文章编号: 1000-5013(2007)03-0292-04

木聚糖酶特征序列与最适 pH 定量关系

张光亚, 方柏山

(华侨大学 工业生物技术福建省高等学校重点实验室, 福建 泉州 362021)

摘要: 研究 G/11 家族木聚糖酶特征序列与其最适 pH 的定量构效关系, 建立逐步回归方程, 模型的相关系数为 0.975, 显著水平 $p < 0.0001$. 利用该方程, 计算各木聚糖酶的 pH 值, 结果表明, 仅有 2 个木聚糖酶的最适合 pH 计算值与实测值相差大于 0.6, 而有 6 个木聚糖酶最适 pH 计算值与测定值相差小于 0.1, 平均绝对百分比误差为 5.91%, 平均绝对误差为 0.26 个 pH 单位, 所得结果优于用二肽进行逐步回归所得模型的结果. 通过对计算结果和已知晶体结构的木聚糖酶的比较发现, 在 1YNA 中, 这 6 个氨基酸分别是 Pro, Leu, Val, Val, Val 和 Tyr, 前 3 个氨基酸残基位于第 7 个转角处, 更适合作为突变的位点.

关键词: 木聚糖酶; 最适 pH; 逐步回归; 定量构效关系; 特征序列

中图分类号: Q 814; Q 501; Q 539 文献标识码: A

近年来, 有关蛋白质结构与功能关系的研究受到了国内外研究者的重视[1]. 随着酶蛋白在工农业和医药上的应用日益广泛, 人们试图利用对蛋白质构效关系的认识, 进而设计具有新特性的蛋白质. 目前, 对蛋白质结构与功能关系的研究大多为定性描述, 而对蛋白质定量构效关系(Quantitative Structure Activity Relationship, QSAR)的研究比较缺乏[2]. QSAR 技术在研究小分子物质和一些多肽上取得了不少成功的例子[3-4], 但由于蛋白质分子相对于小分子物质和多肽而言, 其分子量更大, 空间结构的测定存在较大困难, 描述其结构的相关参数难以获取, 因而给蛋白质 QSAR 研究带来了很大困难. 分子生物学技术的迅猛发展, 使蛋白质序列信息的获取比较容易, 这为建立蛋白质定量序列-功能关系的研究提供了契机. 本文利用结构较为简单, 用于纸浆漂白更有优势的 G/11 家族木聚糖酶的序列信息及对应的最适 pH, 从该家族的特征序列出发, 利用逐步回归建立了序列及其最适 pH 的模型, 获得了优于文[5]报道的, 以二肽量进行逐步回归的结果.

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 和程序 PDB 数据库(http://www.rcsb.org/pdb/), 前者是一个非冗余的专家库. 前 21 个木聚糖酶最适 pH 实验值取自文[5], 1F5J 的序列和 pH 值来自文[6]. G/11 家族木聚糖酶两个特征序列 PS00776 和 PS00777 来源于 PROSITE 数据库(http://us.expasy.org/prosite/), 其调和序列分别为[PSA]-[LQ]-x-E-Y-Y-[LIVM](2)-[DE]-x-[FYWHN] 和[LIVMF]-x(2)-E-[AG]-[YWG]-[QRFGS]-[SG]-[STAN]-G-x-[SAF]. 其中, [ ]中的氨基酸表示在该位点可以是其中的任何一个, x 表示该位点可以是 20 种天然氨基酸中任一个. 可见, 对于某个特定序列, 对这两个区域进行突变的可能选择比较多. 已有研究证实[7], 对蛋白质保守区进行突变可提供足够的突变体, 并从中筛选到活性显著提高(K_{cat}/K_m 提高了 10 000 倍)的酶.

1.2 序列的编码

收稿日期: 2006-11-10

作者简介: 张光亚(1975-), 男, 讲师, 博士研究生, 主要从事酶工程与生物信息学的研究; 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: 国务院侨办科研基金资助项目(05Q0018)

1.2.1　*Z* 标度　Hellberg 等[8]对氨基酸的 29 个物理化学性质进行主成分分析, 得到了 3 个显著的主成分, 并将相应主成分得分矢量作为新的氨基酸描述子, 称为主性质, 即 *Z* 标度, 从而成功地建立了多肽的定量构效关系模型. 这一标度目前得到了广泛应用[9]. 20 种氨基酸的 *Z* 标度, 如表 1 所示.

表 1　20 种氨基酸的 *Z* 标度

Tab. 1　Descriptor *Z* scales for amino acids

氨基酸	z_1	z_2	z_3	氨基酸	z_1	z_2	z_3
Ala (A)	0.07	− 1.73	0.09	Leu (L)	− 4.19	− 1.03	− 0.98
Arg (R)	2.88	2.52	− 3.44	Lys (K)	2.84	1.41	− 3.14
Asn (N)	3.22	1.45	0.84	Met (M)	− 2.49	− 0.27	− 0.41
Asp (D)	3.64	1.13	2.36	Phe (F)	− 4.92	1.30	0.45
Cys (C)	0.71	− 0.97	4.13	Pro (P)	− 1.22	0.88	2.23
Gln (Q)	2.18	0.53	− 1.14	Ser (S)	1.96	− 1.63	0.57
Glu (E)	3.08	0.39	− 0.07	Thr (T)	0.92	− 2.09	− 1.40
Gly (G)	2.23	− 5.36	0.30	Trp (W)	− 4.75	3.65	0.85
His (H)	2.41	1.74	1.11	Tyr (Y)	− 1.39	2.32	0.01
Ile (I)	− 4.44	− 1.68	− 1.03	Val (V)	− 2.69	− 2.53	− 1.29

1.2.2　*序列的编码*　将字符序列编码为数字序列的方法有两种, 本文使用较常用的 *Z* 标度方法编码氨基酸序列, 编码过程如图 1 所示. 1 个含 3 个氨基酸的短肽, 可用 9 个变量进行描述. 变量用 $x_{i,j}$ 表示, 其中, i 表示位点, i= 1, 2, 3, …, i; j 表示 *Z* 标度在位点处的 3 个主成分得分矢量, j = 1, 2, 3. 从 G/11 家族木聚糖酶两个特征序列 PS00776 和 PS00777 的调和序列可看出, 在 PS00776 中第 4 位 E, 第 5 位 Y 和第 6 位 Y, PS00777 第 4 位 E 和第 10 位 G 在所有的序列中均存在, 作为常数项. 因此, 这 5 位的氨基酸未进行转换, 剩下 18 个位点的氨基酸按照它们在一级结构上的顺序进行排列, 分别编号为 1, 2, …, 18. 用 *Z* 标度表征这 18 个肽序列的结构特征, 共产生 54 个变量. 22 个木聚糖酶最适 pH 为目标函数.

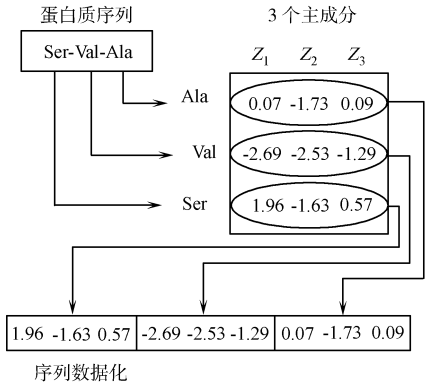

图 1　*Z* 标度描述序列结构特征的流程图

Fig. 1　Flow chart of structural description of the sequences by *Z* scales

1.3　计算步骤

整个计算过程分为 5 个步骤. (1) 从 22 条序列中分别找出上述两个特征序列, 总长度为 23 个氨基酸残基. (2) 找出其中的 5 个保守氨基酸残基并剔除, 剩 18 个氨基酸残基. (3) 对字符序列进行数字编码, 每条序列均产生 54 个变量, 所得到数字矩阵为 22 × 54. (4) 用数据处理系统 (DPS) 软件将 54 个变量与其对应最适 pH 进行逐步回归, 确定显著相关的变量. (5) 根据确定的变量, 建立回归方程.

2　结果与分析

2.1　回归方程的建立

t 检验结果表明, 有 6 个变量均达到极显著水平 (p < 0.006), 所得回归方程 (系数仅保留 2 位小数) 为 pH_{opt} = 0.60 + 1.28 × $x_{1,3}$ − 0.73 × $x_{2,1}$ + 0.49 × $x_{3,1}$ − 1.99 × $x_{5,3}$ + 0.31 × $x_{9,1}$ − 1.48 × $x_{17,3}$. 其中, pH_{opt} 表示最适 pH, 模型的相关系数为 0.975, 显著水平 p < 0.000 1. 说明 pH_{opt} 与这些变量存在很好的相关性. 利用该回归方程, 计算出各木聚糖酶的最适 pH, 如表 2 所示. 表中, ID 为木聚糖酶 ID 号, pH_{opt} 为最适 pH 值, pH_{pre1}, pH_{pre2}, pH_o 分别为文[7] 和本文的计算结果以及实验值. 由表 2 可知, 仅有 2 个木聚糖酶 (P33557 和 P55328) 最适 pH 的计算值与测定值相差大于 0.6, 而有 6 个木聚糖酶 (P45705, P55334, P35809, P36217, P36218 和 1F5J) 最适 pH 计算值与测定值相差小于 0.1, 平均绝对百分比误差为 5.91%, 平均绝对误差为 0.26 个 pH 单位, 所建立的回归方程具有相当精度. 文[5] 用二肽含量进行逐步回归所得模型的平均绝对百分比误差为 9.36%, 平均绝对误差为 0.39 个 pH 单位. 由此可见, 本

文所得结果优于文[5].

2.2 回归方程的作用

从上述方程可知,在所研究的 18 个位点中有 6 个位点对木聚糖酶最适 pH 的影响较大,而其他位点影响较小,这意味着可能存在着大量的中性突变. 根据上述方程的调和序列,可计算出 pH_{opt} 的最大和最小值分别为 16.09 和 -8.98. 尽管这种计算的结果并不完全具有生物学意义,但至少可以说明通过蛋白质工程的手段对该 6 个位点进行突变,可使得木聚糖酶有较宽的 pH 忍受范围. 在上述 6 个位点中,有两个位点(3 和 17)的保守性很低(调和序列中用 x 表示的位点),另两个位点(5 和 9)也可以有多种氨基酸的存在,分别为[LIVM]和[LIVMF]. 由此可见,对于某一条特定木聚糖酶氨基酸序列,在这几个位点进行单位点和多位点突变可以有很多种. 6 个变量中 $x_{2,1}$ 的 t 检验值最大(11.03),显著水平 $p < 0.000\ 01$. 通过检查序列发现,所有的碱性木聚糖酶($pH_{opt} > 5.0$)[5] 在该位点均为 Leu(Z_1 值为 -4.19),而 $pH_{opt} \leqslant 3.0$ 的木聚糖酶在该位点均为 Gln(Z_1 值为 2.18). 计算结果和已知晶体结构的木聚糖酶进行比较,如图 2 所示. 利用 Swiss-PdbViewer 软件(http://au.expasy.org),呈现了上述方程中 6 个位点的氨基酸残基. 在 1YNA 中,这 6 个氨基酸分别是 Pro, Leu, Val, Val, Val 和 Tyr,前 3 个氨基酸残基位于第 7 个转角处,后 3 个氨基酸残基分别位于第 8,第 13,第 14 个 β-折叠上,而这转角和 β-折叠的柔性比其他二级结构(α-螺旋)更强,更适合作为突变的位点.

表 2 G/11 家族木聚糖酶

Tab.2 Xylanase in family G/11

ID	pH_{opt}	基于二肽		基于序列	
		pH_{pre1}	$\Delta(pH_{opt} - pH_{pre1})$	pH_{pre2}	$\Delta(pH_{opt} - pH_{pre2})$
P45705	8.00	7.52	0.48	7.93	0.07
P29127	8.00	8.19	-0.19	7.82	0.18
P55332	6.50	6.31	0.19	6.77	-0.27
O43097	6.50	5.82	0.68	6.88	-0.38
P00694	6.50	6.49	0.01	6.71	-0.21
P26515	6.50	6.04	0.46	6.12	0.38
P55334	6.5	7.04	-0.54	6.49	0.01
P35811	6.30	6.05	0.25	6.05	0.25
Q06562	6.00	5.83	0.17	5.65	0.35
P35809	5.90	5.37	0.53	5.99	-0.09
P55333	5.50	5.83	-0.33	5.19	0.31
P09850	5.50	5.48	0.02	5.26	0.24
P29126	5.50	5.99	-0.49	5.68	-0.18
P36217	5.25	5.36	-0.11	5.19	0.06
P48793	5.00	4.69	0.31	5.19	-0.19
P18429	5.00	5.48	-0.48	5.26	-0.26
P48824	4.50	4.82	-0.32	4.72	-0.22
P36218	4.25	4.68	-0.43	4.31	-0.06
P55328	3.50	3.34	0.16	2.83	0.67
P55329	3.00	3.34	-0.34	2.83	0.17
P33557	2.00	3.68	-1.68	2.83	-0.83
1F5J	5.60			5.58	0.02

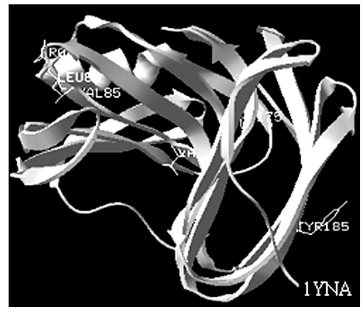

图 2 木聚糖酶的三级结构

Fig.2 3D structures of xylanase

3 结束语

最近,有研究者[10] 构建了 G/11 家族木聚糖酶氨基酸组成与其最适 pH 的神经网络模型,取得了较好拟合和预测效果. 本文结果略逊于该文献用神经网络构建的模型. 但基于氨基酸组成难以给出确切的位点信息,因此,仅适用于对产生的序列进行虚拟筛选. 文[5]利用二肽含量得出的回归方程,尽管能在一定程度上给出位点信息,但计算精度不如本文结果. 而且,利用本方程,可以根据突变目的,预先明确地选择特定位点,以及将该位点某个氨基酸突变成何种氨基酸,从而有利于实现突变目的. 使得突变更具目的性,可提高突变成功率. 总之,在给定的某个序列中,本模型不但能提供突变位点的信息,而且可以对突变后的特性有预先的判断,这对指导定点突变(定向进化)更具意义. 同时,如果能将该方法与随机突变(定向进化)所得序列-特性的数据相结合,在获得足够序列-特性数据的基础上,可用偏最小二乘回归法建立全长序列与特性的定量模型,使得序列中各位点的信息都能在方程

中得以体现, 再以此来指导突变, 可进一步提高成功率. 因此 Claes 等[11] 认为, 将传统用于工程的数学方法及数据挖掘工具与现代分子生物学手段相结合, 可以大大推进蛋白质工程的发展, 降低研究费用, 以及克服高通量筛选可能存在的不可靠性. 本文对此进行了尝试.

参考文献:

[1] LUO Yi, JIANG Xi-lin, LAI Lu-hua. Modeling protein backbone structure based on C_α guiding coordinates[J]. Protein Eng, 1992, 5: 147-152.

[2] LIANE S U, HUMBERTO G D, EUGENIO U. Proteins markovian 3D-QSAR with spherically-truncated average electrostatic potentials[J]. Bioorgan Med Chem, 2005, 13: 3641-3647.

[3] 梅　虎, 周　原, 孙立力, 等. 一种新的氨基酸描述子及其在肽 QSAR 中的应用[J]. 物理化学学报, 2004, 20 (8): 821-825.

[4] LIU Lin, MURRAY M, BURCHER E. Structure-activity studies of bufokinin, substance P and their C-terminal fragments at bufokinin receptors in the small intestine of the cane toad Bufo marinus[J]. Biochem Pharmacol, 2002, 63: 217-224.

[5] LIU Liang-wei, LI Xiang-qian, LI Xin, et al. Computational analysis of responsible dipeptides for optimum pH in G/11 xylanase[J]. Biochem Bioph Res Co, 2004, 321: 391-396.

[6] CEPELJNIK T, KRIZAJ I, MARINSEK L R. Isolation and characterization of the *Pseudobutyrivibrio xylanivorans* Mz5T xylanase XynT-the first family 11 endoxylanase from rumen butyrivibrio-related bacteria[J]. Enzyme Microb Technol, 2004, 34: 219-227.

[7] CASTLE L A, SIEHI D L, GORTON R, et al. Discovery and directed evolution of a glyphosate tolerance gene[J]. Science, 2004, 304: 1151-1154.

[8] HELLBERG S, SJOSTROM M, SKAGERBERG B, et al. Peptide quantitative structure-activity relationships, a multivariate approach[J]. J Med Chem, 1987, 30: 1126-1135.

[9] ANDERSSON P M, SJOSTROM M, LUNDSTEDT T. Preprocessing peptide sequences for multivariate sequence-property analysis[J]. Chemometr Intell Lab Sys, 1998, 42: 41-50.

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

[11] CLAES G, SRIDHAR G, JEREMY M. Putting engineering back into protein engineering: Bioinformatics approaches to catalyst design[J]. Curr Opin Biotech, 2003, 14: 366-370.

Quantitative Relationship of Characteristic Sequence and Optimum pH in G/11 Family Xylanases

ZHANG Guang-ya, FANG Bai-shan

(Laboratory of Industrial Biotechnology of Fujian Province University, Huaqiao University, Quanzhou 362021, China)

Abstract: The quantitative structure-activity relationship (QSAR) of xylanase in G/11 family was studied based on the the stepwise regression, the correlation coefficient of the model was 0.975 and model reached a significant level ($p < 0.000\ 1$). The optimal pH of xylanases was calculated based on the model and the results was shown as below: there were only two xylanses with the deviation between the calculated pHs the optimum pHs larger than 0.6, while there 6 xylanases with the deviation less than 0.1 the mean absolute percent error was 5.91%, the mean absolute error were 0.26 pH unit, respectively. It was superior when compared with the reported stepwise regression model based on dipeptide composition. Meanwhile, by comparing the calculating results with the crystal structure of xylanse, in 1YNA, the six amino acids were Pro, Leu, Val, Aal and Tyr, the first three amino acids were located in the seventh turn, which meant that they were suit for mutation.

Keywords: xylanase; optimum pH; stepwise regression; quantitative sequence-property relationship; characteristics sequence

(责任编辑: 黄仲一)