木聚糖酶氨基酸组成与其最适 pH 的神经网络模型

A Model for Amino Acid Composition and Optimum pH in G/11 Xylanase Based on Neural Networks

张光亚,方柏山*

ZHANG Guang-Ya and FANG Bai-Shan*

华侨大学工业生物技术研究所,泉州 362021

Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, *China*

摘 要 籍均匀设计(UD)方法,构建了 G/11 家族木聚糖酶氨基酸组成和最适 pH 的神经网络(NNs)模型。当学习速率为 0.09、动态参数为 0.4、Sigmoid 参数为 0.98,隐含层结点数为 10 时,该模型对最适 pH 的拟合和预测平均绝对百分比误差可分别达到 3.02%和 4.06%,均方根误差均为 0.19 个 pH 单位,平均绝对误差分别为 0.11 和 0.19 个 pH 单位。该结果比文献报道的用逐步回归方法好。

关键词 木聚糖酶,均匀设计,神经网络,氨基酸组成,最适 pH

中图分类号 Q55 **文献标识码** A **文章编号** 1000-3061(2005)04-0658-04

Abstract In this paper, a prediction model for amino acid composition and optimum pH of xylanase in G/11 family was established in terms of an artificial neural networks based on uniform design. Results showed that the calculated and predicted pHs fitted the optimum pHs of xylanase very well and the MAPEs (Mean mean Absolute Percent Error) were 3.02% and 4.06%, the MSEs (Mean Square Error) were 0.19 and 0.19 pH unit, the MAE (Mean Absolute Error) were 0.11 and 0.19 pH unit, respectively. It was better in fittings and predictions compared with the reported model based on stepwise regression.

Key words xylanase, uniform design, neural networks, amino acid composition, optimum pH

木聚糖酶(EC3.2.1.8)是一种重要的工业用酶,可广泛应用于食品、饲料、医药和造纸等行业。自发现木聚糖酶可用于纸浆漂白以后,对木聚糖酶研究和开发越来越受到人们的重视[1,2]。用于纸浆漂白的木聚糖酶应耐热和耐碱,目前有两种解决方法:一是从极端环境中筛选木聚糖酶产生菌株[3];另一种方法是对木聚糖酶进行遗传改造。后者是一个更好的选择。木聚糖酶分别属于 F/10 和 G/11 家族,由于 G/11 家族的木聚糖酶分子更小,因此被认为用于纸浆漂白更有优势,而且其结构较为简单,更适合作为理论研究的分子模型[4]。

人工神经网络(ANN)是一种平行分散处理模式,其构建思想来源于人类大脑神经运作的模拟。ANN 不但具有较好的模式识别能力,而且可以克服统计等方法的限制。最重要的是它具有学习能力,可随时依据数据资料进行自适应学习、训练,调整其内部的存储权重参数以任意精度逼近一个非线性函数。神经网络作为一种人工智能技术,应用于众多学科领域。目前,已发展了几十种神经网络,在这众多神经网络模型中,应用最广泛的是多层感知机神经网络。该网络

Received: November 29, 2004; Accepted: January 13, 2005.

This work was supported by the grants from the National Natural Sciences Foundation of China (No. 20276026, 20446004) and the Science and Technology Foundation of Fujian Province of China (No. 2003I020).

* Corresponding author. Tel: 86-595-22691560, E-mail: fangbs@hqu.edu.cn

国家自然科学基金资助项目(No. 20276026, 20446004)、福建省科技计划重点项目(No. 2003I020)。

的研究始于 50 年代,但一直进展不大,直到 1985 年,Rumelhart 等人提出了误差反向传递学习算法(即 BP 算法)。BP 算法不仅有输入层节点、输出层节点,还可以有 1 个或多个隐含层节点。输入信号,先向前传播到隐含层节点,经作用函数变换后,再把隐节点的输出信号传播到输出节点,最后给出输出结果。由于其具有高度非线性映射的能力,通用性好且较为成熟,现已在发酵控制[5]、生物信息学[6]等领域得到广泛的应用。

本文利用 G/11 家族木聚糖酶的序列信息及对应的最适 pH,基于均匀设计,建立了氨基酸组成和最适 pH 之间的神经网络模型。该模型具有较高的拟合和测试精度,比文献报道的用逐步回归方法好,可望用于木聚糖酶和指导其它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

G/11 家族木聚糖酶的序列来源于 Swiss-Prot Release 44.4 (31/8/2004),Swiss-Prot 是一个非冗余的专家库。木聚糖酶最适 pH 实验值取自文献[7],22 个木聚糖酶 ID 号及最适 pH 见表 1 第 1、2 列。木聚糖酶的氨基酸组成分析由 Bioedit 软件完成。以各木聚糖酶的 20 种氨基酸组成百分比作为神经网络的输入,其对应最适 pH 为神经网络的输出。

表 1 G/11 家族木聚糖酶

Table 1 Xylanase in family G/11

ID	pH_{obs}	pH_{exp1}	MAPE1	pH_{exp2}	MAPE2	pH_{pre}	MAPE
P45705	8.00	6.91	13.63	7.52	6.00	7.54	5.76
P29127	8.00	7.40	7.50	8.19	2.38	7.77	2.88
P55332	6.50	5.96	8.31	6.31	2.92	6.49	0.08
O43097	6.50	6.43	1.08	5.82	10.46	6.50	0.04
P00694	6.50	6.64	2.15	6.49	0.15	6.51	0.21
P26515	6.50	7.05	8.46	6.04	7.08	6.56	0.85
P55334	6.50	7.10	9.23	7.04	8.31	6.50	0.03
P45796	6.50	5.71	12.15	6.67	2.62	6.50	0.01
P25811	6.30	6.24	0.95	6.05	3.97	6.35	0.75
Q06562	6.00	6.34	5.67	5.83	2.83	5.90	1.70
P35809	5.90	4.21	28.64	5.37	8.98	5.93	0.46
P55333	5.50	5.02	8.73	5.83	6.00	5.52	0.39
P09850	5.50	7.05	28.18	5.48	0.36	5.41	1.55
P29126	5.50	6.13	11.45	5.99	8.91	5.50	0.04
P36217	5.25	5.30	0.95	5.36	2.10	5.26	0.21
P48793	5.00	5.06	1.20	4.69	6.20	4.94	1.17
P18429	5.00	6.35	27.00	5.48	9.60	5.12	2.44
P48824	4.50	5.35	18.89	4.82	7.11	4.48	0.52
P36218	4.25	5.19	22.12	4.68	10.12	4.31	1.42
P55328	3.50	3.50	0.00	3.34	4.57	3.00	14.19
P55329	3.00	3.50	16.67	3.34	11.33	3.13	4.35
P33557	2.00	3.50	75.00	3.68	84.00	2.51	25.48
Average			14.00		9.36		2.93

ID, the accession number of xylanase in Swiss-Prot; pH_{obs}, the optimum pH found in the literature at which the relative xylanase has the maximum activity; pH_{exp1} and pH_{exp2}, the calculated pH according to reference 7; pH_{pre}, the calculated pH according to our model; MAPE, mean absolute percent error.

1.2 神经网络的典型结构

最常用神经网络是前馈神经网络,网络训练的方法为 BP 算法。近年来,人们在实际应用中发现前馈神经网络存在一些缺陷,主要表现在:(1) BP 算法收敛速度慢、容易陷入局部极小、数值稳定性差、参数难以调整。(2) 网络构造时,隐层神经元个数的选择难,若隐层神经元个数选得太少,网络容错性差;若隐层神经元个数选得太多,则学习及训练时间长且精度未必高。(3) 网络训练时初始值选择难,而初始值对训练是否陷入局部最小和是否收敛关系很大[8]。因此,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,这缺乏理论依据。为了克服上述弊端,本文采用均匀设计的方法来优

化神经网络的拓扑结构和选择适当的运行参数。

1.3 均匀设计法

均匀设计由我国数学家方开泰教授所创造[9],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 :$\mathrm{MAPE}=\frac{1}{n}\sum_{t=1}^{n}\frac{|y_t-\hat{y}_t|}{|y_t|}$,

(2)均方根误差 :$\mathrm{MSE}=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}$,

(3)平均绝对误差 :$\mathrm{MAE}=\frac{1}{n}\sum_{t-1}^{n}|y_t-\hat{y}_t|$

式中,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 基于均匀设计的 BP 神经网络的优化

为了科学地确定神经网络中连接权的初始值、最佳的隐含层神经元的个数、学习速度等参数,本文选择一个隐含层的神经网络,对学习速率、动态参数、Sigmoid 参数和隐含层结点数 4 个因素 9 水平进行均匀设计,所得的均匀设计表和训练结果如表 2 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 0.001,最大迭代次数设为 1000 次。

表 2 均匀设计表

Table 2 Uniform design U9(9^4)

Levels	Four factors				MAPE
	Sigmoid parameter	The learning rate	Momentum parameter	The neuron numbers of the hidden layer	
1	0.9	0.1	0.6	8	3.40
2	0.93	0.15	0.35	13	3.05
3	**0.98**	**0.09**	**0.4**	**10**	**2.93**
4	0.91	0.3	0.45	5	3.03
5	0.96	0.25	0.5	17	2.99
6	0.94	0.4	0.65	11	3.11
7	0.95	0.07	0.55	3	16.78
8	0.97	0.2	0.8	6	2.99
9	0.92	0.08	0.7	15	4.23

由表 2 可见,当学习速率为 0.09、动态参数为 0.4、Sigmoid 参数为 0.98,隐含层结点数为 10 时,模型对 pH 值拟合的平均绝对百分比误差为 2.93%,均方根误差为 0.19 个 pH 单位,平均绝对误差为 0.11 个 pH 单位。具有很好的拟合效果。后续训练及测试均采用上述参数。

2.2 BP 神经网络的测试

对神经网络而言,由于训练样本集的大小有限,网络训练后对训练集外的输入的响应如何,直接决定了网络的性能。为了检验所建立的神经网络的可靠性,从上述 22 组数据中任取 3 组作为测试样本,其余 19 组作为训练样本,对 BP 神经网络模型进行测试,共进行了 60 次测试,限于篇幅,仅给出其中较好的 7 组数据,如表 3 所示。在 60 次测试中,拟合的效果总体好于测试的效果,其平均绝对百分比误差分别为 3.18%和 16.47%。而采用不同样本进行训练和测试,所得的结果也相差较大,尤其是测试的结果存在比较明显的差异,所得 MAPE 值最大为 38.2%,最小仅为 3.45%。而方案 22 无论是训练还是测试都有良好的性能,所得训练和测试的 MAPE 分别为 3.02%和 4.06%,总和为 7.08%;方案 51 的训练和测试的 MAPE 分别为 3.26%和 3.45%,总和为 6.71%。为了进一步比较测试的结果,从上述 60 组测试中,取 MAPE 总和小于 10%的 7 组较好数据,分别计算其 MSE 和 MAE 值,结果如表 4 所示。7 组训练及测试的 MSE 和 MAE 值均小余 0.5 个 pH 单位,而方案 22 结果最好,可作为后续使用的模型。

表 3 训练与测试的绝对平均百分比误差

Table 3 MAPEs of training and testing

Run No.	MAPE of training/%	MAPE of testing/%	Sum/%	Run No.	MAPE of training/%	MAPE of testing/%	Sum/%
15	**2.94**	**5.76**	**8.69**	50	**3.01**	**6.72**	**9.73**
16	**3.04**	**5.50**	**8.54**	51	**3.26**	**3.45**	**6.71**
17	**3.19**	**6.65**	**9.83**	58	**3.20**	**4.35**	**7.55**
22	**3.02**	**4.06**	**7.08**				

表 4 10 组方案的训练和测试 MSE 和 MAE

Table 4 MSEs and MAEs of training and testing

Run No.	15	16	17	**22**	50	51	58
MSE of training	0.20	0.19	0.20	**0.19**	0.19	0.20	0.19
MSE of testing	0.26	0.31	0.37	**0.19**	0.71	0.31	0.29
MAE of training	0.13	0.11	0.13	**0.11**	0.12	0.14	0.12
MAE of testing	0.22	0.24	0.33	**0.19**	0.42	0.21	0.26

3 讨论

目前,对木聚糖酶进行遗传改造的方法主要有理性(定点突变法)和非理性(定向进化法)两种方法。由于大部分定点突变技术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[10]。而定向进化却存在筛选容量过大,筛选过程复杂且费用昂贵、费时费力等缺陷[11]。因此,筛选往往成为定向进化的瓶颈[12]。

随着计算机技术的不断发展,将计算机用于筛选已有文献报道[13,14]。最近,Liu[7] 等利用逐步回归的方法首次建立了单个氨基酸和二肽与木聚糖酶最适 pH 之间的数学模型,其拟合的平均绝对百分比误差分别为 14.00%和 9.36%(见表 1 第 3~6 列),本文拟合的平均绝对百分比误差仅为 2.93%(见表 1 第 7、8 列),同时该模型也具有良好的预测效果。可见本文结果更理想。这说明木聚糖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,而相对传统的数理统计方法而言,BP 神经网络可以求解非线性问题,还具有较强的容错能力,且判别精度一般不受样本中噪声的影响。利用木聚糖酶的晶体数据,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高通量预筛选,从而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})[14] 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 BP 神经网络的结构进行了优化,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,神经网络的检测效果还会有所改善。同时,如果在进行神经网络训练之前对样本集进行优化和选择,如采用主成分分析法,可望进一步提高拟合和测试的效果。而且本文只考虑了 20 种氨基酸的频率分布,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。

REFERENCES(参考文献)

[1] Xie FH(解复红), Li WP(李文鹏), Zhang KQ(张克勤). Advance of alkaline and thermophilic xylanase. *China Biotechnology*(中国生物工程杂志),2003,**23**(7):72 – 75

[2] Badhan AK, Chadha BS, Sonia KG *et al*. Functionally diverse multiple xylanases of thermophilic fungus *Myceliophthora* sp. IMI 38709. *Enzyme and Microbial Technology*, 2004, **35**(5): 460 – 466

[3] Khasin A, Alchanati I, Shoham Y. Purification, and characterization of a thermostable xylanase from *Bacillus stearothermophilus* T-6. *Appl Environ Microbiol*, 1993, **59**: 1725 – 1730

[4] Sapag A, Wouters J, Lambert C *et al*. The endoxylanases from family 11: computer analysis of protein sequences reveals important structural and phylogenetic relationships. *J Biotechnol*, 2002, **95**: 109 – 131

[5] Fang BS(方柏山), Chen HW(陈宏文), Xie XL(谢小兰) *et al*. The medium optimization of xylitol fermentation based on neural networks and genetic algorithms. *Chinese Journal of Biotechnology*(生物工程学报),2000,**16**(5): 648 – 650

[6] Zhang LZ(张立震), Tang HW(唐焕文). A method of protein structure class prediction based on subsequence distribution. *Computers and Applied Chemistry*(计算机与应用化学), 2003, **20**(3): 251 – 256

[7] Liu LW, Li XQ, Li X *et al*. Computational analysis of responsible dipeptides for optimum pH in G/11 xylanase. *Biochemical and Biophysical Research Communications*, 2004, **321**: 391 – 396

[8] Hsiao TCR, Lin CW, Chiang HHK. Partial least-squares algorithm for weights initialization of backpropagation network. *Neurocomputing*, 2003, **50**: 237 – 247

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

[10] Andreas S, Jean HJ. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. *Analytical Biochemistry*, 2004, **324**: 285 – 291

[11] Daniel NB, Christopher AV, Stephen LM. *De novo* design of biocatalysts. *Current Opinion in Chemical Biology*, 2002, **6**: 125 – 129

[12] Voigt CA, Kauffman S, Wang ZG. Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv Protein Chem*, 2000, **55**: 79 – 160

[13] Voigt CA, Mayo SL, Arnold FH *et al*. Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci USA*, 2001, **98**: 3778 – 3783

[14] Robert JH, Jorg B, Marie LA *et al*. Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci USA*, 2002, **99**: 15926 – 15931