生物物理学报　第十九卷　第二期　二○○三年六月
ACTA BIOPHYSICA SINICA　Vol.19　No.2　Jun. 2003

基于支持向量机和贝叶斯方法的蛋白质四级结构分类研究

张绍武，　潘　泉，　张洪才，　张云龙，　王海瑜
（西北工业大学自动控制系，　陕西　西安　710072）

摘要：用支持向量机和贝叶斯两种方法对蛋白质四级结构进行分类研究。结果表明，基于支持向量机的分类结果最好，其 10CV 检验的总分类精度、正样本正确预测率、Matthes 相关系数和假阳性率分别为 74.2%、84.6%、0.474、38.9%；基于贝叶斯的分类结果没有支持向量机的分类结果好，但其 10CV 检验的假阳性率最低(15.9%)。这些结果说明同源寡聚蛋白质一级序列包含四级结构信息，同时特征向量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。

关键词：支持向量机；贝叶斯；蛋白质四级结构；亚基
中图分类号：Q617　　**文献标识码**：A　　**文章编号**：1000-6737(2003)02-0171-05

蛋白质所具有的功能在很大程度上取决于其空间结构，因此对蛋白质空间结构的研究有着极其重要的意义。目前，研究蛋白质空间结构主要有 X 射线晶体学方法、多维核磁共振方法和基于结构知识的蛋白质结构预测方法。X 射线晶体学方法是迄今为止研究蛋白质结构最有效的方法，但是所需的蛋白质晶体难以培养，且晶体结构测定的周期较长。多维核磁共振方法可以直接测定蛋白质在溶液中的构象，但由于对样品的需要量大、纯度高，被测定的蛋白质的分子量一般不超过 2 万等，因而也受到很大限制。1961 年 Anfinsen 等[1]通过实验证明蛋白质的氨基酸序列决定其三维结构，另外随着人类基因组计划(HGP)在世界范围内的顺利展开，人类已获得了大量的蛋白质序列，而且其增长速度异常迅速，因而利用蛋白质的一级结构所提供的氨基酸序列信息来进行高级结构预测是目前比较流行且较经济的方法。

1958 年 Bernal[2]首次提出了蛋白质四级结构概念，四级结构被看成是蛋白质一级结构、二级结构和三级结构的延伸，是指寡聚蛋白质中亚基的种类、数目、空间排布以及亚基之间的相互作用。寡聚蛋白质广泛地参与物质代谢、信号传导、染色体复制等各种生命活动，从生物大分子功能进化的角度来讲，相对于单体蛋白质，寡聚蛋白质具有许多优势[3,4]。首先，寡聚蛋白质具有更加复杂的结构，可执行更为复杂的功能；其次，可通过相同和不相同亚基之间的协同作用，实现对酶活性的调节；第三，可把中间代谢途径中的各种酶分子集合在一起，提高催化效率，避免中间产物的浪费；第四，亚基的缔合增加了整个分子的稳定性。因此对蛋白质四级结构的研究有重要的生物学意义。

Robert[5]用决策树的方法对四级结构进行了分类研究，得到了较好的结果。本文采用支持向量机和贝叶斯两种方法从蛋白质的一级结构出发对四级结构进行分类研究。

支持向量机(SVM)[6,7]是一类新型的机器学习方法，由于出色的学习性能，该技术已成为当前国际机器学习的研究热点。已被成功地应用于基因微阵列表达模式的分类[8]、蛋白质家族的分类[9]、转录起始点的识别[10]等方面。

1　材料和方法

1.1　数据库

数据库由 1639 个同源寡聚蛋白质序列构成，其中有 914 个同源二聚体和 725 个非同源二聚体。该数据库是由 Robert[5]从 SWISS-PROT 数据库中挑选出。

收稿日期：2002-12-03
基金项目：西北工业大学博士创新基金
作者简介：张绍武，1964 年生，副研究员，博士生，
电话:(029)8495954，E-mail：zsw9957_cn@sina.com

1.2 支持向量机

支持向量机是 Vapnik 等[6,7]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。另外，由于支持向量机算法是一个凸优化问题，因此局部最优解一定是全局最优解，可防止过学习。这些特点是其它学习算法，如神经网络学习算法所不及的。对于分类问题，支持向量机算法可简述为将输入空间中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)，使两类样本(可推广到多类样本)在此特征空间中线性可分，并寻找样本在此特征空间中的最优线性分类超平面。其判别函数为

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{k}\alpha_i\gamma_i k(x,x_i)+b\right)$$

$k(x_i,x_j)$称为核函数，核函数的选取应使其为特征空间的一个点积，既存在函数 Φ，使 $\Phi(x_i)\times\Phi(x_j)=k(x_i,x_j)$。已证明，核函数 $k(x_i,x_j)$只要满足 Mercer[11]条件即可满足上述要求。常用的核函数有：

多项式核函数(polynomial function)

$k(x_i,x_j)=(x_i\times x_j+1)^d$

径向基核函数（radial basis function, RBF）

$k(x_i,x_j)=\exp(-g\parallel x_i-x_j\parallel^2)$

Sigmoid 核函数 (sigmoid function)

$k(x_i,x_j)=\tanh[b(x_i\times x_j)+c]$

本文用 Joachims[12]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 http://ais.gmd.de/~thorsten/svm_light/ 网址上下载。

1.3 Bayes 方法

若每一样本用 n 维向量 x_j^ρ表示, 则

$x_j^\rho=[x_{j1}^\rho,x_{j2}^\rho,\cdots,x_{jn}^\rho]^{\mathrm{T}}$, $j=1,2,\cdots,N_\rho$; $\rho=1,2,\cdots,l$。N_ρ 表示第 ρ 类样本的数目，l 表示类别总数。每类样本的平均向量为： $X^\rho=[x_1^\rho,x_2^\rho,\cdots,x_n^\rho]$

$$x_i^\rho=\frac{1}{N_\rho}\sum_{j=1}^{N_\rho}x_{ji}^\rho \qquad i=1,2,\cdots,n \quad \rho=1,2,\cdots,l$$

每类样本的协方差阵为：

$$\boldsymbol{C}^\rho=\begin{bmatrix}C_{11}^\rho & C_{12}^\rho & \cdots & C_{1n}^\rho\\ C_{21}^\rho & C_{22}^\rho & \cdots & C_{2n}^\rho\\ \vdots & \vdots & \vdots & \vdots\\ C_{n1}^\rho & C_{n2}^\rho & \cdots & C_{nn}^\rho\end{bmatrix}$$

$$C_{ji}^\rho=\frac{1}{N_\rho-1}\sum_{s=1}^{N_\rho}(x_{sj}^\rho-x_j^\rho)(x_{si}^\rho-x_i^\rho) \qquad \text{且 } C_{ji}^\rho=C_{ij}^\rho$$

若测试样本以 X 向量表示，则 X^ρ 和 X 之间的相似程度可用下列 Bayes 函数表示[13,14]：

$F(X,X^\rho)=D^2(X,X^\rho)+\ln|\boldsymbol{C}^\rho|$

其中 $D^2(X,X^\rho)=(X-X^\rho)^T(\boldsymbol{C}^\rho)^{-1}(X-X^\rho)$

于是对给定的测试样本可以根据最小 Bayes 决策函数进行判定：

$F(X,X^\tau)=\min\{F(X,X^1),F(X,X^2),\cdots,F(X,X^l)\}$

若 $\tau=\rho$, ($\rho=1,2,\cdots,l$)，则将待测试样本判为 ρ 类。

1.4 特征向量提取

Nishikawa 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：$x_j^\rho=[x_{j,1}^\rho,x_{j,2}^\rho,\cdots,x_{j,20}^\rho]^T$, 式中 ρ 表示蛋白质的类别，j 表示每类蛋白质的样本个数，$x_{j,i}^\rho$ (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。

1.5 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 检验方法，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 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蛋白集，而其余的 k–1 个子集合作为训练集，此过程循环 k 次。总预测精度(Q)、正样本正确预测率(TPR)、假阳性率(FPR)和 Matthes 相关系数(MCC)分别定义为：

$Q=(TP+TN)/N$; $TPR=TP/(TP+FN)$; $FPR=FP/(FP+TN)$

$MCC=(TP\times TN-FP\times FN)/$

$$\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}$$

其中 N 为样本总数，TP 为正确分类的正样本数，TN 为正确分类的负样本数，FP 为本来是负样本却被划分为正样本数（即假正样本数），FN 为本来是正样本却被划分为负样本数（即假负样本数）。

2 结果与讨论

2.1 支持向量参数的选取

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。首先选取多项式核函数，取 d=1, 2, …, 10，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 g=0.05，然后选惩罚系数 C=1，10，100，500，1000，10000，100000，1000000，用 10-fold cross validation (10CV)检验方法对蛋白质四级结构进行分类，其分类结果见表 1。

Table 1 The confusion matrix of 10-fold cross validation test with different C using RBF kernel (g=0.05) support vector machine

	C=1	C=10	C=100	C=500	C=1000	C=10000	C=100000	C=1000000
TP	810	772	772	772	773	778	772	772
TN	383	443	443	443	443	443	443	443
FP	342	282	282	282	282	282	282	282
FN	104	142	142	142	141	136	142	142

从表 1 我们可看出当 C 值大于 1 后，C 值对分类结果影响不大，故我们就选取 SVM^{light} 程序的默认值 C＝1000。

当 C 值确定后，我们选取不同的 g 值进行分类，发现有些 g 值使运算发散，我们选取了一些不使运算发散的 g 值进行分类，其分类结果见表 2。

从表 2 可以看出 g 值的选取对分类结果有一定的影响，以 g=0.05 的分类结果较好。

Table 2 The confusion matrix of 10-fold cross validation test with different g using RBF kernel (C=1000) support vector machines

	g=0.03	g=0.05	g=0.06	g=0.1	g=0.5
TP	730	773	788	846	268
TN	477	443	426	339	725
FP	248	282	299	386	646
FN	184	141	126	68	0

2.2　支持向量机和贝叶斯方法分类结果

基于支持向量机和贝叶斯方法分类结果及与 Robert 基于决策树方法的分类结果比较见表 3、4。

从表 3、4 我们可以看出，基于支持向量机的分类结果最好，其 10CV 检验的总分类精度 74.2%、正样本正确预测率 84.6%和 Matthes 相关系数 0.474 分别比决策树的总分类精度 69.9%、正样本正确预测率 78.1%和 Matthes 相关系数 0.386 提高 4.3%、6.5%、0.088；10CV 法的假阳性率 38.9%比决策树的假阳性率 40.2%降低 1.3%。基于贝叶斯的分类结果没有支持向量机和决策树的分类结果好，但其 10CV 检验的假阳性率 15.9%最低，分别比支持向量机的假阳性率 38.9%和决策树方法的假阳性率 40.2%低 23%、24.3%。这些结果说明同源寡聚蛋白质一级序列包含四级结构信息，同时特征向量看来能表示埋藏在缔合亚基作用部位接触表面的基本信息。

Table 3 The confusion matrix of Decision-tree method, support vector machines (SVM) method and Bayes method

	Decision tree	SVM (g=0.05, C=1000)		Bayes	
		10CV	Jackknife	10CV	Jackknife
TP	714	773	783	319	326
TN	433	443	518	610	640
FP	292	282	207	115	85
FN	200	141	131	595	588

Robert 用决策树的方法预测蛋白质四级结构

Table 4 Some common performance measures derived from the confusion matrix

	Decision tree	SVM (g=0.05, C=1000)		Bayes	
		10CV	Jackknife	10CV	Jackknife
Sensitivity	0.781	0.846	0.857	0.349	0.357
Specificity	0.597	0.611	0.714	0.841	0.883
Positive predictive rate	0.709	0.733	0.791	0.735	0.793
Negative predictive rate	0.684	0.759	0.798	0.506	0.521
Overall accuracy	0.699	0.742	0.794	0.567	0.411
TP rate	0.781	0.846	0.857	0.349	0.357
FP rate	0.402	0.389	0.286	0.159	0.117
Mathews correlation coefficient	0.386	0.474	0.580	0.214	0.274

时，不仅考虑了蛋白质序列信息，而且还考虑了氨基酸的物理、化学和生化特性，而我们用支持向量机和贝叶斯方法时，仅考虑了蛋白质序列信息，其支持向量机的分类结果就比决策树方法的分类结果好，从而说明支持向量机方法用于蛋白质四级结构分类是一种非常有效的方法。另外，基于贝叶斯方法的假阳性率最低，如果将支持向量机方法和贝叶斯方法通过信息融合理论在决策层进行融合，再考虑氨基酸的物理、化学和生化特性，一定会得到更好的分类效果。

参考文献：

[1] Anfinsen CB, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain [J]. *Proc Natl Acad Sci USA*, 1961,47: 1309~1314.

[2] Klotz IM, Darnall DW, Langerman NR. The protein, 3rd edition[M]. New York: Academic Press, 1975,1:293~411.

[3] 阎隆飞, 孙之荣. 蛋白质分子结构[M]. 北京: 清华大学出版社, 2000.

[4] Price NC. Assembly of multi-subunit structure[M]. New York: Oxford University Press, 1994.

[5] Robert G. Prediction of quaternary structure from primary structure[J]. *Bioinformatics*, 2001,17:551~556.

[6] Vapnik V. The nature of statistical learning theory[M]. New York: Springer, 1995.

[7] Vapnik V. Statistical learning theory[M]. New York: Wiely, 1998.

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

[9] Jaakkola T, Diekhans M, Haussler D. Using the fisher kernel method to detect remote protein homologies[A]. Proceedings of the 7th international conference on intelligent systems for molecular biology[C]. Menlo Park, CA: AAAI Press, 1999. 149~158.

[10] Zien A, Ratsch G, Mika S, et al. Engineering support vector machine kernels that recognize translation initiation sites[J]. *Bioinformatics*. 2000,16:799~807.

[11] Courant R, Hilbert D. Methods of mathematical physics[M]. New York: Wildy-Interscience, 1953.

[12] Joachims T. Making large-scale SVM learning practiceal[A]. In: Scholkopf B, Burges C, and Smola A. (eds), Advances in kernel methods-support vector learning[M]. Cambridge, MA:MIT Press, 1999.

[13] Duda RO, Hart PE. Pattern classification and scene analysis [M]. New York: John Wiley & Sons, 1973.

[14] Liu W, Chou KC. Prediction of protein structure classes by modified mahalanobis discriminant algorithm [J]. *J Protein Chem*, 1998,17:209~217.

[15] Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition [J]. *J Biochem*, 1986,99:152~162.

[16] Klein P. Prediction of protein structural class by discriminat analysis biochem[J]. *Biophys Acta*, 1986,876:205~275.

[17] Chou KC, Maggiora GM. Domain structural prediction [J]. *Protein Engineering*, 1998,11:523~538.

[18] Chou KC. A key driving force in determination of protein structural classes [J]. *Biochemical and Biophysical Research Communication*, 1999,264:216~224.

CLASSIFICATION OF QUATERNARY STRUCTURE USING SUPPORT VECTOR MACHINES AND BAYES METHODS

ZHANG Shao-wu, PAN Quan, ZHANG Hong-cai, ZHANG Yun-long, WANG Hai-yu

(Department of Automatic Control, Northwestern Polytechnical University, Shanxi Xi'an 710072, China)

Abstract: The quaternary structure was classified using support vector machine method and Bayes method. It was found that the result of using support vector machine is the best, using 10-fold cross-validation test, the overall accuracy, true positive rate, Mattew's correlation coefficient and false negative rate are 74.2%, 84.6%, 0.474, 38.9% respectively; the result of Bayes method is not so good as that of the support vector machine method, the false negative rate of using 10-fold cross-validation test is the smallest. Those results show that the primary sequences of homo-oligomeric proteins contain quaternary information. The feature vectors appear to capture essential information about the composition and hydrophobicity of the residues in the surface patches that are buried in the interfaces of associated subunits. And they also show that the support vector machines is a specially effective method.

Key Words: Support vector machines; Bayes; Protein quaternary structure; Subunits