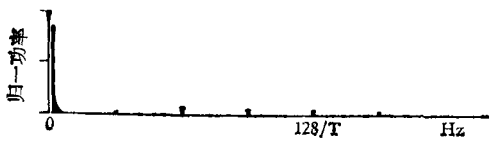

图 8　周期为 T 的方波的自功率谱
（线性归一化谱）

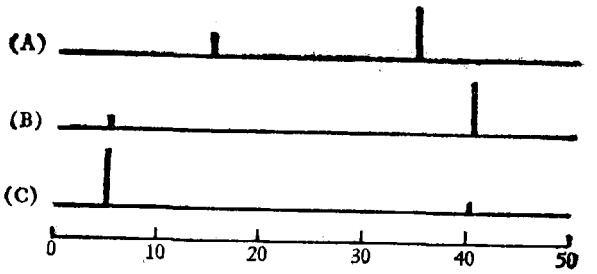

图 9　含有两种频率成分的信号的自功率谱

A	15 Hz	0.2V	C	5Hz	0.4V
	35 Hz	0.3V		40Hz	0.2V
B	5 Hz	0.2V			
	40 Hz	0.4V			

图示为线性归一化谱，每条曲线的最大峰值为 1

再进行分工，最内层的一个控制同一种类型的蝶形，另一个则保证各种类型的蝶形全被做到．第一个参加蝶形运算的数的地址是容易被确定的，其它的地址可以利用它们之间的间距的变化规律得到．

2. 相关函数程序

众所周知，功率谱密度与相关函数之间是一对傅里叶变换对．当我们求得自功率谱 $S_{xx}(n)$ 及互功率谱 $S_{xy}(n)$ 后，通过傅里叶逆变换处理就可以得到自相关函数和互相关函数．

在应用上述各程序之前，原始数据序列经过了汉明窗处理，有些情况还进行补零处理，这些在此不赘述．

三、信号处理结果示例

图 6—9 为一些实际信号经 APPLE II PLUS 系统处理的结果．各图的意义请参阅图注中的说明．

参 考 文 献

[1] 沈兰荪等：《电子技术应用》，7，28，1984.

［本文于1985年2月25日收到］

计算机模式识别技术在蛋白质二级结构预测中的应用

何东明　　陈农安
（中国科学院上海生物化学研究所）

目前，计算机技术在生化领域中的应用正在受到重视，研究工作已经取得了一些可喜的结果．我们在将计算机模式识别技术应用于蛋白质二级结构预测中做了一点工作，现简介如下．

我们从 Chou-Fasman 法中得到构成蛋白质的 20 个氨基酸中的每一个氨基酸出现于二级结构中的 α-螺旋、β-折叠和无规卷曲的概率参数 P_α、P_β 和 P_c；在预测方法上采用了与 Chou-Fasman 法完全不同的，适合于计算机处理的电位函数法：即现有 α-螺旋、β-折叠和无规卷曲这三类模式，对每类模式用电位函数分别构造一个判别函数，以此作为分类的依据来预测蛋白质的二级结构．本工作主要分三部分：(1) 问题的提出及训练模式的提取．(2)判别函数的构成．(3) 牛胰蛋白酶抑制剂和牛胰核糖核

酸酶的二级结构的"预测". 我们共有三个程序(每个程序中包含有若干个子程序),第一个程序用于提取训练模式,第二个程序用于构造判别函数,第三个程序用于预测蛋白质的二级结构.

一、问题的提出及训练模式的提取

蛋白质由 20 个氨基酸组成,蛋白质有其一级结构(氨基酸排列顺序)和高级结构(包括二、三、四级结构);一级结构是形成其高级结构的最重要的因素,由蛋白质的氨基酸排列顺序能够推测出蛋白质的高级结构[1]. 由于大量蛋白质高级结构的阐明逐步揭示出一些结构规律,使人们掌握了一些可供运用的结构规则. 所以,近年来在蛋白质结构研究中兴起了一个新的领域:从一级结构预测蛋白质的高级结构,Chou-Fasman 法就是其中的一种比较好的方法[2,3].

Chou-Fasman 法是从 29 种已知结构的蛋白质中统计出 20 种氨基酸中的每一种采取 α-螺旋、β-折叠和无规卷曲结构的概率参数 P_α、P_β 和 P_c,并建立了一套预测蛋白质二级结构的规则. 虽然这一方法取得了良好的效果,但由于在预测中主要是由人根据多种因素经逻辑判断得出结论,显然工作量是较大的. 为了解决这一问题,我们抛弃了 Chou-Fasman 法中的预测规则,利用它的概率参数,采用了模式识别技术.

我们以四个相邻的氨基酸片段作为窗口,求出其 P_α、P_β 和 P_c 的平均值 $\langle P_\alpha\rangle$、$\langle P_\beta\rangle$ 和 $\langle P_c\rangle$,分别作为三维采样模式的第一、第二和第三分量. 我们按这一方法从 5 个蛋白质 (α-Chymotrypsin[4]、Cytochrome b_5[5]、Ribonuclease S[6]、Elastase[7] 和 Pancreatic Trypsin Inhibitor[8]) 中提取了 150 个训练模式,50 个来自 α 螺旋,50 个来自 β 折叠,50 个来自无规卷曲.

二、判别函数的构成[9]

电位函数法是模式识别技术中的一个算法,基本思想是:把每一个样本看成是在样本空间中的一个电荷,因而在几何位置上聚集的一类样本在样本空间形成一个能量场. 同一类样本聚集的地方出现一个能量高地,所以当有 M 类样本模式时就有 M 个能量高地,两两高地之间由电位峡谷隔开,这就是判别各类模式的标准.

要实现这种算法就应该有一个能描述电位分布的函数,我们选择的是指数衰减函数:$D(X, X_k) = \exp\{-\alpha\|X - X_k\|^2\}$,这里,$\alpha$ 是一系数,X_k 是某一特定样本点,$\|X - X_k\|$ 是欧氏距离. 具体算法如下:

设有 n 个样本点分别属于 M 类,初始电位 $D_0^{(1)}(X) = D_0^{(2)}(X) = \cdots\cdots = D_0^{(M)}(X) = 0$.

设在第 $k+1$ 步时,某样本点 X_{k+1} 属于第 i 类,如果 $D_k^{(j)}(X_{k+1}) > D_k^{(j)}(X_{k+1})$,$j = 1, 2, \cdots M$,$j \neq i$,则电位不需修改,即 $D_{k+1}^{(j)}(X) = D_k^{(j)}(X)$,$i = 1, 2, \cdots, M$.

如果有一 l,使 $D_k^{(j)}(X_{k+1}) \leqslant D_k^{(j)}(X_{k+1})$,则修改电位,即 $D_{k+1}^{(j)}(X) = D_k^{(j)}(X) + D(X, X_{k+1})$,$D_{k+1}^{(j)}(X) = D_k^{(j)}(X) - D(X, X_{k+1})$,$D_{k+1}^{(j)}(X) = D_k^{(j)}(X)$,$j = 1, 2, \cdots, M$,$j \neq i$,$j \neq l$. 这里下标 $k = 1, 2, \cdots, n-1$;这是一次循环,需经多次循环直至每一个属于第 i 类的样本点 X_{k+1},使 $D_k^{(j)}(X_{k+1}) > D_k^{(j)}(X_{k+1})$,$j = 1, 2, \cdots, M$,$j \neq i$,都成立时才停止 ($i = 1, 2, \cdots, M$). 由此得到每一类的判别函数 $d^{(j)}(X) = D_k^{(j)}(X)$.

三、牛胰蛋白酶抑制剂和牛胰核糖核酸酶二级结构的"预测"

我们利用所得的判别函数对牛胰蛋白酶抑制剂和牛胰核糖核酸酶的二级结构进行了"预测",所得结果与晶体结构分析的结果对比如表 1 所示.

由表 1 可见,预测结果是比较准确的,其准确度大致与 Chou-Fasman 法相仿.

我们认为,取足够多的不同样本点所训练出来的判别函数是能够比较准确地推测出一个球蛋白的二级结构的.

表 1

	α 螺旋		β 折叠		无规卷曲	
	预 测	晶体分析	预 测	晶体分析	预 测	晶体分析
牛胰蛋白酶抑制剂	1—8	3—6	17—24	16—24	9—16	7—15
	45—52	46—56	29—36	27—36	25—28	25—26
			53—56	无	37—44	37—45
牛胰核糖核酸酶	5—12	3—13	41—48	41—48	1—4	无
	25—36	24—35	无	60—65	13—24	14—23
	49—60	50—59	69—76	69—76	37—40	36—40
			81—88	79—87	61—68	66—68
			97—112	96—110	77—80	77—78
			117—124	116—124	89—96	88—95
					113—116	111—115

对戚正武教授的支持，李世武和宣建成两同志所给予的帮助，一并致谢.

参 考 文 献

[1] 王大成：《生物化学与生物物理进展》，1981 年，1 期，26 页.

[2] Chou, P. Y. and Fasman, G. D.: *Ann. Rev. Biochem.*, **47**, 251, 1978.

[3] Chou, P. Y. and Fasman, G. D.: *Adv. Enzymology*, **47**, 45, 1978.

[4] Blow, D. M.: *Biochem. J.*, **112**, 261, 1969.

[5] Mathews, F. S., Levine, M. and Argos, P.: *J. Mol. Biol.*, **64**, 449, 1972.

[6] Wyckoff, H. W., et al.: *J. Biol. Chem.*, **245**, 305, 1970.

[7] Shotton, D. M. and Watson, H. C.: *Nature*, **225**, 811, 1970.

[8] Chou, P. Y. and Fasman, G. D.: *Biochemistry*, **13**, 222, 1974.

[9] Tou, J. T. and Gonzalez, R. C.: *Pattern Recovnition Principles*, 1975.

[本文于 1984 年 3 月 26 日收到]

(上接第80页)

银量，最后按下式计算过氧化物酶的活性：

$$A = \frac{50 \times 2.53 \times C}{15 \times V \times W} = \frac{2.53 \times C}{3 \times V \times W}$$

式中 A：20℃ 下过氧化物酶的活性(1 克试样所含酶在 1 分钟内氧化愈创木酚的微克分子数).

C：从标准曲线上找出的银浓度(mg/12ml 显色液).

V：显色用的试样提取液毫升数.

W：试样重量(克).

50：试样提取液毫升数.

15：酶作用时间(分).

2.53：20℃下 1 毫克银存在时，15 分钟所氧化的愈创木酚微克分子数.

三、方法精密度

取同一试样，用同样条件重复测定多次，结果见表 1.

表 1：大米中过氧化物酶的活性

测定次数	测定结果 x_i	均差 $x_i - \bar{x}$	$(x_i - x)^2$	标准差 S
1	0.66	−0.02	0.0004	
2	0.69	+0.01	0.0001	
3	0.66	−0.02	0.0004	0.029
4	0.67	−0.01	0.0001	
5	0.72	+0.04	0.0016	
6	0.73	+0.05	0.0025	
7	0.66	−0.02	0.0004	
8	0.66	−0.02	0.0004	
$n = 8$	$\bar{x} = 0.68$		$\Sigma(x_i - \bar{x})^2 = 0.0059$	$S = \sqrt{\frac{0.0059}{7}}$

表 1 说明此法测定的结果波动性小，准确性好.

[注1]《植物生物化学分析方法》[苏] X. H. 波钦诺克著.

[本文于 1984 年 1 月 9 日收到]