基于均匀设计的主成分分析-支持向量机模型及其在几丁质酶最适 pH 建模中的应用

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

林 毅*，蔡福营，袁宇熹，张光亚

LIN Yi*，CAI Fu-Ying，YUAN Yu-Xi and ZHANG Guang-Ya

华侨大学生物工程与技术系，工业生物技术福建省高校重点实验室，泉州 362021

Department of Bioengineering & Biotechnology，Huaqiao University，Key Laboratory of Industrial Biotechnology of Fujian Province University，Quanzhou 362021，*China*

摘 要 采用主成分分析法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，籍均匀设计，构建了几丁质酶氨基酸组成和最适 pH 的数学模型。当惩罚系数 C 为 10，epsilon 值为 0.7，Gamma 值为 0.5，模型对 pH 值拟合的平均绝对百分比误差为 3.76%，同时具有良好的预测效果，预测的平均绝对误差为 0.42 个 pH 单位。该方法比用 BP 神经网络方法效果更佳。

关键词 主成分分析，支持向量机，几丁质酶，最适 pH，均匀设计

中图分类号 Q811.4 **文献标识码** A **文章编号** 1000-3061(2007)03-0514-06

Abstract The principal component analysis (PCA) was applied to the data processing in training sets, the new principal components were then used as input data for support vector machine model. A prediction model for optimum pH of chitinase was established based on uniform design. When The regularized constant C, epsilon and Gamma were 10, 0.7 and 0.5 respectively, the calculated pHs fitted the reported optimum pHs of chitinase very well and the MAPEs (Mean Absolute Percent Error) was 3.76%. At the same time, the predicted pHs fitted the reported optimum pHs well and the MAE (Mean Absolute Error) was 0.42 pH unit. It was superior in fittings and predictions compared to the model based on back propagation (BP) neural network.

Key words principal component analysis (PCA), support vector machine (SVM), chitinase, optimum pH, uniform design

几丁质酶(chitinase，EC321144)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。它存在于包括人、细菌、真菌、病毒、线虫、昆虫以及鱼等不同的物种体内。几丁质酶在工业上有重要的应用，主要是降解几丁质为低聚物，低聚物在医药和食品行业都有重要的应用。此外几丁质酶还有杀虫活性和抗病作用[1]。工业应用的几丁质酶的最适 pH 和温度分别为：pH 4～8，30～70℃[2]。用于杀虫的几丁质酶需要耐偏碱的 pH 环境，同时要求在较高的温度条件下保持热稳定性，这是因为昆虫肠的 pH 环境偏碱，而田间施用时由于阳光照射下会形成高

Received：October 26，2006；Accepted：December 20，2006.

This work was supported by the grants from the National Natural Sciences Foundation of China (No. 40601046), the Program for New Century Excellent Talents in Fujian Province University, and the Natural Science Foundation of Fujian (No. B0510011).

* Corresponding author. Tel：+86-595-22692031；Email：lyhxm@hqu.edu.cn

国家自然科学基金(No.40601046)、福建省高等学校新世纪优秀人才支持计划资助基金和福建省自然科学基金(No.B0510011)资助。

温环境。从搜集的几丁质酶数据(如表 1)可以明显看出,可用的几丁质酶大多是中性偏酸,最适温度大多在 50℃ 以下。近 20 年来有两种方法可获得耐碱耐热的几丁质酶：一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[3],后者是一个更好的选择。近年来,伴随着理性定向进化[4]和非理性定向进化[5]技术的发展,又提出了一种半理性的定向进化[6]技术,所谓半理性的定向进化就是在未知蛋白质三维结构情况下对蛋白质某个推测的位点进行定点突变。

本文利用几丁质酶的序列信息及其最适 pH 数据,基于主成分分析的支持向量机(Support Vector Machine,SVM),建立了几丁质酶氨基酸组成和其最适 pH 之间的数学模型。所得模型具有较高的拟合和测试精度,可用于几丁质酶定向改造过程中的虚拟筛选。

1　材料与方法

1.1　数据来源

几丁质酶氨基酸序列数据均来源于 Swiss-Prot (Release 49.2 of 07-Mar-2006),Swiss-Prot 是一个非冗余库。26 个几丁质酶 ID 号及其最适 pH 和最适温度如表 1 所示：

表 1　几丁质酶

Table 1　Chitinase

ID	pH_{opt}	T_{opt}	ID	pH_{opt}	T_{opt}
P32823	8	50	Q05638	7.3	55
BAC53628	6	30	BAC99074	6	40
AAK69033	5.8	35	BAA88833	5.5	60
AAF23368	8	40	BAA88834	5.5	70
AAO22144	6	70	BAC76622	6	50
AAC23715	6	37	BAA88835	4	60
JC7996	6.5	65	AAY99632	4.5	55
BAA34922	6	45	AAM93195	4.3	40
AAK69033	5.8	35	AAL01886	6	45
AAA98644	7	20	AAL46648	6.5	50
AAK26395	4	40	AAG12973	8	35
Q9FRV1	5	40	BAA36460	6	60
AAC09387	5.5	45	2DBTC	5	55

ID, the accession number of chitinase in Swiss-Prot; pH_{obs}, the optimum pH found in the literature at which the relative chitinase has the maximum activity; T_{opt}, the optimum tempreture found in the literature at which the relative chitinase has the maximum activity.

Nakashima[7], Klein[8] 和 Chou 等[9,10] 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,鉴于几丁质酶的分子量差别很大,所以我们用几丁质酶的 20 种氨基酸组成和氨基酸残基数为输入数据,其对应最适 pH 为支持向量机的输出数据。这样几丁质酶蛋白序列可表示为如下特征向量：

$$X = [\vec{x}_i \mid x_1, x_2, \cdots, x_{20}, x_{21}]^T \quad (1)$$

式中 X 表示蛋白质序列的特征值,$\vec{x}_i$ 表示蛋白质序列中氨基酸的特征向量,x_i 表示第 T 个蛋白质序列第 $i(i = 1, \cdots, 20)$ 种氨基酸出现的频率数,x_{21} 表示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,对于学术用途者可免费从以下网址下载：http://www.cs.ucl.ac.uk/staff/M.Sewell/winsvm/winSVM.exe。

1.2　基于主成分分析的支持向量机

1.2.1　支持向量机：支持向量机是 Vapnik[11] 等根据统计学习理论提出的一种新的机器学习方法,它基于结构风险最小(Structural Risk Minimization, SRM)原则,有完备的理论基础,近年来引起越来越多的关注,在分类、回归估计和密度估计等方面有很好的应用结果,相对于神经网络来说有更好的推广能力[12]。核函数的引入[13],使得这种方法很容易处理非线性问题,因此 SVM 有很好的应用前景。其判别函数为

$$f(x, a_i) = \sum_{i=1}^{n} (a_i - \hat{a}_i) k(x, x_i) + b \quad (2)$$

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

多项式核函数

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (3)$$

径向基核函数

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (4)$$

Sigmoid 核函数

$$k(x_i, x_j) = \tanh[b(x_i \cdot x_j) + c] \quad (5)$$

但是 SVM 有个缺点：随着样本数目的增大,所需的计算时间和空间存储资源都会成几何级数增加。为了提高 SVM 运算速度和精度,采用基于主成分分析的支持向量机[14]。其基本思想是针对复杂的系统问题,首先利用主成分分析对多变量参数矩阵进行处理,由于主成分分析的实质是 n 维空间的坐标旋转,并不改变样本数据结构,得到的主成分是

原变量的线性组合且两两不相关,能够最大程度地反映原变量所包含的信息,在以一定标准选取前 k 个较重要的主成分之后,原来的多维问题就得以简化。以精简过的参数矩阵作为支持向量机的输人量时,在训练样本数并未减少的基础上消除了支持向量机输入间的相关性,同时减少了支持向量机的输人数,简化了支持向量机的结构,可从整体上提高支持向量机的性能。

1.2.2 主成分分析:主成分分析(Principal Component Analysis, PCA)又称主分量分析,是 Hotelling 于 1933 年首先提出的[15]。主成分分析就是利用降维的思想,把多指标转化为少数几个不相关的综合指标的一种多元统计分析方法。通过主成分分析,把主成分分析的变量作为支持向量机的输入数据,既可减少支持向量机的输入变量,加快收敛,又起到了主成分过滤噪音的目的。具体操作过程为,(1) 对原始数据进行标准化处理,(2) 建立相关矩阵;(3) 计算特征值及特征向量;(4) 建立主成分方程,计算主成分荷载及主成分得分;(5) 根据主成分分析结果构建支持向量机的输入层数据。

1.2.3 均匀设计法:均匀设计由方开泰[16]所创造,它是将数论和多元统计相结合的一种新颖的试验方法,其核心思想是用确定性方法寻找空间中均匀分布的点集来代替 Monte Carlo 中的随机数。它通过提高试验点'均匀分散'的程度,使试验点具有更好的代表性及能用较少的试验获得较多的信息。

为了定量比较拟合和测试效果,特定义以下三个特征指标:

(1) 平均绝对百分比误差:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}, \quad (6)$$

(2) 均方根误差:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

(3) 平均绝对误差:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (8)$$

式中 y_i 和 $\hat{y}_i$ 分别表示实际值和拟合值(或预测值)。

2 结果与分析

2.1 20 种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见表 2。分析表中数据,

表 2 主成分特征值及累计方差贡献率

Table 2 Principal component and their explaining variances

Component	Eigen values	Variance/%	Cumulative/%
Pr1	5.21	24.79	24.79
Pr2	4.19	19.97	44.76
Pr3	2.84	13.51	58.27
Pr4	2.19	10.45	68.71
Pr5	1.5	7.14	75.85
Pr6	1.07	5.12	80.96
Pr7	0.92	4.4	85.36
Pr8	0.8	3.8	89.16
Pr9	0.6	2.86	92.01
Pr10	0.434	2.07	94.08
Pr11	0.38	1.8	95.88

选择显著水平 $\alpha = 95\%$,则只挑选前 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 个变量之间的关系为(限于篇幅,仅写出前 3 个主成分):

Pr1 = 0.200A + 0.063C + 0.116D0.146E − 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 + 0.159 NUM

Pr2 = − 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 + 0.0.301 NUM

Pr3 = − 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 + 0.260 NUM

各氨基酸与 11 个主成分之间的关系如表 3 所示,为简单起见,相关系数保留 1 位小数,且仅列出相关系数绝对值≥0.2 的氨基酸。与 MMPD 中所报道的几丁质酶三级结构进行了比较,发现主成分分

表 3 几丁质酶氨基酸与各主成分的关系

Table 3 Meaning of principle components in chitinase

	Amino acids (positive)	Amino acids (negative)
Pr 1	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
Pr 2	0.4D 0.4E 0.3K 0.2Y	0.3A 0.3C 0.3G 0.3Q
Pr 3	0.2D 0.2Q 0.4T 0.4V	0.4F 0.4H 0.3L 0.3P 0.3R
Pr 4	0.3C 0.3E 0.2G 0.2K 0.2N 0.5W 0.2Y	0.2A 0.3I 0.4L 0.2M 0.2V
Pr 5	0.2D 0.6G 0.5M	0.2A 0.3F 0.3N 0.2V
Pr 6	0.3L 0.6S 0.2W	0.2I 0.3M 0.2T 0.4Y
Pr 7	0.4C 0.4F 0.2I 0.4V	0.3A 0.2H 0.3L 0.3N 0.2W 0.2Y

析的前 7 个主成分所代表的几丁质酶的二级结构分别为转曲、转角(turn)、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，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，有三个参数，分别为：惩罚系数 C 值，epsilon 和 Gamma 值。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，所以我们采用均匀设计方法来优化参数，设计的 4 因素 10 水平均匀设计表如表 3 所示。

表 4　均匀设计表

Table 4　Uniform design U10(103)

Run no.	Three factors			MAPE9/%	MAE	MSE
	C	epsilon	gamma			
R1	1	0.4	0.001	13.79	0.77	1.08
R2	100	0.1	0.03	8.99	0.52	0.86
R3	1,000,000	0.5	0.1	5.44	0.32	0.58
R4	10,000	0.3	0.3	3.83	0.22	0.4
R5	100,000	0.8	0.01	11.37	0.65	0.98
R6	1,000	1	0.04	8.19	0.48	0.82
R7	0.001	0.6	0.02	9.99	0.58	0.92
R8	**10**	**0.7**	**0.5**	**3.76**	**0.22**	**0.41**
R9	0.01	0.2	0.05	7.45	0.44	0.77
R10	0.1	0.9	0.2	4.09	0.24	0.43

The optimum values are highlighted.

计算结果显示，当 C 值为 10，epsilon 为 0.7，Gamma 值为 0.5 时对 pH 值预测的平均绝对百分比误差为 3.76%，均方根误差为 0.41 个 pH 单位，平均绝对误差为 0.22 个 pH 单位。具有比神经网络更好的拟合效果(如图 1 所示)。后续训练及测试均采用上述参数(C = 10，epsilon = 0.7，gamma = 0.5)。

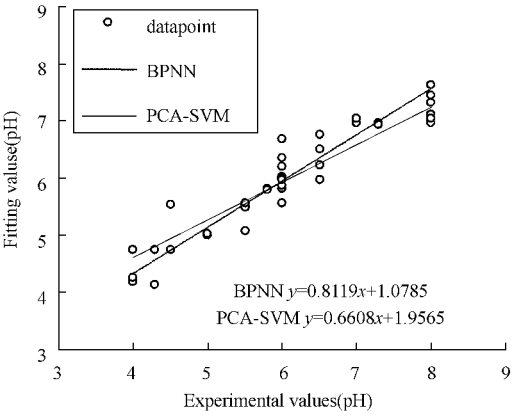

图 1　结构优化后支持向量机的拟合值

Fig. 1　The fitting values of support vector machine optimized

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，训练后对训练集外的输入的响应如何，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。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。由于篇幅有限，取测试结果的一部分如图 5 所示。

从测试结果中，可以看出主成分分析-支持向量

表 5　BPNN、SVM 和 PCA-SVM 的测试结果

Table 5　Results of the cross-validations of BPNN，SVM and PCA-SVM

Run NO.		MAPE			MAE		
		BPNN	SVM	PCA-SVM	BPNN	SVM	PCA-SVM
1st	Training	0.04	0.04	0.04	0.25	0.28	0.22
	Testing	0.04	0.04	0.12	1.66	2.24	0.93
2nd	Training	0.04	0.04	0.04	0.25	0.04	0.22
	Testing	0.04	0.04	0.00	0.02	0.23	0.01
3rd	Training	0.04	0.04	0.04	0.25	0.00	0.22
	Testing	0.04	0.04	0.00	0.04	0.01	0.00
6th	Training	0.04	0.04	0.04	0.26	0.03	0.22
	Testing	0.04	0.04	0.00	0.43	0.16	0.00
12th	Training	0.04	0.04	0.04	0.22	0.18	0.22
	Testing	0.04	0.04	0.03	1.20	0.89	0.17
Average	Training	0.05	0.04	0.04	0.29	0.22	0.22
	Testing	0.23	0.15	0.08	1.27	0.84	0.42

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 0.04 和 0.08,预测结果的平均绝对误差最大值为 0.40,最小值为 0.00,其变化范围较大。预测结果中共有 11 个样本平均绝对百分比误差近似于零值,有 19 个样本的平均绝对百分比误差的训练和测试结果都小于 5%,并且其绝对误差在 0.22 个 pH 单位以内,26 个测试样本的预测结果的平均绝对误差为 0.42 pH 单位,这些数字可以表明主成分分析-支持向量机模型在预测几丁质酶最适 pH 中得到了很好的预测效果。

2.4 BP 神经网络、支持向量机、主成分分析-支持向量机的比较

近年研究表明,支持向量机模型的预测结果要好于人工智能神经网络,在此做了一个比较。参考我们以前的研究结果[17],选择一个隐含层的神经网络,BP 神经网络的训练误差仍设为 0.01,运算次数为 1000。同样用均匀设计方法($U_{10}(10^3)$)优化 BP 神经网络的四个参数:学习速率、动态参数、Sigmoid 参数和隐含层结点数。当学习速率、动态参数、Sigmoid 参数和隐含层结点数分别为 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,测试结果如图 5 所示。其 26 个样本的训练和测试平均绝对百分比误差的平均值为 0.05 和 0.23,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 0.04 和 0.15,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 个测试样本的预测结果的平均绝对误差为 1.27 个 pH 单位,高于支持向量机模型的 0.84 个 pH 单位,更高于主成分分析 – 支持向量机的预测结果 0.42 个 pH 单位。从图 3 可以看出,用 BP 神经网络预测几丁质酶最适 pH 值,结果比较差,其预测结果不稳定,而支持向量机模型得出

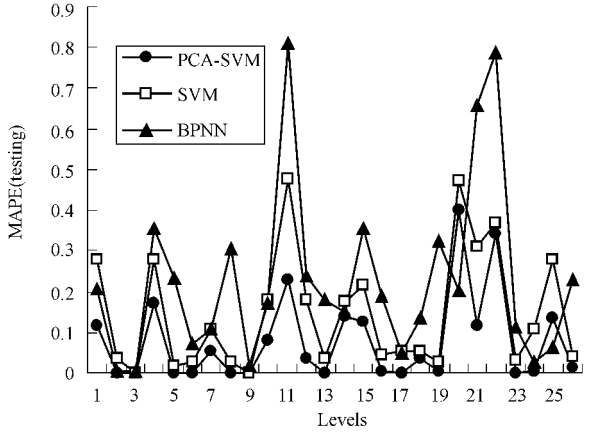

图 2 BPNN、SVM 和 PCA-SVM 的比较

Fig.2 The comparison between BPNN model SVMand PCA-SVM model

BPNN: back-propagation neural network; SVM: support vector machine; PCA-SVM: principal component analysis- support vector machine; MAPE: mean absolute error. It is the results of all the 26 runs in cross-validation.

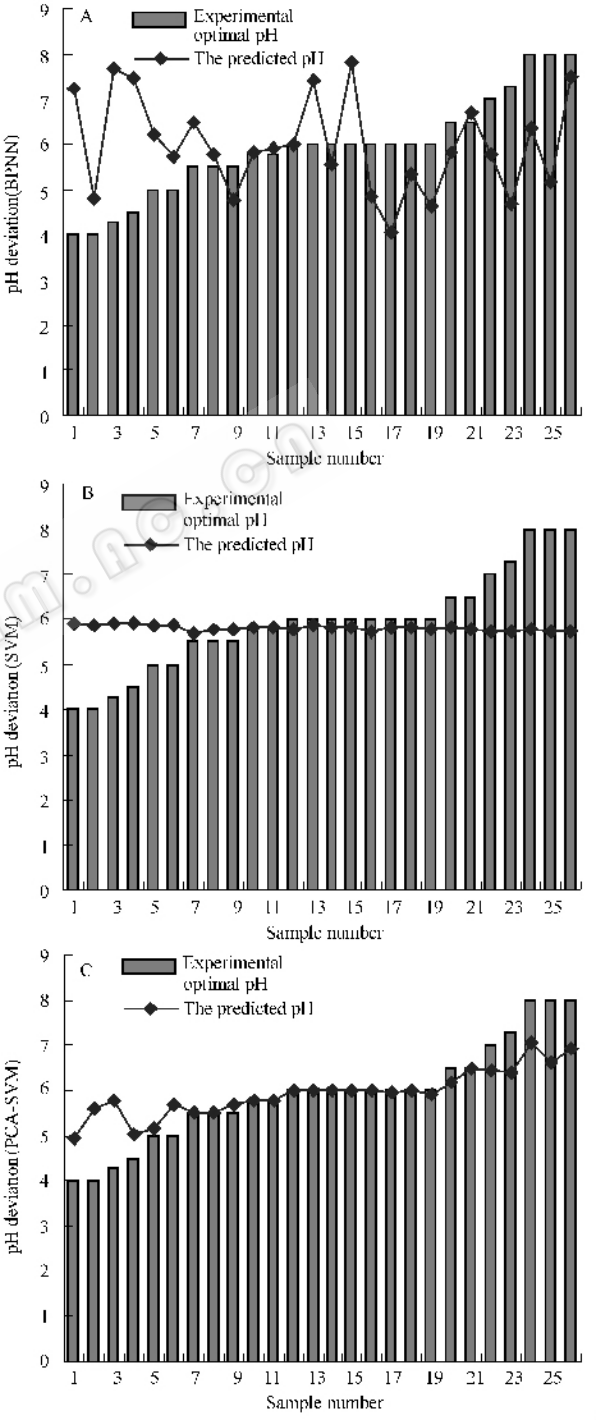

图 3 几丁质酶最适 pH 预测值与实验值的比较

Fig. 3 Comparison of experimental and predicted pH values. The results of the three models in cross-validation were shown in Fig. 3. A is the result of BPNN, B is the result of SVM. And C is the result of PCA-SVM.

的结果要好的多，其预测结果浮动较小，并且大部分预测值接近于真实的实验值。经过主成分分析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，在本试验课题中大大提高了模型的运算速度和测试精度。

3　讨论

几丁质酶结构与功能、性质的关系错综复杂，使用传统的回归分析往往不能满足要求。支持向量机建模以其独特的非线性能力得到广泛应用，但其自身存在缺陷。本文利用主成分分析法对样本集进行预处理，减少支持向量机的输入数目，集合了所有参数的信息量，同时消除输入因子的相关性并简化模型结构，并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。本研究中，主成分分析 – 支持向量机模型优于支持向量机模型和 BP 神经网络模型，说明几丁质酶的氨基酸组成和其最适 pH 间的关系非常复杂，用简单的线性模型可能并不能得到令人满意的结果。而主成分分析能够有效的简化数据变量，使输入数据变量减少到 11 个，经均匀设计优化后的拓扑结构更为简单，使得程序运行的速度明显加快，支持向量机的执行效率有了较大的提高，得到的支持向量机模型也有较高的精度。而一般的支持向量机方法的输入变量数目为 21 个，拓扑结构较为复杂，运行速度较慢。

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 pH 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，可达到大大降低文库丰度，减轻筛选工作量，提高效率，节省费用之目的。更重要的是，一方面它降低突变文库的冗余度，就可以在更大的范围内搜索到有用序列；另一方面，它可充分利用计算机的速度，在比单纯实验（10^{14}）更大的范围内（10^{80}）进行筛选，因此，可显著提高获得性质更优良突变酶的几率。最后需要指出的是，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，支持向量机的检测效果还会有所改善。而且本文仅考虑了 20 种氨基酸的频率分布和氨基酸的个数，排除了其它影响因素，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。而且所得结果仍需要实验进一步验证。

REFERENCES（参考文献）

[1] Anton P，Bussink，Marco van Eijk，*et al*. The Biology of the gaucher cell：the cradle of human chitinase. *A Survey of Cell Biology*，2006，**252**（2）：71 ~ 128.

[2] Jiang HB（蒋红彬），Zhang Y（张瀛），Jiang QL（蒋千里），*et al*. Advances in the research of chitinase. *Shandong Science*（山东科学），2000，**13**：41 ~ 45.

[3] Xiao YC（肖业臣），Luo XB（罗晓斌），Feng PF（冯佩富），*et al*. Advances in insect chitinase research. *Biotechnology*，2003，**232**（13）：38 – 40.

[4] Chang MC，Lai LP，Wu LM. Biochemical characterization and site-directed mutational analysis of the double chitin-binding domain from chitinase 92 of Aeromonas hydrophila JP101. *FEMS Microbiology Letters*，2004，**232**：1 ~ 61.

[5] Svendsen A. Enzyme Functionality：Design，Engineering，and Screening. Boca Raton：CRC Press，2004，pp.1 – 712.

[6] Wang FY（王凡业），Xue WY（薛文漪）. Using a new approach to engineer enzyme activity-sem i-rational design. App lied Chemical Industry，2006，**35**（8）：634 ~ 636

[7] 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.

[8] Klein P. Prediction of protein structural class by discriminate analysis. *Biochem Biophys Acta*，1986，**874**（2）：205 – 275.

[9] Chou KC，Maggiora GM. Domain structure prediction. *Protein Eng*，1998，**11**（7）：523 – 538.

[10] Chou KC. A key driving force in determination of protein structural classes. *Biochem Biophys Res Commun*，1999，**264**（1）：216 – 224.

[11] Vapnik VN. Statistical Learning Theory. New York：Wiley，1998，pp.1 – 736.

[12] Wong WT，Hsu SH. Application of SVM and ANN for image retrieval. *European Journal of Operational Research*，2006，**173**：938 – 950.

[13] Cortes C，Vapnik V. Support vector machine networks. *Machine Learning*，1995，**20**：273 – 297.

[14] Zhao GS（赵广社），Zhang XR（张希仁），Research of support vector machine classified method based on principal componet analysis. *Computer engineering and Application*（计算机工程与应用），2004，**37**（3）：37 – 39.

[15] Jolliffe IT. Principal Component Analysis. New York：Springer，2002，1 – 487.

[16] Fang KT. The uniform design：application of number- theoretic methods in experimental design. *Acta Math Appl Sin*，1980，**3**：363 – 372.

[17] Zhang GY（张光亚），Fang BS（方柏山）. A uniform design based BP-neural network model for amino acid composition and optimum pH in G/11 xylanase. *Chinese Journal of Biotechnology*（生物工程学报），2005，**21**（4）：658 – 661.