，由 20 种氨基酸的组合计算得到 400 种二肽的含量特征[d_1d_2,…,d_{400}]。

氨基酸残基指数数据库 AAindex(http://www.genome.ad.jp/dbget/aaindex.html)收集了反映氨基酸不同的物理化学性质的各种残基指数，其 7.0 版中收录了 516 种残基指数。我们利用其中的数据来计算蛋白质序列的氨基酸残基指数相关系数。首先将蛋白质序列映射为数值序列。假设一条蛋白质序列由 L 个氨基酸残基构成，则可表示为：R_1，R_2，…，R_i，…，R_L。其中，R_i 表示第 i 个位置的氨基酸残基。采用氨基酸残基指数数值将蛋白质序列映射为数值序列：h_1，h_2，…，h_i，…，h_L。其中，h_i 对应于 R_i 的某一种残基指数值。计算之前先将数值归一化处理:

$$\hat{h}_i=\frac{h_i-\bar{h}}{\sigma} \tag{2}$$

其中，$\bar{h}$ 和 σ 分别为为 20 种氨基酸的残基指数的均值和标准差，这样归一化处理后的数值序列为:

$$\hat{h}_1,\hat{h}_2,...,\hat{h}_i,...,\hat{h}_L \tag{3}$$

通过下列自相关函数计算序列顺序之间的相关性:

$$\tau_m=\frac{\omega}{L-m}\sum_{j=1}^{L-m}\hat{h}_j\hat{h}_{j+m},m=1,2,...,\varphi \tag{4}$$

其中，φ为相关系数的阶数，$\varphi<L$，如φ=1 时为第一阶序列顺序相关系数，反映了序列中所有连续的氨基酸残基之间的相关性，其他阶次依次类推。ω是我们引入的一个加权系数，为了调节相关系数特征在整个特征矢量中的比重。

结合上述三类特征以后，一条蛋白质序列可以表示为下列特征向量:

$$x_j=[f_{j1},f_{j2},...,f_{j20},d_{j1},d_{j2},...,d_{400},\tau_{j1},\tau_{j2},...,\tau_{j\varphi}] \tag{5}$$

前 20 个为氨基酸组成，中间 400 个为二肽组成，后面为某一种氨基酸残基指数的φ阶相关系数。在本文测试中，对于序列长度不到φ个氨基酸残基的蛋白质序列，将序列长度自动补齐为φ个残基，且补齐的残基在转换为数值序列时其编码值取 0。这样，由一条蛋白质序列得到一个 20+400+φ维的特征向量作为分类器输入。

考虑到 OMPs 和其它蛋白质在溶解性、疏水结构等方面的区别，我们选择了 4 种氨基酸残基指数进行了计算，分别为：改进的 Kyte-Doolittle 疏水值(Modified Kyte-Doolittle hydrophobicity scale, MKD-Hyd)[17]，Ponnuswamy 疏水值(Ponnuswamy hydrophobicity scales, Pon-Hyd)[18]，平均极性(Mean polarity, MP)[19]，溶剂化自由能(Solvation free energy, SFE)[20]，各数值如表 1 所示。

表 1 四种氨基酸指数值

Table 1 Scales of four types of amino acid

Amino acid	Index scales			
	MKD-Hyd	Pon-Hyd	MP	SFE
A	1.10	0.85	−0.06	0.67
R	−5.10	0.20	0.84	−2.1
N	−3.50	−0.48	−0.48	−0.6
D	−3.60	−1.10	−0.80	−1.2
C	2.50	2.10	1.36	0.38
Q	−3.68	−0.42	−0.73	−0.22
E	−3.2	−0.79	−0.77	−0.76
G	−0.64	0	−0.41	0
H	−3.20	0.22	0.49	0.64
I	4.50	3.14	1.31	1.9
L	3.80	1.99	1.21	1.9
K	−4.11	−1.19	−1.18	−0.57
M	1.90	1.42	1.27	2.4
F	2.80	1.69	1.27	2.3
P	−1.90	−1.14	0	1.2
S	−0.50	−0.52	−0.50	0.01
T	−0.70	−0.08	−0.27	0.52
W	−0.46	1.76	0.88	2.6
Y	−1.3	1.37	0.33	1.6
V	4.2	2.53	1.09	1.5

1.3 SVM 分类算法

我们采用 SVM 方法来对输入的各类蛋白质特征向量进行分类。SVM 是基于统计学习理论中的经验风险最小化原则的一种机器学习方法，将其用于模式分类的观点可简单地阐明如下：首先，无论是否为线性的，选择相应的核函数，均可将输入向量映射到一个高维的 Hilbert 空间；其次，用最优化理论方法寻求最优分界面(超平面) 将二类模式分开。对于 SVM 的基本形式，可用一个权重矩阵 w 和一个偏差值 b 来描述，则 SVM 线性判别式为

$$f(x)=sgn(w^Tx+b) \tag{6}$$

对于线性可分的训练样本, SVM 可以找到与最近的训练样本具有最大欧氏距离的超平面。对于线性不可分的训练样本，总的错误率可用松弛变量 ξ_i 来表示，计算超平面等于解下列方程基本优化问题:

$$\left.\begin{aligned}&\min\ V(w,b,\xi)=\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\xi_i\\&\text{st}\quad \forall_{i=1}^{n}: y_i[w^Tx_i+b]\geqslant 1-\xi_i\\&\qquad \forall_{i=1}^{n}: \xi_i>0\end{aligned}\right\} \tag{7}$$

$\Sigma_{i=1}^{n}\xi_i$ 是所有训练样本错误率的上界，常数 C 起控制对错分样本惩罚程度作用，其决策函数形式为:

$$f(x)=sgn(\sum_{i=1}^{n}y_ia_iK(x,x_i)+b) \tag{8}$$

$K(x,x_i)$为核函数, 2 种典型的核函数如下:

$$K(x_i,x_j)=(x_i^Tx_i+1)^d \tag{9}$$

$$K(x_i,x_j)=\exp(-r\left\|x_i-x_j\right\|^2) \tag{10}$$

(9)式为多项式核函数，当 d=1 时变为线性核函数，(10)式为 RBF 核函数。参数 C 和核函数的选择对预测性能有很大的影响。

1.4 性能评价标准

我们在数据集上进行十倍交叉验证测试和独立性测试，并且测试了不同的惩罚因子 C 和核函数下的预测性能，发现所有测试中，采用 RBF 核函数能够获得最好的预测结果，但 C 的取值有所不同。这里使用三种指标来评价预测结果：敏感性(Sensitivity)，特异性(Specificity)，总精度(Overall accuracy)，分别定义为:

$$\text{sensivity}=\frac{\text{TP}}{\text{TP+FN}} \tag{11}$$

$$\text{specificity}=\frac{\text{TN}}{\text{TN+FP}} \tag{12}$$

$$\text{overall accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{13}$$

这里, TP、TN、FP、FN 分别表示预测结果中真阳性、真阴性、假阳性、假阴性的数量。

2 结果和讨论

2.1 数据集上测试结果

我们在 MCP1087 数据集上进行了下列测试: (1)十倍交叉验证测试。将数据集分为 OMP 和 GP、OMP 和 TMH、OMP 和非 OMP(TMH 和 GP)三组数据，分别进行十倍交叉验证测试，测试结果如表 2 所示。表中氨基酸残基含量特征和二肽含量特征分别用 AA 和 Dipe 表示。我们测试了不同阶次和加权值的相关系数特征下的识别结果，发现采用 80 阶相关系数对三类数据的预测结果都是最好的。(2)独立性数据集测试。将数据集分为训练集和测试集，其中训练集数据为: OMPs(158), TMHs(156), GPs(473); 测试集数据为: OMPs(50), TMHs(50), GPs(200)。利用训练集数据训练 SVM，对测试集数据的 OMPs/非 OMPs 进行识别，结果如表 3 所示。

表 2 结果显示，在 OMPs 和 GPs、OMPs 和 TMHs 以及 OMPs 和非 OMPs 三类数据的识别中，只利用氨基酸残基含量特征，其识别精度均是最低的。在这一特征的基础上加入二肽含量特征后, SVM 对于三组数据的识别精度分别提高了 4.09%、2.66%和 3.13%。因为二肽含量特征考虑了连续的氨基酸残基之间的相关性信息，因此提高了预测精度，特别是在识别 OMPs 和 TMH 两类数据上获得了最高的精度 97.09%。氨基酸含量特征和氨基酸残基相关系数相结合，普遍提高了在 OMPs 和 GPs 中识别 OMPs 的精度。这是因为 OMPs 和 GPs 的重要区别在于其周围环境是脂质环境，从而不同物理化学性质的氨基酸残基在 OMPs 序列中的位置和非 OMPs 序列存在差异，这可以反映在不同阶次的残基指数的区别上，尤其是在疏水性上的区别是这两类蛋白的重要差别之一，因而氨基酸含量和 80 阶 Ponnuswamy 疏水值相关系数的组合特征在该类数据的识别上取得了最好的结果，对阳性(OMPs)和阴性(GPs)数据的识别总精度达到 96.59%，分别比氨基酸含量特征、氨基酸含量和二肽含量组合特征的识别精度提高了 6.25%和 2.16%。但是，在氨基酸含

表 2 数据集 **MCP1087** 上十倍交叉验证测试结果

Table 2 Results of 10-fold cross validated tests on the MCP1087 dataset

Features	φ&ω	Overall accuracy/ % (*C*=10)		
		OMPs/ GPs	OMPs/TMHs	OMPs/non-OMPs
AA		90.34	94.43	91.53
AA+Dipe		94.43	97.09	94.66
AA+KD-Hyd	φ=80, ω=10	96.14	96.37	95.57
AA+Pon-Hyd	φ=80, ω=5	95.11	95.64	94.29
	φ=80, ω=10	96.59	95.40	96.13
AA+MP	φ=80, ω=5	94.55	95.40	94.57
	φ=80, ω=10	96.48	95.16	95.76
AA+SFE	φ=80, ω=10	96.36	95.40	95.58
AA+Dipe+MKD-Hyd	φ=80, ω=1	94.43	96.85	94.84
	φ=80, ω=20	96.59	93.46	96.69
AA+Dipe+Pon-Hyd	φ=80, ω=0.5	95.68	96.85	96.32
	φ=80, ω=20	96.48	93.95	96.96
AA+Dipe+MP	φ=80, ω=1	94.20	96.85	94.84
	φ=80, ω=15	96.25	95.64	96.32
	φ=80, ω=20	96.02	96.13	96.59
AA+Dipe+SFE	φ=80, ω=1	94.20	96.85	94.84
	φ=80, ω=20	96.48	94.67	96.78

表 3 数据集 **MCP1087** 上识别 **OMPs** 和非 **OMPs** 的独立性数据集测试结果

Table 3 Results of independent dataset tests for OMPs/non-OMPs discrimination on the MCP1087 dataset

Features	Results (*C*=100)		
	Sensitivity	Specificity	Overall
AA	88%(44/50)	89.6%(224/250)	89.33%
AA+Dipe	84%(42/50)	94.8%(237/250)	93%
AA+Dipe+MKD-Hyd (φ=80, ω=20)	92%(46/50)	95.2%(238/250)	94.67%
AA+Dipe+Pon-Hyd (φ=80, ω=20)	88%(44/50)	98%(245/250)	96.33%
AA+Dipe+MP (φ=80, ω=20)	94%(47/50)	97.6%(244/250)	97%
AA+Dipe+SFE (φ=80, ω=20)	94%(47/50)	98%(245/250)	97.33%

量和二肽含量特征的基础上引入残基指数相关系数特征后，在 OMPs 和 TMHs 中识别 OMPs 的精度反而降低，这可能是因为 OMPs 和 TMHs 两类蛋白均是膜蛋白，所处的环境均为脂质环境，物理化学性质具有相似性，从而氨基酸残基相关系数的差异较小，引入这类特征反而影响了 SVM 的分类性能。在识别 OMPs 和非 OMPs 的交叉验证试验中，采用氨基酸残基含量、二肽含量和 80 阶 Ponnuswamy 疏水值相关系数的组合特征取得了最高的识别精度 96.96%，比仅采用氨基酸含量特征、氨基酸含量和二肽含量组合特征的识别精度分别提高了 5.43%和 2.30%，而与其他相关系数特征的组合方法也不同程度提高了识别精度，这说明引入残基指数相关系数特征能够有效地提高 SVM 分类器对 OMPs 和非 OMPs 的识别能力。

表 3 的结果显示，在识别 OMPs 和非 OMPs 的独立性测试中，利用氨基酸含量、二肽含量、残基指数相关系数的组合特征作为 SVM 输入特征的识别方法，其敏感性、特异性和总体预测精度均高于只利用前两个特征的识别方法。其中，前两个特征和溶剂化自由能特征组合的方法取得了最好的结果，总体识别精度达到了 97.33%，分别比采用氨基酸含量特征、氨基酸含量和二肽含量组合特征的方法提高了 8.0%和 4.33%，这进一步说明了氨基酸残基指数相关系数特征在提高识别 OMPs 的精度上的正面作用。

在和本文相同的数据集上，Park 等人以氨基酸含量和经过筛选的二肽模体为特征，利用 SVM 分类算法，在交叉验证试验中获得了 94.0%的识别精度[11]。Garrow 等人提取蛋白质的氨基酸含量特征，并设计了一个氨基酸性质特征计算公式，联合这两类特征，采用 K 最近邻分类算法来识别 OMPs，在交叉验证测试中获得了 94.4%的识别精度[12]。而本文

方法最高取得了接近 97%的识别精度，表明采用本文提出的组合特征和 SVM 结合的识别方法，通过选择合适的残基指数相关系数阶次和权值，能够进一步提高对 OMPs 的识别精度。

为了进一步验证本文的联合特征和支持向量机结合的预测方法的预测性能，我们对数据集进行了更严格的筛选，采用 BlastClust 工具(http://toolkit.tuebingen.mpg.de/blastclust)剔除了序列相似度大于 20%的序列，得到数据集包含了 118 条 OMPs，186 条 TMHs 和 673 条 GPs，共 977 条蛋白质序列，称为 MCP977 数据集。在该数据集上进行十倍交叉验证测试，结果如表 4 所示。

表 4 的结果显示，采用更严格的 MCP977 数据集，预测结果整体上和采用 MCP1087 数据集上的预测结果相似，氨基酸组成特征的预测精度在各种预测中仍然是最低的，而且各类特征识别 OMPS 和 TMHs 的精度均不同程度下降，但是氨基酸组成和二肽组成联合特征仍然取得了最好的预测精度。这一变化进一步显示出 OMPs 和 TMHs 由于处于相似的脂环境，增加了机器学习方法区分它们的难度。

在 MCP977 数据集上，包含氨基酸残基指数相关系数特征的各种联合特征识别 OMPs 和 GPs，以及识别 OMPs 和非 OMPs 的精度均超过 95%，其中包含疏水特征的联合特征矢量 AA+Dipe+MKD-Hyd 和 AA+Dipe+Pon-Hyd 分别取得了 96.96%和 97.13%的最高精度，这一精度甚至高于在数据集 MCP1087 上取得的结果。这些结果表明本文采用的基于联合特征矢量的 SVM 预测器具有很高的鲁棒性，能够有效的预测 OMPs。

2.2 相关系数特征的阶次和权值对预测性能的影响

由 MCP1087 数据集上的交叉验证测试结果可以看出，残基指数相关系数的阶次和加权值对预测结果有较大的影响，为了得到更好的预测性能，需要选择合适的阶次和加权系数。我们以氨基酸含量、二肽含量和 SFE 残基指数相关系数组合特征为例，计算了在采用不同阶次和不同加权系数 SFE 特征的情况下，对 MCP1087 数据集中的 OMPs 和非 OMPs 数据的十倍交叉验证测试结果，结果如表 5 所示。对表中数据所作的三维曲面图如图 1 所示。

表 4 数据集 **MCP977** 上十倍交叉验证测试结果

Table 4 Results of 10-fold cross validated tests on the MCP977 dataset

Features	φ&ω	Overall accuracy/%(*C*=10)		
		OMPs/GPs	OMPs/TMHs	OMPs/non-OMPs
AA		90.40	90.76	91.91
AA+Dipe		94.43	94.72	94.26
AA+Dipe+MKD-Hyd	φ=80, ω=20	96.96	92.74	96.41
AA+Dipe+Pon-Hyd	φ=80, ω=20	96.84	93.07	97.13
AA+Dipe+MP	φ=80, ω=20	95.82	91.75	96.12
AA+Dipe+SFE	φ=80, ω=20	96.20	91.75	95.80

表 5 氨基酸含量和二肽含量特征与不同阶次和权值的 **SFE** 相关系数特征组合在 **MCP1087** 数据集上识别 **OMPs**/非 **OMPs** 的十倍交叉验证测试结果

Table 5 Results of 10-fold cross validated tests on the MCP1087 dataset for OMPs/non-OMPs discrimination using combined features of amino acid composition, dipeptide composition and SFE correlation coefficients with different orders and weights

φ	ω								
	1	5	10	15	20	25	30	35	40
30	94.84	95.4	95.67	95.58	95.76	96.22	96.13	96.22	96.13
40	94.84	95.58	95.67	95.95	96.04	95.86	95.67	95.58	95.76
50	94.84	95.76	95.95	95.67	95.86	96.13	95.76	95.58	95.49
60	94.84	95.67	96.04	96.04	96.41	96.13	95.76	95.76	95.40
70	94.84	95.49	95.86	96.04	96.41	96.04	95.58	95.21	95.12
80	94.84	95.67	95.95	96.41	96.78	96.32	96.04	95.49	95.03
90	94.84	95.67	95.86	96.13	96.22	95.76	95.76	95.40	95.40
100	94.75	95.40	95.95	96.13	96.13	95.58	95.30	95.30	95.12
120	94.75	95.49	95.86	96.13	95.86	95.40	95.03	94.48	94.20

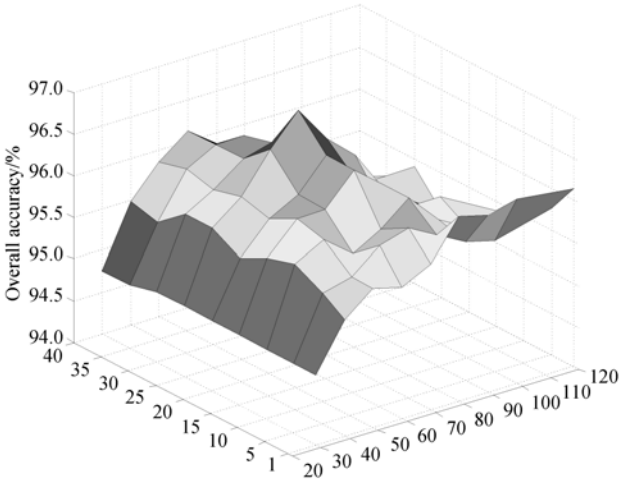

图 1　不同阶次和加权系数值的 SFE 特征对 SVM 识别 OMPs/非 OMPs 的影响

Fig. 1　Influences on discrimination accuracy using different SFE features with different orders and weights

由表 5 的结果和图 1 可以看出，阶次相同的情况下，识别精度普遍先随着权值的增加而提高，达到最高值后又随着权值增加而下降。在不加权的情况下(ω=1)，残基指数相关系数特征与另外两个特征相比，其数值很小，对预测结果影响不大，不能反映该特征对预测性能的影响。当相关系数获得合适的加权，使得数值大小与其它特征的数值大小达到可比较的程度，便能够有效地提高识别精度，这说明本文采取加权处理方式是行之有效的。另外，不同阶次的相关系数反映了蛋白质序列中不同距离上的氨基酸残基的相关模式，这些相关性随着距离不同而变化，这里采用 80 阶的相关系数取得了最好的预测结果。针对不同类型的蛋白质，阶次的取值应该是不同的，需要根据实际计算结果进行选择。

2.3　基因组内识别 OMPs

目前已知三维结构的 OMPs 数量仍然很少，采用有效的方法在基因组筛选 OMPs，能够大大加快对 OMPs 认识进程。一些常规的方法，如利用 BLAST 和 PSI-BLAST 工具搜索同源序列，已经被应用于在基因组内寻找 OMPs，但这两种方法依赖于已知的同源体的数量。由于已知的 OMPs 数量仍然很少，因此上述方法的应用受到限制，并且不能发现新的非同源的 OMPs。为了验证本文方法在基因组内识别 OMPs 的性能，我们下载了五个细菌基因组的完整蛋白质数据，并去除了序列相似度大于 30%的冗余序列，利用 AA+Dipe+SFE(φ=80，ω=20)组合特征和 SVM 算法(以 MCP1087 数据集作为训练集，取 C=100)在这些数据中识别可能的 OMPs，结果如表 6 所示。同时，我们将 Gromiha 等人的方法[15]在这五个基因组内的识别结果也列入表中，作为比较。

表 6 的结果显示，本方法具有很高的特异性，在五个随机选取的细菌基因组内识别出的 OMPs 数量均未超过基因组蛋白质数量的 5%，这与 OMPs 在基因组内数量较少的特点一致。和 Gromiha 等人的方法相比，我们的方法在其中两种基因组内预测到更多的 OMPs，而在另外三种基因组内预测到更少的 OMPs，预测的全部 OMPs 占五个基因组全部蛋白质数量的 3.45%，低于 Gromiha 等人的方法预测的 3.54%，因此我们的方法具有与之相当或者更好的特异性。

在整个 *Escherichia coli* 基因组内的预测结果显示，我们的方法能够正确识别其中已知结构的 11 个家族的 OMPs。在 PSORT-B 2.0 数据库中(http://db.psort.org/)，经试验确认的 *Escherichia coli* OMPs 数据为 57 条，我们的方法将其中 56 条预测为 OMPs，准确率为 98.2%。我们采用 BLAST 工具在最近完成测序的 *Escherichia coli E24377A* 基因组内(包含 4755 条蛋白质序列)搜索这 57 条 OMPs 序列的同源体，结果在 80%以上的序列相似度搜索条件下得到 27 条蛋白质序列。我们的方法将这 27 条序列中的 25 条预测为 OMPs，占 92.6%；并且在该基因组内预测到 210 条 OMPs 序列，其中有 185 条序列不能通过

表 6　联合特征在五个细菌基因组内识别 OMPs 的结果

Table 6　Results of OMPs discrimination in five bacterial genomes using combined features

Genomes	Total proteins	Total identified OMPs	Ratio of OMPs	Gromiha *et al*'s method
Bacillus anthracis Ames 0581	5309	140	2.64%	4.75%(252/5309)
Bacillus licheniformis DSM 13	4196	119	2.84%	3.41%(143/4196)
Clostridium tetani E88	2373	91	3.83%	4.97%(118/2373)
Escherichia coli K12	4237	193	4.56%	2.05%(87/4237)
Shigella flexneri 2a	4182	158	3.78%	2.85%(119/4182)
Overall	20297	701	3.45%	3.54%(718/20297)

BLAST 同源性搜索得到，这其中包含着可能的新的 OMPs。此外，我们在 *Escherichia coli HS* 基因组内(包含 4384 条蛋白质序列)进行了相同的搜索，得到 24 条蛋白质序列，本文方法将其中 23 条序列预测为 OMPs，占 95.8%；并且在该基因组内预测到的 OMPs 数量为 192，其中有 169 条序列不能通过 BLAST 同源搜索得到。上述结果说明本方法比基于同源性搜索的方法具有更好的预测可能的新 OMPs 的能力。我们还从 PDB 数据库(http://www.rcsb.org/pdb/home/home.do)下载了 SCOP Classification 标记为 Transmembrane beta-barrels 的 59 个家族 94 条蛋白质数据，去除序列不完整的两条数据后，得到 92 条数据，其中 91 条数据被正确识别，准确率为 99.0%。在 PDB 数据库中搜索 Classification 标记为 Outer Membrane Protein 的蛋白质数据，共得到 10 个家族的蛋白质数据，本方法正确识别其中 9 个家族为 OMPs，唯一没有识别为 OMP 的一个家族(PDB ID:1bhy)，其 PDB 数据库中三维结构图显示为α/β型结构，而不是β桶结构，这些正例的识别结果则表明本文方法同时具有很高的敏感性。

3 结论

本研究提出了一种新的蛋白质序列组合特征来预测 OMPs，即在提取氨基酸含量特征和二肽含量特征的基础上，增加氨基酸残基指数相关系数特征来提高识别性能。通过选择合适阶次的残基指数相关系数特征，并且对该特征赋予合适的权值，和氨基酸含量以及二肽含量特征组合，能够提高本文所采用的 SVM 分类算法对 OMPs 的识别精度。该方法在数据集交叉验证测试和独立性数据集测试中分别取得了最高 96.96%和 97.33%的识别精度，优于多种现有的 OMPs 识别方法。并且，本文方法在基因组内搜索 OMPs，具有很高的敏感性和特异性，能够用于在基因组内筛选 OMPs。

REFERENCES

[1] Forst D, Welte W, Wacker T, *et al*. Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat Struct Biol*, 1998, **5**: 37–46.

[2] Schulz GE. β-Barrel membrane proteins. *Curr Opin Struct Biol*, 2000, **10**: 443–447.

[3] Wimley WC. The versatile b barrel membrane protein. *Curr Opin Struct Biol*, 2003, **13**: 404–411.

[4] Schulz GE. The structure of bacterial outer membrane proteins. *Biochim Biophys Acta*, 2002, **1565**: 308–317.

[5] Gnanasekaran TV, Peri S, Arockiasamy A, *et al*. Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. *Bioinformatics*, 2000, **16**: 839–842.

[6] Wimley WC. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci*, 2002, **11**: 301–312.

[7] Martelli PL, Fariselli P, Krogh A, *et al*. A sequence-profile-based HMM or predicting and discriminating b-barrel membrane proteins. *Bioinformatics*, 2002, **18**: S46–S53.

[8] Bagos PG, Liakopoulos TD, Spyropoulos IC, *et al*. A hidden Markov model method, capable of predicting and discriminating b-barrel outer membrane proteins. *BMC Bioinformatics*, 2004, **5**: 29.

[9] Natt NK, Kaur H, Raghava GPS. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, 2004, **56**: 11–18.

[10] 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.

[11] Park KJ, Gromiha MM., Horton P, *et al*. Discrimination of outer membrane proteins using support vector machines, *Bioinformatics*, 2005, **21**(23): 4223–4229.

[12] Gromiha MM, Suwa M. Influence of amino acid properties for discriminating outer membrane proteins at better accuracy, *Biochim Biophys Acta*, 2006, **1764**: 1493–1497.

[13] Zhai Y, Saier MH. The beta barrel finder (BBF) program, allowing identification of outer membrane beta barrel proteins encoded within prokaryotic genomes. *Protein Sci*, 2002, **11**: 2196–2207.

[14] Bigelow HR, Petrey DS, Liu JF, *et al*. Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res*, 2004, **32**: 2566–2577.

[15] Gromiha1 MM, Yabuki Y, Kundu S, *et al*. TMBETA-GENOME: database for annotated beta-barrel membrane proteins in genomic sequences, *Nucleic Acids Res*, 2007, **35**(Database issue): D314–D316.

[16] Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, 2005, **21**: 961–968.

[17] Juretic D, Lucic B, Zucic D, *et al*. Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. *Theoretical and Computational Chemistry*, 1997, **5**: 405–445.

[18] Ponnuswamy PK, Gromiha MM. Hydrophobic characteristics of folded proteins. *Prog Biophys Mol Biol*, 1993, **59**: 57–103.

[19] Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol and neutral aqueous solution. *Biochemistry*, 1988, **27**: 1664–1670.

[20] Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. *Nature*, 1986, **319**: 199–203.