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

针对基因识别问题，本文基于 DNA 序列的 3 周期这一性质，首先给出了 DNA 序列功率和信噪比的快速算法并讨论了不同物种基因类型的阈值确定方法；在此基础上，本文建立了基于背景噪声抑制和频谱平滑的 SNR 频谱预处理模型，经过预处理后的频谱不仅大幅度抑制了背景噪声，同时保留了 SNR 频谱的模式特征。在编码序列识别上，本文对经典的 EPND 预测算法进行了改进，使用改进的 EPND 算法对经过预处理后频谱进行基因识别，实验结果显示这种基因识别模型具有优异的基因识别性能，比传统直接使用基于滑动窗口 DFT 的 EPND 识别算法在敏感度、特异性等评价指标上提高了 2%-12%左右(不同指标提高程度不同)；最后本文将提出的基因识别模型用于 6 个未知 DNA 序列(gene6)的编码区预测。

问题 1：针对 Voss 映射，本文首先给出了使用快速 Fourier 变换计算功率谱的快速算法，其时间复杂度为$O(N\log N)$，再对 N/3 频率点的功率和信噪比给出了时间复杂度为$O(N)$的快速算法；最后针对实数映射，类似地，本文也给出了时间复杂度为$O(N)$的功率谱和信噪比快速计算公式；另外，对于 Z-curve 和 Voss 映射下的频谱和信噪比的关系，本文从理论和实验两个方面进行了详细地探讨，并得出结论：$P_Z = 4P_I$，$R_Z = \frac{4}{3}R_I$。

问题 2：对于 SNR 阈值的确定，本文给出了均值平均，概率平均和线性分类器三种方法确定 SNR 阈值并对分类错误进行了讨论，给出了分类评价指标，对比了以上三种方法的优劣，并针对外显子判别正确率与外显子序列长度的关

系以及信噪比阈值与物种类型关系做了一系列讨论。

问题 3：由于随机背景噪声等原因的影响，使得很多基因识别算法准确率偏低，鉴于此，本文建立了基于背景噪声抑制和频谱平滑的基因识别 SNR 频谱预处理模型，经过预处理后的频谱不仅大幅度抑制了背景噪声，同时保留了 SNR 频谱的模式特征。在编码序列识别上，本文对经典的 EPND 预测算法进行了改进，使用改进的 EPND 算法对经过预处理后频谱进行基因识别，在评估测试中，这种基因识别模型表现出了优异的基因识别性能；最后对问题中的 6 个未知 DNA 序列(gene6)的编码区进行预测。另外，本文对比分析了滑动窗口长度对频谱影响并提出了合理选取滑动窗口长度的策略；还对比分析了不同编码方式对频谱的影响，这些分析的结论都在一定程度上促进改善了基因识别算法的正确性。

延拓问题，本文主要围绕频谱分析展开，但频谱分析方法存在一些固有缺陷，譬如无法识别极短编码序列。在延拓问题解答中文章给出了几种其他识别基因的特征指数，最后，对于“基因突变”的探测和发现，本文也给出了一些展望性想法。

关键字:基因识别；频谱分析；噪声抑制；EPND；滤波；信噪比

目录

1. 问题总结

DNA是生物遗传信息的载体，其化学名称为脱氧核糖核酸。DNA分子是一种长链聚合物，DNA序列A、G、T、C 四个碱基按一定顺序排列的链状结构。其中带有遗传讯息的DNA片段称为基因。在真核生物的DNA序列中，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。随着世界人类基因组工程计划的顺利完成，通过物理或数学的方法从大量的DNA序列中获取丰富的生物信息，对生物学、医学、药学等诸多方面都具有重要的理论意义和实际价值，也是目前生物信息学领域的一个研究热点。对给定的DNA序列，识别出其中的编码序列(即外显子)称为基因预测，基因预测问题的常用方法一类基于统计学[r],一类基于信号处理与分析[r]。A题中的问题建立在使用信号处理与分析方法进行基因预测的基础上，在利用离散Fourier变换(DFT)对数值化映射后的基因序列进行频谱分析中，DNA序列信号频谱3-周期性被认为是用来区分编码区和非编码区的一个重要特征。基于信号处理与分析方法的基因识别技术还存在一些问题，对“2012全国研究生数学建模竞赛A题”中提出的问题的总结如下：

问题一 功率谱与信噪比的快速算法

a. 针对采用离散 *Fourier* 变换(*DFT*)计算 DNA 序列的功率谱或信噪比计算量较大的问题，对 *Voss* 映射，探求功率谱与信噪比的某种快速计算方法。

b. 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。

c. 针对实数映射，如 $A \to 0, C \to 1, G \to 2, T \to 3$ ，探求其功率谱与信噪比的快速计算公式。

问题二 对不同物种类型基因的阈值确定

对于具有代表性的不同物种基因序列类，研究其阈值确定方法和阈值结果。并分析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。

问题三 基因识别算法的实现

a. 针对现有基因识别算法的不足设计新的基因识别算法探测、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），并适当的分析、评估其准确率。

b. 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。

问题四 延展性研究

a. 对于某些 DNA 序列而言，其部分编码序列（外显子），尤其是短的编码序列，可能不具有频谱或者信噪比显著性，采用单一的频谱或者信噪比作为判别特征会影响基因识别的正确率，针对这一问题提出一些识别基因编码序列的其它特征指数，并对其进行分析。

b. 利用频谱或信噪比方法发现基因编码序列中可能存在的突变。

2. 模型假设

DNA 序列中除了 A、C、G、T 之外，没有符号错误。
DNA 序列中只包含外显子和内含子。

3. 问题分析与求解

3.1. 问题一 功率谱与信噪比的快速算法

3.1.1 问题分析

对于问题一中的a问的分析

在使用*Voss*映射利用离散*Fourier*变换(*DFT*)对数值化映射后的基因序列进行频率谱和信噪比的计算中，*DFT*的复杂度为$O(N^2)$，对于很长的DNA序列使用*DFT*的计算量将会很大，计算功率谱时，一种可能的降低*DFT*复杂度的方法是使用复杂度为$O(N\log N)$的快速*Fourier*变换*FFT*来代替*DFT*。特别地，对于N/3频率点处的功率和信噪比可以找到时间复杂度更优的为$O(N)$的快速算法。

对于问题一中的b问的分析

Z-curve与 Voss是两种不同的数值映射,Voss算法是将1个长度为N的DNA序列表示为4个二进制数字序列；Z-curve算法将DNA序列转换成与其等价的三维表达式。通过对Z-curve映射的分析，可知Z-curve映射与Voss法映射之间存在线性关系，因此可以推断 Z-curve与 Voss两种映射的频谱和信噪比之间可能存在一定比例关系，b问的求解将详细叙述它们之间的这种比例关系。

对于问题一中的c问的分析

对于实数映射，和a问类似，计算其功率谱可以采用时间复杂度为$O(N\log N)$的FFT算法，而对于N/3频率点处的频率和信噪比则可以找到时间复杂度更优的为$O(N)$的快速算法。

基于以上的分析、理解，我们对问题一做了如下的求解。

3.1.2 问题求解

对问题一中 a 问的求解如下:

快速 Fourier 变换(FFT)----复杂度为$O(N\log N)$频谱快速算法

题中给出的基于 *Voss* 映射，计算指定 DNA 序列的功率谱对应的 DFT 变换计算公式如下：

$$U_b[k]=\sum_{n=0}^{N-1}u_b[n]e^{-j\frac{2\pi nk}{N}},\quad k=0,1,\cdots,N-1;b\in I \tag{3-1}$$

其中$I=\{A,T,G,C\}$表示 DNA 序列中的四种核苷酸的符号序列，对于上述公式，使用 FFT 进行演变，过程如下：

用 N 次单位根W_N来表示$e^{-j\frac{2\pi nk}{N}}$，假设待变换序列长度$n=2^r$。根据上面单位根的对称性，求级数$U_b[k]=\sum_{n=0}^{N-1}W_N^{kn}x_n$时，可以将求和区间分为两部分：

$$U_b[k]=\sum_{n=2i}W_N^{kn}u_b[n]+\sum_{n=2i+1}W_N^{kn}u_b[n]=\sum_i W_{\frac{N}{2}}^{ki}u_b[2i]+W_N^K\sum_i W_{\frac{N}{2}}^{ki}u_b[2i+1]$$

如上述公式所示，FFT 将一个 N 点变换就分解成两个 N/2 点变换(分治算法)，计算的复杂度从 DFT 的$O\left(N^2\right)$降为$O\left(N\log N\right)$。可见使用 FFT 能够显著的减少计算 DNA 序列功率谱的复杂度。

N/3 频率点处的功率----复杂度为$O(N)$快速算法

在DNA序列$\{u[n],\ n=0,1,2,\cdots N-1\}$中，若N为3的倍数，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：

$$P[\frac{N}{3}]=\sum_{b\in I}\left|U_b[\frac{N}{3}]\right|^2=\sum_{b\in I}\left|\sum_{n=0}^{N-1}u_b[n]\cdot e^{-j\frac{2\pi n\cdot\frac{N}{3}}{N}}\right|^2=\sum_{b\in I}\left|\sum_{n=0}^{N-1}u_b[n]\cdot e^{-j\frac{2\pi}{3}n}\right|^2$$

$$=\sum_{b\in I}\left|x_b\cdot e^{-j\frac{2\pi}{3}}+y_b\cdot e^{-j\frac{2\pi}{3}}+z_b\cdot e^{j\frac{2\pi}{3}}\right|^2=\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)$$

继续对上述公式进行推导：

$$(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)=(x_b,y_b,z_b)\begin{pmatrix}1&-1/2&-1/2\\-1/2&1&-1/2\\-1/2&-1/2&1\end{pmatrix}\begin{pmatrix}x_b\\y_b\\z_b\end{pmatrix}=X_b^TMX_b$$

因此，得出序列在 N/3 处的功率谱为：

$$P[\frac{N}{3}]=\sum_{b\in I}X_b^TMX_b \tag{3-2}$$

N/3 频率点处信噪比----复杂度为$O(N)$快速算法

假设 DNA 序列的长度为 N，DNA 序列的四种核苷酸 A，C，T，G 的出现次数为N_A N_C N_G和N_T。因此指示序列$u_b(n)$的功率谱为：

$$|U_b|^2=\sum_{k=0}^{N-1}|U_b[k]|^2=N.N_b \tag{3-3}$$

所以全部序列的总功率 E 为：

$$E=\sum_{b\in I}|U_b|=N.N_A+N.N_C+N.N_G+N.N_T=N^2 \tag{3-4}$$

由于指示序列$u_b(n)$的 DFT 计算公式为：

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

其反 DFT 计算公式为：

$$u_b[n]=\frac{1}{N}\sum_{n=0}^{N-1}U_b[k]e^{j\frac{2\pi nk}{N}},n=0,1,...,N-1 \tag{3-5}$$

根据 Parseval 定理，可以得到

$$\sum_{n=0}^{N-1}|u_b[n]|^2=\frac{1}{N}\sum_{k=0}^{N-1}|U_b[k]|^2 \tag{3-6}$$

观察上式，易得出此等式的左边等于N_b，即：

$$|U_b|^2=\sum_{k=0}^{N-1}|U_b[k]|^2=N.N_b \tag{3-7}$$

因此，DNA 序列的总功率谱 E能由以下式子得到：

$$E=\sum_{b\in I}|U_b|^2=N.N_A+N.N_C+N.N_G+N.N_T=N^2 \tag{3-8}$$

由式 3-8 和题目中给出的信噪比定义式可得:

$$R=\frac{P[\frac{N}{3}]}{\bar{E}}=\frac{P[\frac{N}{3}]}{E/\mathrm{N}}=\frac{\sum_{b\in I}X_b^TMX_b}{N} \tag{3-9}$$

对问题一中b问的求解如下：

Voss 映射与 Z-curve 映射信噪比关系

对于两种线性映射关系

$$\begin{pmatrix} x[n] \\ y[n] \\ z[n] \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix}, \beta_i \text{为列向量}$$

其中 $u_i[n]$ 为 Voss 映射序列， x[n],y[n],z[n]为 Z-curve 映射序列

文献[2]中证明了:

(引理) 如果列向量 β_i 满足以下两个条件

(i) $\forall i(1 \le i \le 4), \|\beta_i\|^2 \equiv c1$， c1 为常量;

(ii) $\forall i, j(1 \le i, j \le 4, i \ne j)$,内积 $\langle \beta_i, \beta_j \rangle \equiv c2$ c2 为常量。

则 x[n],y[n],z[n]序列信噪比 R_Z 与 $u_i[n]$ 序列信噪比 R_I 满足关系

$$R_S = \frac{c_1 - c_2}{c_1} R_I \tag{3-10}$$

由 Voss 映射序列与 Z-curve 映射序列的线性关系

$$\begin{pmatrix} \Delta x[n] \\ \Delta y[n] \\ \Delta z[n] \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix} \tag{3-11}$$

可得

$$\begin{cases} \beta_1 = \{1,\ 1,\ 1\ \} \\ \beta_2 = \{-1,1,-1\} \\ \beta_3 = \{1,-1,-1\} \\ \beta_4 = \{-1,-1,1\} \end{cases} \tag{3-12}$$

易见

(i) $\forall i(1 \le i \le 4)$ $\|\beta_i\|^2 \equiv 3 = c_1$，即满足上述假设(i)。

(ii) $\forall i, j(1 \le i, j \le 4, i \ne j)$， $\langle \beta_i, \beta_j \rangle \equiv -1 = c_2$ 即满足上述假设(ii)。

从而由引理可得 Z-curve 映射序列信噪比 R_Z 与 Voss 映射序列信噪比 R_I 之间存在比例关系:

$$R_Z = \frac{4}{3} R_I \tag{3-13}$$

为了进一步验证（3-13）式中结论的正确性，我们使用 A 题给出的基因数据 gene6 中的第一个序列的从 1 到 2497 段子序列，然后计算在两种映射情况下

信噪比的值，下图 3-1 是 Z-curve 映射与 Voss 映射的信噪比：

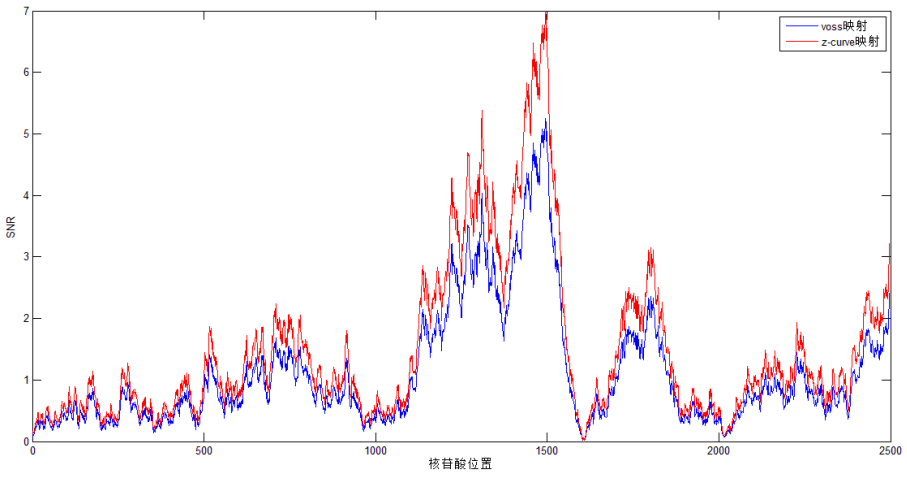

图 3-1　Z-curve 映射与 Voss 映射的信噪比

(注：红色为 Z-curve 映射，蓝色为 Voss 映射)

由上图两种映射对应的信噪比曲线，可以看出采用 Z-curve 映射对应的信噪比值大概是采用 Voss 映射对应的功率谱值的 1-2 倍左右，因此也验证了（3-13）式中 $R_Z = \frac{4}{3} R_I$ 的结论。

Voss 映射与 Z-curve 映射频谱关系

假设一个 DNA 序列 $u(n)$ 长度为 N，$u(n)$ 的滑动窗口，M 点 DFT 定义为：

$$U(m,k) \triangleq \sum_{n=0}^{M-1} u(n+m) e^{-j2\pi nk/M} \tag{3-14}$$

窗口开始的点设 m=0，P，….，（N-1）/P（若（N-1）/P 整数，$u(n)$ 取 0），P 为窗口滑动的次数，如果 P=1，则窗口每次滑动一个核苷酸的距离，而如果 P=3，则窗口的位移是密码子的基数。为了造成 $u(n)$ 周期 3 的情况，我们设 $M = 3L$（L 为正整数）然后设频率 k 为 $L = M/3$，则等式（3-14）可变换为：

$$U(m) \triangleq U(m,L) = \sum_{n=0}^{M-1} u(n+m) e^{-j2\pi n/3} \tag{3-15}$$

继续对该公式进行 P=3 的多项展开，使 $u(m+n) \equiv u_m(n)$ 我们能重新得到

$U(m)$ 的表达式如下：

$$U(m) = \sum_{r=0}^{2} \sum_{n=r,r+3,\ldots}^{\left[\frac{N-1}{3}\right]} u_m(3n+r) e^{-j2\pi r/3} \triangleq \sum_{r=0}^{2} U_{m_r} e^{-j2\pi r/3} \tag{3-16}$$

接着计算 Voss 映射的 DNA 频谱 P_S，首先找到 4 个 Voss DFT 序列的多相组成，如下所示：

$$U_{bm} \triangleq U_{bm0} + U_{bm1}e^{-j2\pi/3} + U_{bm2}e^{-j4\pi/3}, b \in I \tag{3-17}$$

根据上文的假设条件，DNA 总频谱 P_S 表达式为：

$$P_S = |U_{Am}|^2 + |U_{Tm}|^2 + |U_{Gm}|^2 + |U_{Cm}|^2 \tag{3-18}$$

由于 $|U_{bm}|^2 = U_{bm}U_{bm}^*$，*表示复数。可以推出：$|U_{bm}|^2 = 1/2\sum_{r=0}^{2}[U_{bm_r} - U_{bm_q}]^2$

将上述公式带入(3-18)，P_S 的计算公式为：

$$P_S = 1/2\sum_{b\in I}\sum_{r=0}^{2}[U_{bm_r} - U_{bm_q}]^2 \quad (q = (r+1) \bmod 3) \tag{3-19}$$

对于 Z-curve 映射，首先分别找到 $x(n), y(n)$ 以及 $z(n)$ 对应的多相表达式，以 $x(n)$ 为例，我们能写成：

$$U_m = \sum_{n=0}^{M-1} x(n)e^{-j2\pi n/3} = 2\sum_{n=0}^{M-1}[x_A(n) + x_G(n)]e^{-j2\pi n/3} - \sum_{n=0}^{M-1} e^{-j2\pi n/3} \tag{3-20}$$

由于 M=3L，公式（3-20）中的第二个总和为 0，通过多相表达式，我们能得到

$$U_m = 2\sum_{r=0}^{2}(U_{Am_r} + U_{Gm_r})e^{-j2\pi r/3} \tag{3-21}$$

由于 $|U_{bm}|^2 = U_{bm}U_{bm}^*$，能得到 $|U(m)|^2 = 2\sum_{r=0}^{2}[U_{Am_r} + U_{Gm_r} - U_{Am_q} - U_{Am_q}]^2$

其中 $q = (r+1) \bmod 3$，同理可得 $y(n)$ 和 $z(n)$ 的频谱。

则 Z-curve DNA 频谱表达式为：

$$P_Z = 2\sum_{b\in I^1}\sum_{r=0}^{2}[U_{Am_r} + U_{bm_r} - U_{Am_q} - U_{bm_q}]^2 \tag{3-22}$$

其中 $q = (r+1) \bmod 3$，I^1 是核苷酸集合 I 的子集，即 $I^1 = \{C, G, T\} \subset I$，重新整理有：

$$P_Z = 4P_S + 4\sum_{r=0}^{2}(U_{Am_r} - U_{Am_q})\sum_{b\in I}(U_{Am_r} - U_{Am_q}) \tag{3-23}$$

通过观察，可以得到对于 $r \in \{0,1,2\}$，$[U_{Ar} + U_{Gr} + U_{Cr} + U_{Tr}]$ 等于处于窗口中

的第 r 个密码子位置所包含的基因数的总和。这是因为处于窗口中的每个密码子，第 r 个位置永远都是被一个核苷酸占据的，这个数值是个常量而且等于窗口长度的 1/3，通过这个结论，我们能得出：

$$P_Z = 4P_S \tag{3-24}$$

为了进一步验证（3-24）式中结论的正确性，我们使用 Z-curve 映射和 Voss 映射对人类线粒体基因（NC_012920_1.fasta）做功率谱曲线，两个功率谱曲线如下图 3-2，3-3 所示：

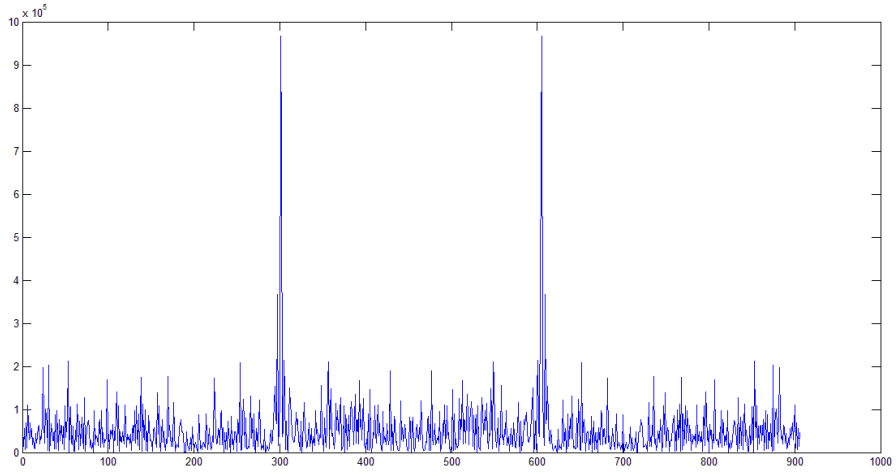

图 3-2 采用 Voss 映射对应的功率谱图

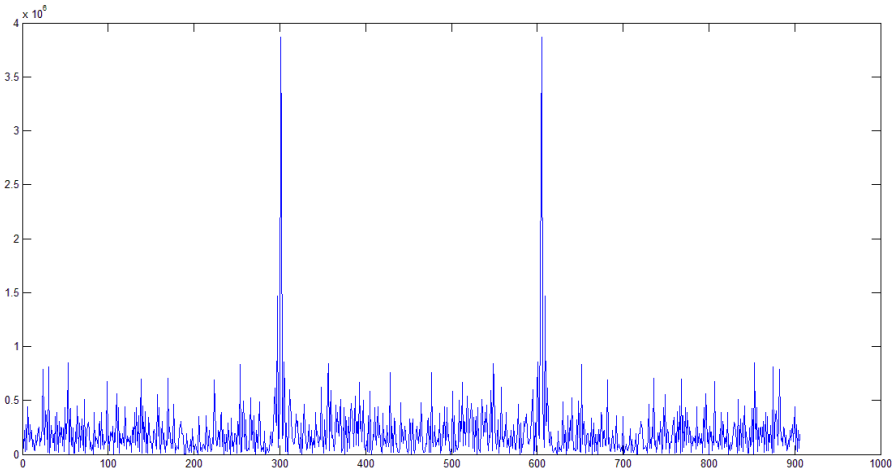

图 3-3 采用 Z-curve 映射对应的功率谱图

由上图两种映射对应的功率谱曲线，可以看出采用 Z-curve 映射对应的功率谱的值大概是采用 Voss 映射对应的功率谱值的 4 倍左右，因此也验证了（3-24）式中 $P_Z = 4P_S$ 的结论。

对问题一中 c 问的求解如下:

N/3 频率点处的功率----复杂度为$O(N)$快速算法

对于实数映射$A\to 0,C\to 1,G\to 2,T\to 3$，给定一段 DNA 序列片段$u(n)$，长度为N，对指示序列做离散 Fourier 变换(DFT):

$$U[k]=\sum_{n=0}^{N-1}u[n]e^{\frac{-j2\pi nk}{N}},k=0,1,...,N-1 \tag{3-25}$$

整个 DNA 序列$u(n)$的总功率谱序列$\{P_R[k]\}$为:

$$P_R[k]=\left|U[k]\right|^2,\ k=0,1,\cdots N-1 \tag{3-26}$$

进一步可以得到 DNA 序列$u(n)$的总功率谱的平均值为:

$$\overline{E}=\frac{\sum_{k=0}^{N-1}P_R[k]}{N} \tag{3-27}$$

同样，我们也可以得出此实数映射的信噪比:

$$R_R=\frac{P_R[\frac{N}{3}]}{\overline{E}}=\frac{\left|U[\frac{N}{3}]\right|^2}{\overline{E}} \tag{3-28}$$

在DNA序列$\{u[n],n=0,1,2,\cdots N-1\}$中，四种核苷酸$A,T,C,G$出现在该序列的Position1：0, 3, 6, …与Position2：1, 4, 7, …以及Position3：2, 5, 8, …三个位置上的和分别记为x_b,y_b和z_b，即

$$x_b=\sum_{i\in position1}u(i),y_b=\sum_{i\in position2}u(i),z_b=\sum_{i\in position3}u(i)$$

则N/3处的总功率谱值即为:

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

$$=\left|x_b+y_b\cdot e^{-j\frac{2\pi}{3}}+z_b\cdot e^{j\frac{2\pi}{3}}\right|^2=(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)$$

$$=(x_b,y_b,z_b)\begin{pmatrix}1&-1/2&-1/2\\-1/2&1&-1/2\\-1/2&-1/2&1\end{pmatrix}\begin{pmatrix}x_b\\y_b\\z_b\end{pmatrix}=X_b^TMX_b \tag{3-29}$$

N/3 频率点处的信噪比----复杂度为$O(N)$快速算法

$$R=\frac{P[\frac{N}{3}]}{\bar{E}}=\frac{P[\frac{N}{3}]}{E/\mathrm{N}}=\frac{\sum_{b\in I}X_b^T MX_b}{N_C+4N_G+9N_T}$$

N_C,N_G,N_T分别代表 C，G，T 三种核苷酸在 DNA 序列中出现的频数。

3.2. 问题二 对不同物种类型基因的阈值确定

3.2.1 问题分析

对于SNR阈值的确定，本文给出了均值平均，概率平均和线性分类器三种方法确定SNR阈值并对分类错误进行了讨论，给出了分类评价指标，对比了以上三种方法的优劣，并针对外显子判别正确率与外显子序列长度的关系以及信噪比阈值与物种类型关系做了一系列讨论。

3.2.2 问题求解

阈值判别效果的评价指标

现假设所选定的信噪比分类阈值为R_0，将$R\geq R_0$作为外显子的判别，$R<R_0$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度和专一性两种指标来表示:

敏感度: $S_n=\frac{T_P}{T_P+F_N}$ (3-30)

专一性: $S_p=\frac{T_N}{T_N+F_P}$ (3-31)

上式中， 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（AC）定义为:

$$AC=(S_n+S_p)/2 \quad (3-32)$$

信噪比阈值的确定方法

A．均值平均法

对于给定的基因数据，设所有外显子的信噪比均值为m_1，所有内含子的信噪比均值为m_2，一种简单确定阈值的方法是将它们作算术平均，即令:

$$R_0=\left(m_1+m_2\right)/2 \quad (3-33)$$

B．概率平均法

本模型基于对给定基因数据外显子和内含子信噪比统计量均服从正态分布的假设。并设所有外显子的信噪比标准差为σ_1，所有内含子的信噪比标准差为σ_2。则阈值R_0的估计值可由(3-34)式给出:

$$R_0 = \frac{(m_1.\sigma_2 + m_2.\sigma_1)}{(\sigma_1 + \sigma_2)} \tag{3-34}$$

推导过程如下：

对于标准正态分布随机变量x落在[-∞，a]区间$(a>0)$的概率是

$$P(a) = \Phi(a)$$

随机变量x落在$[-a，+\infty]$区间$(a>0)$的概率是

$$P(a) = 1 - \Phi(-a) = \Phi(a)$$

由于显然有$m_1 < R_0 < m_2$

对于给定阈值$\hat{R}_0$，随机变量x落在外显子$\hat{R}_0$阈值左侧内的概率

$$P_1 = \Phi(\frac{x - m_1}{\sigma_1})$$

同理，随机变量x落在内含子$\hat{R}_0$阈值右侧内的概率

$$P_2 = \Phi(\frac{m_2 - x}{\sigma_1})$$

理想的阈值估计值$\hat{R}_0$应该是使$P_1 = P_2$的阈值，即

$$\hat{R}_0 = \frac{(m_1.\sigma_2 + m_2.\sigma_1)}{(\sigma_1 + \sigma_2)} \tag{3-35}$$

C．线性判别分类器方法

信噪比阈值判定可以转化为一个典型的二分类问题，因而可以使用线性判别分类器进行分类，而对应决策函数的零点将成为信噪比判定的阈值。事实上这种类型的分类器有很多，包括Fisher分类器，神经网络，SVM等等。这里使用Ho-Kashyap算法(最小均方误差算法的一种)，算法的本质就是将求解判别线性不等式方程组转化为求欠定方程组的最优解，而后者可以使用最小二乘法求解。

阈值确定方法的评价与分析

为评价3种阈值确定方法的性能，并将其与固定阈值为2的方法比较，我们使

用上述方法分别对A题中Genes100，genes200数据中Homo sapiens(人类)，Mus musculus(鼠类)，Mammal(哺乳动物)三种类型的基因的阈值进行计算，阈值判断的正确率统计如下表3-1所示。

表3-1 不同算法对应的阈值的正确率对比

基因种类	均值平均法		概率平均法		线性判别分类器方法		$R_0=2$
	R0	*AC*	R0	*AC*	R0	*AC*	*AC*
genes100，Homo sapiens	1.8858	0.7375	1.1577	0.7750	1.0039	0.7750	0.7250
genes100，Mus musculus	1.6065	0.7712	1.0884	0.8058	1.0402	0.8136	0.7199
genes200，Mammal	1.6957	0.7789	0.8161	0.7778	1.2582	0.8166	0.7481

分别分析上表中不同算法阈值R0对应的正确率AC，如下图3-4所示：

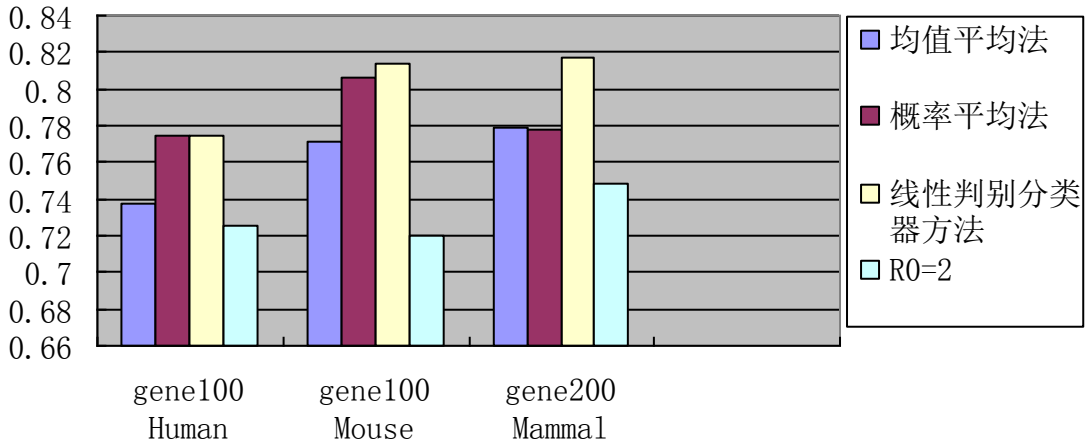

图3-4 不同算法对应的阈值的正确率

对表3-1和图3-4中的数据分析可知：

a) 对于两种均值算法和线性分类器方法的判别正确率均比常值2为阈值进行判别时的正确率高。
b) 概率平均法的正确率比直接均值平均法的效果要好。
c) 两类均值算法的正确率比线性分类器方法的判别正确率低。
d) 对于测试数据，三种算法的信噪比阈值均比题目中给出的常数阈值2小，因此以2为信噪比阈值来判别外显子和内含子并不总是合适的。

外显子判别正确率与外显子序列长度的关系讨论

对A题给出的数据gene200和gene100中Mammal和Mus musculus的DNA序列，按长度将外显子分成3类：长度≤100bp为短外显子，100bp＜长度≤300bp为中等长外显子，300bp＜长度≤500bp为较长外显子，长度＞500bp为长外显子，使用线性分类器确定阈值方法，并得到判别正确率统计如表3-2所示。

表3-2　外显子长度与正确性关系

外显子类型	gene200 Mammal		gene100 Mus musculus	
	R0	AC	R0	AC
短外显子<100	1.0532	0.7037	1.0231	0.7046
中等长度外显子>=100 && <300	1.6141	0.9078	1.1914	0.8735
较长外显子>=300 && <500	1.7544	0.9472	2.0834	0.9465
长外显子>=500	2.8271	0.9953	2.7664	0.9907

由表3-2可见：对于Mammal和Mus musculus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别正确率也随之增加，短外显子的判别正确率比较低。

信噪比阈值与物种类型关系讨论

对A题给出的genes200中Mammal和Genes100中Homo sapiens和Mus musculus，以及题目和附件中提到的酵母菌，拟南芥和人类线粒体的DNA序列的信噪比阈值使用均值平均法进行确定，统计结果如下表所示：

表3-3 不同物种基因的阈值及其正确率

基因种类	均值平均法	
	R0	AC
genes100，Homo sapiens	1.8858	0.7375
genes100，Mus musculus	1.6065	0.7712
genes200，Mammal	1.6957	0.7789
拟南芥	2.0765	0.7553
酵母菌	4.6758	0.7843
线粒体	12.0872	0.7555

由表3-3可以得出以下结论：

a) 不同物种生物，信噪比阈值是不同的，因此信噪比阈值应该至少基于物种类型；

b) DNA序列相近的生物，信噪比阈值往往也接近，譬如人类、老鼠和哺乳动物；

c) 高等生物信噪比阈值通常趋于高于低等生物；

譬如 $R_0(\text{酵母菌}) > R_0(\text{拟南芥}) > R_0(\text{Mammal})$，而人类线粒体序列信噪比阈值较高除了其固有性质造成以外，还可能其属于和细菌一样的低等生物(尽管我们都知道线粒体是参与有氧呼吸的细胞器，并非一种生物)，而这一点有趣地和生物学中“内共生”学说不谋而合。

3.3. 问题三 基因识别算法的实现

3.3.1 问题分析

由于 DNA 序列随机噪声等原因,现有基因识别算法很难“精确地”确定基因外显子区间的两个端点。针对这一问题，首先，需要探讨序列映射、滑动窗口长度、噪声抑制滤波对频谱的影响，并探求对频谱的优化。其次针对优化了的频谱，设计基因识别算法识别出对应的外显子和内含子区域，并设计算法的评估指标，使用 A 题中给出的 Genes100 和 Genes200 中已经标注的 DNA 数据

对算法进行验证和评估。然后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，并分析预测的结果。

基于以上的分析、理解，我们对问题三做了如下的求解。

3.3.2 模型建立与求解

本文首先设计算法的评估指标，然后探讨了不同序列映射、不同的滑动窗口长度、滤波对频谱的影响，并探求对频谱的优化。其次针对优化了的频谱，设计基因识别算法识别出对应的外显子和内含子区域，使用 A 题中给出的 Genes100 中已经标注的 DNA 数据对算法进行验证和评估。

a) 算法评价指标

为了检验算法对 DNA 编码段预测的结果，本文采用了敏感度 S_n 和特异性 S_p，准确率 AC 三个指标：敏感度指在预测的结果中，预测正确的外显子个数在真实的外显子总数中的比例，特异性指在预测的结果中，预测正确的内含子个数在真实的内含子总数中的比例。准确率综合这二者指标的评估指标，用于综合反映整体的指标。

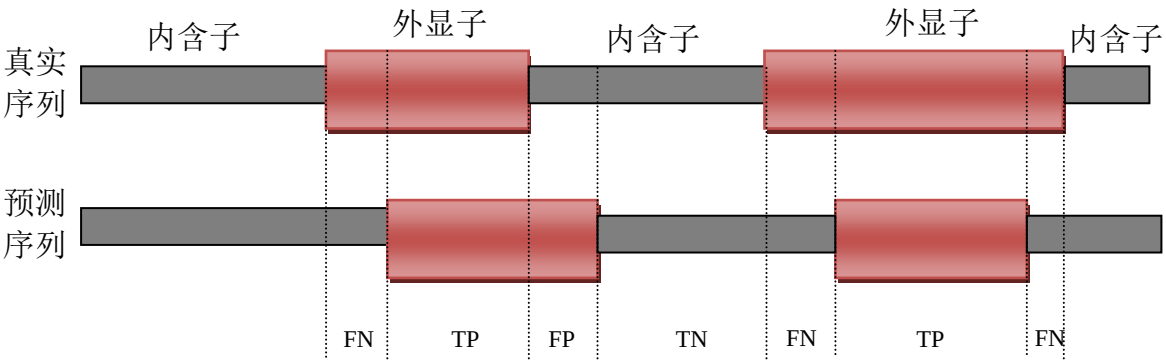

图 3-5 真实 DNA 序列与预测序列之间的关系

上图中，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。三个指标的定义式如下：

$$\text{敏感度：}\ S_n = \frac{T_P}{T_P + F_N} \tag{3-36}$$

$$\text{特异性：}\ S_p = \frac{T_N}{F_P + T_N} \tag{3-37}$$

$$\text{准确率：}\ AC = \frac{S_n + S_p}{2} \tag{3-38}$$

b) DNA 序列映射方法的选择

DNA 映射方法有多种，对不同映射方法进行 DFT，产生的频谱可能会对

DNA 编码序列的预测结果造成不同的影响，因此在建模的第一步我们需要确定对 DNA 序列进行映射的方法，本文中，我们主要对比采用 Voss 映射，Z-curve 映射，以及复数映射时基因预测的效果。

复数映射

复数映射假设表示碱基的三维四面体向量参数是+1 或-1，通过把基本四面体投影到一个选定的平面，生成的映射维数可以进一步降低为二维，即 $A=1+j$，$C=-1+j$，$G=-1-j$ 和 $T=1-j$。复数映射反映出碱基的一些数学性质。例如，根据实轴对称，表现出 A-T 和 C-G 配对的互补原则，嘌呤-嘧啶配对（即 A-C，G-T）具有相同的虚部。复数映射可以把 DNA 字符串转换成一条或四条序列。

文献[3]中指出复数映射中噪声干扰严重，无法直接判别外显子的位置。其效果明显不如 Voss 和 Z-curve 映射的效果。由上文可知 Z-curve 和 Voss 具有相同的频谱图。因此，在本文提出的模型中采取的数值映射方式是 Voss 映射。

c) 滑动窗口的选择

本文对 A 题附件 Genes100 第一个 DNA 片段数据进行不同滑动窗口长度的 DFT 并绘制频谱图如下图 3-6 所示. 不同滑动窗口长度频谱图，自上而下滑窗长度分别为 $\frac{L}{10},\frac{L}{6},\frac{L}{2},4L,8L$和L；L 为平均外显子长度。

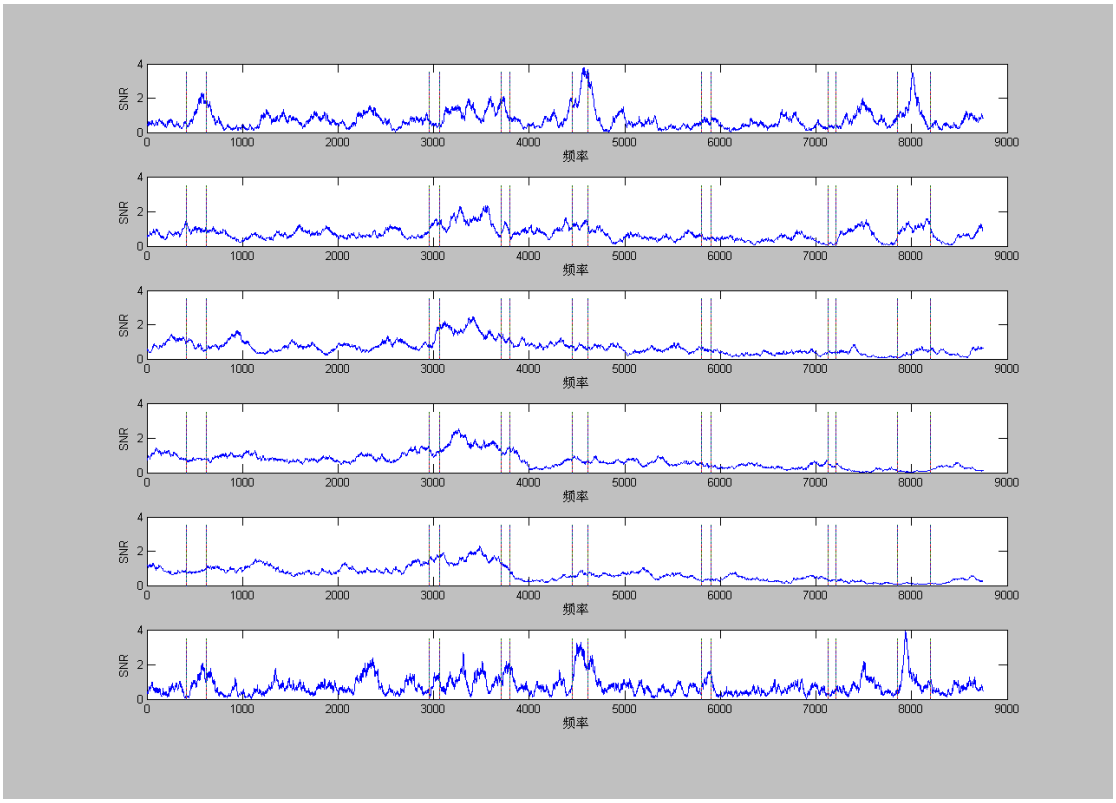

图 3-6 不同滑动窗口长度的 DNA 片段的 DFT 频谱图

由图 3-6 可以看出，使用短窗口让频谱振荡加剧，这会使很多内含子序列被错误判定为外显子序列。而使用长窗口，短的外显子将无法捕获。因此一个

合适的滑动窗口长度将对基因识别产生比较大的影响，结合实验的数据，本文提出两个经验性的滑动窗口长度选择策略。

策略一：设定滑动窗口长度 M 为$[\frac{L}{2},2L]$之间的一个值，L 为平均外显子长度。

策略二：使用不同窗口长度进行滑动 DFT 频谱，然后对频谱进行加权平均。

d) 滤波处理

平滑滤波

滑动窗口频谱曲线图(以信噪比 SNR 为纵轴)呈现高频震荡，直观上表现为频谱的“毛边”，这将会影响之后基因识别算法的判断，这里使用平滑滤波预处理。平滑预处理算法可以简单描述成对于滑动窗口频谱曲线的每一点使用其相邻 K 点(包括该点)的信噪比(R)的均值替代为该点的信噪比。

滑动窗口频谱为: $R(n;\frac{M}{3}),\mathrm{n}=1,2,\cdots \mathrm{N},\quad M$为滑动窗口长度

使用平滑处理的滑动窗口的频谱曲为：

$$R(n;\frac{M}{3})=\sum_{i=n-\frac{k-1}{2}}^{n+\frac{k-1}{2}}R(i;\frac{M}{3})/K,\mathrm{n}=1,2,\cdots \mathrm{N}$$

下图 3-7 是针对附件 Genes100 第一个 DNA 片段数据的滑动窗口频谱曲线平滑处理结果，K 取不同值的结果对比:

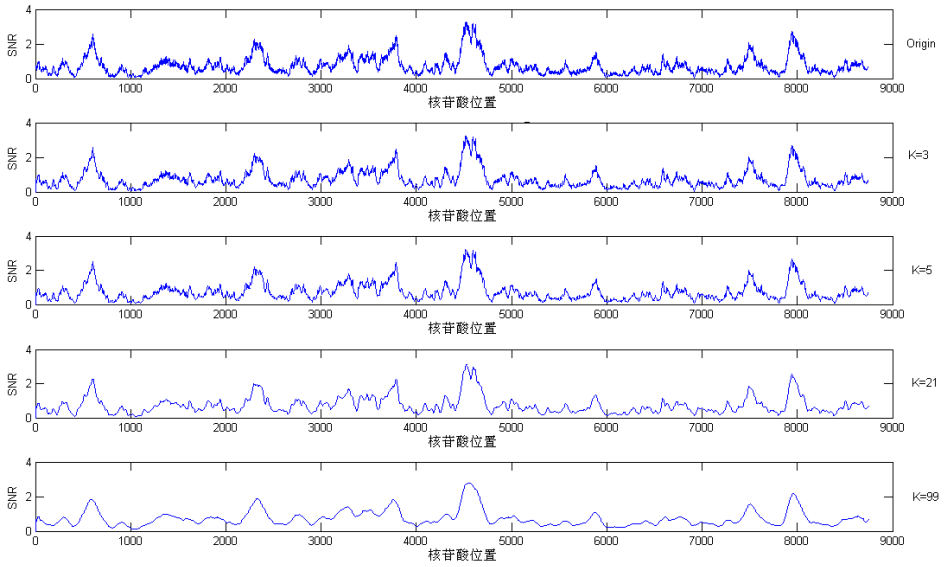

图 3-7 DNA 片段滑动窗口频谱曲线平滑处理结果

由上图 3-7 可见，滑动平滑滤波很好地平滑了频谱曲线，消除了频谱毛边，合适的 K 值(譬如 K=21)不但很好地消除了频谱的毛边，也很好的保持了频谱的低频轮廓。这将对基因识别算法的准确性产生积极意义。

随机噪声抑制

由于 DNA 序列随机噪声的影响，频谱的特征会不同程度上被削弱或淹没。

因而抑制背景噪声可以成为提高基因算法准确率的一种有效手段，本文提出一种滤波算法抑制序列随机噪声以增强频谱特征。

注意到外显子在 N/3 出现频谱峰值而内含子没有这样的特征，因而在滑动窗口频谱曲线图滑窗滑动到外显子区域时往往会出现 SNR 峰值。基于这样的事实，如果对整个 DNA 的 N/3 中心频率进行陷波处理，将可以剥离出噪声序列。再对 DNA 序列和噪声序列进行滑动窗口 DFT，所得 SNR 结果之差可以认为是抑制噪声后的信号功率的信噪比。算法流程如下。

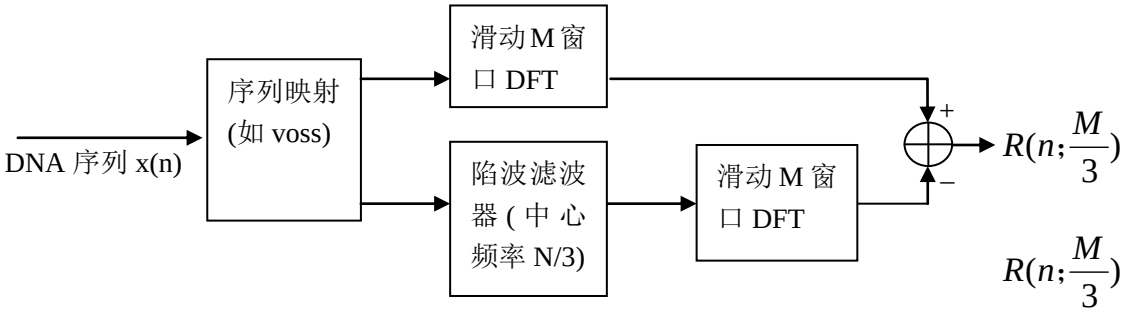

图 3-8. 背景噪声抑制算法流程

算法中的陷波滤波器可以使用二阶 Notch 滤波器，针对附件 Genes100 第一个 DNA 片段数据的随机噪声抑制结果如下。

下图 3-9 是 Genes100 第一个 DNA 片段数据频谱结果，虚线框为外显子位置，图 3-9 中图 1 为原始滑动窗口 DFT 频谱图(窗口大小设置为 200)，图 2 为抑制噪声算法处理后的滑动窗口 DFT 频谱图，图 3 为图 2 经平滑处理后的频谱图。

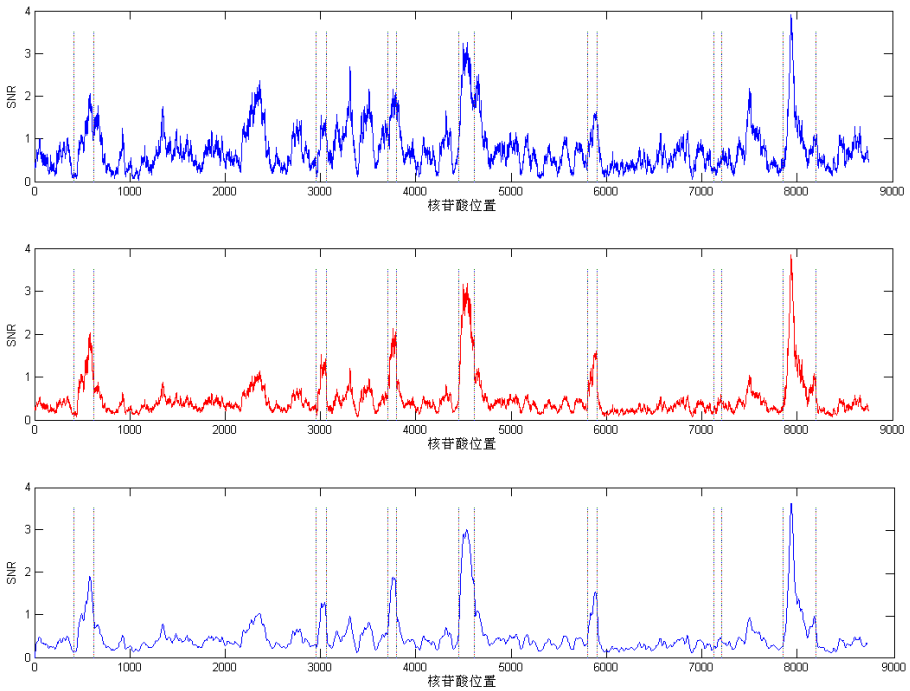

图 3-9 DNA 片段数据频谱结果

(注:图 1，图 2，图 3 为图中依次从上到下的 3 个子图)

由图 3-9 可以看出，结合滑动窗口长度的合理选取以及平滑处理，该算法有效抑制了背景噪声，并保留了 SNR 峰值特征，这将为后面的基因识别算法提

供了易于进行准确识别的频谱。

e) 基因识别算法

选择合适的序列映射、合适的滑动窗口长度并进行噪声抑制滤波和滑动平滑之后将得到比较好的可供进行基因识别的算法的 SNR 频谱。基因识别算法将根据得到的频谱识别出对应的外显子和内含子区域。通常使用 EPND 算法(由 Yin 和 Yau 于 2007 年提出)。

EPND 算法简述如下

1. 计算 SNR 曲线中每个位置的斜率，公式如下：

$$\mathrm{SL(i)}=\frac{SNR(i)-SNR(i-50)}{50}, i=51,52\cdots,N \qquad (3\text{-}40)$$

若在位置 i 处的斜率 $\mathrm{SL(i)}>0$，且 $\mathrm{SNR(i)}\geq R_0$（R_0为SNR阈值），则判定此位置的核苷酸为外显子，否则，判定为内含子。

2. 消除假阳性和假阴性：若外显子区域长度少于 50bp[1]，且左右两边都有内含子，这个区域就判定为内含子区域。另外，若一个内含子区域长度小于 50bp，且夹在 2 个外显子中间，这个区域就判定为外显子区域。

由于 EPND 算法对长 DNA 序列预测正确率不高，尤其是当一个外显子跟着一段很长外显子的时候，外显子序列往往不能被很好地预测。鉴于此本文对 EPND 算法进行了一点改进，改进后的 EPND 算法被实验证明可以更准确地预测外显子和内含子。

改进的 EPND 算法简述如下:

1. 将 DNA 序列分割成 2000bp 的子序列。
2. 将每一个 2000bp 的子序列等分成 k 段，这将产生 k+1 个端点 $\mathrm{P_1}$，$\mathrm{P_2},\cdots,\mathrm{P_{K+1}}$。
3. 对子序列每一段$[\mathrm{P_i}, \mathrm{P_{K+1}}]$，i=1,2,…，K 使用 EPND 算法；单个核苷酸将被多次计算，如果在大多数的计算中，某一个位置核苷酸被预测为属于外显子区域则将该核苷酸判定为外显子核苷酸，反之判定为内含子核苷酸。

为比较不同k值情况下改进的EPND的效果，使用本文提出的算法对Genes100中包含的 92 个 Mus musculus 的 DNA 序列进行预测，并对预测结果求平均 S_n，平均 S_p 和平均 AC，值如下表所示:

[1] EPND 算法忽略长度小于 50bp 的外显子，这样处理带来的误差并不大，因为已经被大量实验验证长度小于 50bp 的外显子非常少(Long et al.,1995;Deutsch and Long，1999).

表 3-4 改进的 EPND 与 EPND 的比较

基因识别算法	平均 S_n	平均 S_P	平均 AC
DFT[2]+EPND	69.37%	95.06%	82.21%
DFT+改进的 EPND(k=2)	68.56%	94.17%	79.36%
DFT+改进的 EPND(k=4)	72.54%	95.22%	83.88%
DFT+改进的 EPND(k=6)	69.60%	94.54%	82.07%
DFT+改进的 EPND(k=8)	67.80%	94.20%	81.00%

由表 3-4 数据可以得到以下结论：

a） 改进的 EPND 比 EPND 的效果要好。

b） 在 k=4 的情况下，改进的 EPND 的效果最好。

f) 算法评估

使用本文提出的基因预测算法对 Genes100 中包含的 92 个 Mus musculus 的 DNA 序列进行预测，并对预测结果求平均 S_n，平均 S_p，平均 AC，值如下表所示:

表 3-5 Genes100 中 92 个 Mus musculus 的 DNA 序列预测结果指标的平均值

基因识别算法	平均 S_n	平均 S_P	平均 AC
DFT+EPND	69.37%	95.06%	82.21%
DFT+改进的 EPND	72.54%	95.22%	83.88%
DFT+滤波+改进的 EPND	81.25%	96.30%	88.77%

由表 3-5 中的数据可以看出在 DFT 上增加抑制背景噪声滤波使用改进的 EPND 的预测算法能够显著的提升预测结果的敏感度，同时也能够提供预测结果的特异性。

[2] 实验中 DFT 均指滑动窗口 DFT，并固定窗口长度 M=200

3.3.3 基因识别算法流程

使用本文提出的基因识别算法预测 DNA 序列编码区的流程如下图 3-10 所示：

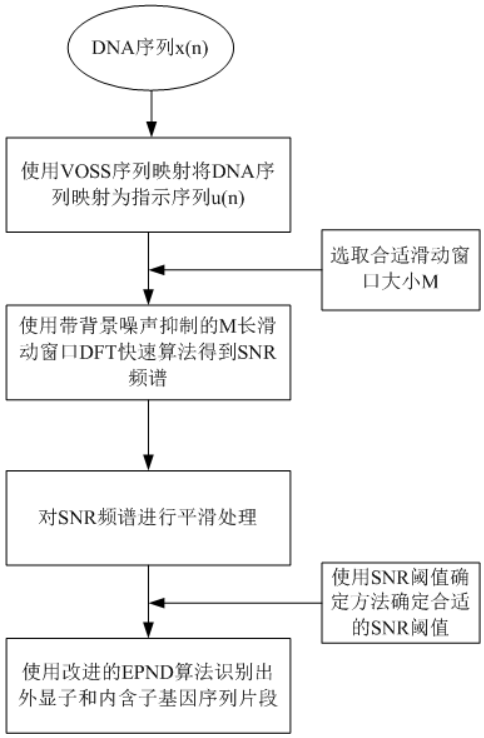

图 3-10 本文基因识别算法识别基因流程

3.3.4 对六个未注释的 DNA 序列编码区的预测

使用本文中提出的基因预测算法预测 A 题中 genes6 中的六段未标注 DNA 序列的编码区域。预测的结果如下表 3-6 至 3-11 所示。图 3-11 至 3-16 中两个子图从上至下称为图 1，图 2。图 1 是基于背景噪声抑制的 SNR 频谱，图 2 是在图 1 的结果上使用平滑滤波得到的 SNR 频谱。预测出的外显子区域在图中使用虚线框框出。

第一段的未知 DNA 序列预测结果

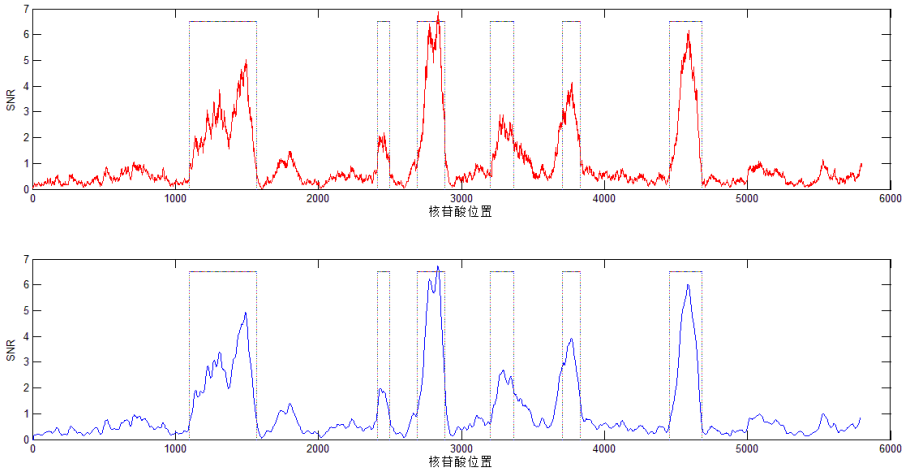

图 3-11 对 genes6 中第一段的编码区预测结果

图 3-11 是对 genes6 中第一段 DNA 序列的预测，预测的结果如下表所示：

表 3-6 对 genes6 中第一段 DNA 序列的预测结果

外显子(共 6 个)	开始位置	结束位置
第 1 个	1099	1569
第 2 个	2415	2497
第 3 个	2693	2885
第 4 个	3204	3365
第 5 个	3708	3833
第 6 个	4456	4680

第二段的未知 DNA 序列预测结果

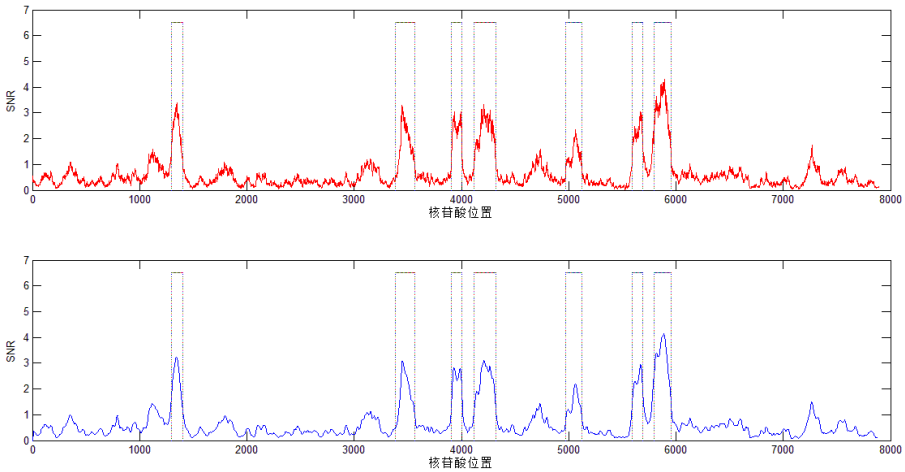

图 3-12 对 genes6 中第二段的编码区预测结果

图 3-12 是对 genes6 中第二段 DNA 序列的预测，预测的结果如下表所示：

表 3-7 对 genes6 中第二段 DNA 序列的预测结果

外显子(共 7 个)	开始位置	结束位置
第 1 个	1295	1406
第 2 个	3387	3565
第 3 个	3903	4007
第 4 个	4114	4319
第 5 个	4691	4747
第 6 个	5589	5687
第 7 个	5793	5957

第三段的未知 DNA 序列预测结果

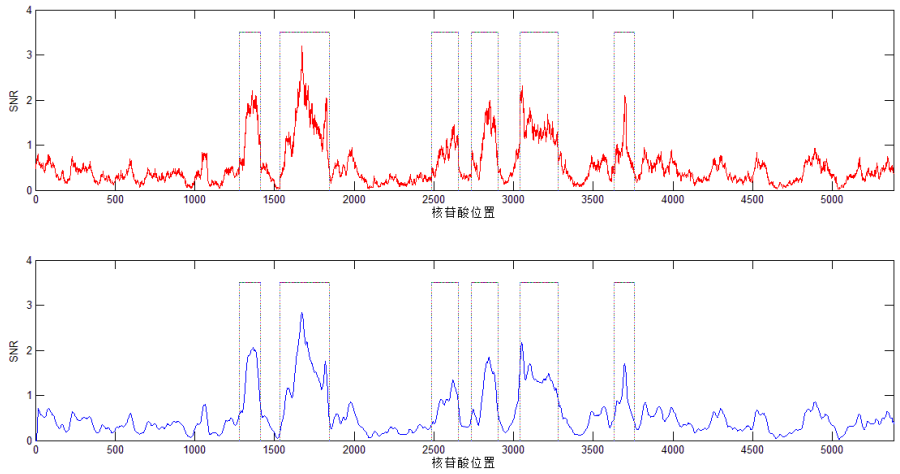

图 3-13 对 genes6 中第三段的编码区预测结果

图 3-13 是对 genes6 中第三段 DNA 序列的预测，预测的结果如下表所示：

表 3-8 对 genes6 中第三段 DNA 序列的预测结果

外显子(共 6 个)	开始位置	结束位置
第 1 个	1278	1411
第 2 个	1533	1845
第 3 个	2485	2653
第 4 个	2736	2901
第 5 个	3038	3277
第 6 个	3628	3755

第四段的未知 DNA 序列预测结果

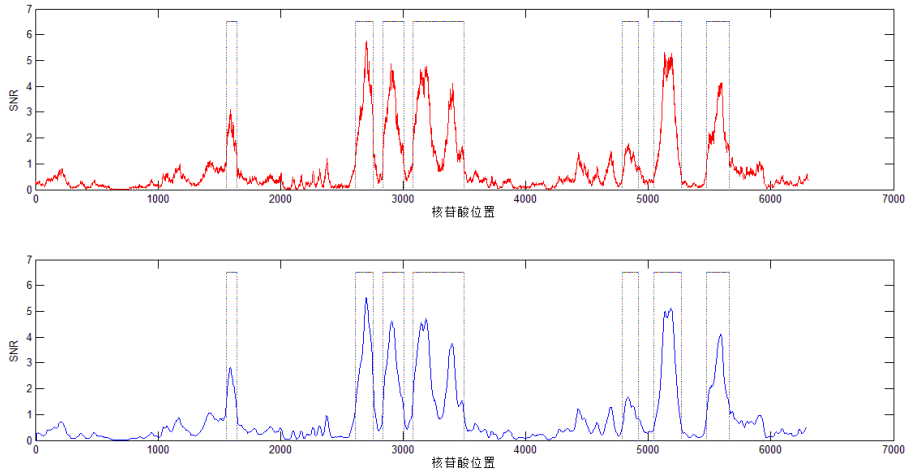

图 3-14 对 genes6 中第四段的编码区预测结果

图 3-14 是对 genes6 中第四段 DNA 序列的预测，预测的结果如下表所示：

表 3-9 对 genes6 中第四段 DNA 序列的预测结果

外显子(共 7 个)	开始位置	结束位置
第 1 个	1559	1643
第 2 个	2608	2756
第 3 个	2834	3004
第 4 个	3078	3495
第 5 个	4797	4971
第 6 个	5045	5271
第 7 个	5474	5658

第五段的未知 DNA 序列预测结果

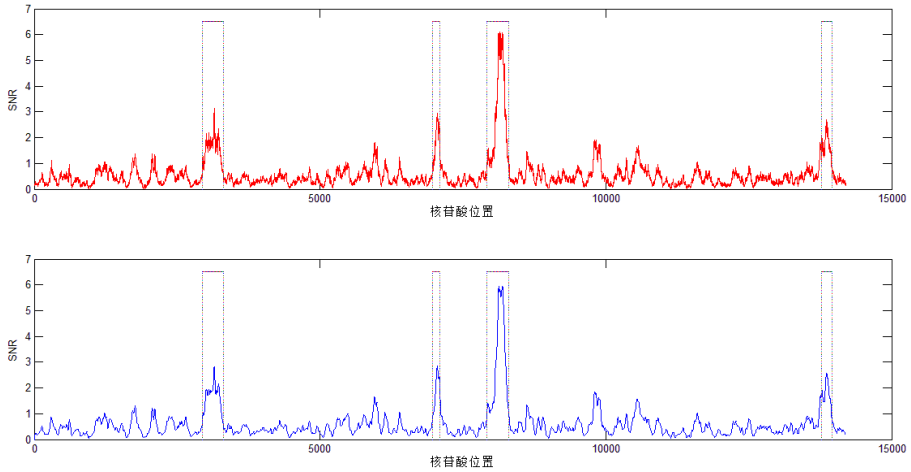

图 3-15 对 genes6 中第五段的编码区预测结果

图 3-15 是对 genes6 中第五段 DNA 序列的预测，预测的结果如下表所示：

表 3-10 对 genes6 中第五段 DNA 序列的预测结果

外显子(共 4 个)	开始位置	结束位置
第 1 个	2944	3360
第 2 个	6964	7101
第 3 个	7912	8302
第 4 个	13775	13951

第六段的未知 DNA 序列预测结果

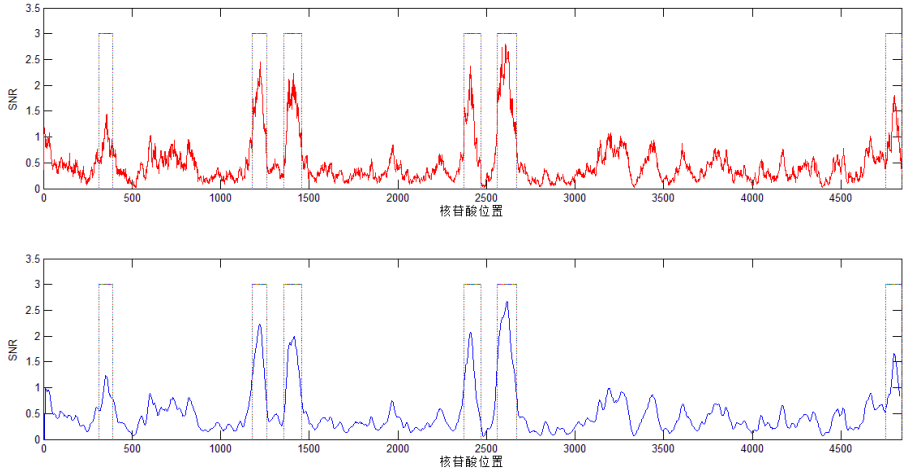

图 3-16 对 genes6 中第六段的编码区预测结果

图 3-16 是对 genes6 中第六段 DNA 序列的预测，预测的结果如下表所示：

表 3-11 对 genes6 中第六段 DNA 序列的预测结果

外显子(共 6 个)	开始位置	结束位置
第 1 个	305	390
第 2 个	1180	1261
第 3 个	1357	1457
第 4 个	2373	2468
第 5 个	2559	2669
第 6 个	4753	4842

3.4. 问题四 延展性研究

3.4.1 其他特征指数探究

本文主要围绕频谱分析展开的，其主要判别指标是频率或信噪比，但这样的方法却可能存在固有缺陷，譬如虽然本文的基因识别模型在测试中表现优异，但对于极短编码外显子(<50bp)事实上是无法被有效识别的,这一方面是由EPND算法的导致的，但从问题 2 的求解过程以及问题 3 的实验研究中我们却可以深刻地感受到对于几段编码外显子，其频谱往往不具有显著特征，下面就结合一些文献给出一些其他的特征指数。

a.密码子使用频率

基因在编码生成蛋白质的过程中，每三个碱基组成一个密码子，转换成一个氨基酸。碱基包含A、T、C 和G 四种，氨基酸包含有20 种。因此，密码子共有4×4×4=64 种不同的组合对应20 种不同的氨基酸。对于一条基因编码序列，假设长度L 为3 的整数倍，将每三个相连的碱基作为一个密码子，统计出各个密码子的数量，比如密码子AAA 的个数记作 n_{AAA}，则该密码子的使用频率为

$f_{AAA}=\frac{3n_{AAA}}{L}$,定义密码子使用频率特征向量为 $F_1=(f_R)$，其中R属于密码子集合，

$R \in \{AAA, AAC, AAG, AAT, ACA, ACC, ..., TTT\}$。

b.碱基组成成分

对于DNA 序列中不同的片段，编码区和非编码区中所包含的碱基成分是不相同的，根据生物学理论可知蛋白质编码区中的C、G 含量较高，而非编码区中的A、T 含量较高。因此，分析A、T、C 和G 的含量是一个非常重要的统计特征。对于DNA 序列S={A，C，T，A，G，A，T，A，C，G，…}，长度为L，其中碱基A 的个数为n_A，其他三种碱基的个数分别为n_C n_G和n_T，则四种碱基对应的含量可表示为$P_A = n_A / L, P_C = n_C / L, P_G = n_G / L, P_T = n_T / L$,把由碱基成分组成的向量$F_2 = (P_A \quad P_C \quad P_G \quad P_T)$作为一个统计特征。

c.碱基位置的相对性

碱基组成成分只反映了碱基的百分含量，没有体现碱基出现的周期性，即碱基出现位置的相关性。对于DNA 序列S，长度为L，其中碱基A 出现的位置分别记做$K_A = \left(k_{A,1}, k_{A,2}, k_{A,3}, ..., k_{A,n+1}\right), 1 \le k_{A,i} \le L, i \in [1, n+1]$。如果取相连的两个位置之间的差$h_{A,i} = k_{A,i+1} - k_{A,i}, i \in [1, n]$，这个差值体现了该碱基出现位置的相关性。

定义$H_R = \left(h_{R,1}, h_{R,2}, ..., h_{R,i}, ...\right)$，其中$i \in [1, n], R \in \{A, C, G, T\}, H_R$的二阶中心矩为:

$V_{H_R} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(h_{R,j} - \overline{h_R}\right)^2}$ 其中$\overline{h_R} = \sum_{i=1}^{n} h_{R,i} / n, R \in \{A, C, G, T\}$，定义特征向量$F_3 = \left(V_{H_A} \; V_{H_C} \; V_{H_G} \; V_{H_T}\right)$。

上述三种与基因相关的特征量均与外显子的长度没有明显的关系。我们知道基因的外显子具有周期3的性质，对于长序列的外显子此性质较为明显，对于短序列的外显子则不明显，利用延长打乱傅利叶方法计算基因序列功率谱在N/3处的幅值$P_x\left(N/3\right), P_y\left(N/3\right), P_z\left(N/3\right)$作为算法中的统计特征量。定义特征向量$F_4 = \left(\log[P_x\left(N/3\right)] \; \log[P_y\left(N/3\right)] \; \log[P_z\left(N/3\right)]\right)$。

对于上述F_1, F_2, F_3, F_4四种特征向量，可以使用Fisher分类器，神经网络，SVM等分类方法识别基因外显子，通过训练样本算出4个特征向量，并利用这些数据训练统计判别函数，再利用该判别函数来识别未知的DNA序列。

由于F_1, F_2, F_3均不受外显子的长度的影响，因此使用上述四种特征向量基于统计学习方法的算法均能够克服短的编码序列可能不具有频谱3周期的问题。

3.4.2 基因突变的探测和发现

基因突变的形式包括 DNA 序列中单个核苷酸的替换和删除或插入。无论对

于外显子还是内含子，单个核苷酸的改变都能影响 DNA 编码性质。故有时候单个核苷酸的突变对整个 DNA 序列编码影响很大。正因为如此，能否掌握预测基因编码序列突变情况的方法，对人类未来了解自身进化等方面有着里程碑式的意义。然而仅仅根据频谱和信噪比来定为基因编码序列中可能存在的突变可能比较困难。

本文提出了一种借助标准基因库基于频谱差异的方法比来探测基因的突变。这种方法将单个的核苷酸改变扩散为频谱中 M 个点的变化，从而可以使用基因检测识别算法检测出来。如下图 3-17，当滑动窗口(窗口长度为 M)滑过突变核苷酸点(设坐标序号为 P)时，在 SNR 频谱上[P-M/2,P+M/2]频段都将受到突变的扰动而发生改变。

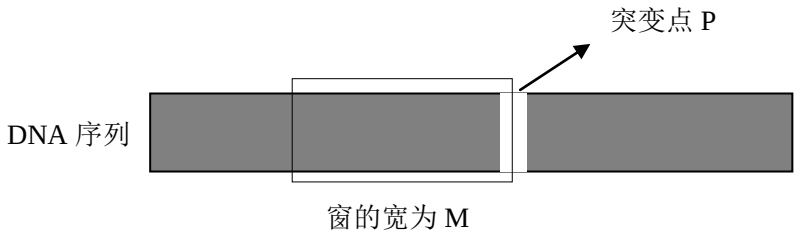

图 3-17 DNA 序列中单个核苷酸变化

如上图 3-17 所示，当检测某个 DNA 片段时时，将待测片段频谱与标准库对应片段 DNA 频谱作差值。这时候相当于噪声被抑制而突变信号特征被凸显，这样就可能使用之前的基因识别算法检测出这一段突变频谱的起始点 x_1 和终点 x_2 (类似于一段外显子片段)而其中点就是突变点 P。

4. 模型总结

模型的优点

由于随机背景噪声等原因的影响，使得很多基因识别算法准确率偏低，本文建立的基于背景噪声抑制和频谱平滑的 SNR 频谱预处理模型，经过预处理后的频谱不仅大幅度抑制了背景噪声，同时保留了 SNR 频谱的模式特征。在编码序列识别上，本文对经典的 EPND 预测算法进行了改进，使用改进的 EPND 算法对经过预处理后频谱进行基因识别，实验结果显示这种基因识别模型具有优异的基因识别性能。

模型的缺点

本文主要围绕频谱分析展开的，其主要判别指标是频率或信噪比，但这样的方法却可能存在固有缺陷，譬如虽然本文的基因识别模型在测试中表现优异，但对于极短编码外显子(<50bp)事实上是无法被有效识别的，这一方面是由 EPND 算法的导致的，另一方面极短编码外显子频谱往往不具有显著特征，由于时间仓促，这方面没有做非常深入的探究。

5.参考文献

[1]马宝山，朱义胜，用于基因预测的自适应滤波器的仿真研究[J],系统仿真学报，2007,9(24):5620～5623

[2] J Shao, X Yan, S Shao 2012 SNR of DNA sequences mapped by general affine transformations of the indicator sequences Journal of Mathematical Biology

[3]饶妮妮,邱丽君, DNA 序列数值映射方法的研究[J], 生物医学工程学杂志 ,2005, 22(4) ：681～685

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