文章编号 1000-5013(2005)02-0191-04

木聚糖酶氨基酸组成与最适温度的模型

张光亚 方柏山

(华侨大学材料科学与工程学院,福建 泉州 362021)

摘要 运用基于均匀设计(UD)的神经网络(NNs)构造法,构建 F/10 家族木聚糖酶氨基酸组成和最适温度的数学模型. 当学习速率为 0.1、动态参数为 0.6、Sigmoid 参数为 0.9,隐含层结点数为 7 时,该模型对最适温度的拟合和预测的平均绝对百分比误差分别为 6.61 %和 1.78 %,均方根误差分别为 5.43 ℃和 2.00 ℃,平均绝对误差分别为 4.13 ℃和 1.46 ℃.

关键词 均匀设计,BP 神经网络,木聚糖酶,氨基酸组成,最适温度

中图分类号 Q 556⁺.203 **文献标识码** A

木聚糖酶能水解木聚糖主干链内部的 β1,4 糖苷键,产生低聚木糖或带有侧枝的寡聚糖. 木聚糖酶在纸浆和造纸工业上,有着非常广泛的应用前景. 但是,目前工业上多数用于纸浆漂白的木聚糖酶,其最适温度在 45 ℃以下. 这引起了国内外学者对木聚糖酶及其基因的关注[1~3]. BP 神经网络是一种基于误差反向传播算法的神经网络,由于其高度非线性映射的能力,通用性好且较为成熟,现已在蛋白质高级结构等方面得到了广泛应用[4~6]. 本文利用现有属于 F/10 家族木聚糖酶的序列信息及最适温度,利用基于均匀设计的 BP 神经网络,建立了氨基酸组成和最适温度之间的数学模型. 它具有较高的拟合和测试精度,对于木聚糖酶的理性改造具有重要意义.

1 材料与方法

1.1 数据来源

通常,最适温度高的木聚糖酶,其热稳定性也较好[7],而且最适温度与酶的活性相关,故而以最适温度作为标准. F/10 家族的木聚糖酶的序列来源于 Swiss-Prot Release 44.4(2004-08-31),木聚糖酶最适温度的数据来源于文[8]. 23 个木聚糖酶的 ID 号及最适温度,如表 1 所示. 表中,ID 为 Swiss-Prot 登陆号,T_{obs} 指最适温度的实验值,T_{pre1} 为文[8]用逐步回归拟合的结果,T_{pre2} 为本文用 BP 神经网络拟合的结果. 所有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 中氨基酸组成百分比作为神经网络的输入层,其对应最适 pH 值为神经网络的输出层.

1.2 神经网络的典型结构[9]

人工神经网络按结构与训练算法的不同,可分为几种不同类型,最常用的是前馈网络. 该网络常用误差反向传播算法(BP 算法)进行. 近年来,人们在实际应用中发现前馈神经网络存在许多缺陷. 本文采用均匀设计法,来寻找合适的神经网络结构.

1.3 均匀设计法[10]

均匀设计是将数论和多元统计相结合,创新出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,从而代替 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,以及能用较少的试验获得较多的信息.

收稿日期 2004-10-29

作者简介 张光亚(1975-),男,讲师,博士研究生,主要从事酶与生物信息学的研究. E-mail:zhgyghh@hqu.edu.cn

基金项目 华侨大学科研基金资助项目(03HZR7)

表 1 使用的 F/10 家族木聚糖酶

ID	T_{obs}	T_{pre1}	MAPE1	T_{pre2}	MAPE2
Q60042	102	84.14	17.51	98.48	3.45
Q60041	90	91.44	1.60	93.89	4.32
Q60037	90	84.08	6.58	98.41	9.34
Q12603	85	85.91	1.07	90.91	6.95
P40942	80	81.00	1.25	85.89	7.37
P23360	80	72.81	8.99	77.18	3.52
P51584	75	67.43	10.09	86.59	15.46
P07528	70	69.73	0.39	77.32	10.45
P23556	70	81.02	15.74	69.35	0.93
P40944	70	75.15	7.36	71.57	2.24
P36917	70	71.67	2.39	66.66	4.77
P10478	70	76.46	9.23	73.93	5.61
P56588	67	57.84	13.67	56.51	15.66
P07986	65	66.16	1.78	64.21	1.22
P40943	65	67.14	3.29	66.58	2.43
P45703	60	50.87	15.22	58.13	3.11
P33559	60	49.62	17.30	55.38	7.70
P26541	60	68.67	14.45	57.63	3.95
P14768	55	60.96	10.84	54.82	0.34
P26223	55	58.2	5.82	54.89	0.20
Q00177	52	49.91	4.02	50.45	2.99
O59859	40	52.89	32.23	55.55	38.88
P29417	40	44.81	12.03	47.07	17.67
平均值			9.25		7.33

为了定量比较拟合和测试效果,特定义 3 个特征指标,分别是平均绝对百分比误差 MAPE、均方根误差 MSE 和平均绝对误差 MAE. 即

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

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

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

在上式指标中,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 神经网络优化

由于 BP 神经网络在多数情况下无法确定连接权的初始值,以及最佳的隐含层神经元的个数,而网络恰恰对隐含层中神经元的个数很敏感. 神经元太少导致网络不适应;太多则容易溢出. 为提高 BP 算法的性能,避免局部极小,希望增加隐含层神经元个数,故较低的学习速度更适应连接权的调整. 为此,本文采用均匀设计方法设计出一个 BP 网络,以构造一个用于预测最适温度的神经网络模型. 训练样本来自 23 个木聚糖酶的氨基酸组成,以及对应的最适温度. 本次使用的 BP 神经网络包含一个隐含层,对学习速率、动态参数、Sigmoid 参数和隐含层结点数 4 个因素的 9 水平进行均匀设计,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 0.01,最大迭代次数设为 1 000 次. 计算结果显示,当学习速率为 0.1、动态参数为 0.6、Sigmoid 参数为 0.9,隐含层结点数为 7 时,所得的样本误差为 0.951 %. 同时,模型对温度预测的平均绝对百分比误差为 7.3 %,MSE 值为 6.01 ℃,MAE 值为 4.6 ℃. 后续训练及测试均采用上述参数.

表 2　均匀设计表

Sigmoid 参数	学习速率	动态参数	隐含层节点数	Sigmoid 参数	学习速率	动态参数	隐含层节点数
0.90	0.10	0.6	7	0.94	0.40	0.65	11
0.93	0.15	0.35	13	0.95	0.07	0.55	2
0.98	0.09	0.40	9	0.97	0.20	0.80	4
0.91	0.30	0.45	3	0.92	0.08	0.70	15
0.96	0.25	0.50	17				

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响应,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 23 组数据中任取 3 组作为测试样本,其余 20 组作为训练样本. 对 BP 神经网络模型进行检验,共进行了 60 次检验,相关的操作由神经网络软件完成. 限于篇幅,此处仅列出较好的几次,如表 3 所示. 表中, T_{MAPE}, P_{MAPE}分别为训练和测试的平均误差,Σ为 总和. 由表可知,在60次检验中,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 7.68 %和 13.04 %. 采用不同样本进行训练和测试,所得的结果也相差较大.

表 3　训练与测试的绝对平均百分比误差

方案号	T_{MAPE}/ (%)	P_{MAPE}/ (%)	Σ (%)	方案号	T_{MAPE}/ (%)	P_{MAPE}/ (%)	Σ (%)
10	8.95	6.08	15.03	38	7.15	8.08	15.24
23	6.53	7.18	13.7	41	7	7.38	14.38
30	6.55	6.06	12.61	45	6.61	1.78	8.39
32	5.53	7.28	12.81	46	7.65	5.77	13.41
36	5.82	5.38	11.2	52	8.37	0.83	9.2

尤其是测试的结果存在比较明显的差异,所得 MAPE 值最大为 34.02 %,最小仅为 0.83 %. 方案 45 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6.61 %和1.78 %,总和为 8.39 %. 为了进一步比较测试的结果,从上述 60 组测试中,取 10 组较好的数据,分别计算其 MSE 和 MAE 值,如表 4 所示. 除方案 32 以外,其它 9 组训练及测试的 MSE 和 MAE 值均小于 7 ℃,而方案 45 和 52 的结果相差较小,均可作为后续使用的模型.

表 4　10 组方案的训练和测试 MSE 和 MAE

方案号	10	23	30	32	36	38	41	45	46	52
训练 MSE	6.55	5.35	5.45	4.25	5.21	5.45	6.05	5.43	5.54	6.38
测试 MSE	6.82	6.47	5.44	11.42	3.43	6.91	6.97	2.00	5.05	0.64
训练 MAE	5.33	3.76	3.92	3.41	3.61	4.37	4.58	4.13	4.52	5.26
测试 MAE	5.39	5.52	3.85	7.12	3.32	5.93	5.49	1.46	4.56	0.62

3　讨论

目前,对现有木聚糖酶进行遗传改造的方法,主要有定向进化法、定点突变法和计算机辅助法[7]. 前者属于非理性的蛋白设计范畴,采用随机的方法. 因此,存在筛选容量太大,筛选过程较复杂且费用较昂贵. 后两者属于理性蛋白设计范畴,它们利用了蛋白质分子的空间结构的数据,尽管也取得了一定的成功[11],但很大程度上依赖于经验. 最近,Liu 等人[8]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温度之间的数学模型. 其拟合的平均绝对百分比误差分别为 9.25 %和 3.94 %. 然而,文[8]利用二肽经逐步回归所得的方程,用于计算表 1 数据的最大和最小值,分别为 150.46 ℃和 - 29.06 ℃. 由此可见,本文预测的结果更理想. 这说明木聚糖酶的氨基酸组成和其最适温度间的关系非常复杂,用简单的线性模型可能得不到令人满意的结果. 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,可寻找出潜在的,可提高该酶最适温度的单个氨基酸残基(一般位于蛋白分子的表面). 结合本文所得的数学模型,可用计算机对突变体进行预筛选. 这样理性地减少突变库大小,以达到大大减少实验工作量,提高效率、节省费用,而且使得对木聚糖酶的改造更具有目的性. 同时,可以利用

实验后的数据,重新进行下一轮的拟合和预测,直到获得所需的结果为止. Voigt 等[12]利用计算机辅助的方法,将 β内酰胺酶的 7×10^{23} 可能突变体序列减少到 172 800 个,同时获得抗性提高了 1 280 倍的突变体. 尽管本文采用了均匀设计的方法对 BP 神经网络结构进行了优化,但在各因素各水平的选择上仍带有一定的随意性. 如果经过精心选择,网络的检测效果还会有所改善. 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,这仅仅是一种最简单的情形.

参 考 文 献

1 Xue Yiming, Mao Guizhong, Shao Weilan. Expression of xlyanase B gene of *Thermotoga maritime* in *Escherichia coli* [J]. Food and Fermentation Industries, 2003, 29(11): 20～25

2 Badhan A K, Chadha B S. Functionally diverse multiple xylanases of thermophilic fungus *Myceliophthora* sp. IMI 38709[J]. Enzyme and Microbial Technology, 2004, 35(5): 460～466

3 Kulkarni N, Shendye A. Molecular and biotechnological aspects of xylanases[J]. FEMS Microbiol. Rev., 1999, 23: 411～456

4 Fang Baishan, Chen Hongwen, Xie Xiaolan. The medium optimization of xylitol fermentation based on neural networks and genetic algorithms[J]. Chinese Journal of Biotechnology, 2000, 16(5): 648～650

5 Vladimir B, George R, Margo H, et al. Prediction of MHC class Ⅱ binding peptides using an evolutionary algorithm and artificial neural network[J]. Bioinformatics, 1998, 14(2): 121～130

6 Zhao Ziqiang, Huang Dashan, Sun Binyi. Human face recognition based on multi-features using neural networks committee[J]. Pattern Recognition Letters, 2004, 25(12): 1 351～1 358

7 Vieille C, Zeikus GJ. Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability[J]. Microbiol. Mol. Biol. Rev., 2001, 65(1): 1～43

8 Liu Lianwei, Wang Mingli, Shao Weilan, et al. A novel model to determine the dipeptides responsible for optimum temperature in F/10 xylanase[J]. Process Biochemistry, 2005, 40(3): 1 389～1 394

9 方柏山. 生物技术过程模型优化与控制[M]. 广州:暨南大学出版社, 1997. 185～212

10 方开泰. 均匀设计与均匀设计表[M]. 北京:科学出版社, 1994. 21～85

11 Ossi T, Kirsikka E. A combination of weakly stabilizing mutations with a disulfide bridge in the helix region of *Trichoderma* reesei endo-1,4-xylanase Ⅱ increases the thermal stability through synergism[J]. Journal of Biotechnology, 2001, 88 (1): 37～46

12 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., 2001, 98:3 778～3 783

A Model for Amino Acid Composition and Optimum Temperature in F/10 Xylanase

Zhang Guangya Fang Baishan

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

Abstract By using uniform design and neural network construction, the authors established a mathematical model for amino acid composition and optimal temperature of xylanases in F/10 family. As compared with the method of stepwise regression adopted by previous literature, the present model showed better results in fitting and prediction of optinal temperature. Under the conditions including learning rate of 0.1 and dynamic parameter of 0.6 and sigmoid parameter of 0.9 and knot number in the hidden layer of 7, the results of the present model in fitting and prediction of optimal temperatare can be shown respectively by mean absolute percent error of 6.6% and 1.78%, mean square error of 5.43 ℃ and 2.00 ℃, and mean absolute error of 4.13 ℃ and 1.46 ℃.

Keywords uniform design, BP neural network, xylanase, amino acid composition, optimum temperature