employ the BP (back propagation) algorithm to train NN by use of the Neural Network Toolbox in MATLAB software package. In this paper, two three-story NN are created to input the extracted DNA character vectors as samples into them. After the training, characters are extracted from the 20 unclassified artificial sequence samples and 182 natural sequence samples to form the character vectors as input of the two NN for clustering. The results shows: the clustering method presented in this paper can classify the DNA sequences in quite high accuracy and precision. It is quite feasible to apply the artificial neural network to DNA sequence clustering.

DNA 分 类 模 型

杨 健, 王 驰, 杨 勇
指导老师: 王 鸣

(北京大学, 北京 100871)

编者按: 本文将DNA 序列的碱基的组合看作"文章"的关键词, 用逐步优选法对关键词进行优选并用分层分类的方法进行分类. 从理论上说, 这一方法可以提取较好的特征, 而且分类也较精细. 这一模型有一定创造性, 分析问题比较精细而贴近实际, 思路清楚, 叙述通顺简练.

摘要: 本模型充分利用了所给数据的特点, 运用统计、最优化等数学方法, 从已知样本序列中提炼出能较好代表两类特征的关键字符串, 据此提出量化的分类标准, 能较好的对任给DNA 序列进行分类. 首先, 从已知样本序列中用广度优先法选出所有重复出现的字符串, 并计算其标准化频率及分散度. 然后, 利用样本数据结合最小二乘法确定两类字符串各自的优先级函数, 并且逐步优化其参数使之达到稳定, 提高了可信度. 最后, 根据优先级函数找出关键词, 然后确定权数, 用层次分析法对未知样本进行分类, 并定出显著水平, 从而得到了一个比较通用的分类方法. 经过检验, 此方法对 21—40 号待测样本进行了很好的分类, 对后面的 182 个DNA 序列进行同样的操作, 也有较好的效果.

1 问题的重述(略)

2 模型假设

(1) 假定待分类样本 21—40 中既不属于A 类也不属于B 类的样本百分比不超过 5%.
(2) 假设 keyword 的重要性与 t 和 s 有确定的关系, 且只与 t 和 s 有关(t, s 定义见下).
(3) 假设不代表A、B 类特征的字符串在DNA 序列中是均匀分布的.

3 模型的分析

从所给的DNA 序列观察发现, 很多字符串重复出现的频率很高, 而且有些字符串在A 类和B 类中出现的次数有很明显的差距, 这暗示把某些字符串作为A, B 两类的一个分类标准. 所以应对A、B 两类已知样本做统计分析, 找出其中可能代表该类特征的字符串. 因为每个字符串重要性可能不一样, 所以对这些字串的重要性排序, 选出最能代表该类特征的一部分字串. 然后用这些字串作为标准判断验证A, B 两类, 看所选的标准的准确性, 最后用于任何一个DNA 序列的分类.

4 定义与符号说明

A 类样本: 编号为 1—10 的DNA 序列

B 类样本: 编号为 11—20 的DNA 序列

词(word): 由 a, c, t, g 组成的在样本中重复至少两次的字符串

关键词(keyword): 能代表A 类或B 类的特征由 a , c, t, g 组成的词

分散度(s): 指某一类中包含某个word 的DNA 序列的个数

出现次数(l): 某一字符串在DNA 序列中的出现次数

序列长度(n):DNA 序列的长度

字符串长(m): 字符串的长度

标准化频率(t): $t = l\frac{4^m}{n}m$ 标准化了的词的出现次数

优先级函数(f): 衡量词重要性的指标, 是 s 和 t 的函数

权值(D): 衡量DNA 序列类别特征的量化指标

5 模型的建立与求解

(1) Keywords 的选择

选择 keyword 是所有工作的基础, 能否对基因序列进行有效的分类在很大程度上依赖 keyword 选择的好坏, 所以我们的原则是所选的 keyword 一定要能代表A 类和B 类基因序列的尽量全面的特征, 并且所选的 keyword 应对两类基因有好的区分度, 以利于分类

第一步: 选择 word(广度优先法). 以A 类样本为例:

① 计算字长为 1 的word 的出现次数, 并置 i= 1;

② 分别以字长为 i 的各结点为根结点 对每一个根结点, 若其出现次数大于 1, 在其后分别加入‘a’、‘c’、‘g’、‘t’, 组成长度为 i+ 1 的新字符串, 计算其出现次数及出现位置

例如: i= 1 时, 字长为 1 的word‘a’、‘c’、‘g’、‘t’可做根结点 对于‘g’, 在其后分别加入‘a’、‘c’、‘g’、‘t’, 生成‘ga’、‘gc’、‘gg’、‘gt’四个长度为 2 的新串

③ 若②中找到了新词, 则置 i= i+ 1, 转到②, 否则结束 程序流程图如图 1:

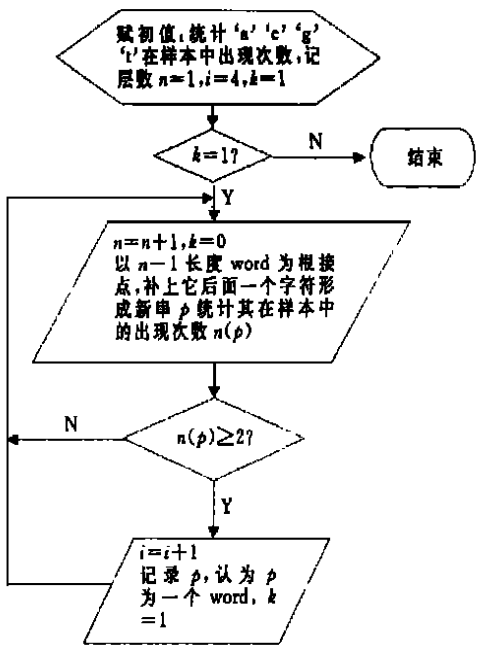

图 1

第二步: 选择 keywords(逐步优选法)

现在我们已经有了A、B 类的所有的word, 下面要在这些word 中选出好的keyword.

①选择典型代表

我们认为, 对于每一个word, 标准化频率 t 和分散度 s 越大, 它越具有代表性, 所以我们先在所有的word 中人工粗选出一些明显具有类别特征的word 作为初选关键词, 选择时,

不妨多选出一些, 尽量不要丢失信息, 也就是‘囫囵吞枣’, 我们在具体操作时, 对A、B 类各选出了 30 个左右.

②进行初步优选:

我们假定 keyword 的重要性只与 t 和 s 有关, 而且这种关系在一定的范围内有确定的解析表达式, 我们想找出这个函数关系 f.

显然, f 是 s 的增函数, 也是 t 的增函数, 所以假设 f 具有形式:

$$f(s, t) = s^{\alpha} t^{\beta} \tag{1}$$

其中 α 和 β 分别称为 s 和 t 的影响力因子.

对于某个word, 由(1)式求出的值称为它的 f 值.

接下来我们用上面选出的 30 多个word 结合最小二乘法来确定 α 和 β. 以A 类为例, 设A 类中粗选的word 分别有值:

$$t_1, t_2, \cdots, t_{39} \text{ 和 } s_1, s_2, \cdots, s_{39}$$

为了运用最小二乘法对(1)式取对数有:

$$\log f = \alpha \log s + \beta \log t \tag{2}$$

我们期望好的 keyword 的 f 值都较大, 而且稳定在某个下限之上, 所以把A 类的 39 个word 的 f 的期望值定为 50.0(这个值只具有相对意义, 具体定为多少对模型的优劣没有影响). 然后对这 39 个样本运用最小二乘法, 即: 求 α 和 β, 使得

$$g(\alpha, \beta) = \sum_{i=1}^{39} (\alpha \log s_i + \beta \log t_i - \log 50)^2 \tag{3}$$

达到最小.

利用MATLAB 提供的优化工具箱处理此最小二乘问题, 得到

$$\alpha_A = 0.4625 \qquad \beta_A = 0.8057$$

同样处理B 类的 26 个word 同样处理得到:

$$\alpha_B = 0.2217 \qquad \beta_B = 1.3060$$

这样, 我们就得到了A 类B 类的优先级函数的表达式:

$$f_A = s^{\alpha_A} t^{\beta_A} \tag{4}$$

$$f_B = s^{\alpha_B} t^{\beta_B} \tag{5}$$

这个函数只是一个初步的函数, 在下面的步骤中还会进一步优化函数的参数值.

将A 类B 类的所有word 的 s 和 t 值分别代入(4)(5)式, 得到每个word 对应的 f 值, 即优先级, 将每类的word 按优先级大小排序, 选择优先级最高的 16 个word 作为每类初步优化后的 keywords.

③ 对上面得到的 keywords 重复第②步的方法, 用最小二乘法求出每类新的 α 和 β 值, 这样就得到了进一步优化之后的优先级函数, 用此函数算出每个word 新的优先级, 再选择优先级最高的 16 个作为进一步优化之后的 keyword.

④重复几遍第③步, 直到 α 和 β 的值达到稳定. 此时的结果是:

$$\alpha_A = 0.4585 \quad \beta_A = 0.8851$$

$$\alpha_B = 0.2347 \quad \beta_B = 1.3769$$

至此, 选择 keyword 的工作结束, 而且得到的 keyword 是按优先级大小排序的.

由上面的工作步骤可以看出, 这 32 个 keyword 基本上代表了A 类B 类的特征, 而且A

类B 类的 keyword 基本没有重复, 有很好的区分度 Keywords 见表 1.

表 1 选择的关键字

A 组			B 组		
关键字	出现次数	分散度	关键字	出现次数	分散度
agga	32	9	taa	50	10
ga	109	10	ttta	52	10
ggcgg	28	7	aa	119	10
ggc	69	9	tat	148	10
ggaggc	11	8	at	131	10
aa	115	8	ttaa	34	10
agg	65	4	tatt	36	10
ggaa	34	10	ta	62	10
ggcgga	17	10	att	80	10
a	318	10	a	325	10
cgg	80	10	ttat	42	10
cgga	42	10	attt	44	10
ggagg	26	9	tta	105	10
gga	93	10	tttttt	58	10
gg	198	10	t	552	7
g	425	10	ttt	193	10

图 2 为每个word 的优先级对分散度、标准化频率的图象 其中 keyword 用圈标出 从图中看出, 用逐步优选法求出的 keyword 确实分别代表了A、B 类的重要特征

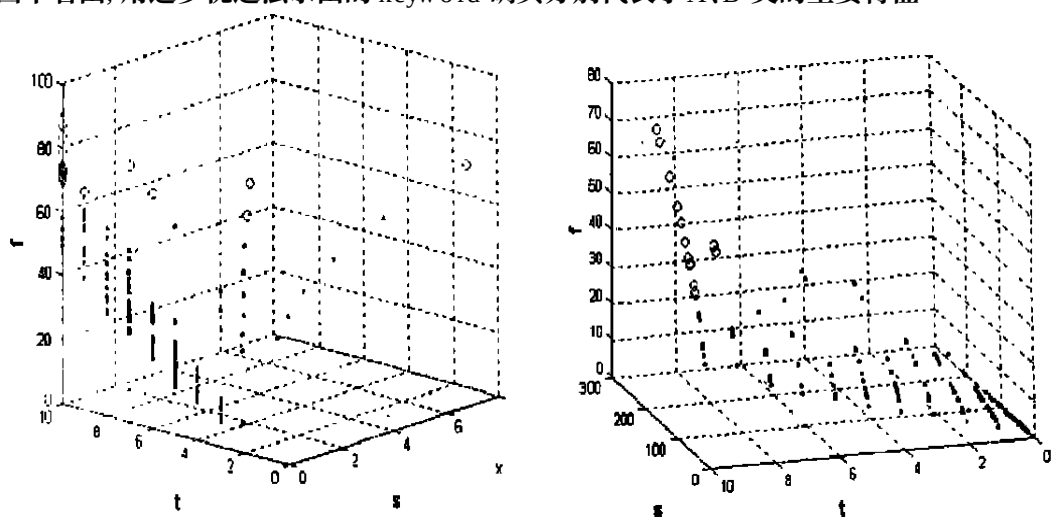

图 2

图 3 为逐步优选法的流程图 需要注意的是,word 的 f 值不仅是提取 keyword 的标准, 而且也反映了每个 keyword 的重要程度, 这对于以后的基因分类标准也具有重要的价值

(2) 加权系数的确定

现在来提取A 类和B 类基因序列的主要特征以便提出分类标准

对于某条基因序列 P , 我们假定它相对于每一类的权值 D 只与此类 keyword 在 P 中的出现次数 l 和 keyword 的优先级(即 f)有关, 我们希望找到这个关系

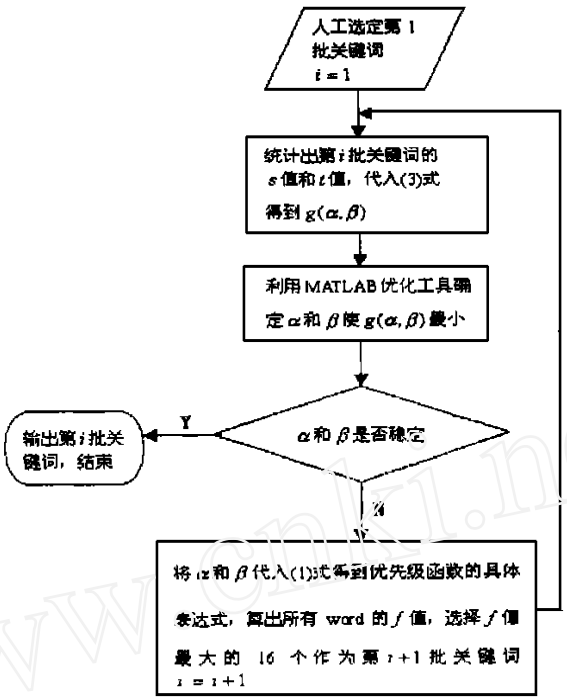

图 3

显然, 一个合理的假定是: P 的权值D 是 l 和 f 值的某种组合.

设某条基因序列 P 的权值具有表达式:

$$D_A(P) = \sum_{i=1}^{16} f_{A,i}^{\lambda_{A,i}} x_{A,i}(P) \tag{6}$$

$$D_B(P) = \sum_{i=1}^{16} f_{B,i}^{\lambda_{B,i}} x_{B,i}(P) \tag{7}$$

式(6)中, $f_{A,i}$ 是A 类第 i 个 keyword 的优先级函数值

$x_{A,i}$是序列 P 中具有A 类第 i 个 keyword 的个数

λ_A 称为**A 类 keyword 的影响力因子**

式(7)中的变量、参数同理理解

称$D_A(P)$ 为基因序列 P 的D_A 值, $D_B(P)$ 为 P 的D_B 值

下面来确定 λ_A 和 λ_B, 以 λ_A 为例

我们期望对于 10 个A 类样本, 它们的D_A 值都较大且稳定在某个定值(设为 d)周围, 另外, 为了提高分类效率我们把 λ_A 推广为 $\lambda_{A,k}$, 表示只与前 k 个 keyword 有关, 它的作用在第 3 点详细讨论　借用最小二乘法的思想, 令

$$h(\lambda_{A,k}) = \sum_{i=1}^{10} \left(\sum_{j=1}^{k} f_{A,j}^{\lambda_{A,j,k}}(P_i) - d \right)^2 \tag{8}$$

利用MATLAB 的最优化工具箱, 就可以确定 $\lambda_{A,k}$的值使 $h(\lambda_{A,k})$ 达到极小(与上面的分析相同, d 只具有相对意义, 它的选择对模型好坏没有影响).

3 层次分类法

现在要确定一个分类标准值 r, 当待分类样本 P 对某类的权值 > r 时就将 P 分为该类. 但通过上面确定 $\lambda_{A,k}$、$\lambda_{B,k}$的过程可以看出, 由A、B 类的已知样本确定 r 值会使标准过高, 所以用待分类的 20 个样本确定 r. 一个合理的假设是这 20 个样本中既不属于A 也不属于B 的样本数很少(≤5%). 用下面的程序经过几次实验即可得出较合适的 r 值.

另一个问题是选取多少个keyword. 选择越多的keyword 标准越严, 所以在算法中将 k 从 16 逐一递减到 7(k < 7 时 $\lambda_{A,k}$为负值, 标准无判断力), 逐层判别 P 是否属于A、B 类. 计算方法:

对于待分类样本 P, 先取定一个 r 值. 使用最严格的标准, 即 k = 16, 计算$D_A(P)$, 若大于 r, 则认为 P 属于A 类, 并记下此时的 k 值和$D_A(P)$, 否则逐步降低 k 值并与 r 比较, 直到 k = 7 为止. 若此时$D_A(P)$ 仍然小于 r, 则认为 P 不属于A. 同理可判断 P 是否属于B.

上述算法中 r 的确定:

在保证不漏掉一个待分类样本的条件下, 使被同时归为A, B 两类的样本数最小(不超过待分类样本总数的 5%).

据此标准确定的 r 值为 4.565(表 2).

表 2 λ_A、λ_B 的计算值

	1	2	3	4	5	6	7	8
λ_A	-3.5033	-1.7439	-1.1226	-0.9562	-0.6998	-0.2840	0.6556	0.7058
λ_B	-0.8553	-0.1686	-0.0716	0.0007	0.0301	0.0584	0.2300	0.2651
	9	10	11	12	13	14	15	16
λ_A	0.7953	0.9480	1.1425	1.1628	1.2812	1.3328	1.4940	1.5398
λ_B	0.3250	0.3353	0.3513	0.4042	0.4272	0.4915	0.5050	0.5220

程序流程图如图 4:

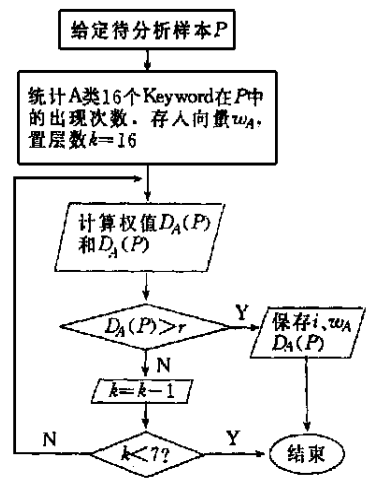

图4

4 方法总结

对于未知样本P, 用本模型对其分类的操作方法总结如下:

统计A、B 类的keyword 在P 中的出现次数 l_A, l_B; 用层次分类法对①中的数据进行分析, 求出权值D_A、D_B 及层次数 k_A、D_B.

比较这两组数据, 确定 P 属于哪一类. 还可以定义显著性水平(以A 为例):

a) 层数为 16 且 $k_A - k_B > 1$ 的定为高度显著属于A (***);

b) 层数为 15——14 且 $k_A - k_B > 1$ 或 $k_A = k_B$ 且$D_A - D_B > 1$ 的定为属于显著A (**);

c) 层数为 13——7 且 $k_A - k_B > 1$ 的定为较显著属于A (*);

d) 其余认为不显著;

若 P 对A、B 两类均不显著, 将其归入另类.

5　模型的检验及效果

① 对A、B 类已知样本: 最小二乘法保证A 类样本的D_A 值,B 类样本的D_B 值足够大, 这些样本必能正确归类; ② 对 20 组待测样本可以看出,A、B 类区分度很大

A 类 (11 个): 23　25　27　29　32　34　35　36　37　39; B 类 (9 个): 21　24　26　28　30　31　33　38　40

③ 对 182 组未知样本: (略)

可以看出有少部分样本不能很好归类　原因一是已知样本数量太少、长度太短, 无法有效检索到一些长度较长的 keyword; 二是模型有系统误差; 还可能有部分样本本来就不属于A 和B.

模型的稳定性: 分别把A ,B 两组中的序列任意改变一些字符, 再利用本文的方法进行分类, 经过多次实验, 新的分类结果均和原来相同, 说明本模型的算法有很高的稳定性

6　模型的改进

在选择 keyword 时, 数量上可以增加(本模型对每组选了 16 个 keyword), 这样更全面的概括A ,B 两组的特点, 使的分组更充分.

本模型是利用算术加权的方法综合个 keyword 的分类重要性, 可以探索其它的加权法, 如几何加权法, 对加权数的选择也可以进一步改进

在分类的过程中采用学习法: 如果某个DNA 序列用本文的结果判断, 发现对A 组(或B 组)的特点高度一致, 就可以把次序列列入A 组作为一个分类序列, 然后对新的A ,B 组重复本文的过程, 这样可以得到更好的分类法

7　模型的评价和推广

(1) 模型的优点

① 本模型中有较多的新想法和新的标准, 创建了一些比较独特使用的函数, 比如字符数字转化函数 yj m (见附录).

② 分类标准直观明了, 容易用计算机完成

③ 本模型的算法容易推广到实际的DNA 的序列分析中, 具有一定的实际应用价值

④ 关键词选择时考虑较全面(涉及到出现次数, 频率, 理论概率各个方面).

⑤ 对任何一个给定的DNA 序列采取多层分类法, 能够保证一定的置信度

⑥ 灵活的利用掩码‘m ’, 使得程序简洁, 编程量大大减少.

(2) 模型的不足之处

① 所给分类数据太有限, 导致采用 keyword 分类的方法, 判别与已知样本长度相差太大的未知序列时准确率有所限制

② 一些新的想法缺乏足够的理论根据, 所以有些问题的解决的带有一定的主观性

③ 本模型纯粹从数学的角度来分类, 缺乏生物学背景

(3) 模型的推广与应用

由于本模型主要是根据 20 个长度比较短的DNA 序列的特征归纳出的分类方法, 如果直接用本模型的结果用于实际的科研中, 可能会有很大的局限性, 但是本模型用 keyword 作为分类标准的思想是一个可以推广的比较好的想法, 具体的推广思路: 从自然DNA 序列

任意选出比较多的(为了保证较高的准确性), 利用 keyword 作为分类标准, 然后利用本文提供的加权系数的确定方法就可以定出一个具体的定量标准 具有一定实用价值

参考文献:

[1] 李 涛, 贺勇军等. MATLAB 工具箱应用指南——应用数学篇. 电子工业出版社.
[2] 袁亚湘. 最优化方法. 科学出版社.
[3] 张乃孝, 裘宗燕. 数据结构——c+ + 与面向对象的途径. 高教出版社.
[4] 汪仁官. 概率论引论. 北京大学出版社.
[5] 陈家鼎, 孙山泽等. 数理统计学讲义. 高教出版社.

The Grouping of DNA Sequences Model

YANG Jian, WANG Chi, YANG Yong

(Peking University, Beijing 100871)

Abstract: In this paper, a method to classify the DNA sequences is proposed. Mathematical methods such as statistics and optimization are used to build the model. The data is analysed sufficiently and the "critical words" is got, which can represent the characteristics of each group. According to this, a quantitative standard for grouping is brought forward. This model can properly classify the given data through testing. First, the strings which appear repeatedly (called words) in the given data are scanned out. The standard frequency and dispersion for each word are calculated. Second, using the Least Squares method, the priority function is fixed. Through stepwise optimization, the coefficients are made stable. Third, the key words are selected out and calculate the weight according to the priority function. At last, using the "analyse hierarchy process", the undetermined data is classified. This method can classify the undetermined data (No. 21—No. 40) fairly well, it can also give good result for the last 182 sequences.

DNA 序列的分类

韩铁平, 余 杭, 刘 威
指导老师: 杨启帆

(浙江大学, 杭州 310027)

编者按: 本文借助于计算机符号处理的能力来把握序列中不同碱基的丰度特征, 从而进行了利用数理统计方法的分类研究. 而后引入相关度分类判别算法及反馈机制来比较碱基的相对位置, 在既定方向上颇具新意地把工作推向深入. 不足之处在于, 未能使用相关度工具对各类样本分别进行分析; 此外, "纯数学"必须与其他学科紧密结合才会有优秀的建模工作, 本文虽然对编码氨基酸的三联体进行初步探讨, 着墨处自是轻淡许多.

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