文章编号: 1000-5013(2007)01-0055-04

一种预测木聚糖酶最适温度的 PCANN 模型

张光亚, 葛慧华, 方柏山

(华侨大学 材料科学与工程学院, 福建 泉州 362021)

摘要: 采用主成分分析法对样本数据集进行预处理,将得到的新样本数据集输入神经网络,构建 F/10 家族木聚糖酶氨基酸组成和最适温度的主成分分析神经网络(PCANN)模型. 结果表明,当学习速率为 0.07、动态参数为 0.8、Sigmoid 参数为 0.96,隐含层结点数为 5 时,模型对温度拟合的平均绝对百分比误差为 4.97%,绝对误差为 3.03 ℃. 同时,方法具有良好的预测效果,预测的平均绝对百分比误差为 4.68%,平均绝对误差为 3.55 ℃.

关键词: 主成分分析; BP 神经网络; 木聚糖酶; 最适温度; 虚拟筛选

中图分类号: Q 550.3; Q 811.211; TP 183 **文献标识码**: A

设计具有新特性的蛋白主要有两种方法,即理性设计(定点突变法)和非理性设计(定向进化法). 定点突变在一次循环中仅能对一个位点进行突变,当靶目标超过 3 个时,其效率急剧下降[1]. 而且,由于对定点突变结果的预测能力有限,对每个突变体要用实验一一验证,方可判断是否达到所需目标,从而降低了工作效率,限制了筛选通量,制约了突变的序列范围. 定向进化虽然取得了不少成功的例子[2-3],但其受限于实验所能筛选的序列的数量(高通量筛选可达 10^7). 因此,筛选往往成为定向进化的瓶颈[4]. 计算机技术的迅猛发展使其开始应用到筛选过程,其最大筛选的序列库可达 10^{80},极大增加了获得新蛋白的几率[5]. 目前已有一些应用成功的例子[6-7],算法和数学模型是其核心. 本文利用 F/10 家族木聚糖酶信息及其对应的最适温度,采用主成分分析的 BP 神经网络,建立了预测其最适温度的模型. 该模型具有较高的拟合和预测精度,可望用于对木聚糖酶的改造过程中的计算机虚拟筛选.

1 材料与方法

1.1 数据来源

F/10 家族木聚糖酶的序列来源于 Swiss-Prot Release 44.4(2004-08-31);木聚糖酶最适 pH 值的数据来源于文[8],最适温度均为实验所得数据. 23 个木聚糖酶 ID 号及最适温度,如表 1 所示. 木聚糖酶的氨基酸组成分析由 Bioedit 软件完成. 以各主成分得分值经标准化处理后的数据作为神经网络的输入层,各木聚糖酶对应的最适温度为输出层,采用单隐含层的 BP 神经网络. 主成分分析及神经网络由 DPS 软件完成. 表中,ID 为 Swiss-Prot 的登录号, T_{obs} 为文献报道的最适温度, T_{pre1} 为文[8]计算的最适温度, T_{pre2} 为 PCANN 模型计算的最适温度, $\delta_{MAPE,1}$, $\delta_{MAPE,2}$ 分别为文[8]和 PCANN 模型的平均绝对百分误差.

1.2 基于主成分分析的神经网络

1.2.1 主成分分析 主成分分析(Principal Component Analysis, PCA)也称主分量分析,是 Hotelling 于 1933 年首先提出的. 它利用降维的思想,把多指标转化为少数几个不相关的综合指标. 通过主成分分析,把变量作为神经网络的输入参数,既可减少神经网络的输入变量,加快网络的收敛,又起到了主成分

收稿日期: 2006-05-13

作者简介: 张光亚(1975-),男,讲师,博士研究生,主要从事生物信息学和酶工程的研究; 通信作者:方柏山(1957-),男,教授,博士生导师. E-mail:bsfang@hqu.edu.cn.

基金项目: 国务院侨务办公室科研基金资助项目(05Q0018)

过滤噪音的目的.具体操作过程有 5 个步骤. (1) 对原始数据(氨基酸组成百分比)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神经网络的输入层数据.

表 1 F/10 家族木聚糖酶

Tab. 1 Xylanase in family F/10

ID	T_{obs}	T_{pre1}	$\delta_{MAPE,1}$	T_{pre2}	$\delta_{MAPE,2}$	ID	T_{obs}	T_{pre1}	$\delta_{MAPE,1}$	T_{pre2}	$\delta_{MAPE,2}$
Q60042	102	84.14	17.51	91.83	9.97	P56588	67	57.84	13.67	63.17	5.72
Q60041	90	91.44	1.60	89.38	0.69	P07986	65	66.16	1.78	68.17	4.87
Q60037	90	84.08	6.58	92.04	2.26	P40943	65	67.14	3.29	67.19	3.37
Q12603	85	85.91	1.07	88.16	3.71	P45703	60	50.87	15.22	61.06	1.76
P40942	80	81.00	1.25	80.23	0.28	P33559	60	49.62	17.30	58.08	3.19
P23360	80	72.81	8.99	78.38	2.03	P26541	60	68.67	14.45	60.58	0.96
P51584	75	67.43	10.09	78.37	4.49	P14768	55	60.96	10.84	53.92	1.97
P07528	70	69.73	0.39	74.15	5.93	P26223	55	58.20	5.82	56.19	2.17
P23556	70	81.02	15.74	69.01	1.42	Q00177	52	49.91	4.02	44.55	14.33
P40944	70	75.15	7.36	68.59	2.01	O59859	40	52.89	32.23	48.24	20.61
P36917	70	71.67	2.39	70.89	1.28	P29417	40	44.81	12.03	46.22	15.56
P10478	70	76.46	9.23	74.03	5.76	平均值			9.25		4.97

1.2.2 均匀设计法 由于 BP 神经网络在多数情况下无法确定最佳隐含层神经元个数,而网络恰恰对隐含层中神经元个数很敏感.太少的神经元导致网络不适应,太多又容易溢出.同时,选择合适的神经网络拓扑结构和参数至关重要.在实际应用中,一些研究者往往根据自己的经验来选择参数.为了克服上述弊端,本文采用均匀设计来优化神经网络的拓扑结构和选择适当的运行参数.均匀设计由方开泰[9]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.为了定量比较拟合和测试效果,特定义以下 3 个特征指标.即平均绝对百分比误差 δ_{MAPE},均方根误差 δ_{MSE}和平均绝对误差 δ_{MAE},它们分别为

$$\delta_{MAPE} = \frac{1}{n}\cdot\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|},$$

$$\delta_{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2},$$

$$\delta_{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|.$$

其中,y_t 和 $\hat{y}_t$ 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后得到的特征值及累计方差贡献率,如表 2 所示.表中,P_C 为主成分,E_V 为特征值,ϕ_{TA}为贡献率,ϕ_{CU}为累计贡献率.分析表中数据,选显著水平 α= 95 %,则只选择前 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 20 个变量之间的关系(限于篇幅,仅写出前 3 个主成分)为

$$\begin{aligned}P_{C,1} = &-0.317A - 0.166C + 0.108D + 0.351E + 0.290F - 0.278G + \\ &0.185H + 0.252I + 0.271K + 0.219L + 0.197M - 0.154N + 0.020P - \\ &0.244Q + 0.133R - 0.279S - 0.258T - 0.156V + 0.075W + 0.216Y, \qquad (1)\\ P_{C,2} = &0.100A + 0.261C - 0.135D + 0.109E + 0.176F - 0.011G + 0.366H - \\ &0.260I - 0.262K + 0.040L - 0.032M - 0.310N - 0.009P + 0.165Q + \\ &0.445R - 0.023S - 0.114T - 0.058V + 0.404W - 0.292Y, \qquad (2)\\ P_{C,3} = &-0.081A + 0.383C - 0.300D - 0.049E + 0.089F + 0.191G - 0.006H + \\ &0.026I - 0.221K - 0.051L + 0.334M + 0.331N + 0.038P - 0.188Q + \\ &0.180R + 0.140S - 0.109T - 0.474V - 0.230W + 0.242Y. \qquad (3)\end{aligned}$$

式(1) (3)也说明了各氨基酸与主成分之间的关系. 相关系数越大,表明对该主成分的影响越大. 根据 Liu 等[10]的研究结果,F/ 10 家族主成分分析的前 7 个主成分所代表的该家族木聚糖酶的二级结构,分别为转曲、螺旋、折叠、转角、折叠、螺旋和转角,这与该家族的结构特征基本吻合. 本文主成分分析结果与其基本相符,但略有差异.

表 2　主成分及其累计方差贡献率

Tab. 2　Principal components and their accumulative square errors

P_C	E_V	ϕ_{VA}/ (%)	ϕ_{CU}/ (%)	P_C	E_V	ϕ_{VA}/ (%)	ϕ_{CU}/ (%)	P_C	E_V	ϕ_{VA}/ (%)	ϕ_{CU}/ (%)	P_C	E_V	ϕ_{VA}/ (%)	ϕ_{CU}/ (%)
1	6.7	33.6	33.6	6	0.9	4.6	80.0	11	0.4	1.8	95.3	16	0.1	0.3	99.6
2	2.8	14.0	47.6	7	0.8	4.1	84.1	12	0.3	1.3	96.6	17	0.0	0.2	99.8
3	2.5	12.3	59.9	8	0.7	3.5	87.6	13	0.2	1.1	97.8	18	0.0	0.1	99.9
4	1.8	9.2	69.1	9	0.6	3.2	90.8	14	0.2	0.9	98.7	19	0.0	0.1	100.0
5	1.3	6.3	75.4	10	0.5	2.7	93.5	15	0.1	0.7	99.3	20	0.0	0.0	100.0

2.2　BP 神经网络拓扑结构的优化

将主成分分析方法所得的新样本数据集,作为神经网络的输入层,最适温度作为输出层,选择 3 层 BP 神经网络. 本次使用的 BP 神经网络包含一个隐含层,对学习速率(η)、动态参数(k)、Sigmoid 参数(S)和隐含层结点数(N)共 4 个因素 9 水平进行均匀设计,所得的均匀设计表如表 3 所示. 为了避免过度拟合而导致测试效果较差,将允许误差设为 0.005,最大迭代次数设为 1 000 次. 计算结果显示,当学习速率为 0.07、动态参数为 0.8、Sigmoid 参数为 0.96,隐含层结点数为 5 时,所得模型对温度拟合的平均绝对百分比误差为 4.97 %,均方根误差为 4.00 ℃,平均绝对误差为 3.07 ℃,具有很好的拟合效果(表 1). 所得神经网络的拓扑结构为"11-5-1",后续训练及预测均采用上述参数.

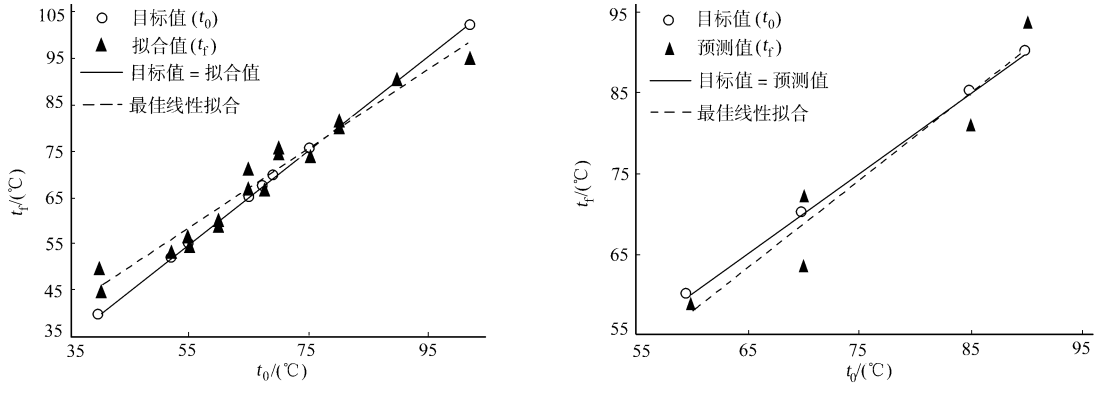
表 3　均匀设计表 $U_9(9^4)$

Tab. 3　Uniform design $U_9(9^4)$

水平	S	η	k	N	δ_{MAPE}	水平	S	η	k	N	δ_{MAPE}	水平	S	η	k	N	δ_{MAPE}
1	0.90	0.09	0.65	11	5.71	4	0.93	0.08	0.70	8	5.44	7	0.96	0.07	0.80	5	4.97
2	0.91	0.20	0.50	10	8.51	5	0.94	0.15	0.55	7	5.92	8	0.97	0.10	0.60	4	5.23
3	0.92	0.40	0.35	9	6.11	6	0.95	0.30	0.40	6	5.42	9	0.98	0.25	0.45	3	5.31

2.3　BP 神经网络模型预测

为了检验所建立的神经网络的可靠性,将上述 PCA 方法所得的新样本数据集从中随机取出 5 组作为测试集,其余 18 组作为学习集. 利用拓扑结构经优化后的 BP 神经网络模型进行学习和预测,共进行了 35 次,取预测结果较好的一次进行了回归分析,如图 1 所示. 由图 1 可知,PCA-BPNN(主成分分析-



(a)　$y = 0.846\,8x + 1.192\,9$　　(b)　$y = 1.088\,2x - 7.721\,1$

图 1　PCANN 的拟合和预测值

Fig. 1　Fitting and predicting value of PCANN

BP 神经网络)方法拟合数值的最佳线性拟合和"拟合值等于目标值"的理性曲线基本重合,而预测结构又有一定的偏差. 由此可见,拟合的效果优于预测的效果,其平均绝对百分比分别为 5.06 %和 4.68 %,均方根误差分别为 4.08 ℃和 3.95 ℃,平均绝对误差分别为 2.98 ℃和 3.55 ℃. 所建立的神经网络模型

具有令人满意的拟合和预测效果.

3 结束语

木聚糖酶结构与功能、性质的关系错综复杂,使用传统回归分析所得的数学模型往往不能满足要求.本文利用主成分分析对样本集进行预处理,在保留数据最大量信息的前提下,消除BP网络输入间的相关性,减少神经网络的输入数,使输入层的节点数从20减少到11个,简化了网络结构.而且,再经均匀设计优化,使得神经网络的拓扑结构更为合理,同时,程序运行的速度也明显加快,使神经网络的执行效率有了较大的提高,得到的神经网络模型也有较高精度.

参考文献:

[1] ANDREAS S, JEAN H J. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round[J]. Analy Biochem, 2004, 324: 285-291.

[2] LIU Xiang-mei, QÜ Yin-bo, FAN Yin, et al. Studies on the key amino acid residues responsible for the alkali-tolerance of the xylanase by site-directed or random mutagenesis[J]. J Mol Catalysis B: Enzy, 2002, 18(4-6): 307-313.

[3] 徐卉芳, 张先恩, 张治平, 等. 大肠杆菌碱性磷酸酶的体外定向进化研究[J]. 生物化学与生物物理进展, 2003, 30(1): 181-186.

[4] VOIGT C A, KAUFFMAN S, WANG Zhen-gang. Rational evolutionary design: The theory of *in vitro* protein evolution[J]. Adv Protein Chem, 2001, 55: 79-160.

[5] ROBERT J H, JORG B, MARIE L A, et al. Combining computational and experimental screening for rapid optimization of protein properties[J]. Proc Natl Acad Sci USA, 2002, 99: 15 926-15 931.

[6] VOIGT C A, MAYO S L, ARNOLD F H, et al. Computational method to reduce the search space for directed protein evolution[J]. Proc Natl Acad Sci USA, 2001, 98: 3 778-3 783.

[7] RICHARD F, AJOY R, SRIDHAR G, et al. Optimizing the search algorithm for protein engineering by directed evolution[J]. Protein Eng, 2003, 16(8): 589-597.

[8] LIU Liang-wei, WANG Mei-li, SHAO Wei-lan, et al. A novel model to determine the dipeptides responsible for optimum temperature in F/10 xylanase[J]. Process Biochem, 2005, 40(3): 1 389-1 394.

[9] 方开泰. 均匀设计与均匀设计表[M]. 北京: 科学出版社, 1994: 363-372.

[10] LIU Liang-wei, ZHANG Jing, CHEN Bin, et al. Principle component analysis in F/10 and G/11 xylanase[J]. Bioch Biophy Res Co, 2004, 322(1): 277-280.

A Principal Component-Artificial Neural Network Model for Predicting Optimum Temperature in F/10 Xylanases

ZHANG Guang-ya, GE Hui-hua, FANG Bai-shan

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

Abstract: The principal component analysis was first applied to the data processing in training sets, and then the obtained new principal components were used as input parameters of BP neural networks. A prediction model for optimum temperature of xylanases in F/10 family was established based on uniform design. When the learning rate, momentum parameter, Sigmoid parameter and the neuron numbers of the hidden layer was 0.07, 0.8, 0.96 and 5, respectively, the calculated temperatures fitted the reported optimum temperatures very well. The mean absolute percent error was 4.97%. At the same time, the predicted temperatures fitted the reported optimum temperatures well and the mean absolute error was 3.55 ℃. It was superior in fittings and predictions compared to the reported model based on stepwise regression.

Keywords: principal component analysis; BP neural networks; xylanase; optimum temperature; virtual screening

(责任编辑: 黄仲一)