第 31 卷第 1 期 数学的实践与认识 Vol. 31 No. 1
2001 年 1 月 MATHEMATICS IN PRACTICE AND THEORY Jan. 2001

DNA 序列的分类模型

汤诗杰, 周 亮, 王晓玲
指导老师: 孙广中

(中国科技大学, 合肥 230026)

编者按: 本文提出了DNA 序列分类的三种模型, 其一, 基于A、G、T、C 四种碱基出现的频率; 其二利用了同一碱基在序列中的间隔, 这一信息是单纯考虑频率所不能包含的; 在第三种模型中, 作者把DNA 序列视为一个信息流, 考虑每增加一个字符所带来的信息增量. 尽管文中信息量的定义方式仍可讨论, 但本文思想新颖活跃, 有其独特之处. 本文最后的分类方法, 是以上三种的综合使用.

摘要: 本文针对DNA 序列分类这个实际问题, 提出了相应的数学模型. 为了很好的体现DNA 序列的局部性和全局性的特征, 我们给出了衡量分类方法优劣的标准, 即在满足一定限制条件的情况下, 是否能充分反映序列的各方面特性.

依据我们提出的判别标准, 单一标准的分类是无法满足要求的. 我们的方法是侧重点不同的三种方法的综合集成. 这三种方法分别体现了序列中元素出现的概率, 序列中元素出现的周期性, 序列所带有的信息含量. 利用这个方法, 完成了对未知类型的人工序列及自然序列的分类工作. 最后, 对分类模型的优缺点进行了分析, 并就模型的推广作了讨论.

1 问题的提出(略)

2 问题的分析

这是一个比较典型的分类问题, 为了表述的严格和方便, 我们用数学的方法来重述这个问题. 已知字母序列 $S_1, S_2, S_3 \cdots\cdots S_{40}$, $S_i = x_{i1}x_{i2}x_{i3}\cdots x_{in_i}$, 其中 $x_j \in \{a, t, c, g\}$; 有字符序列集合 A, B, 满足 $A \cap B = \phi$, 并当 $1 \le i \le 10$ 时, $S_i \in A$; 当 $11 \le i \le 20$ 时, $S_i \in B$. 现要求考虑当 $21 \le i \le 40$ 时, S_i 与集合 A 及集合 B 的关系.

在这里, 问题的关键就是要从已知的分好类的 20 个字母序列中提取用于分类的特征. 知道了这些特征, 我们就可以比较容易的对那些未标明类型的序列进行分类. 下面我们将首先对用于分类的标准问题进行必要的讨论.

3 分类的标准及评价

首先, 我们提取的特征应该满足以下两个条件:

(1) 所取特征必须可以标志A 组和B 组. 也就是说, 我们利用这些特征应该可以很好的区分已经标示分类的 20 个序列. 这是比较显然的一个理由.

(2) 所取特征必须是有一定的实际意义的. 这一点是决不能被忽视的. 比如, 如果不考虑模型的实际意义, 我们就可以以序列的开头字母为分类标准: 已知在B 类中的十个序列都是以 gt 开始的, 而已知在A 类中 10 个序列没有以 gt 开始的, 甚至以 g 开始的都没有. 显然这是满足上面的第一个条件的. 如果仅因此就认为这种特征是主要的, 并简单的利用这个特征将所有待分类的序列分成两类, 显然是不甚合理的.

对于这样的一个复杂的分类问题, 需要考虑的因素很多, 也是就说, 可供我们使用的分类特征有许多. 如何从众多的因素中提取分类的主要因素, 是我们处理这个问题的困难之处. 上面的第一个条件是我们的分类方法所必须满足的, 可以看作是个限制条件; 而第二个条件是我们在设计分类方法时必须考虑到的, 可以看作是对分类方法优劣的一种衡量, 是某种意义下的目标函数.

4 模型的建立及分析

由上面的分析可知, 由于DNA 序列本身的复杂性, 我们很难在不知道确切的分类标准的情况下, 使用单一的方法来处理这个分类问题. 由于,DNA 序列同时具有局部性和全局性的特征, 我们尝试综合使用几种设计思想不同的方法来处理这个问题, 以使该分类方法具有好的分类性能和相当的健壮性.

下面我们先从不同的角度出发, 提出三种侧重点不同的分类方法, 第一种从频率角度出发, 第二种从字母出现的周期性的角度出发, 第三种从序列所带的某方面的信息量出发, 并给出它们单独使用时的分类结果. 我们认为, 这三方面综合考虑, 可以较好的体现出序列各个方面的特征, 最后, 从这三种方法出发, 得到一个综合系统的分类方法, 并利用它得到了最终的 182 个序列的分类结果.

方法 1 基于字母出现频率

不同段的DNA 中, 每个碱基出现的概率并不相同, 从生物理论中, 我们知道, 编码蛋白质的DNA 中 G、C 含量偏高, 而非编码蛋白质的DNA 中A、T 含量偏高. 因此, A、G、T、C 的频率中会含有很多的信息, 下面给出A、B 组的频率统计. 见表 1, 表 2 (略).

由统计的数字可以看出,A 组的碱基构成与B 组的碱基构成有较大的不同. A 组的G 含量较高,B 组的T 含量较高. 为做定量化的分析, 引入数学中的内积概念, 即将A、T、G、C 的频率分别作为四维向量的四个分量(P_A, P_G, P_T, P_C), 现在我们得到两组向量A_i、B_i $(i = 1, 2, 3\cdots\cdots 10)$, 然后将未知的序列 21～40 作为一个新的向量C, 要将它归入A 组或B 组, 我们可以尝试在Hilbert 空间中将向量归一化后求C 与A 组和B 组的平均距离. 记$\overline{C}$、$\overline{A_i}$、$\overline{B_i}$为归一化后的向量. 为此, 我们计算内积和 $\sum_{i=1}^{10}\overline{C}\cdot\overline{A_i}$ 与 $\sum_{i=1}^{10}\overline{C}\cdot\overline{B_i}$, 其中内积定义为欧氏度量引导出的内积$(c_1, c_2, c_3, c_4)\cdot(a_1, a_2, a_3, a_4) = c_1a_1 + c_2a_2 + c_3a_3 + c_4a_4$. 即

$$\text{内积} = \frac{(P_A, P_G, P_T, P_C)_A \cdot (P_A, P_G, P_T, P_C)_{未知}}{|A| \cdot |未知|}$$

内积小的两个序列, 我们可以认为它们的相关性小, 而内积大的序列, 我们就认为其相关性大. 因此, 如果 $\sum_{i=1}^{10}\overline{C}\cdot\overline{A_i} > \sum_{i=1}^{10}\overline{C}\cdot\overline{B_i}$, 则认为C 应归入A 类, 否则认为它应归入B 类.

计算结果如表 3 所示.

由此, 我们找到了区分C 组的一种方法, 这种比较 $\sum_{i=1}^{10}\overline{C}\cdot\overline{A_i}$ 和 $\sum_{i=1}^{10}\overline{C}\cdot\overline{B_i}$ 的方法, 我们可以归纳为一个目标函数$F_1(l)$, 即

$$F_1(l) = \frac{\sum_{i=1}^{10}\overline{C}\cdot\overline{A_i}}{\sum_{i=1}^{10}\overline{C}\cdot\overline{B_i}}$$

表 3

未知的序号	与A 组的内积	与B 组的内积	属于的类别	未知的序号	与A 组的内积	与B 组的内积	属于的类别
1	0. 815781	0. 938814	B	11	0. 852231	0. 920957	B
2	0. 926922	0. 803952	A	12	0. 866976	0. 853967	A
3	0. 939727	0. 656827	A	13	0. 860955	0. 917122	B
4	0. 788524	0. 937135	B	14	0. 961689	0. 67678	A
5	0. 948194	0. 772073	A	15	0. 960322	0. 739089	A
6	0. 801201	0. 930121	B	16	0. 904282	0. 747578	A
7	0. 953019	0. 76695	A	17	0. 944724	0. 723664	A
8	0. 746071	0. 968035	B	18	0. 75862	0. 954652	B
9	0. 931007	0. 613193	A	19	0. 885631	0. 811837	A
10	0. 897774	0. 844082	A	20	0. 75584	0. 941	B

方法一讨论　这种方法是从概率统计的角度分析问题, 通过对每个字母出现频率的计算, 找出A ,B 两类DNA 链中的频率特性, 建立四维向量空间, 然后对待求分类的序列统计频率, 与已知分类的向量进行内积运算, 找出量化的关联性, 从而将其分类　但这种方法也有其局限性, 在统计字母出现的频率时, 忽略了字母所在位置以及各个字母之间的相互关系, 造成用这种方法对已知分类的序列进行检验时, 个别频率特性不明显的序列不太容易分类　所以, 这种方法虽然有其科学性, 但还不够完善, 不能完全体现序列的所有特征

方法二　基于字母出现周期性

在以上进行了基于字母出现频率的分类之后, 我们认为, 一个序列所含的信息远不止每个字母出现的频率, 还有字母出现和它前后若干个字母的相关联性, 字母在序列中出现的规律性等等　前一个问题我们留到下面讨论, 现在我们想办法处理后一个问题

对于某单个字母, 以 a 为例, 假设它在序列中第 $t_1, t_2, \cdots\cdots, t_{k+1}$ 个位置出现, 我们试图找出这些数字之间的关联　首先, 可以认识到考查 t_i 的分布及绝对值是意义不大的, 因为序列是一大段DNA 中的一个片断, 片断的起始段不同会导致 t_i 的不同　于是为了抵消 t_i 的线性位移, 考虑下面一组值

$$s_i = t_{i+1} - t_i \qquad i = 1, 2, \cdots, k$$

即字母 a 出现的间距

可以看出, 序列 $s_1, s_2, \cdots\cdots, s_n$ 的大小包含的信息是 a 的"稠密度", 也可看成一个与频率有关的量, 前面已经处理过　所以我们可以考虑序列 $s_1, s_2, \cdots\cdots, s_n$ 的波动幅度, 幅度越小, 说明 s_i ($i = 1, 2, \cdots\cdots, k$) 的值越趋于统一, 即 a 的出现周期性越大　而表征波动幅度的量在统计中是中心矩　现求 s_i 的二阶中心矩, 即方差

$$\mathrm{Var}_a(s_1, s_2, \cdots\cdots, s_n) = \sqrt{\frac{1}{n-l}\sum_{i=1}^{n}(s_i - \bar{s})^2}, \quad \bar{s} = \frac{\sum_{i=1}^{n} s_i}{n}$$

同理, 可以求出Var_g、Var_t、Var_c。

由所得数据知, 对Var_g 与Var_t, 上述方法对A、B 组的区分率很高, 就有良好的可分辨性　为了强调这种特征的显著性, 我们用 $F_2 = \mathrm{Var}_g / \mathrm{Var}_t$ 作为这种方法的目标函数

由图 1 可以看出点与原点连线的斜率在A 组中和B 组中有显著差别, 根据这个特征,A 组和B 组可以很好地区分开来, 并且较好地弥补了方法一中的不全面之处

方法二讨论　这种方法是从序列中相邻相同字母之间的距离即字母出现的周期性着手

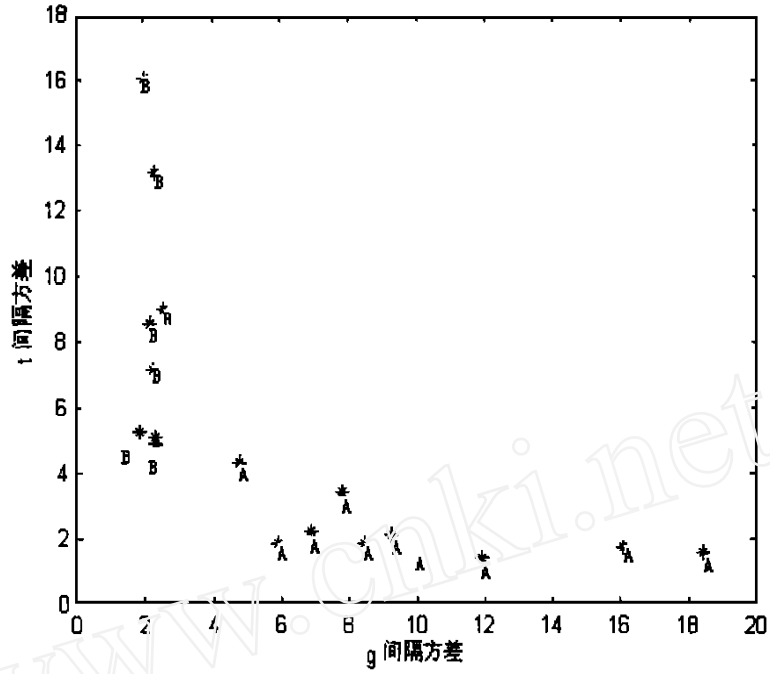

图 1

分析的. 它统计了每个字母在序列中两次出现的间隔, 并且用方差度量这种间隔的波动大小, 由此找到了一个能较好区分 A, B 组的目标函数, 综合地考虑了序列全局和局部的性质.

方法 3 基于序列熵值

我们可以把一串 DNA 序列看成一个信息流, 这与生物学的基础知识是相应的. 关于 A、B 的分类, 可以考虑其单位序列所含信息量(即熵)的多少. 从直观上来看, 我们可以认为, 重复得越多, 信息量越少. . 这是我们通过观察 A、B 组的特点而归纳出的方法.

设序列为 $L = (a_1, a_2, a_3, \cdots\cdots, a_n)$; 前 m 个字符所带的信息量为 $f_m(l)$, 记

$$g_m(l) = f_m(l) - f_{m-1}(l),$$

即 $g_m(l)$ 为加上第 m 个字母之后所增加的信息量. 然后, 由 $g_m(l) = f_m(l) - f_{m-1}(l)$, 得 $f_n(l) = \sum_{i=1}^{n} g_i(1)$, 则 $f_n(l)$ 为整个序列所带的信息量. $F_3(l) = \frac{f_n(1)}{|1|}$ 即为单位长度所带的信息量. 现在的问题就归结为如何找出一个合适的 $g_m(l)$.

我们有理由认为: g 具有以下性质:

性质 1: $g_m(l) > 0$. 即任意加上一个字符, 它或多或少带有一定信息量;

性质 2: 第 m 个字符(或者是以它结尾的较短序列)与前面的序列(信息流)重复得越多, $g_m(l)$ 的值必然越小;

性质 3: 第 m 个字符(或者是以它结尾的较短序列)如果和与它靠得越近的重复, $g_m(l)$ 的值越小; 和与它离得越远的重复, $g_m(l)$ 的值越大;

性质 4: $f_0(l) = 0$.

对此, 我们可以构造如下函数:

$$g_m(l) = \frac{b}{b + t_1\sigma_1 + t_2\sigma_2 + \cdots\cdots + t_p\sigma_p}$$

其中 b 为防止分母为零而设的一个小正数;

$$\sigma_i = \sum_{t=1}^{m} a^t \delta_{it};$$

$$\delta_{it} = \begin{cases} 1 & \text{以第 } m-t \text{ 个字符结尾的 } i \text{ 字串且与以第 } t \text{ 个字符结尾的 } i \text{ 字串完全相同} \\ 0 & \text{否则} \end{cases}$$

a 为一个小于 1 的数, 其存在体现了 g 的性质 3, 即如果越近的位置出现重复, 认为字串信息量越少, 反之较多.

σ_i 的表达式中, t 表示两个相同字串之间的距离, i 表示字串长度, 这个表达式定量的给出距离和信息量之间的关系.

又由于长度不同的字串重复对信息量的影响是不同的, 所以必须在 σ_i 前乘上一个权值 t_i, 由概率统计的知识可知, 这种影响是呈指数上升的, 则可选择一适当的常数 $c > 1$, 使得 $t_i = c^{i-1}$, 这个表达式定量的给出长度和信息量之间的关系.

可以认为, 字串长度太大的重复非常少见, 则可将 p 取为某一固定的正数. 那么, 给出 a、b、c、p 四个参数, 就可以把 f_n 严格确定下来. 通过反复上机搜索, 我们认为, 取 $p = 6$, 即只检查长度为 1 到 6 的字串即可.

另外, 取 $a = 0.392$, $b = 0.1$, $c = 3$ 可以将 A、B 组 $F_3(l)$ 值分得较开, 并可以用来处理未知数据.

方法三讨论　这种方法从序列的信息量(熵) 入手, 认为当序列中有大量的重复元素时, 信息量就会比重复少的序列所含有的信息少. 所以, 其侧重点是是序列前后的重复性, 也就是序列元素的相关性. 从所给的 A, B 两类中可以很清楚地看到 B 中序列重复量大, 所含的信息明显少于 A 组, 而这个特征就被我们定义的熵函数凸显出来. 将 DNA 序列看成一个信息流的方法由于其在实际问题中的广泛背景, 将会是一个很有价值的想法, 统计学和信息论的一套非常成熟的强大工具也会在 DNA 研究中发挥巨大的作用.

综合模型的建立

以上我们分别用三种方法得出了分类方案, 这三种方案分别基于三种不同的方面对问题进行分析. 第一种方法主要考虑的是单个字母出现的频率; 第二种方法主要考虑每个字母的出现是否具有周期性; 而第三种方法则考虑的是每条 DNA 所蕴含的信息量. 我们将这三种方法对 A、B 组自身进行了检验, 都得到了较令人满意的结果, 但因为每个模型都只突出考虑序列某一方面的特征, 所以, 总有一些不尽如人意的地方, 于是, 我们认为应该把三种方法综合起来考虑, 使序列各方面的特征都能得到体现, 以使分类更加科学.

下面就是我们将几种方法综合考虑得到最后结果.

以上我们用三种方法得到了三个目标函数: $F_1(l), F_2(l), F_3(l)$, 这三个目标函数可以作为分类的判别标准. 将它们看成定义在序列空间 $L = \{l \mid l$ 是由 a, g, t, c 四个字母组成的序列$\}$ 上, 作用于实轴上的函数. 现在, 我们必须找到一个函数 F, 使得 F 可以体现序列的各个特征.

由于 $F_1(l), F_2(l), F_3(l)$ 的值域范围差别很大, 为了有效的比较这三个函数, 我们必须将它们归一化, 将 $\xi_i = f_i(l)$ ($i = 1, 2, 3$, 以下同) 看成一定义在 L 空间上的随机变量, A, B 为 L 的子集, 则将 f_i 归一化得

$$g_i = \frac{\xi_i - E\xi_i}{\mathrm{Var}\xi_i} \tag{1}$$

而现有样本点 $f_i(11), f_i(12), \cdots\cdots, f_i(120)$ 利用距估计方法估计得:

$$E\xi_i = \frac{1}{n}\sum_{j=1}^{n} f_i(l_i)$$

$$\mathrm{Var}\xi_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(f_i(l_i) - ES_i)^2}$$

代入(1)即得 g_i

现估计 g_i 投射 L 的点到实轴上后, $g_i(A)$ 和 $g_i(B)$ 的分界点 x_i, 其中

$$g_i(A) = \{g(a) \mid a \in A\}$$

$$g_i(B) = \{g(b) \mid b \in B\}$$

以 g_1 为例, A 的 10 个样本点和 B 的 10 个样本点不能被一个分界点分开, 有极大似然估计的思想, 分界点应该把尽可能多的点分开, 即

$$x_i \in (-0.276758, 0.482296)$$

由于 $g_i(l)$ 的分布未知, 故只能假设其满足较均匀的分布, 则 A, B 的分界点的最好估计为 $\frac{Eg_i(A) + Eg_i(B)}{2}$, 而 $Eg_i(A) + Eg_i(B)$ 的矩估计为 $\sum_{i=1}^{20} g(l_i) = 0$ (由 g 的定义). 恰好 $0 \in (-0.276758, 0.482296)$, 则 $x_1 = 0$ 是分界点的最佳估计.

同理, $x_2 = 0, x_3 = 0$ 分界是 g_2, g_3 对应分界点的最佳估计.

令 $F = a_1g_1 + a_2g_2 + a_3g_3$, 则其分界点为 $x = a_1 \times 0 + a_2 \times 0 + a_3 \times 0 = 0$

由 F 的构造方法知, F 作用到 A 样本上大于零, 作用到 B 样本上小于零. 我们确定适当的权值, 以此作为 A, B 的分类法即可. 根据不同的实际情况, 可以相应调节这三个权值, 以体现分类中的不同因素所在的比重. 在下面的计算中, 我们简单的取 $a_1 = 1, a_2 = -1, a_3 = 0.5$. 得到的结果如表 4, 表 5 所示:

表 4

	序号	目标函数值		序号	目标函数值		序号	目标函数值		序号	目标函数值
	1	1.80288		6	1.75355		11	-1.38528		16	-2.60295
A	2	1.75894	A	7	1.25115	B	12	-1.22372	B	17	-0.0165438
	3	2.5887		8	1.41371		13	-0.940004		18	-1.31022
组	4	0.27582	组	9	1.9011	组	14	-0.93612	组	19	-2.6043
	5	2.1781		10	1.97282		15	-2.27465		20	-3.603

表 5

序号	目标函数值	类别	序号	目标函数值	类别
21	-1.96454	B	31	-1.06638	B
22	0.873279	A	32	-0.668504	B
23	2.32887	A	33	-0.877053	B
24	-1.48005	B	34	2.60904	A
25	1.21328	A	35	1.69535	A
26	-1.184	B	36	1.22298	A
27	1.22569	A	37	1.83991	A
28	-3.71616	B	38	-3.01466	B
29	2.69272	A	39	0.499763	A
30	0.550393	A	40	-2.77993	B

由以上数据可以看出, 我们构造的目标函数具有较好的区分度　对于A 组, 目标函数值都大于零; 而对B 组, 目标函数值都小于零　也就是说, 用这种方法, 对A、B 组样本的区分率已达到了 100%.　正如前面所说, 这种方法综合了序列中的许多信息　因此, 我们完全可以采用这个标准来区分C 组　表 5 是对C 组区分的结果

对 20 个未标明分类的人工序列的分类结果为:

A 类: 22, 23, 25, 27, 29, 30, 34, 35, 36, 37, 39　　B 类: 21, 24, 26, 28, 31, 32, 33, 38, 40

同样的, 我们利用这种方法对所给的 182 个自然序列进行了分类, 结果如下所示(略).

5　模型的评价及推广

在我们的模型基础上提出的分类方法可以很好的验证已知的 20 个序列, 并且很好的完成了对未知类型序列的分类　我们认为这种模型, 同时考虑了序列中元素的局部性质和序列的全局性质, 具有相当的实际背景　当我们知道分类标准的更多信息时, 我们可以很方便的调整模型中的参数, 使之符合新的情况, 具有很好的自学习性　但这个模型比较复杂, 在实际计算中参数选择需要花费大量计算时间进行搜索

我们在模型中使用的基于信息流的方法中, 如果选取更为合适的熵函数, 一定可以使它更加符合实际情况; 在三种方法综合的时候, 所取的权值也是可以采用更为有效的方法选取, 如应用层次分析法; 还可以选取其他分类方法加入　这些都是本模型可以改进的地方.

参考文献:

[1]　姜启源　数学模型(第二版).　高等教育出版社, 1992.
[2]　刘郁强等 . 序列空间方法 . 广东科技出版社, 1996.
[3]　刘祖洞 . 遗传学(第二版 . 高等教育出版社, 1991.
[4]　姜　丹, 钱玉美 . 信息理论与编码 . 中国科学技术大学出版社, 1992.
[5]　王玲玲等 . 常用统计方法 . 华东师范大学出版社, 1994.
[6]　陆　璇 . 应用统计 . 清华大学出版社, 1999.

The Classified Model for DNA Sequences

TANG Shi-jie,　ZHOU Liang,　WANG Xiao-ling

(University of Science and Technology of China, Hefei　230026)

Abstract:　Classifying the DNA sequences is a practice problem in biology.　In this paper, a mathematics model is established for the classifying of DNA sequences　Since there are both locality and globality in the DNA sequences, we discuss the criterion about whether the classified method is good or not.　That is whether the method bases on all properties that the DNA sequences have.

So a classified method with a single standard is not enough for the problem.　Here is a synthesis method on three different classified ways　The three ways base on varied property that DNA sequences have.　The first is the frequency of occurrences of the element in the DNA

sequences. The second is the periodic property of the DNA sequences. The third is that amount of information of the sequences. By using this method, we classify the nature sequences and artifical sequences. At last, we analyze the characteristic in this model and consider the generalization of this model.

关于DNA 序列分类问题的模型

冯 涛, 康喆雯, 韩小军
指导老师: 贺明峰

(大连理工大学, 大连 116024)

编者按: 本文以统计方法提取样本特征, 以之作为BP 神经网络的输入, 用MATLAB 中相应算法进行训练. 然后用于解决本分类问题, 得到了较准确的结果. 本文提取特征时考虑较为全面, 在此基础上正确地运用了神经网络方法, 发挥了神经网络适用于非线性问题、具有自适应能力的优点. 思路清楚, 文字简练.

摘要: 本文提出了一种将人工神经元网络用于DNA 分类的方法. 作者首先应用概率统计的方法对 20 个已知类别的人工DNA 序列进行特征提取, 形成DNA 序列的特征向量, 并将之作为样本输入BP 神经网络进行学习. 作者应用了MATLAB 软件包中的Neural Network Toolbox (神经网络工具箱) 中的反向传播(Back propagation BP) 算法来训练神经网络. 在本文中, 作者构造了两个三层BP 神经网络, 将提取的DNA 特征向量集作为样本分别输入这两个网络进行学习. 通过训练后, 将 20 个未分类的人工序列样本和 182 个自然序列样本提取特征形成特征向量并输入两个网络进行分类. 结果表明: 本文中提出的分类方法能够以很高的正确率和精度对DNA 序列进行分类, 将人工神经元网络用于DNA 序列分类是完全可行的.

1 问题重述(略)

DNA 序列由四个碱基A、T、C、G 按一定规律排列而成. 已知所给人工序列 1- 10 属于A 类, 11- 20 属于B 类. 本题中, 我们的主要工作有两个:

1) 提取A、B 两类特征;

2) 以所提取A、B 两类特征为依据, 把 20 个人工序列及 182 个自然序列分为A、B 两类(可能存在同时不具有A、B 两类特征, 不能归为A、B 中任一类的序列).

在本题中, 先以序列 1- 20 为依据, 提取出A、B 两类序列的统计特征, 然后运用神经网络中的BP 网络对未知序列进行了分类识别.

2 模型建立的理论依据

神经网络是近年来发展的一种大规模并行分布处理的非线性系统[1], 其主要特点有:

1) 能以任意精度逼近任意给定连续的非线性函数;

2) 对复杂不确定问题具有自适应和自学习能力;

3) 具有较强的容错能力和信息综合能力, 能同时处理定量和定性的信息, 能很好地协调多种输入信息的关系.

传统的分类识别方法, 对于一般非线性系统的识别很困难, 而神经网络却为此提供了一

第 31 卷第 1 期 数学的实践与认识 Vol. 31 No. 1
2001 年 1 月 MATHEMATICS IN PRACTICE AND THEORY Jan. 2001

DNA 序列分类的数学模型

吕金翅, 马小龙, 曹 芳
指导老师: 陶大程

(中国科学技术大学, 合肥 230026)

编者按: 本文能从生物学背景提出不同的三种判别模型 建模的分析和文字叙述条理清楚, 模型一对 21—40 和 182 样本均进行了分类, 分类正确率较高

摘要: 本文从三个不同的角度分别论述了如何对DNA 序列进行分类的问题, 依据这三个角度分别建立了三类模型

首先, 从生物学背景和几何对称观点出发, 建立了DNA 序列的三维空间曲线的表达形式 建立了初步数学模型- 积分模型, 并且通过模型函数计算得到了 1 到 20 号DNA 序列的分类结果, 发现与题目所给分类结果相同, 然后我们又对后 20 个DNA 序列进行了分类

然后, 从人工神经网络的角度出发, 得到了第二类数学模型- 人工神经网络模型 并且选择了三种适用于模式分类的基本网络, 即感知机模型, 多层感知机(BP 网络) 模型以及LVQ 矢量量化学习器, 同时就本问题提出了对BP 网络的改进(改进型多层感知机), 最后采用多种训练方案, 均得到了较理想的分类结果 同时也发现了通过人工神经网络的方法得到的分类结果与积分模型得到的分类结果是相同的(前四十个).

最后, 我们对碱基赋予几何意义:A. C. G. T 分别表示右 下 左 上 用DNA 序列控制平面上点的移动, 每个序列得到一个游动曲线, 提取游动方向趋势作为特征, 建立起了模型函数, 同时也得到了后二十个DNA 序列的分类结果, 而且发现结果与上述两个模型所得到的分类结果几乎相同(其中有一个不同, 在本模型中表示为不可分的). 此模型保留的信息量更多, 而且稳定性更强

1 问题的重述(略)

2 基本假设及模型建立:

第一类数学模型: 积分模型

DNA 序列是一种用 4 种字母符号(A、T、G、C) 表达的一维链 在这条链上不仅包含有制造人类全部蛋白质的信息(也就是基因), 还有按照特定的时空模式把这些蛋白质装配成生物体的四维调控信息(三维空间和一维时间), 找到这些信息的编码方式和调节规律是人类基因组研究的首要科学问题 下面我们首先将着手从几何学的角度来分析DNA 序列 鉴于自然界对称这一朴素原理, 我们的模型始于对 4 种碱基对称性的考察 图 1.1(略) 从纯化学的角度, 我们可以将碱基进行两类划分: (1) 按双环或单环结构, 可分为: 嘌呤碱基R (A 或G) 与嘧啶碱基Y (C 或T) (2) 按环中对应位置上是否存在氨基或酮基, 可分为: 氨基碱基M (A 或C) 与酮基碱基K (G 或T) 从生物学的角度, 在双螺旋结构中, 按碱基对形成氢键的数目或强弱, 碱基又可分: 强氢键碱基S (G 或C) 与弱氢键碱基W (A 或T), 这一种划分既包含了化学的也包含了DNA 双螺旋的结构信息在内

参照基本粒子理论中的做法, 我们利用三维 Euclid 空间中的对称几何图形——立方体 G 来表示碱基的上述三种对称性 如图 1.2 所示, 以 G 的中心为坐标原点建立三维直角坐