

文章编号: 1000-5013(2011)02-0194-04

使用伪氨基酸组成和 BP 神经网络预测 类弹性蛋白多肽的相变温度

黄凯宗, 张光亚

(华侨大学 化工学院, 福建 泉州 362021)

摘要: 根据获得的 16 条 ELP 序列及相变温度的数据, 利用伪氨基酸组成方法提取其序列特征值. 将伪氨基酸组成中的相关系数部分作为类弹性蛋白的特征向量, 从类弹性蛋白序列出发, 利用最小中位方差回归, 找出与其序列相关系数的最佳阶数. 运用均匀设计法, 分别对支持向量机与 BP 神经网络参数进行优化. 结果表明: BP 神经网络获得的预测模型最佳, 相变温度绝对误差为 0.39 °C, 均方根误差为 0.89 °C.

关键词: 类弹性蛋白; 相变温度; 伪氨基酸组成方法; 支持向量机; BP 神经网络

中图分类号: Q 516.02

文献标志码: A

类弹性蛋白多肽(Elastin-Like Polypeptides, ELPs)是一种具有弹性功能且对环境非常敏感的生物高分子, 它由五肽重复序列单元构成. 如果环境温度低于 ELP 的相变温度, 则该多肽在水溶液中是高度可溶的, 聚合物链就保持无序结构, 且相当伸展; 反之, 当环境温度高于相变温度时, 这一含水的多肽链结构就会瓦解, 并开始聚集, 形成一个富含 ELPs 的聚集物^[1]. 利用类弹性蛋白的可逆相变特性, 使其在蛋白纯化、药物载体、组织工程等方面得到广泛的应用^[2]. Urry 等^[3]认为, 相变温度是关于 ELP 序列、多肽链长度、Xaa 种类摩尔分数的函数. Chilkoti 等^[4]利用重组基因进行克隆表达, 得到了在序列和多肽链长均能精确控制的 ELP. 他们用非线性回归分析描述了 ELP 序列链长及浓度与相变温度的关系, 但所得到的模型仅能预测 3 种 ELP 文库的相变温度. 本文根据获得的 16 条 ELP 序列及相变温度的数据, 利用伪氨基酸组成方法提取其序列特征值, 采用 BP 神经网络、支持向量机方法、最小中位方差回归预测 ELP 的相变温度值.

1 材料与方法

1.1 试验数据来源

文中所用的数据取自于文献[5].

1.2 伪氨基酸组成

伪氨基酸组成包含 $20 + \lambda$ 个变量, 最早由 Chou 等^[6]提出. 由于文中所涉及的 ELP 氨基酸组成极为相似, 而且种类很少, 为了减少输入变量数目, 对其略作调整, 仅取其后的 λ 个变量, 即氨基酸相关系数. ELP 相关系数的阶数 λ 从 1 取到 10, 氨基酸相关系数计算参见文献[7].

1.3 均匀设计

在运行时, 支持向量机(SVM)^[8]和 BP 神经网络^[9]都需要选择参数, 以达到最佳效果. 因此, 采用均匀设计法(UD)^[10]来选择适当的运行参数. 定义 3 个特征指标^[11], 即平均绝对百分比误差 δ_{MPAE} 、均方根误差 δ_{MSE} 和平均绝对误差 δ_{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.

收稿日期: 2009-09-21

通信作者: 张光亚(1975-), 男, 副教授, 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(20806031); 福建省自然科学基金资助项目(2009J01030)

2 结果与分析

2.1 氨基酸相关系数的阶数的选择

根据文献[6],氨基酸相关系数的阶数(λ)是伪氨基酸组成一重要参数. 文献数据的相变温度呈离散分布,使用最小中位方差回归会更为精确^[11-12],且运行过程中无需调整参数.

参数 λ 经最小中位方差(Least Median of Squares Regression,LMSQ)回归检测,获得的平均绝对百分比误差 δ_{MPAE} 、均方根误差 δ_{MSE} 和平均绝对误差 δ_{MAE} 关系,如表 1 所示.由表 1 可知,当 $\lambda=8$ 时, δ_{MAE} 为 3.04, δ_{MSE} 为 5.73, δ_{MPAE} 为 40.91%.即拟合所得 ELP 相变温度准确率最高,因此取 $\lambda=8$.

表 1 氨基酸相关系数的阶数对特征指标的影响

Tab.1 Effect of the order of correlation coefficient for amino acids on characteristic index

特征指标	λ									
	1	2	3	4	5	6	7	8	9	10
δ_{MAE}	3.15	3.11	3.10	3.15	3.14	3.63	3.18	3.04	3.55	3.22
δ_{MSE}	6.08	6.08	6.074	6.06	6.01	6.32	5.98	5.73	6.445	5.93
$\delta_{\text{MPAE}}/\%$	42.43	41.10	41.70	42.44	42.31	48.86	42.78	40.91	47.72	43.33

当 $\lambda=8$ 时,执行最小中位方差回归得到 ELP 的相变温度拟合模型为

$$y = 0.49x_2 - 2.02x_1 + 0.46x_3 - 2.31x_4 + 1.30x_5 - 2.04x_6 + 0.48x_7 + 0.31x_8 - 0.18x_9 - 1.07x_{10} + 80.49.$$

(1)

其中: $x_1 \sim x_8$ 分别为伪氨基酸组中相关系数; $x_9 \sim x_{10}$ 分别为 ELP 的相对分子质量、ELP 每一单体的 Xaa 数量;ELP 浓度对 ELP 相变温度没有影响,故为其相关系数零.

从模型(1)可见,第 1,第 4 和第 6 个相关系数对相变温度有较大的负面影响,而第 5 个相关系数则有较大的正面影响;伪氨基酸组的相关系数对 ELP 的相变温度影响较大.当 ELP 浓度较高时,其浓度在一定范围变化对相变温度几乎不影响.这与 Chilkoti 等^[4]的实验结果较为一致.

使用最小中位方差回归获得的拟合值与实测值关系,如图 1 所示.由图 1 可知,一些拟合值非常好,而另外一些预测值与实测值差距比较大,从而导致其回归直线的斜率偏离较大.

2.2 利用支持向量机预测相变温度

表 2 支持向量机运行参数的选择

Tab.2 Selection of running parameters of SVM

方案	C	ϵ	γ	δ_{MAE}	δ_{MSE}	$\delta_{\text{MPAE}}/\%$
1	10	0.5	5.00	8.72	11.44	110.02
2	0.10	0.80	1.00	8.72	11.44	110.02
3	0.005	0.40	0.09	7.62	10.10	96.20
4	1.00	0.10	0.50	4.69	5.91	59.15
5	5.00×10^4	0.20	0.100	5.56	6.76	70.15
6	0.01	0.01	0.90	7.33	9.26	92.55
7	100.00	1.00×10^{-5}	0.30	1.85	3.31	23.39
8	1.00×10^4	0.05	0.01	3.29	4.53	41.56
9	50.00	0.15	0.001	6.83	8.75	86.24
10	0.50	0.60	0.005	8.72	11.44	110.02
11	0.05	1.00×10^{-4}	0.03	7.12	9.14	89.74
12	5.00	0.005	0.07	4.14	6.65	52.26
13	500	1.00	0.05	8.72	11.44	110.02
14	5 000	0.70	0.10	8.72	11.44	110.02
15	1 000	0.001	1.50	3.24	5.66	40.92
默认值	1.00	0.001	0.01	6.61	8.84	83.41

与 实际 测量 值的 相关 系数 达 0.93,模型 预测 的 结果 较好.

2.3 利用神经网络预测相变温度

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,直接决定网

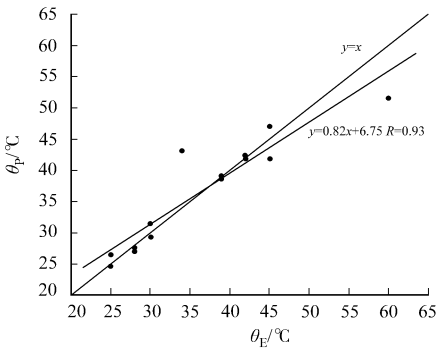
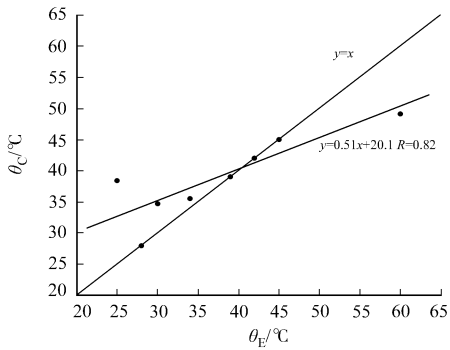


图 1 利用最小中位方差回归的拟合值与实测值关系

Fig. 1 Relationship between experimental and fitted transition temperature obtained by LMSR

图 2 使用支持向量机获得的预测值与实测值关系

Fig. 2 Relationship between experimental and predicted transition temperature obtained by SVM

络的性能. 为了检验所建立的神经网络的可靠性, 对其进行 3 因素 9 水平交叉验证, 结果如表 3 所示.

从表 3 可知, 3 个特征值变化幅度较大, 神经网络对运行参数比较敏感. 在 9 组验证中, 采用默认参数获得的特征值最好. 即隐含层节点数(n)为 6, 学习速率(v)为 0.3, 动态参数(σ)为 0.2 时, 准确率最高, 其 δ_{MAE} , δ_{MSE} 和 δ_{MPAE} 值均最小, 分别为 0.39, 0.89 和 4.86%.

用 BP 神经网络建立的相变温度模型. 通过该模型对实际测得的数据进行预测, 结果如图 3 所示. 从图 3 可知, 模型预测的结果与实际测量值的相关系数达 0.99.

表 3 神经网络运行参数的选择

Tab. 3 Selection of running parameters of BP neural network

水平	因素			δ_{MAE}	δ_{MSE}	$\delta_{MPAE}/\%$
	n	v	σ			
1	8	0.25	0.35	1.08	2.02	13.21
2	10	0.15	0.55	0.61	1.04	7.68
3	13	0.40	0.50	1.32	2.12	16.43
4	17	0.20	0.65	2.14	3.19	27.10
5	5	0.30	0.70	11.66	21.38	147.17
6	11	0.09	0.80	0.71	1.57	9.00
7	15	0.08	0.40	0.78	1.38	9.73
8	3	0.10	0.45	0.45	1.05	5.74
9	6	0.07	0.60	0.53	0.89	6.62
默认值	6	0.30	0.20	0.39	0.89	4.86

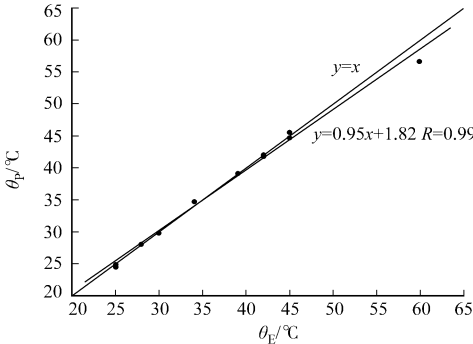


图 3 使用 BP 神经网络获得的

预测值与实测值关系

Fig. 3 Relationship between experimental and predicted transition temperature obtained by BP neural network

3 讨论

由图 1~3 可知, BP 神经网络所建立的预测相变温度的精度, 比使用支持向量机和最小中位方差回归建立的相关性要好, 可作为后续使用的模型.

当实测的 ELP 相变温度为 60 ℃(此时 ELP 的序列最短浓度最高), 与 3 种算法所预测(回归的结果是拟合的)出来相变温度值均差距较大. 这可能是因为当序列较短时, ELP 浓度与长度的变化对相变温度影响更大^[4], 而 ELP 的序列组成对相变温度影响较小.

与传统的拟合方法预测 ELP 的相变温度相比, 基于支持向量机和神经网络对相变温度进行预测, 不用通过预测相变温度具体形式, 就可以直接从数据中得到相变温度与 ELP 序列、分子量、Xaa 组成、浓度之间的关系. 同时, 只要能加以一定的先验知识, 还能够更大范围地反映它们之间的关系, 其应用的范围也将更为广阔.

文中基于 Chou 等提出的伪氨基酸概念^[6], 考虑到 ELP 的氨基酸组成极为相似, 构造了一种 λ 维的伪氨基酸组成来表示蛋白质序列. 采用 BP 神经网络、支持向量机方法、最小中位方差回归预测 ELP 的相变温度值. 结果表明, 当 $\lambda=8$ 为氨基酸相关系数的阶数最佳运行参数时, 使用 BP 神经网络所建立的

相变温度预测模型为最佳.

参考文献:

[1] URRY D W. Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers[J]. Phys Chem (B), 1997, 101(51): 11007-11028.

[2] CHOW D, NUNALEE M L, CHILKOTI A, et al. Peptide-based biopolymers in biomedicine and biotechnology[J]. Mater Sci Eng R Rep, 2008, 62(4): 125-155.

[3] URRY D W, LUAN C H, PARKER T M, et al. Temperature of polypeptide inverse temperature transition depends on mean residue hydrophobicity[J]. J Am Chem Soc, 1991, 113(11): 4346-4348.

[4] MEYER D E, CHILKOTI A. Quantification of the effects of chain length and concentration on the thermal behavior of elastin-like polypeptides[J]. Biomacromolecules, 2004, 5(3): 846-851.

[5] OLSON S D. Mathematical models for analysis of tissue regeneration in articular cartilage[D]. North Carolina State: North Carolina State University, 2009.

[6] CHOU Kuo-chen. Prediction of protein cellular attributes using pseudo amino acid composition[J]. Proteins; Structure, Function, and Bioinformatics, 2001, 43(3): 246-255.

[7] SHEN Hong-bin, CHOU Kuo-chen. PseAAC: A flexible web-server for generating various kinds of protein pseudo amino acid composition[J]. Analytical Biochemistry, 2008, 373(2): 386-388.

[8] VANPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

[9] 黄永恒, 曹平, 汪亦显. 基于 BP 神经网络的岩土工程预测模型研究[J]. 科技导报, 2009, 27(6): 61-64.

[10] 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(3): 363-372.

[11] 张光亚, 葛慧华, 方柏山. 一种预测木聚糖酶最适温度的 PCANN 模型[J]. 华侨大学学报: 自然科学版, 2007, 28(1): 55-58.

[12] ROUSSEUW P J. Least median of squares regression[J]. Journal of the American Statistical Association, 1984, 79(388): 871-880.

[13] STEELE J M, STEIGER W L. Algorithms and complexity for least median of squares regression[J]. Discrete Applied Mathematics, 1986, 14(1): 93-100.

Using Pseudo-Amino Acid Composition and BP
Neural Network to Predict the Transition
Temperature of Elastin-Like Peptides

HUANG Kai-zong, ZHANG Guang-ya

(College of Chemical Engineering, Huaqiao University, Quanzhou 362021, China)

Abstract: Elastin-like peptides (ELP) is one of the multi-peptides which has been widely used. Transition temperature is the most convenient parameters for quantificational description of the ELP properties. It is of great importance to explore the relationship between the transition temperature and the sequence characteristics, the number of Xaa of each monomer and the concentration of ELP. In this article, the best order of the correlation coefficient for pseudo-amino acid composition was obtained by using Least Median of Squares Regression from sequence. The uniform design was used to optimize the running parameters and leave-one out cross-validation was carried out to evaluate the model of back propagation neural network (BPNN) and support vector machines, respectively. The results showed that the predicted model obtained by BPNN was the best, of which the mean absolute error and root mean squared error was 0.39 °C and 0.89 °C, respectively.

Keywords: elastin-like peptides; transition temperature; pseudo-amino acid composition; support vector machines; back propagation neural network