参赛密码 _____

（由组委会填写）

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：

基于信号处理分析方法原理，本文针对基因识别问题，提出了频谱与信噪比的快速算法，并研究了阈值确定方法，以此为基础设计了准确的快速基因识别算法，并用于判别基因突变等问题。

Voss 映射下直接使用核苷酸的频数计算 N/3处的功率谱值和信噪比，避免了 DFT 运算，减少了计算量，该方法也适用于 Z-curve、实数等映射，通过理论推导和实验验证，得出 Z-curve 映射与 Voss 映射的信噪比关系为 Rzcurve/Rvoss=4/3。

提出了一种使外显子和内含子总判决正确率最大的最优化阈值确定方法，对 3 种不同类型 DNA 序列计算了最优阈值。基于统计分析理论，对 Genes100 和 genes200 基因序列数据做了统计分析，得出 Genes100 中人与小鼠的基因序列数据对应频谱阈值为 0.3，探测率约为 86%，正确率约为 91.3%；genes200 中哺乳动物类的基因序列数据对应频谱阈值为 0.25，探测率约为 82.7%，正确率约为 71.2%。

基于 Z-curve 映射，将可变滑动窗口法与小波变换法相结合，以抑制随机噪声，实现精确基因识别。利用 NC_012920_1(人线粒体全基因组)验证了其优势，相对于固定滑动窗口法，敏感性从 0.8549 提高到 0.8919，专一性从 0.9924 提高到 0.9941，并用此法预测了 gene6 的编码区域。研究了 Genes100.mat 中#9DNA 序列加入、删除和替换单个核苷酸对应的频谱变化。结果表明可利用频谱或信噪比方法发现基因突变。

关键词：基因识别、快速算法、最优阈值、小波变换、可变滑动窗

一、 问题重述

在目前基因预测研究中，采用信号处理与分析方法来发现基因编码序列受到广泛重视。在DNA序列研究中，首先把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；然后根据碱基的3-周期特性计算出DNA序列的功率谱或信噪比；最后利用频谱或信噪比，探测、预报一个尚未被注释的完整DNA序列的所有基因编码序列（外显子）片段。

问题 1、 功率谱与信噪比的快速算法

对于很长的 DNA 序列，在计算其功率谱或信噪比时，离散 Fourier 变换(DFT)的总体计算量仍然很大，会影响到所设计的基因识别算法的效率。能否对 Voss 映射，探求功率谱与信噪比的某种快速计算方法；

在基因识别研究中，为了通过引入更好的数值映射而获取 DNA 序列更多的信息，除了 Voss 映射外，实际上还有许多不同的数值映射方法，如著名的 Z-curve 映射。试探讨 Z-curve 映射的频谱与信噪比和 Voss 映射下的频谱与信噪比之间的关系；

此外，能否对实数映射，如： $A \to 0$, $C \to 1$, $G \to 2$, $T \to 3$，也给出功率谱与信噪比的快速计算公式。

问题 2、对不同物种类型基因的阈值确定

对特定的基因类型的 DNA 序列，将其信噪比 R 的判别阈值取为 R_0=2，带有一定的主观性、经验性。对不同的基因类型，所选取的判别阈值也许应该是不同的。附件中给出了来自于著名的生物数据网站的几个基因序列数据，另外也给出了带有编码外显子信息的 100 个人和鼠类的，以及 200 个哺乳动物类的基因序列的样本数据集合。大家还可以从生物数据库下载更多的数据，找你们认为具有代表性的基因序列，并对每类基因研究其阈值确定方法和阈值结果。此外，对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。

问题 3、 基因识别算法的实现

我们的目的是要探测、预报尚未被注释的、完整的 DNA 序列的所有基因编码序列（外显子）。目前基因识别方面的多数算法结果还不是很充分。例如前面所列举的某些基因识别算法，由于 DNA 序列随机噪声的影响等原因，还很难“精确地”确定基因外显子区间的两个端点。

对此，有没有更好的解决方法。请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（gene6）的编码区域的预测。

问题 4、 延展性研究

在基因识别研究中，还有很多问题有待深入探讨。比如

（1）采用频谱或信噪比这样单一的判别特征，也许是影响、限制基因识别正确率的一个重要原因。人们发现，对某些 DNA 序列而言，其部分编码序列（外显子），尤其是短的（长度小于 100bp）的编码序列，就可能不具有频谱或者信噪比显著性。能否总结，甚至独自提出一些识别基因编码序列的其它特征指数，并对此做相关的分析。

（2）“基因突变”是生物医学等方面的一个关注热点。基因突变包括 DNA 序列中单个核苷酸的替换，删除或者插入等。那么，能否利用频谱或信噪比方法去发现基因编码序列可能存在的突变。

二、 基本假设

1）DNA序列只有合法符号A、C、G、T，没有错误符号；
2）DNA 序列具有 3-周期性；
3）DNA序列中的四个碱基等概率出现；
4）第二问提供的数据已经具有统计特性；
5）在统计内含子区域时，将整个非编码区域均认为是内含子区域。

三、 变量说明

$u_b[n]$： Voss 映射下 DNA 序列的指示序列；

$U_b[k]$：指示序列的离散 Fourier 变换（DFT）；

$P[k]$：整个序列的功率谱；

N_b：四个核苷酸 A、C、G、T 的频数；

$\bar{E}$：功率谱的平均值；

R：DNA 序列的信噪比；

α：一维映射的系数；

T_p：被正确判为外显子的个数；

T_N：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；

S_n：敏感性；

S_p：专一性；

$\psi_{a,b}(t)$：小波变换窗口函数；

$W_\psi f(a,b)$：小波函数的功率谱。

四、问题1分析及模型建立与求解

4.1 基于 Voss 映射的快速算法

对 Voss 映射，能够找出比传统 DFT 更快速的计算功率谱与信噪比的快速计算方法。

利用功率谱分析探测 DNA 序列编码区的主要特征信号三周期性，需要计算 1/3 频率点的傅里叶频谱。当处理长 DNA 序列时，DFT 计算量仍然很大，针对该问题，提出了只计算 1/3 频率点处的傅里叶频谱快速预测 DNA 序列编码区的方法。理论分析和实验证明，该方法的计算速度比使用傅里叶变换或快速傅里叶变换的方法快，计算准确性保持不变，不需要一个训练组或现有数据库的信息。

一个基因组序列可以看作是由A,T,C,G四种碱基所构成的符号序列，在对基因组序列进行计算分析之前，先将其转化为数值序列。Voss 法是应用最为广泛且较早提出的1种DNA序列数值化表示方法。该方法将1个长度为N的DNA序列表示为4个二进制数字序列。

该算法的主要优点是不会引入相关；可以证明任何维数小于4的表示方法其本身就会引入相关[1]。

采用 Voss 映射，由公式

$$u_b[n]=\begin{cases}1, & S[n]=b\\ 0, & S[n]\neq b\end{cases},\quad n=0,1,2,\cdots N-1 \tag{4-1-1}$$

$$U_b[k]=\sum_{n=0}^{N-1}u_b[n]e^{-j\frac{2\pi nk}{N}},\quad k=0,1,\cdots,N-1 \tag{4-1-2}$$

$$P[k]=|U_A[k]|^2+|U_T[k]|^2+|U_G[k]|^2+|U_C[k]|^2,\ k=0,1,\cdots N-1 \tag{4-1-3}$$

其中 $b\in I$，$I=\{A,T,G,C\}$。

（1）计算功率谱平均值

设 DNA 序列的四个核苷酸 A、C、G、T 的频数为 N_A,N_C,N_G,N_T，根据 DFT 形式下的 Parseval 定理[2]：

$$\sum_{n=0}^{N-1}|u[n]|^2=\frac{1}{N}\sum_{k=0}^{N-1}|U[k]|^2 \tag{4-1-4}$$

可以计算功率谱的平均值

$$\begin{aligned}\overline{E} &= \frac{1}{N}\sum_{k=0}^{N-1} P[k] \\ \overline{E} &= \frac{1}{N}\sum_{k=0}^{N-1}(|U_A[k]|^2 + |U_T[k]|^2 + |U_G[k]|^2 + |U_C[k]|^2) \\ &= \sum_{n=0}^{N-1}(|u_A[n]|^2 + |u_T[n]|^2 + |u_G[n]|^2 + |u_C[n]|^2)\end{aligned} \qquad (4\text{-}1\text{-}5)$$

根据 Voss 映射的特点，对于$|u_A[n]|,|u_T[n]|,|u_G[n]|,|u_C[n]|$，同时只有一个为 1，其余三个为 0，有

$$|u_A[n]|^2 + |u_T[n]|^2 + |u_G[n]|^2 + |u_C[n]|^2 = 1$$

则功率谱的平均值$\overline{E} = N$。对于每种核苷酸，$|U_b|^2 = \sum_{k=0}^{N-1} |U_b[k]|^2 = N \cdot N_b$。

（2）计算$P[\frac{N}{3}]$

在*DNA*序列$\{S[n],\ n = 0,1,2,\cdots N-1\}$中，将核苷酸符号$b \in I = \{A,T,G,C\}$出现在该序列的0,3,6,... N－3与1,4,7,…N－2以及2,5,8,…N－1等位置上的频数分别记为x_b, y_b和z_b，则$\frac{N}{3}$处的总功率谱值即为[3]

$$\left|U_b\left[\frac{N}{3}\right]\right|^2 = (x_b, y_b, z_b)\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}\begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} = X_b^T M X_b \qquad (4\text{-}1\text{-}6)$$

其中$X_b = (x_b, y_b, z_b)^T$，M是二次型系数矩阵。

证明如下：

$$\begin{aligned} P[\frac{N}{3}] &= \sum_{b\in I}\left|U_b[\frac{N}{3}]\right|^2 = \sum_{b\in I}\left|\sum_{n=0}^{N-1} u_b[n]\cdot e^{-j\frac{2\pi n\cdot\frac{N}{3}}{N}}\right|^2 = \sum_{b\in I}\left|\sum_{n=0}^{N-1} u_b[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2 \\ &= \sum_{b\in I}\left|x_b + y_b\cdot e^{-j\frac{2\pi}{3}} + z_b\cdot e^{j\frac{2\pi}{3}}\right|^2 \\ &= \sum_{b\in I}(x_b^2 + y_b^2 + z_b^2 - x_b y_b - x_b z_b - y_b z_b) \\ &= \sum_{b\in I}(x_b, y_b, z_b)\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}\begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} = \sum_{b\in I} X_b^T M X_b \end{aligned} \qquad (4\text{-}1\text{-}7)$$

序列$\frac{N}{3}$处的功率谱为

$$P[\frac{N}{3}]=\sum_{b\in I} X_b^T M X_b \qquad (4\text{-}1\text{-}8)$$

如果 $x_b=y_b=z_b$，上式计算结果为0，说明内含子没有3-周期特性。

（3）根据前两步计算的值求 DNA 序列的信噪比

$$R=\frac{P[\frac{N}{3}]}{\bar{E}}=\frac{1}{N}P[\frac{N}{3}] \qquad (4\text{-}1\text{-}9)$$

对于每个指示序列 $u_b[n]$，$b\in I$，$I=\{A,T,G,C\}$，

$$R_b=\frac{X_b^T M X_b}{N_b} \qquad (4\text{-}1\text{-}10)$$

总的 DNA 序列的 SNR 可以表示为

$$R=\frac{N_A}{N}R_A+\frac{N_C}{N}R_C+\frac{N_G}{N}R_G+\frac{N_T}{N}R_T \qquad (4\text{-}1\text{-}11)$$

（4）仿真分析

根据以上原理，以编号为 *BK006948.2* 的酵母基因 DNA 序列中的外显子（区间为[81787，82920]）为例，四类核苷酸的出现频率为

$$X_A=(x_A,y_A,z_A)^T=(135,\ 123,\ 82)^T,\quad X_C=(x_C,y_C,z_C)^T=(72,\ 52,\ 120)^T$$
$$X_G=(x_G,y_G,z_G)^T=(97,\ 90,\ 65)^T,\quad X_T=(x_T,y_T,z_T)^T=(74,\ 113,\ 111)^T$$

$$N=1134,\ N_A=340,\ N_C=244,\ N_G=252,\ N_T=298$$

每种核苷酸的 SNR 为 $R_A=6.8147, R_C=15.0164, R_G=3.3690, R_T=4.8557$，总 SNR 为 $R=7.2989$。

以同一个 DNA 序列中的内含子（区间为[96362, 97550]，长 1191*bp*）为例，四类核苷酸的出现频率为

$$X_A=(x_A,y_A,z_A)^T=(117,\ 118,\ 110)^T,\quad X_C=(x_C,y_C,z_C)^T=(75,\ 73,\ 70)^T$$
$$X_G=(x_G,y_G,z_G)^T=(87,\ 66,\ 71)^T,\quad X_T=(x_T,y_T,z_T)^T=(118,\ 139,\ 145)^T$$

$$N=1189,\ N_A=345,\ N_C=218,\ N_G=224,\ N_T=402$$

每种核苷酸的 SNR 为 $R_A=0.1652, R_C=0.0872, R_G=1.6116, R_T=1.5000$，总 SNR 为 $R=0.8747$。

可以看出外显子与内含子 SNR 有明显差异。

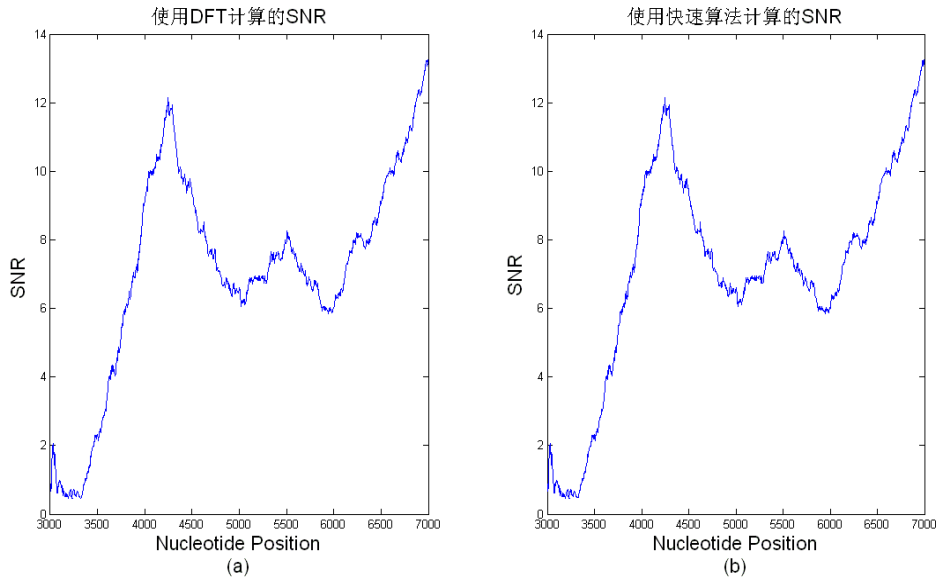


图 4-1　(a) 使用 DFT 计算的 SNR
(b) 使用快速算法计算的 SNR（人类线粒体基因，NC_012920_1.fasta）

由上图的仿真结果可见，使用 DFT 和使用快速算法计算的 SNR 完全一致，但快速算法计算的时间更少：DFT 时间为 16.465065s，而快速算法时间仅为 1.469127s。

（5）快速算法的优点：

由于 N/3 处的功率谱和信噪比可以直接从核苷酸出现频数计算得到，从而不需要进行 DFT 运算，减少了计算量；使用这个公式计算时不要求 N 是 3 的倍数。

通过理论分析和实验证实，利用本文提出的基于傅立叶技术的快速预测方法对基因组序列的编码区进行预测可取得良好的效果。该方法的显著优点是运算速度比利用FFT的方法快（由于不需要计算所有点的傅里叶频谱，而只计算f =1/3点的频谱），容易应用，不需要基因组序列的任何先验知识；并且可同时实现基因的预测和定位。预测出编码区的大概位置，为进一步用实验方法精确定位编码区打下基础。正如文献[4]所指出的，通常难以用一种方法将各种生物DNA序列的编码区预测问题全部解决，需要多种方法融合，才能达到准确预测和定位编码区的目的。

4.2 Z-curve 与 Voss 映射下频谱与信噪比关系

Voss、Z-Curve法的预测结果几乎相同，Z-Curve实际上是对Voss法进行了一种变换后得到的，而且都具有明确的生物学意义。

Z-curve 映射为

$$\begin{pmatrix}\Delta x[n]\\ \Delta y[n]\\ \Delta z[n]\end{pmatrix}=\begin{pmatrix}1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix}\begin{pmatrix}u_A[n]\\ u_C[n]\\ u_G[n]\\ u_T[n]\end{pmatrix} \qquad (4\text{-}2\text{-}1)$$

其中，$\Delta x[n]=x[n]-x[n-1]$，$\Delta y[n]=y[n]-y[n-1]$，$\Delta z[n]=z[n]-z[n-1]$，设 $x[-1]=0$，$y[-1]=0$，$z[-1]=0$，$u_A[n],u_C[n],u_G[n],u_T[n],n=0,1,\cdots N-1$ 是 Voss 变换的四个指示序列。

定义 Z-curve 映射的总功率谱

$$P_Z[k]=\left|\Delta X[k]\right|^2+\left|\Delta Y[k]\right|^2+\left|\Delta Z[k]\right|^2 \qquad (4\text{-}2\text{-}2)$$

其中 $\Delta X[k],\Delta Y[k]$ 和 $\Delta Z[k]$ 分别表示数字序列 $\Delta x[n],\Delta y[n]$ 和 $\Delta z[n]$ 的离散傅立叶变换。

对于 Voss 映射，存在 $u_A[n]+u_T[n]+u_G[n]+u_C[n]=1,$ 对于所有 n，这四个指示序列是线性相关的，对四个指示序列的和进行 DFT，可以得到

$$U_A[k]+U_T[k]+U_G[k]+U_C[k]=\sum_{n=0}^{N-1}(u_A[n]+u_T[n]+u_G[n]+u_C[n])e^{-j\frac{2\pi nk}{N}},\ k=0,1,\cdots,N-1$$

$$U_A[k]+U_T[k]+U_G[k]+U_C[k]=\begin{cases}0, & k\neq 0\\ N, & k=0\end{cases} \qquad (4\text{-}2\text{-}3)$$

可以看出这也是线性相关的，那么计算时可以减少一个指示序列，通过一个系数矩阵，将四个指示序列转化成三个新的指示序列。Z-curve 映射就是把四个指示序列降维到三个指示序列，去除了冗余量，减少了计算量。

Z-curve 映射的总功率谱 $P_Z[k]$ 与 Voss 映射的总功率谱 $P_{voss}[k]$ 的关系为

$$P_Z[k]=\begin{cases}4P_{voss}[k], & k\neq 0\\ 4P_{voss}[0]-N^2, & k=0\end{cases},\quad \sum_{k=0}^{N-1}P_z[k]=3\sum_{k=0}^{N-1}P_{voss}[k],\quad \overline{E}_z=3\overline{E}_{voss} \qquad (4\text{-}2\text{-}4)$$

Z-curve 映射的信噪比 R_z 与 Voss 映射的信噪比 R_{voss} 关系为

$$\frac{R_z}{R_{voss}}=\frac{4}{3} \qquad (4\text{-}2\text{-}5)$$

不考虑 $k=0$ 时的功率谱值。

$P_Z[k]$ 与 $P_{voss}[k]$ 关系证明如下：

$$
\begin{aligned}
P_Z[k] &= |\Delta X[k]|^2 + |\Delta Y[k]|^2 + |\Delta Z[k]|^2 \\
&= (U_A[k] - U_C[k] + U_G[k] - U_T[k])^2 + (U_A[k] + U_C[k] - U_G[k] - U_T[k])^2 \\
&\quad + (U_A[k] - U_C[k] - U_G[k] + U_T[k])^2 \\
&= 3(U_T^2[k] + U_C^2[k] + U_G^2[k] + U_A^2[k]) \\
&\quad - 2(U_T[k]\cdot U_C[k] + U_T[k]\cdot U_A[k] + U_T[k]\cdot U_G[k] + U_A[k]\cdot U_C[k] + U_A[k]\cdot U_G[k] + U_C[k]\cdot U_G[k]) \\
&= 3(U_T^2[k] + U_C^2[k] + U_G^2[k] + U_A^2[k]) \\
&\quad - [(U_A[k] + U_T[k] + U_G[k] + U_C[k])^2 - (U_T^2[k] + U_C^2[k] + U_G^2[k] + U_A^2[k])] \\
&= \begin{cases} 4P[k], & k \neq 0 \\ 4P[0] - N^2, & k = 0 \end{cases}
\end{aligned}
$$

对编号为 *BK006948.2* 的酵母基因 DNA 序列中的外显子(区间为[81787，82920]，长 1134*bp*)分别作 Z-curve 映射下的功率谱和 Voss 映射下的功率谱，

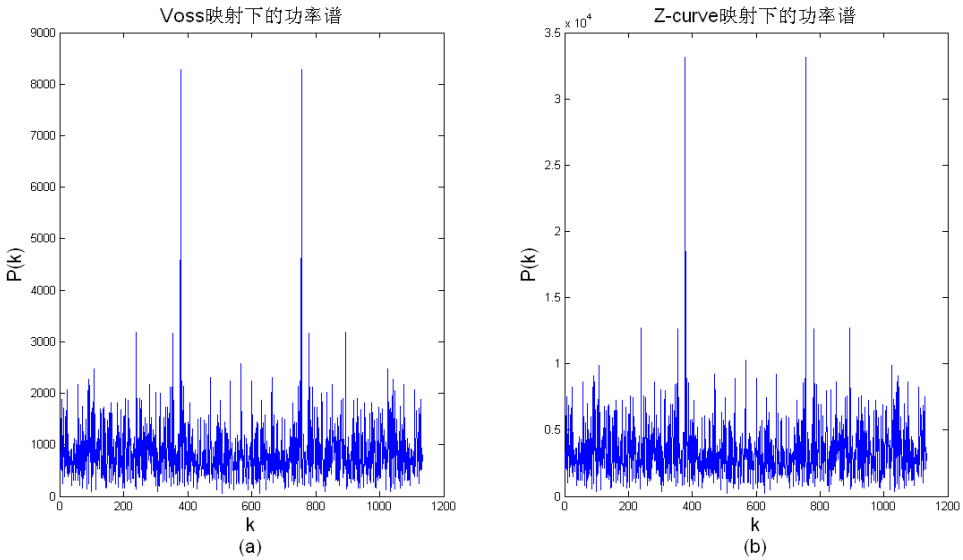


图 4-2　*BK006948.2* 外显子(区间为[81787，82920]，长 1134*bp*)
(a) Z-curve 映射下的功率谱和 (b)Voss 映射下的功率谱

从图中可以明显看到在频率的 1 / 3 处的功率谱存在峰值，该峰值也证明了该序列中存在外显子。

信噪比计算得：

$$R_z = 9.7319,\ R_{voss} = 7.2989,\ \frac{R_z}{R_{voss}} \approx \frac{4}{3}$$

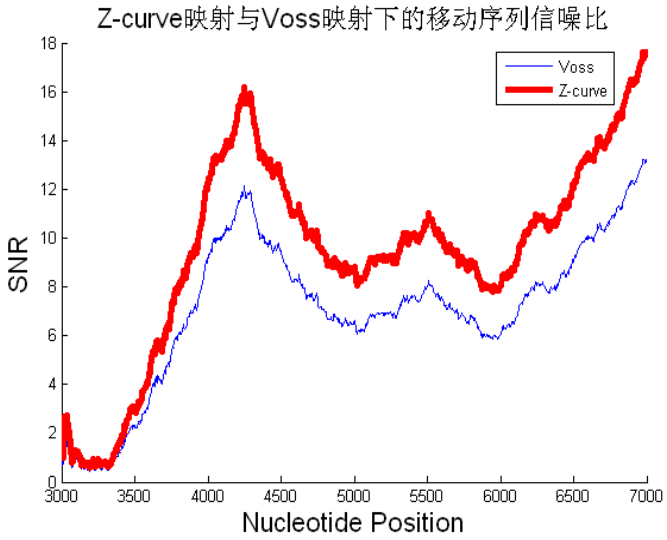


图 4-3　Z-curve 与 Voss 映射下的移动序列信噪比
（人类线粒体基因，NC_012920_1.fasta）

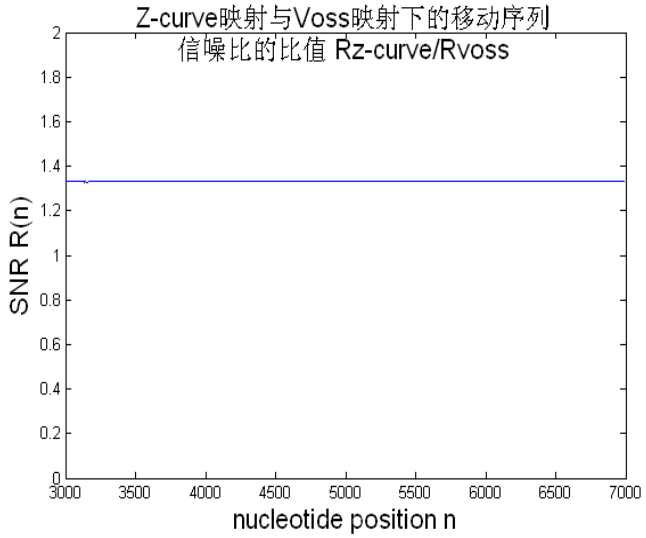


图 4-4　Z-curve 与 Voss 映射下的移动序列信噪比比值 Rz-curve/Rvoss
（人类线粒体基因，NC_012920_1.fasta）

仿真实验验证了 $\frac{R_z}{R_{voss}}=\frac{4}{3}$。

当采用Voss法对序列进行数值映射时，DNA序列中有三种不同的碱基被映射为同一数值，使得序列中原来变化的信号部分(即交流分量)经映射后，变为不变的部分(即直流分量)，因此采用这类方法得到的频谱图的直流分量被增大，而交流分量被削弱。

Voss 映射功率谱的直流分量很大也可从仿真结果看出，$\frac{P_{voss}[0]}{P_{voss}}=0.2546$，而 Z-curve 映射功率谱的直流分量很小，$\frac{P_z[0]}{P_z}=0.0062$，Z-curve 映射对直流分量有

抑制作用，这点可以从 Z-curve 映射性质得到。

从 Z-curve 映射的系数矩阵中看出，每行系数的和为 0。

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

在计算 $P_z[0]$ 时相互抵消，减少了直流分量，加强了 3-周期性，提高了信噪比。

4.3 实数映射的快速算法

对于实数映射，也能找出功率谱与信噪比的快速计算公式。

实数映射将 Voss 映射变换为一维，如 $A\to 0, C\to 1, G\to 2, T\to 3$。设

$$\alpha=[\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4]^T=[0\ 1\ 2\ 3]^T$$

$$x[n]=0\cdot u_A[n]+1\cdot u_C[n]+2\cdot u_G[n]+3\cdot u_T[n]=\alpha^T\cdot\left[u_A[n]\ u_C[n]\ u_G[n]\ u_T[n]\right]^T \quad (4\text{-}3\text{-}1)$$

（1）计算功率谱平均值

根据 Parseval 定理

$$\begin{aligned}\sum_{k=0}^{N-1}|X(k)|^2 &= N\cdot\sum_{n=0}^{N-1}|x[n]|^2 = N\cdot\sum_{n=0}^{N-1}\left|\alpha^T\cdot\left[u_A[n]\ u_C[n]\ u_G[n]\ u_T[n]\right]^T\right|^2\\ &= N\cdot(\alpha_1^2N_A+\alpha_2^2N_C+\alpha_3^2N_G+\alpha_4^2N_T)\end{aligned} \quad (4\text{-}3\text{-}2)$$

其中 DNA 序列的四个核苷酸 A、C、G、T 的频数为 N_A, N_C, N_G, N_T。

功率谱平均值

$$\overline{E}=\frac{1}{N}\sum_{k=0}^{N-1}|X(k)|^2=\alpha_1^2N_A+\alpha_2^2N_C+\alpha_3^2N_G+\alpha_4^2N_T \quad (4\text{-}3\text{-}3)$$

（2）计算 $P[\frac{N}{3}]$

计算 $\frac{N}{3}$ 处的功率谱为

$$P[\frac{N}{3}]=\left|X[\frac{N}{3}]\right|^2=\left|\sum_{n=0}^{N-1}x[n]\cdot e^{-j\frac{2\pi n\cdot\frac{N}{3}}{N}}\right|^2=\left|\sum_{n=0}^{N-1}x[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2$$

$$=\left|\sum_{j=1}^{4}\alpha_j\sum_{n=0}^{N-1}u_j[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2 \qquad (4\text{-}3\text{-}4)$$

$$=\left|\sum_{j=1}^{4}\alpha_j(x_j+y_j\cdot e^{-j\frac{2\pi}{3}}+z_j\cdot e^{j\frac{2\pi}{3}})\right|^2$$

$$=\alpha^T X^T MX\alpha$$

其中 $X=\begin{pmatrix}x_A & x_C & x_G & x_T\\ y_A & y_C & y_G & y_T\\ z_A & z_C & z_G & z_T\end{pmatrix}=\begin{pmatrix}X_1 & X_2 & X_3 & X_4\end{pmatrix}$，$M=\begin{pmatrix}1 & -1/2 & -1/2\\ -1/2 & 1 & -1/2\\ -1/2 & -1/2 & 1\end{pmatrix}$。

x_b, y_b 和 z_b 为核苷酸($b\in I=\{A,T,G,C\}$)出现在该序列的 0,3,6,...与 1,4,7,…以及 2,5,8,…等位置上的频数。

（3）根据以上两步得到的值计算 DNA 序列的信噪比

$$R=\frac{P[\frac{N}{3}]}{\overline{E}}=\frac{P[\frac{N}{3}]}{\alpha_1^2N_A+\alpha_2^2N_C+\alpha_3^2N_G+\alpha_4^2N_T}=\frac{\alpha^T X^T MX\alpha}{\alpha_1^2N_A+\alpha_2^2N_C+\alpha_3^2N_G+\alpha_4^2N_T} \qquad (4\text{-}3\text{-}5)$$

（4）仿真分析

根据以上原理，以人类线粒体基因，NC_012920_1.fasta 为例，对传统 DFT 和所提快速算法进行仿真对比。

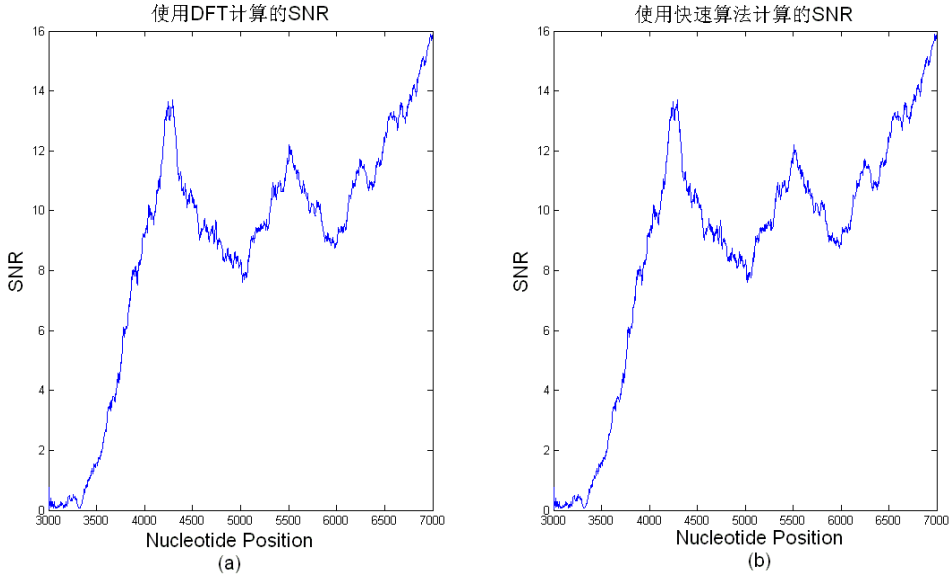


图 4-5　(a) 使用 DFT 计算的 SNR
(b) 使用快速算法计算的 SNR（人类线粒体基因，NC_012920_1.fasta）

由仿真可见，使用 DFT 和使用快速算法计算的 SNR 完全一致，但快速算法计算的时间更少。DFT 时间为 4.530696s，而快速算法时间为 1.720497s。

（5）快速算法的优点：

由于 N/3 处的功率谱和信噪比可以直接从核苷酸出现频数计算得到，不需要进行 DFT 运算，减少了计算量；使用这个公式计算时不要求 N 是 3 的倍数。

五、问题2分析及模型建立与求解

5.1 问题 2 分析

对于真核生物，其基因结构较复杂，许多基因是断裂基因，间断成外显子（exon）和内含子（intron），并且外显子在序列中长度比例极小[5]。对不同的基因类型，所选取的判别阈值应该是不同的。

对序列 $x(n)$ 进行DFT：

$$X(k)=DFT[X(n)]=\sum_{n=0}^{N-1}x(n)\exp(-2\pi ink/N) \tag{5-1-1}$$

将上式的相乘做和部分用矩阵来表示：

$$X=Wx=\begin{bmatrix}1 & 1 & \cdots & 1\\ 1 & \omega & \cdots & \omega^{(N-1)}\\ \vdots & \vdots & \cdots & \vdots\\ 1 & \omega^{(N-1)} & \cdots & \omega^{(N-1)(N-1)}\end{bmatrix}\begin{bmatrix}x(0)\\ x(1)\\ \vdots\\ x(N-1)\end{bmatrix} \tag{5-1-2}$$

其中，$\omega=\exp(-2i\pi/N)$。

这样可以通过下式计算得到序列在N/3点处的频谱：

$$\begin{aligned}S(N/3)&=\begin{bmatrix}1 & \omega^{N/3} & \cdots & \omega^{(N-1)N/3}\end{bmatrix}\begin{bmatrix}x(0)\\ x(1)\\ \vdots\\ x(N-1)\end{bmatrix}\\ &=\sum_{n=0}^{N-1}x(n)\omega^{nN/3}\end{aligned} \tag{5-1-3}$$

可以在不需要计算矩阵中的其他部分的情况下仅通过求解上式来得到序列在N/3的频率特性，在待测序列很长时，这种方法可大大提高运算效率。

从问题 1 的第二小问可看出，对同样的数据进行处理，Z-curve 映射法要比 Voss 映射法计算的信噪比值差异更大，更便于区分，所以针对这一小题采用 Z-curve 映射。

采用固定的滑动窗步长来预测与定位蛋白编码序列时，滑动窗步长取得太长或太短都会严重影响到蛋白编码区的定位精度和预测蛋白编码区的计算复杂度。本文采用一种具有可变滑动窗步长的功率谱估计方法，能有效提高蛋白编码区定位精度，减小计算复杂度。

现假设所选定的信噪比分类阈值为 R_0，即 $R\geq R_0$ 作为外显子的判别，$R<R_0$

则作为内含子的判别。通过阈值判别外显子与内含子的效果可用敏感性和专一性来表示[7-8]。

$$\text{敏感性 } S_n = T_p / (T_p + F_N) \tag{5-1-4}$$

$$\text{专一性 } S_p = T_N / (T_N + F_P) \tag{5-1-5}$$

式中：T_p表示被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。最后，阈值判别的总正确率A_c定义为

$$A_c = \frac{S_n + S_p}{2} \tag{5-1-6}$$

探测率=正确探测的编码序列数/已知编码区序列数；

缺失率=缺失的编码序列数/已知编码序列数；

正确率=探测到的编码序列数/（正确探测的编码序列数+误探测的编码序列数）；

错误率=误探测到的编码序列数/（正确探测的编码序列数+误探测的编码序列数）。

这里采用一种判别正确率较高的阈值确定的最优化方法：

设所有外显子的信噪比值组成的集合为S_1，所有内含子信噪比值组成的集合为S_2。欲寻求的最优分类阈值为R_0。设$R_i^{(1)} \in S_1$，$R_j^{(2)} \in S_2$，求解阈值R_0的优化模型为

$$\max \sum_i \text{sgn}(R_i^{(1)} - R_0) + \sum_j \text{sgn}(R_0 - R_j^{(2)}) \quad (a < R_0 < b) \tag{5-1-7}$$

其中$[a,b]$为信噪比值域区间。即在基因外显子、内含子信噪比样本集上，优化模型求得使判别正确率达到最大的阈值解R_0。

影响信噪比阈值和基因外显子判别正确率的因素：

在进行DNA序列信噪比的统计时，注意到部分外显子的信噪比并不显著，同样地也存在信噪比较大的内含子，这是造成误判的根源。产生这一现象的原因至少有序列长度影响和不同基因类型的影响。

5.2 问题2仿真分析

基于Z-curve映射，根据生物体的已知先验生物信息，适当选择滑动窗的长度M，运行速度更快，且精度更高。编程流程图如下所示：

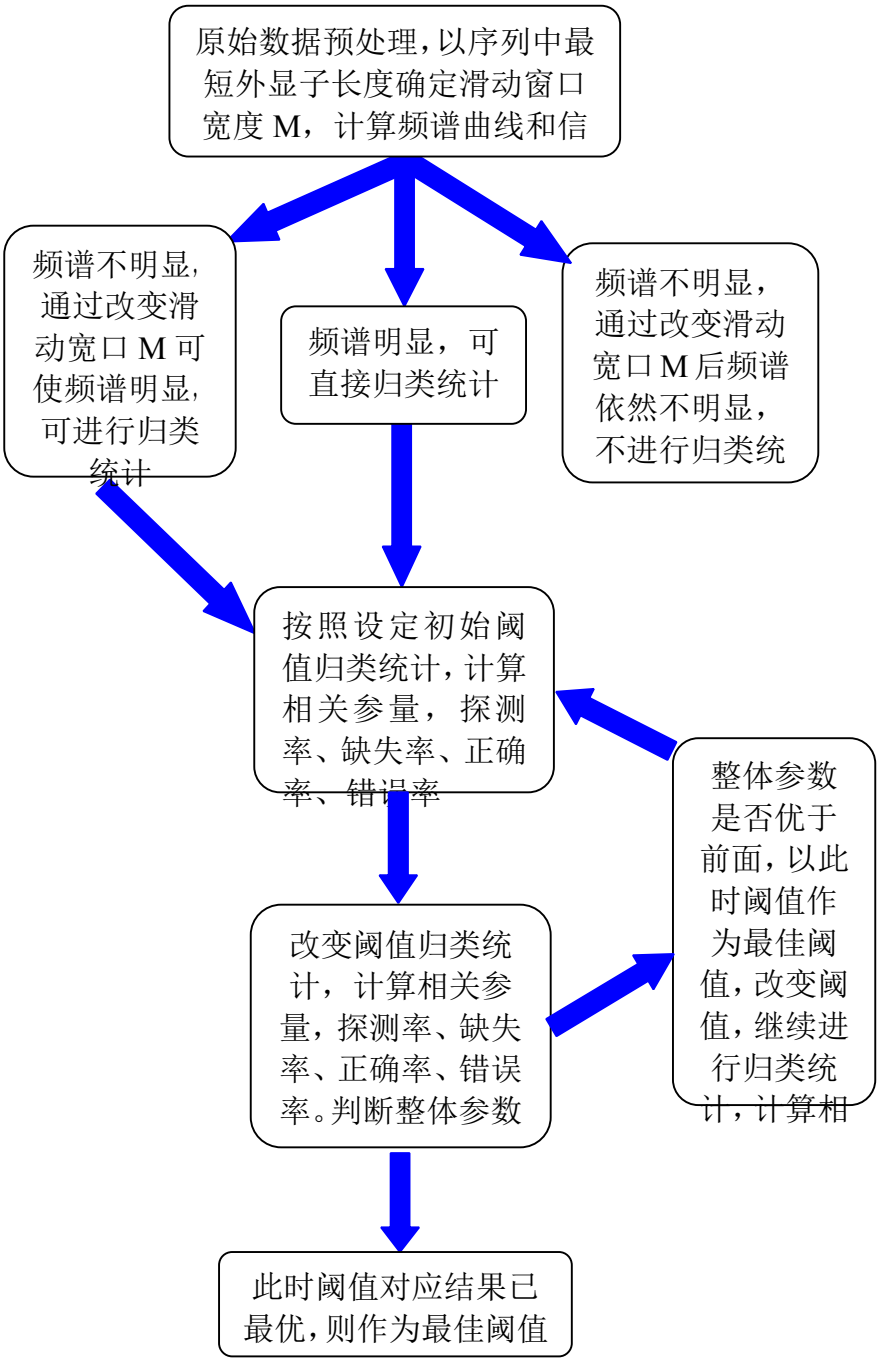
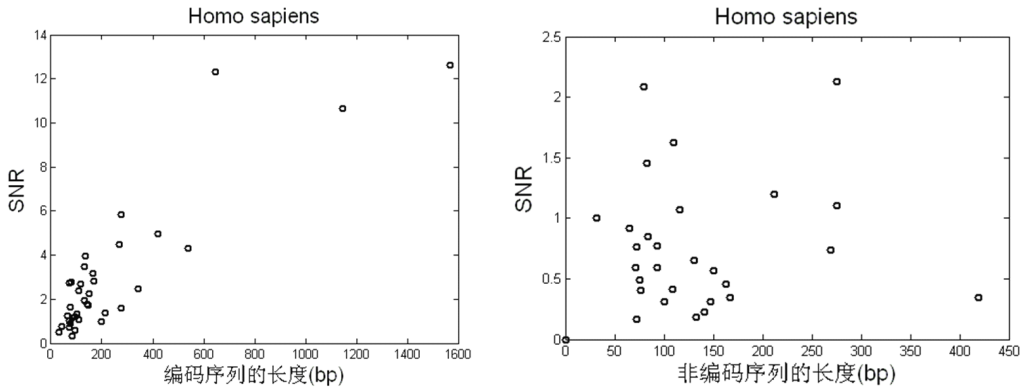


图 5-1　基于 Z-curve 映射编程流程图

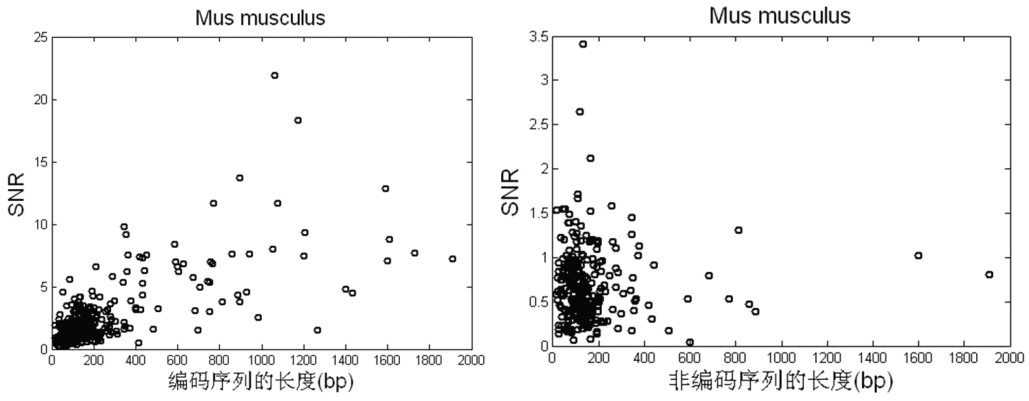
（1）信噪比统计实例

对 100 个人和鼠类的三种基因序列（Homo sapiens，Mus musculus，Rattus norvegicus）进行分析，这些基因序列已经带有编码外显子信息，对每个外显子区间与内含子区间单独统计信噪比。



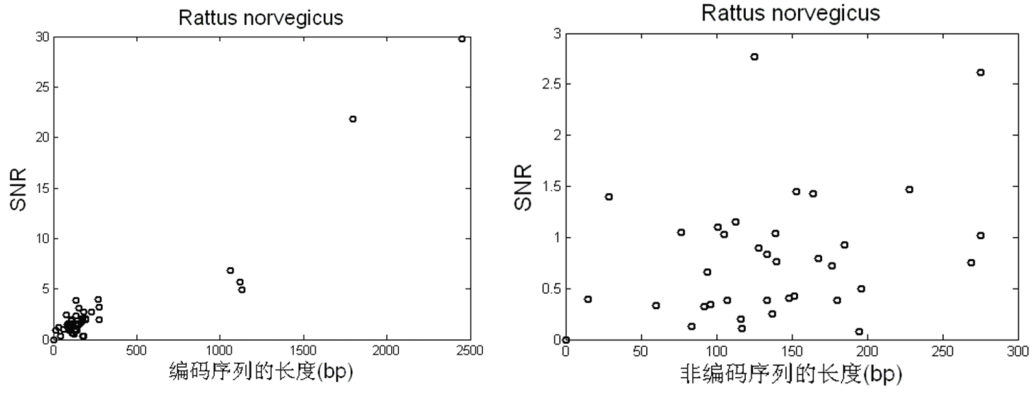
（a） （b）

图 5-2 Homo sapiens（a）外显子的信噪比（b）内含子的信噪比



（a） （b）

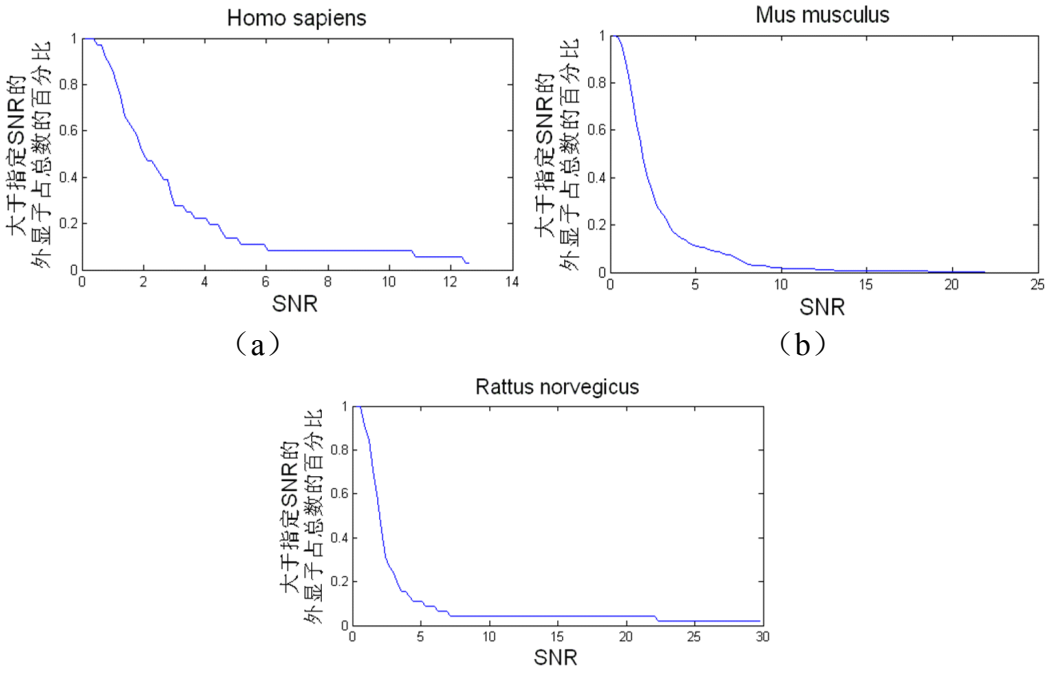
图 5-3 Mus musculus （a）外显子的信噪比（b）内含子的信噪比



（a） （b）

图 5-4 Rattus norvegicus（a）外显子的信噪比（b）内含子的信噪比

从图中看出，当序列长度比较长时，信噪比比较大，3-周期性比较强，容易被识别出来。而当序列长度比较短时，信噪比比较小，不容易识别，而且三种基因的外显子大多比较短（例如小于 200bp）。三种基因的内含子信噪比基本上都在一个范围内（例如 SNR < 2）。

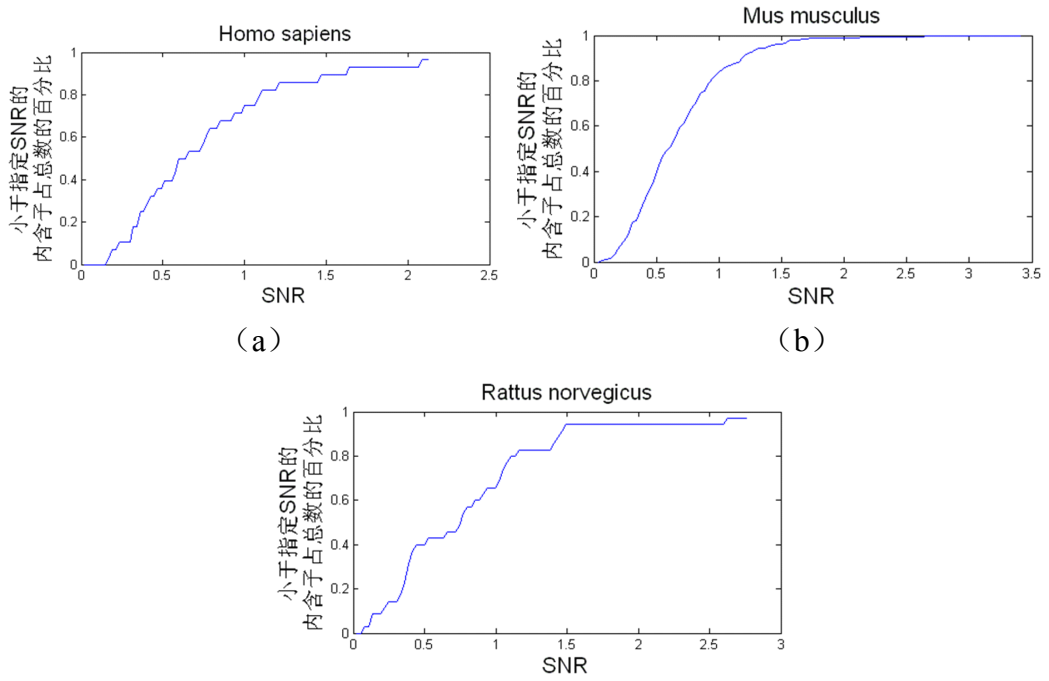


（a）　　（b）

（c）

图 5-5　大于指定 SNR 的外显子占总数的百分比

(a) Homo sapiens (b) Mus musculus (c) Rattus norvegicus

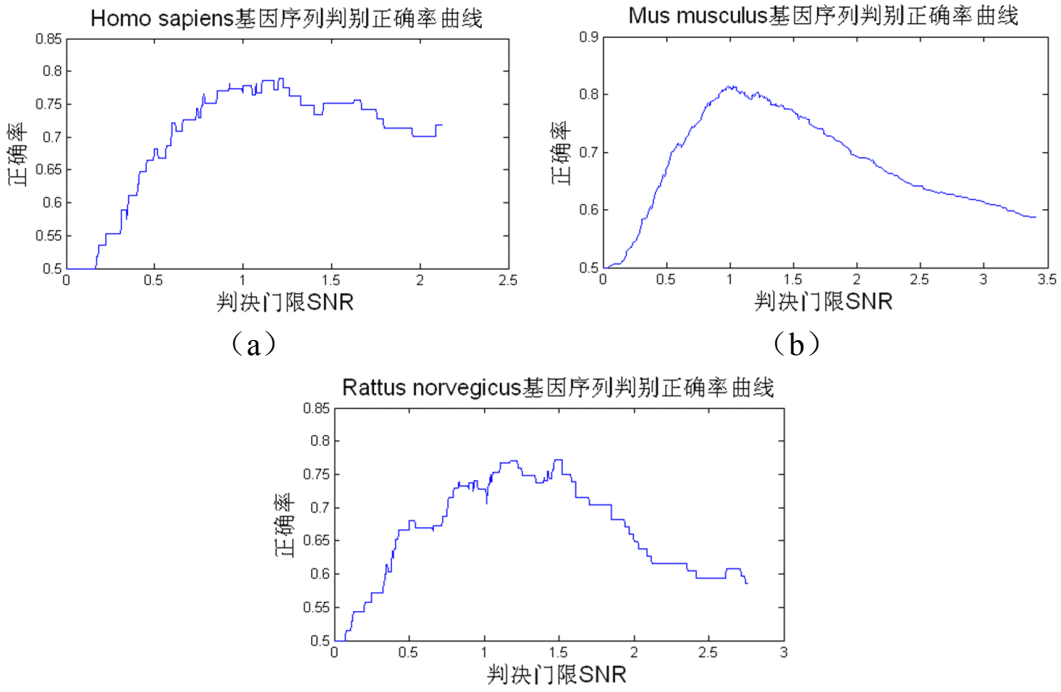


（a）　　（b）

（c）

图 5-6　小于指定 SNR 的内含子占总数的百分比

(a) Homo sapiens (b) Mus musculus (c) Rattus norvegicus



（a）　　（b）

（c）

图 5-7　基因序列判别正确率曲线

(a) Homo sapiens (b) Mus musculus (c) Rattus norvegicus

表 5-1　阈值确定最优方法统计

	外显子数量	内含子数量	最优 SNR	正确率
Homo sapiens	36	28	1.2005	0.7897
Mus musculus	357	275	0.9863	0.8141
Rattus norvegicus	45	35	1.4707	0.7714

分析：三个阈值都小于 2，可以看到如果阈值为 2 时准确率将明显小于最优阈值的准确率。对于不同种的基因，最优判决阈值 SNR 有一定的差异，对于不同类型的基因选择不同的阈值是有依据的。

影响阈值与判决正确率的因素：基因序列随着外显子长度的增加，外显子序列的信噪比和判决正确率也随之增加，当外显子长度比较小时，信噪比比较小，判决正确率将减小。

（2）对几类基因实例的频谱阈值确定及分析

为了确定频谱阈值，本文对题目提供的基因序列数据 Genes100.mat 和 genes200.mat 进行了统计分析。为了能够从统计结果中得到预期结果，首先对统计设定规则。由于题目提供的数据具有很大的随机性，本文采用以下方法对两组数据进行归类并进行统计分析。

以 Genes100.mat 中数据为例进行说明。由于提供的数据随机性大，采用滑动窗口方法计算频谱曲线时具体将其窗口宽度设为多少要视情况而定，窗口宽度 M 取得过大或者过小，都无法得到预期的结果。在以下各种计算结果中，窗口宽度 M 均采用最小外显子宽度（或者最小外显子宽度值附近）作为初值。

注：从图 5-8 到图 5-15，粗虚线代表阈值，一段段的粗实线是已知外显子位置，实曲线为频谱或信噪比曲线。

1、可以直接进行统计的序列

采用序列数据 Genes100#15AccessionNO:AF074912 计算相应的频谱曲线和信噪比移动曲线，（Genes100 表示数据源时来自 Genes100.mat，#15 表示数据序号，即 Genes100.mat 的第 15 个数据，AccessionNO:AF074912 表示序列数据的编号），Genes100#15AccessionNO:AF074912 频谱图如图 5-8 所示，Genes100#15AccessionNO:AF074912 信噪比如图 5-9 所示。

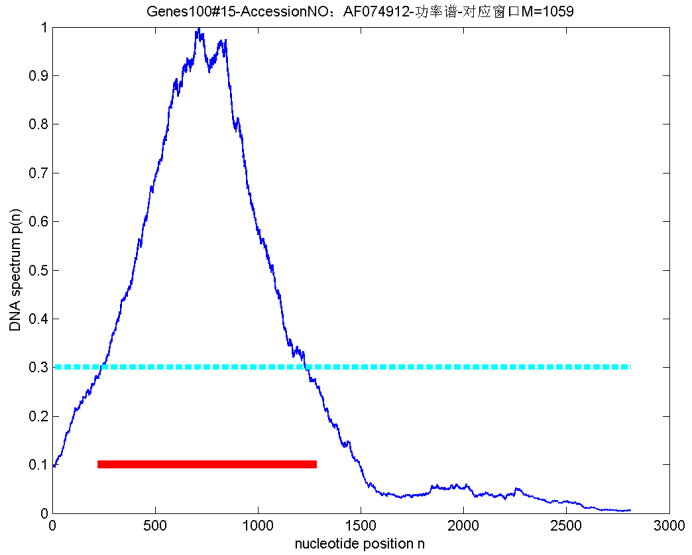


图 5-8　Genes100#15AccessionNO:AF074912 频谱图

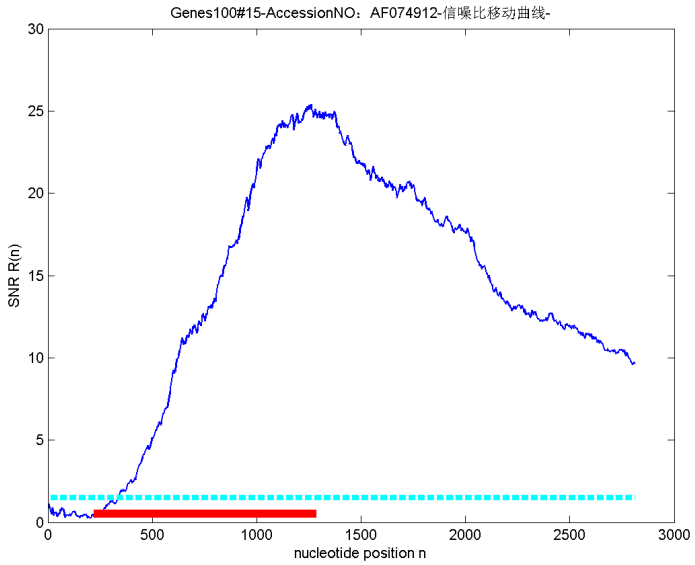


图 5-9　Genes100#15AccessionNO:AF074912 信噪比

从图 5-8Genes100#15AccessionNO:AF074912 频谱图可以直接判定设置的滑动窗口宽度是合适的，同时以图 5-9Genes100#15AccessionNO:AF074912 信噪比来辅助判断，由此可以统计相关结果，统计结果如表 5-2 所示。

表 5-2　Genes100#15AccessionNO:AF074912 统计结果

序号	窗口宽度 M	已知编码序列数	正确探测的编码序列数	误探测的编码序列数	缺失的编码序列数	探测率	缺失率	正确率	错误率
15	1059	1	1	0	0	1	0	1	0

2、改变滑动窗口可以进行统计的序列

采用序列数据 Genes100#81AccessionNO:U93050 计算相应的频谱曲线和信噪比移动曲线，Genes100#81AccessionNO:U93050 窗口长 M=69 时的频谱图如图 5-10 所示，Genes100#81AccessionNO:U93050 窗口长 M=183 时的频谱图如图 5-11 所示，Genes100#81AccessionNO:U93050 信噪比如图 5-12 所示。

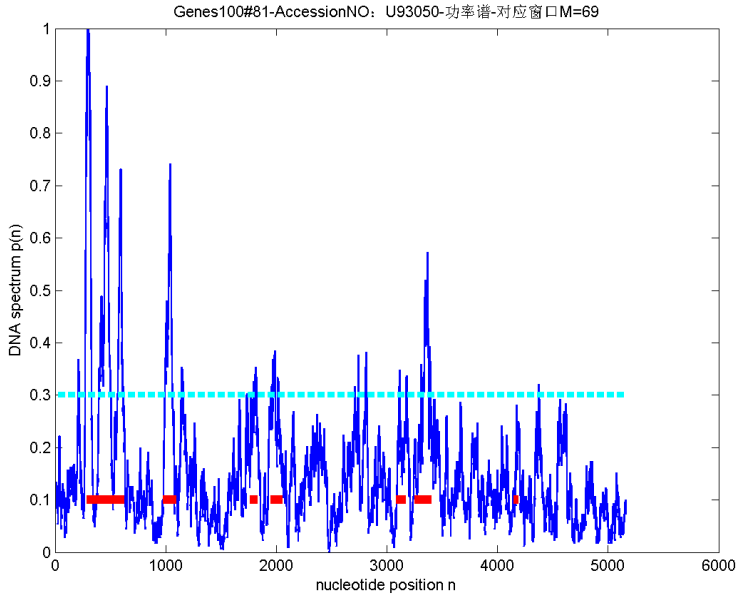


图 5-10　Genes100#81AccessionNO:U93050 窗口长 M=69 时的频谱图

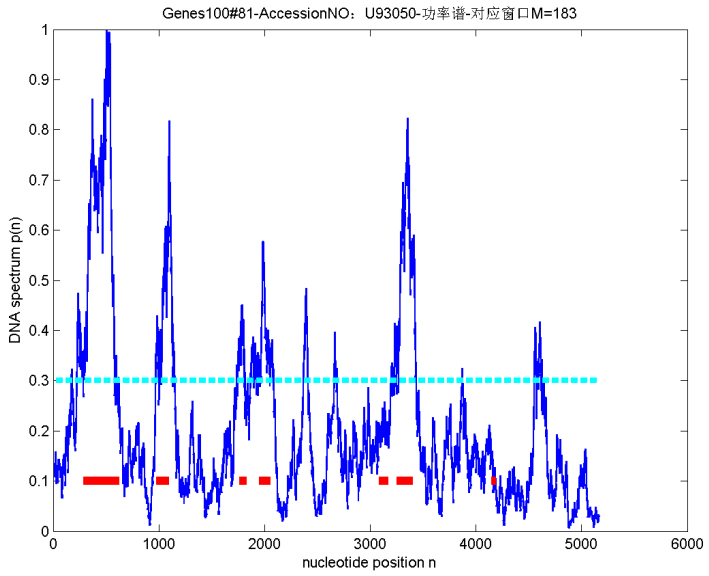


图 5-11　Genes100#81AccessionNO:U93050 窗口长 M=183 时的频谱图

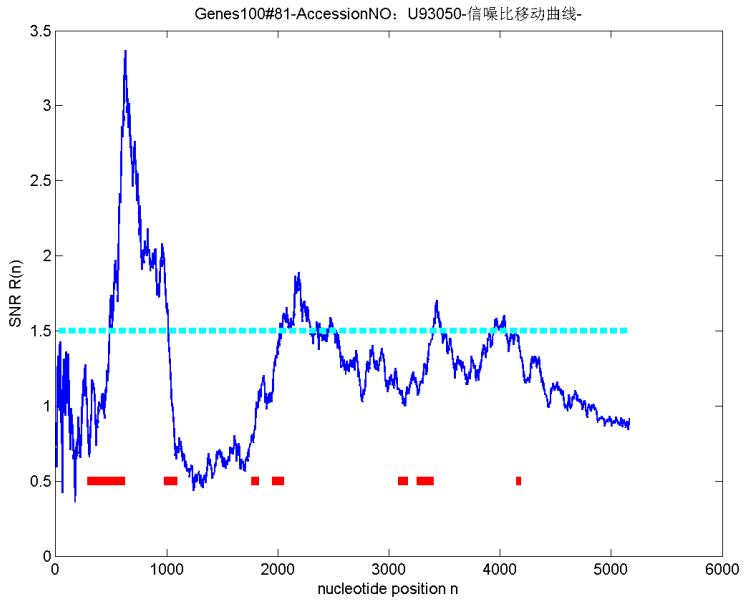


图 5-12　Genes100#81AccessionNO:U93050 信噪比

从图 5-10Genes100#81AccessionNO:U93050 窗口长 M=69 时的频谱图可知并不能很清晰的判断出外显子位置。改变滑动窗口宽度 M，重新计算器频谱曲线。从图 5-11Genes100#81AccessionNO:U93050 窗口长 M=183 时的频谱图可知，此时已经可以较清晰的进行统计，同时以图 5-12Genes100#81AccessionNO:U93050 信噪比来辅助判断，由此可以统计相关结果，统计结果如表 5-3 所示。

表 5-3　Genes100#81AccessionNO:U93050 统计结果

序号	窗口长度 M	已知编码序列数	正确探测的编码序列数	误探测的编码序列数	缺失的编码序列数	探测率	缺失率	正确率	错误率
81	183	7	5	3	2	0.714285714	0.285714286	0.625	0.375

3、改变滑动窗口也无法准确进行统计的序列

采用序列数据 Genes100#41AccessionNO:U84903 计算相应的频谱曲线和信噪比移动曲线，Genes100#41Accession NO:U84903 窗口长 M=69 时的频谱图如图 5-13 所示，Genes100#41Accession NO:U84903 窗口长 M=129 时的频谱图如图 5-14 所示，Genes100#41Accession NO:U84903 信噪比如图 5-15 所示。

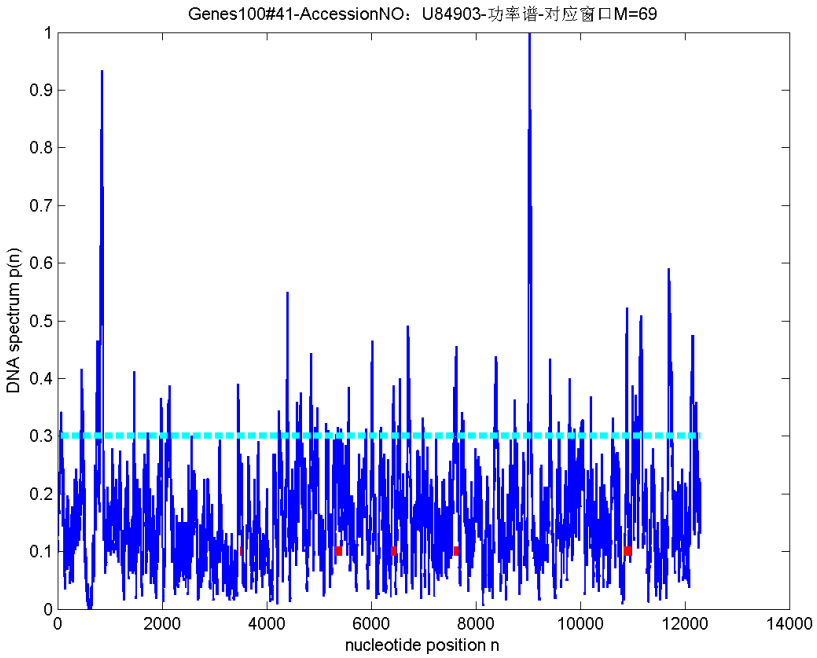


图 5-13　Genes100#41Accession NO:U84903 窗口长 M=69 时的频谱图

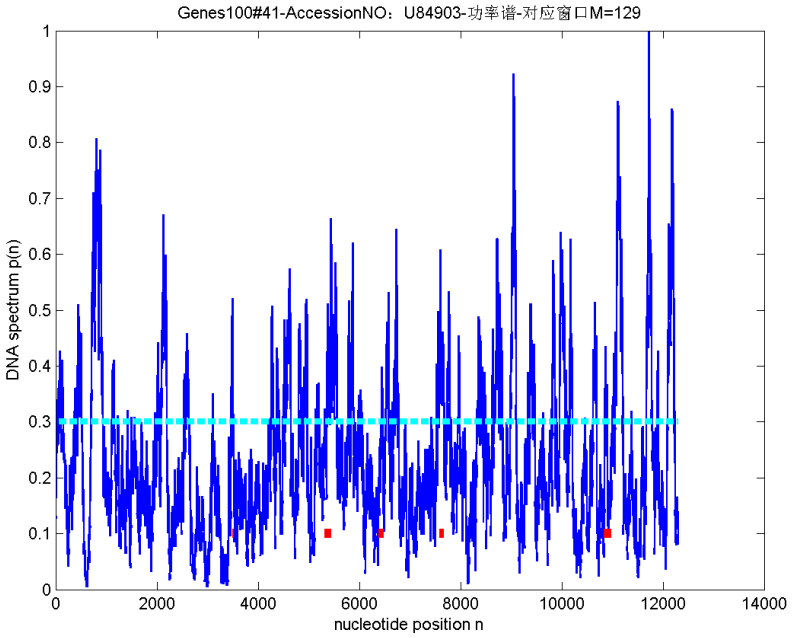


图 5-14　Genes100#41Accession NO:U84903 窗口长 M=129 时的频谱图

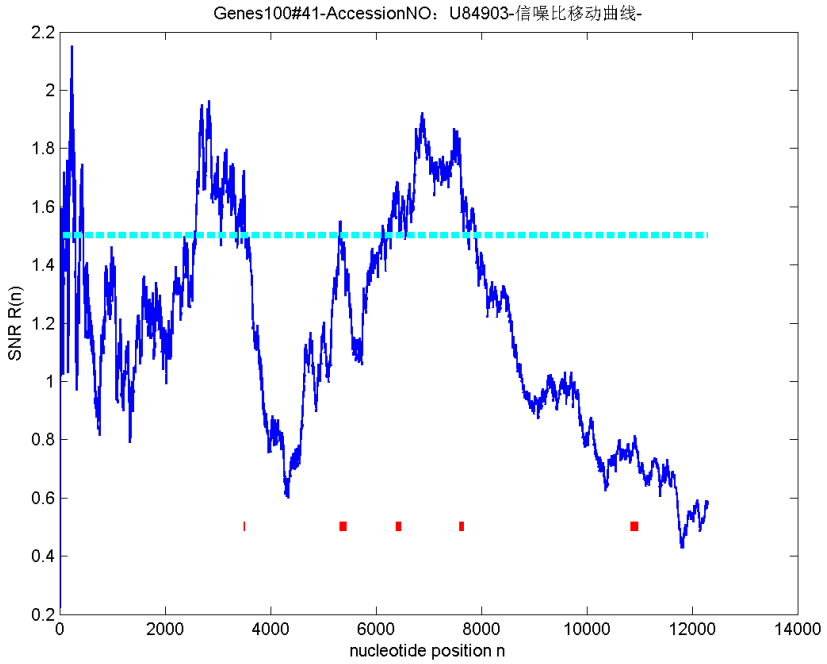


图 5-15　Genes100#41Accession NO:U84903 信噪比

从图 5-13Genes100#41Accession NO:U84903 窗口长 M=69 时的频谱图可知并不能很清晰的判断出外显子位置。改变滑动窗口宽度 M，重新计算器频谱曲线。从图 5-14Genes100#41Accession NO:U84903 窗口长 M=129 时的频谱图可知，此时同样不能很清晰的判断出外显子位置。这类基因序列数据将不进行统计分析。

5.3 对预测结果的评估

根据以上统计规则对 Genes100.mat 和 genes200.mat 进行统计分析，结果如表 5-4Genes100 统计结果和表 5-5genes200 统计结果所示。

注：表中的序号对应所在数据组中的位置。

表 5-4Genes100 统计结果中给出了 10 组统计数据，总共有 72 组数据。同样地，表 5-5genes200 统计结果中给出了 10 组统计数据，总共有 153 组数据。详细数据见附件 genes100 统计.xls 及 genes200 统计.xls。从统计数据可知，对 Genes100 中的人与小鼠的基因序列数据，设置频谱阈值为 0.3 是合理的，此时探测率约为 86%，正确率约为 91.3%。对 genes200 中的哺乳动物类的基因序列数据，设置频谱阈值为 0.25 是合理的，此时探测率约为 82.7%，正确率约为 71.2%。

表 5-4　Genes100 统计结果

序号	窗口长度 M	已知编码序列数	正确探测的编码序列数	误探测的编码序列数	缺失的编码序列数	探测率	缺失率	正确率	错误率
1	255	7	5	2	2	0.714285714	0.285714286	0.714285714	0.285714286
2	645	1	1	0	0	1	0	1	0
3	129	7	7	1	0	1	0	0.875	0.125
4	603	3	2	0	1	0.666666667	0.333333333	1	0
6	1563	1	1	0	0	1	0	1	0
7	1143	1	1	0	0	1	0	1	0
8	129	4	3	1	1	0.75	0.25	0.75	0.25
9	1003	2	1	0	1	0.5	0.5	1	0
11	429	1	1	0	0	1	0	1	0
12	255	4	2	1	2	0.5	0.5	0.666666667	0.333333333
平均值						0.860213783	0.167563995	0.913254578	0.114523208

表 5-5　genes200 统计结果

序号	窗口长度 M	已知编码序列数	正确探测的编码序列数	误探测的编码序列数	缺失的编码序列数	探测率	缺失率	正确率	错误率
1	183	3	3	1	0	1	0	0.75	0.25
2	603	6	4	4	2	0.6666667	0.3333333	0.5	0.5
3	813	6	1	1	5	0.1666667	0.8333333	0.5	0.5
5	255	4	3	3	1	0.75	0.25	0.5	0.5
6	813	2	1	0	1	0.5	0.5	1	0
7	813	2	2	0	0	1	0	1	0
8	255	3	3	4	0	1	0	0.4285714	0.5714285
9	255	10	9	4	1	0.9	0.1	0.6923077	0.3076923
10	255	8	7	4	1	0.875	0.125	0.6363636	0.3636364
11	129	3	2	3	1	0.6666667	0.3333333	0.4	0.6
平均值						0.8279886	0.1720113	0.7119536	0.2880463

六、问题3分析及模型建立与求解

6.1 问题 3 分析

直接傅立叶变换存在的主要问题：对一个长的DNA序列进行直接FFT变换得到频谱图，观察N／3处是否有峰值存在，这种情况下，我们只能推测出这段序列中包含能够编码蛋白质的外显子区，但却不能精确定位外显子所在的位置。

前面所提的快速算法，由序列的傅立叶变换可以得到该序列的功率谱，但可看到图中有很大的背景噪声，在我们需要观察的峰值旁边有较大的其他次峰值存在，这是因为待测序列存在长程 l/f 相关，容易对 1/3 峰值的确定造成干扰。由于 DNA 序列随机噪声的影响等原因，还很难“精确地”确定基因外显子区间的两个端点。

虽然也是在傅立叶变换的基础上进行分析，但在变换的过程中却能根据待测信号的不同做出相应的调整。小波变换采用加窗分析的思想，所使用的窗口可以根据所测信号的不同进行改变。理想的窗函数在时域和频域分析应该满足的形式，即当我们面对的非平稳信号中包含很多高频和低频部分时，我们希望所使用的分析窗口可以随着待测信号的频率变化调整大小，小波变换就是在这种要求下提出的。

小波变换是一种新的变换分析方法，具有多分辨分析的特点，被誉为分析信号的显微镜。小波变换在一定的滤波尺度下可有效地除去随机涨落引起的高频噪声。所以我们将小波变换引入到 DNA 序列编码区的预测中，建立基于小波变换的 DNA 序列编码区预测方法。

小波变换的具体实现是通过窗口 $\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})$ 对信号进行变换

$$W_{\psi}f(a,b)=\frac{1}{\sqrt{|a|}}\int_{-\infty}^{+\infty}f(t)\psi(\frac{t-b}{a})dt \qquad (6\text{-}1\text{-}1)$$

式中，b为时间平移参数，通过它来确定小波函数的中心位置，$a\in R,a\neq 0$，a为尺度参数，由它确定了小波函数的中心位置。正是由于引入这两个参数，才使得小波函数的窗口实现可调。

当小波满足窗口条件和容许性条件时，可以得出小波的均值为0，且是具有带通性质的窗口，通过对小波函数分析，在平移之前分析窗所占的矩形面积是 $[t^{*}\pm\frac{\Delta\varphi}{2},\omega^{*}\pm\frac{\Delta\hat{\varphi}}{2}]$，而经过平移之后分析窗所占的矩形区域为：

$$[at^{*}\pm\frac{a\Delta\varphi}{2},\frac{\omega^{*}}{a}\pm\frac{\Delta\hat{\varphi}}{2a}] \qquad (6\text{-}1\text{-}2)$$

首先选择一个小波基，使其与待测信号进行积分，得出一个该时段的相关系数C。该系数C可以理解为所该区间的待测信号与所选小波基在波形的相似评价。然后调整平移因子b（图6-2），使小波基与不同时段的待测信号进行积分，直到该小波滑过整个待测信号的时段。重新调整小波的伸缩因子a（图6-3），也就是改变

小波的时频分辨率，重复上面的过程。通过不断变换伸缩因子a，就可以完成不同小波基下待测信号的小波系数。

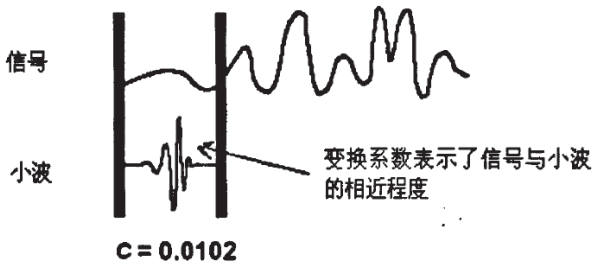


图6-1　计算小波变换系数示意图

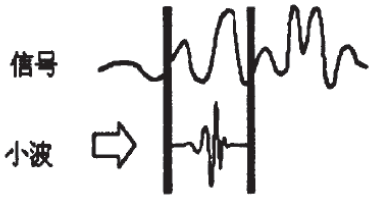


图6-2　向右平移小波

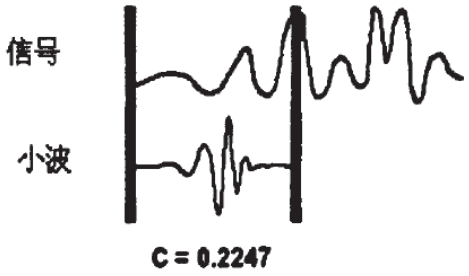


图6-3　调整参数a，重新计算小波变换系数

由小波变换的定义式，有

$$\begin{aligned} W_f(a,b) &= \left\langle f(t),\psi_{a,b}(t)\right\rangle = \int_{-\infty}^{+\infty} f(t)\psi_{a,b}^{*}(t)dt \\ &= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})dt \qquad (a>0, f\in L^2(R)) \end{aligned} \tag{6-1-3}$$

其中$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})$，并设$f(t)=f(k\Delta t), t\in(k,k+1)$，则

$$\begin{aligned} W_f(a,b) &= \sum_k \int_k^{k+1} f(t)|a|^{-1/2}\psi^{*}(\frac{t-b}{a})dt \\ &= \sum_k \int_k^{k+1} f(k)|a|^{-1/2}\psi^{*}(\frac{t-b}{a})dt \\ &= |a|^{-1/2}\sum_k f(k)\left(\int_{-\infty}^{k+1}\psi^{*}(\frac{t-b}{a})dt-\int_{-\infty}^{k}\psi^{*}(\frac{t-b}{a})dt\right) \end{aligned} \tag{6-1-4}$$

小波变换可以通过以上来实现。

小波函数的功率谱为$\left|W_f(a,b)\right|^2$，则其平均功率为

$$W_f(a)=\frac{\int_{b_0}^{b_1}\left|W_\psi f(a,b)\right|^2 db}{b_1-b_0} \tag{6-1-5}$$

式中，b_0和b_1是指伸缩因子为a时窗口的左右两个端点。

小波变换的原理就是通过小波基函数与所分析信号进行相似性比较，所以小波基函数选择的好坏直接决定了我们能否很好地观测到待测信号不同的频率分量。在实际分析的过程中应该根据待测信号的一些特性来选择小波。

这里我们选用了Mexico草帽小波，通过变换伸缩因子a得到不同尺寸下的小波系数，然后观察该窗口中的频谱。

$$\psi(t)=\frac{2}{\sqrt{3}}\pi^{-1/4}(1-t^2)e^{-t^2/2} \tag{6-1-6}$$

波形如下图所示：

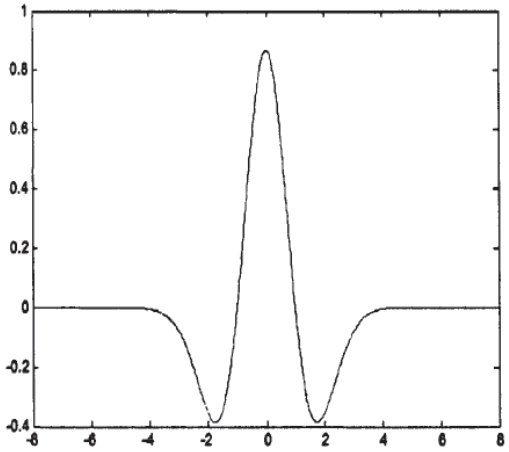


图6-4　墨西哥草帽小波波形图

6.2 问题3建模求解、分析

如图6-5所示为小波变换来处理DNA序列的基本流程，本问题的整体解决思路如图6-6采用Z-curve映射滑动窗口的小波变换法的计算流程图所示，首先验证采用小波变换法可以很好的解决DNA序列随机噪声的影响，可以较“精确地”确定基因外显子区间的两个端点。采用敏感性和专一性来评价该方法的准确率。

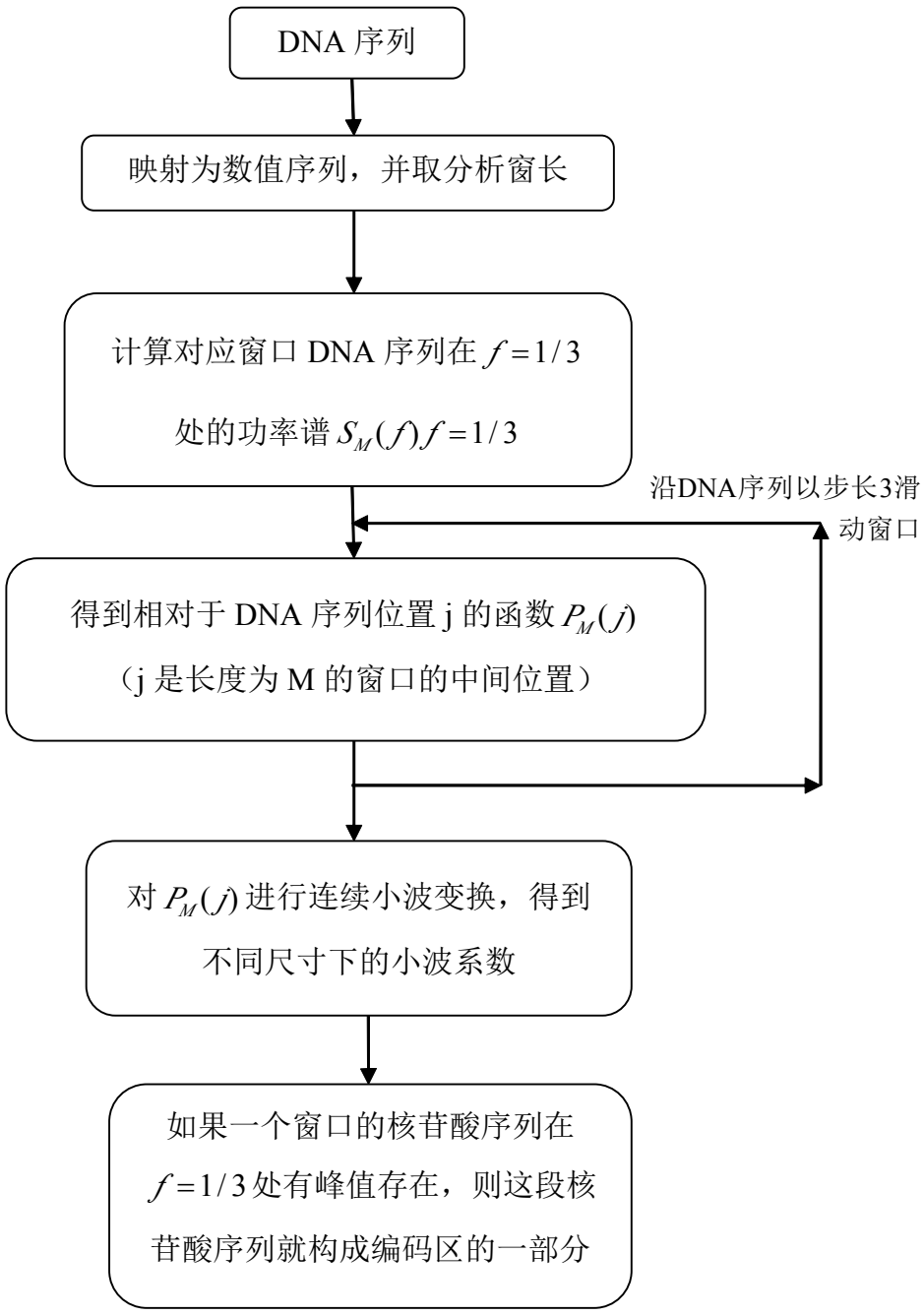


图6-5　使用小波变换来处理DNA序列流程

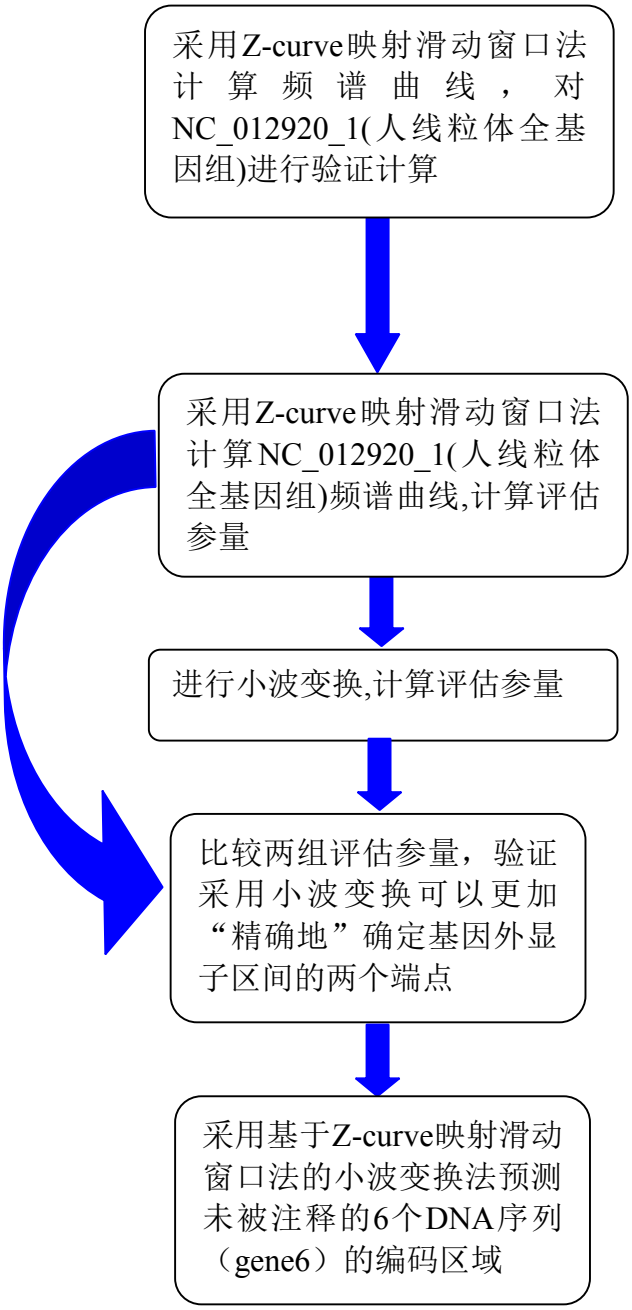


图 6-6　采用 Z-curve 映射的计算流程图

首先验证小波变换能更精确地确定基因外显子区间的两个端点。

采用 Z-curve 映射滑动窗口法直接计算频谱曲线，这里窗口宽度值是基于最小外显子宽度选取了“最佳”窗口宽度值。固定滑动窗口长度快速算法计算出的频谱图如图 6-7 所示，以此为基础采用小波变换计算频谱图如图 6-8 所示。

注：从图 6-7 到图 6-8，粗虚线代表阈值，一段段的粗实线是已知外显子位置，实曲线为频谱或信噪比曲线。

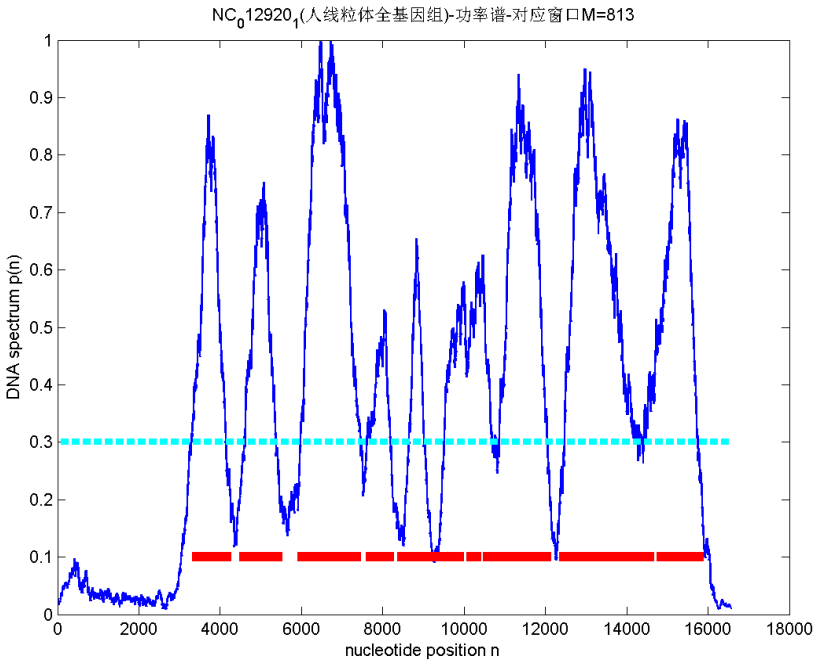


图 6-7　固定滑动窗口长度快速算法计算频谱图

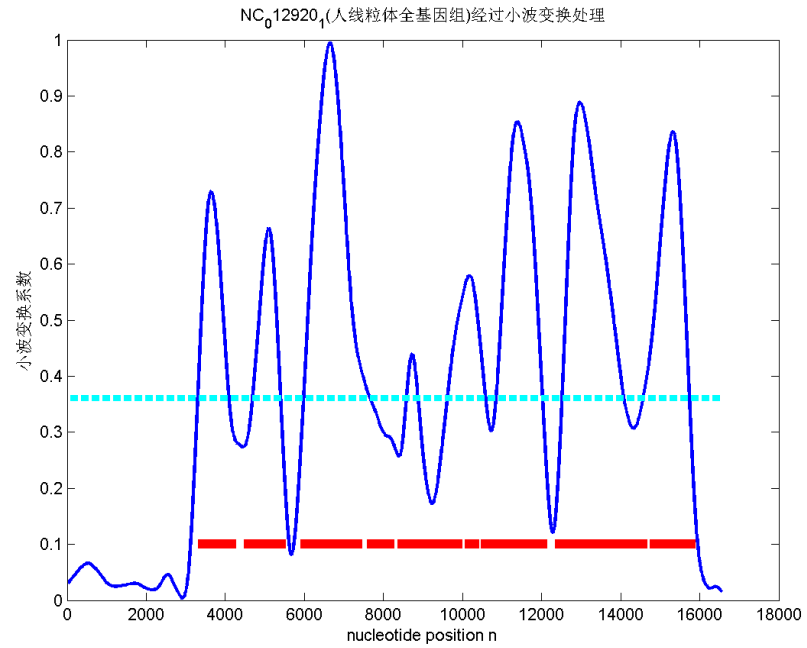


图 6-8　采用小波变换计算频谱图

对图 6-7 中固定滑动窗口长度快速算法计算频谱图和图 6-8 中采用小波变换计算频谱图进行计数，分别计算出 T_p、T_N、F_p、F_N的数目，进而计算出敏感性和专一性参数。由表 6-1 中结果可知，采用小波变换法后，可以有效的消除 DNA 序列随机噪声的影响，可以更“精确地”确定基因外显子区间的两个端点。接下来将采用这种方法来预测 gene6 中 6 个未被注释的 DNA 序列的编码区域。

表 6-1　固定滑动窗口长度快速算法与小波变换计算频谱对比表

统计	NC_012920_1(人线粒体全基因组)	总 16569
	滑动窗口	小波变换
T_p	9687	10107
T_N	5164	5178
F_p	74	60
F_N	1644	1224
S_n	0.8549	0.8919
S_p	0.9924	0.9941

下面以 gene6.mat 中第一个 DNA 序列为例来说明预测过程，其他 DNA 序列采用相同的方法来预测。gene6.mat 中第一个 DNA 序列的计算结果如图 6-9 直接计算频谱曲线和 6-10 小波变换计算频谱曲线所示。根据此计算结果可以得到每个 DNA 序列的编码区域。

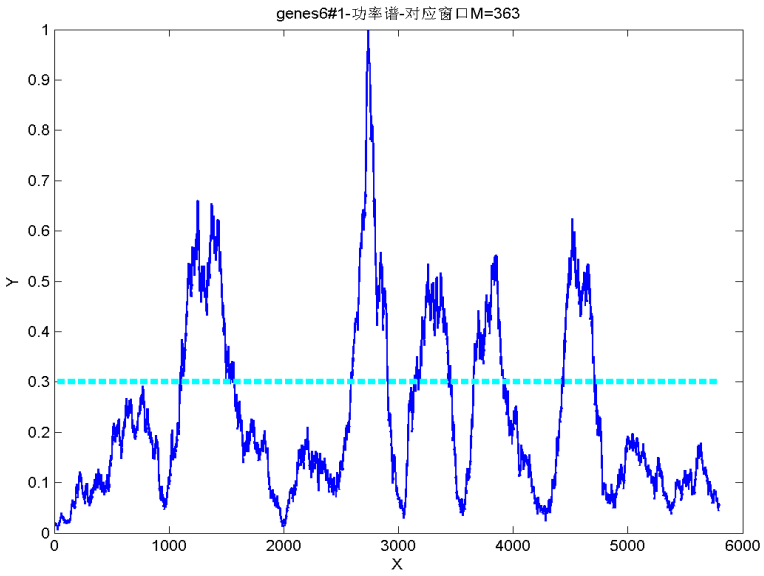


图 6-9　genes6 直接计算频谱曲线

编码区域预测结果如表 6-2 所示，表中#1 表示 DNA 序列的编号，总共 6 个，M 为相应的最佳滑动窗口宽度值，lev 是 Matlab 中和小波变换有关的参数，数值区域里每一行的两个数据表示一段基因外显子区间。

此方法的优点：

1、计算量小，计算速度快：因为在加窗的傅里叶变换中，仅计算了一个频点的频谱，当序列结构复杂而窗口长度又小时，效果越明显。

2、使用小波变换通过变换伸缩因子来观察傅里叶变换后的不同频率分量，放大我们想要的频率分量，抑制背景噪声，使周期特性更加明显。

但是这种方法也存在缺陷：就是当所测序列的长度不是3的整数倍时，得到的频谱会有很大的噪声，甚至彻底将周期性淹没。

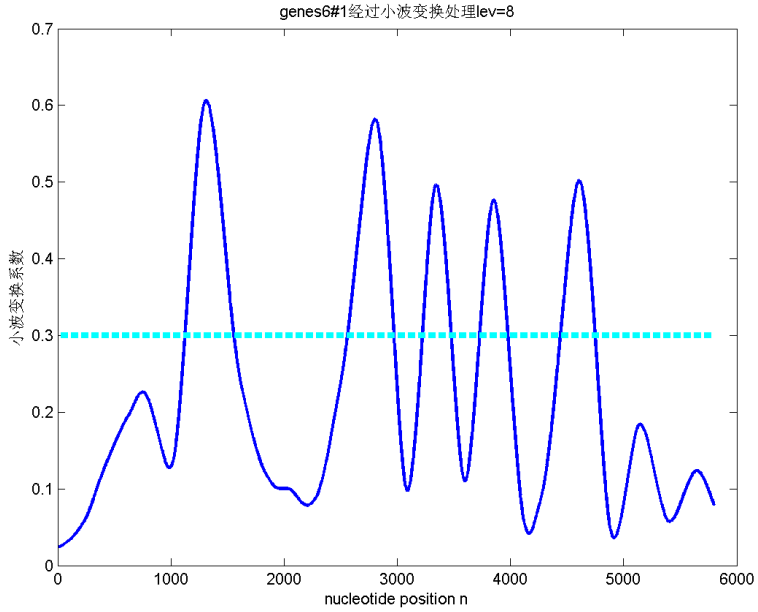


图 6-9　genes6 小波变换计算频谱曲线

表6-2 编码区域预测结果

DNA序列编号	起始点	结束点	DNA序列编号	起始点	结束点
#1	M=363	lev=8	#4	M=903	lev=10
	1124	1555		1145	1716
	2560	2970		2493	3740
	3223	3475		4734	5836
	3729	3976	#5	M=513	lev=10
	4442	4747		2910	3321
#2	M=813	lev=8		7873	8567
	1325	1748	#6	M=723	lev=8
	4035	5101		596	1262
	5508	6375		1384	1951
#3	M=2103	lev=9		2239	3272
	228	1082		4196	4662
	2295	2689			
	3670	5255			

七、问题4分析及模型建立与求解

7.1 短编码序列识别分析

（1）基于多种特征量的基因识别方法

改进傅立叶方法虽然识别性能有一定改善, 但是提高的精度并不是很大, 特别当序列长度小于130bp时预测精度不高, 其中一个主要原因是傅立叶变换只能探测序列的周期特性, 例如周期3性质, 但对于比较短的序列, 周期性不容易探测到, 而且对于复杂的DNA序列, 在识别算法中只考虑周期性也是不全面的。

大多数基因识别算法的核心是编码测量(coding measures), 即对于一定长度的序列, 计算出一些跟蛋白质编码功能相关的特征向量。一般可以用这些蛋白质编码区的特征统计量建立外显子的数学模型。常用的编码测量包括密码子使用, 碱基成分以及序列的傅立叶变换等。

1. 密码子使用：序列中 64 种密码子的使用频率；
2. 组成成分：每种碱基在序列中所占的百分比；
3. 六方体；给定长度的序列中所有六方体出现的频率；
4. 阅读框：给定长度的序列中最长密码子区域；
5. 信息熵：将碱基序列看作信息流, 定义序列的信息熵；
6. 相关性：根据每种碱基在序列中出现的位置定义相关性；
7. 傅里叶分析：对序列进行傅里叶变换, 把傅里叶系数作为特征量；
8. 序列周期性：DNA 序列具有多种周期性, 例如周期 3, 周期 10-11 等

都可以作为基因序列的特征量。

可结合基因序列的多种特性, 比如碱基组成成分, 碱基位置相关性, 周期3特性和密码子使用偏好性等, 进一步提高较短序列基因识别算法的精度。实现一种基于多种特征量的基因识别算法, 能够有效提高长度小于90bp的基因序列的预测精度。

（2）AR 模型法

参数模型法是现代谱估计技术的主要内容，包括IvIA(moving-average)模型、AR(auto-regressive)模型和ARMA模型，其目标是旨在努力改善谱估计的分辩率, 并提高其平滑性。ARMA模型虽然综合了MA模型和AR模型的特征，是一个极-零模型，易于反映功率谱中的峰值和谷值，但考虑到预测蛋白编码区问题的需要, 以及计算模型参数的可实现性，本文引入AR模型谱估计来预测与定位DNA序列中的蛋白编码区。

参数模型法的思路是：

1、假定所研究的过程 $x(n)(n=0,1,\cdots,N-1)$ 是由一个输入 $u(n)$ 激励一个线性系统 $H(z)$ 的输出，如下图所示：

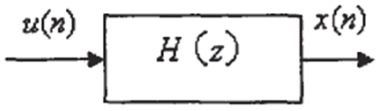


图7-1　参数模型

2、由已知的 $x(n)$，或其自相关函数 $r_x(m)$ 来估计 $H(z)$ 的参数。

3、由 $H(z)$ 的参数来估计 $x(n)$ 的功率谱。

根据信号与系统理论[9]，不论 $x(n)$ 是确定性信号还是随机信号，对于上图的线性系统，$u(n)$ 和 $x(n)$ 之间总有如下的输入与输出关系：

$$x(n)=-\sum_{k=1}^{p}a_k x(n-k)+\sum_{k=0}^{q}b_k u(n-k) \tag{7-1-1}$$

对上式两边取Z变换，并假定 $b_0=1$，可得

$$H(z)=\frac{1+\sum_{k=0}^{q}b_k z^{-k}}{1+\sum_{k=1}^{p}a_k z^{-k}} \tag{7-1-2}$$

为了保证 $H(z)$ 是一个稳定的且是最小相位系统，$H(z)$ 的零点和极点都应在单位圆内。

如果 $b_1,b_2,\cdots,b_q$ 全为0，则

$$x(n)=-\sum_{k=1}^{p}a_k x(n-k)+u(n) \tag{7-1-3}$$

$$H(z)=\frac{1}{1+\sum_{k=1}^{p}a_k z^{-k}} \tag{7-1-4}$$

上两式给出的模型称为自回归模型，简称AR模型，它是一个全极点模型，其现在的输出是现在的输入和过去p个输出的加权和。

假定 $u(n)$ 是一个方差为 σ^2 的白噪声序列，由随机信号通过线性系统的理论可知[10]，输出序列 $x(n)$ 的AR模型功率谱为：

$$S_x(e^{j\omega})=\frac{\sigma^2}{\left|1+\sum_{k=1}^{p}a_k e^{-j\omega k}\right|^2} \tag{7-1-5}$$

从上式可知，要估计出 $x(n)$ 的AR模型功率谱，主要是计算模型参数 $\{a_k\}$ 和方差 σ^2。常用的模型参数计算方法包括：Yule-Walker法、Burg法，改进的协方

差法、无拘束最小均方法以及序列估计法等。文献[11]对这些方法均进行了全面的介绍。

与周期图法比较，AR模型估计出的功率谱有一系列好的性质，包括：

（a）AR谱比周期谱平滑得多；

（b）周期图谱估计的分辨率反比于信号的长度，而AR谱估计的分辨率不受信号长度的限制，其分辨率比周期图谱估计的分辨率高；

（c）周期图法估计的功率谱存在“栅栏效应”，这种效应可能屏蔽一些有用的弱信号；在整个频率范围内，AR模型估计的功率谱和真是谱相跟随，二者存在较好的匹配性，特别是在峰点的跟随程度更高。

AR模型谱估计的不足有：

（a）“谱线分裂”现象，即在本来应只有一个谱线的位置附近分裂成两个谱线，因此，产生虚假信号。

（b）AR谱估计的质量受到阶次p的影响：p选得过低，谱太平滑，反映不出谱峰；p选得过大，可能会产生虚假的峰值。

大量研究表明，通过算法的改进和阶次的合理选择可以克服AR模型谱估计的上述缺陷。由于AR模型易于反映功率谱中的峰值，因此，它适合应用于DNA序列中蛋白编码区的预测问题。

AR模型法预测DNA序列中蛋白编码区的重要依据仍然是其周期3-性质。因此，利用AR模型估计出基因组DNA序列的功率谱，再通过 $f=1/3$ 处是否存在峰值来判断该序列是否是蛋白编码序列。预测具体步骤如下：

1、将DNA符号链映射成为DNA数值序列，可采用Voss映射或Z-Curve映射等等。

2、合理选择模型的阶次，可以有效地提高谱估计质量。在不发生谱分裂现象的前提下，AR模型的阶次应该尽可能地选择高，以确保获得最大的分辨率。

3、根据已知DNA数值序列，求解AR模型参数。建议采用Burg法和改进的协方差法来求解DNA序列的AR模型参数。

4、估计DNA序列的AR功率谱。利用第三步计算出的AR模型参数，根据前面的公式估计DNA序列的功率谱 $S_x(f)$。

5、定义信噪比，确定信噪比阈值，预测蛋白编码区。本文定义DNA序列的AR谱在 $f=1/3$ 处的功率谱幅度 $S_x(1/3)$ 为信号，其AR谱在整个频率范围内的平均值 $\bar{S}_x$ 为噪声，即：

$$\bar{S}_x=\frac{1}{N}\sum_{k=1}^{N}10*\lg[S_x(k)] \qquad (7\text{-}1\text{-}6)$$

其中，N为DNA序列的长度，信噪比定义为：

$$P=\frac{10*\lg[S_x(1/3)]}{\bar{S}_x}(dB) \qquad (7\text{-}1\text{-}7)$$

当P大于一定的阈值时，判断这段DNA序列为蛋白编码序列，否则，判断这段DNA序列为非编码区。明显地，P值越大，预测方法的敏感性越高，敏感性越

高，则预测蛋白编码区结果的准确性就越高。

确定P的阈值通常需要知道DNA序列的先验生物特征。

综上所述，在小尺度序列的情况下，只有AR模型法能够有效的预测DNA序列中的蛋白编码区，而FFT法失效。

7.2 基因突变分析

基因突变包括 DNA 序列中单个核苷酸的替换，删除或者插入等，仍可利用频谱或信噪比方法去发现基因编码序列可能存在的突变。

采用Genes100.mat中数据#9-AccessionNO：AF037313分别研究了DNA序列中加入、删除和替换对应的频谱变化。加入单个核苷酸对应的频谱曲线如图7-2所示，删除单个核苷酸对应的频谱曲线如图7-3所示，替换单个核苷酸对应的频谱曲线如图7-4、图7-5所示。由图7-2加入单个核苷酸对应的频谱曲线和图7-3删除单个核苷酸对应的频谱曲线可以明显看出，加入和删除单个核苷酸会导致其频谱发生较大的变化。如图7-4、图7-5所示，替换单个核苷酸对应的频谱曲线变化不太明显，但是经放大后仍然可以分辨出DNA序列发生了变化。

因此，根据本文所做的初步分析可知，利用频谱或信噪比方法去发现基因编码序列可能存在的突变在理论上是可行的。当然，具体的处理方法还有待进一步研究。

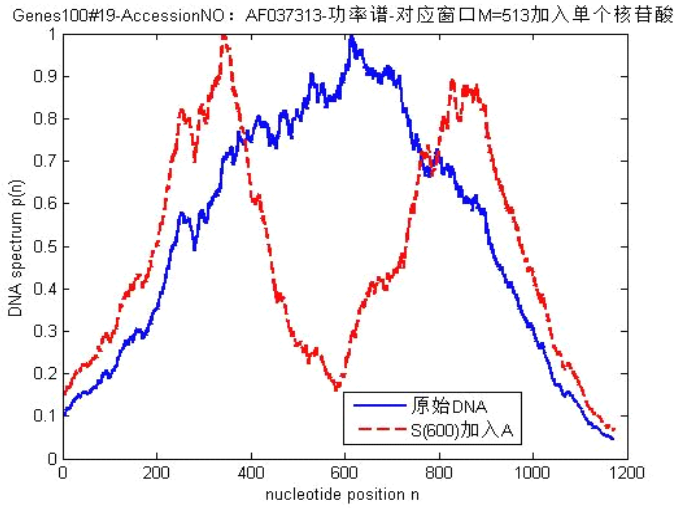


图 7-2　加入单个核苷酸前后频谱对比图

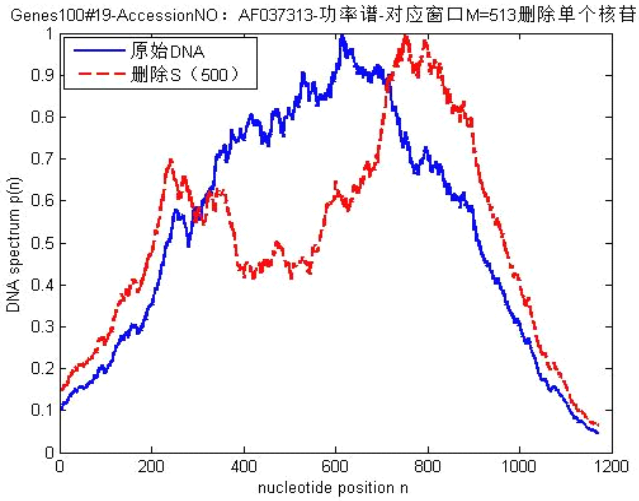


图 7-3　删除单个核苷酸前后频谱对比图

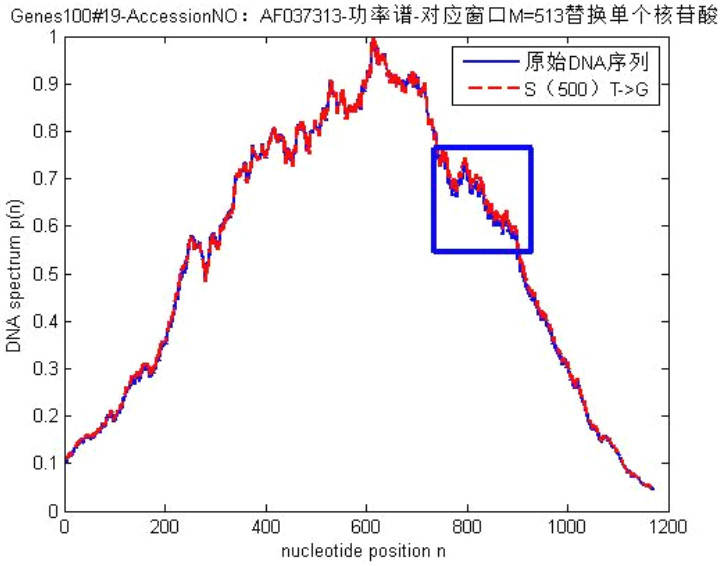


图 7-4　　替换单个核苷酸前后频谱对比图

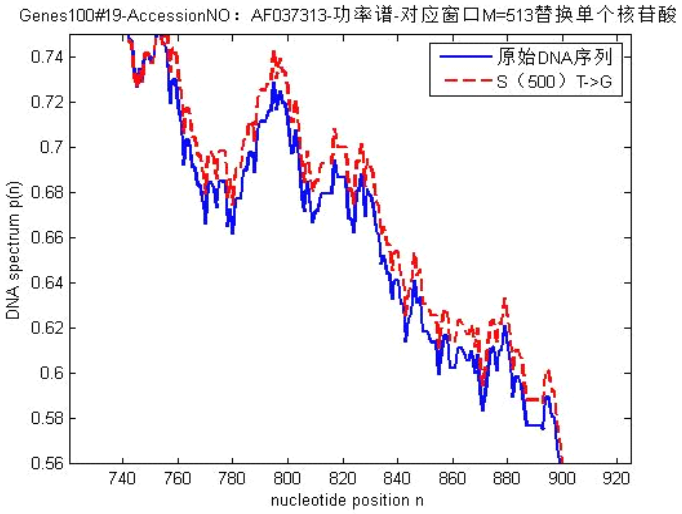


图 7-5　替换单个核苷酸前后频谱对比局部图

八、 总结与展望

本文在准确分析题目的基础上，主要完成了以下工作：

1、直接采用核苷酸出现频数计算信噪比和 N/3 处的功率谱，不需要进行 DFT 运算；并且不要求 DNA 序列长度 N 是 3 的倍数。通过理论分析和实验证实了 Voss 映射下该方法优于 DFT 算法。

2、通过严密的理论推导得出 Z-curve 映射的频谱与信噪比和 Voss 映射下的频谱与信噪比之间呈线性关系，频谱间满足 $P_Z[k]=\begin{cases}4P[k], & k\neq 0\\ 4P[0]-N^2, & k=0\end{cases}$，信噪比间满足 $\frac{R_z}{R_{voss}}=\frac{4}{3}$。

3、基于改进的核苷酸频数计算方法，提出了实数映射的功率谱与信噪比快速计算公式。

4、基于统计方法，采用一种判别正确率较高的阈值确定最优化方法得到了信噪比阈值确定方法，并对多个实例得出了最优阈值。

5、引入统计分析的概念，以探测率、缺失率、正确率、错误率等参数作为评价指标，对题目提供的大量基因序列数据 Genes100 和 genes200 做了统计分析，最终得到相应的频谱阈值。从统计数据可知，Genes100 中的人与小鼠的基因序列数据对应频谱阈值为 0.3，此时探测率约为 86%，正确率约为 91.3%；genes200 中的哺乳动物类的基因序列数据对应频谱阈值为 0.25，此时探测率约为 82.7%，正确率约为 71.2%。

6、基于对问题 1、2 的研究，创造性的提出了基于 Z-curve 映射可变滑动窗口的小波变换法，来实现基因识别。该方法较好的解决了 DNA 序列随机噪声对基因识别造成的影响，能够较精确地确定基因外显子区间的两个端点。在对 NC_012920_1（人线粒体全基因组）（NC_012920_1.fasta）验证的基础上，采用了这种识别算法预测了题目给定的 6 个未被注释的 DNA 序列（gene6）的编码区域。

7、总结了识别基因编码序列的多种特征指数。

8、采用 Genes100.mat 中数据#9-AccessionNO：AF037313 研究了 DNA 序列中加入、删除和替换对应的频谱变化。

关于如何准确的识别基因编码序列，很多新颖的方法不断涌现，由于专业背景知识不足和时间关系，还有诸多问题尚待解决。文中采用基于 Z-curve 映射可变滑动窗口的小波变换法精确识别基因序列，此处的可变滑动窗口并不是随信号实时可变的，因此还可以进一步优化为随时可变滑动窗口。

此外，文中采用的所有方法仍是基于 3-周期特性，当部分编码序列（外显子）尤其是较短的编码序列，就可能不具备这一特征，因此仅仅采用这一特征指数还不能较好的实现精确基因识别。目前，基于多特征识别的研究还不是很完善，有待在今后的学习中进一步研究。

文中针对基因突变问题，仅做了最基本的分析，发现 DNA 序列中单个核苷酸的替换、删除或者插入会导致频谱特性发生很大变化，相关问题还需要今后进一步的深入研究。

参考文献

[1] VOSS R F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences[J]. Physics Review Letter, 1992, 68(25) : 3805 -3808.

[2] 吴镇扬 数字信号处理 北京: 高等教育出版社, 2004

[3] Jianfeng Shao, Xiaohua Yan, Shuo Shao SNR of DNA sequences mapped by general affine transformations of the indicator sequences[J]. Math. Biol. July 2012

[4] Fickett J W. The gene identification problem: An overview for developers[J]. Comp. Chem., 1996, 20: 103-118.

[5] T.A.布朗著, 袁建刚, 周严,强伯勤译 基因组 科学出版社, 2002

[7] WangZ, Chen Y, Li Y. A brief review of computational gene prediction methods [J]. Geno Prot Bioinfo, 2004, 2:4-8

[8] Burset M, Guigo R. Evaluation of gene structure prediction programs[J]. Genomics, 1996, 34:353-367.

[9] A.S. Oppenheim, A. S. Willsky, S. H. Nawab. Signals and Systems(Second Edition). 北京: 电子工业出版社, 2002

[10]张旭东 离散随机信号处理 北京: 清华大学出版社, 2005

[11]J．G Proakis, D. G Manolakis. Digital Signal Processing. 2 edition, Macmillan Publishing Company, 1992: 886-896