

文章编号: 1000-5013(2008)02-0236-05

# PCA-SVM 模型在几丁质酶最适温度建模中的应用

林 毅, 蔡福营, 袁宇熹, 张光亚

( 华侨大学 材料科学与工程学院, 福建 泉州 362021)

摘要: 采用主成分分析法(PCA)对样本数据集进行预处理,将得到的新样本数据集输入支持向量机(SVM),藉助均匀设计(UD),构建几丁质酶氨基酸组成和最适温度的数学模型.当径向基核函数的 3 个参数,惩罚系数  $C$  为 10,  $\epsilon$  为 0.5,  $\gamma$  为 5 时,模型对温度拟合的平均绝对百分比误差为 5.06%,预测的平均绝对误差为 1.83  $^{\circ}\text{C}$ ,说明具有良好的预测效果且优于神经网络的预测结果.

关键词: 几丁质酶; 数学模型; 主成分分析; 支持向量机; 最适温度

中图分类号: Q 556 $^{+}$ .2; Q 141

文献标识码: A

几丁质酶(Chitinase, EC3.2.1.14)是能够催化水解 N-乙酰-D-葡萄糖胺糖苷键的酶.在自然界中,几丁质酶在碳和氮的循环中扮演着重要的角色,它存在于多种物种体内,包括人、细菌、真菌、病毒、线虫、昆虫、鱼等.几丁质酶在工业上有重要的应用,主要是降解壳聚糖为低聚物,此外,几丁质酶还有杀虫活性和抗菌作用<sup>[1]</sup>.工业应用的几丁质酶的最适温度为 30~70  $^{\circ}\text{C}$ , pH 值为 4~8<sup>[2]</sup>.近年来,有两种方法可获得耐碱、耐热的几丁质酶.一种方法是通过从极端环境中筛选几丁质酶产生菌株;另一种方法是对几丁质酶进行遗传改造.随着理性定向进化<sup>[3]</sup>和非理性定向进化技术<sup>[4]</sup>的发展,又提出了一种半理性的定向进化<sup>[5-6]</sup>技术.本文利用几丁质酶的序列信息及最适温度,利用主成分分析的支持向量机,建立了氨基酸组成和最适温度之间的数学模型.

## 1 材料与方法

几丁质酶蛋白质序列数据,均来源于 NCBI(美国国立生物技术信息中心)上的蛋白质数据库.26 个几丁质酶 ID 号分别为 P32823, BAC53628, AAK69033, AAF23368, AAO22144, AAC23715, JC7996, BAA34922, AAK69033, AAA98644, AAK26395, Q9FRV1, AAC09387, 2DBTC, Q05638, BA C99074, BAA88833, BAA88834, BAC76622, BAA88835, AAY99632, AAM93195, AAL01886, AAL46648, AAG12973, BAA36460. Nakashima<sup>[7]</sup>, Klein<sup>[8]</sup>和 Chou 等<sup>[9-10]</sup>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.鉴于几丁质酶的相对分子量差别很大,所以用几丁质酶的 20 种氨基酸组成和氨基酸残基数作为输入数据,对应的最适温度为支持向量机的输出数据,则几丁质酶蛋白序列的特征向量可表示为

$$X = [x_i^T | x_1, x_2, \dots, x_{20}, x_{21}]^T. \quad (1)$$

上式中,  $X$  表示蛋白质序列的特征值,  $x_i^T$  为蛋白质序列中氨基酸的特征向量,  $x_i$  为第  $T$  个蛋白质序列第  $i$  ( $i = 1, \dots, 20$ ) 种氨基酸出现的频率数,  $x_{21}$  为蛋白质序列氨基酸的个数,  $T$  表示蛋白质序列的个数,特征向量中元素的顺序按照 20 种氨基酸的字母顺序排列.所有几丁质酶的氨基酸组成分析由 Bioedit 软件完成,主成分分析由 SPSS10.0 完成,支持向量机是由 Thorsten Joachims 用 C 语言编写的.

收稿日期: 2007-07-06

作者简介: 林 毅 (1976-), 男, 副教授, 主要从事应用与环境微生物的研究. E-mail: lyhxm@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(40601046); 福建省高等学校新世纪优秀人才计划支持项目(2006); 福建省自然科学基金资助项目(B0510011)

2 结果与分析

2.1 氨基酸组成的主成分分析

原始数据经主成分分析(PCA)后得到的特征值及累计方差贡献率,如表 1 所示.表中, $P_c$  为主成分, $E_v$  为特征值,  $\varphi_{VA}$  为贡献率,  $\varphi_{cu}$  为累计贡献率. 分析表 1 中的数据,选择显著水平  $\alpha=95\%$ ,只挑选前

表 1 主成分特征值及累计方差贡献率

Tab.1 Principal component and their explaining variances

$P_c$	$E_v$	$\varphi_{VA}/\%$	$\varphi_{cu}/\%$	$P_c$	$E_v$	$\varphi_{VA}/\%$	$\varphi_{cu}/\%$
1	5.21	24.79	24.79	12	0.30	1.42	97.31
2	4.19	19.97	44.76	13	0.18	0.83	98.14
3	2.84	13.51	58.27	14	0.14	0.67	98.81
4	2.19	10.45	68.71	15	0.08	0.38	99.19
5	1.50	7.14	75.85	16	0.06	0.27	99.46
6	1.07	5.12	80.96	17	0.05	0.25	99.72
7	0.92	4.40	85.36	18	0.03	0.16	99.87
8	0.80	3.80	89.16	19	0.02	0.10	99.98
9	0.60	2.86	92.01	20	0.01	0.02	100.00
10	0.434	2.07	94.08	21	0.00	0.00	100.00
11	0.38	1.80	95.88				

11 个主成分即可代表原始数据中蕴涵的绝大部分信息. 11 个主成分和原 21 个变量之间的关系(限于篇幅,仅写出前 3 个主成分)为

$$P_{c1}=0.200A+0.063C+0.116D+0.146E-0.212F-0.004G+0.267H-0.319I-0.271K-0.059L-0.225M-0.327N+0.311P+0.034Q+0.344R-0.265S+0.204T+0.194V+0.150W-0.236Y+0.159n$$

$$P_{c2}=-0.326A-0.333C+0.386D+0.363E-0.0342F-0.289G+0.117H+0.115I+0.302K+0.128L-0.012M-0.035N+0.053P-0.317Q+0.043R-0.143S-0.133T+0.101V+0.044W+0.186Y+0.301n$$

$$P_{c3}=-0.014A-0.038C+0.162D+0.114E-0.363F+0.029G-0.351H-0.060I-0.030K-0.341L+0.045M+0.100N-0.269P+0.249Q-0.312R-0.008S+0.369T+0.350V+0.019W+0.105Y+0.260n.$$

上式中, $n$  表示几丁质酶含有的氨基酸个数向量.

各氨基酸与 11 个主成分之间的关系,如表 2 所示. 为简单起见,表 2 中氨基酸正负相关系数保留 1

表 2 几丁质酶氨基酸与各主成分的关系

Tab.2 Meaning of principle components in chitinase

正相关氨基酸							负相关氨基酸						
$P_{c1}$	0.2A	0.3H	0.3P	0.3R	0.2T	0.2V	0.2F	0.3I	0.3K	0.2M	0.3N	0.3S	0.2Y
$P_{c2}$		0.4D	0.4E	0.3K	0.2Y			0.3A	0.3C	0.3G	0.3Q		
$P_{c3}$		0.2D	0.2Q	0.4T	0.4V			0.4F	0.4H	0.3L	0.3P	0.3R	
$P_{c4}$	0.3C	0.3E	0.2G	0.2K	0.2N	0.5W	0.2Y	0.2A	0.3I	0.4L	0.2M	0.2V	
$P_{c5}$			0.2D	0.6G	0.5M			0.2A	0.3F	0.3N	0.2V		
$P_{c6}$			0.3L	0.6S	0.2W			0.2I	0.3M	0.2T	0.4Y		
$P_{c7}$		0.4C	0.4F	0.2I	0.4V		0.3A	0.2H	0.3L	0.3N	0.2W	0.2Y	
$P_{c8}$		0.3I	0.3N	0.3P	0.5Q	0.2W		0.2A	0.2D	0.2S	0.4T	0.2Y	
$P_{c9}$	0.2C	0.2G	0.2H	0.2N	0.2P	0.2S	0.3Y	0.3A	0.2M	0.2V	0.6W		
$P_{c10}$			0.3C	0.4H	0.3L	0.3T			0.4A	0.2G	0.5P		
$P_{c11}$	0.2A	0.2C	0.4Q	0.2H	0.2M	0.3Q	0.2V	0.2Y		0.2G	0.4I	0.6T	

位小数,且仅列出相关系数绝对值大于等于 0.2 的氨基酸.与蛋白质结构数据库(MMPD)中所报道的几丁质酶蛋白质三级进行了比较,发现主成分分析的前 7 个主成分所代表的几丁质酶的二级结构,分别

为无规卷曲、转角、折叠、转角、转角、螺旋和折叠. 这与几丁质酶结构特征基本吻合, 但略有差异, 可能与所选择的样本有关.

### 2.2 支持向量机模型结构的优化

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<sup>[11-13]</sup>. 选取多项式和 Sigmoid 二核函数, 通过计算发现, 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, 有 3 个参数, 分别为惩罚系数  $C$ ,  $\varepsilon$  和  $\gamma$ , 常规的参数选取方法是“一对多”策略, 就是先确定一个值, 令这个值不变再确定另外一个值, 最后找出一组最优的参数, 不过这样的方法很笨拙, 而且也体现不出各因子之间的交互影响. 本文设计的 3 因素 15 水平均匀设计表来优化参数, 如表 3 所示. 表 3 中,  $e_{RMS}$ ,  $e_{MAP}$ ,  $e_{MA}$  分别为平均绝对百分比误差, 均方极误差, 平均绝对误差,  $N$  为运算次数.

表 3 均匀设计表  
Tab. 3 Uniform design

$N$	$C$	$\varepsilon$	$\gamma$	$e_{RMS}$	$e_{MAP}$	$e_{MA}$
1	10.00	0.50	5.00	0.05	1.80	4.35
2	0.10	0.80	1.00	0.06	2.58	5.93
3	0.005	0.40	0.09	0.11	4.74	9.23
4	1.00	0.10	0.50	0.07	3.00	6.89
5	50 000.00	0.20	0.10	0.11	4.53	9.05
6	0.01	0.01	0.90	0.06	2.66	6.08
7	100.00	0.000 01	0.30	0.07	3.22	7.46
8	10 000.00	0.05	0.01	0.20	8.61	11.75
9	50.00	0.15	0.001	0.23	9.94	12.34
10	0.50	0.60	0.005	0.22	9.26	11.94
11	0.05	0.000 1	0.03	0.17	6.97	10.91
12	5.00	0.005	0.07	0.13	5.26	9.67
13	500.00	1.00	0.05	0.14	5.99	10.23
14	5 000.00	0.70	0.10	0.11	4.53	9.05
15	1000.00	0.001	1.50	0.06	2.30	5.34

计算结果显示, 当  $C$  值为 10,  $\varepsilon$  为 0.5,  $\gamma$  值为 5 时, 对温度预测的平均绝对百分比误差为 5%, 均方根误差为 1.8, 平均绝对误差为 4.35, 具有比神经网络更好的拟合效果(图 1). 后续训练及测试均采用上述参数( $C=10$ ,  $\varepsilon=0.5$ ,  $\gamma=5$ ).

### 2.3 主成分分析-支持向量机模型预测

对支持向量机而言, 由于训练样本集的大小有限, 训练后对训练集外输入的响应如何, 直接决定了支持向量机的性能. 对预测结果的评价基于两种较为客观和严格的检验方法, 一种是 Jackknife 检验, 另一种为  $k$ -fold cross validation 检验. 在 Jackknife 检验方法中, 每一种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, 依次取出一个子集作为测试集, 而其余的  $k-1$  个子集作为训练集, 此过程循环  $k$  次. 由于数据量较少, 为了提高检验的灵敏度, 采用 Jackknife 检验方法, 每次从 26 组数据中取出 25 个序列作为训练数据, 留出一个作检测, 依次循环, 共进行 26 次循环测试. BP 神经网络(BPNN)、支持向量机(SVM)和主成分分析-支持向量机(PCA-SVM)的测试结果, 如表 4 所示.  $n$  为循环次数. 由于篇幅所限, 只列出 5 组较好的结果和 5 组较差的结果. 从表 4 中可以看出, 主成分分析-支持向量机模型的拟合值总体上要好于预测值, 训练和测试的平均绝对百分比误差分别为 0.05 和 0.24, 训练和测试的平均绝对误差为 1.83 和 9.94. 没有经主成分分析优化过的支持向量机模型, 其训练和测试的平均绝对百分比误差分别为 0.06 和 0.26, 训练和测试的平均绝对百分比误差分别为 2.61 和 10.76. 显然, 主成分分析在支持向量机模型的数据优化中起了重要的作用.

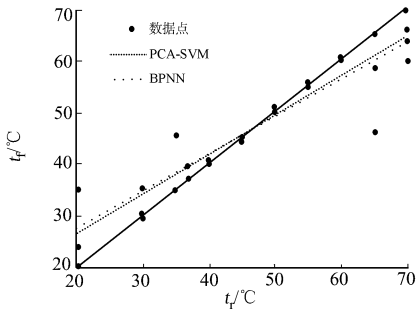


图 1 结构优化后支持向量机的拟合值  
Fig. 1 The fitting temperature of support vector machine being optimized

表 4 3 种模型的测试结果

Tab.4 Results of the cross validations of three models

<i>n</i>	方式	<i>e</i> <sub>MAP</sub>			<i>e</i> <sub>MA</sub>		
		BPNN	SVM	PCA-SVM	BPNN	SVM	PCA-SVM
1	训练值	0.06	0.06	0.05	2.66	2.59	1.82
	测试值	0.01	0.07	0.05	0.59	3.39	2.72
2	训练值	0.08	0.06	0.05	3.68	2.46	1.68
	测试值	0.51	0.58	0.60	15.44	17.51	18.06
7	训练值	0.07	0.05	0.05	3.02	1.89	1.79
	测试值	0.54	0.31	0.31	34.78	20.10	20.04
9	训练值	0.08	0.07	0.05	3.23	2.67	1.91
	测试值	0.46	0.34	0.04	16.06	11.74	1.25
10	训练值	0.06	0.05	0.03	2.45	2.73	1.53
	测试值	2.17	1.35	1.40	43.43	27.02	27.91
20	训练值	0.07	0.06	0.05	2.75	2.68	1.96
	测试值	0.08	0.23	0.22	4.66	13.97	13.35
21	训练值	0.08	0.06	0.05	3.14	2.65	1.90
	测试值	0.01	0.16	0.15	0.42	8.72	8.05
22	训练值	0.06	0.06	0.05	2.39	2.53	1.82
	测试值	0.47	0.18	0.20	18.99	7.23	7.82
25	训练值	0.07	0.06	0.05	2.92	2.55	1.84
	测试值	0.47	0.36	0.37	16.33	12.52	13.05
26	训练值	0.06	0.06	0.05	2.42	2.68	1.96
	测试值	0.15	0.23	0.22	8.91	13.97	13.35
平均值	训练值	0.07	0.06	0.05	3.00	2.61	1.83
	测试值	0.36	0.26	0.24	14.60	10.76	9.94

2.4 3 种预测模型的比较

参考文献[ 14] 的研究结果, 选择 1 个隐含层的神经网络, BP 神经网络的训练误差仍设为 0.01, 其运算次数为 1 000, 用均匀设计方法优化 BP 神经网络的 4 个参数: 学习速率、动态参数、Sigmoid 参数和隐含层结点数. 当 4 个参数分别为 0.09, 0.4, 0.98 和 10 时, BP 神经网络具有最佳的拟合结果. 后续训练及测试均采用上述参数. 同样, 用 Jackknife 检验方法来检验 BP 神经网络的测试结果, 测试结果如表 4 所示. 其 26 个样本的训练和测试的平均绝对百分比误差的平均值为 0.07 和 0.36, 而支持向量机训练和测试的平均绝对百分比误差的平均值为 0.06 和 0.26, 都比主成分分析-支持向量机模型的结果略差些, 如图 2 所示. 26 个测试样本预测结果的平均绝对误差为 14.6, 高于支持向量机模型的 10.76, 更高于主成分分析-支持向量机的预测结果 9.94.

从图 2 可以看出, 用 BP 神经网络预测几丁质酶最适温度, 结果比较差, 其预测结果不稳定; 而支持向量机模型的出的结果要好的多, 其预测结果浮动较小, 且大部分预测值接近于真实的实验值. 经过主成分分析处优化输入数据后的支持向量机, 其预测值明显比没用主成分分析优化数据的支持向量机模型更接近实验值. 这大大提高了模型的运算速度和测试精度.

3 结束语

本文利用主成分分析法对样本集进行预处理, 利用均匀设计对其拓扑结构进行了优化, 大大提高了支持向量机的学习速率和性能. 利用几丁质酶的晶体数据, 结合多序列比对等手段, 可寻找出有利和不利于提高该酶最适温度的可能位点, 然后有目的地利用仿真软件进行随机突变. 利用基于本文所得数学模型的计算机软件进行虚拟筛选, 可减轻筛选工作量, 提高效率. 尽管本文采用了均匀设计的方法对支持向量机的结构进行了优化, 但在各因素水平的选择上仍带有一定的随意性. 如果经过精心的选择, 支

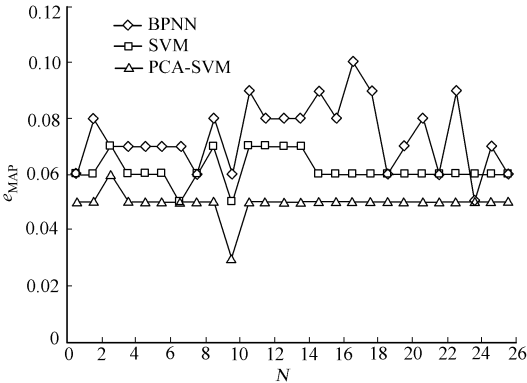


图 2 3 种模型的比较

Fig.2 The comparison between model

持向量机的检测效果还会有所改善. 由于本文仅考虑了 20 种氨基酸的频率分布和氨基酸的个数, 排除了其他影响因素, 这是一种最简单的情形. 同时, 样本中噪声的影响也不可忽视, 对于进一步提高该模型质量的相关研究仍需要逐步深入, 所得结果仍需要实验进一步验证.

## 参考文献:

- [ 1 ] BU SSIVK A P, VAN EIJK M. The Biology of the gaucher cell: The cradle of human chitinase[ J]. A Survey of Cell Biology, 2006, 23( 2 ) : 71-128.
- [ 2 ] 蒋红彬, 蒋千里. 几丁质酶的研究概况[ J]. 山东科学, 2000, 13( 3 ) : 41-45.
- [ 3 ] CHANG Ming Chun, LAI Pei Lin, WU Mei Li. Biochemical characterization and site directed mutational analysis of the double chitin binding domain from chitinase 92 of *Aeromonas hydrophila* JP101[ J]. FEMS Microbiology Letters, 2004, 232( 5 ) : 1-61.
- [ 4 ] ALLEN S. Enzyme functionality: Design, engineering, and screening[ M]. New York: Marcel Dekker, 2004: 1-712.
- [ 5 ] MILDVAN A S. Inverse thinking about double mutants of enzymes[ J]. Biochemistry, 2004, 43( 2 ) : 14517-14520.
- [ 6 ] ROBERT J H, JORG B, MARIE L A. Combining computational and experimental screening for rapid optimization of protein properties[ J]. Proc Natl Acad Sci, 2002, 99( 3 ) : 15926-15931.
- [ 7 ] NAKASHIMA H, NISHIKAWA K, OOI T. The folding type of a protein is relevant to the amino acid composition[ J]. J Biochem, 1986, 99( 1 ) : 153-162.
- [ 8 ] KLEIN P. Prediction of protein structural class by discriminate analysis[ J]. Biochem Biophys Acta, 1986, 874( 2 ) : 205-275.
- [ 9 ] CHOU K C, MAGGIORA G M. Domain structure prediction[ J]. Protein Eng, 1998, 11( 7 ) : 523-538.
- [ 10 ] CHOU K C. A key driving force in determination of protein structural classes[ J]. Biochem Biophys Res Commun, 1999, 264( 1 ) : 216-224.
- [ 11 ] VAPNIK V. Statistical learning theory[ M]. New York: Wiley, 1998: 1-736.
- [ 12 ] CORTES C, VAPNIK V. Support vector machine networks[ J]. Machine Learning, 1995, 20( 4 ) : 273-297.
- [ 13 ] FANG K T. The uniform design application of number theoretic methods in experimental design[ J]. Acta Math Appl Sin, 1980, 66( 3 ) : 363-372.
- [ 14 ] 张光亚, 方柏山. 木聚糖酶氨基酸组成与其最适 pH 的神经网络模型[ J]. 生物工程学报, 2005, 21( 4 ) : 658-661.

# A Uniform Design Based PCA-SVM Model for Predicting Optimum Temperature in Chitinase

LIN Yi, CAI Fu-ying, YUAN Yu-xi, ZHANG Guang-ya

( College of Material Science and Engineering, Huaqiao University, Quanzhou 362021, China )

**Abstract:** The principal component analysis was applied to the data processing in training sets, the new principal components were then used as input data of support vector machine model. A prediction model for optimum temperature of chitinase was established based on uniform design. When the regularized constant  $C$ ,  $\epsilon$  and  $\gamma$  were 10, 0.5 and 5, respectively, the calculated temperature fitted the reported optimum temperature of chitinase very well and the mean absolute percent error (MAPEs) was 5.06%. At the same time, the predicted temperature fitted the reported optimum temperature well and the mean absolute error (MAE) was 1.83 °C. It was superior in fittings and predictions compared to the model based on back propagation neural network.

**Keywords:** principle component analysis; support vector machine; chitinase; optimum temperature

(责任编辑: 黄仲一 英文审校: 陈国华)