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

标系, 使 G 的三组对面分别与三条坐标轴相垂直　分别与 X, Y, Z 轴相交的 G 的三组对面称为嘧啶/嘌呤面, 酮基/氨基面, 弱氢键/强氢键面　在 G 的六个面中各引一条对角线, 使相对面的对角线两两相互垂直, 如图 1.2 所示　在嘌呤面对角线的两端分别标上 A 和 G; 在嘧啶面对角线的两端分别标上 C 和 T, 如图 1.2 所示　显然, 此时上述碱基的三种对称关系全部自动成立　而且, 六条对角线刚好是正四面体 $ACGT$ 的六条棱

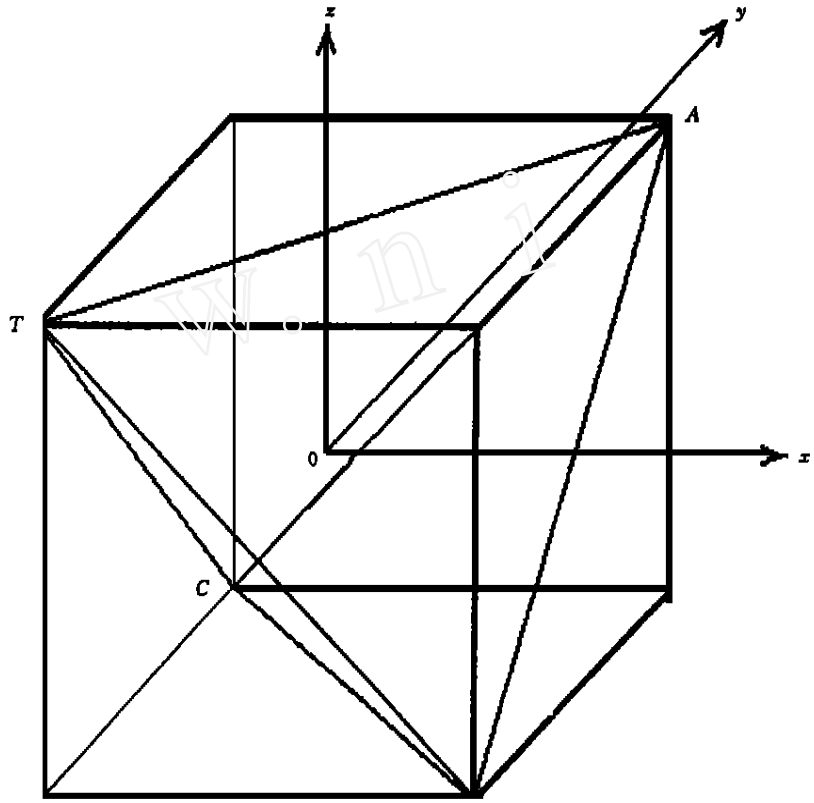

图 1.2　用立方体表示碱基的三种对称性

现在考察一个长为 L 的单链 DNA 序列, 阅读方向不限　从第一个碱基开始, 依次考察此序列, 每次只考察一个碱基　当考察到第 n 个碱基时 ($n = 1, 2, \cdots, L$), 统计一下从 1 到 n 这个子序列中四种碱基各自出现的次数, 并以 A_n、C_n、G_n、T_n 分别表示 4 种碱基 A、C、G、T 出现的次数, 如图 1.3 所示　显然它们都是非负整数　根据正四面体的对称性我们可以证明, 正四面体内存在唯一的一个点 P_n 与这四个非负整数一一对应　在图 1.3 所示建立的坐标系之下, 点 P_n 的坐标可用四个非负整数来表达

$$X_n = 2(A_n + G_n) - n,\ Y_n = 2(A_n + C_n) - n,\ Z_n = 2(A_n + T_n) - n,\ X_n, Y_n, Z_n \in [-n, n],\ n = 1, 2, \cdots, L;$$

其中 X_n, Y_n 和 Z_n 为点 P_n 的三个坐标分量　当 n 从 1 到 L 时, 我们依次得到 $P_1, P_2, \cdots, P_L$ 共 L 个点　将相邻两点用适当的曲线连接所得到的整条曲线, 就成为表示此 DNA 序列的 P-曲线　可以证明, P-曲线与所表示的 DNA 序列是一一对应的, 也就是说, 给定一定 DNA 序列, 存在唯一的一条 P-曲线与之对应; 反之, 给定一条 P-曲线, 可以找到唯一的一个 DNA 序列与之对应　换言之, P-曲线很大程度上包含了 DNA 序列的内蕴信息　P-曲线

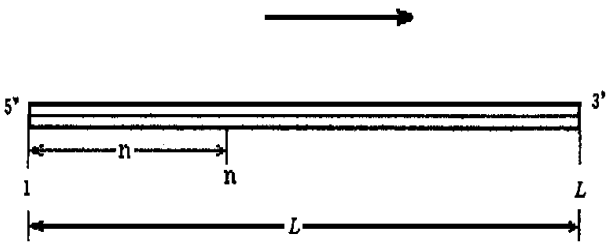

图 1. 3 DNA 序列示意图

是与符号DNA 序列等价的另一种几何表现形式 我们的核心想法就是通过对 P-曲线的研究来挖掘DNA 序列的内蕴信息

P-曲线的三个分量都具有明确的生物学意义: X_n 表示嘌呤/嘧啶碱基沿序列的分布 当从 1 到 n 这个子序列中嘌呤碱基多于嘧啶碱基时, $X_n > 0$; 否则 $X_n < 0$; 当两者相等时 $X_n = 0$. 同样, Y_n 表示氨基/酮基碱基沿序列的分布 当在子序列中氨基碱基多于酮基碱基时, $Y_n > 0$; 否则, $Y_n < 0$; 当两者相等时 $Y_n = 0$. Z_n 表示强/弱氢键碱基沿序列的分布 当弱氢键碱基多于强氢键碱基时, $Z_n > 0$; 否则, $Z_n < 0$; 当两者相等时 $Z_n = 0$. 由概率论中的结论: 如果任何一种分布均不能由其他两种分布的线性叠加表示出来, 则这三种分布是相互独立的给定的DNA 序列唯一的决定了这三种分布; 这三种分布唯一的描述了DNA 序列

我们对 P_n 的三个坐标分量分别积分, 发现 Y_n、Z_n 两个方向上并没有什么区别, 而在 X_n 方向上, A 组均大于零, B 组均小于零

$$f(t) = \int X_n(t)\,\mathrm{d}t$$

这表明在整个序列上不同结构的碱基对所占的成分, 即A 组嘌呤的含量较大, B 组嘧啶的含量较大

以"X 方向分量大于/小于零"为标准对给出的序列 21～ 40 进行分类, 得到如下结果:

A 类: 2, 3, 5, 7, 9, 14, 15, 17, 19; B 类: 1, 4, 6, 8, 10, 11, 12, 13, 16, 18, 20

第二类数学模型: 神经网络模型

由于神经网络具有运用已知认识新信息, 解决新问题, 学习新方法, 预见新趋势, 创造新思维的能力, 所以我们将神经网络处理问题的方法介入进来, 处理模式分类的问题

在本题中, 采用如下几种方案: 1. 单层感知机; 2. 双层感知机; 3. 改进型双层感知机

4. LVQ 矢量量化学习

对于每种算法我们又采用了三种统计方案, 即:

1. 统计 a c g t 在DNA 序列中出现的次数(共有 4 种)

2. 统计 a c g t 的两两组合在DNA 序列中出现的次数(共有 4^2 种不同的组合)

3. 统计 a c g t 的三三组合在DNA 序列中出现的次数(共有 4^3 种不同的组合)

所以总共可以得到 12 种模式分类模型

下面给出详细讨论, 但只列出 12 种方案中的四种, 因为剩下八种只是在统计方案上有所不同, 其训练实质和学习实质以及最后的模拟实质是相同的, 所以不需要一一罗列

第一方案(单层感知机)

1. 综述:

单层感知机是一个具有单层计算神经元的神经网络, 并由线形域值单元组成　原始的 Perceptron 算法只有一个输出节点, 它相当于单个神经元　当它用于两类模式的分类时, 相当于在高维样本空间中, 用一个超平面将两类样本分开.　F. Rosenblatt 也已证明, 如果两类模式是线形可分的(指存在一个超平面将它们分开), 则算法一定收敛　感知器特别适用于简单的模式分类问题, 也可用于基于模式分类的学习控制和多模态控制中.

2. 修正方案:

首先分析问题实质, 即采用一个单一神经元解决简单分类问题: 将 n 个输入矢量分为两类, 其中一部分为 1, 另一部分为 0.　最后确定网络结构(图 1.4):

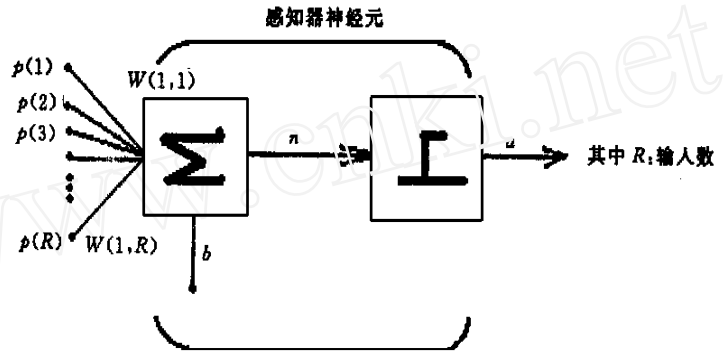

图 1.4

3. 训练算法: (采用单层感知机的经典算法, 这里略去)

判定网络收敛的标准有两种: 一是平均平方误差, 二是误差平方和　这里采用第二种.

学习结束后的网络将学习样本模式以连接权的形式分布记忆下来　当给网络提供一输入模式时, 网络将按上式计算出输出值 y_k, 并可根据 y_k 为 1 或 0 判断出这一输入模式属于记忆中的哪一种模式.

4. 训练和模拟结果:

a) 从 20 个已知结果的DNA 序列中随机选取不同的 4 个序列(向量)进行训练, 再对 20 个序列(向量)进行重新模拟, 其正确率为 90% , 发现出错的原因在于, 第 4 个和第 17 个序列在这几种统计方式下具有相似性.

b) 每次从 20 个已知结果的DNA 序列中随机选取不同的 4 个序列(向量)进行训练, 共进行两次, 再对 20 个序列(向量)进行重新模拟, 其正确率为 95% , 依然发现出错的原因在于, 第 4 个和第 17 个序列在这几种统计方式下具有相似性.

c) 每次从 20 个已知结果的DNA 序列中随机选取不同的 4 个序列(向量)进行训练, 共进行三次, 再对 20 个序列(向量)进行重新模拟, 其正确率为 95% , 依然发现出错的原因在于, 第 4 个和第 17 个序列在这几种统计方式下具有相似性.

5. 结论: 数据为线性不可分的, 所以单层网络不能实现完全识别.

6. 优缺点分析: 以上采用的是单个神经元的网络进行分类, 其优点是运算速度快, 但模式分类正确率较低.

第二方案(双层感知机, 即BP 网络)

1. 综述: BP 神经网络, 由于含有隐藏层, 所以可实现非线性分类　BP 算法属于 δ 算法,

是一种监督式的学习算法

2. 算法推导: (略)

3. 网络结构(图 1.5):

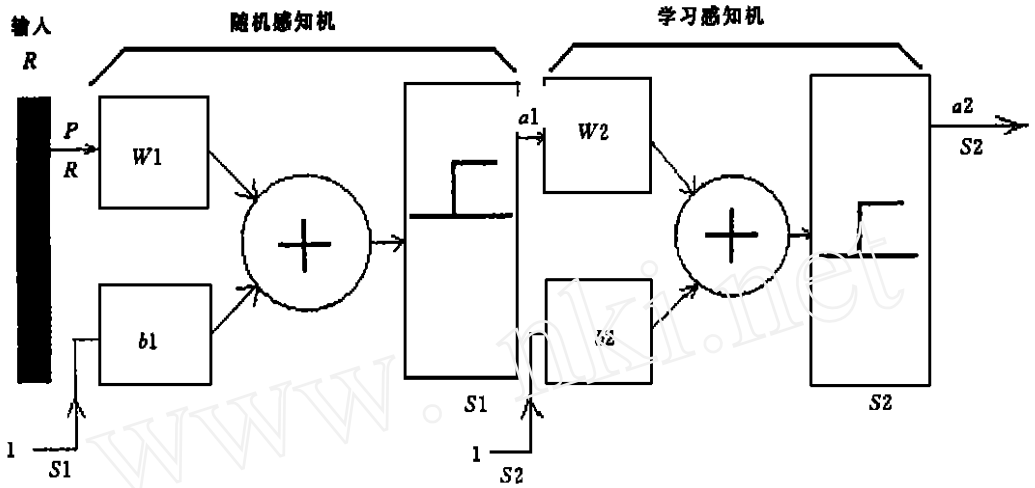

图 1.5

4. 训练算法: 由于其训练过程与学习过程相似, 所以这里不再赘述

5. 训练和模拟结果: 与第一方案相似, 只是分类正确率有所提高

7. 结论: 本题所给数据是线性不可分的, 而且通过简单的模式分类也很难行得通, 所以即使用多(双)层网络也难以实现完全识别

8. 优缺点分析: 以上采用的是多个神经元的带有一个隐藏层的网络进行分类, 其优点是运算速度较快, 且模式分类正确率较高, 但依然存在不可完全识别的问题

第三方案(改进型双层感知机)

1. 综述: 为了改进上述算法的不可完全识别的缺点, 现在对网络进行改进, 其目的是使网络可以对所有向量进行正确的分类

2. 改进的方案: 以提取更多的 1 分类信息为原网络结构与BP 神经网络相似, 但随机感知机层的响应函数采用 sigmoid 函数

3. 训练算法: 采用与BP 网络相同的训练算法

4. 训练和模拟结果: (分类正确率有所提高, 这里略去)

5. 结论: 数据是线性不可分的, 而且通过简单的模式分类也很难行的通, 所以只是简单改进网络结构, 是很难实现完全识别的. 所以下面将采用其它方法(LVQ 矢量量化学习)进行模式识别

6. 优缺点分析: 以上采用的是改进型多个神经元的带有一个隐藏层的网络(也就是改进型BP 神经网络)进行分类, 其优点是运算速度较快, 且模式分类正确率较高, 但依然存在不可完全识别的问题

第四方案(LVQ 学习向量量化)

1. 综述: 学习向量量化(LVQ)是在监督状态下对竞争层进行训练的一种学习算法. 竞争层将自动学习对输入向量进行分类, 这种分类的结果仅仅依赖于输入向量之间的距离. 如果两个输入向量之间特别相近, 竞争层就把他们分在同一类

2. 训练算法: (采用经典算法这里略去)

3. 训练和模拟结果: (分类正确率有所提高, 这里略去)

4. 要想从网络角度和学习算法上调整, 使得对已有的数据进行正确分类, 必须进行大规模学习, 但是如果对所有的样本进行训练再检策网络分类能力, 其可信服程度就大大降低了. 所以最后将采用改进网络输入的办法, 即结合生物学结论

5. 优缺点分析: 可靠性较高, 但算法复杂度较大

第五方案:

仅从神经网络结构上的角度考虑, 我们发现很难找到一个很好的网络, 所以将结合生物学重建神经网络

引用生物学的结论, 我们将输入模式变为 100 * 4, 其中 4 表示从 20 个已知样本中随机抽取 4 个样本. 100 表示(A + G) 含量的输入序列

采用BP 神经网络结构. 训练方案采用方案二中的误差逆传播算法

训练和模拟结果:

a) 从 20 个已知结果的DNA 序列中随机选取不同的 4 个向量进行训练, 再对 20 个向量进行重新模拟, 其正确率为 95% (较单层感知机有所改进, 但与BP 网络和LVQ 向量量化学习是相同的), 发现出错的原因是由于学习不充分造成的. 其本质是第 4 组数据和第 17 组数据可分性不好, 所以反应到网络上其可学习性又较大; 但如果学习不足, 则会导制误判, 所以应加大学习力度

b) 每次从 20 个已知结果的DNA 序列中随机选取不同的 4 个向量进行训练, 共进行两次, 在对 20 个向量进行重新模拟, 其正确率为 100%. 这次的结果充分说明了上述问题

结论: 目前的方法已很好的解决了分类的问题, 所以如果加大训练力度可以对其它数据进行正确率更高的分类. 我们对网络进行了 100 次随机抽取, 每次抽取的结果均进行训练, 最后对 40 个数据进行模拟, 发现前 20 个数的输出完全正确, 而且发现误差曲线也是十分好的, 所以有理由认为这个结论的正确性

模拟结果序列 21～40 为:

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

第三类数学模型: 二维随机游动模型

以四种碱基分别代表复平面上四个不同的方向, 顺序读取DNA 序列, 得到一条由原点出发的每次向相应方向移动单位长度的轨迹. 发现曲线明显地向两个相反的方向收敛(图 1.6) (略). 我们依此建立如下的数学模型:

设DNA 序列长为L, 记A_n, G_n, C_n, T_n 为 1 到 n 这个子序列中碱基A, G, C, T 所出现的次数, 令 P_n 为复平面上的点, 且

$$P_n = A_n + G_n i - T_n - C_n = (A_n - T_n) + i(G_n - C_n) = r_n e^{i\theta_n},$$

其中

$$r_n = \sqrt{(A_n - T_n)^2 + (G_n - C_n)^2}, \theta_n = Arg P_n, \bar{\theta} = \frac{1}{L}\sum_{k=1}^{L}\theta_k$$

假设 $n = 0$ 时, $A_n = G_n = C_n = T_n = 0$, 当 n 从 0 到 L 时, 在复平面上便得到了 $L + 1$ 个点, 并且得到了从原点出发的一条游动轨迹

鉴于幅角信息的突出重要地位, 我们依此对DNA 序列进行分类, 为了避免那种螺旋轨

迹我们假设DNA 序列可分类, 当且仅当$\exists\ p\in N$, s t 当$n>p$ 时 $\sum_{i=1}^{n}\theta_i$ 保持定号.

模型一: 对 20 个参数已知的DNA 序列, 分别求出其相应的游动方程 $P_n=(A_n-T_n)+i(G_n-C_n)$, 设 $\theta_{j,k}$ 为第 i 类第 j 个DNA 序列的A rgP_K

$$\bar{\theta}_j=\frac{1}{L}\sum_{k=1}^{L}\theta_{j,k},\quad j=1,2,\cdots,10,\quad i=1,2$$

在每一类中求出 $\theta_{\min}^{i}=\min_{1\le j\le 10}\theta_j^i$, $\theta_{\max}^{i}=\max_{1\le j\le 10}\theta_j^i$, 从而得到每个类的辐角特征区间$[\theta_{\min}^{i},\theta_{\max}^{i}]$. 如果$[\theta_{\min}^{1},\theta_{\max}^{1}]\cap[\theta_{\min}^{2},\theta_{\max}^{2}]=\varnothing$, 则对任意DNA 序列, 若可分类, 则满足 $\bar{\theta}\in[\theta_{\min}^{i},\theta_{\max}^{i}]$的属于第 i 类; 否则, 不可分类.

显然, 这时存在着不可分类的情形, 这主要是由于我们从DNA 序列样本中提取了两类游动在辐角上的趋势信息并将作为我们进行分类的标准. 这一点, 在模型二中得到了改进. 而实际上L 总有限, 前面关于可分类的假设是基于对游动辐角变化总体趋势的一种控制, 对于有限而言, 对此也有刻画即$\exists\ p\in N$ s t 当$n>p$, 辐角保持后续信息.

模型二、上面模型一提取了DNA 序列的最本质的辅角特征, 这里我们假设各类的DNA 序列的 $\bar{\theta}$ 在如下变换后满足正态分布. 首先辐角值可以与复平面中的圆周上的点建立自然的对应关系, 并且圆周挖去一点之后同胚于实直线, 为方便起见, 投影后的点仍用原来的字母表示, 从$\{\theta_j^i:1\le j\le 10\}$可得均值 μ_i 和方差 σ_i 及在第 i 类的概率密度函数为 $p_i(\theta)=\frac{1}{\sqrt{2\pi}\sigma_i}e^{-(\theta-\mu_i)^2}$.

任给一个DNA 序列, $\bar{\theta}$ 它属于第 i 类的概率:

$$P_i(\bar{\theta})=\lim_{\theta\to 0+}\frac{\int_{\bar{\theta}}^{\bar{\theta}+\theta}p_i(\theta)\,d\theta}{\int_{\bar{\theta}}^{\bar{\theta}+\theta}[p_1(\theta)+p_2(\theta)]\,d\theta}=\frac{p_i(\bar{\theta})}{p_1(\bar{\theta})+p_2(\bar{\theta})}$$

以概率 0.5 为阀值, 如 $p_i(\bar{\theta})>0.5$, 则属于第 i 类.

下面再用区间估计法给出结果在统计意义上的可信度, 设 n 个相互独立的样本 $X_i\sim N(a,\sigma^2)$, $i=1,2,\cdots,n$, 令 $Z=(X_1+X_2+\cdots+X_n)/n$, 则 $Y=(Z-a)/(\sigma^2/n)^{1/2}\sim N(0,1)$, 但 σ^2 未知, 必须先把它估计出来, 用 $Sn^2=[(X_1-Z)^2+(X_2-Z)^2+\cdots+(X_n-Z)^2]/(n-1)$ 代替 σ^2, $(Z-a)/(Sn^2/n)^{1/2}=(Z-a)(\sigma^2/n)^{-1/2}/(Sn^2/\sigma^2)^{1/2}=Y/(Sn^2/\sigma^2)^{1/2}$, 因 $Y\sim N(0,1)$, $(Sn^2/\sigma^2)^{1/2}=\{[(X_1-Z)/\sigma]^2+[(X_2-Z)/\sigma]^2+\cdots+[(X_n-Z)/\sigma]^2\}/(n-1)\sim\chi^2(n-1)$, 因而 $t=(Z-a)/(Sn^2/n)^{1/2}\sim t(n-1)$, 这里要求 Y 与 $(Sn^2/\sigma^2)^{1/2}$ 相互独立. 于是给定 α 后, 查表 $t(n-1)$ 可得 t^*, 使得 $Pr(|t|\le t^*)=1-\alpha$, 即 $Pr(|Z-a|/(Sn^2/n)^{1/2}\le t^*)=1-\alpha$, 从而我们便得到了 a 的 $1-\alpha$ 水平上的置信区间为$[Z-t^*Sn/n^{1/2},\ Z+t^*Sn/n^{1/2}]$. 现在共有 10 个已知样本点$X_1,X_2,\cdots,X_{10}$, 为了保证 Y 与 $(Sn^2/\sigma^2)^{1/2}$ 相互独立, 现将这 10 个样本点等分成两组这样便得到 $Z=(X_1+X_2+\cdots+X_5)/5$, $Z'=(X_6+X_7+\cdots+X_{10})/5$, $Y=(Z-a)/(\sigma^2/5)^{1/2}$, $S_5^2=[(X_6-Z')^2+(X_7-Z')^2+\cdots+(X_{10}-Z')^2]/(5-1)$, $t=(Z-a)/(S_5^2/5)^{1/2}$, 依前所述给定 α, 我们可得 a 的 $1-\alpha$ 水平上的置信区间为$[Z-t^*S_5/5^{1/2},\ Z+t^*S_5/5^{1/2}]$.

由该模型可以看出曲线的趋向正代表着序列中所含对应元素的整体含量和分布. 当基因序列中所含的非特征随机信息较多时, 会导致游动曲线螺旋摇摆情形, 从而导致前进距离

变短, 但是由随机信号在各方向上的平均性, 总体前进方向并未受到影响, 故我们只提取方向而忽略距离作为特征信息

我们从不同角度, 提取序列整体上和局部之间的特征, 建立了以上三种数学模型　三种模型各有优劣, 但他们在特征提取, 模式识别和分类上的都具有一定的普适性和优越性

参考文献:

[1] 郝柏林, 刘寄星　理论物理与生命科学　上海科学技术出版社

[2] 金冬燕, 金　奇, 侯云德　核酸和蛋白质的化学合成与序列分析　科学出版社

The Mathematical Models on the Classification of The DNA Sequences

LU Jin-chi,　MA Xiao-long,　CAO Fang

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

Abstract: This paper deals with the problem of how to classify the DNA sequences from three different angles and accordingly establishes three kinds of models.

Firstly, on the point of biological background and geometrical symmetry, we established a descriptive model of 3-dimensional space curve on the DNA sequence, by which we got a rudimentary mathematical model-Calculus model. Through the integration of the model function, we have acquired the classification results of the DNA sequences from 1 to 20, and found them identical to the classification results given by the problem. Then we classified the latter 20 DNA sequences.

Then, on the view of the artificial neural networks, a second model - The Artificial neural networks model was established. We chosen three kinds of basic networks, which well fit into the classification at last. And by the same time, we proposed the improvement of the BP network, and finally procured comparatively ideal classification results by various training programmes. also, we found the results identical to what we have got by Calculus model.

By the end, we endowed A, C, G, T with geometrical meaning: A indicates right, while C as down, G as up, T as left. We got a mobile curve from each sequence with the points of the plain moving according to the controlling of the DNA sequence. By following the feature of the moving direction, the model function was established. By the way we acquired the classification results of the latter 20 DNA sequences and found them practically identical to the results of the two above models (One of results differently showed in this model is regarded as indivisible). This model contains more information, and is more stable.