蛋白质中三联氨基酸数与二级结构数的模型研究

朱尔一
(厦门大学 化学化工学院,现代分析科学重点实验室,福建 厦门 361005)

摘要:蛋白质的一级结构或序列与二级结构的关系在蛋白质结构研究中是很重要的,通过建立模型的方法来研究这种关系. 在文献中已有的模型(蛋白质一级结构中的二联氨基酸与蛋白质二级结构的模型)的基础上,建立了蛋白质一级结构中的三联氨基酸个数与蛋白质二级结构个数模型. 该模型能够较准确地反映蛋白质的一级结构或序列与蛋白质的二级结构的关系,比较适合应用于氨基酸序列长度变化较大的建模数据,同二联氨基酸与二级结构模型比较,由于三联氨基酸含有更多氨基酸之间的耦合信息,该模型的拟合精度更高. 由于蛋白质一级结构中的三联氨基酸的种类数很大(为 4 200),用以建模的变量数就很大,同时从 DSSP 数据库得到的样本量也很大(为 11 600),用以建模的数据量很大. 研究结果表明,PLS 变量筛选法是一种建立大数据模型有效的方法,可有效地处理变量数为 4 200,样本数为 11 600 这样大数据量的建模问题.

关键词:蛋白质二级结构预测; 偏最小二乘法变量筛选; 海量数据建模; 三联氨基酸
中图分类号:O 604 **文献标识码**:A **文章编号**:0438-0479(2009)05-0704-05

蛋白质的一级结构或序列与二级结构的关系在蛋白质结构研究中是很重要的. 利用蛋白质的一级结构或序列预测蛋白质的二级结构的研究方法[1]有 Chou-Fasman、GOR、基于疏水性方法和人工神经网络法等很多种,其中有一类方法是研究蛋白质中 20 种氨基酸数量与二级结构数量的关系的,如 Chou-Fasman 法[1]是研究 20 种单个氨基酸形成 α 螺旋,β 折叠及无规则卷曲 3 种二级结构的倾向性的,既建立蛋白质序列中 20 种单个氨基酸出现的频率与二级结构含量的关系模型的方法. 以后 Chou[2]考虑蛋白质序列中 20 种氨基酸之间的两两耦合作用,建立了二联氨基酸在蛋白质序列中出现的频率与二级结构含量的关系模型,该模型同单个氨基酸与二级结构的关系模型比较,模型的拟合精度有很大提高.

本文在这类方法的基础上,考虑建立三联氨基酸在蛋白质序列中的数量与二级结构的数量关系模型,简称三联氨基酸模型. 与二联氨基酸模型比较,二联氨基酸模型中自变量数为 210(二联氨基酸种类数为 20 ×20 = 400,如果认为每一种二联氨基酸中前后排列顺序不同的为同一种,例如 A-B 与 B-A 相同,则共有 210 种),而三联氨基酸模型中自变量数为 4 200(不考虑排列顺序,A-B-C 与 C-B-A 相同),三联氨基酸模型自变量中包含了更多的氨基酸之间的耦合作用信息,因此三联氨基酸模型或许能更精确地反映蛋白质的序列与二级结构的关系. 另外,三联氨基酸模型中自变量数很大,建模的过程中容易出现过拟合问题,采用一般多元回归的方法,根本无法建模,本研究采用偏最小二乘(PLS)变量筛选法[3-4]建模,使得最终模型中的自变量数尽可能小,模型的预报稳定性尽可能的高.

1 蛋白质二级结构预测模型

本文 PLS 方法中的模型为一般多元线性模型

$$Y = XB + E \tag{1}$$

其中 X 中包含蛋白质序列中各种三联氨基酸数据,三联氨基酸的种类数为 20 ×20 ×20 = 8 000,如果考虑每一种氨基酸与临近的氨基酸相互作用与前后次序无关,如认为 A-B-C 与 C-B-A 相同,则三联氨基酸种类数为 4 200,即维数. Y 中包含蛋白质中各种二级结构数据,即 α 螺旋,β 折叠等,Y 变量数为二级结构的种类数. 模型(1)中 B 为待定系数矩阵,E 为误差阵.

以上模型(1)中数据 X 变量和 Y 变量可以采用两种形式,1) 对于每一个蛋白质序列,都可得到其各种三联氨基酸的出现频率或成分,即

三联氨基酸的出现频率 = 三联氨基酸的出现个数/(序列长度 - 2),

将这个量作为 X 变量,而该蛋白质序列中各种二级结

收稿日期:2009-05-05
基金项目:福建省自然科学基金(X0750052),近海海洋环境科学国家重点实验室(厦门大学)开放项目资助
Email:ryzhu@xmu.edu.cn

构含量,即二级结构的出现个数与序列长度的商作为 Y 变量,即模型(1)为三联氨基酸的出现频率或成分与二级结构含量的模型. 2) 直接用蛋白质序列中各种三联氨基酸的出现个数作为 X 变量,而各种二级结构的出现个数作为 Y 变量,即模型(1)为三联氨基酸的个数与二级结构个数的模型. 文献[2,5]采用的是第 1 种形式,本文中采用第 2 种形式,因为要处理的蛋白质序列数据中序列长度数变化很大(变化范围 50～4 000),采用第 2 种形式的模型更能适应这种变化,实际数据处理说明,对同一套数据,采用第 2 种形式建立的模型精确度比采用第 1 种形式要高.

2　建模方法

模型(1)中自变量数很大,采用普通的多元线形回归的方法根本无法建模,而采用 PLS 方法可以,因为 PLS 回归方法是隐变量回归法,回归过程中虽然 X 变量数很大,但隐变量数很小. 本文采用 PLS 变量筛选法建立模型,PLS 变量筛选法是在 PLS 回归法基础上作变量筛选的,其特点为:根据 PLS 法建模过程中的一些信息,对原始自变量进行筛选,在不损失模型的预报能力的条件下,除去一部分冗余变量或影响不显著的变量,得到更简单的回归模型.

2.1　PLS 回归法

PLS 法[6]是一种研究 2 个数据块或矩阵 X 和 Y 相关关系的方法. 在该法中对数据矩阵 X 实施序列的正交变换

$$t_i = X r_i \qquad (i = 1,2,\cdots,h) \tag{2}$$

其中 h 为隐变量的个数,在变换过程中,使得到的矢量 t_i 与对数据矩阵 Y 变换得到的矢量 $u_i = Yq_i$ 的协方差为最大值. 具体 PLS 正交变换算法见文献[6].

式(2)可写为矩阵的形式

$$T = XR \tag{3}$$

式(2)中 h 也是式(3)矩阵 T 和 R 的列数,h 一般由预报残差平方和(PRESS)值为最低确定. PRESS 定义如下:

$$\mathrm{PRESS} = \sum (y_i - \hat{y}_{i,-i})^2 \tag{4}$$

其中 $\hat{y}_{i,-i}$ 为第 i 个样本不参加建模时得到的模型对该样本的预报值,PRESS 可以用来交叉检验(Cross validation)所建立数学模型的有效性,PRESS 值越小,表示模型的预报能力越强.

对于 Y 中某一变量 y,PLS 法回归实际模型为

$$y = Tv + e \tag{5}$$

将式(3)代入式(5)可得

$$y = XRv + e = Xb + e \tag{6}$$

因此 PLS 回归法的模型系数由下面公式计算求得:

$$b = Rv \tag{7}$$

其中隐变量的个数或矩阵 T 中变量的个数 h 远小于矩阵 X 中变量的个数 n. PLS 回归法是一种隐变量回归法,在本文处理的问题中隐变量数一般小于 30,因此能有效地避免过拟合.

2.2　PLS 变量筛选方法

在 PLS 变量筛选法[3-4]中,首先用 PLS 法对含有全部变量的数据处理,建立一个预报稳定性较高的模型,在此基础上利用其中回归系数等有关信息来进行变量筛选,删除影响不大的变量. 主要采用以下判据来删除影响不大的变量:

$$\Delta E_i = b_i^2 / \; l_i^T R \; (T^T T)^{-1} \; R^T l_i \tag{8}$$

以上 ΔE_i 表示当删除第 i 个变量时,PLS 回归模型的拟合误差增加值. 上式中 T 为 PLS 法得到的正交矩阵,矩阵 $(T^T T)^{-1}$ 为对角矩阵,较容易计算,而 R 是 PLS 正交分解得到的矩阵,矢量 l_i 为第 i 个分量为 1,其余分量为 0 的一种特殊矢量,b_i 为第 i 个变量对应的回归系数. 在 PLS 变量筛选法中主要是删除那些 ΔE_i 值很小对应的变量.

2.3　PLS 变量筛选法具体操作过程

用 PLS 法对全部变量数据进行处理,建立一个预报稳定性较高的模型,在此基础上进行变量筛选,从所有自变量中选出最优的变量子集,再用 PLS 法对挑选出的变量建立线性模型,得到简单实用的预报分类模型.

为便于模型之间的比较,本文采用模型交叉检验相关系数 CR 来衡量模型的预报准确率,CR 定义如下:

$$\mathrm{CR} = \sqrt{1 - \mathrm{PRESS}/S_y} \tag{9}$$

其中 $S_y = \sum (y_i - \bar{y}_i)^2$ 为变量 y 的总方差. CR 越接近于 1,说明模型的预报能力越强,越可靠.

具体变量筛选过程为:1) 首先用 PLS 法对含有全部变量数据进行处理,建立回归模型,即求得每个变量的回归系数,及 ΔE_i 值(对于每一个变量,可根据式(8)计算得到一个该变量对应的 ΔE_i 值) 及模型的 PRESS 值;2) 每次删除那些 ΔE_i 值小于某一个特定数值 P 的所有变量(P 值根据实际问题中模型 PRESS 值变化情况而设定);3) 再用 PLS 法建立剩下变量的数据的回归模型,即求得剩下每个变量的回归系数和 ΔE_i 值,及模型的 PRESS 值,返回 2);4) 直到剩下变量的 ΔE_i 值都大于某一个设定值,或模型的 PRESS值升高超过给定的值,停止,最终选出CR值

ID: 1R3N Length: 438

Amino acid sequence:

GTLNLPAAAPLSIASGRLNQTILETGSQFGGVARWGQESHEFGMRRLAGTALDGAMRDWFTNECESLGCKVKVDKI
GNMFAVYPGKNGGKPTATGSHLDTQPEAGKYDGILGVLAGLEVLRTFKDNNYVPNYDVCVVVWFNEEGARFARSC
TGSSVWSHDLSLEEAYGLMSVGEDKPESVYDSLKNIGYIGDTPASYKENEIDAHFELHIEQGPILEDENKAIGIVTGV
QAYNWQKVTVHGVGAHAGTTPWRLRKDALLMSSKMIVAASEIAQRHNGLFTCGIIDAKPYSVNIIPGEVSFTLDFR
HPSDDVLATMLKEAAAEFDRLIKINDGGALSYESETLQVSPAVNFHEVCIECVSRSAFAQFKKDQVRQIWSGAGHDS
CQTAPHVPTSMIFIPSKDGLSHNYYEYSSPEEIENGFKVLLQAIINYDNYRVIRGH

Secondary structure sequence:

CSCCCCCCCCCCCCTTHHHHHHHHHHHHHHTTEECCCSSSTTCCEECCCTTSHHHHHHHHHHHHHHHHHHTCEEEEBTT
SCEEEEECCSSCSSCEEEEECCCCCSSBCSSTTHHHHHHHHHHHHHHHHHTTCCCSSCEEEEECSCSSCSSBSSTTHHH
HHHTTSSCHHHHHTCBBSSCSSCCBHHHHHHHHTTCCCSBCCSTTTSCCSEEEEEEECSSSHHHHTTCSEEEEEEECEE
EEEEEEEECCCEETTTSCGGGCCCHHHHHHHHHHHHHHHHHHHTTCEEECCCEEEESCCTTEECSEEEEEEEEESCH
HHHHHHHHHHHHHHHHHTTCCTTCCCEEEEEEEEECCEECCHHHHHHHHHHHHHTTSCGGGEEEEESSCCTHHHH
TTTSCEEEEECCGGGCCSSTTCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHC

图 1 DSSP 数据库中的二级结构信息

Fig. 1 Information of secondary structure in DSSP

较高而变量数 n 又较小的预报模型.

3 数 据

本文所用数据取自于互联网(http://www.cmbi.kun.nl/gv/pdbfinder/overview.html 或 http://swift.cmbi.ru.nl/gv/dssp/)上的蛋白质二级结构数据库 DSSP(Database of secondary structure of protein)[7],DSSP 是一个二级数据库,除了二级结构以外,DSSP 中还包括蛋白质的几何特征及溶剂可及表面等. DSSP 二级结构区分得比较细致,共分 7 种二级结构,其编码含义如下:H 代表 α螺旋,E 代表 β折叠,G 和 I 分别代表 3-螺旋和 π螺旋,B 代表孤立的 β桥,T 代表氢键转折,S 代表弯曲. 该数据库中含有已知二级结构的蛋白质序列 35 000 多条. 从 DSSP 中提取的典型的二级结构信息就如图 1 所示.

对于以上类型信息,用长度为 3 的窗口对整个氨基酸序列扫描,对每一个氨基酸序列来说,都可得到 4 200种三联氨基酸的出现个数,可作为 X 变量(大部分 X 变量的值为 0),而对二级结构序列直接统计,得到各种二级结构的个数作为 Y 变量.

4 结果与讨论

从 DSSP 数据库中获取 10 000 多个已知二级结构的蛋白质序列,其中除去相同的序列,并删除序列中含较多 X 符号的序列(蛋白质序列中 X 表示不确定氨基酸),共得到 11 600 蛋白质序列用以建模. 对 α螺旋,β折叠等一些不同二级结构分别建模,采用 PLS 变量筛选法所得模型的结果见表 1.

表 1 各种二级结构预测模型的 CR 值

Tab. 1 CR values of the prediction model for various kinds of secondary structure

二级结构	变量数	模型的 CR 值
αhelix	3778	0.9872
βsheet	3880	0.9760
H-bond	2715	0.9819
Bend	2769	0.9735
Radom coil	3553	0.9880

表中第 2 列为 PLS 变量筛选法处理后,对不同二级结构模型中的原始 X 变量数目,第 3 列为模型的 CR 值,表 1 中的 5 种二级结构模型的 CR 值都大于 0.97,说明得到的这些模型质量很高.

建模所用数据量大,代表性强,模型适应范围大. 另外我们用 CB513 蛋白质数据样本集[8]作为检验样本,代入以上所得模型,对其进行检验,结果见图 2,3.

图 2,3 的结果为 α螺旋和 β折叠 2 种二级结构的实际个数与根据其蛋白质序列预报的二级结构个数的关系图,对于 α螺旋和 β折叠 2 种结构,二级结构的实际个数与预报的个数的相关系数分别为 0.939 和 0.903.

为了对模型进行分析,对于表 1 中的 5 种二级结构模型,列出部分模型的回归系数,由于模型的回归系

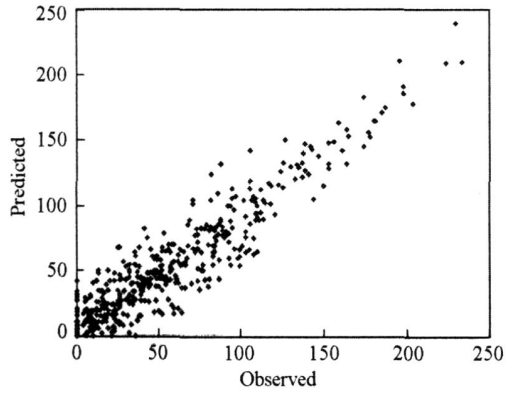

图 2　来自 CB513 数据库中的样本 α 螺旋的预测数和实际数的关系

Fig. 2　Relation of the observed and predicted number of α-helix in the protein sequence from CB513 dataset

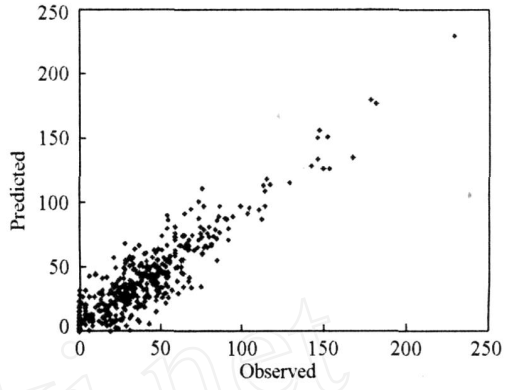

图 3　来自 CB513 数据库中的样本 β 折叠的预测数和实际数的关系

Fig. 3　Relation of the observed and predicted number of β-sheet in the protein sequence from CB513 dataset

表 2　5 种二级结构模型的部分回归系数

Tab. 2　Part of Regression coefficients of the models for 5 kinds of secondary structure

α helix		β sheet		H-bond		Bend		Radom coil	
EAM	6.7884	TWY	6.2610	FPH	2.7543	MEY	3.0435	IQY	3.6023
IMK	6.5402	NNT	5.3291	HAN	2.5015	GDH	2.5220	CGS	3.2414
QKY	6.0207	EWT	5.1155	GCP	2.2023	IMI	2.5057	CIS	3.1410
CLK	5.9501	GRW	4.8745	GMP	2.1990	YLY	2.2692	KPR	3.0521
LNM	5.8858	EVW	4.8598	AHC	2.1122	WGY	2.2285	NNS	2.8226
…	…	…	…	…	…	…	…	…	…
IEV	- 3.8487	AHS	- 3.3889	CMM	- 0.9778	DVH	- 1.2332	CHF	- 1.1509
RFT	- 4.8320	LNR	- 3.4385	CEM	- 0.9959	MAM	- 1.2625	CWH	- 1.1761
FQV	- 5.5873	INT	- 3.8430	MVS	- 1.0474	AVM	- 1.3900	MLW	- 1.2293
PES	- 6.4545	IMK	- 4.8485	LFW	- 1.2231	ECG	- 1.5420	AVM	- 1.7107
NNT	- 7.4344	EAM	- 4.8861	MDQ	- 1.3726	QRW	- 1.9224	EAF	- 2.1557

数的数量很大(2 700～3 900,见表 1),这里只列出绝对值较大的前 10 个回归系数,见表 2.

表 2 中的三联字母代表三联氨基酸单字母三联码,旁边的数值为该三联氨基酸对应的回归系数,从模型的回归系数分析可知,模型系数为正值代表对应的三联氨基酸有形成该结构的趋势,而模型系数为负值代表对应的三联氨基酸有解析该结构的趋势,系数的值越大则趋势越强. 从表 2 中的数据可知,三联氨基酸 EAM,IMK,QKY,CLK,LNM 等有形成 α 螺旋结构的趋势,或很容易形成 α 螺旋结构,三联氨基酸 NNT,PES,FQV,RFT 和 IEV 有解析 α 螺旋结构的趋势,或不容易形成 α 螺旋结构. 而三联氨基酸 TWY,NNT,EWT,GRW,EVW 等有形成 β 折叠结构的趋势,三联氨基酸 EAM,IMK,INT,LNR 和 AHS 有解析 β 折叠结构的趋势. 对其他 3 种二级结构的回归系数分析,也能得到类似的结果.

5　结　论

三联氨基酸的个数与二级结构个数的模型(见式(1),其中三联氨基酸的个数为 X 变量,二级结构个数为 Y 变量),能够较准确地反映蛋白质的一级结构或序列与蛋白质的二级结构的关系. 与三联氨基酸的频率或成分与二级结构含量的模型相比较,该模型更能适应序列长度变化较大的蛋白质序列数据,同二联氨基酸与二级结构模型相比较,由于三联氨基酸含有更多氨基酸之间的耦合信息,该模型的拟合精度更高. 通过建模分析,对所得三联氨基酸的个数与二级结构个数模型的回归系数分析,可得到各种三联氨基酸对应于各种二级结构形成趋势和解析趋势的信息,有助于进一步获取关于蛋白质一级结构或序列与蛋白质的二级结构的关系方面的新知识.

研究结果还表明,PLS 变量筛选法是一种建立海量数据模型有效的方法,可有效地处理变量数为 4 200,样本数为 11 600 这样大数据量的建模问题.

参考文献:

[1] 孙啸,陆祖宏,谢建明. 生物信息学基础[M]. 北京:清华大学出版社,2005:249 - 261.

[2] Chou Kuochen . Using pair - coupled amino acid composition to predict protein secondary structure content [J]. Journal of Protein Chemistry,1999,18(4):473 - 480.

[3] 朱尔一,林燕. 利用偏最小二乘法的一种变量筛选法[J]. 计算机与应用化学,2007,24(6):471 - 475.

[4] 朱尔一,林燕. 偏最小二乘变量筛选法在毒品来源分析中的应用[J]. 分析化学,2007,35(7):973 - 977.

[5] Chen Chao,Tian Yuanxin,Zou Xiaoyong,et al. Prediction of protein secondary structure content using support vector machine[J]. Talanta,2007,71(5):2069 - 2073.

[6] 朱尔一,扬芃原. 化学计量学技术及应用[M]. 北京:科学出版社,2001:100 - 107.

[7] Wolfgan Kabsch,Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical feature[J]. Biopolymer,1983,22(12):2577 - 2637.

[8] Cuff J A,Barton G I. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction[J]. Protein,1999,34(4): 509 - 519.

The Model Study Between the Number of Tria-coupled Amino Acid and the Number of Protein Secondary Structure

ZHU Er-yi

(College of Chemistry and Chemical Engineering, Key Laboratory of Analytical Sciences, Xiamen University, Xiamen 361005, China)

Abstract: The relation between protein sequence and protein secondary structure is very important, which has been studied by the method of building the model. Based on the models (between pair-coupled amino acid and protein secondary structure) in literature, the models between the number of tria-coupled amino acid in protein sequence and the number of protein secondary structure have been built. The models are more accurately reflect the relation between protein sequence and protein secondary structure. The models are more suitable to deal with the data in which the length of protein sequence varies a lot. Comparing with the models between pair-coupled amino acid and protein secondary structure, the models contain more information about coupling effect among varies kinds of amino acids, and therefore are of the higher fitting accuracy. The data set in the research is very large, because the kinds of tria-coupled amino acid in protein sequence are very big (4 200) and the number of samples from DSSP database is also very large (11 600). The results indicate that the PLS variable selection method is effective to deal with the huge data modeling problem in which the number of variables is 4 200 and the number of samples is 11 600.

Key words: protein secondary structure prediction; PLS variable selection; huge data modeling; tria-coupled amino acid