嗜热和常温蛋白模式识别的研究

A Study on the Pattern Recognition of Thermophilic and Mesophilic Proteins

张光亚,方柏山*

ZHANG Guang-Ya and FANG Bai-Shan*

华侨大学工业生物技术研究所,泉州 362021

Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, *China*

摘 要 采用主成分分析、偏最小二乘回归和 BP 神经网络三种方法对嗜热和常温蛋白进行模式识别。结果表明,三种方法对训练集拟合的平均正确率分别为 92%、95%和 98%,对测试集进行预测的平均正确率分别为 60%、72.5%和 72.5%,对嗜热蛋白预测正确率最高为 75%,常温蛋白最高为 85%。构建了数学模型并对其生物学意义进行了解释,建立了一种基于序列的识别嗜热和常温蛋白的新方法。

关键词 模式识别,主成分分析,偏最小二乘回归,BP 神经网络,热稳定性

中图分类号 Q811.4 **文献标识码** A **文章编号** 1000-3061(2005)06-0960-05

Abstract Pattern recognition of thermophilic and mesophilic proteins were studied through principle component analysis, partial least-square regression and BP neural network. The results showed that the fitting accuracy of the three methods was 92%, 95% and 98%, respectively. And the forecasting accuracy was 60%, 72.5% and 72.5%, respectively. The best forecasting accuracy for thermophilic proteins was 75%, and for mesophilic proteins was 85%. A mathematical model was established and the biological meaning of it was expatiated on, a new method to discriminate the thermophilic and mesophilic proteins based on their sequences was established here.

Key words pattern recognition, principle component analysis, partial least-square regression, BP neural network, thermostability

蛋白质的热稳定性一直是生物物理和生物技术领域研究的热点[1],这主要是由于蛋白在高温下易失活,对其在某些极端条件下进行工业生产中的应用造成了困难,成为拓展其应用领域的瓶颈。因此,如何提高酶蛋白的热稳定性一直是分子生物学、生物工程学和化学工业等所关注的重要研究课题之一[2]。尤其是这种热稳定特性能否在氨基酸水平上进行检测,尽管有研究发现嗜热蛋白中某些氨基酸含量与常温蛋白存在差异,但目前尚存异议[3]。嗜热酶作为生物催化剂却有许多优势,因而它成为相关研究的一个重点。

模式识别是信息科学领域中的一个特别分支,它诞生于 20 世纪 20 年代,在 60 年代初迅速发展成一门独立的学科。成为当代高科技研究和应用的重

Received: April 15, 2005; Accepted: July 11, 2005.

This work was supported by a grant from the National Sciences Foundation of China (No. 20276026) and the Science Foundation of Overseas Chinese Affairs Office of the State Council of China (No. 05QZR06).

* Corresponding author. Tel: 86-595-22691560; E-mail: fangbs@hqu.edu.cn

国家自然科学基金资助项目(No. 20276026)和国务院侨办科研基金资助项目(No. 05QZR06)。

要领域之一。模式识别是通过对已知类别样本(训练样本)的分类学习,使识别系统增长分类知识,形成分类能力,从而具备对未知类别样本的类别预报(模式识别)能力和控制对象类别优化(模式优化)能力[4]。模式识别基本上是由三个相互关联而又有明显区别的过程组成的,即数据生成、模式分析和模式分类。建模是模式识别的重要手段之一,有两类统计建模方法在模式识别领域得到广泛应用:一类是各种多元回归方法,另一类是人工神经网络方法。模式识别已在许多方面得到了成功的应用,但在生物工程领域的应用还不多[5]。而将其应用到嗜热和常温蛋白的识别尚未见报道。

本文利用主成分分析(PCA)、偏最小二乘回归(PLSR)和主成分分析的 BP 神经网络(PCANN)三种常见模式识别的方法对嗜热和常温蛋白进行模式识别,取得了较好的拟合和预测效果,建立了一个基于序列的识别嗜热和常温蛋白新方法,具有一定的理论和实际意义。

1 材料和方法

1.1 数据来源

用于训练的 76 组(152 条)嗜热和常温蛋白的序列信息来源于 Swiss-Prot,Swiss-Prot 是一个非冗余的专家库。首先用‘thermo’和‘pyro’作为关键词进行查询,从返回的结果中剔除推测的(putative)、可能的(probable)、片段(fragment)以及一些高度相似的序列,最后共获得 76 条不同种类的嗜热蛋白序列,然后分别以这些蛋白的名称或酶的 EC 编号寻找其对应的常温蛋白,共计 76 条。用于测试的 20 组嗜热和常温蛋白的 PDB ID 来源于文献[6],根据 ID 号从 PDB 下载其序列。训练集和测试集蛋白序列的 ID 号见表 1。所有蛋白氨基酸组成分析由 Bioedit 软件完成。PCA 由 SPSS10.0 完成,PLSR 和 BP 神经网络(PCANN)由 DPS 软件完成[7],作图软件为 origin7.0。

表 1 训练和测试样本的 ID 号

Table 1 Accession numbers of the training and testing samples

Sample	Thermophilic protein	Mesophilic protein
Training	Q9V0L2, O93730, Q9UY47, Q58549, Q9V2I6, O58362, Q8ZVE4, Q9V1I5, Q9V1I6, Q8ZUA0, Q8ZV07, Q8ZU95, Q8ZW80, Q8ZW90, Q8ZW59, Q8U0A6, Q8U0A5, Q8ZU97, P81413, Q9X1B7, Q9HIY2, O58111, P95474, Q47950, Q9HHC4, P77916, Q8DL74, O59605, Q9V1P1, Q9UXW3, Q9HHB6, Q9V0T9, Q9WY82, Q8ZTZ0, O58097, Q8TZI8, Q8U039, Q9UYR1, Q8ZY36, Q8TH25, Q9V1R3, Q9YB30, Q9UZ09, Q8ZYU6, P19514, Q8ZZX3, Q8U0F3, Q8ZU24, Q8ZW35, Q8U0C0, O32450, Q8ZVB2, O59488, Q8U381, Q9WY74, Q8TZL3, O57765, Q8RBA4, Q8U4I9, Q8U3K8, O58429, Q8U263, Q51742, O58665, Q9V0N0, P58202, Q8U0G6, P61883, Q8U111, O58050, O57979, Q9UY56, Q8ZZK5, Q8U3Z2, Q8U491, Q8U4A0	P53001, P36333, Q9VF36, P54570, P71295, P53582, Q82XS4, P68729, Q88QQ6, P59308, P36839, Q8G5F3, P31102, P63609, P05194, P43904, P34003, O34347, O89033, O04928, Q8FC88, P36561, P37306, Q43314, P44121, P21189, P77488, P40370, O34425, P11537, P56091, Q97FQ7, Q9KRB5, P14742, Q9LVI8, Q8FBC3, Q9HTE9, Q91Z53, P60757, Q82WM3, P14891, Q38929, P05793, Q9I6E0, P00817, P29364, P46086, P32895, P09151, P30127, P15977, P22133, P17109, P59286, P52085, P17443, P10902, Q9I4W9, Q9HX21, Q9HUP3, Q99JR6, P39207, P68739, Q8FAE1, Q9I3C3, P38787, Q9NXJ5, Q8KEX0, P00558, P32662, P35558, P00496, Q05728, Q9UUB4, Q9Y0Y2, P35421
Testing	1zin, 1tmy, 1aj8:b, 1tfe, 1yna, 1gtm:b, 1hdg:o, 2prd, 1ldn:c, 1bdm:b, 3mds:a, 1xgs:b, 3pfk, 1php, 1ebd:a, 1ril, 1caa, 1thm, 1lnf:e, 1btm:a	1aky, 3chy, 1csh, 1efu:b, 1xnb, 1hrd:b, 1gad:o, 1ino, 1ldg, 4mdh:b, 1qmn:a, 1mat, 2pfk:a, 1qpg, 1lpf:a, 2rn2, 8rxn, 1st3, 1npc:e, 1ypi:a

1.2 主成分分析(PCA)

主成份分析(Principal Component Analysis,PCA)也称主分量分析,是 Hotelling 于 1933 年首先提出的。主成分分析就是利用降维的思想,把多指标转化为少数几个不相关的综合指标的一种多元统计分析方法。具体操作过程为:(1) 对原始数据进行标准化处理;(2) 建立相关矩阵;(3) 计算特征值及特征向量;(4) 建立主成分方程,计算主成分荷载及主成分得分。

1.3 偏最小二乘回归(PLSR)[8]

1983 年 Herman,Harald 和 Wold 发表第一篇关于 PLSR 和多变量标准化文章,首次正式提出偏最小二乘回归理论。偏最小二乘回归是多元线性回归典型相关分析和主成分分析的集成和发展。其思路是首先从自变量集合 X 中提取成分 t_h ($h = 1,2,\cdots\cdots$),各成分相互独立,然后建立这些成分与自变量 X 的回归方程。PLSR 可以有效地实现观测数据信息的浓缩处理,将解释变量空间和反应变量空间分解成几个少数的解释潜变量和反应潜变量,相应地建立它们之间的回归关系。并且,解释潜变量与反应潜变量之间的关系比较稳定、明确,不会随着潜变量选

取标准的不同而变化。它具有计算量小、收敛快、简单、稳健等优点。其建模的过程为:(1) 数据的标准化处理;(2) 第一成分 t_1 的提取;(3) 第二成分 t_2 的提取;(4) 第 h 成分 t_h 的提取;(5) 建立偏最小二乘回归模型。

1.4 BP 神经网络(BPNN)

神经网络具有对噪声数据的高承受能力,并行处理能力以及非线性映射能力等特性。这些特性推动了神经网络在模式识别中的应用。BP 神经网络是目前应用最广、研究最深入的一种神经网络。典型的 BP 神经网络由 3 层构成:输入层、隐含层和输出层。

本文采用三层网络结构,取 Sigmoid 函数为神经元非线性函数。同时,为了减少输入层的神经元个数,并消除样本集中样本之间可能存在的一定相关性,在进行神经网络训练前对样本集进行优化和选择,以便加快收敛速度。在此,采用主成分分析的神经网络(PCANN)方法,即,根据主成分分析结果,计算各主成分分值,取该分值作为神经网络的输入层数据。

2 结果与分析

2.1 基于 PCA 的嗜热和常温蛋白的模式识别

主成分分析是模式识别中较为广泛应用的一种线性映射。将所得数据进行主成分分析,由第一主成分 PC_1 和第二主成分 PC_2 构筑的映射平面如图 1(A)所示。可见嗜热蛋白和常温蛋白能较清晰地分布在两个区域,对测试集拟合的平均正确率为 92%。其中:

$$PC_1 = 0.273A + 0.253C + 0.047D - 0.295E - 0.112F + 0.038G + 0.257H - 0.342I - 0.425K + 0.23L + 0.055M - 0.111N + 0.194P + 0.364Q + 0.203R + 0.047S - 0.008T - 0.152V - 0.082W - 0.288Y \quad (1)$$

(式中,A、C、D……Y 表示组成蛋白质的 20 种氨基酸,下同。)

$$PC_2 = -0.267A + 0.282C + 0.179D - 0.235E + 0.109F - 0.209G + 0.158H + 0.172I + 0.056K - 0.047L + 0.041M + 0.399N - 0.066P + 0.156Q - 0.326R + 0.338S + 0.349T - 0.338V + 0.001W + 0.011Y \quad (2)$$

从图 1 中可知,当某个蛋白经主成分分析后的 PC_1 和 PC_2 之间满足关系 $PC_2 > -1.5PC_1 + 6$ 时,则该蛋白为常温蛋白;当二者满足 $PC_2 < -1.5PC_1 + 6$ 时,则为嗜热蛋白。即满足:

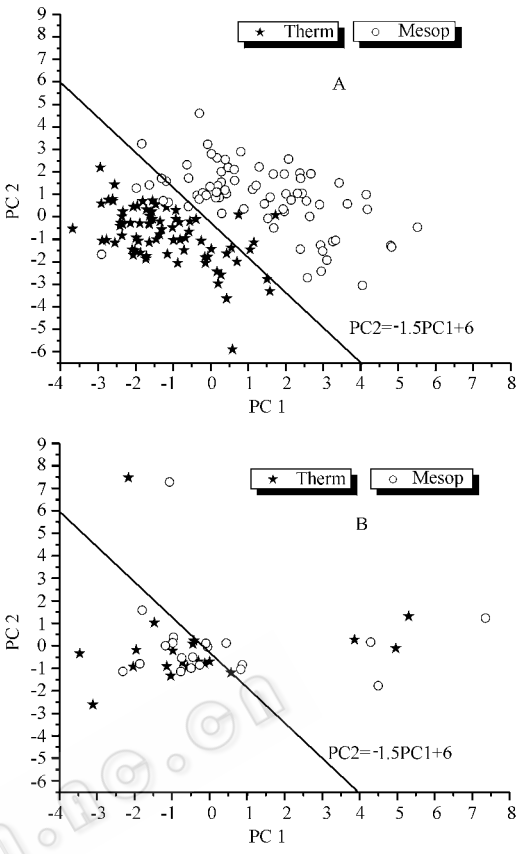

图 1 主成分 PC1 和 PC2 分类图

Fig. 1 The map of classification between principle component PC1 and PC2

A: the training sample; B: the testing sample.

$$0.678E + 0.059F + 0.152G + 0.341I + 0.582K + 0.021R + 0.566V + 0.421Y + 6 > 0.142A + 0.661C + 0.249D + 0.544H + 0.298L + 0.123M + 0.232N + 0.225P + 0.701Q + 0.408S + 0.337T + 0.124W \quad (3)$$

时,为嗜热蛋白。从式(3)可知,当某蛋白质中 Glu、Lys、Val 和 Tyr 的含量较高,而 Gln、Cys、His 和 Ser 的含量较低时,上述不等式成立,即该蛋白为嗜热蛋白,这与一些对嗜热和常温蛋白的研究结果基本吻合。如:Kumar[9] 认为,随着最适温度的升高,Ile、Tyr、Lys 和 Glu 含量增高,而 Gln 和 Cys 含量降低。Thompson 和 Eisenberg[10] 发现,嗜热蛋白中含有更多的 Glu、Val、Arg 和 Gly,更少的 Gln、Ser、Asp 和 Lys。可见,上述数学模型能在一定程度上解释嗜热蛋白热稳定性的机制。为了验证模型的正确性,利用上述方程对测试集的 40 组数据进行了计算,结果如图 1(B)所示,该模型对嗜热蛋白预测的正确率为 75%,而对常温蛋白预测的效果较差,正确率仅为 45%。预测的平均正确率为 60%。导致预测准确

率较低的原因可能与前 2 个主成分的荷载较低有关(分别为 27.7% 和 17.7%),但文献[11]报道的前三个主成分荷载分别为 22.9%、14.9%和 11.1%,其利用前两个主成分进行模式识别却取得了很好的效果。

2.2 基于 PLSR 的嗜热和常温蛋白的模式识别

令嗜热蛋白为 1,常温蛋白为 0,利用偏最小二乘回归提取训练集样本的成分。用前两个潜变量的得分(t_1 和 t_2)作二维映照图,如图 2 所示。可知用 PLSR 法已经能大体将嗜热和常温蛋白分开,但也存在一定的"死区"。对训练集拟合的平均正确率为 95%。偏最小二乘回归建立了如下方程:

$$y = 0.462 - 0.001A - \mathbf{0.116C} - 0.041D + 0.018E - 0.001F + 0.011G - 0.033H + 0.035I + 0.013K - 0.002L - 0.025M - 0.025N + 0.012P - \mathbf{0.069Q} + 0.028R - 0.055S - \mathbf{0.070T} + 0.053V - 0.012W + \mathbf{0.069Y} \quad (R = 0.854) \quad (4)$$

通过分析训练集拟合结果,可定义:当 $y > 0.48$ 时,该蛋白可被认为是嗜热蛋白,而当 $y < 0.48$ 时,该蛋白为常温蛋白。从上述方程可知,对 y 值影响最大的是 Cys 的含量(与 y 值呈负相关),说明 Cys 含量越高,y 值越小,是常温蛋白的可能性越大。Cys 被认为是热稳定性氨基酸,由于它在高温下会发生氧化,研究者[12]发现它在嗜热蛋白中含量低于常温蛋白,上述方程恰好解释了这一现象。另外,Gln 和 Thr 也和 y 值呈负相关,也意味着在嗜热蛋白中这两种氨基酸的含量也较低。Thr 能通过氢键与蛋白分子表面的水分子相互作用,但在高温下氢键容易断裂,这会导致蛋白在高温下不稳定,因此,嗜热蛋白中 Thr 的含量较低[13]。这也从上述方程中得到了反映。而和 y 值呈正相关且相关系数相对较大的氨基酸是 Tyr,这也与 Kumar[12] 等的研究结果相吻合,他们认为 Tyr 在嗜热蛋白中的含量更高。

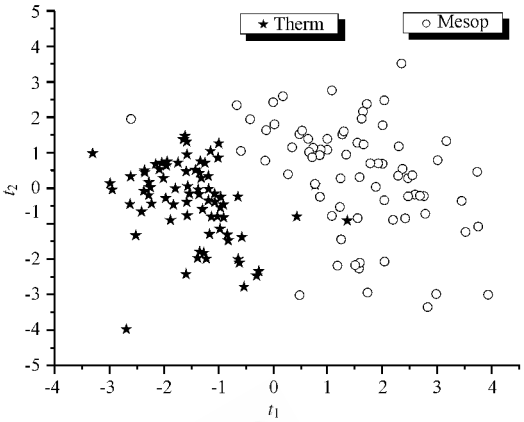

图 2 PLSR 主成分映照图

Fig. 2 The two dimensional map of t_1 and t_2 of PLSR

为了验证模型的正确性,利用上述方程对测试集的 40 组数据进行了计算,结果见表 2,表中加粗部分为预测错误的结果。可知该模型对嗜热蛋白预测的正确率为 65%,对常温蛋白预测的正确率为 80%,预测的平均正确率为 72.5%。

2.3 基于 PCANN 的嗜热和常温蛋白的模式识别

为了减少神经网络输入数据之间的相关性,同时减少输入层的神经元个数,采用主成分分析法对

表 2 PLSR 和 PCANN 的预测值和原分类值

Table 2 The predicted value by PLSR, PCANN and the primary classification value

PDB ID	Actual value	Predicted value		PDB ID	Actual value	Predicted value	
		PLSR	PCANN			PLSR	PCANN
1zin	1	**0.22**	**0.00**	1aky	0	−0.09	0.00
1tmy	1	0.73	1.00	3chy	0	0.47	**0.98**
1aj8 :b	1	1.04	1.00	1csh :b	0	0.00	0.00
1tfe	1	0.79	0.99	1efu :b	0	**0.62**	0.24
1yna	1	**0.02**	**0.03**	1xnb	0	−0.33	0.00
1gtm :b	1	0.80	1.00	1hrd :b	0	**0.65**	**1.00**
1hdg :o	1	0.53	**0.03**	1gad :o	0	0.17	0.00
2prd	1	1.04	1.00	1ino	0	**0.70**	**1.00**
1ldn :c	1	0.69	1.00	1ldg :c	0	0.36	0.00
1bdm :b	1	0.53	0.95	4mdh :b	0	0.12	0.00
3mds :a	1	0.50	1.00	1qmn :a	0	−0.34	0.00
1xgs :b	1	1.12	1.00	1mat	0	0.34	0.00
3pfk	1	0.49	**0.01**	2pfk :a	0	**0.50**	0.00
1php	1	0.81	1.00	1qpg	0	0.44	0.00
1ebd :a	1	0.64	0.72	1lpf :a	0	0.47	0.02
1ril	1	**0.03**	**0.01**	2rn2	0	−0.02	0.00
1caa	1	**−0.15**	**0.01**	8rxn	0	−0.14	0.00
1thm	1	**−0.25**	**0.01**	1st3	0	−0.21	0.00
1lnf :e	1	**0.08**	0.93	1npc :e	0	−0.17	0.00
1btm :a	1	**0.21**	**0.01**	1ypi :a	0	0.41	0.00

原始数据进行处理,选择前 13 个主成分,可使选择显著水平 α 达到 88%(通常 $\alpha > 85\%$即可),该 13 个主成分即可代表原始数据中蕴含的绝大部分信息。以各主成分的得分作为神经网络输入层,同样令嗜热蛋白为 1,常温蛋白为 0,作为神经网络的输出层,本文采用单隐含层的 BP 神经网络。隐含层的神经元个数设为 9,即神经网络的拓扑结构为"13-9-1"。神经网络运行的其它参数分别为:学习速率为 0.1、动态参数为 0.6、Sigmoid 参数为 0.9,允许误差设为 0.005,最大迭代次数设为 1000 次。以训练集的 152 组数据对神经网络进行训练后,将拟合值与原分类值进行比较后发现,取得了较好的拟合效果,正确率为 98%。

在训练好的基础上,对测试集的 40 组数据进行了预测,预测结果与实际结果如表 2 所示。该神经网络模型对嗜热蛋白预测正确率为 60%,而对常温蛋白预测正确率为 85%,平均正确率为 72.5%。

3 小结

主成分分析法(PCA)、偏最小二乘回归(PLSR)和主成分分析的 BP 神经网络(PCANN)三种方法对训练集拟合的效果以及对测试集预测的效果如表 3 所示。三种方法拟合的效果均优于预测的,而尤以 PCANN 拟合的效果最佳。三种方法对预测的效果则存在一定的差异,如对嗜热蛋白 PCA 预测的效果最好,正确率达 75%,对常温蛋白则 PCANN 预测的效果最好,正确率达 85%,而 PLSR 则在整体上表现较好。在实际应用中可综合考虑使用上述三种方法。

表 3 三种方法拟合和预测效果的比较

Table 3 Summary of the fitting and predicting results of the three methods

Method	Training sample/%			Testing sample/%		
	Therm	Mesop	Average	Therm	Mesop	Average
PCA	94.7	89.5	92.1	75.0	45.0	60.0
PLSR	96.1	94.7	95.4	65.0	80.0	72.5
PCANN	96.1	100.0	98.1	60.0	85.0	72.5

对嗜热和常温蛋白进行模式识别,提供了一个基于序列的识别嗜热和常温蛋白的方法。从理论上,它可探讨蛋白质热稳定性的分子机制,而且在一定程度上可以定量表示这种机制,建立数学模型 在实际应用上,它可在提高蛋白热稳定性所进行的分子改造过程中,作为一种实验开始前的预测方法,对从原始目标序列出发而随机产生的序列进行识别,若达到所需目的,则可通过实验对原始序列进行改造,从而使对分子的改造更具有目的性,同时亦可提高改造成功率,提高研究效率,节省研究经费。

REFERENCES(参考文献)

[1] Atomi H. Recent progress towards the application of hyperthermophiles and their enzymes. *Current Opinion in Chemical Biology*, 2005, **9**:1 – 8

[2] Wang YB(王耀兵), Nagata S (永田进一). Participation of ions and solutes on the thermostability of α-amylase. *Chinese Journal of Biotechnology*(生物工程学报),2004, **20**(1):104 – 110

[3] Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res*, 2001, **29** 1608 – 1615

[4] Dipti PM, Srimanta P. Advances in pattern recognition. *Pattern Recognition Letters*, 2005, **26**:395 – 398

[5] Chen M (陈敏), Liu WH(刘万卉), Wang JX(王静馨). Application of pattern recognition on optimum control of zinc yeast fermentation process. *Chinese Journal of Biotechnology*(生物工程学报), 1996,**12**(3):367 – 370

[6] Seung PP, Young JY. Protein thermostability: structure-based difference of animo acid between thermophilic and mesophilic proteins. *Journal of Biotechnology*, 2004, **111**:269 – 277

[7] Tang QY(唐启义), Feng MG (冯明光), Practical Statistics and DPS Data Processing System. Science Press, Peking, China, 2002

[8] Wang WS (王文圣), Ding J (丁晶), Zhao YL(赵玉龙) *et al*. Study on the long term prediction of annual electricity consumption using partial least square regressive model. *Chin Soc for Elec Eng* (中国电机工程学报), 2003, **23**(10):17 – 21

[9] Kumar S, Nussinov R. How do thermophilic proteins deal with heat? *Cell Mol Life Sci*, 2001, **58**:1216 – 1233

[10] Thompson MJ, Eisenberg D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J Mol Biol*, 1999, **290**:595 – 604

[11] Dorn ED, McDonald GD, Storrie-Lombardi MC *et al*. Principal component analysis and neural networks for detection of amino acid biosignatures. *Icarus*, 2003, **166** 403 – 409

[12] Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostalility. *Protein Eng*, 2000, **13**:179 – 191

[13] Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett*, 2000, **470**:65 – 69