第九届“华为杯”全国研究生数学建模竞赛

题 目　　基因识别问题及其算法实现

摘　　要

基因识别问题是当前生物信息学的一个最基础的问题，目前采用信号处理与分析方法来发现基因编码序列受到了广泛重视。本文主要基于 DNA 序列的信噪比特征来进行基因识别。

首先，我们分别简化了 Voss 映射、Z-curve 映射以及实数映射下对应的功率谱和信噪比的计算公式，证明了 Z-curve 映射与 Voss 映射下对应的功率谱和信噪比相差非零常数倍。同时，我们还对三种映射的信噪比曲线进行了分析比较，并利用实例验证了碱基的 3-周期现象与映射的选取无关。

其次，我们先通过一些统计数据，强调了不同物种类型应当有不同的基因阈值。同时，我们提出了基于位置判定和基于距离平方的两种最优化方法，并给出了一系列的评价指标，对代表性生物的阈值判定结果作了分析比较。

然后，我们给出了基于 DNA 序列信噪比特征的基因识别算法，强调了这种算法的合理性。同时，我们还通过实例验证了算法的有效性，并作了相应误差分析。另外，我们还标出了附件中的六组基因数据外显子的具体位置。

最后，我们探讨了位于内含子和外显子之间的剪切位点的识别问题。根据剪切位点区域的生物特性，我们引入了统计学中的隐马尔可夫过程，对剪切位点的供点区域建立相应模型，得到隐马尔可夫模型的重估参数，用以识别 DNA 序列中的剪切位点。

关键词：信噪比；3-周期性；阈值；基因识别；剪切位点

一、问题重述

DNA 是生物体遗传信息载体，其化学名称为脱氧核糖核酸。DNA 是一种长链聚合物，由腺嘌呤(A),鸟嘌呤(G),胸嘧啶(C),胸腺嘧啶(T)按一定顺序组成。其中带有遗传讯息的 DNA 片段称为基因。基因通常被划分为许多间隔的片段，其中编码蛋白质的部分，即编码序列片段，称为外显子，不编码的部分称为内含子。随着人类基因组计划的实施和顺利完成，基因预测成为生物信息学中最基础，也是最首要的问题。对基因预测中，由于用统计预测方法在未知基因的准确率明显下降，采用信号处理与分析方法来发现基因编码序列也受到广泛重视。基于此，题目中特别介绍数字序列映射、DNA 序列信号 3 周期特性的信噪比（SNR）概念和基因识别的两种方法，在这两种方法中，其中一种是一是固定长度窗口滑动法，另一是移动信噪比曲线识别法。最后提出以下 4 个问题：

1.功率谱与信噪比的快速算法:

(1) 对 Voss 映射，探求功率谱与信噪比的某种快速计算方法；

(2) Z-curve 映射和 Voss 映射下的频谱与信噪比之间的关系；

(3) 对实数映射给出功率谱与信噪比的快速计算公式；

2.对不同物种类型基因的阈值确定

(1) 研究其阈值确定方法和阈值结果;

(2) 按照频谱或信噪比特征将编码与非编码区间分类的有效性，以及分类识别时所产生的分类错误作适当分析;

3. 基因识别算法的实现

对所设计的基因识别算法的准确率做出适当评估，并将算法用于对附件中给出的 6 个未被注释的 DNA 序列的编码区域的预测;

4. 延展性研究

对你们自己认为有价值的其它相关问题展开探讨

二、问题假设

1. 本文中用于验证分析的所有 DNA 序列都是正常序列，未发生基因突变；
2. 本文中用于验证分析的所有 DNA 序列中没有无用的序列；
3. 本文中用于验证分析的所有 DNA 序列中若外显子和内含子的位置已知，则它们的位置完全无误；

4. DNA 序列中的外显子和内含子是紧邻的，它们之间不含有其它序列

三、问题分析

针对问题 1，关键在于如何根据离散 Fourier 变换以及 DNA 指示序列的性质得到不同映射下功率谱和信噪比的快速计算方法。对于不同的映射，它们所反映的 DNA 的总体信息也不同，对映射之间的优劣比较有助于我们更好地理解 DNA 信息。

针对问题 2，由于物种的多样性，不同物种的基因结构存在显著差异，反映在基因阈值上也应当有所差异。合理地确定某一物种的基因阈值有助于基因的识别分类。

针对问题 3，由于 DNA 序列随机噪声的影响等原因，目前基因识别方面的多数算法结果还不是很充分，我们的目的在于根据 DNA 序列的信噪比特征进行基因识别，并采用改进算法来增加预测准确度。

针对问题 4，我们选取了位于外显子和内含子之间的剪切位点进行识别分析。剪切位点在基因转录中起着关键作用，通过对前体 RNA 进行剪接去除内含子是大多数真核基因表达的关键步骤，因此, 对剪切位点的精确预测直接关系到外显子和基因突变点的定位。

四、问题的求解分析

4.1 功率谱与信噪比的快速算法

在 DNA 序列研究中，采用信号处理与分析方法来发现基因编码序列受到广泛的欢迎[1,2]。这种方法首先需要把 A、T、G、C 四种核苷酸的符号序列，根据一定的规则映射成相应的数值序列，以便于对其作数字处理。本节首先针对 Voss 映射、Z-curve 映射和实数映射等三种常见的映射，分别给出对应的功率谱与信噪比快速算法。同时，我们通过实例对这三种映射进行了分析比较。

4.1.1 Voss 映射

令 $I=\{A,T,G,C\}$，长度为 N 的任意 DNA 序列，可表达为

$$S=\{S[n]\,|\,S[n]\in I, n=0,1,2,\cdots N-1\},$$

即 A、T、G、C 的符号序列 $S:S[0],S[1],\cdots,S[N-1]$。现对于任意确定的 $b\in I$，令：

$$u_b[n]=\begin{cases}1, & S[n]=b\\ 0, & S[n]\neq b\end{cases}, n=0,1,2,\cdots N-1,$$

称之为 Voss 映射，于是生成相应的 0-1 序列

$$\{u_b[n]\}:u_b[0],u_b[1],\cdots,u_b[N-1](b\in I)。$$

为研究 DNA 编码序列（外显子）的特性，对指示序列分别做离散 Fourier 变换(DFT)

$$U_b[k]=\sum_{k=0}^{N-1}u_b[n]e^{-j\frac{2\pi nk}{N}},\ b\in I\ ,k=0,1,\cdots,N-1$$

以此可得到四个长度均为 N 的复数序列 $\{U_b[k]\},b\in I$ 。计算每个复序列 $\{U_b[k]\}$ 的平方功率谱，并相加则得到整个 DNA 序列 S 的功率谱序列 $\{P[k]\}$，其中

$$P[k]=|U_A[k]|^2+|U_T[k]|^2+|U_G[k]|^2+|U_C[k]|^2,\ k=0,1,\cdots N-1。$$

经相关研究证实，外显子序列的功率谱曲线在频率 $k=N/3$ 处，有较大的频谱，这种现象称为碱基的 3-周期。

记 DNA 序列 S 的总功率谱的平均值为

$$\bar{E}=\frac{\sum_{k=0}^{N-1}P[k]}{N},$$

而将 DNA 序列在特定位置，即 $k=N/3$ 的功率谱值，与整个序列 S 的总功率谱的平均值的比率称为 DNA 序列的“信噪比”，即

$$R=\frac{P[\frac{N}{3}]}{\bar{E}}。$$

为了简化 DFT 运算，我们首先给出以下几个命题[3,4,5]。

命题 1：在 DNA 序列 $\{S[n],n=0,1,2,\cdots N-1\}$ 中，设其长度 N 为 3 的倍数。将核苷酸符号 $b\in I=\{A,T,G,C\}$ 出现在序列 $\{0,3,6,\cdots,3k,\cdots\}$，$\{1,4,7,\cdots,3k+1,\cdots\}$ 以及 $\{2,5,8,\cdots,3k+2,\cdots\}$ 等位置上的频数分别记为 x_b,y_b 和 z_b，则在 $N/3$ 处有

$$\left|U_b[\frac{N}{3}]\right|^2=(x_b,y_b,z_b)\begin{pmatrix}1 & -1/2 & -1/2\\ -1/2 & 1 & -1/2\\ -1/2 & -1/2 & 1\end{pmatrix}\begin{pmatrix}x_b\\ y_b\\ z_b\end{pmatrix}=X_b^TMX_b\,,$$

证明：DNA 序列在 $N/3$ 的功率谱峰值为

$$\begin{aligned}\left|U_b[\frac{N}{3}]\right|^2&=\left|\sum_{n=0}^{N-1}u_b[n]\cdot e^{-j\frac{2\pi n\cdot\frac{N}{3}}{N}}\right|^2=\left|\sum_{n=0}^{N-1}u_b[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2\\&=\left|x_b+y_b\cdot e^{-j\frac{2\pi}{3}}+z_b\cdot e^{j\frac{2\pi}{3}}\right|^2\\&=\left(x_b-\frac{1}{2}(y_b+z_b)\right)^2+\frac{3}{4}(z_b-y_b)^2\\&=\left(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b\right)\\&=(x_b,y_b,z_b)\begin{pmatrix}1 & -1/2 & -1/2\\ -1/2 & 1 & -1/2\\ -1/2 & -1/2 & 1\end{pmatrix}\begin{pmatrix}x_b\\ y_b\\ z_b\end{pmatrix}\\&=X_b^TMX_b\end{aligned}$$

命题 2：我们假定 DNA 序列长度为，*A*、*T*、*G*、*C* 四种核苷酸在 DNA 出现的次数为 N_A， N_T， N_G， N_c。可以得到

$$|U_b|^2=\sum_{k=0}^{N-1}|U_b[k]|^2=N\cdot N_b\text{。}$$

证明：由指示序列分别做离散 *Fourier* 变换(*DFT*)可得

$$U_b[k]=\sum_{n=0}^{N-1}u_b[n]e^{-j\frac{2\pi nk}{N}},\quad k=0,1,\cdots,N-1\text{。}$$

再由 *Fourier* 逆变换可以得到

$$u_b[n]=\frac{1}{N}\sum_{n=0}^{N-1}U_b[k]e^{j\frac{2\pi nk}{N}},n=0,1,\ldots,N-1\,,$$

根据帕斯瓦尔定理：

$$\sum_{n=0}^{N-1}|u_b[n]|^2=\frac{1}{N}\sum_{k=0}^{N-1}|U_b[k]|^2,$$

显然，等式左边等于 N_b，因此可得

$$|U_b|^2=\sum_{k=0}^{N-1}|U_b[k]|^2=N\cdot N_b。$$

综合命题 1 与命题 2，我们得到 Voss 映射下 DNA 序列功率谱和信噪比的快速算法。

命题 3：设 DNA 序列长度为 N 为 3 的倍数，则有

$$P[\frac{N}{3}]=\sum_{b\in I}\left|U_b[\frac{N}{3}]\right|=\sum_{b\in I}X_b^TMX_b$$

DNA 序列的信噪比可简化为

$$R=\frac{\sum_{b\in I}X_b^TMX_b}{N}。$$

证明：直接由命题 2，

$$E=\sum_{b\in I}|U_b|^2=N\cdot N_A+N\cdot N_T+N\cdot N_G+N\cdot N_C=N^2。$$

那么 DNA 序列的总功率谱的平均值为

$$\overline{E}=\frac{E}{N}=\frac{N^2}{N}=N,$$

结合命题 1，得到

$$R=\frac{P[\frac{N}{3}]}{\overline{E}}=\frac{\sum_{b\in I}X_b^TMX_b}{N}。$$

4.1.2 Z-curve 映射：

设 DNA 序列的四个指示序列 $\{u_b[n]\}$，$b\in I=\{A,C,G,T\}$，$n=0,1,2,\cdots N-1$ 的积累序列 b_n ($n=0,1\ldots N-1$) 为 $b_n=\sum_{i=0}^{n-1}u_b[i]$。则定义三个序列 $x[n],y[n],z[n]$：

$$\begin{cases}x[n]=2(A_n+G_n)-n\\ y[n]=2(A_n+C_n)-n\\ z[n]=2(A_n+T_n)-n\end{cases}$$

若令

$\Delta x[n]=x[n]-x[n-1]$， $\Delta y[n]=y[n]-y[n-1]$， $\Delta z[n]=z[n]-z[n-1]$，

于是我们得到 Z-curve 映射

$$\begin{pmatrix}\Delta x[n]\\ \Delta y[n]\\ \Delta z[n]\end{pmatrix}=\begin{pmatrix}1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix}\begin{pmatrix}u_A[n]\\ u_C[n]\\ u_G[n]\\ u_T[n]\end{pmatrix}$$

令

$$F=\begin{pmatrix}1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix}=\begin{pmatrix}\alpha_1^T\\ \alpha_2^T\\ \alpha_3^T\end{pmatrix}=\begin{pmatrix}\beta_1 & \beta_2 & \beta_3 & \beta_4\end{pmatrix},$$

显然，我们得到

$$\left\langle \beta_i,\beta_j\right\rangle \equiv -1,\forall i,j\left(1\le i,j\le 4,i\ne j\right),$$

$$\left\|\beta_i\right\|^2\equiv 3,\forall i\left(1\le i\le 4\right),$$

其中$\langle\square,\square\rangle$为向量的内积。

命题 4：设 DNA 序列长度为 N 为 3 的倍数，在 Z-curve 映射下，成立

$$P_Z[\frac{N}{3}]=4P_Z[\frac{N}{3}],$$

并且 DNA 序列的信噪比为 $R_Z=\frac{4}{3}R$。

证明：因为 $\Delta x[n]=\alpha_1^T\cdot\left(u_A[n],u_C[n],u_G[n],u_T[n]\right)^T$，从而有

$$\begin{aligned}\left|\Delta X\left(\frac{N}{3}\right)\right|^2&=\left|\sum_{n=0}^{N-1}\Delta x[n]\cdot e^{-j\frac{2\pi n\cdot\frac{N}{3}}{N}}\right|^2=\left|\sum_{n=0}^{N-1}\Delta x[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2\\&=\left|\sum_{j=1}^{4}\alpha_{1j}\left(\sum_{n=0}^{N-1}u_j[n]\cdot e^{-j\cdot\frac{2\pi n}{3}}\right)\right|^2\\&=\left|\sum_{j=1}^{4}\alpha_{1j}\cdot\left(x_j+y_j\cdot e^{-j\frac{2\pi}{3}}+z_j\cdot e^{j\frac{2\pi}{3}}\right)\right|^2\\&=\left|\left(\sum_{j=1}^{4}\alpha_{1j}x_j\right)+\left(\sum_{j=1}^{4}\alpha_{1j}y_j\right)e^{-j\frac{2\pi}{3}}+\left(\sum_{j=1}^{4}\alpha_{1j}x_j\right)e^{j\frac{2\pi}{3}}\right|^2\\&=\alpha_1^TX^TMX\alpha_1\end{aligned}$$

同理可求

$$\left|\Delta Y\left(\frac{N}{3}\right)\right|^2=\alpha_2^T X^T MX\alpha_2,$$

$$\left|\Delta Z\left(\frac{N}{3}\right)\right|^2=\alpha_3^T X^T MX\alpha_3。$$

在$k=N/3$时，

$$\begin{aligned}P_Z[\frac{N}{3}]&=\left|\Delta X[\frac{N}{3}]\right|^2+\left|\Delta Y[\frac{N}{3}]\right|^2+\left|\Delta Z[\frac{N}{3}]\right|^2\\&=\sum_{i=1}^{3}\alpha_i^T X^T MX\alpha_i\end{aligned}$$

令$\alpha=(1,1,1,1)^T$，并定义

$$X^T MX=\begin{pmatrix}X_1^T\\X_2^T\\X_3^T\\X_4^T\end{pmatrix}M\left(X_1,X_2,X_3,X_4\right)=\begin{pmatrix}X_1^T MX_1 & \cdots & X_1^T MX_4\\ \vdots & \ddots & \vdots\\ X_1^T MX_1 & \cdots & X_1^T MX_4\end{pmatrix}\square H,$$

因为

$$X_1+X_2+X_3+X_4=\begin{pmatrix}x_A+x_C+x_G+x_T\\y_A+y_C+y_G+y_T\\z_A+z_C+z_G+z_T\end{pmatrix}=\begin{pmatrix}\frac{N}{3} & \frac{N}{3} & \frac{N}{3}\end{pmatrix}^T,$$

并且

$$\begin{aligned}MX\alpha&=M\left(X_1,X_2,X_3,X_4\right)(1,1,1,1)^T=M\left(X_1+X_2+X_3+X_4\right)\\&=\begin{pmatrix}1 & -1/2 & -1/2\\-1/2 & 1 & -1/2\\-1/2 & -1/2 & 1\end{pmatrix}\begin{pmatrix}N/3\\N/3\\N/3\end{pmatrix}=0\end{aligned},$$

容易得到

$$\begin{aligned}\alpha^T H\alpha=\alpha^T X^T MX\alpha&=(1,1,1,1)H(1,1,1,1)^T=\sum_{i,j=1}^{4}X_i^T MX_j\\&=\sum_{i=1}^{4}X_i^T MX_i+2\sum_{i<j}X_i^T MX_j=0\end{aligned},$$

DNA 序列在$k=N/3$的功率谱值

$$P_z[\frac{N}{3}]=\alpha_1 H\alpha_1^T+\alpha_2 H\alpha_2^T+\alpha_3 H\alpha_3^T=3\left(\sum_{i=1}^{4}X_i^T MX_i\right)-2\sum_{i<j}X_i^T MX_j。$$

事实上，由于

$$\alpha_{1i}^2+\alpha_{2i}^2+\alpha_{3i}^2=\|\beta_i\|^2=3,$$

当$i,j=1,2,3,4$并且$i\neq j$，则$P_Z[N/3]$可简化为

$$\begin{aligned}P_Z[\frac{N}{3}]&=\alpha_1 H\alpha_1^T+\alpha_2 H\alpha_2^T+\alpha_3 H\alpha_3^T\\&=4\left(\sum_{i=1}^{4}X_i^T MX_i\right)-\left[\left(\sum_{i=1}^{4}X_i^T MX_i\right)+2\sum_{i<j}X_i^T MX_j\right],\\&=4P_Z[\frac{N}{3}]\end{aligned}$$

由于

$$P_Z[k]=\left|\Delta X[k]\right|^2+\left|\Delta Y[k]\right|^2+\left|\Delta Z[k]\right|^2,$$

根据帕斯瓦尔定理：

$$\sum_{n=0}^{N-1}\left|\Delta x[n]\right|^2=\frac{1}{N}\sum_{k=0}^{N-1}\left|\Delta X[k]\right|^2,$$

记 DNA 序列四个碱基A,C,G,T出现的次数分别为N_A,N_C,N_G,N_T，则有

$$\begin{aligned}\sum_{k=0}^{N-1}\left|\Delta X[k]\right|^2&=N\sum_{n=0}^{N-1}\left|\Delta x[n]\right|^2\\&=N\cdot\sum_{n=0}^{N-1}\left|\alpha_1^T\cdot(u_A[n],u_C[n],u_G[n],u_T[n])^T\right|^2,\\&=N\cdot\left(\alpha_{11}^2N_A+\alpha_{12}^2N_C+\alpha_{13}^2N_G+\alpha_{14}^2N_T\right)\end{aligned}$$

类似地，我们可得

$$\sum_{k=0}^{N-1}\left|Y[k]\right|^2=N\cdot\left(\alpha_{21}^2N_A+\alpha_{22}^2N_C+\alpha_{23}^2N_G+\alpha_{24}^2N_T\right),$$

$$\sum_{k=0}^{N-1}\left|Z[k]\right|^2=N\cdot\left(\alpha_{31}^2N_A+\alpha_{32}^2N_C+\alpha_{33}^2N_G+\alpha_{34}^2N_T\right),$$

把上面三个式子相加，得

$$E_Z=3N(N_A+N_C+N_G+N_T)=3N^2,$$

因此有

$$R_Z=\frac{P_Z(N/3)}{E_Z/N}=\frac{4P_Z[N/3]}{3N}=\frac{4}{3}R。$$

4.1.3 实数映射

令 $I=\{A,T,G,C\}$，长度为 N 的任意 DNA 序列可表达为

$$S=\{S[n]\,|\,S[n]\in I,n=0,1,2,\cdots N-1\}\,,$$

对 DNA 序列进行实数映射

$$A\to 0,C\to 1,G\to 2,T\to 3\,,$$

例如，假设给定的一段 *DNA* 序列片段为 *S = ATCGTACTG*，则所生成的序列分别为：

Voss 映射序列：

$\{u_A[n]\}$：$\{1,0,0,0,0,1,0,0,0\}$；　$\{u_G[n]\}$：$\{0,0,0,1,0,0,0,0,1\}$；

$\{u_C[n]\}$：$\{0,0,1,0,0,0,1,0,0\}$；　$\{u_T[n]\}$：$\{0,1,0,0,1,0,0,1,0\}$。

实数映射序列：

$\{u_A'[n]\}$：$\{0,0,0,0,0,0,0,0,0\}$；　$\{u_G'[n]\}$：$\{0,0,0,2,0,0,0,0,2\}$；

$\{u_C'[n]\}$：$\{0,0,1,0,0,0,1,0,0\}$；　$\{u_T'[n]\}$：$\{0,3,0,0,3,0,0,3,0\}$。

从例子得出以下关系式：

$$U_A'[k]=0\,,\ U_C'[k]=U_C[k]\,,$$

$$U_T'[k]=3U_T[k]\,,\ U_G'[k]=2U_G[k]\,。$$

实数映射的功率谱为

$$\begin{aligned}P_R[k]&=\left|U_A'[k]\right|^2+\left|U_T'[k]\right|^2+\left|U_G'[k]\right|^2+\left|U_C'[k]\right|^2\\&=\left|3U_T[k]\right|^2+\left|2U_G[k]\right|^2+\left|U_C[k]\right|^2\\&=9\left|U_T[k]\right|^2+4\left|U_G[k]\right|^2+\left|U_C[k]\right|^2\end{aligned}\,,$$

实数映射的总功率谱为

$$\begin{aligned}\sum_{k=0}^{N-1}P_R[k]&=\sum_{k=0}^{N-1}\left(9\left|U_T[k]\right|^2+4\left|U_G[k]\right|^2+\left|U_C[k]\right|^2\right)\,,\\&=9N\cdot N_T+4N\cdot N_G+N\cdot N_C\end{aligned}$$

因此，得到实数映射的总功率谱平均值为

$$\overline{E_R} = \frac{\sum_{k=0}^{N-1} P_R[k]}{N} = 9N_T + 4N_G + N_C \text{。}$$

再由命题 1，

$$\begin{aligned} P_R[\frac{N}{3}] &= \left|U'_A[\frac{N}{3}]\right|^2 + \left|U'_T[\frac{N}{3}]\right|^2 + \left|U'_G[\frac{N}{3}]\right|^2 + \left|U'_C[\frac{N}{3}]\right|^2 \\ &= \left|3U_T[\frac{N}{3}]\right|^2 + \left|2U_G[\frac{N}{3}]\right|^2 + \left|U_C[\frac{N}{3}]\right|^2 \quad , \\ &= 9\left|U_T[\frac{N}{3}]\right|^2 + 4\left|U_G[\frac{N}{3}]\right|^2 + \left|U_C[\frac{N}{3}]\right|^2 \\ &= 9X_T^T MX_T + 4X_G^T MX_G + X_C^T MX_C \end{aligned}$$

从而得到实数映射条件下的信噪比为

$$R_R = \frac{9X_T^T MX_T + 4X_G^T MX_G + X_C^T MX_C}{9N_T + 4N_G + N_C} \text{。}$$

4.1.4 几种映射的验证与比较

为了验证 Voss 映射下的快速计算公式，我们选取了家鼠的一段 DNA 序列（AF077860，来自于 NCBI 数据库[8]），先后用 DFT 算法和快速计算公式计算了该序列的信噪比，如图 1 所示。结果显示，两种方法得到的信噪比完全一样。

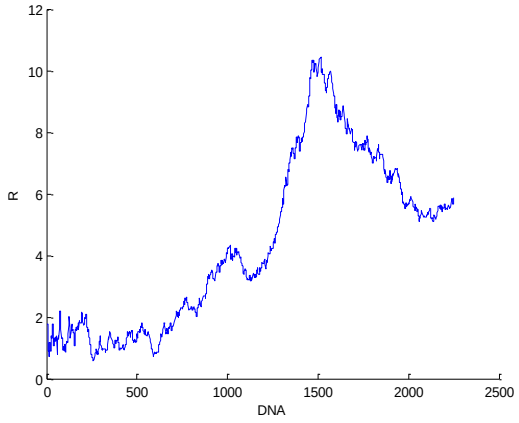

图 1A：DFT 算法（DNA: AF077860）

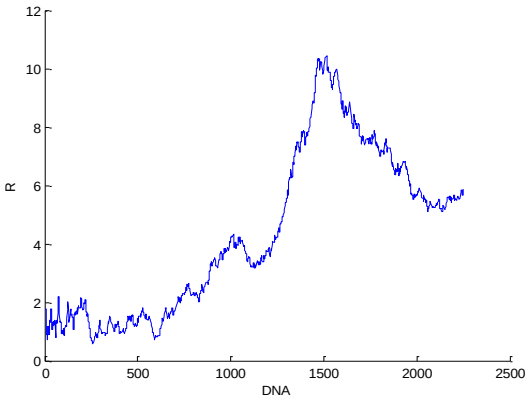

图 1B：快速计算公式（DNA: AF077860）

为了比较 Voss 映射、Z-curve 映射和实数映射，我们分别绘制了三种映射下的信噪比曲线，并计算了一些特殊点的信噪比值，如图 2 所示和表 1 所示。

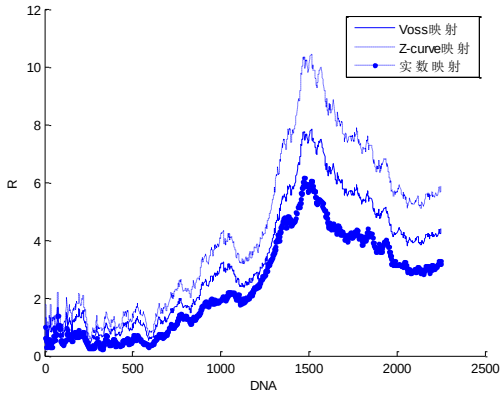

图 2 三种映射下信噪比曲线比较（DNA: AF077860）

位置	450	900	1350	1800	2250
R	1.1467	2.4444	5.5156	5.3367	4.4116
R_Z	1.5289	3.2593	7.3541	7.1156	5.8821
R_R	0.5580	1.7338	4.6322	3.9857	3.2670
R_Z/R	1.3333	1.3333	1.3333	1.3333	1.3333
R/R_R	2.0550	1.4098	1.1907	1.3390	1.3504

表 1 特殊点三种映射信噪比值比较

由图 2 和表 1 我们可以得出：

- 对于同一个 DNA 序列，由 Voss 映射和 Z-curve 映射得到的信噪比图像峰形完全一致，而实数映射信噪比图像与前两个映射图像峰形走势大致相同；
- Voss 映射的信噪比曲线介于 Z-curve 映射曲线与实数映射曲线之间，实数映射的波形变化较为平缓，Z-curve 映射与 Voss 映射的波形变化相同。

由三种映射信噪比曲线的比较，我们得到一些启发，即：给定一个 DNA 序列，对于不同的映射，外显子的功率谱是否仍然具有 3-周期性？我们仍然对家鼠的 DNA 序列（AF077860）中同一个外显子和内含子进行比较。

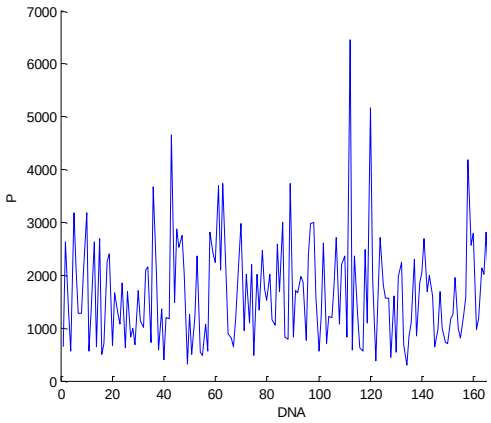

图 3A：Voss 映射下外显子功率谱线

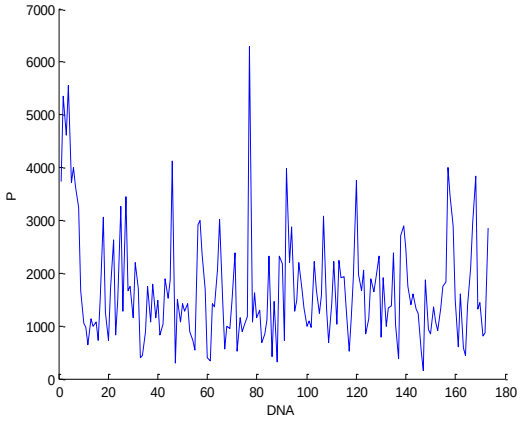

图 3B：Voss 映射下内含子功率谱线

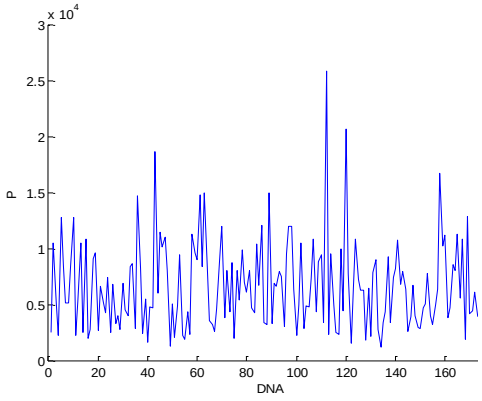

图 4A：Z-curve 映射下外显子功率谱

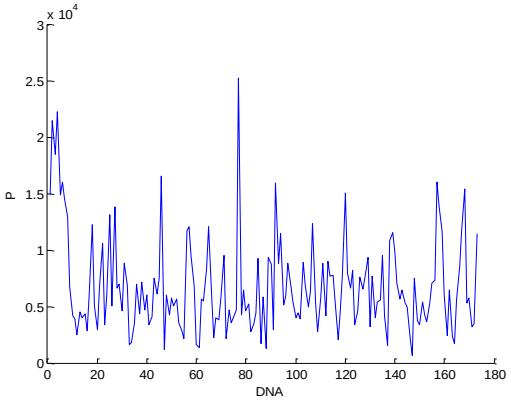

图 4B：Z-curve 映射下内含子功率谱

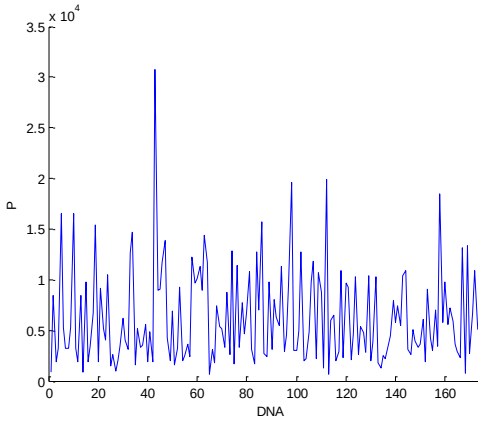

图 5A：实数映射下外显子功率谱

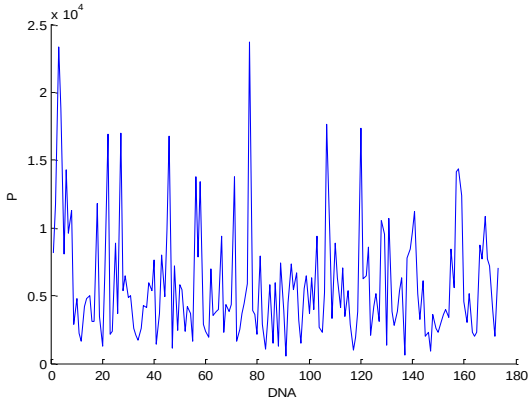

图 5B：实数映射下内含子功率谱

以上功率谱图表明，对这三种映射，内含子均不存在 3-周期性，而外显子均存在 3-周期性。因此，DNA 序列的 3-周期性与这三种映射的选取是无关的。同时，我们认为，上述关于三种映射的结论不仅仅对于家鼠成立，对于所有物种都应该成立。在做基因识别时，原则上这三种映射方法都可以采用，但是由

于实数映射的波形变化较为平缓，在基因识别中出现误差的可能性比较大，因此，在基因识别中采用 Z-curve 映射或 Voss 映射有助于减少识别误差。而由前一节的公式推导，Z-curve 映射与 Voss 映射的功率谱和信噪比均成正比例关系，即相差非零常数倍，我们认为这两种映射在基因识别中区别不大，在下文中，我们均采用 Voss 映射来进行阈值确定和基因识别。

4.2 不同物种类型的基因阈值确定

4.2.1 不同物种类型基因的外显子和内含子信噪比分析

一般现有的文献通常将基因阈值取为 2，然而由于不同物种的基因具有其独特性，如果基因阈值都取为 2 则不具有科学性。为了说明这一点，我们选取了人、家鼠、黑腹果蝇、拟南芥以及酵母等 4 种不同的生物 DNA 序列外显子和内含子（来自于 NCBI 数据库），对它们的内含子和外显子信噪比进行了统计分析，如表 2 所示。

基因种类	外显子			内含子		
	数量	*R* 均值	*R* 标准差	数量	*R* 均值	*R* 标准差
人	632	2.3479	2.8701	531	0.7263	0.4279
家鼠	430	2.7034	3.3293	378	0.9649	1.2547
黑腹果蝇	416	1.9625	5.2058	428	0.8537	1.0385
拟南芥	452	1.8738	1.9610	396	0.7945	0.9923
酵母	474	6.8256	5.2346	563	0.6984	1.4879

表 2 外显子和内含子信噪比统计分析

由表 2 我们得知：不同种类基因的信噪比有着较大的差距，特别是外显子的信噪比的标准差远远大于内含子的标准差。因此，对于不同种类的基因，我们需要寻找其特定的基因阈值。

4.2.2 基因阈值的确定方法

4.2.2.1 基于位置判定的最优化方法

对于已有的基因数据，不妨设有 n 个外显子，m 个内含子。通过 Voss 映射，我们计算出所有外显子和内含子的信噪比值。记

$S_1 = \{R_i^1\}_{i=1,2,\cdots,n}$，其中 R_i^1 为外显子信噪比值，

$S_2 = \{R_j^2\}_{j=1,2,\cdots,m}$，其中 R_i^1 为内含子信噪比值，

我们引入符号函数 $\mathrm{sgn}(x)$，即

$$\mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases},$$

并且记

$$N_1 = \sum_{i=1}^{n} \mathrm{sgn}(R_i^1 - R_0),$$

$$N_2 = \sum_{j=1}^{m} \mathrm{sgn}(R_0 - R_j^2),$$

则 N_1， N_2 分别代表外显子信噪比值大于 R_0 的个数以及内含子信噪比值小于于 R_0 的个数。自然地，我们希望 $N_1 + N_2$ 最大，因此求解阈值的优化模型为

$$\max\left(\sum_{i=1}^{n} \mathrm{sgn}(R_i^1 - R_0) + \sum_{j=1}^{m} \mathrm{sgn}(R_0 - R_j^2) \right),$$

其中 $\min\{R_j^2\}_{j=1}^{m} \leq R_0 \leq \max\{R_i^1\}_{i=1}^{n}$。

4.2.2.2 基于距离平方的最优化方法

不妨设某个生物的基因阈值为 R_0，则第 i 个外显子信噪比 R_i^1 到阈值 R_0 的距离平方为

$$d_i^1 = (R_{i`}^{1`} - R_0)^2,$$

所有外显子信噪比值到阈值的距离平方和为

$$d_1 = \sum_{i=1}^{n} (R_i^1 - R_0)^2,$$

从而得到平均距离平方

$$\bar{d}_1 = \frac{d_1}{n} = \frac{1}{n} \sum_{i=1}^{n} (R_i^1 - R_0)^2 。$$

类似地，所有内含子信噪比值到阈值的平均距离平方为

$$\bar{d}_2 = \frac{d_2}{m} = \frac{1}{m}\sum_{j=1}^{m}(R_j^2 - R_0)^2 \text{ 。}$$

我们要使 $\bar{d}_1$ 和 $\bar{d}_2$ 都达到最小值，这样就得到基于距离平方的优化模型

$$\min\left(\frac{1}{n}\sum_{i=1}^{n}(R_i^1 - R_0)^2 + \frac{1}{m}\sum_{j=1}^{m}(R_j^2 - R_0)^2\right),$$

其中 $\min\{R_j^2\}_{j=1}^{m} \le R_0 \le \max\{R_i^1\}_{i=1}^{n}$。

4.2.3 阈值评价指标与结果分析

我们首先给出阈值判别效果的评价指标。

设基因阈值为 R_0，当一段序列的信噪比 $R \ge R_0$ 时，判断该序列为内含子；反之，则判断为内含子。令 T_P 为正确判为外显子的个数，T_N 表示被正确判为内含子的个数，F_N 表示被错误地判为内含子的个数，F_P 表示被错误地判为外显子的个数，定义敏感度 $S_n = T_P/(T_P + T_N)$，专一度 $S_p = T_N/(T_N + F_P)$，阈值判别的总正确率 $A_c = (S_n + S_p)/2$。

运用上面提出的阈值确定最优化方法，我们分别对五种生物基因的阈值进行了测定，并计算了阈值判断的正确率，同时也对以 2 为阈值进行了比较，见表 3。

基因种类	距离平方最优化		位置判定最优化		以 2 为阈值	
	R_0	A_c	R_0	A_c	R_0	A_c
人类	1.1853	0.7762	1.0545	0.7854	2	0.7124
家鼠	1.6107	0.8273	1.4222	0.8097	2	0.7203
黑腹果蝇	1.4143	0.7826	1.3468	0.7284	2	0.6811
拟南芥	1.5304	0.7463	1.4147	0.7827	2	0.6628
酵母	1.4834	0.8475	1.5628	0.8826	2	0.7029

表 3 阈值判断结果及其正确率统计

由上表的数据，我们可以得到以下结论：

- 阈值为 2 时判别正确率明显低于前两种方法，前两种优化方法的判断正确率可以达到 80%以上，而取阈值为 2 时的正确率只有 70%左右。
- 对于不同的基因种类，前两种优化方法正确率总体较高，并且没有明显差异，最大也只相差 0.06%左右，并且预测出来的阈值差异并不大。这说明两种预测方法基本上差不多。

综上所述，采用两种最优化方法都可以有效地提高阈值判定准确率，但从运算求解复杂度考虑，基于位置判定的最优化方法更容易一些。

下面我们考虑影响阈值判定正确率的因素。我们认为，影响判定正确率的主要因素有以下几个方面：

- 在对外显子和内含子信噪比统计分析中，我们发现部分外显子的信噪比并不显著，同时，也有一些信噪比较大的内含子。这些例外的存在直接导致了判定结果出现误差。
- 考虑到我们外显子和内含子的样本有可能来自于同一个物种不同类型的基因，基因种类的不同也有可能导致判定结果出现误差。
- 另外我们注意到，在外显子和内含子的样本中，每个序列的长度一般都不相同，有时长度差异还非常大。例如，有的外显子序列的长度达到上千甚至上万个，而有的外显子长度则只有几十个。这种长度的差异可能会影响阈值的判定。

4.3 基于信噪比特征的基因识别算法

4.3.1 外显子和内含子信噪比曲线分析

由第一节我们知道，外显子的功率谱通常具有 3-周期性，而内含子则不具有。现在，我们希望探究外显子和内含子的信噪比曲线具有什么样的特征，并且期望通过这种特征提出基于信噪比特征的基因识别算法。

为此，我们各挑选了家鼠的一个外显子序列和一个内含子的序列，分别绘出了它们的信噪比曲线。

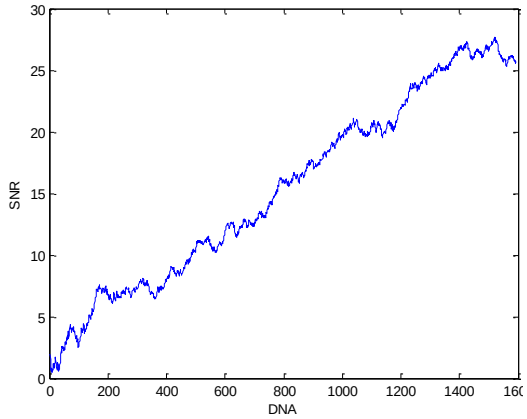

图 6 外显子信噪比曲线

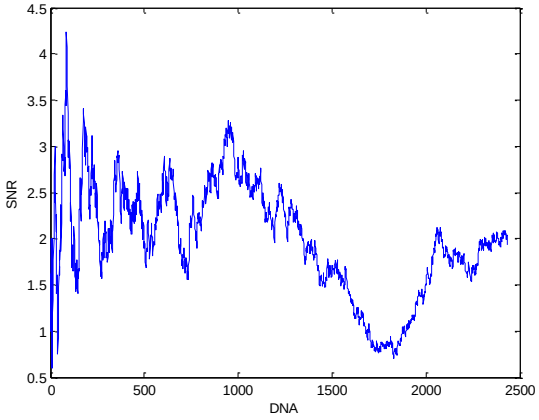

图 7 内含子信噪比曲线

由图 6 和图 7 我们可以看到，外显子的信噪比曲线呈上升趋势，而内含子的信噪比曲线则无明显规律。基于这种性质，我们给出基因识别的一种模糊法则：

- 如果 DNA 序列的信噪比曲线某一段呈上升趋势，那么这一段对应的 DNA 序列就有可能是外显子区域；如果成下降趋势或无规则的波动，则有可能是内含子区域。

注 实际上我们在进行区域分类时还要考虑阈值，而不仅仅是曲线的走势。这在我们下面提出的基因识别算法中有所体现。

4.3.2 基于信噪比特征的基因识别算法

通过查阅文献，我们得知，一般而言外显子和内含子区域的核苷酸数目都大于 50。利用这个性质，可以提高我们区域分类的准确度。对于一个长度为 N 的 DNA 序列，令 S_k 为该序列中从第一个位置到第 k 个位置的子序列，为了判断该序列第 k 个位置的核苷酸究竟属于外显子还是内含子，我们采用以下算法：

1. 令 $k=1$；
2. 计算从 0 到 k 处四种碱基出现的频数，具体地说，即第一节中定义的 $X_b=(x_b,y_b,z_b)^T$；
3. 计算 $k/3$ 处的功率谱值，即

$$P[\frac{k}{3}]=\sum_{b\in I}X_b^T M X_b\ ,$$

以及信噪比，即

$$R(k)=\frac{P(k/3)}{k}=\frac{\sum_{b\in I}X_b^T M X_b}{k}。$$

4、让 k 增加 1，重复步骤 2 和 3，直到 $k=N$；

5、从 $k=51$ 起，计算每一个位置对应的斜率，即

$$SL(k)=\frac{R(k)-R(k-50)}{50}；$$

6、判断每一个点究竟是属于外显子还是内含子，用如下判断法则：如果该点的斜率大于 0 并且对应的 $R(k)\geq 2$，就断定它属于外显子；否则，属于内含子。

7、经验法修正局部误差。如果 DNA 序列的一个内含子区域的核苷酸数目少于 50，并且它两端紧邻外显子，则将此区域修正为外显子；反之，如果一个外显子区域的核苷酸数目少于 50，并且它两端紧邻内含子，则将此区域修正为内含子。

8、分段法修正局部误差。将 DNA 序列平均分成 m 个片段，将这些片段的起点分别记为，从 P_2 到 P_{m+1} 重复步骤 1 到步骤 7，再从 P_3 到 P_{m+1} 重复步骤 1 到步骤 7，……，从 P_m 到 P_{m+1} 重复步骤 1 到步骤 7。

下面是该算法的流程图：

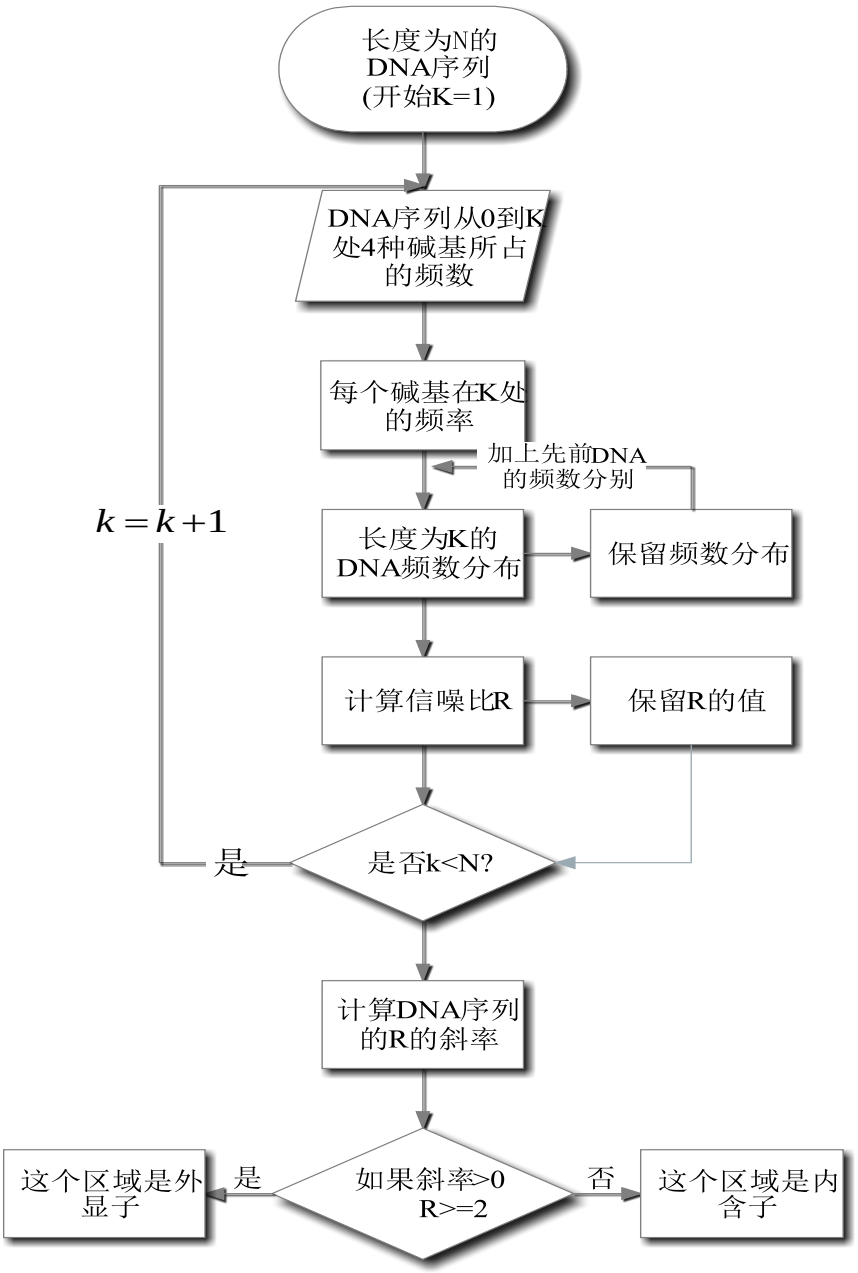

图 8 算法流程图

4.3.3 算法验证与误差分析

为了验证前面提出的算法，我们选取了家鼠的一段 DNA 序列（AF077860，来自于 NCBI 数据库），取其信噪比阈值为 1.4222（这是表 3 中基于位置判定最优化得到的阈值结果）。

通过算法的步骤 1 到步骤 4，我们绘制了信噪比曲线，如下图所示

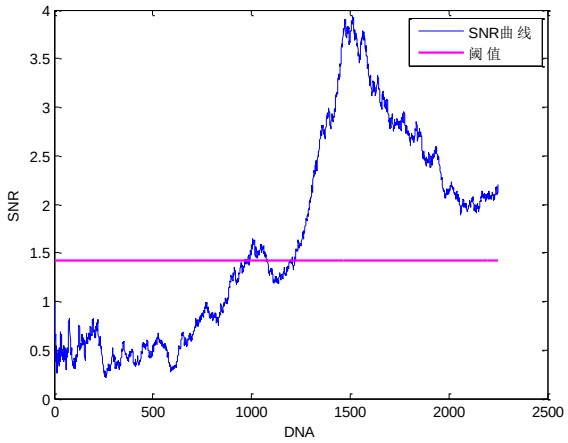

图 9　DNA 信噪比曲线

通过算法步骤 7，即利用经验法修正局部误差后，我们得到了图 11 所示的外显子分布图，图中上面的线段代表正确的外显子分布，下面的线段代表预测的外显子分布。在此基础上，通过算法步骤 8，即利用分段法修正局部误差后，我们得到了图 12 所示的外显子分布图。

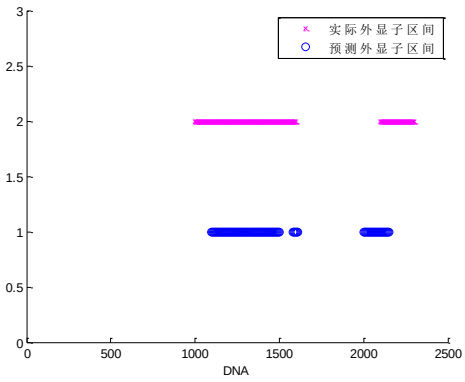

图 10 经验法修正后外显子分布图（图中上面线段代表实际外显子区间，下面线段代表预测外显子区间）

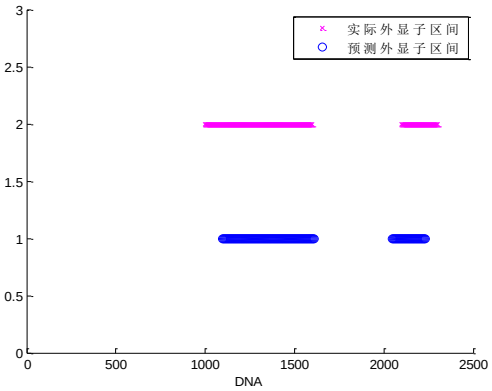

图 11 分段法修正局部误差(图中上面线段代表实际外显子区间，下面线段代表预测外显子区间）

比较上面两图，我们可以得知：局部误差的修正明显改进了预测的准确度。具体地说，通过计算，图 11 对应的 $A_c = 0.7263$，而图 12 对应的 $A_c = 0.8571$，预测准确度提高了将近 13%。因此，通过以上分析，我们的模型是有效的并且具有较高的精确度。

下面我们对算法误差做分析。从算法本身来说，首先我们做了一个经验假设，即假定外显子和内含子区域的核苷酸数目都大于 50，但是在实际中情况并不总是这样，例如黑腹果蝇的就存在长度小于 50 的外显子和内含子序列。其次，

从图 7、图 8 中内含子和外显子信噪比曲线中可以直接看出，存在某些区域，内含子的曲线成上升趋势，而外显子的曲线呈下降趋势，这就有可能导致误判。最后，由第二节的讨论，我们注意到不同生物类型有可能有不同的基因阈值，而基因阈值是否准确判定将直接影响到基因识别。

4.3.4 基因编码区域预测

利用上面的算法，我们对附件中的六组基因数据进行了预测，表 4 列出了外显子在 DNA 序列中的具体位置，图 10 到图 15 则分别是六个基因组的信噪比曲线。

基因数据	外显子位置
第一组	[1528,1614]　[1724,1812]　[1870,1945] [2831,2935]　[3800,3873]　[4802, 4870]
第二组	[1838,1909]　[1988,2042]　[3489,2571]
第三组	[55,105]　[283,345]　[1107,1165]　[1683,1763]
第四组	[1846,2085]　[3545,3611]　[5159,5264]
第五组	[5998,6092]　[10573,13663]
第六组	[1343,1401]

表 4 基因数据的外显子位置

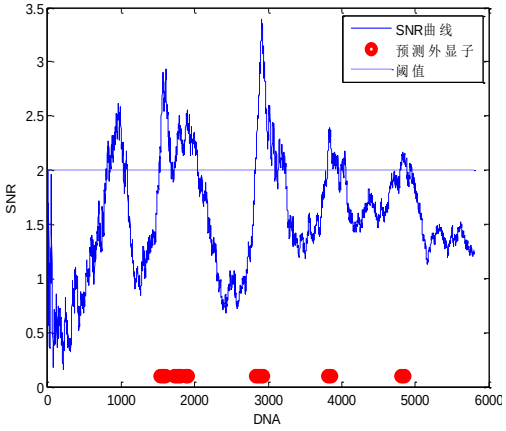

图 12 第一组数据外显子

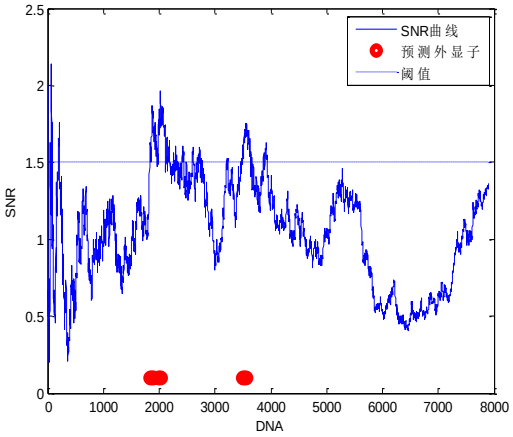

图 13 第二组数据外显子预测

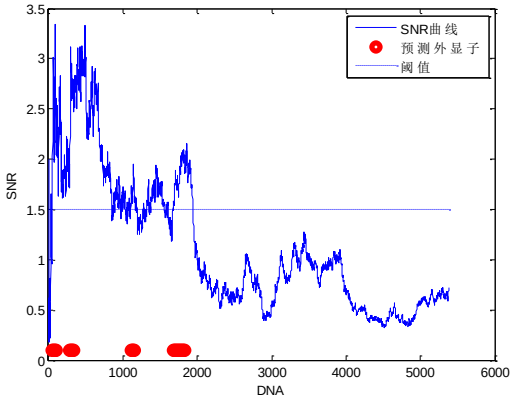

图 14 第三组数据外显子预测

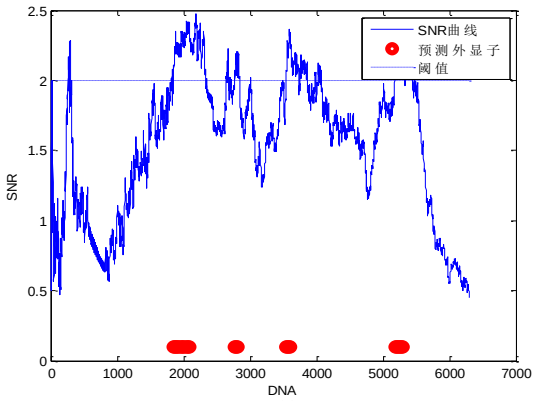

图 15 第四组数据外显子预测

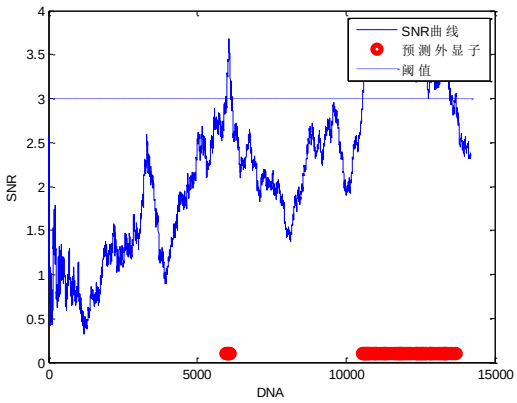

图 16 第五组数据外显子预测

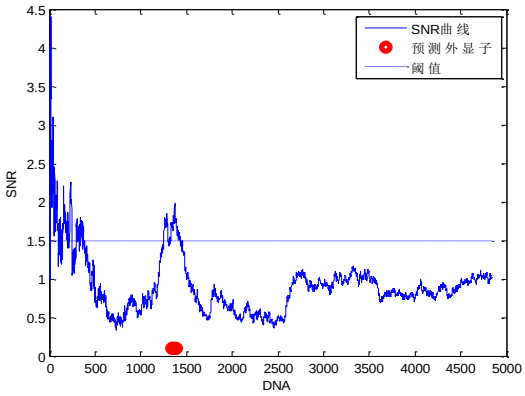

图 17 第六组数据外显子预测

4.4 基于隐马尔科夫模型的剪切位点识别

剪切位点是内含子右末端和相邻外显子左末端的边界，在基因转录中起着关键作用。通过对前体RNA进行剪接去除内含子是大多数真核基因表达的关键步骤，因此, 对剪接位点的精确预测直接关系到外显子和基因突变点的定位。同时由于剪切位点区域在生物进化上比较保守，发生缺失、插入等碱基变异的概率较小。我们根据剪切位点区域的生物特性，引入统计学中的隐马尔可夫过程，对剪切位点的供点区域建立相应模型，通过自学习后，得到隐马尔可夫模型的重估参数，用以识别DNA序列中的剪切位点。

4.4.1 模型建立与求解

我们在此处引入一种简便的隐马尔科夫模型

设 $(X_n)_{n\geq 1}$ 是概率空间 $(\Omega,\mathcal{F},\mathcal{F}_n,\mathrm{P})$ 上的时齐马尔科夫链，其状态空间是 $S=\{s_1,s_2,\ldots,s_m\}$ ，$(Y_n)_{n\geq 1}$ 是 $(\Omega,\mathcal{F},\mathcal{F}_n,\mathrm{P})$ 上与马尔科夫 $(X_n)_{n\geq 1}$ 相关联的一个随

机过程，其状态空间为 $O=\{o_1,o_2,\ldots o_K\}$，且符合下式：

$$P\left(Y_n=y_n\middle|X_n=x_n,X_{n-1},\ldots,X_1,Y_1,\ldots Y_{n-1}\right)=P\left(Y_n=y_n\middle|X_n=x_n\right)。$$

马尔科夫链$\left(X_n\right)_{n\geq1}$一般观察不到，称为隐过程，而过程 $\left(Y_n\right)_{n\geq1}$ 是由于马尔科夫链$\left(X_n\right)_{n\geq1}$的运动而产生的可观察过程。隐马尔科夫模型是，给定观察过程$\left(Y_n\right)_{n\geq1}$ 的一样轨道 $\left(y_1,y_2,\ldots y_T\right)$，求出隐过程$\left(X_n\right)_{n\geq1}$的一概率最优轨道$\left(x_1^*,x_2^*,\ldots x_T^*\right)$，即$\left(x_1^*,x_2^*,\ldots x_T^*\right)$满足下面等式：

$$\begin{aligned}&P\left\{X_1=x_1^*,X_2=x_2^*,\ldots,X_T=x_T^*;Y_1=y_1,Y_2=y_2,\ldots Y_T=y_T\right\}\\&=\max_{x_1,x_2\ldots,x_T}P\left\{X_1=x_1,X_2=x_2,\ldots X_T=x_T;Y_1=y_1,Y_2=y_2,\ldots,Y_T=y_T\right\}\end{aligned}$$

通常我们采用 Viterbi 算法求解隐过程的概率最优轨道。为了说明 Viterbi 算法，我们先引入一些记号。

设$\left(X_n\right)_{n\geq1}$的初分布为

$$\pi_i=P\left\{X_1=s_i\right\},i=1,\ldots,m,$$

$\left(X_n\right)_{n\geq1}$的转移概率为

$$p(i,j)=P\left\{X_n=s_j\middle|X_{n-1}=s_i\right\},i,j=1,\ldots,m;n=2,\ldots,T,$$

观察条件概率为

$$b_j\left(k\right)=P\left\{Y_n=o_k\middle|X_n=s_j\right\},j=1,\ldots,m;k=1,\ldots,K;n=1,\ldots,T,$$

令

$$\delta_n\left(i\right)=\max_{x_1,\cdots,x_{n-1}}P\left\{X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=s_i;Y_1=y_1,\ldots,Y_n=y_n\right\},$$

由马尔科夫性，可知

$$\begin{aligned}\delta_{n+1}\left(i\right)&=\max_{x_1,\cdots,x_{n-1}}P\left\{X_1=x_1,\ldots,X_n=x_n,X_{n+1}=s_i;Y_1=y_1,\ldots,Y_{n+1}=y_{n+1}\right\}\\&=\max_{x_1,\cdots,x_{n-1}}\left\{\max_jP\left\{X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=s_j,X_{n+1}=s_i;Y_1=y_1,\ldots,Y_{n+1}=y_{n+1}\right\}\right\}\\&=\max_{x_1,\cdots,x_{n-1}}\left\{\max_jP\left\{X_1=x_1,\ldots,X_{n-1}=x_{n-1},X_n=s_j;Y_1=y_1,\ldots,Y_{n+1}=y_{n+1}\right\}p\left(j,i\right)b_i\left(y_{n+1}\right)\right\}\\&=\max_j\delta_n(j)p\left(j,i\right)b_i\left(y_{n+1}\right)\end{aligned}$$

设$\left(y_1,y_2,\ldots y_T\right)$是观察过程$\left(Y_n\right)_{n\geq1}$的一样本轨道，Viterbi 算法如下：

(1) 初始化

$$\delta_1(i)=\pi_i b_i(y_1), i=1,\ldots,m,$$

$$\psi_1(i)=0,$$

(2) 迭代

$$\delta_{n+1}(i)=\max_j\{\delta_n(j)p(j,i)b_i(y_{n+1})\},$$

$$\varphi_{n+1}(i)=Arg\max_{j=1,\ldots,m}\{\delta_n(j)p(j,i)\}, i=1,\ldots,m; n=1,\ldots,T-1,$$

(3) 结束

$$P^*=\max_{i=1,\ldots,m}\{\delta_T(i)\},$$

$$x_T^*=Arg\max_{i=1,\ldots,m}\{\delta_T(i)\},$$

(4) 后推

$$x_n^*=\varphi_{n+1}(x_{n+1}^*), n=T-1,\ldots,1,$$

$(x_1^*,\ldots,x_T^*)$ 即为所求的概率最优轨道。

4.4.2 剪切位点的识别

我们取隐过程$(X_n)_{n\geq 1}$的状态空间为$S=\{M,D,I\}$，其中'M'表示主状态，'D'表示缺失状态，'I'表示插入状态，观察过程$(Y_n)_{n\geq 1}$的状态空间为$S=\{_,a,c,g,t\}$，其中'_'表示缺失状态，'a'，'c'，'g'，'t'，表示 DNA 序列的四种碱基。

首先我们从基因数据库中随机选取 500 条含供点的序列作为学习集，用 Viterbi 算法和极大似然估计对隐马尔科夫模型进行估计，得到和剪切位点区域相合模型。在此模型下，对给定的 DNA 序列$b_1b_2\ldots b_L$，设其在隐过程轨道$x_1x_2\ldots x_T$下对应观察过程轨道为$y_1y_2\ldots y_T$，按下式计算其概率

$$\begin{aligned}&P\{X_1=x_1,\ldots,X_T=x_T;Y_1=y_1,\ldots,Y_T=y_T\}\\&=P\{Y_1=y_1|X_1=x_1\}\prod_{n=3}^{T}P\{X_n=x_n|X_{n-1}=x_{n-1}\}P\{Y_n=y_n|X_n=x_n\}\end{aligned}$$

然后根据概率$P\{X_1=x_1^*,X_2=x_2^*,\ldots,X_T=x_T^*;Y_1=y_1,Y_2=y_2,\ldots Y_T=y_T\}$的大小来判别供点的位置。通常剪切位点处的概率较大，这说明此模型反映了剪切位点的一种统计结构。

参考文献

[1]Burge, C., Karlin, S., Prediction of complete gene structures in human genomic DNA [J]. J. Mol. Biol. 1997:268, 78–94;

[2] Anastassiou, D., Frequency-domain analysis of biomolecular sequences. Bioinformatics [J] Bioinformatics. 2000:16, 1073–1081;

[3] Kotlar, D., Lavner, Y.,Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions.[J] Genome Res. 2003:13, 1930–1937;

[4] Berryman, M. J., Allison, A., Review of signal processing in genetics [J]. Fluctuation and Noise Letters. 2005:5(4), 13-35;

[5] Sharma, S. D., Shakya，K., Sharma，S. N., 2011. Evaluation of DNA Mapping Schemes for Exon Detection. International Conference on Computer, Communication and Electrical Technology– ICCCET. 2011, 18th & 19th;

[6] Yin, C., Yau, S.S.-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence [J] Journal of Theoretical Biology. 2007:247, 687–694;

[7] Shao J.F., Yan X.H. and Shao S. SNR of DNA sequences mapped by general affine transformations of the indicator sequences [J] Journal of Mathematic Biology. 2012:7,3-15;

[8] http://www.ncbi.nlm.nih.gov/;