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：

本文主要利用基因编码序列的频谱 3-周期性，使用信号处理和分析的手段处理 DNA 序列，运用已知编码区域的基因样本，建立了基因编码区域识别的数学模型，并应用此模型对 6 个未知的 DNA 样本进行了编码区域预测。最后讨论了基于信号处理的 DNA 频域分析法在检测基因突变领域上的应用。

对于问题一，我们推导了基因序列的 Voss 映射、Z-curve 和实数映射在计算 N/3 频谱峰值和总功率谱平均值中的快速算法。使用碱基在子序列出现的频数进行计算，避免了 DFT 的繁杂运算，特别的是当连续的滑动窗口或移动序列时，出现频数可以由之前频数简单处理得到，使得频谱与信噪比的求解为常数复杂度。讨论了各数值映射关系的优劣，并决定在建模中使用 Voss 映射。

对于问题二，我们定义了识别灵敏度和识别特征度两个指标量来定量的描述识别算法的优劣。对于基因识别算法模型中的重要参量——判别阈值，采用“大范围、小步进”的方式遍历搜索。对于每一个判别阈值分别求出相应的灵敏度和特征度，并分别给出了具有工程意义的 90%、80%、70%特征度意义下，灵敏度最优估计阈值。最终将获得阈值在同一物种 DNA 样本中进行编码区域识别验证，取得了预期的效果，说明了阈值确定方法的合理性。

对于问题三，我们结合了已有的识别算法，针对识别序列破碎、端点模糊等问题，在计算过程中加入一些滤波、填补、检测调整等辅助方式，获得了较为精确的外显子识别算法，并应用新识别算法对已知编码区序列进行验证，取得十分良好的效果。最终应用新算法于 6 个未被注释的 DNA 序列的编码区域预估中，预测出相应的编码区域。

对于拓展性研究，我们探讨了 DNA 频域分析法在检测基因突变领域上的应用。通过对 DNA 序列中单个核苷酸进行替换、删除、插入等操作，根据 DNA 序列频谱的变化，观察 $P[\frac{N}{3}]$ 幅值的衰减或者产生的杂散谱幅值，大概分析出基因突变的核苷酸位置。对接下来深入研究基因突变检测具有指导性的意义。

关键词： 频谱 3-周期性 阈值确定 编码区预测 基因突变识别

基因识别问题及其算法实现

1　问题来源

对大量、复杂的基因序列的分析，传统生物学解决问题的方式是基于分子实验的方法，其代价高昂。诺贝尔奖获得者 W.吉尔伯特（Walter Gilbert，1932—；【美】，第一个制备出混合脱氧核糖核酸的科学家）1991 年曾经指出：“现在，基于全部基因序列都将知晓，并以电子可操作的方式驻留在数据库中，新的生物学研究模式的出发点应是理论的。一个科学家将从理论推测出发，然后再回到实验中去，追踪或验证这些理论假设。” 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，怎么去识别出其中的编码序列（即外显子），也称为基因预测，是一个尚未完全解决的问题，也是当前生物信息学的一个最基础、最首要的问题。

研究表明，在基因外显子序列的功率谱曲线中，在 $\frac{N}{3}$ 频率处具有较大的频谱峰值(Peak Value)，反映了在基因外显子片段上，四种核苷酸符号在序列的三个子序列上分布的“非均衡性”。通常认为这种现象源于编码基因序列“密码子”（coden）使用的偏向性（bias）。虽然目前对此现象产生的“机理”还不是十分地清楚，但是频谱的 3-周期性被普遍认为是可用于识别基因编码序列（外显子）的一个重要的特征信息。

因此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也受到广泛重视。

2　问题描述

采用信号处理与分析的手段进行基因预测，首先需要对于 DNA 序列进行数值映射，根据一定的规则将 DNA 映射成相应的数值序列，以便于对其进行数字处理。根据不同需要，可对基因序列进行 Voss、Tetrahedron、Z-Curve、Complex…等数值映射，通常将其称为 DNA 序列的指示序列(indicator Sequence)。

对于数值化映射得到的指示序列，对其进行离散 *Fourier* 变换(*DFT*)，可以得到指示序列的功率谱结构。大量实验表明，外显子序列的功率谱曲线在频率 $k=\frac{N}{3}$ 处，具有较大的频谱峰值(*Peak Value*)，而内含子则没有类似的峰值。这种统计现象被称为碱基的 3-周期(*3-base Periodicity*)。

为了定量获取频谱峰值(*Peak Value*)的相对高度，通常我们将 *DNA* 序列在特定位置，即 $k=\frac{N}{3}$ 处的功率谱值，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”（*Signal Noise Ratio，SNR*），即

$$R=\frac{P[\frac{N}{3}]}{\bar{E}} \tag{2-1}$$

其中 $\bar{E}$ 为 *DNA* 序列的总功率谱的平均值

$$\bar{E}=\frac{\sum_{k=0}^{N-1}P[k]}{N} \tag{2-2}$$

DNA 序列的信噪比值 R 的大小，既表示频谱峰值(*Peak Value*)的相对高度也反映编码或非编码序列 3-周期性的强弱。

信噪比 R 大于某个适当选定的阈值 R_0（比如 $R_0=2$），是 *DNA* 序列上编码序列片段（外显子）通常满足的特性，而内含子则一般不具有该性质。所以通过比对信噪比 R 和阈值的相对大小，可以判别该 DNA 片段是否为编码序列片段。

频谱与信噪比概念的引入，最终目的是要探测、预报一个尚未被注释的完整的 *DNA* 序列的所有基因编码序列（外显子）片段。一个典型的基于序列频谱 3—周期性的的基因预测方法流程图如图 1-1 所示。

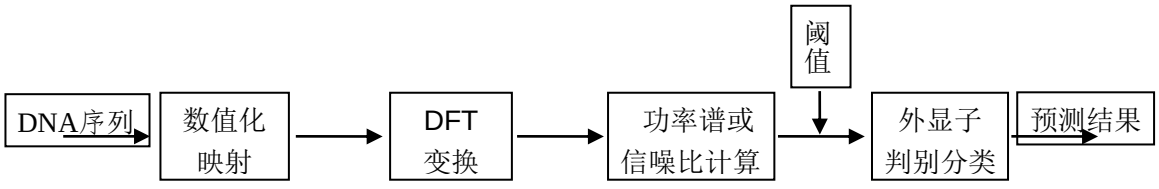

图 2-1 基于序列频谱 3—周期性的的基因预测方法流程图

在利用信号处理手段识别基因编码序列的方法中，还有一些环节的若干问题有待研究，本文主要尝试解决以下几个问题：

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，除了上面介绍的 Voss 映射外，实际上人们还研究过许多不同的数值映射方法。例如，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？

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$\mathrm{R}_0 = 2$，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：http://www.ncbi.nlm.nih.gov/guide/ 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。

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。对于某些基因识别算法，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，你的建模团队有没有更好的解决方法？请对你们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。

4. 延展性研究

“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呢？

3　数学模型

3.1　问题分析

采用信号处理与分析的手段进行基因预测，首先需要对于 DNA 序列进行数值映射，根据一定的规则将 DNA 映射成相应的数值序列，以便于对其进行数字处理。则数值映射的选择直接关系到后期信号处理的难易程度，及识别算法的精度。在建模中我们对 Voss、Tetrahedron、Z-Curve、Complex…等数值映射进行查阅和对其功率谱进行了详细计算分析。注意到 DNA 序列由 ACGT 四种核苷酸组成，且对于某个位置的核苷酸必为这四种核苷酸的一种，则可知其排列的自由度不会超过 3。在 4.1 小节的详细分析中我们发现碱基排列的自由度为 3，所以对于自由度小于 3 的 Complex、EIIP、Real number 等映射，因为自由度信息损失并不可取；而对于 Voss 和 Z-Curve 映射，二者在所需计算的$\frac{k}{3}$谱线峰值信噪比数值上只有一个常数系数差异，本质相同。所以，在实际的建模中，我们采用了比较容易处理的 Voss 数字序列映射。

由生物学知识可知，在基因 DNA 表达遗传信息过程中，编码序列中三个碱基一组转录为相应 RNA，再翻译为相应氨基酸。由于蛋白质中氨基酸使用具有非均匀性，则三碱基组-密码子的使用具有偏向性。而对于不编码蛋白质序列的内含子或基因间“无意义”DNA 序列，则认为密码子的使用无偏向性。所以在本模型中 DNA 序列频谱的 3-周期性被认为是用于识别基因编码序列(外显子）的唯一的特征信息。对于目前发现的基因序列的其他规律，例如 tRNA 信息岛识别、密码子频率等，未加以讨论。

频谱峰值特征的发现，或者频谱与信噪比概念的引入，其最终目的是要探测、预报一个尚未被注释的完整的 DNA 序列。已经有一些研究者提出了识别基因的算法（如参见[6]及其后面的文献）。目前利用信噪比的基因识别算法通常有两种：一是固定长度窗口滑动法；另一是“移动序列”信噪比曲线识别法。研究算法原理发现，窗口滑动法可以较为快速的搜索基因所在区域，而“移动序列”在识别基因外显子端点上具有较大优势。在本次建模实验中，我们主要是针对两种算法各自的特点优势，设计出了一套精确度更高的基因编码区域识别算法。

对于一套算法优劣的判别，我们需要选择合适的指标。对于基因编码序列的识别，一般希望算法可以将尽量多的真实外显子识别出来，同时希望尽量少的将内含子错误的识别为外显子。针对这两个想法，查阅相关资料后，定义灵敏度表征外显子被识别出的多寡；定义特征度表征识别出外显子正确率的多寡。

3.2　模型假设

1. 从基因库下载得到的 DNA 序列没有碱基丢失或碱基添加，且在编码区是没有基因突变，是正确无误的
2. DNA 序列频谱的 3-周期性在编码区、非编码区序列有着显著差异，使

用该方法区分外显子、内含子在绝大多数情况下是正确的，

3. 编码序列长度和非编码序列长度均为 3 的倍数
4. 编码序列长度和非编码序列长度均大于 30
5. 不同物种基因中碱基种类、比例、排列顺序等方面不同，然而同一类型物种，或者亲缘关系较近的两物种，在上述诸方面中具有相似性。

3.3 模型建立

3.3.1 基于固定长度滑动窗口上频谱曲线的基因识别方法:

对一个 DNA 序列 S 和它的指示序列 $\{u_b[n]\}$，$b \in I$，$n = 0,1,2,\cdots N-1$。取长度 M（通常取为 3 的倍数，例如 M=99, 129, 255, 513 等）作为固定窗口长度。

对任意 n（$0 \le n \le N-1$），在以 n 为中心的长度为 M 的序列片段[$n-\frac{M-1}{2}$，$n+\frac{M-1}{2}$]上（当 n 接近序列的两端时，窗口实际有效长度可能会小于 M），作四个指示序列的离散 *Fourier* 变换(*DFT*)

$$U_b[k] = \sum_{i=n-\frac{M-1}{2}}^{i=n+\frac{M-1}{2}} u_b[i] e^{-j\frac{2\pi ik}{M}}, \quad k = 0,1,\cdots,M-1$$

并求出它在 $\frac{M}{3}$ 处总频谱 $p(n;\frac{M}{3})$，即

$$P[\frac{M}{3}] = \left|U_A[\frac{M}{3}]\right|^2 + \left|U_T[\frac{M}{3}]\right|^2 + \left|U_G[\frac{M}{3}]\right|^2 + \left|U_C[\frac{M}{3}]\right|^2 \triangleq p(n;\frac{M}{3})$$

把这样得到的频谱值 $p(n;\frac{M}{3})$，$n = 0,1,2,\cdots N-1$，经过标准化处理（即除以最大频谱值 $\max\limits_{0 \le n \le N-1}\{p(n;\frac{M}{3})\}$），并画出其频谱曲线

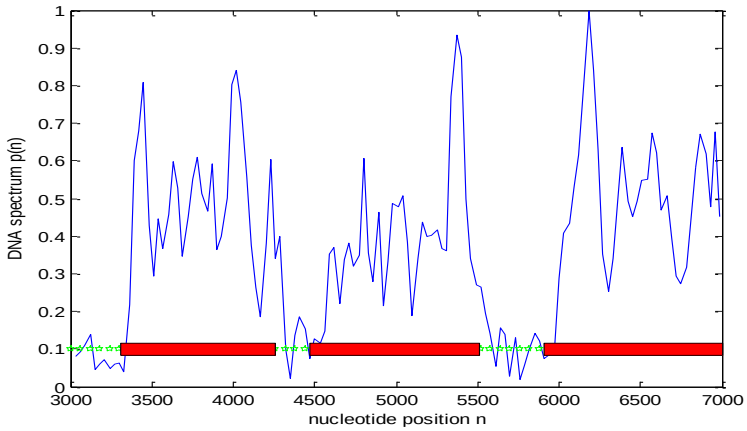

图 3-3-1 固定长度滑动窗口的频谱 $p = p(n;\frac{M}{3})$ 曲线（人类线粒体基因，NC_012920_1.fasta）

图中红色水平细线条是 DNA 序列实际的基因外显子的区间。滑动窗口频谱 $p(n;\frac{M}{3})$ 曲线的峰与基因外显子区间具有“对应”关系。

在实际的定量建模中，选取信噪比判别阈值 R_0。对于信噪比大于 R_0 的序列点位判别为外显子区域，记为 1；对于信噪比小于 R_0 的序列点位判别为非编码序列，记为 0。则得到了由 1,0 组成的外显子判别特征序列 S_1。考虑到外显子长度、内含子长度大于 30 的基本假设，置该序列中的小于 30 个的且两边均为连“1”的连“0”为“1”得到 S_2，然后在核查序列 S_2 中连“1”段（外显子），若长度小于 30，则将该段舍弃置“0”，得到最终的外显子判别序列 S_3。判别示意图如图 3-3-2 所示

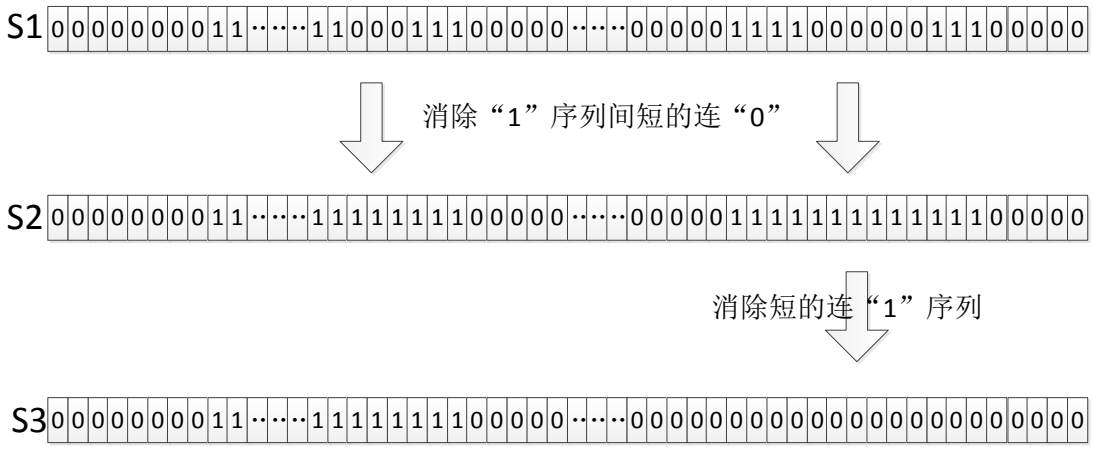

图 3-3-2 阈值判别外显子算法

3.3.2　基于 DNA 序列上“移动序列”信噪比曲线的基因识别方法：

设已知 DNA 序列 S 和它的指示序列 $\{u_b[n]\}$，$b \in I$，$n = 0,1,2,\cdots N-1$。对任意 n（$0 < n \le N-1$），通常 n 取 3 的倍数并逐渐增大。在 n 的左边一个长度为 n 的序列片段[0，n-1]上，相应的子序列 $S_{0\sim n-1}$ 称为 DNA 序列 S 的“**移动子序列**”，作该移动子序列对应的四个指示序列的离散 *Fourier* 变换(*DFT*)

$$U_b[k] = \sum_{i=0}^{i=n-1} u_b[i] e^{-j\frac{2\pi ik}{M}}, \quad k = 0,1,\cdots,n-1$$

并求出移动子序列 $S_{0\sim n-1}$，$n = 0,1,\cdots,N-1$ 上的信噪比 $R[n]$

$$R[n] = \frac{P[\frac{n}{3}]}{\overline{E}[n]} = \frac{\left|U_A[\frac{n}{3}]\right|^2 + \left|U_T[\frac{n}{3}]\right|^2 + \left|U_G[\frac{n}{3}]\right|^2 + \left|U_C[\frac{n}{3}]\right|^2}{\overline{E}[n]}, \quad 0 < n \le N-1$$

其中 $\overline{E}[n]$ 为移动子序列 $S_{0\sim n-1}$ 的功率谱的平均值 $\overline{E}[n] = \frac{\sum_{k=0}^{n-1} P[k]}{n}$。在坐标系中画出移动序列 $S_{0\sim n-1}$ 的信噪比曲线 $R[n]$（称为信噪比移动曲线（SNR walk curve），见图 3-3-3）

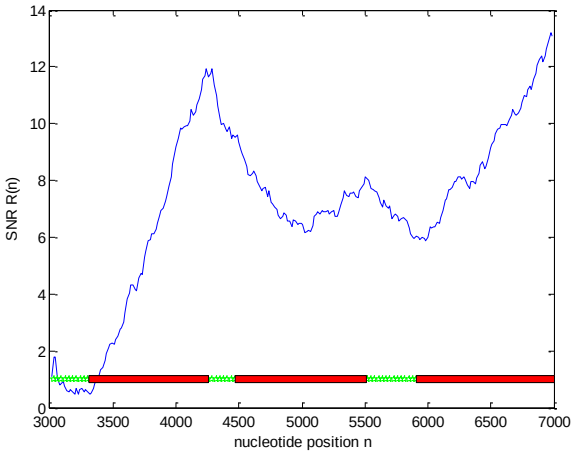

图 3-3-3　DNA 移动序列其指示序列的信噪比曲线。（人类线粒体基因，NC_012920_1.fasta）

图中红色水平细线条是 DNA 序列实际的基因外显子的区间。DNA 序列的信噪比移动曲线的峰、谷与基因外显子区间的端点也具有较“明显的”的对应关系。

在实际的定量建模中，对 DNA 移动序列得到的信噪比曲线取差分得到差分

序列$\{dR_n\}$，对于差分序列$\{dR_n\}$做映射$D_n=\begin{cases}1, dR_n>0\\0, dR_n\le 0\end{cases}$得到信噪比差分特征序列$\{D_n\}$，则在$\{D_n\}$序列中$0\to1$处为信噪比曲线的波谷，序列中$1\to0$处为信噪比曲线的波峰。对$\{D_n\}$做类似于“滑动窗口法”中外显子判别特征序列{$S_1$}的处理则可以取出毛刺影响的外显子端点位置，在此不再赘述。

4　问题求解

4.1　功率谱与信噪比的快速算法

4.1.1　VOSS 映射方式

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）

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

将得到的四个长度均为 N 的复数序列$\{U_b[k]\},b\in I$，分别求功率谱，相加后得到整个 DNA 序列 S 的功率谱序列

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

外显子序列在统计上具有碱基 3-周期性，因此在其频率谱上表现为在频率$k=\frac{N}{3}$处，具有较大的频谱峰值。在基于功率谱与信噪比对基因序列进行探测的算法中，主要只关注$P[\frac{N}{3}]$处的值。

则由 DNA 序列的信噪比的定义

$$R=\frac{P[\frac{N}{3}]}{E}$$

其中 E 表示 DNA 序列 S 的总功率平均值为

$$E=\frac{1}{N}\sum_{k=0}^{N-1}P[k]=\frac{1}{N}\sum_{k=0}^{N-1}\sum_{b\in I}|U_b[k]|^2=\sum_{b\in I}(\frac{1}{N}\sum_{k=0}^{N-1}|U_b[k]|^2)=\sum_{b\in I}\sum_{n=0}^{N-1}|u_b[n]|^2=N$$

在*DNA*序列$\{S[n],\ n=0,1,2,\cdots N-1\}$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

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

$$=\sum_{b\in I}\left|x_b+y_b\cdot e^{-j\frac{2\pi}{3}}+z_b\cdot e^{j\frac{2\pi}{3}}\right|^2=\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b) \qquad (4\text{-}1)$$

通过统计不同核苷酸的三个子序列中的出现频数,可以计算出DNA序列的信噪比

$$R=\frac{P[\frac{N}{3}]}{N} \qquad (4\text{-}2)$$

其中$P[\frac{N}{3}]$采用频数统计算法，这样就大大减少了计算序列FFT的运算量，此种快速算法具有很好的实际效果。

当采用滑动窗口和移动序列算法时，考察序列在前后几次计算中有很大部分的重叠，此种快速算法可以通过增减不重叠部分的三个子序列出现的编码频数，对$P[\frac{N}{3}]$进行计算,避免了多次计算大部分重叠的长序列,大大减小运算复杂度，降低冗余计算量。

4.1.2 Z-curve 映射方式

设DNA序列S的四个指示序列$\{u_b[n]\}$，$b\in I=\{A,C,G,T\}$，$n=0,1,2,\cdots N-1$的累积序列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$x[n],y[n],z[n]$:

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

接着，若令$x[-1]=0$，$y[-1]=0$和$z[-1]=0$，以及$\Delta x[n]=x[n]-x[n-1]$,

$\Delta y[n]=y[n]-y[n-1]$和$\Delta z[n]=z[n]-z[n-1]$，于是我们得到 Z-curve 映射

$$\begin{pmatrix}\Delta x[n]\\\Delta y[n]\\\Delta z[n]\end{pmatrix}=\begin{pmatrix}1&-1&1&-1\\1&1&-1&-1\\1&-1&-1&1\end{pmatrix}\begin{pmatrix}u_A[n]\\u_C[n]\\u_G[n]\\u_T[n]\end{pmatrix} \qquad (4\text{-}3)$$

定义其功率谱

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

由（4-3)、(4-4）式可以推出

$$P_Z[k]=3\sum_{b\in I}|U_b[k]|^2-\sum_{b\in I}(U_b[k]\sum_{\substack{q\in I\\q\neq b}}U_q^*[\mathrm{k}]) \qquad (4\text{-}5)$$

由于指示序列满足

$$\sum_{b\in I} u_b[\mathrm{n}] = I, \text{其中 } I = (1,1,1,...,1) \qquad (4\text{-}6)$$

对（4-6）做离散 DFT 可知

$$\sum_{b\in I} U_b[\mathrm{k}] = H[k], \text{其中 } H[k] = \begin{cases} N, k = 0 \\ 0, k = 1,..N-1 \end{cases} \qquad (4\text{-}7)$$

则将(4-7)代入（4-5）化简得

$$P_Z[k] = 4\sum_{b\in I} |U_b[k]|^2 - H[k]\sum_{b\in I} U_b[k]$$

进而得到

$$P_Z[k] = \begin{cases} 4P[0] - N^2, k = 0 \\ 4P[k], k = 1,...,\mathrm{N}\text{-}1 \end{cases} \qquad (4\text{-}8)$$

其中 P[k]为 Voss 映射下的功率谱函数。

$$E_Z = \frac{1}{N}\left\{4\sum_{k=1}^{N-1} P[k] + 4P[k] - N^2\right\} = \frac{1}{N}\left\{4\sum_{k=0}^{N-1} P[k] - N^2\right\} = 3N$$

按照信噪比的定义可以得到

$$R_Z = \frac{P_Z[\frac{N}{3}]}{E_Z} = \frac{4}{3}R \qquad (4\text{-}9)$$

式 4-9 中的 R 为 Voss 映射下的信噪比，该式反应了 Z-curve 和 Voss 映射下的信噪比关系。

4.1.3 实数映射

对于 DNA 序列 S 进行实数映射 $A \to 0, C \to 1, G \to 2, T \to 3$,得到实数序列 $u[n]$，

$$u[n] = \begin{cases} 0, S[n] = A \\ 1, S[n] = C \\ 0, S[n] = G \\ 0, S[n] = T \end{cases} \qquad (4\text{-}10)$$

对实数序列 $u[n]$ 做离散 Fourier 变换（DFT）

$$U[k] = \sum_{n=0}^{N-1} u\ [n] e^{-j\frac{2\pi nk}{N}}, k = 0,1,...,\mathrm{N}\text{-}1 \qquad (4\text{-}11)$$

计算复序列 $\{U[k]\}$ 的平方功率谱即为序列 S 的功率谱序列 $\{P[k]\}$：

$$P[k]=\left|U[\mathrm{k}]\right|^2=\left|\sum_{n=0}^{N-1}u[n]e^{-j\frac{2\pi nk}{N}}\right|^2,k=0,1,...N-1 \qquad (4\text{-}12)$$

在*DNA*序列$\{S[n]，n=0,1,2,\cdots N-1\}$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

$$P[\frac{N}{3}]=\left|\sum_{n=0}^{N-1}u[n]e^{-j\frac{2\pi n\frac{N}{3}}{N}}\right|^2=\left|(x_C+2x_G+3x_T)+(y_C+2y_G+3y_T)\cdot e^{-j\frac{2\pi}{3}}+(z_C+2z_G+3z_T)\cdot e^{j\frac{2\pi}{3}}\right|^2$$

$$\begin{aligned}&=(x_C+2x_G+3x_T)^2+(y_C+2y_G+3y_T)^2+(z_C+2z_G+3z_T)^2\\&-(x_C+2x_G+3x_T)(y_C+2y_G+3y_T)\\&-(y_C+2y_G+3y_T)(z_C+2z_G+3z_T)\\&-(z_C+2z_G+3z_T)(x_C+2x_G+3x_T)\end{aligned} \qquad (4\text{-}13)$$

用 E 表示 DNA 序列 S 的总功率平均值为

$$E=\frac{1}{N}\sum_{k=0}^{N-1}P[k]=\frac{1}{N}\sum_{k=0}^{N-1}\left|U\ [\mathrm{k}]\right|^2=\sum_{n=0}^{N-1}|u\ [\mathrm{n}]|^2 \qquad (4\text{-}14)$$

$$=\left(x_C+y_C+z_C\right)+4\left(x_G+y_G+z_G\right)+9\left(x_G+y_G+z_G\right)$$

将求得的$P[\frac{N}{3}]$、E带入公式$R=\frac{P[\frac{N}{3}]}{N}$可以求得相应的信噪比。

4.2 不同物种类型基因的阈值确定问题

4.2.1 问题重述

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$\mathrm{R}_0=2$，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。能否针对每类基因给出阈值确定方法和阈值结果，并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。

4.2.2 问题分析

在基于固定长度滑动窗口上频谱曲线的基因识别方法中，信噪比判别阈值R_0的选定直接关系到对基因片段外显子、内含子区域的判断。

较低的R_0将导致更多的基因序列被识别为外显子，即识别出的外显子区域变长变多，此时实际的一些内含子区域可能也被识别为外显子区域。若把对编码序列的检验视为任务目标，则R_0偏低“虚警率”上升，即错误的将内含子识别为外显子的几率增加。与之类似，较高的R_0将导致被识别为外显子的基因片段变短变少，“漏警率”上升，即识别出编码区域遗漏真实外显子区域的几率增加。合适的R_0是综合考虑“虚警率”“漏警率”的一种折衷，为定量的分析问题，需要定义合适的指标评价R_0选取的优劣。

考虑识别算法预测出的外显子区域、内含子区域与实际基因中外显子区域与内含子区域的关系，可以归纳为以下四种情况：

1. 预测为外显子区域，实际外显子区域，记录其序列长度为 TP
2. 预测为外显子区域，实际内含子区域，记录其序列长度为 FP
3. 预测为内含子区域，实际外显子区域，记录其序列长度为 FN
4. 预测为内含子区域，实际内含子区域，记录其序列长度为 TN

定义针对某一固定阈值的识别灵敏度（sensitivity）为

$$S_n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

特征度（specificity）为

$$S_p = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

不难看出，灵敏度S_n表征的是“漏警率”的信息，S_n越高，识别算法效果越好，阈值选取越合理；特征度S_p表征的是“虚警率”的信息，S_p越高，识别算法效果越好，阈值选取越合理。对于不同的灵敏度阈值R_0，S_n、S_p随之变化，但遗憾的是二者并不能同时的增减。

考虑工程实际需求，一般将特征度作为容忍指标，我们选取≥90%、≥80%、≥70%三个特征度S_p指标分别作为先决条件。为了获取合适的信噪比阈值R_0，使用已知编码区域的有代表性的基因序列进行训练，将R_0在一个较大的区间进行小间距滑动（例如在 0.5~3 区间以 0.01 步进递增），则将获得一系列的$\{S_n, S_p\}$序列对，根据预定的特征度S_p指标，在满足条件的$\{S_n, S_p\}$对中选取使S_n最大的$\{S_n, S_p\}$对，则与之对应的R_0为对于该训练序列最好的判别阈值。

在实际中我们使用附件提供的 100 个人和鼠类基因序列的一部分进行判别阈值确定的训练，得到判别阈值R_0。然后再对剩余部分使用得到的R_0进行基因识别，得到相应的$\{S_n, S_p\}$，进而对该阈值确定方法进行评价。最后，对不同的物种类别，确定不同的阈值。

4.2.3 阈值确定方法及结果

采用基于固定长度滑动窗口上频谱曲线的基因识别的方法，选取 100 个人和鼠类 DNA 样本中的奇数序号的 50 个样本作为训练样本集。首先统计样本集

中基因编码外显子长度，得到该物种类别外显子长度的分布，如图所示(Y 表示外显子序列长度大于 X 值的概率)。

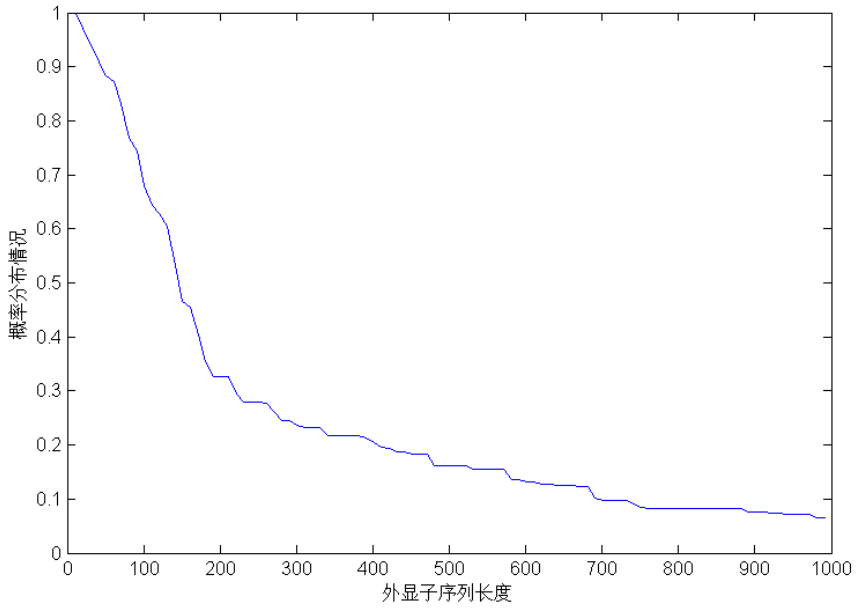

图 4-2-1 外显子长度分布情况

从图中可以看出，人和鼠类基因中外显子长度为 100bp~200bp 占 40%，100bp 以下的短序列为 30%，200bp 以上的 30%。由 3.4 节所述，当滑动窗长度与外显子长度接近时，识别效果较好，故选择 147bp 作为滑动窗的长度。

采用滑动窗口法对基因序列进行信号处理，得到频谱曲线如图所示，可以看出所得到的曲线的毛刺较多，高频波动明显。为了帮助识别，在实际实现中，算法中采用了一个四阶巴特沃兹滤波器对高频随机噪声进行处理。

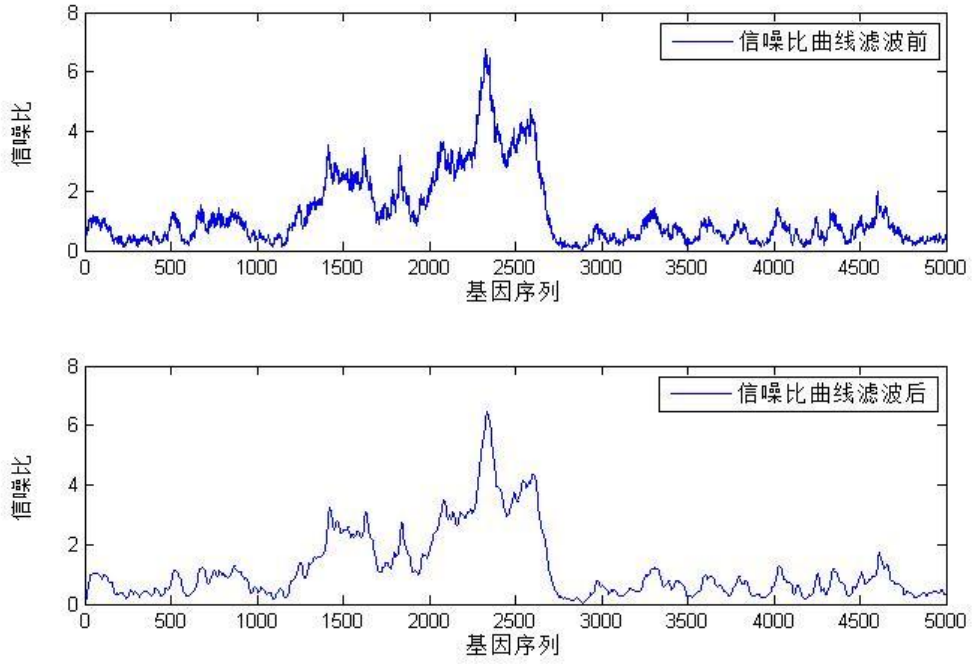

图 4-2-2 滤波器效果图

上图为直接使用滑动窗口法得到的频谱曲线，下图为加入滤波器后的频谱曲线。滤波后会造成信噪比曲线细节的丢失，并可能造成信噪比幅值的降低，但是对于接下来结合采用的移动序列方法有重要的作用。在修正固定滑动窗口方法中寻找出的外显子边界时，需要对曲线进行差分，这会加重信噪比曲线的毛刺，从而增大误差。因此对于计算过程中得到的信噪比曲线进行一次滤波是需要并且有效的。

针对频谱曲线，将其中 K/3 处 SNR 大于阈值处判别为外显子，小于阈值处判别为内含子。依次选取阈值R_0为 0.5,0.52,0.54……2.0，得到相应的S_n、S_p随R_0的变化曲线。

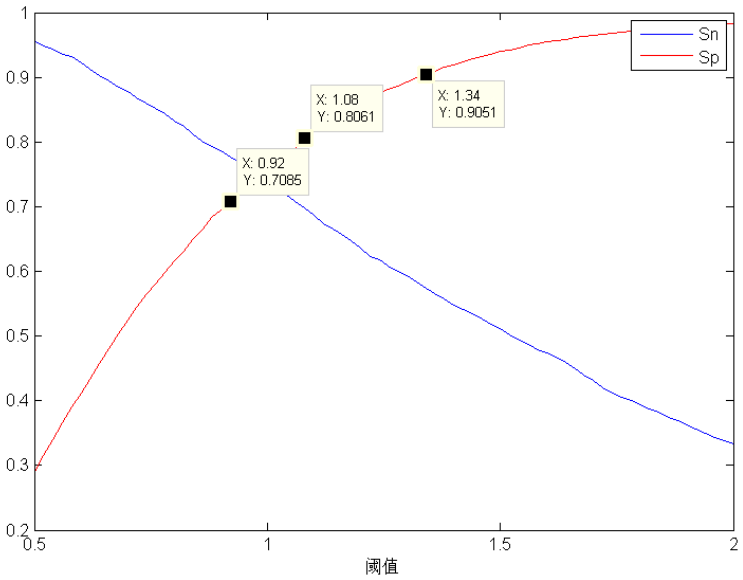

图 4-2-3 S_n、S_p随R_0的变化曲

进而分别得到特征度 90%，特征度 80%，特征度 70%意义下，灵敏度最高的S_{n_90}、S_{n_80}、S_{n_70}，及相应的阈值 R_{90}、R_{80}、R_{70}。软件运行结果如表所示

表 4-2-1 特征度与高灵敏度的阈值关系

	R	Sn
特征度 90%	1.34	0.574
特征度 80%	1.08	0.6982
特征度 70%	0.92	0.7773

然后，我们将 100 个人和鼠类 DNA 样本中偶数序号的 50 个样本分为三份，分别为 5、15、30 个样本，使用得到的 R_{90} 阈值对其进行基因编码序列识别，得到相应的S_n、S_p如表所示。

表 4-2-2 样本中S_n、S_p的值

样本数	Sn	Sp
5	0.7041	0.7822
15	0.5576	0.9139
30	0.5983	0.9088

可以看出当样本数达到一定数值的时候，测试得到的 Sn 和 Sp 都与之前相近，说明该阈值确定方法是行之有效的。

最后对所给的不同物种类型基因分别确定信噪比判别阈值 R_0，及相应的特征度 90%意义下的灵敏度S_n值。

表 4-2-3 对于不同类型生物的阈值和 Sn

物种	R	Sn
哺乳动物	1.26	0.5899
酵母菌类	1.06	0.4974

4.3 基因识别算法

4.3.1 基因识别算法的实现

识别算法主要采取固定长度的滑动窗口法与移动窗口法相结合的方式进行基因识别，在计算过程中加入一些滤波、填补、检测调整等辅助方式，从而得到较为精确的外显子识别算法。

基于滑动窗口的信噪比方法可以找出外显子片段的位置，如图 4-3-1 所示为仅仅采用这种方法求解出的外显子片段。由统计可以知道，极少数的外显子片段长度小于 50bp。对于信噪比曲线的振荡会在一段大于阈值的片段中产生低于阈值的狭缝，往往这种极短的狭缝内的序列仍然是外显子。算法中会检测并且填补这些细碎的片段，使得外显子片段更加完整。

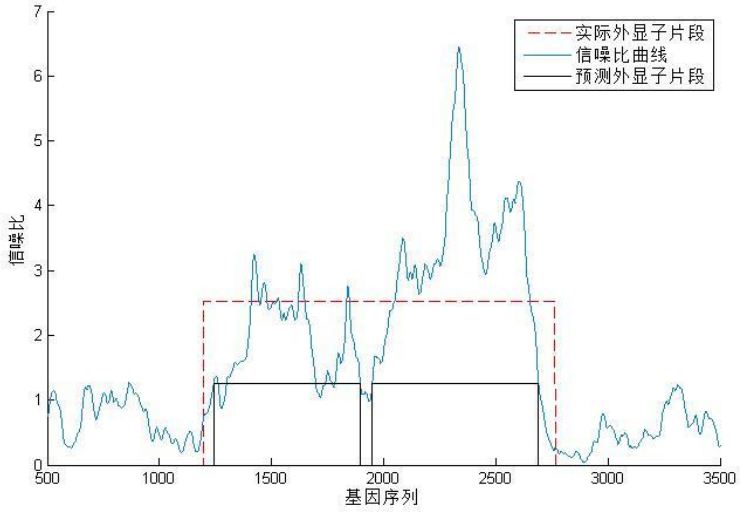

图 4-3-1 基于固定长度的滑动窗口法预测（AF223321）

基于固定长度的滑动窗口方法，对于预测外显子具有方向性，能够估计出外显子片段的大概位置，但是这种方法本身对于外显子边界并不敏感，因此对其边界的估计有较大的误差。然而，移动序列方法对于探测外显子边界有较大的优势，由于移动序列的一短固定为总 DNA 序列的起始，因此该方法对于移动的前端基因数据十分敏感。当前端位置 k 处仍为外显子的时候，R[k]曲线通常是上升趋势，直到 R[k]在 k_m 处达到局部峰值之后开始下降，则 k_m 处即为外显子片段的末尾段。

由于移动序列方法对前端敏感的特点，为了找到外显子两端的边界，实验中采用正向和逆向两个方向的移动序列方法对外显子边界进行预测。正向时，移动序列一端取定为 1，另一端逐步增大向后推进；反向时，移动序列取定基因段的末尾，另一端逐步减小向前推进。

实际计算中，先对移动序列的信噪比曲线 R[k]进行滤波，再对滤波后的数据进行差分。差分以后的曲线 $R'[k]$ 中，对于原移动序列 R[k]的上升段，$R'[k]$ 为正值；原移动序列的下降段，差分函数 $R'[k]$ 为负值。接着，将 $R'[k]$ 变为二值函数 B[k]，即 $R'[k]$ 大于 0 的部分置为 1，小于 0 的部分置为 0。对于 B[k]进行一次差分，得到的 $B'[k]$ 反应了 B[k]的上升沿和下降沿位置。将反向移动序列曲线整合入正向移动序列曲线的统一坐标，对于正向移动序列曲线，我们只关注 B[k]的下降沿，即原信噪比曲线 R[k]到峰值开始下降的位置，表征的是外显子的末端；同理，对于反向移动序列曲线，我们关注 B[k]的上升沿，即外显子的起始端。

基于固定长度的滑窗方法所找出的外显子片段，结合移动序列法，在边界上通过移动序列寻找的边界进行修正。将修正后的外显子片段（如图 4-3-2 所示）与图 4-3-1 中进行对比，可以明显发现此方法的有效性。

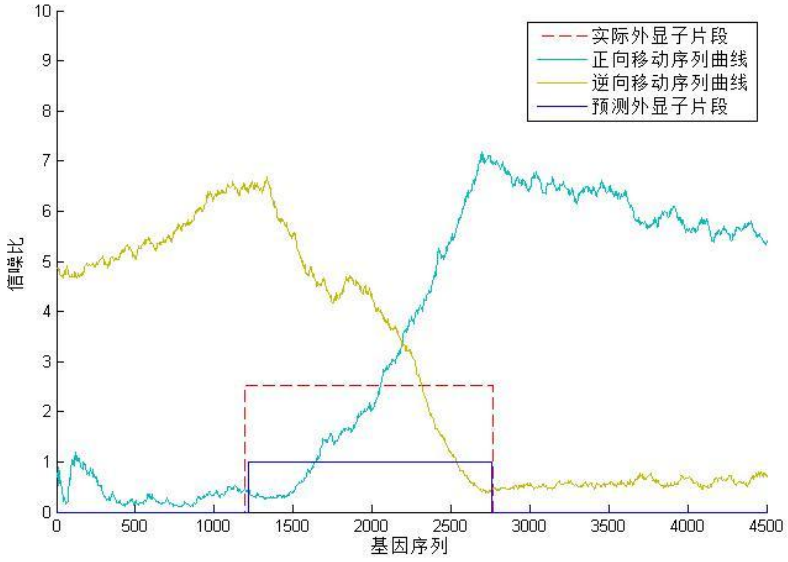

图 4-3-2 固定长度滑窗法与移动序列结合的优化方式

4.3.2　对未知 DNA 序列编码区间的预测

对于未知的 DNA 序列编码区间的预测，首先需要根据 DNA 物种类别选取适当的判别阈值 R_0。根据分子生物学的观点，不同物种 DNA 序列中 C-G 碱基所占的比例不同，同一物种 DNA 序列中 C-G 碱基所占的比例在一个较小区间内波动，不同物种亲缘性越近 DNA 序列中 C-G 碱基越接近。一般来讲，C-G 比例差异比较大的两个长 DNA 序列一定不属于同一物种；而 C-G 比例差异较小的两个长 DNA 序列有可能为同一物种或亲缘关系较近，也有可能不为同一物种且亲缘关系较远。由于所给待预测基因的物种未知，则估且假设碱基中 C-G 比例接近的 DNA 属于同一物种或亲缘较近物种的 DNA。计算所给的几种典型物种类型 DNA 中 C-G 碱基比例如下表所示：

表 4-3-1 不同生物的 C-G 碱基比例

类型	C-G 碱基比例
酿酒酵母类	0.3890
线虫粘粒类	0.3742
拟南芥植物类	0.3831
人线粒体类	0.4436
人和鼠类	0.5247
哺乳动物类	0.5032

再分别对未知序列求取碱基序列中的C-G比例依次为0.5253，0.5483，0.3896，0.5952，0.4614，0.5268。则结合物种分类方法及问题二中得到的各物种类型阈值表，选择合适阈值对 DNA 序列进行基因预测。

1. 未知基因序列一

长度 5800bp 序列中 C-G 比例为 0.5253 选取阈值 R_0 为 1.28

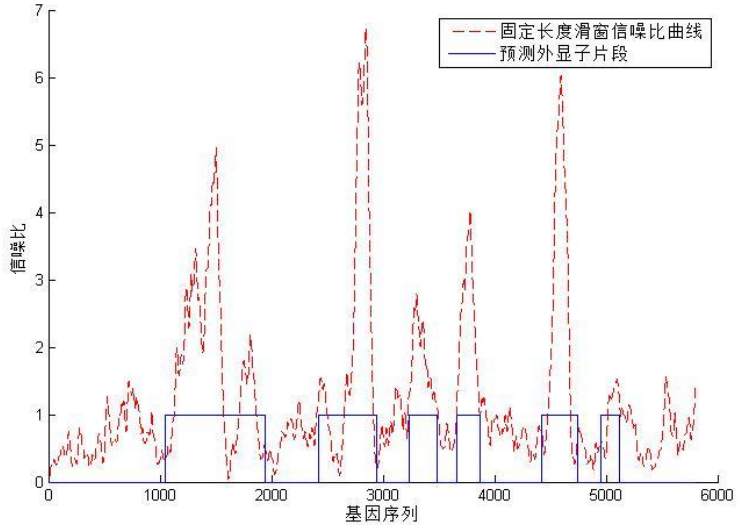

图 4-3-3 未知序列一外显子预测

基因编码区识别为：1066..1932，　2503..2937，　3247..3483，
3661..3954，　4456..4776，　4969..5127

2. 未知基因序列二

长度 7894bp 序列中 C-G 比例为 0.5483 选取阈值 R_0 为 1.28

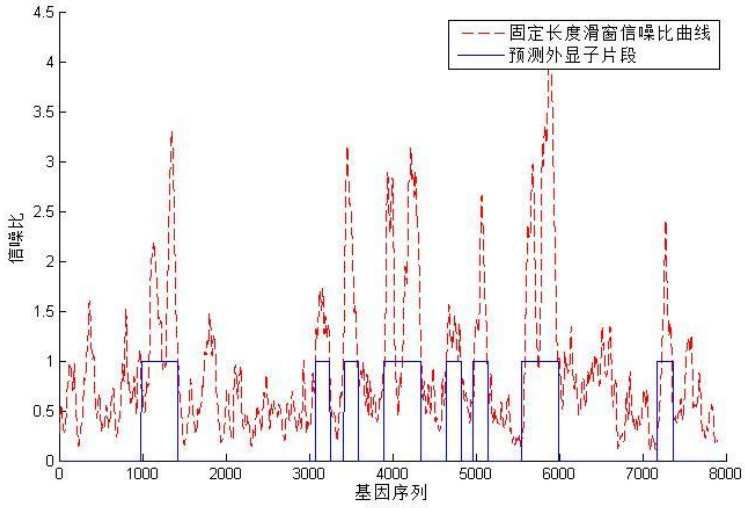

图 4-3-4 未知序列二外显子预测

基因编码区识别为：949..1377， 3115..3276， 3388..3585，
3967..4296， 4621..4827， 4984..5178，
5599..6072， 7204..7404

3. 未知基因序列三

长度 5383bp 序列中 C-G 比例为 0.3896 选取阈值 R_0 为 1.22

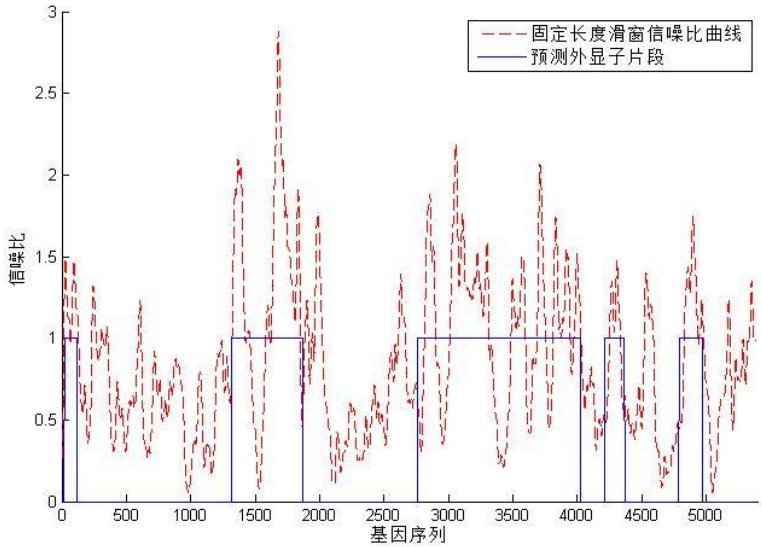

图 4-3-5 未知序列三外显子预测

基因编码区识别为：28..111， 1330..1878， 2758..4119，
4198..4344， 4804..4992

4. 未知基因序列四

长度 6301bp 序列中 C-G 比例为 0.5952 选取阈值 R_0 为 1.44

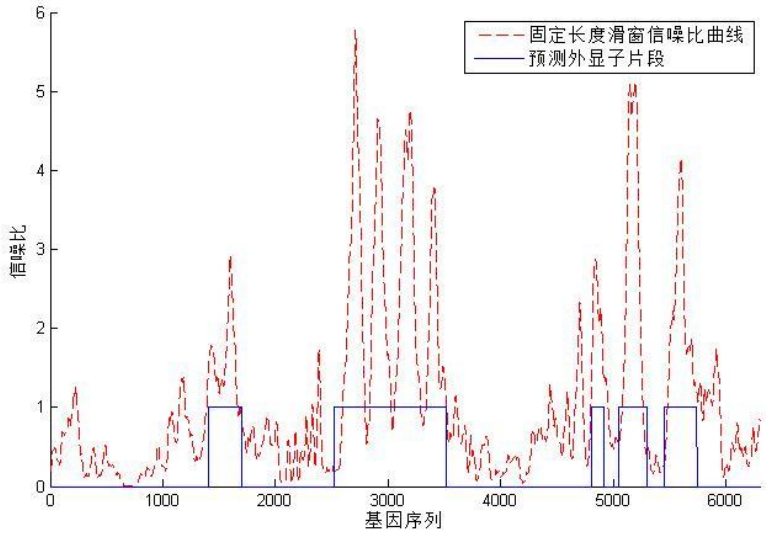

图 4-3-6 未知序列四外显子预测

基因编码区识别为：1423..1704， 2620..3495， 4840..4929，
5077..5322， 5446..5709

5. 未知基因序列五

长度 14206bp 序列中 C-G 比例为 0.4614 选取阈值 R_0 为 1.26

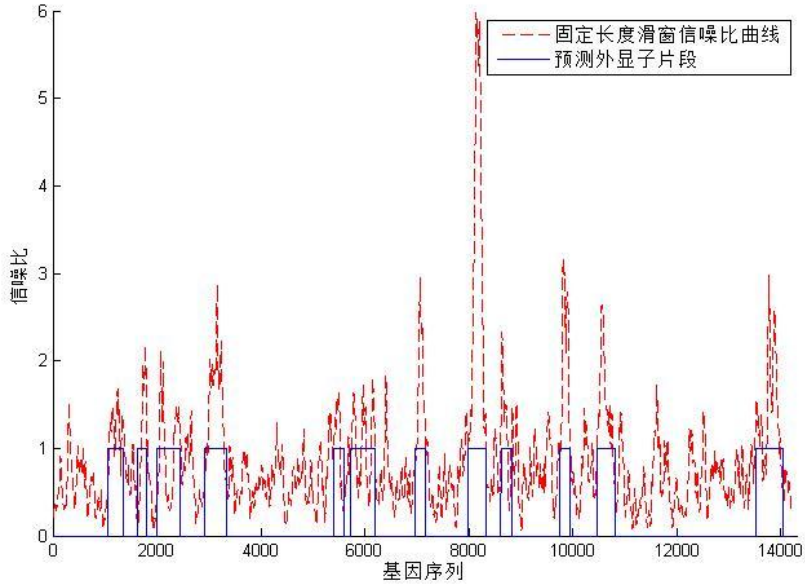

图 4-3-7 未知序列五外显子预测

基因编码区识别为：1072..1383， 1636..1827， 1987.. 2511，
2944..3375， 5389..5622， 5764..6234
6976..7230， 8008..8345， 8617..8835，
9742..9948， 10435..10806，13498.. 13986

6. 未知基因序列六

长度 4842bp 序列中 C-G 比例为 0.5268 选取阈值 R_0 为 1.28

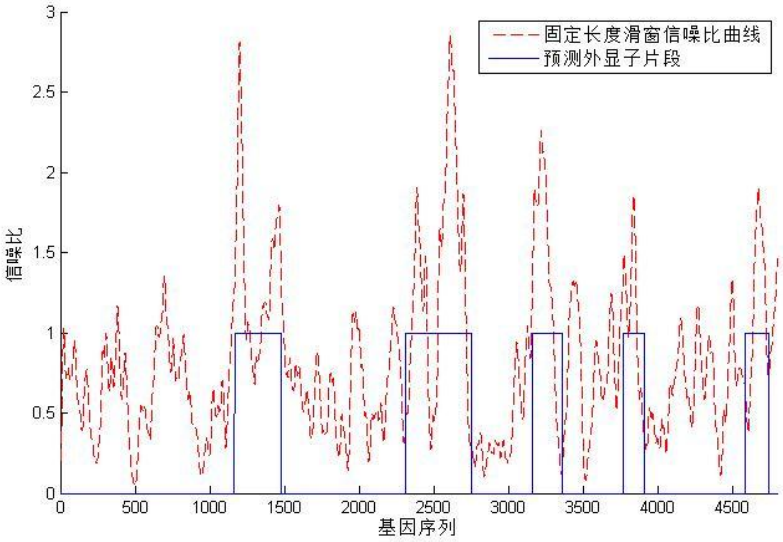

图 4-3-8 未知序列六外显子预测

基因编码区识别为：1180..1494， 2305..2775， 3154..3354，
3790..3912， 4597..4749

4.4　信号处理与基因编码序列突变的识别

众所周知，目前发现的大部分 DNA 编码区间都具有频率 3-周期性。基因的某些突变会改变 DNA 编码的 3-周期性，从而改变了外显子基因段频谱的形状，例如削弱 $P[\frac{N}{3}]$ 处的信噪比 $P[\frac{N}{3}]$ 的值，或者在其他频率位置产生不规律的次峰值。对于单个核苷酸的替换，在频谱分析中是难以判断出来的，然而对于核苷酸的插入和删除，很多情况下是可以观察频谱分析出来。图 4-4-1 所示为四种不同的突变，突变的核苷酸位于距离序列开头 $\frac{1}{6}$ 个外显子片段长度的位置。

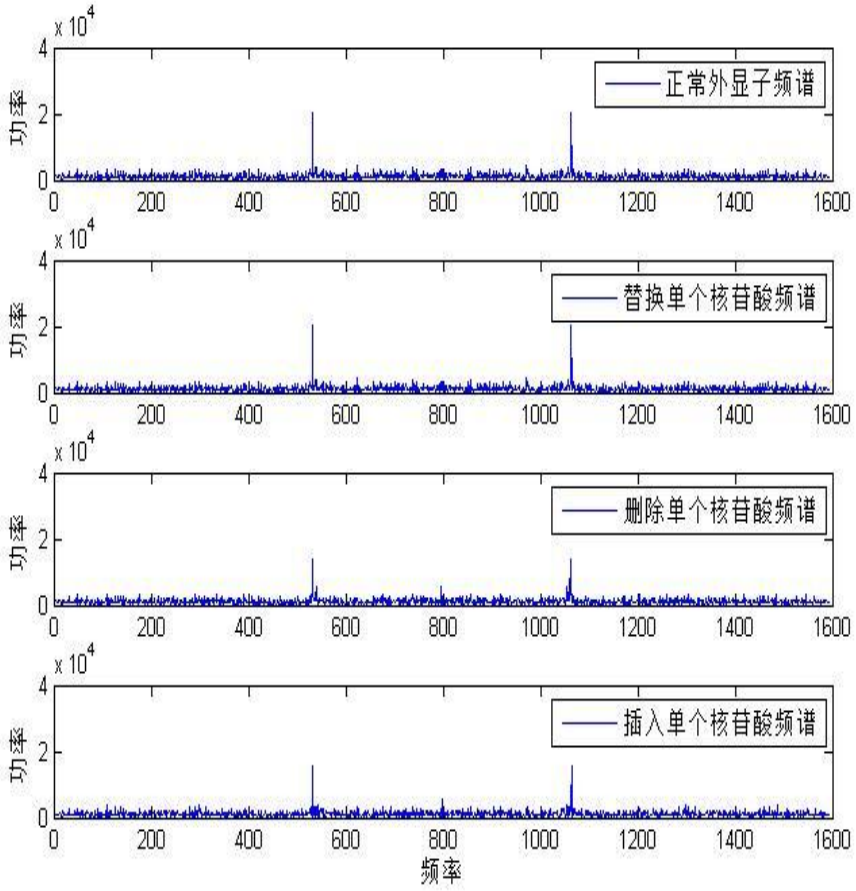

图 4-4-1 序列长度 $\frac{1}{6}$ 处核苷酸突变的频谱变化（AF019045 外显子片段）

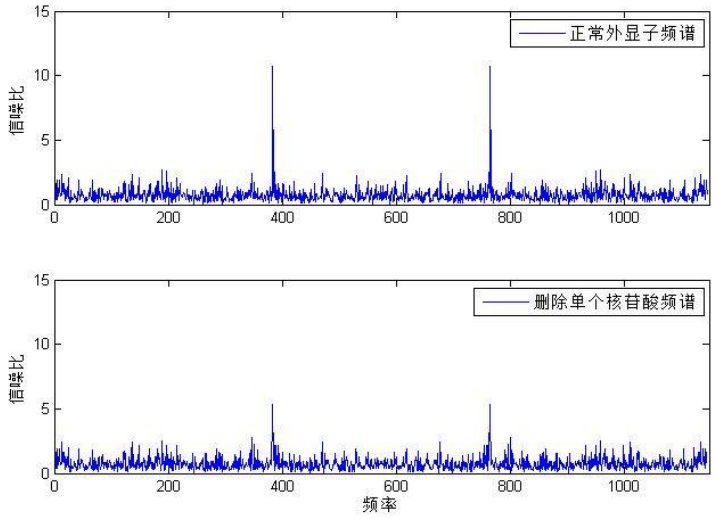

图 4-4-2 序列长度 $\frac{1}{6}$ 处核苷酸突变的频谱变化（U76254 外显子片段）

由两幅图容易看出，单个核苷酸的替换很难从频谱中发现，删除跟插入单个核苷酸两种情况比较类似，都会改变频谱的形状。如图 4-4-1，频谱中心频点 $P[\frac{N}{2}]$ 处出现了一个小峰值，这表明由于增减一个核苷酸以后，使得整个编码序列具有微弱的 2-周期性。另一种情况如图 4-4-2 所示，由于核苷酸的删除使得 $P[\frac{N}{3}]$ 处的值大幅降低，能量被分散到其他频点处，序列的 3-周期性被削弱。

由于基因删除、插入所造成的频谱变化受到突变在 DNA 序列中的位置影响，对于图 4-4-1 中的外显子序列，当突变位置从 $\frac{1}{2}$ 向 $\frac{1}{8}$ 改变时， $P[\frac{N}{2}]$ 处幅值先增后降。

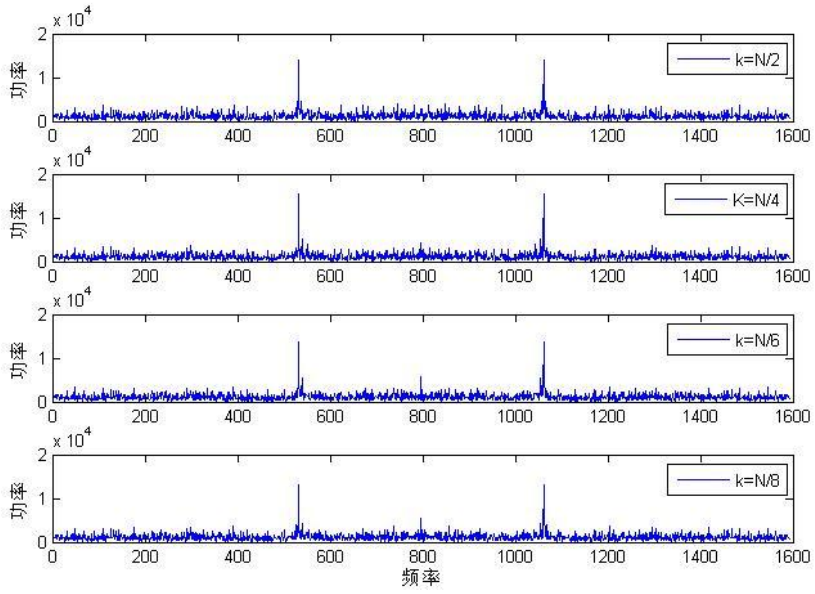

图 4-4-3 删除单个核苷酸的频谱随位置变化情况（AF019045 外显子片段）

由对称性，只考虑 k 到$\frac{1}{2}$的情况。如图 4-4-3 所示，$P[\frac{N}{2}]$的幅值随着删除核苷酸的位置而变化。通过比较频谱$P[\frac{N}{2}]$的幅值（或者$P[\frac{N}{2}]$与$P[\frac{N}{3}]$的相对值），可以粗略的估计出基因突变的位置。

图 4-4-4 显示了 AF019045 外显子片段的频谱随删除的核苷酸位置系数变化的曲线（被删除的核苷酸位置实际为 kN 的整数部分），受到样本中核苷酸数量的限制，曲线存在较大的随机噪声，然而其整体趋势还是很明显的。可以通过对比曲线大致找到发生突变的位置，大大减小搜索范围。对于统计特性略微不同的外显子片段，其突变后的频谱有所不同，但是都会导致$P[\frac{N}{3}]$减小，或者产生杂谱，$P[\frac{N}{3}]$减小的程度和杂谱的幅值都能够反映出突变的位置。由于时间关系，这里只做定性的分析，留待下一步研究。

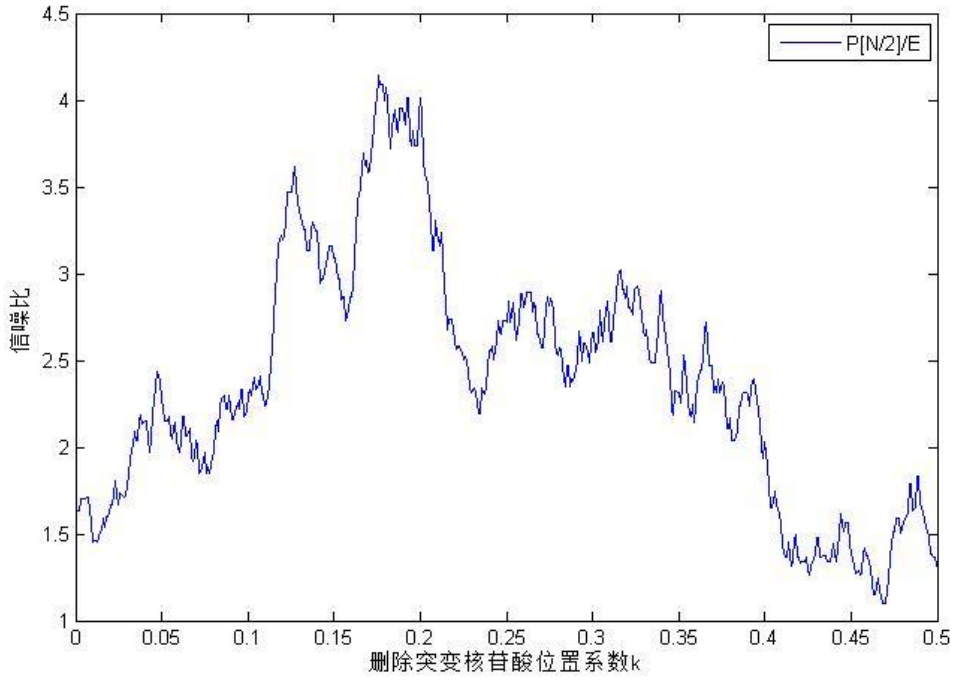

图 4-4-4 信噪比随删除位置系数 K 变化曲线

5　模型的总结与改进

1．判别阈值确定模型

总结：考虑到相同或亲缘相近物种在基因序列核苷酸排列上具有的相似性，选取同一物种的多个 DNA 序列进行阈值训练。以定量的识别灵敏度和识别特征度作为阈值优劣的评判标准，得到一定识别特征度意义下的灵敏度最高的最优阈值估计。该模型的精确程度在很大意义上取决于用于训练的 DNA 样本的代表性。所选样本自身的频谱 3-周期性是否显著、样本集中各序列的核苷酸排布相似度即物种的亲缘程度、训练样本集个数的多寡等多方面因素将直接影响阈值确定的准确度。

思考：目前的编码区域评判标准为一单阈值二分区间，即只有高于阈值作为外显子和低于阈值作为内含子两种分类。考虑是否可以引入双阈值三分区间的分类标准。两个阈值一高一低将信噪比区间分为三份：对于高于高阈值的区域将其归类为编码区域；低于低阈值的区域归为非编码区域；对于介于两个阈值之间的编码区域则保持与前一区域相同的判别分类。

4. 编码序列识别模型

总结：在识别模型中，对某些曲线先用信号处理方法进行滤波来减小随机噪声，采用“滑动窗口法”搜寻出编码序列大概位置，针对识别序列零碎，采用探测填补等辅助方式；针对端点模糊，创造性的引入双向“移动序列法”，在编码大概区域稍大的临域内精确地搜寻编码区域的端点，获得了更为精确的编码序列识别效果

思考：目前的识别算法是基于 DNA 序列 Fourier 变换的幅频特性——频谱 3-周期性设计得到的。猜想编码区域与非编码区域 DNA 序列的 Fourier 变换的相频曲线也有较大不同，也可以提取出作为识别编码序列的特征信息。相信合理设计的结合了 Fourier 变换的幅频特性和相频特性的编码序列识别算法将有更好的灵敏度和特征度。

参考文献

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

【2】Anastassiou, D., 2000. Frequency-domain analysis of biomolecular sequences. Bioinformatics 16, 1073–1081.

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

【4】Berryman, M. J., Allison, A., 2005. Review of signal processing in genetics. Fluctuation and Noise Letters. 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. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of Theoretical Biology. 247, 687–694.