基于支持向量机的蛋白质同源寡聚体分类研究*

张绍武[1)]** 潘 泉[1)] 陈润生[2)] 张洪才[1)]

（[1)]西北工业大学自动控制系，西安 710072；[2)]中国科学院生物物理研究所，北京 100101）

摘要 基于支持向量机和贝叶斯方法，从蛋白质一级序列出发对蛋白质同源二聚体、同源三聚体、同源四聚体、同源六聚体进行分类研究，结果表明：基于支持向量机，采用“一对多”和“一对一”策略，其分类总精度分别为77.36%和93.43%，分别比基于贝叶斯协方差判别法的分类总精度50.64%提高26.72和42.79个百分点. 从而说明支持向量机可用于蛋白质同源寡聚体分类，且是一种非常有效的方法. 对于多类蛋白质同源寡聚体分类，基于相同的机器学习方法（如支持向量机），采用“一对一”策略比“一对多”效果好. 同时亦表明蛋白质同源寡聚体一级序列包含四级结构信息.

关键词 支持向量机，贝叶斯协方差判别法，分类，同源二聚体，同源三聚体，同源四聚体，同源六聚体

学科分类号 Q617

蛋白质四级结构按亚基的种类和数目可分为同源寡聚体和异源多聚体，例如同源二聚体，同源三聚体，同源四聚体和同源六聚体等. 对于蛋白质四级结构一般是通过生物学实验测定的，但需要昂贵的仪器，且实验中还会遇到一些困难. 随着人类基因组计划（HGP）在全世界范围内的顺利展开，人类已获得了大量的蛋白质序列，因而与实验方法相结合，利用蛋白质一级结构序列信息预测蛋白质空间结构将扮演重要的角色.

Garian[1]用决策树的方法对蛋白质的同源二聚体和非同源二聚体进行了分类研究. 关于多类同源寡聚体的分类研究目前还未见有报道，本文拟采用支持向量机[2,3]和贝叶斯协方差判别法[4,5]对蛋白质同源二聚体、同源三聚体、同源四聚体和同源六聚体进行分类研究.

支持向量机（SVM）[2,3]是近年来国际上新兴的一种机器学习方法，由于出色的学习性能，该技术已成为当前的研究热点. 且已被成功地应用于生物信息学中基因微阵列表达模式[6]、蛋白质家族[7]、转录起始点[8]、蛋白质亚细胞定位[9~11]、蛋白质折叠[12]等方面.

1 材料与方法

1.1 数据库

数据库由1 568条同源寡聚蛋白质序列构成，其中包含914个同源二聚体（homo-dimer，2EM），139个同源三聚体（homo-trimer，3EM），407个同源四聚体（homo-tetramer，4EM）和108个同源六聚体（homo-hexamer，6EM）四类蛋白质一级结构序列. 为了排除膜蛋白和其他特异蛋白，该数据库仅限于原核生物和细胞质的同源寡聚蛋白内，可从SWISS-PROT数据库中选出.

1.2 支持向量机

支持向量机（support vector machine，SVM）是Vapnik等[2,3]根据统计学习理论提出的一种新的机器学习方法，其最大特点是根据Vapnik的结构风险最小化原则，尽量提高学习的泛化能力，即由有限的训练集样本得到的小误差仍能够保证对独立的测试集小的误差. 应用支持向量机进行分类研究的基本思想可简述为：首先将输入空间的样本通过某种非线性函数关系映射到一个特征空间中（维数可能较高），在此特征空间中构造最优分类超平面，使两类样本（可推广到多类样本）在此特征空间中可分. 映射函数仅与低维输入向量和特征空间的点积有关，此映射函数点积可用一核函数来替代，从而可避免“维数灾难”，因而可解决高维特征问题. 支持向量机总是寻找全局最优解，且可防止过学习. 其判别函数为

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} a_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b\right)$$

$k(\boldsymbol{x}_i, \boldsymbol{x}_j)$称为核函数，核函数的选取应使其为特征空间的一个点积，即存在函数$\boldsymbol{\Phi}$，使$\boldsymbol{\Phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\Phi}(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. 业已证明，核函数$k(\boldsymbol{x}_i, \boldsymbol{x}_j)$只要满足Mercer条件即可满足上述要

* 西北工业大学博士创新基金资助.

** 通讯联系人.

Tel：029-8495954，E-mail：shaowuzhang@hotmail.com

收稿日期：2003-03-28，接受日期：2003-04-26

求[13]. 常用的核函数有：

多项式核函数

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d$$

径向基核函数

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\gamma \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2)$$

Sigmoid 核函数

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \tanh[b(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + c]$$

本文用 Joachims[14]编写的 SVM^{light}程序，该程序是由 C 语言编写的，对于学术用途者可免费从以下网址下载：http://ais.gmd.de/~thorsten/svm_light/

1.3 特征向量提取

Nakashima 等[15]，Klein[16]和 Chou 等[17,18]研究表明，蛋白质的折叠信息与氨基酸组成有明显的关联性，这样蛋白质序列可表示为如下特征向量：

$$\boldsymbol{x}_j^\rho = [\boldsymbol{x}_{j,1}^\rho, \boldsymbol{x}_{j,2}^\rho, \cdots, \boldsymbol{x}_{j,20}^\rho]^T$$

式中 ρ 表示蛋白质的类别，j 表示每类蛋白质的样本个数，$\boldsymbol{x}_{j,i}^\rho$（$i=1$，2，…，20）表示 ρ 类蛋白质第 j 个蛋白质序列中 i 种基本氨基酸出现的频率数，特征向量中元素的顺序按照 20 种基本氨基酸的字母顺序排列.

1.4 分类系统检验

对分类结果的评价基于两种检验方法，一种是 Jackknife 检验，另一种为 k-fold cross-validation 检验，这两种检验是较为客观和严格的方法. 在 Jackknife 检验方法中，每一蛋白质依次从数据库中取出作为测试蛋白，而剩余的蛋白质作为训练集. 在 k-fold cross-validation（k-CV）检验方法中，随机将数据库分为 k 个子集合，依次取出一个子集作为测试集，而其余的 $k-1$ 个子集合作为训练集，此过程循环 k 次.

总分类精度 Q，每类样本的分类精度 Q_i 和 Matthews 相关系数 MCC 分别定义为：

$$Q = \sum_{i=1}^{\rho} p(i)/N \qquad Q_i = p(i)/obs(i)$$

$$MCC(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}$$

上式中 N 为样本总数，ρ 为样本类别数，obs（i）为类别 i 的样本数，p（i）为 i 类样本的正确分类数，n（i）为非 i 类样本的正确分类数，u（i）为 i 类样本中被错分为其他类别的样本数，o（i）为其他类别的样本被错分为 i 类样本的数目.

2 结果与讨论

机器学习方法用于多类问题，一般将其分解为两类问题进行处理，主要有“一对多”（one-versus-rest，1-v-r）和“一对一”（all-versus-all，a-v-a）2 种策略. 下面我们基于支持向量机方法采用此 2 种策略对四类蛋白质同源寡聚体进行分类研究.

2.1 核函数和其参数的选取

由于支持向量机的核函数及其参数的选取对分类结果有一定的影响，我们对此进行了研究. 选取多项式和 Sigmoid 二核函数，通过计算我们发现运算不是速度慢就是发散，因而不对其进行详细研究. 对于径向基核函数，我们先确定 $\gamma=0.04$，然后选惩罚系数 C = 1，10，100，1 000，10 000，100 000，1 000 000，在 10CV 检验下，采用“一对多”策略对四类蛋白质同源寡聚体进行分类，其结果为：当 C 取值大于 10 后，对四类蛋白质同源寡聚体的分类精度没有太大影响，因而我们选 SVM^{light}程序的默认值 $C=1\,000$.

当 C 值确定后，选取 γ = 0.5，0.1，0.05，0.04，0.03，0.02，在 10CV 检验下，采用“一对多”策略对四类蛋白质同源寡聚体进行分类，其结果为：γ 值对分类结果有一定的影响，在一定范围内（0.1～0.02）对总分类精度影响不大，但对每类样本的分类精度有影响，另外当 γ 小于 0.02 后运算发散，经过综合考虑我们选 $\gamma=0.04$.

2.2 结果

基于贝叶斯协方差判别法和采用“一对多”和“一对一”2 种策略构建支持向量机分类器对四类蛋白质同源寡聚体进行分类，其结果见表 1.

Table 1 Classifying results for protein homo-oligomers sequences using SVM and Bayes covariant discriminant methods by Jackknife test

	Bayes		SVM		
			1-v-r		a-v-a
	Q_i/%	MCC	Q_i/%	MCC	Q_i/%
2EM	42.67	0.28	89.93	0.58	99.23
3EM	62.59	0.45	57.55	0.69	74.82
4EM	65.36	0.29	64.13	0.57	96.81
6EM	47.22	0.26	46.30	0.59	55.56
Q/%	50.64	–	77.36	–	93.43

从表1可知：Jackknife检验下，基于支持向量机采用“一对多”策略的总分类精度、同源二聚体、同源三聚体、同源四聚体和同源六聚体的分类精度分别为77.36%，89.93%，57.55%，64.13%，46.3%；采用“一对一”策略的总分类精度、同源二聚体、同源三聚体、同源四聚体和同源六聚体的分类精度分别为93.43%，99.23%，74.82%，96.81%，55.56%；基于贝叶斯协方差判别法的总分类精度、同源二聚体、同源三聚体、同源四聚体和同源六聚体的分类精度分别为50.64%，42.67%，62.59%，65.36%，47.22%. 基于支持向量机，采用“一对多”和“一对一”策略的分类总精度分别比基于贝叶斯协方差判别法的分类总精度提高26.72和42.79个百分点，从而说明对于蛋白质多类同源寡聚体，支持向量机是一非常有效的分类方法.

“一对一”策略的分类精度明显高于“一对多”策略，其Jackknife检验的总分类精度、同源二聚体、同源三聚体、同源四聚体和同源六聚体的分类精度分别比“一对多”策略提高16.07、9.3、17.27、32.68、9.26个百分点. 从而说明在蛋白质同源寡聚体分类问题上，基于相同的分类方法（支持向量机）采用“一对一”策略较大家常用的“一对多”策略效果好，从而亦验证了Ding的结论[12].

2.3 分类系统的可信度指数

用机器学习方法对蛋白质同源寡聚体进行分类，其分类的可信度是非常重要的. Rost等[19~21]用神经网络方法进行分类研究时，引入了一个可信度指数（reliability index，*RI*）来评价分类系统，该指数定义为网络输出最大值与次最大值之差. 这一简单的思想同样可以引入支持向量机分类系统[9]，在支持向量机多类分类系统中采用“一对多”策略，*RI*定义为：

$$RI = \begin{cases} 0 & \text{if} \quad diff < 0.2 \\ \mathrm{INT}(5diff) & \text{if} \quad 0.2 \leqslant diff < 1.8 \\ 9 & \text{if} \quad diff \geqslant 1.8 \end{cases}$$

式中“INT”表示取整数，“*diff*”表示支持向量机输出最大值与次输出最大值之差. 对蛋白质同源寡聚体分类系统的评价如图1、图2所示.

图1、图2可清晰地反映*RI*值、期望分类精度及蛋白质序列样本之间的关系，*RI*值越大其期望分类精度越高. 从图1可知，具有某*RI*值蛋白质序列的期望分类精度及具有这一*RI*值的蛋白质序列占总蛋白质序列的百分比，例如，*RI* = 5的蛋白质序列的期望分类精度为81.71%，大约11%的蛋白质序列样本的*RI*值等于5. 对于给定的某*RI*阈值，从图2可知期望分类平均精度与蛋白质序列之间的关系，例如，大约62%的蛋白质序列*RI* ≥ 5，*RI* ≥ 5的蛋白质序列的平均分类精度为87.74%.

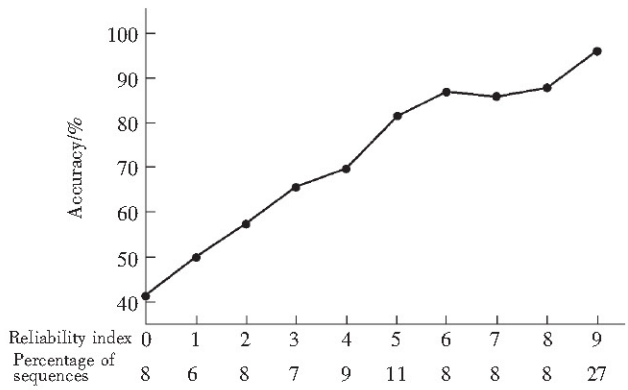

Fig. 1 Expected prediction accuracy with a reliability index equal to a given value

The fractions of sequences that are predicted with *RI* = 0，1，2，…，9 are also given in jackknife test with one-versus-rest

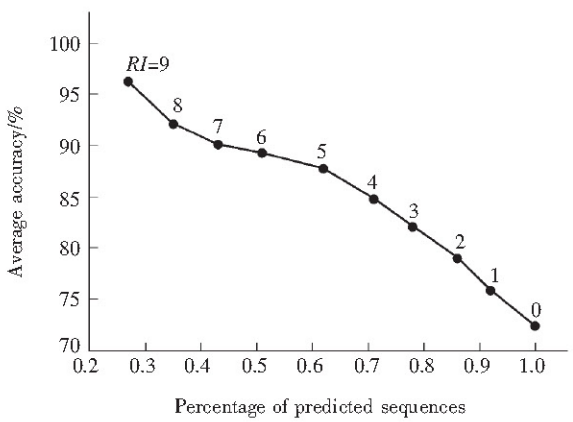

Fig. 2 Average prediction accuracy with a reliablity index above a given cut-off

2.4 样本数对分类结果的影响

为了检验样本数对分类结果有何影响，我们将914个蛋白质同源二聚体序列平分为两个子集，每一子集包含457个蛋白质同源二聚体序列，分别与139个蛋白质同源三聚体序列、407个蛋白质同源四聚体序列、108个蛋白质同源六聚体序列组成2个新的数据库data1、data2. 在10CV检验下，采

用“一对多”策略对其进行分类，其分类结果及与原数据库（data）的分类结果比较见表2.

Table 2 The classifying accuracy for datasets which have different protein numbers in 10CV test using RBF kernel SVM

Data	Q_i/%	Data1	Q_i/%	Data2	Q_i/%
2EM	89.28	2EM	73.26	2EM	77.68
3EM	51.08	3EM	59.88	3EM	58.99
4EM	62.65	4EM	72.48	4EM	73.22
6EM	42.59	6EM	47.22	6EM	45.37
Q/%	75.77	Q/%	69.85	Q/%	70.57

从表2可知，蛋白质的样本数对分类结果有一定的影响. 一般来说样本数愈大，其分类精度越高，例如，914个同源二聚体样本的分类精度为89.28%，当被均分为二个子集后，其精度分别下降为73.26% 和77.68%. 另外类样本间的样本数不平衡对分类结果亦有影响，例如数据库data1、data2中同源二聚体（457个）与同源四聚体（407个）的样本数差别不是太大，其精度差0.78%和4.46%分别比原数据库data中的精度差26.63%低了许多. data1、data2数据库中同源三聚体、同源六聚体与同源二聚体的分类精度差与原数据库data相比亦有不同程度的缩小.

2.5 蛋白质同源寡聚体与单体的分类结果

我们从SWISS-PROT数据库中选出1 652条蛋白质单体序列，这些蛋白质序列仍限于原核生物和细胞质内，与1 568条蛋白质同源寡聚体序列组成一数据库. 采用支持向量机方法对其进行分类，其结果见表3.

Table 3 The classifying results of homo-oligomers and monomers in 10CV test using RBF kernel SVM

Homo-oligomers accuracy%	Monomers accuracy%	Total accuracy%	*MCC*
85.97	80.21	83.01	0.6621

$C=1\,000$，$\gamma=0.08$.

从表3可知支持向量机可以较好地将蛋白质同源寡聚体和单体分开，同时意味着同源寡聚蛋白质一级序列包含四级结构信息.

综合表1、表2和表3可知，同源寡聚蛋白质一级序列包含蛋白质四级结构信息，其特征向量（氨基酸成分）的确表示了埋藏在缔合亚基作用部位接触表面的基本信息. 另外亦说明支持向量机用于蛋白质同源寡聚体分类是一有效的方法，对于蛋白质同源寡聚体分类，基于同样的机器学习方法（如支持向量机）采用“一对一”策略比采用“一对多”策略更为有效.

本文我们仅以氨基酸成分构成的特征向量表示蛋白质同源寡聚体一级序列，而没有考虑氨基酸序列次序及氨基酸物理化学特性的影响，这必将遗失一些包含在蛋白质序列内的信息. 若考虑这些信息，分类精度会得到进一步的改善，这将是我们下一步研究的任务.

参 考 文 献

1 Garian R. Prediction of quaternary structure from primary structure. Bioinformatics, 2001, **17** (6): 551 ~556

2 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995. 1 ~188

3 Vapnik V. Statistical Learning Theory. New York: Wiely, 1998. 1 ~736

4 Chou K C, Elrod D W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun, 1998, **252** (1): 63 ~68

5 Chou K C, Elrod D W. Protein subcellular location prediction. Protein Eng, 1999, **12** (2): 107 ~108

6 Brown M, Grundy W, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA, 2000, **97** (1): 262 ~267

7 Jaakkola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies. Proceedings of the 7th International Conference on Intelligent systems for Molecular Biology. Menlo Park, CA: AAAI Press, 1999. 149 ~158

8 Zien A, Ratsch G, Mika S, *et al*. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics, 2000, **16** (9): 799 ~807

9 Hua S J, Sun Z R. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 2001, **17** (8): 721 ~728

10 Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for prediction of protein subcellular location. Mol Cell Biol Res Commun, 2000, **4** (4): 230 ~233

11 Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem, 2002, **84** (2): 343 ~348

12 Ding C H Q, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 2001, **17** (4): 349 ~358

13 Courant R, Hilbert D. Methods of mathematical physics. New York: Wildy-Interscience, 1953. 1 ~560

14 Joachims T. Making large-scale SVM learning practiceal. In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods-support Vector Learning. Cambridge, MA: MIT Press, 1999. 1 ~376

15 Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem, 1986, **99** (1):

153 ~162
16 Klein P. Prediction of protein structural class by discriminat analysis. Biochem Biophys Acta, 1986, **874** (2): 205 ~275
17 Chou K C, Maggiora G M. Domain structure prediction. Protein Eng, 1998, **11** (7): 523 ~538
18 Chou K C. A key driving force in determination of protein structural classes. Biochem Biophys Res Commun, 1999, **264** (1): 216 ~224
19 Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. J Mol Biol, 1993, **232** (2): 584 ~599
20 Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res, 1998, **26** (9): 2230 ~2236
21 Emanuelsson O, Nielsen H, Brunak S, *et al*. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol, 2000, **300** (4): 1005 ~1016

Classification of Protein Homo-oligomers Using Support Vector Machine*

ZHANG Shao-Wu[1)**], PAN Quan[1)], CHEN Run-Sheng[2)], ZHANG Hong-Cai[1)]

([1)] *Department of Automatic Control, Northwestern Polytechnical University, Xi'an* 710072, China;

[2)] *Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China)

Abstract The homo-dimer, homo-trimer, homo-tetramer and homo-hexamer of protein were classified using both of support vector machine and Bayes covariant discriminant methods. It was found that the total accuracies of "one-versus-rest" and "all-versus-all" are 77.36% and 93.43% respectively using support vector machine in jackknife test, which are 26.72 and 42.79 percentile higher respectively than that of Bayes covariant discriminant method in the same test. These results show that the support vector machine is a specially effective method for classifying the higher protein homo-oligomers from protein primary sequences. Using "all-versus-all" policy is better than "one-versus-rest" policy for classifying homo-oligomers based on the same machine learning method (such as support vector machine). And it was also indicated that the primary sequences of homo-oligomeric proteins contain quaternary information.

Key words support vector machine, Bayes covariant discriminant, classification, homo-dimer, homo-trimer, homo-tetramer, homo-hexamer

* This work was supported by a grant from The Doctor Innovation Grant of Northwestern Polytechnical University.

** Corresponding author. Tel: 86-29-8495954, E-mail: shaowuzhang@hotmail.com

Received: March 28, 2003 Accepted: April 26, 2003