

编码方式对蛋白质二级结构预测精度的影响

阮晓钢, 孙海军

(北京工业大学 电子信息和控制工程学院, 北京 100022)

摘 要: 为了比较在蛋白质二级结构预测中常用的氨基酸序列编码方式的优缺点, 借助前向型 BP 神经网络对正交编码、5 位编码、Codon(2 种)编码和 Profile 编码 5 种氨基酸编码方式进行了对比分析。实验结果显示, 用富含“生物进化信息”的 Profile 编码方式可以得到较高的预测结果, 同时也表明, 充分利用生物本身所具有的生物信息对提高蛋白质二级结构预测精度是非常重要的。

关键词: 氨基酸序列编码; 蛋白质二级结构; 神经网络

中图分类号: Q 518.1

文献标识码: A

文章编号: 0254-0037(2005)03-0229-07

自 20 世纪 60 年代中期人们开始从事蛋白质二级结构预测的研究工作以来^[1], 已提出 20 多种预测方法, 有的利用氨基酸的物理化学性质, 有的是利用序列的同源性, 还有的对已知的蛋白质结构进行统计分析等。自从 1988 年 L. Howard Holley 和 Martin Karplus 提出用人工神经网络来预测蛋白质二级结构以来, 人工神经网络一直是蛋白质二级结构预测的研究领域中的一个非常重要的工具^[2]; 而且, 与其他各种方法相比而言, 应用神经网络已经取得了较高的预测结果。但是用人工神经网络对蛋白质二级结构进行预测所用的氨基酸编码方式却不尽相同, 例如 Lamont 采用的是 Codon(密码子)和正交编码(标准编码)^[3]; 方慧生等采用的是正交编码^[4]; 王龙会等采用的是 5 位编码^[5]; Bohr 等采用的是 Profile 编码^[6]。这些研究显示, 不同的氨基酸编码方式对蛋白质二级结构预测的准确率有着重要影响, 用哪种氨基酸编码方式可以取得较高的预测率, 还需做比较具体深入的研究。

1 编码问题与编码方式

作者将对上述的 5 种常用的氨基酸编码方式进行分析和比较, 分析它们的优缺点和对蛋白质二级结构预测精度的影响。下面将分别介绍这 5 种编码方式的具体编码规则。

1.1 正交(标准)编码方式

在蛋白质二级结构预测研究领域里, 正交编码就是用 20 位二进制数唯一表示某一种氨基酸, 也就是说每一种氨基酸用 20 位二进制数来表示, 并且满足不同氨基酸的编码向量值的正交积为 0。例如可以这样为 20 种氨基酸依次编码, 氨基酸 A: 10000000000000000000; 氨基酸 C: 01000000000000000000; 氨基酸 D: 00100000000000000000;; 氨基酸 Y: 00000000000000000001。很显然, 这 20 个编码向量是正交的, 于是称其为“正交编码”。

1.2 5 位编码方式

蛋白质由 20 种氨基酸组成, 因此用 5 位二进制数可以唯一表示它们, 具体编码为氨基酸 A: 00001, 氨基酸 C: 00010, 氨基酸 D: 00100;; 氨基酸 Y: 10010。

收稿日期: 2003-10-28。

基金项目: 国家自然科学基金资助项目(60234020)。

作者简介: 阮晓钢(1958-), 男, 四川自贡人, 教授, 博士, 博士生导师。

1.3 Codon(密码子)编码方式一(基本型)

所谓密码子编码方式,即把氨基酸“还原”成 3 个碱基的组成形式(碱基转录成氨基酸表见文献^[21]),而碱基再用二进制数来表示.具体的编码过程分为 2 步,即碱基的编码和用碱基来表示氨基酸.

在 Codon 编码方式一里 4 种碱基 A、C、G、T 的编码是用 4 位二进制数来表示的,具体编码过程如下.

- 1) 碱基的编码
- A :1000 ;C :0100 ;G :0010 ;T :0001 ;A or G :1010 ;A or T :1001 ;C or G :0110 ;C or T :0101 ;A or C or T :1101 ;A or C or G or T :1111.

- 2) 用碱基表示氨基酸
- 把氨基酸“还原成”用 3 个碱基来表示的形式,如氨基酸 A 可以有 4 种碱基组合形式,即 GCA、GCC、GCG 和 GCT,组合起来就是 GC(A or C or G or T),于是,氨基酸 A 的编码方式是 0010 0100 1111;同理氨基酸 C 的 Codon 编码为 0001 0010 0101,以此类推.

1.4 Codon(密码子)编码方式二(扩展型)

扩展型 Codon 编码与基本型 Codon 编码不同,在这种 Codon 编码方式里 4 种碱基 A、C、G、T 的编码是用 10 位二进制数来表示的(因此,又叫做扩展型 Codon 编码);具体编码过程如下.

- 1) 碱基的编码
- A :1000000000 ;C :0100000000 ;G :0010000000 ;T :0001000000 ;A or G :0000100000 ;A or T :0000010000 ;C or G :0000001000 ;C or T :0000000100 ;A or C or T :0000000010 ;A or C or G or T :0000000001.

- 2) 用碱基表示氨基酸
- 把氨基酸用 3 个碱基来表示,如氨基酸 A 可以有 4 种碱基组合,即 GCA、GCC、GCG 和 GCT,组合起来就是 GC(A or C or G or T),于是,氨基酸 A 的编码方式是 0010000000 0100000000 0000000001;同理氨基酸 C 的 Codon 编码为 0001000000 0010000000 0000000100,以此类推.

1.5 Profile 编码

Profile 面是指在氨基酸序列的每个位置上 1 个氨基酸类型出现的相对概率.例如:氨基酸 Q 在其对应的行“Q Q...”序列中出现的概率显然是 1.0,因为在“Q Q...”序列中只有氨基酸 Q;再如,氨基酸 S 出现的概率是 0.4,因为在“S S Q A A”序列中 S 有 2 个,所以 S 的出现概率是 2/5 = 0.4,等等.一般认为这种编码带有较高的生物进化信息.其具体的编码规则如表 1 所示.

表 1 Profile 编码表例

Tab.1 Profile encoding

S _{eq}	A _M	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D
Q	Q	Q	Q	Q	Q	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0
S	S	S	Q	A	A	0	0	0	0	0	0	0	0.4	0	0.4	0	0	0.2	0	0	0
E	K	K	K	K	E	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0.2	0	0
P	P	P	P	P	P	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0
E	E	E	G	E	E	0	0	0	0	0	0	0.2	0	0	0	0	0	0	0.8	0	0
D	D	E	D	D	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.8
L	L	L	A	A	A	0	0.4	0	0	0	0	0	0.6	0	0	0	0	0	0.2	0	0.8
L	L	L	I	V	I	0.2	0.4	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0
K	K	K	E	K	H	0	0	0	0	0	0	0	0	0	0	0.2	0	0.6	0	0.2	0

其中 S_{eq} 列代表一个氨基酸序列片段, A_M 列是和测试序列(S_{eq})对比过的相关序列.从表 1 中看出,氨基酸序列片段 QSEPEDLLK 的 Profile 编码是其对应的行.例如氨基酸 Q 的 Profile 编码是 000000000000000001.0000,氨基酸 S 的 Profile 编码是 0000000000.400.400000.2000,其余的以此类推.

2 评测系统

2.1 结构设计

用一个 3 层结构的 BP 神经网络(见图 1)作为比较 5 种氨基酸编码方式的评测系统。在输入层引入了广为使用的“局部滑动窗”,这个“滑动窗”是沿着氨基酸序列“滑动”的窗口,每一次预测都是对“滑动窗”中间位置的氨基酸进行的。例如,对图 1 来说,这一时刻就是对窗中间的氨基酸“R”进行预测,下一时刻将对氨基酸“W”进行预测。作者所用滑动窗的长度为 9 个氨基酸的长度,如图 1 中输入层是 VRKKRWACD 等 9 个氨基酸,假设每种编码方式编码 1 个氨基酸需要 K 个神经元,那么输入层的总的输入神经元个数为 $9 \times K$ 。隐含层神经元个数分别取 2、10、20、40 等几种情况。输出层是由 3 个神经元组成的,以对应 3 种蛋白质二级结构状态 H、E 和 L。这 3 个神经元分别编码成 3 个二进制码,即 H(螺旋) 100、E(折叠) 010、L(卷曲) 001。

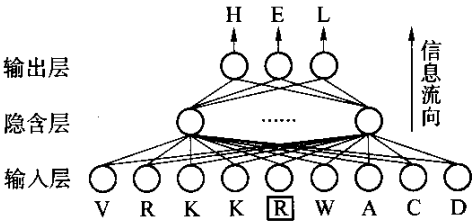


图 1 单神经网络模型
Fig.1 Pattern setting for single neural network

2.2 网络算法

1) 工作算法(working algorithm)

BP 神经网络的工作算法是由给定的输入信号产生输出信号的过程,即

$$WA:X \rightarrow Y$$

其中: X 是输入信号, Y 是输出信号。详细的工作算法如下(网络共有 3 层)。

① 计算输入层的电位向量

$$V^{(1)} = X - \theta^{(1)} \tag{1}$$

② 计算 1 到 2 层的输出向量和 2 到 3 层的膜电位向量

$$O^{(k-1)} = f^{(k-1)}(V^{(k-1)}) \tag{2}$$

$$V^{(k)} = W^{(k)T} O^{(k-1)} - \theta^{(k-1)} \tag{3}$$

③ 计算输出层(第 3 层)的输出向量

$$Y = O^{(L)} = f^{(L)}(V^{(L)}) \tag{4}$$

式中: $V^{(k)}$ 为第 k 层的膜电位; X 为样本输入向量; Y 为第 3 层的输出向量; W 为权值矩阵; θ 为阈值矩阵; f 为神经元作用函数; L 为 BP 神经网络第 L 层。

2) 学习算法(learning algorithm)

BP 神经网络的学习算法包括两大步骤:其一是输入的正向传播过程,即输入数据从输入层逐层向输出层传播,得到输出响应;其二是输出误差的反向传播过程,即输出的误差由输出层开始反向传播到输入层,而网络各层的权值改变量则根据传播到该层的误差大小来决定。也就是从样本中选取输入模式作为 BP 网络的输入,并测试其输出,即测试实际输出与期望输出的一致性,以指导 BP 网络权值和阈值的修改。BP 神经网络学习算法的步骤为

- ① 初始化网络的权值矩阵 W 和神经元阈值矩阵 θ 为小的随机数矩阵。
- ② 提供学习样本,即输入向量 X 和输出向量 Y_n ($n = 1, 2, 3$)。
- ③ 计算网络的实际输出及隐含层各神经元的输出

$$O_{pj} = f \left(\sum_i W_{ij} I_{pi} - \theta_j \right) \tag{5}$$

$$O_{pk} = f \left(\sum_j W_{jk} I_{pj} - \theta_k \right) \tag{6}$$

式中: f 为 Sigmoid 函数, $f(x) = 1 / [\exp(-x) + 1]$; I_{pi} 为训练样本为 p 时节点 i 的输入; O_{pj} 为训练样本为 p

时隐含层第 j 个节点的输出 ; O_{pk} 为训练样本为 p 时输出层第 k 个节点的实际网络输出.

④ 计算误差对各层的影响

$$\delta_{pk} = O_{pk}(1 - O_{pk}) \times (Y_{pk} - O_{pk}) \tag{7}$$

$$\delta_{pj} = O_{pj}(1 - O_{pj}) \times \sum \delta_{pk} W_{jk} \tag{8}$$

式中 : δ_{pk} 和 δ_{pj} 分别是输出层和隐含层反传误差信号 ; Y_{pk} 为训练样本为 p 时网络输出层节点 k 的期望输出.

⑤ 修正权值和阈值

$$W_{jk}(n+1) = W_{jk}(n) + \eta \delta_{pj} O_{pj} + \alpha (W_{jk}(n) - W_{jk}(n-1)) \tag{9}$$

$$\theta_j(n+1) = \theta_j(n) + \eta \delta_{pj} + \alpha (\theta_j(n) - \theta_j(n-1)) \tag{10}$$

式中 : η 为学习率(一般取 0.01 ~ 0.8) ; α 为动量因子(一般取 0.9 左右).

⑥ 计算网络输出误差

$$E = 1/2 \sum (Y_{pk} - O_{pk})^2 \tag{11}$$

⑦ 判断误差是否满足要求 ,若满足则转到⑧ ,否则转到③.

⑧ 训练结束.

2.3 预测准确率的衡量

选用最简单并且被广泛使用的评估公式来表示预测的精度 ,总的预测准确度公式为

$$Q_3 = 100 \sum_{i=1}^3 C_i / N \tag{12}$$

式中 : Q_3 表示总的预测精度 ; C_i 表示每 1 种二级结构状态预测正确的氨基酸总数($i = 1, 2, 3$),其中 $i = 1, 2, 3$ 分别代表 H、E、L ; N 表示总的样本(氨基酸)个数.

每种蛋白质二级结构状态预测的准确度公式为

$$Q_H = 100 C_H / N_H \tag{13}$$

$$Q_E = 100 C_E / N_E \tag{14}$$

$$Q_L = 100 C_L / N_L \tag{15}$$

式中 : Q_H 、 Q_E 和 Q_L 分别代表 3 个蛋白质二级结构状态的预测精度 ; N_H 、 N_E 和 N_L 分别是 3 种二级结构状态的氨基酸总的个数.

3 评测数据库

作者主要是对比研究 5 种常用氨基酸编码方式对蛋白质二级结构预测结果的影响 ,因此 ,样本数据和测试数据应该是相同的 ,同时考虑到 Profile 编码方式是基于同源蛋白质的 ,所以把数据来源定位在 ftp://

ftp.ebi.ac.uk/pub/databases/hssp/. 从 HSSP(蛋白质同源结构数据库)数据库中提取出 36 个蛋白质 ,这 36 个蛋白质文件是文本文件为 1azu ,1bbp _ A ,1cbh ,1fc2 _ A ,1ovo _ A ,1pyp ,1tgs _ I ,2aat ,2ccy _ A ,2mev _ 4 ,2mhu ,2rsp _ A1 ,2tgp _ I ,2utg _ A ,2wrp _ R ,3hmg _ B ,3mt ,4cpv ,4acn ,4cpp ,4hir ,9wga _ A ,1cm ,1cse _ I ,1cdt _ A ,1cc5 ,7rsa ,1r09 _ 2 ,1fdl _ H ,1eca ,2gbp ,2hmr _ A ,5er2 _ E ,2stv ,5ldh ,3ebx.

表 2 列出 HSSP 数据库文件的主要结构 ,序号列代表氨基酸序列中氨基酸的序号 ; A_A 列是氨基酸序列中具体的氨基酸 ,它将作为神经网络的输入 ; S_S 列是氨基酸对应的二级结构状态 ,它将作

表 2 HSSP 数据库文件的主要格式

Tab.2 Format of database HSSP

序号	A_A	S_S	A_M	P_A
1	A	S	AAAAAAAAAA	1.0
2	V	S	AHSAAAAAVV	0.2
3	K	S	KKKKKKKPPK	0.8
4	H	H	AAHHCGAHHA	0.4
5	W	H	WWPPPAWWWA	0.5
6	D		DEPPPPFLLLL	0.1
7	G	E	GGGGHHHHIL	0.4
8	P	E	PPPPPPPEPPP	0.9
9	R	E	RARDNNSTSS	0.2
10	C		CCCCCCCPC	0.9

为神经网络的教师信号。由于要预测的蛋白质结构是 3 状态的,即 H 螺旋、E 折叠和 L 薄片,而 HSSP 数据库中的二级结构状态是 8 状态的,这种 8 状态的二级结构是 3 状态的结构的细化。于是,先把 8 状态转化成 3 状态的形式。即 H 包括 H(α 螺旋)、C(3_{10} 螺旋)和 I(π 螺旋);E 包括 E(扩展折叠)和 B;L 包括 T(弯曲)、S(薄片)和其他。 A_M 列是和测试序列对比的相关序列; P_A 列给出了在每一个位置一种氨基酸类型出现的相对频率。

从 36 个蛋白质中选取 4 000 个氨基酸后,随机取出 400 个氨基酸作为测试数据集,剩下的 3 600 个作为训练数据集。它们 3 种二级结构所占百分比为:训练样本中 H 占 20%、E 占 31%、L 占 49%,测试样本中 H 占 20%、E 占 31%、L 占 49%。

4 试验与分析

把从公共数据库中提取出来的训练样本数据和测试样本数据分别用这 5 种不同的编码方式进行编码,评测系统的隐含层神经元数分别定为 2、10、20 和 40,这样做是为了比较这 5 种氨基酸编码方式在不同 H (隐含层神经元数)下的预测准确率,之后把学习样本编码数据输入到评测系统中,开始训练评测系统,待系统训练完毕后,再把测试样本编码输入到评测系统中得到实际系统输出的评测结果,结合期望结果,利用公式(12)~(15)得到表 3。

表 3 5 种编码方式预测结果的比较
Tab.3 Comparison of prediction accuracy of five encodings

编码方式	H	Q_3	Q_H	Q_E	Q_L
正交编码	2	0.543 478	0.518 797	0.416 667	0.634 842
5 位编码	2	0.492 754	0.288 026	0.379 310	0.772 499
Codon 编码一	2	0.466 184	0.418 605	0.218 391	0.615 384
Codon 编码二	2	0.516 908	0.501 628	0.333 333	0.606 383
Profile 编码	2	0.625 604	0.496 124	0.609 195	0.755 315
正交编码	10	0.502 425	0.451 128	0.357 143	0.631 016
5 位编码	10	0.478 261	0.333 333	0.252 874	0.696 335
Codon 编码一	10	0.497 584	0.519 379	0.344 827	0.545 454
Codon 编码二	10	0.492 754	0.387 597	0.321 839	0.652 850
Profile 编码	10	0.608 096	0.604 651	0.528 736	0.680 851
正交编码	20	0.525 507	0.458 667	0.434 686	0.571 587
5 位编码	20	0.471 014	0.449 614	0.302 851	0.581 152
Codon 编码一	20	0.471 014	0.356 589	0.367 816	0.600 000
Codon 编码二	20	0.478 261	0.403 101	0.390 805	0.595 745
Profile 编码	20	0.640 097	0.573 647	0.586 207	0.744 681
正交编码	40	0.528 985	0.458 647	0.407 642	0.652 632
5 位编码	40	0.502 415	0.379 845	0.390 805	0.626 263
Codon 编码一	40	0.478 206	0.387 597	0.379 710	0.589 743
Codon 编码二	40	0.454 106	0.392 090	0.352 218	0.564 767
Profile 编码	40	0.683 575	0.612 401	0.666 667	0.776 596

为了更直观地表现表 3,把表 3 拆成 4 张正确率与 H (隐含层神经元个数)的关系曲线图(见图 2~5)。从图 2 中可以清晰地看出用富含生物进化信息的 Profile 编码方式编码神经网络所取得的预测结果(Q_3)要比其他的编码方式所取得的预测结果(Q_3)高,正交编码次之。这说明虽然 Profile 编码比较复杂,但是生物进化信息对提高蛋白质二级结构预测准确率是非常重要的。而像正交编码和 5 位编码,虽然编码简单明了,但是,所携带的生物信息却很少,致使预测结果不如用 Profile 编码所取得的预测结果高。同

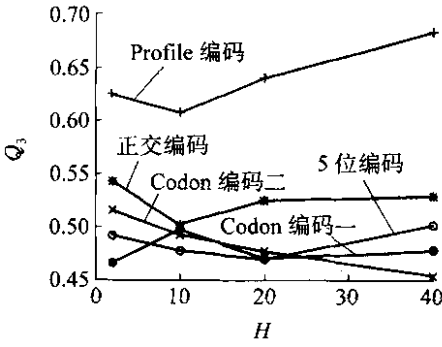


图 2 Q_3 与 H 关系曲线

Fig.2 The relation curve between Q_3 and H

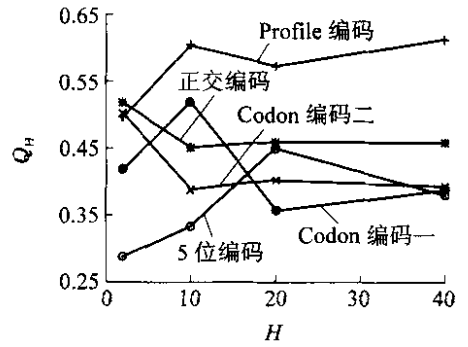


图 3 Q_H 与 H 关系曲线

Fig.3 The relation curve between Q_H and H

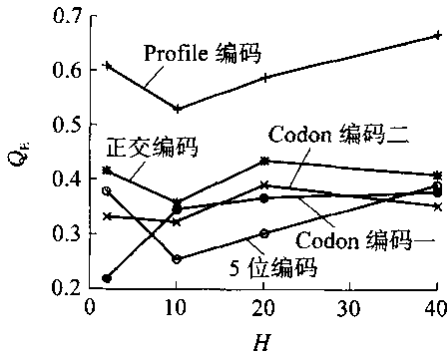


图 4 Q_E 与 H 关系曲线

Fig.4 The relation curve between Q_E and H

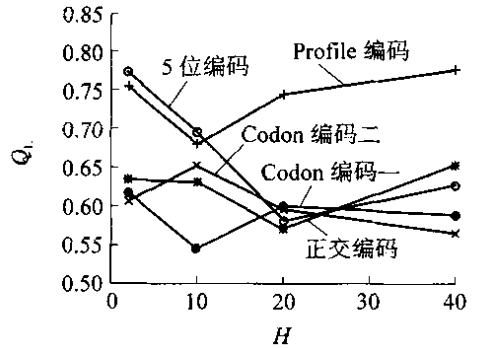


图 5 Q_L 与 H 关系曲线

Fig.5 The relation curve between Q_L and H

时用 Profile 编码得到的预测结果也比文献^[6]所用的单个神经网络预测的结果高。出现这种的结果的原因,是数据集和文献^[6]中所用的数据集不一样,样本数据集较文献^[6]用的数据集小,而且各部分所占的比例也不一样。在作者的数据集中 H(螺旋)占的比例是 20%,E(薄片)占的比例是 31%,L(卷曲)占的比例是 49%,而文献^[6]中 H、E、L 占的比例分别是 30%、23%、47%。另一个值得注意的是,用 Codon 编码所取得的预测结果没有文献^[3]中所取得预测结果好,依文献^[3]所述,用 Codon 编码进行蛋白质二级结构预测所取得的结果总的来说应该比正交编码所取得的结果要好,而且 Q_E 的准确率要较正交编码高,而从图 4 可以看出,结果与文献^[3]所取得的结果相反,出现这个结果是因为:第 1,所用的数据集和文献^[3]所用的数据集不一样,而且所用的数据集远小于文献^[3]中所用的数据集;第 2,由于数据是从 HSP 数据库中提取出来的,而 HSP 数据库是基于同源蛋白质的数据库。图 2~5 中的曲线显示,无论是 Q_3 、 Q_H 、 Q_E 和 Q_L ,Profile 编码基本上都比其他的几种编码方式优越,尤其是 Q_3 和 Q_E 。这表明,虽然 Profile 编码较为复杂,但相对于其他几种编码方式而言 Profile 编码更适合同源性蛋白质的二级结构预测;而且,随着隐含层神经元数目(本文所取的隐含层神经元数目)的增加,用 Profile 编码得到的预测精度基本上是升高的,而其他的编码方式随着隐含层神经元数目的增加,其预测准确率变化规律不是很明显。

5 结束语

借助前向型 BP 神经网络对正交编码、5 位编码、Codon(2 种)编码和 Profile 编码 5 种氨基酸编码方式进行了对比分析。分析表明,正交编码和 5 位编码虽然编码简单,但是由于其携带生物信息量较少,致使其二级结构预测的准确率较低,从表 3 中也发现,用 Codon 编码方式取得的预测结果都没有文献^[2]中用 Codon 编码方式取得的预测结果高。而富含生物进化信息的 Profile 编码方式可以得到较高的预测结果,这

表明 ,Profile 编码更适合同源性蛋白质的二级结构预测和充分利用生物本身所具有的生物信息对提高蛋白质二级结构预测精度是非常重要的。因此 ,如何更加有效地把更多的生物信息进行编码 ,如何更加有效地利用氨基酸之间的生物联系及蛋白质二级结构之间的联系 ,将成为蛋白质二级结构预测研究领域中的重要研究课题。

参考文献：

[1] 来鲁毕. 蛋白质的结构预测和分子设计[M]. 北京 :北京大学出版社 ,1993.
LAI Lu-bi. The Structure of the Protein is Predicted and Designed With the Molecul[M]. Beijing :Peking University Press ,1993. (in Chinese)

[2] HOLLEY L H , KARPLUS M. Neural network applied in prediction of protein secondary structur[J]. Biophysics ,1989 ,86 :152-156.

[3] LAMONT Owen , LIANG Hiew Hong , BELLGARD Matthew. Data representation influences protein secondary structure prediction using artificial neural networks[A]. Seventh Australian and New Zealand Intelligent Information Systems Conference[C]. Perth ,Western Australia :ANZIS ,2001 . 18-21 .

[4] 方慧生 , 相秉仁 , 安登魁. 改进 Madaline 学习算法预测蛋白质二级结构[J]. 中国药科大学学报 ,1996 ,27(6) :366-369.
FANG Hui-sheng , XIANG Bing-ren , AN Deng-kui. Improving study algorithms of Madaline to predict second structure of protein[J]. Journal of China Pharmaceutical University ,1996 ,27(6) :366-369. (in Chinese)

[5] 王龙会 , 石峰. 遗传神经网络及其在蛋白质二级结构预测中的应用[J]. 数学杂志 ,2002 ,24(2) :179-185.
WANG Long-hui , SHI Feng. Protein secondary structure prediction based on genetic meira network[J]. Journal of Mathematics , 2002 ,24(2) :179-185. (in Chinese)

[6] BOHR Henrik , BOHR Jakob , BRUNAK Seren , et al. Protein secondary structure and homology by neural networks[J]. The Ahelices in Rhodopsin ,1988 ,24(1 2) :223-228.

Research on Encode Influencing Protein Secondary Structure Prediction

RUAN Xiao-gang ,SUN Hai-jun

(College of Electronic Information and Control Engineering , Beijing University of Technology , Beijing 100022 ,China)

Abstract : In order to compare orthogonal encoding , five byte encoding , Codon encoding(two) and Profile encoding to find out their virtues and shortcomings , we carefully studied them with artificial neural network . Results indicated that Profile encoding that preserves the redundant evolutionary information got higher prediction performance , and that how to use biologic evolutionary information effectively is very important to improve accuracy of protein secondary structure prediction .

Key words : amino acid sequence encode ; protein secondary structure prediction ; neural network