文章编号:1000-5013(2014)03-0316-05 **doi**:10.11830/ISSN.1000-5013.2014.03.0316

G/11木聚糖酶最适pH值的预测及其与氨基酸位置的关系

林源清,张光亚

(华侨大学 化工学院,福建 厦门 361021)

摘要: 把木聚糖酶全序列均分为N端,中间端(I端)及C端3个部分,并分别以全序列及分段氨基酸的组成作为模型输入值.通过主成分分析(PCA)方法探讨全序列及分段氨基酸组成和最适pH值的相关性,运用均匀设计法分别优化支持向量机和BP神经网络运行参数.研究结果表明:支持向量机获得的预测模型优于神经网络,其中RBF支持向量机是最佳的模型.主成分分析结果显示:I端主成分跟最适pH值相关性最高;相关系数R绝对值为0.68,得到的结果与支持向量机结果一致.

关键词: 木聚糖酶;氨基酸;支持向量机;BP神经网络;均匀设计;主成分分析

中图分类号: Q 55 **文献标志码**: A

木聚糖酶(EC3.2.1.8)是一种重要的工业用酶,可广泛应用于饲料、造纸、食品等行业.木聚糖酶的使用可大大减少造纸工业漂白过程中氯化物的用量,从而有效降低制浆造纸工业对环境的污染[1].用于造纸工业的木聚糖酶需满足耐热和耐碱条件,目前满足所需条件的酶来源于两种途径:一是从极端环境中筛选产酶菌株[2];二是通过基因工程对酶进行遗传改造[3].鉴于菌株筛选耗时长,效率低,基因工程改造越来越受研究者的关注.木聚糖酶可分为F/10和G/11家族,由于G/11家族的木聚糖酶分子较小,而且其结构更为简单,因此比较适合作为理论研究的分子模型[4].对于蛋白质的改造主要有两种策略:一是理性设计(rational design),即定点突变;二是非理性设计(irrational design),定向进化.定点突变目的明确,但需要事先了解蛋白质的结构;定向进化不需事先了解蛋白质的结构,但其筛选困难.本文利用木聚糖酶序列的信息和最适pH值,构建了氨基酸组成和最适pH值关系的模型.旨在探索影响酶最适pH值的氨基酸及其位置,为木聚糖酶的改造提供可靠的信息,以期提高研究效率. 本课题组曾利用木聚糖酶的氨基酸与最适pH值关系构建BP神经网络模型,并且取得较好的预测结果[5].采用均匀设计(UD)方法,构建氨基酸组成和最适pH值关系的模型.

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UniProt(http://www.uniprot.org/),木聚糖酶数据来源于文献[6].73个木聚糖酶ID号及最适pH值,如表1所示.表1中:ID为木聚糖酶在UniProt数据库中的收录号;pH_{opt}为文献中报道的木聚糖酶的最适pH值.

对于最适pH值在一定范围的,取其中间值.木聚糖酶的氨基酸组成分析由自行设计的软件完成.该软件不仅可以计算全段序列的氨基酸组成,还可以计算分段氨基酸组成.主成分分析由MVSP软件完成,神经网络及支持向量机由weka3.6.8软件完成.以各个木聚糖酶中全段序列及分段序列(将酶蛋白序列均分为3段,分别表示为序列的N端,C端及中间端)的20种氨基酸的组成百分比作为神经网络

收稿日期: 2013-09-18

通信作者: 张光亚(1975-),男,教授,主要从事生物信息与生物化工的研究.E-mail:zhgyghh@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(21376103);福建省自然科学基金资助项目(2013J01048)

和支持向量机的输入，其对应的最适 pH 值作为结果输出.

表 1　G/11 木聚糖酶 ID 号及最适 pH 值

Tab. 1　Xylanase ID in family G/11 and the optimum pH value

ID	pH_{opt}	ID	pH_{opt}	ID	pH_{opt}	ID	pH_{opt}
Q14RS0	6.00	P36217	5.00	Q12549	4.00	Q5NDZ1	7.00
P48793	5.00	Q9HFH0	5.00	C0LZ11	3.00	Q59962	6.25
Q9EW89	6.00	Q9UVF9	5.00	Q2I0I8	2.90	C6FGW6	2.60
P81536	6.25	Q6QA21	5.00	A4GG22	5.75	Q8J1V5	6.00
P35809	5.00	P48824	4.50	A3QRI7	6.00	D1FNQ6	6.50
Q7SID8	8.00	P55330	5.50	Q1XGE6	7.00	Q7ZA57	8.00
Q6VAY1	3.20	Q3S401	5.00	Q84F19	6.00	Q9HFA4	6.00
A2I7V2	6.00	Q2PU02	4.60	Q59256	7.00	D3KT79	5.00
Q96W72	3.50	P55332	5.50	A5H0S3	7.00	O77398	6.50
Q92397	2.00	O43097	6.50	B5SYI8	6.00	P26220	6.00
B0FIU1	3.00	B3VSG7	4.75	Q3HLJ4	8.50	D1KJJ7	7.00
D2KPJ0	8.00	Q58G72	7.50	C7F433	6.50	Q8J1V6	6.00
Q6U894	6.25	Q9HGE1	5.50	Q2PGY1	5.00	P17137	5.75
Q71S35	5.50	Q38Q19	7.50	D1GFE6	4.50	B8YQ34	6.25
P45705	8.00	Q4WG11	6.00	Q8J0T4	5.50	Q8J0K5	3.50
Q43993	7.00	Q8RMN7	6.50	Q96TR7	2.00	Q96UV7	4.50
P55328	3.50	B5M0C6	8.50	Q9UW17	4.80	P55333	5.50
P33557	2.00	Q06RH9	7.20	Q06562	6.00	B5A7N4	5.00
Q9UUQ2	2.00						

1.2　均匀设计的支持向量机

在运算时，支持向量机(SVM)[7]模型和 BP 神经网络[8]模型都需要选择参数，以达到最佳拟合结果. 因此，采用均匀设计法(UD)[9]来选择适当的运行参数. 定义两个特征指标[5]，即均方根误差 RMSE 和平均绝对误差 MAE. 模型预测的结果采用常用的“留一法”，即对 n 组数据，每次取 1 组作测试，其他 $n-1$ 组作为训练样本，共进行 n 次循环，使得样本中所有数据都能进行预测.

1.3　主成分分析

主成分分析(principal components analysis，PCA)又称主分量分析，把多指标转化为少数几个综合指标，在许多领域有着有效而广泛的应用[10]，是一种较为客观的综合评价方法. 运用 MVSP 软件，可直接获得 20 个氨基酸变量的主成分荷载和 73 个个案的主成分得分. 利用主成分得分与最适 pH 值进行拟合，拟合结果可在一定程度上综合反映氨基酸组成与最适 pH 值的关系.

2　结果与分析

2.1　基于均匀设计的支持向量机

利用均匀设计法，对两种不同核函数(Linear 和 RBF)的支持向量机运算参数进行优化，10 倍交叉验证结果，如表 2，3 所示. 表 2，3 中：MAE 为平均绝对误差；RMSE 为均方根误差. 限于篇幅，仅列出最优预测结果.

表 2　基于 Linear 核函数的支持向量机预测结果

Tab. 2　Result of SVM prediction based on linear kernel

参数		N 端		I 端		C 端		全段	
C	ε	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.50	0.200 0	1.12	1.52	0.90	1.14	1.05	1.32	1.07	1.38
0.05	0.000 1	1.05	1.36	1.00	1.26	1.06	1.36	1.28	1.60
0.01	0.700 0	1.28	1.60	1.28	1.60	1.28	1.60	0.83	1.12
100 000.00	0.300 0	1.20	1.59	0.99	1.28	0.92	1.23	1.28	1.60

由表 3 可知：在 RBF 核函数支持向量机模型中，以 I 端氨基酸组成作为输入，得到的预测结果最

佳,即 $C=1$,$\varepsilon=0.1$,$\gamma=0.5$ 时,其 MAE 和 RMSE 值均最小,分别为 0.84 和 1.17. 此时,所建立的模型对木聚糖酶最适 pH 值预测准确率最高,故为最佳方案.

表 3 基于 RBF 核函数的支持向量机预测结果

Tab. 3 Result of SVM prediction based on RBF kernel

参数			N 端		I 端		C 端		全段	
C	ε	γ	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	0.100	0.500	1.00	1.40	0.84	1.17	0.86	1.16	0.89	1.18
50	0.150	0.001	1.04	1.41	0.92	1.18	1.04	1.38	0.88	1.19
1 000	0.001	1.500	0.97	1.28	0.97	1.29	1.01	1.31	0.96	1.27

通过比较两种核函数的预测结果,可以得知 RBF 核函数的整体预测结果优于 Linear 核函数. 尽管在 Linear 核函数中,当惩罚值 $C=1$,$\varepsilon=0.005$,其 MAE 为 0.83,是所有预测结果中最小的. 这个结果说明,在 Linear 核函数中运行参数取得了比较理想的结果. 如果对 RBF 核函数进一步优化,可能会取得更好的结果. 根据表 3 中的最优化参数 $C=1$,$\varepsilon=0.1$,$\gamma=0.5$,使用支持向量机法建立最适 pH 值模型. 通过该模型对实际测得的数据(pH_{exp})进行预测,预测结果(pH_{pre})如图 1 所示. 从图 1 可知:该模型预测结果与实际测得结果的相关性为 0.67,说明该模型可行.

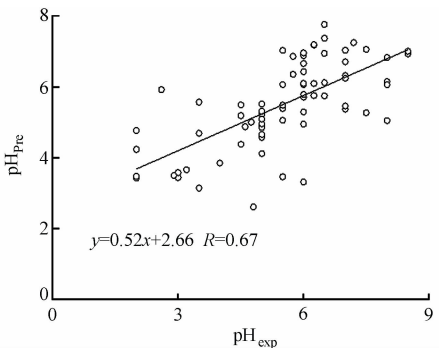


图 1 预测值和实测值的关系

Fig. 1 Relationship between experimental and predicted transition temperature obtained

2.2 基于均匀设计的 BP 神经网络

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,选择一个隐含层的神经网络,对学习速率、动态参数和隐含层结点数 3 个因素 15 水平进行均匀设计,所得的均匀设计表和训练结果,如表 4 所示(仅列出最优预测结果).

由表 4 可知:当学习速率(v)为 0.06,动态参数(MP)为 0.2,隐含层结点数(NHL)为 8 时,以 I 端氨基酸组成为输入的模型,对最适 pH 值拟合的均方根误差为 1.49 个 pH 值单位,平均绝对误差为 1.09 个 pH 值单位,具有很好的拟合效果.

表 4 BP 神经网络的预测结果

Tab. 4 Result of BP neural network

参数			N 端		I 端		C 端		全段	
v	MP	NHL	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.04	0.6	11	1.76	2.36	1.14	1.49	1.30	1.75	1.37	1.81
0.08	0.7	5	1.65	2.27	1.24	1.69	1.49	2.06	1.77	2.38
0.06	0.2	8	1.80	2.43	1.09	1.49	1.31	1.69	1.31	1.72

由表 2~4 可知:构建的 3 种模型中,基于 RBF 核函数的支持向量机模型的整体预测结果最佳;I 端的预测结果在分段预测模型中均最佳;其次是 C 端;最后是 N 端(表 5). 这个结果说明 I 端与木聚糖酶的最适 pH 值相关性最高.

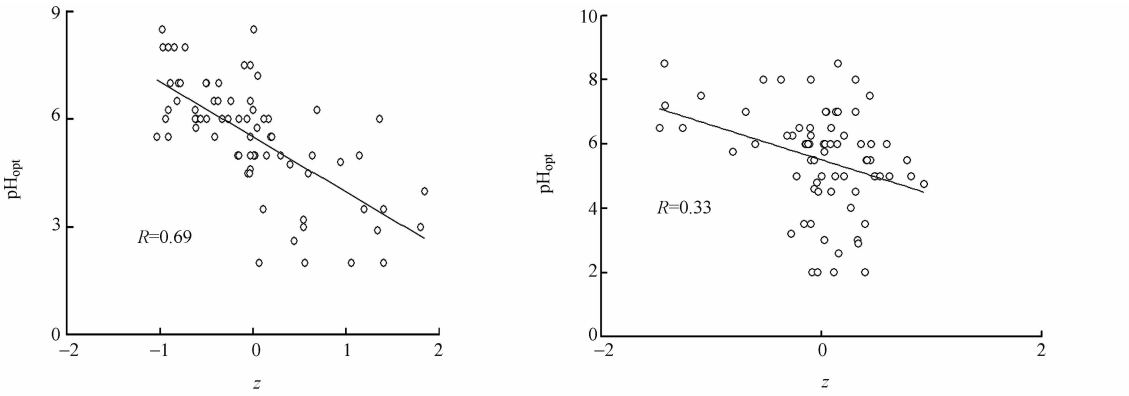
表 5 3 种模型 3 端最佳优化结果

Tab. 5 Optimum result of 3 segments in the three different models

氨基酸位置	线性函数		径向机函数		BP 神经网络	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
I 端	0.90	1.14	0.84	1.17	1.09	1.49
N 端	0.92	1.23	0.86	1.16	1.30	1.75
C 端	1.05	1.36	1.00	1.40	1.65	2.27

2.3 氨基酸与最适 pH 值的相关性

原始数据运用 MVSP 软件做主成分分析(PCA)分析后,可得到 20 个氨基酸变量的主成分和 73 个个案主成分. 根据个案得分主成分(z),做主成分与实测最适 pH 值的相关性图,如图 2 所示.



(a) 全段最佳拟合结果　　(b) N 端最佳拟合结果

(c) I 端最佳拟合结果　　(d) C 端最佳拟合结果

图 2　主成分与最适 pH 值相关性图

Fig. 2　Relationship between principal components and optimum pH value

由图 2 可知：全段序列的相关性最好，R 值为 0.69. N 端、I 端、C 端与最适 pH 值的相关性大小分别为－0.33，－0.68，0.40，其中负值表示负相关. 它们与最适 pH 值影响的大小顺序依次为 I 端、C 端、N 端. 这个结果和前文的支持向量机和 BP 神经网络的结果一致，验证了前文所构建模型的可靠性. 值得注意的是全段序列和 I 端序列的 R 值数值相近，但是方向却相反.

由于 I 端序列氨基酸组成对于木聚糖酶的最适 pH 值影响较大. 因此，仅列出 I 端分析结果. 原始数据运用 MVSP 软件 PCA 分析后，得到 20 个氨基酸变量的 5 个主成分.

各氨基酸与 5 个主成分之间的关系，如表 6 所示. 表 6 中：相关系数只保留一位小数，且仅列出绝对值大于 0.2 的氨基酸. 从表 6 可知：第 1 主成分与丝氨酸(S)相关性最强，相关性高达 0.8，该结果表明丝氨酸(S)是木聚糖酶的关键氨基酸；第 2 主成分与甘氨酸(G)、第 3 主成分与甘氨酸(G)、第 4 主成分与酪氨酸(Y)及天冬酰胺(N)的相关性显著，表明这 3 种氨基酸是木聚糖酶比较重要的氨基酸. Liu 等[11]研究结果表明：G/11 家族主成分分析的前 7 个主成分所代表的是该家族木聚糖酶的 2 级结构，分别为：卷曲、转角、折叠、转角、转角、螺旋和折叠.

表 6　木聚糖酶 20 种氨基酸与各主成分的关系

Tab. 6　Relationship between 20 amino acids and principle components in xylanase

主成分	氨基酸	
	正相关	负相关
Pr1	0.8S 0.2A	0.2T 0.2N 0.3G 0.3R
Pr2	0.6G 0.3N 0.2F 0.2S 0.2K	0.2D 0.2A 0.2R 0.3Y 0.5T
Pr3	0.6G 0.5T 0.3S 0.3Y	0.2A 0.2K 0.2N 0.2I
Pr4	0.5Y 0.3Q 0.3A 0.2R 0.2G	0.2K 0.2F 0.3D 0.3V 0.4T
Pr5	0.6N 0.3S 0.2Y 0.2R 0.2V 0.2P	0.2D 0.2H 0.2F 0.3G 0.3A 0.3Q

3 结束语

构建了不同的最适pH值预测模型,其中基于RBF核函数的支持向量机模型预测木聚糖酶的最适pH值的精度,比使用BP神经网络及Linear核函数的支持向量机模型更好,可做为木聚糖酶模拟的后续使用模型.采用了均匀设计的方法对构建的模型进行了参数优化,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,模型的预测效果还会有所改善.此外,由于木聚糖酶分子量较小、结构比较简单,只有一条多肽链,基于此酶所建立的模型对于其他具有4级结构的复杂酶类是否仍然适用仍有待探讨.

参考文献:

[1] 聂国兴,王俊丽,明红.木聚糖酶的应用现状与研发热点[J].工业微生物,2008,38(1):53-59.

[2] 包怡红,刘伟丰,毛爱军,等.耐碱性木聚糖酶高产菌株的筛选、产酶条件优化及其在麦草浆生物漂白中的应用[J].农业生物技术学报,2005,13(2):235-240.

[3] UMEMOTO H,YATSUNAMI R,INAMI M,et al. Improvement of alkaliphily of bacillus alkaline xylanase by introducing amino acid substitutions both on catalytic cleft and protein surface[J]. Bioscience Biotechnology and Biochemistry,2009,73(4):965-967.

[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]. Journal of Biotechnology,2002,95(2):109-131.

[5] 张光亚,方柏山.木聚糖酶氨基酸组成与其最适pH值的神经网络模型[J].生物工程学报,2005,21(4):658-661.

[6] PAES G,BERRIN J G,BEAUGRAND J. GH11 xylanases: Structure/function/properties relationships and applications[J]. Biotechnology Advances,2012,30(3):564-592.

[7] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag,2000:138-167.

[8] 王轶夫,孙玉军,郭孝玉.基于BP神经网络的马尾松立木生物量模型研究[J].北京林业大学学报,2013,35(2):17-21.

[9] 方开泰.均匀设计-数论方法在试验设计的应用[J].应用数学学报,1980(4):363-372.

[10] 王志江.主成分分析法在地区企业经济效益评价中的应用[J].华侨大学学报:自然科学版,2004,25(3):322-325.

[11] LIU Liang-wei,ZHANG Jue,CHEN Bin,et al. Principle component analysis in F/10 and G/11 xylanase[J]. Biochemical and Biophysical Research Communications,2004,322(1):277-280.

Prediction of Optimum pH of G/11 Xylanases and the Relationship between the Location of Amino Acid and Optimum pH Value

LIN Yuan-qing, ZHANG Guang-ya

(College of Chemical Engineering, Huaqiao University, Xiamen 361021, China)

Abstract: We divided the xylanase sequences into three equally segments named N-terminus, I-terminus and C-terminus. And then, we calculated the amino acid compositions of the whole sequences and the segmented sequences, respectively. The amino acid compositions were used as the input values of these models. The principal component analysis (PCA) method was utilized to analyze the relationship between the amino acid composition and the optimum pH. The uniform design was used to optimize the running parameters of support vector machines (SVM) and neural network (BPNN), respectively. Our results showed the predicted model obtained by SVM was better than that of BPNN, and the SVM model based on RBF kernel was best. The results of PCA showed the correlation between principle component and optimum pH was best in the I-terminus with the $R=-0.68$, which coincided with the result of the SVM.

Keywords: xylanase; amino acid; support vector machine; BP neural network; uniform design; principle component analysis

(**责任编辑**: 陈志贤 **英文审校**: 刘源岗)