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 **DNA 序列表示及基因识别方法研究**

摘 要：

本文就 DNA 序列表示及基因识别算法实现的相关问题进行了研究，取得了以下几方面的成果。

1. 功率谱与信噪比的快速算法

- 针对 Voss 映射，给出了计算基因序列功率谱或信噪比的快速 Fourier 变换和 AR 模型，仿真实验结果表明，计算效率有所提升。经过理论推导，建立了功率谱、信噪比与 DNA 序列中核苷酸出现的频次之间的关系，即为 SNR-F 公式：

$$R=\frac{N_A}{N}\cdot R_A+\frac{N_C}{N}\cdot R_C+\frac{N_G}{N}\cdot R_G+\frac{N_T}{N}\cdot R_T$$

  利用该公式，计算功率谱与信噪比将不再需要离散 Fourier 变换等计算量较大的运算，只需要对 DNA 序列中核苷酸出现的频次进行统计，然后进行简单的数值运算即可，有效提升了功率谱与信噪比的计算效率。
- 推导出了 Z-curve 映射的功率谱与信噪比和 Voss 映射下的功率谱与信噪比之间的数值关系，即为：

$$E_z=4E \text{ 和 } R_z=R$$

  并从理论基础、生物学意义和特征三个方面对 Z-curve 映射和 Voss 映射进行了对比分析，刻画出了两种映射之间更深层次、更全面的关系。
- 经过理论推导，给出了一般的实数映射下功率谱、信噪比的快速计算公式，将其功率谱、信噪比的计算简化为核苷酸出现频次的统计和简单数值运算，极大简化了实数映射下功率谱与信噪比的计算。

2. 对不同物种类型基因的阈值确定

- 本文结合重采样技术，提出了最佳阈值确定算法，能为每一个特定种类的生物推测其最佳阈值。模型能够针对不同生物基因的结构特征，启发式地为其推断出一个最佳的预测阈值。仿真实验结果表明，附件中所给的人和鼠类生物基因预测的最佳阈值为 1.7773，200 个哺乳动物类的基因预测的最佳阈值为 2.18。在合理确定窗口大小的基础上，利用该最佳阈值能显著提高基于功率谱分析方法的基因预测精度，同时还可用来预测该生物目前尚未标注确认的其它基因。

3. 基因识别算法的实现

- 针对基因识别算法的设计与实现问题，本文首先利用基于 AR 模型重采样的基因预测方法对附件中给出的6个未被注释的DNA序列的编码区域进行了预测。然后，结合数字滤波器与信噪比快速计算公式，提出了一种基于 SNR-F 的基因识别模型。该模型克服了现有 Fourier 方法对 DNA 序列长度的限制，并且能够提高实现效率。最后，利用该模型对未被注释的 DNA 序列的编码区域进行了预测。两种预测方法相结合有助于提高基因预测的精度，同时使后期基因识别更具有针对性。

4. 延展性问题

- 针对目前常用的基因识别算法对特征选取的主观性，建立了基因识别特征的动态筛选模型。该模型在训练中充分选取基因的多类特征作为候选特征，构造编码区与非编码区的正负数据集，运用特征筛选方法在数据集中提取主特征，以达到优化特征集、减少冗余度的目的。同时，模型用组合向量的方式实现多类特征的融合，将序列转换成特征空间中的向量，通过利用判别分析的方法达到识别的目的。特征的筛选和组合提高了基因识别算法的合理性和信息利用率，预测精度达到了 98%以上，高于已有算法的预测精度。
- Z-曲线的提出表明利用几何工具可以有效地分析 DNA 序列，受此启发，本文基于改进的基于 DNA 序列的“四线”图，提出了基于改进“四线”图的 DNA 序列突变分析模型，为检测基因突变提供模型基础。

随着人类基因组计划的顺利完成，基因识别已成为生物信息学中最基础、最首要的问题。本文就基因识别方法的相关问题进行了深入探讨，提出了一些新的思路，期待有益于基因识别领域的后续研究。

关键词：基因识别，功率谱，信噪比，AR 模型，阈值，重采样

一、问题背景

DNA 是生物遗传信息的载体，DNA 序列由腺嘌呤（Adenine，A），鸟嘌呤（Guanine，G），胞嘧啶（Cytosine，C），胸腺嘧啶（Thymine，T）这四种核苷酸（nucleotide）符号按一定的顺序连接而成。其中带有遗传讯息的 DNA 片段称为基因（Gene）。其他的 DNA 序列片段，有些直接以自身构造发挥作用，有些则参与调控遗传讯息的表现。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（Coding Sequence）片段，称为外显子（Exon），不编码的部分称为内含子（Intron）。

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。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。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。通过对 DNA 序列进行 Voss 映射，可以发现，对于同一段 DNA 序列，其外显子与内含子序列片段的功率谱通常表现出不同的特性：外显子序列具有频谱 3-周期性而内含子没有。频谱峰值特征的发现，或者频谱与信噪比概念的引入，有助于探测、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。已经有一些研究者提出了识别基因的算法，目前利用信噪比的基因识别算法通常有两种：固定长度窗口滑动法和移动信噪比曲线识别法。

二、模型假设

1．题目中所列数据均真实可靠且具有较强的代表性；
2．在解决前三个问题的过程中不考虑基因突变问题；
3．不考虑软件工具在数据处理及图形绘制中的误差。

三、符号说明

DFT：离散 Fourier 变换

FFT：快速 Fourier 变换

$u_b[n]$：DNA 指示序列

$U_b[k]$：Fourier 变换序列

$P(\ )$：功率谱函数

R：信噪比

σ^2：方差

a^*：a 的共轭

$H(z)$：系统转移函数

ρ：误差功率

四、问题一模型的建立与求解

问题一：

（1）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？

（2）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；

（3）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？

问题分析：

对 Voss 映射，功率谱与信噪比的快速计算方法是多种多样的，从思路上，我们先后尝试了两种方法，分别是引入快速 Fourier 变换和 AR 模型，算法效率有了有效提升，但提升的效果并不显著，这引导我们从理论推导上简化功率谱与信噪比的计算公式，建立功率谱、信噪比与 DNA 序列中核苷酸出现的频次之间的关系，从而彻底简化功率谱与信噪比的计算。

4.1 和 4.2 小节将分别对快速 Fourier 变换和 AR 模型进行描述。在 4.3 小节，本文将重点给出功率谱、信噪比与 DNA 序列中核苷酸出现的频次之间的关系。基于此，4.4 和 4.5 小节将分别讨论 Z-curve 映射和 Voss 映射的关系及实数映射下功率谱与信噪比的快速计算公式。

4.1 快速 Fourier 变换

依据材料所述，对于很长的 DNA 序列，在计算其功率谱或信噪比时，首先要对 DNA 的指示序列做离散 Fourier 变换(DFT)

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-j\frac{2\pi nk}{N}}, \quad k = 0,1,\cdots,N-1 \tag{4.1}$$

但是，上述 DFT 的计算量太大，很难高效进行基因识别。因此，本文首先采用快速 Fourier 变换(FFT)对功率谱或信噪比的计算方法进行改进，FFT 并不是一种新的变换，而是 Cooley 和 Tukey 于 1965 年提出的计算 DFT 的一种快速算法，此算法将 DFT 的运算量减少了几个数量级。

在利用 DFT 计算式(4.1)时，算出全部 N 点 $U_b[k]$ 共需次 N^2 复数乘法和 $N(N-1)$次复数加法，即计算量是与 N^2 成正比的，DFT 直接变换的计算复杂度是 $O(N^2)$ 。FFT 的基本思想是将大点数的 DFT 分解为若干个小点数 DFT 的组合，从而减少运算量。FFT 可以计算出与 DFT 直接计算相同的结果，但只需要 $O(N\log N)$ 的计算复杂度。通常，FFT 要求 N 能被因数分解，但不是所有的快速 Fourier 变换都要求 N 是合数，对于所有的整数 N ，都存在复杂度为 $O(N\log N)$ 的快速算法。因此，对于本题目中的能够被 3 整除的整数 N ，利用 FFT 可将式(4.1)的计算复杂度降至 $O(N\log N)$ ，从而相应提高了功率谱与信噪比的计算效率。

本文不再对FFT的具体算法进行描述，MATLAB工具中也已经提供了进行FFT计算的相关函数。利用MATLAB 2011a，本文对题目中的酵母基因DNA序列频谱3-周期性进行了验证，在剔除 $k=0$ 处（实际上，此处的数值在信号处理与分析系统中的意义为时域数据的直流分量，对于研究DNA编码序列而言为噪声信号）的数据后，所得结果与题目中图3所示一致。

然而，功率谱及信噪比的计算方式并不仅局限于利用Fourier变换一种方法来实现，本文接下来主要讨论不利用Fourier变换进行功率谱及信噪比的计算方法。

4.2 AR 模型

在用DFT算法计算功率谱及信噪比时，其存在固有的缺陷，比如存在泄漏误差和混迭误差，分辨率低，不适于处理短数据，谱线不平滑，起伏剧烈，难以拟合出光滑曲线等。针对经典谱估计的分辨率低和方差性能不好等问题，为此人们提出参数谱方法（现代功率谱估计）。参数谱估计方法是通过观测数据估计参数模型再按照求参数模型输出功率的方法估计信号功率谱。

参数谱估计的主要方法有最大熵谱分析法(AR模型法)、Pisarenko谐波分解法、Prony提取极点法、Prony谱线分解法以及Capon最大似然法等。其中AR模型应用较多，具有代表性。

4.2.1 AR 模型的公式表达

我们首先介绍 AR 模型的一般模型——ARMA 模型。

ARMA 模型[1]功率谱的数学表达式为：

$$P_k(e^{j\omega})=\sigma^2\left|1+\sum_{k=1}^{p}b_k e^{-j\omega k}\right|^2 \Big/ \left|1+\sum_{k=1}^{p}a_k e^{-j\omega k}\right|^2 \tag{4.2}$$

其中 σ^2 是白噪声的方差， $P_k(e^{j\omega})$ 为功率谱密度， a_k 和 b_k 为模型参数。

如果 ARMA 模型的参数 $b_1,b_2,\cdots,b_p$ 全为 0，就演化为 AR 模型：

$$P_k(e^{j\omega})=\sigma^2 \Big/ \left|1+\sum_{k=1}^{p}a_k e^{-j\omega k}\right|^2 \tag{4.3}$$

如果 ARMA 模型的参数 $a_1,a_2,\cdots,a_p$ 全为 0，就演化为 MA 模型：

$$P_k(e^{j\omega}) = \sigma^2 |1+\sum_{k=1}^{p} b_k e^{-j\omega k}|^2 \tag{4.4}$$

在实际中，AR 模型的参数估计比较简单，对其有比较充分的研究，而对于 ARMA 模型，其参数比较复杂，对其算法的研究和改进还在完善中，故本节对功率谱及信噪比的快速计算采用 AR 模型。

4.2.2 AR 模型功率谱的估计方法

在利用 AR 模型进行功率谱估计时，必须计算出 AR 模型的参数和激励白噪声序列的方差。目前，AR 模型中参数的提取算法有很多，主要包括自相关法、Burg 算法、协方差法、改进的协方差法，以及最大似然估计法等。

1. AR 模型 Yule-Walker 方程的建立

AR 模型，又称为自回归模型，是一个全极点的模型，假设 AR 模型的差分方程用下式表示：

$$x(n) = -\sum_{i=1}^{p} a_{p,i} x(n-i) + u(n) \tag{4.5}$$

其中，$u(n)$ 是均值为零、方差为 σ^2 的白噪声序列，p 是 AR 模型的阶数，$a_{p,i}, (i = 0,1,\cdots, p)$ 是 p 阶 AR 模型的参数。

假设 AR 模型的系统转移函数用下式表示：

$$H(z) = \frac{1}{1+\sum_{i=1}^{p} a_{p,i} z^{-i}} \tag{4.6}$$

从而得到 AR 模型的功率谱估计的计算公式：

$$P_x(k) = \frac{\sigma^2}{|1+\sum_{i=1}^{p} a_i W_N^{-ki}|^2} \tag{4.7}$$

要利用 AR 模型进行功率谱估计，必须得到模型参数和白噪声序列的方差。将式(4.5)变形，则有

$$R_x(m) = \begin{cases} -\sum_{i=1}^{p} a_{p,i} r_{x,m-i} & , m \geq 1 \\ -\sum_{i=1}^{p} a_{p,i} r_{x,i} & , m = 0 \end{cases} \tag{4.8}$$

式(4.8)的矩阵形式为

$$\begin{bmatrix} r_{x,x} & r_{x,x} & \cdots & r_{x,x} \\ r_{x,x} & r_{x,x} & \cdots & r_{x,x} \\ \vdots & \vdots & & \vdots \\ r_{x,x} & r_{x,x} & \cdots & r_{x,x} \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.9}$$

式(4.8)和式(4.9)称为 AR 模型的 Yule-Walker 方程[2]。下面我们介绍 AR 模型参数的求解方法。

2. AR 模型参数的求解——自相关法（Levenson-Durbin 递推法）

自相关法的出发点是选择 AR 模型的参数使预测误差功率最小，预测误差功率为

$$\rho = \frac{1}{N}\sum_{n=-\infty}^{\infty} |e(n)|^2 = \frac{1}{N}\sum_{n=-\infty}^{\infty} |x(n) + \sum_{i=1}^{p} a_{p,i}x(n-i)|^2 \tag{4.10}$$

假设信号 $x(n)$ 的数据区在 $0 \le n \le N-1$ 范围，有 p 个预测系数，N 个数据经过冲激响应为 $a_{p,i}, (i = 0,1,\cdots, p)$ 的滤波器，输出预测误差 $e(n)$ 的长度为 $N+P$，因此可得用下式计算:

$$\rho = \frac{1}{N}\sum_{n=0}^{N+P-1} |e(n)|^2 = \frac{1}{N}\sum_{n=0}^{N+P-1} |x(n) + \sum_{i=1}^{p} a_{p,i}x(n-i)|^2 \tag{4.11}$$

显然， $e(n)$ 的长度长于数据的长度，上式中数据 $x(n)$ 的两端需补充零点，这相当于无穷长的信号经过加窗处理，得到长度为 N 的数据。用式(4.11)对系数 $a_{p,i}$ 的实部和虚部求微分的方法使预测误差功率最小，得到

$$\begin{bmatrix} \hat{r}_{x,x}(0) & \hat{r}_{x,x}(-1) & \cdots & \hat{r}_{x,x}(-p+1) \\ \hat{r}_{x,x}(1) & \hat{r}_{x,x}(0) & \cdots & \hat{r}_{x,x}(-p+2) \\ \vdots & \vdots & & \vdots \\ \hat{r}_{x,x}(p-1) & \hat{r}_{x,x}(p-2) & \cdots & \hat{r}_{x,x}(0) \end{bmatrix} \begin{bmatrix} a_{p,1} \\ a_{p,2} \\ \vdots \\ a_{p,p} \end{bmatrix} = -\begin{bmatrix} \hat{r}_{x,x}(1) \\ \hat{r}_{x,x}(2) \\ \vdots \\ \hat{r}_{x,x}(p) \end{bmatrix} \tag{4.12}$$

式中，

$$\hat{r}_{x,x}(m) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1-m} x^*(n)x(n+m) & m = 0,1,\cdots, p \\ \hat{r}^*_{x,x}(-m) & m = -p+1, -p+2, \cdots, -1 \end{cases} \tag{4.13}$$

式(4.12)实质上是 Yule-Walker 方程，因此自相关法也是基于解 Yule-Walker 方程的一种方法。首先由信号的观测数据估计出其自相关函数，再解式(4.12)，得到模型参数，便可求出信号的功率谱，因此该方法也称为解 Yule-Walker 法。但是直接解该方程，需要计算逆在矩阵，计算复杂度较大，因此可利用 Yule-Walker 方程中自相关矩阵的性质，导出 Levenson-Durbin 递推法，这是一种高效的解方程方法。下面简要介绍 Levenson-Durbin 递推法求解 AR 模型参数

的基本流程[2]：

（1）估计观测序列的自相关系数矩阵；

（2）利用 Lenvinson-Durbin 递推算法求解 AR 模型参数。

Lenvinson-Durbin 算法是从低阶开始递推，直到 p 阶，给出了每一阶次的所有参数，这有利于选择合适阶次的 AR 模型。具体操作如下：

$$a_{k,k}=-\left[r_{x,x}(k)+\sum_{l=1}^{k-1}a_{k-1,l}r_{x,x}(k-l)\right]/\sigma_{k-1}^2 \tag{4.14}$$

$$a_{k,i}=a_{k-1,i}+a_{k,k}a_{k-1,k-i}^{*}\quad i=1,2,\cdots,k-1 \tag{4.15}$$

$$\sigma_k^2=(1-|a_{k,k}|^2)\sigma_{k-1}^2 \tag{4.16}$$

由 $k=1$ 开始递推，递推到 $k=p$，依次得到 $\{a_{1,1},\sigma_1^2\}$，$\{a_{2,1},a_{2,2},\sigma_2^2\}$，…，$\{a_{p,1},a_{p,2},\cdots,a_{p,p},\sigma_p^2\}$。求出 AR 模型的各个系数 $a_{p,i},(i=0,1,\cdots,p)$ 以及模型输入白噪声方差 σ^2 后，代入信号功率谱的计算公式：

$$P_{x,x}(\mathrm{e}^{\mathrm{j}\omega})=\sigma_w^2|H(\mathrm{e}^{\mathrm{j}\omega})|^2=\sigma_w^2\left|\frac{1}{1+\sum_{i=1}^{P}a_i\mathrm{e}^{-\mathrm{j}\omega i}}\right|^2 \tag{4.17}$$

即可计算出功率谱。

式(4.16)表明：$\sigma_1^2\geq\sigma_2^2\geq\sigma_3^2\geq\cdots\geq\sigma_p^2$，说明随着阶数增加，预测误差功率将减少或者不变，为此要求 $a_{k,k}\leq 1$，$a_{k,k}$ 称为反射系数。另外，递推公式提供了一种确定模型阶数的实验方法，如模型的阶数不知道，由低阶开始递推，当递推到 M 阶时，预测误差满足允许的值，停止递推，选 AR 模型的阶数为 M。这种递推法效率高，且当阶数变化时，无需从头计算。利用列文森递推法计算功率谱的流程如图 4.1 所示。

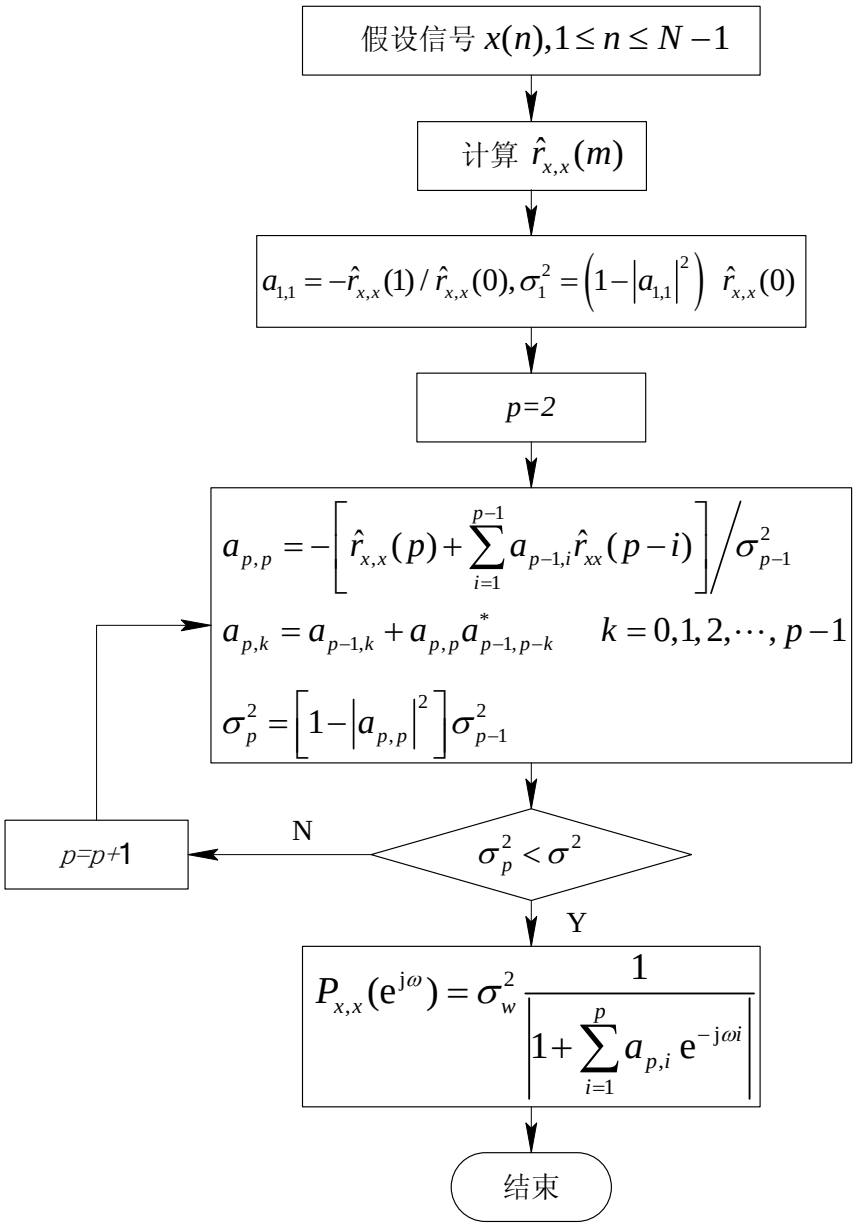

图 4.1 基于列文森递推法计算功率谱的流程图

3 AR 模型参数的求解——伯格（Burg）递推法

AR 模型参数的另外一种常用的求解方法为 Burg 递推法。Burg 算法与自相关法不同，它是使序列 $x(n)$ 的前后向预测误差功率之和

$$\rho_p^{fb}=\frac{1}{N-P}\sum_{n=p}^{N-1}\{|e_p^f(n)|^2+|e_p^b(n)|^2\} \tag{4.18}$$

达到最小。下面简要介绍 Burg 递推法求解 AR 模型参数的基本流程[2]：

（1）利用初始条件 $e_0^f(n)=x(0)$， $e_0^f(n)=x(0)$，求解

$$k_p = \frac{-2\sum_{n=p}^{N-1} e_{p-1}^{f}(n) e_{p-1}^{b*}(n-1)}{\sum_{n=p}^{N-1}(|e_{p-1}^{f}(n)|^2 + |e_{p-1}^{b}(n-1)|^2)} \tag{4.19}$$

（2）根据序列 $x(n)$ 的自相关函数 $\hat{r}_{xx}(0) = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2$， $\rho_0 = \hat{r}_{xx}(0)$，求出阶次 m=1 时的 AR 模型参数 $a_{1,1} = k_1$ 与前后向预测误差功率之和

$$\rho_1 = (1-|k_1|^2)\rho_0 = (1-|k_1|^2)\hat{r}_{xx}(0) \tag{4.20}$$

（3）由式

$$\begin{aligned} e_p^f(n) &= e_{p-1}^f(n) + k_p e_{p-1}^b(n) \quad n = p+1, p+2, \cdots, N-1 \\ e_p^b(n) &= e_{p-1}^b(n-1) + k_{p-1}^* e_{p-1}^f(n) \quad n = p, p+1, \cdots, N-2 \end{aligned} \tag{4.21}$$

前向预测误差 $e_1^f(n)$ 与后向预测误差 $e_1^f(n)$，然后由式(4.19)估计出反射系数 k_2；

（4）由式

$$\begin{aligned} a_{m,k} &= a_{m-1,k} + k_m a_{m-1,m-k} \\ a_{m,m} &= k_m \\ \rho_m^{fb} &= (1-k_m^2)\rho_{m-1}^{fb} \end{aligned} \tag{4.22}$$

的递推关系，求出阶次 $m=2$ 时的 AR 模型参数 $a_{2,1}$， $a_{2,2}$ 以及 ρ_2^{fb}。

（5）重复上述过程，直到阶次 $m=p$，这样就求出了所有阶次的 AR 模型参数，对于 p 阶 AR 模型的输入白噪声方差 $\sigma_W^2 = \rho_p$。利用 Burg 递推法求 AR 模型参数的流程图如图 4.2 所示。

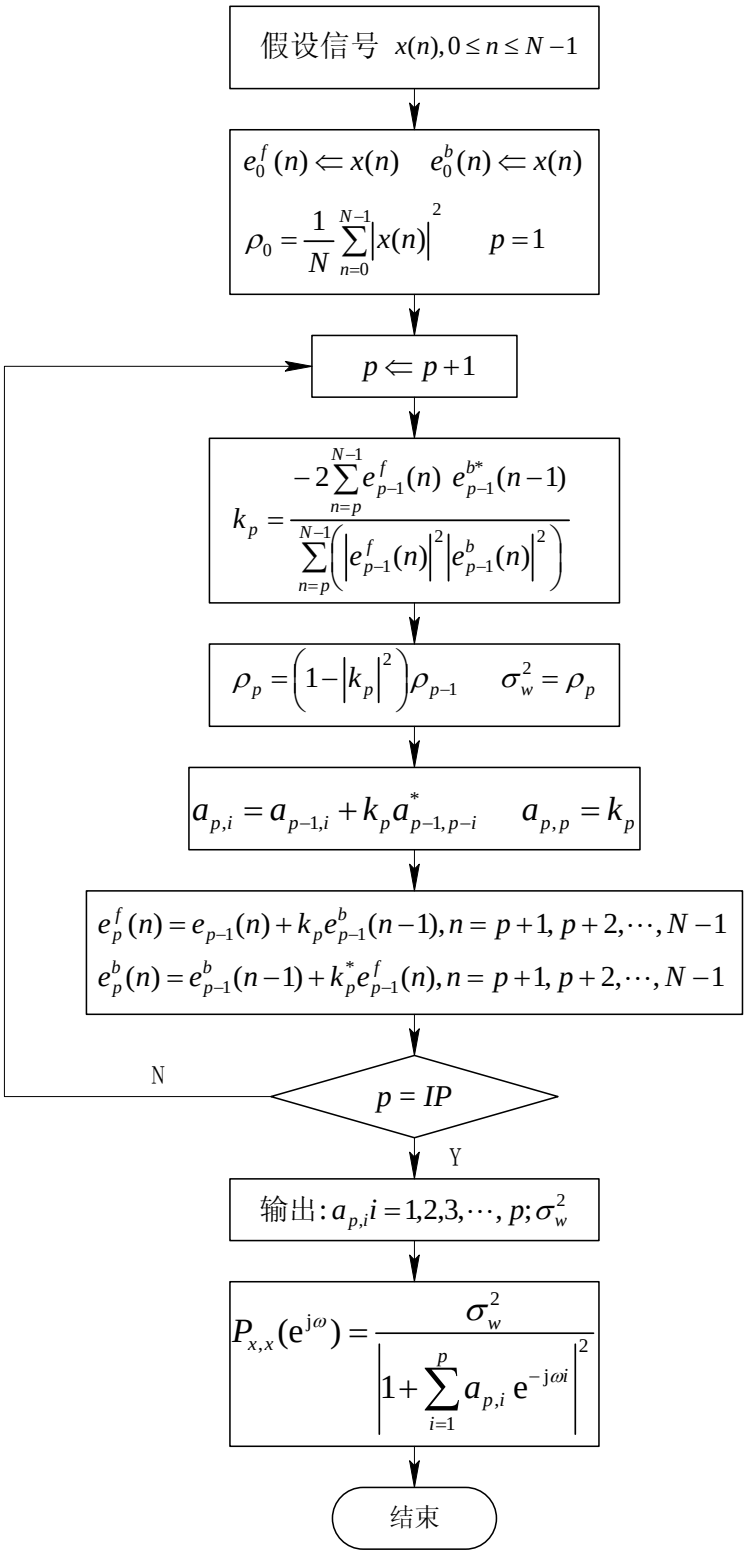

图 4.2 基于 Burg 递推法计算功率谱的流程图

经典功率谱估计的分辨率反比于有效信号的长度，但现代谱估计的分辨率可以不受此限制。这是因为对于给定的 N 点有限长序列 $x(n)$，虽然其估计出的自相关函数也是有限长的，但是现代谱估计的一些隐含着数据和自相关函数的外

推，使其可能的长度超过给定的长度，不象经典谱估计那样受窗函数的影响。因而现代谱的分别率比较高，而且现代谱线要平滑得多。

4.2.3 AR 模型功率谱算法仿真实验

利用 MATLAB 2011a 软件编程实现了 Lenvinson-Durbin 递推法与 Burg 递推法的快速算法，分别对酿酒酵母 ATP1a 的基因序列的功率谱及信噪比进行了计算，算法的源代码见附录一。算法的 FTT 长度设为 256，阶数设为 100，对酿酒酵母 ATP1a 的长度为 1638 的基因序列进行仿真，基于 Lenvinson-Durbin 递推算法的 AR 模型算法程序运行结果见图 4.3，基于 Burg 递推算法的程 AR 模型的程序运行结果见图 4.4。

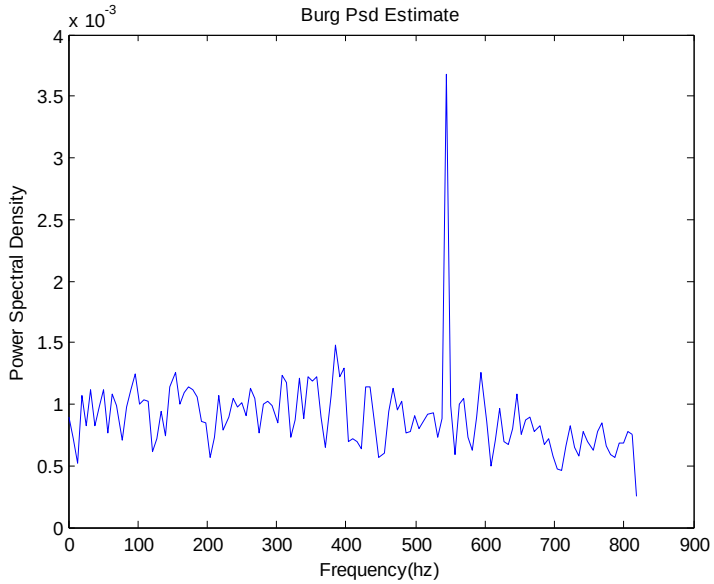

图 4.3 编号 *AB304259.1* 的酵母基因 *DNA* 序列基于 Lenvinson-Durbin 递推法计算的功率谱

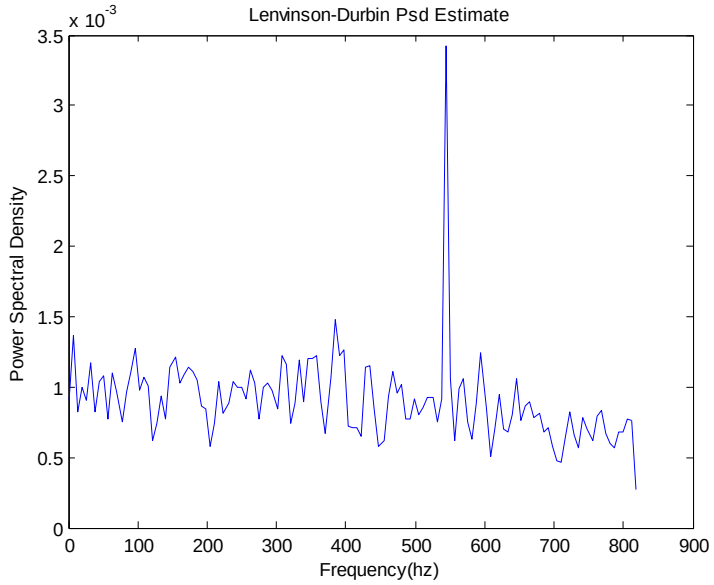

图 4.4 编号 *AB304259.1* 的酵母基因 DNA 序列基于 Burg 递推法计算的功率谱

从图 4.3 和图 4.4 可看出，利用本文给出的计算基因序列的功率谱及信噪比的快速算法对编号为 AB304259.1 的酵母基因 DNA 序列进行识别，发现在频率 f=1/3 HZ 处均有明显的峰值存在，表明了蛋白编码区序列（外显子）的 3-周期性存在。基于 Lenvinson-Durbin 递推算法与基于 Burg 递推算法计算出的信噪比分别为 3.7764 和 4.1。基于 Lenvinson-Durbin 递推算法与基于 Burg 递推算法消耗的时间分别为 0.011 和 0.038。

与直接利用 FFT 变换相比，由于直接基于 FFT 变换的周期图法对信号功率谱的估计是有偏的非一致的估计，信号的功率谱含有很多噪声。而 AR 模型法明显改善了功率谱估计的统计特性，功率谱曲线更加平滑且噪声低。由此说明，AR 模型法在功率谱的计算上显示出了优越性。

我们在 AR 模型基础上，给出了基于 AR 模型的固定长度滑动窗口功率谱的基因识别算法及移动序列功率谱的基因识别算法，源代码见附录二。我们取窗口长度为 99，下图为人和鼠类第 19 个和第 23 个样本的基因序列，使用固定长度滑动窗口功率谱的基因识别图和移动序列功率谱的基因识别结果，图中粗线表示外显子区域。

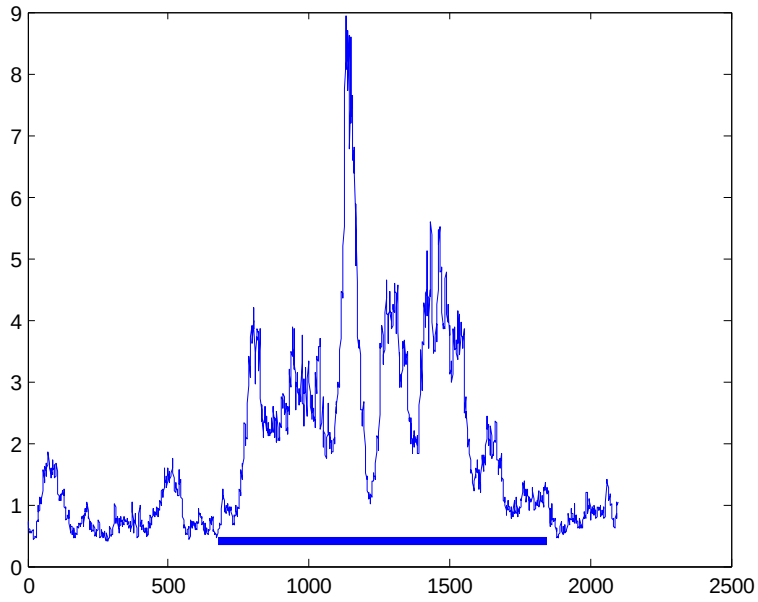

图 4.5 基于 AR 模型的固定长度滑动窗口功率谱人和鼠类第 19 个基因的识别结果

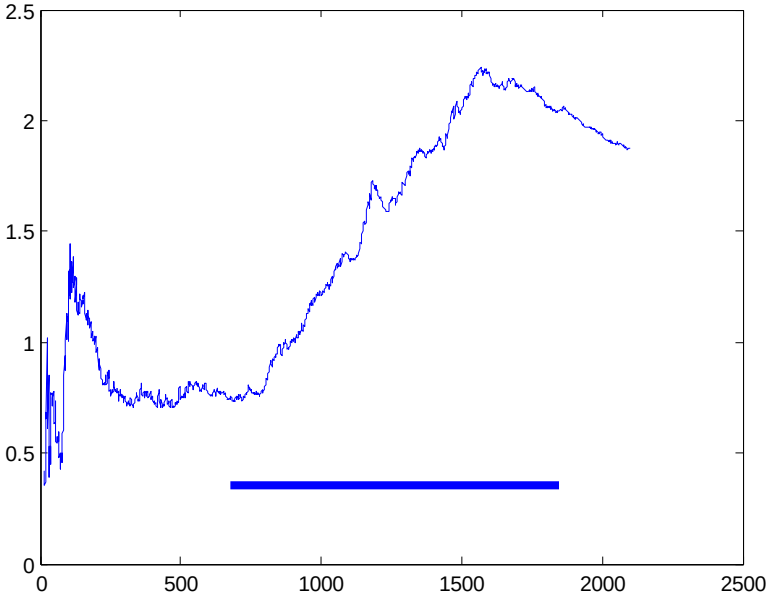

图 4.6 基于 AR 模型的移动序列功率谱人和鼠类第 19 个基因的识别结果

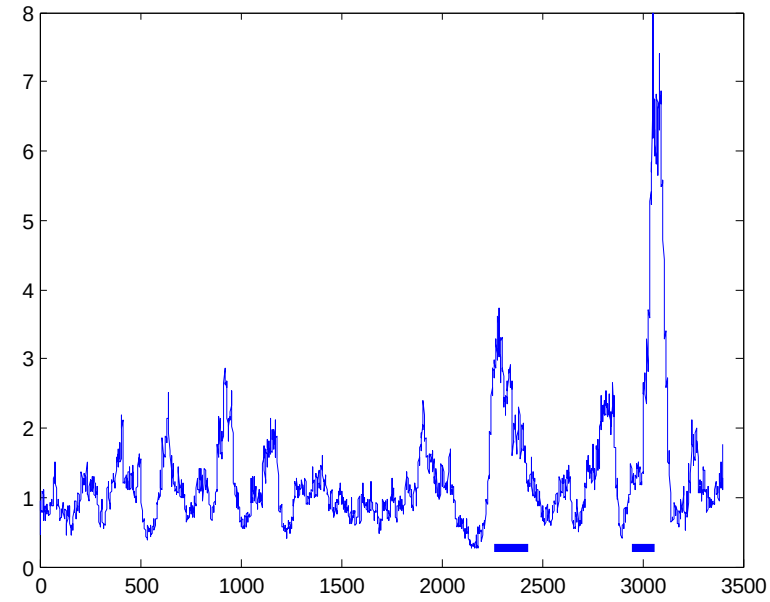

图 4.7 基于 AR 模型的固定长度滑动窗口功率谱人和鼠类第 23 个基因的识别结果

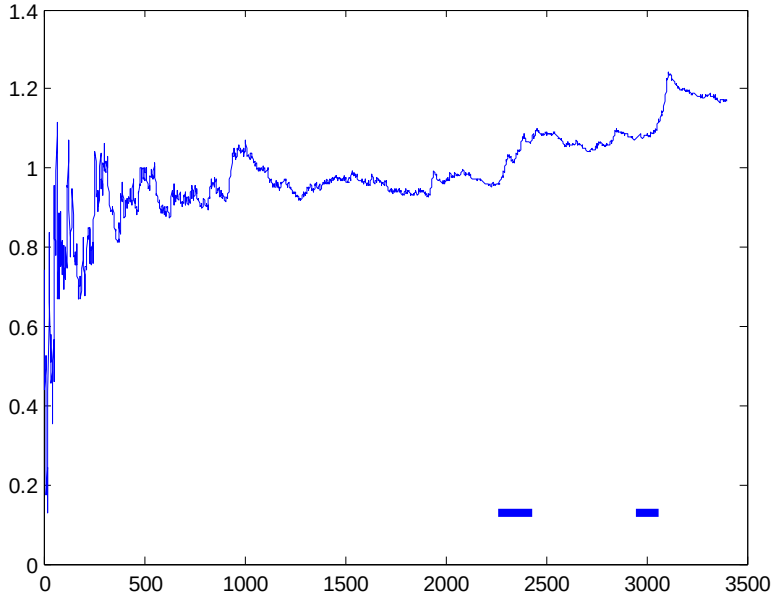

图 4.8 基于 AR 模型的移动序列功率谱人和鼠类第 23 个基因的识别结果

从上面 4 个图可看出，对于比较长的外显子序列，例如第 19 个基因样本，我们给出的基于 AR 模型的固定长度滑动窗口功率谱的基因识别算法及移动序列功率谱的基因识别算法均能很好进行识别。但对于很短的外显子序列，如第 23 个样本，固定长度滑动窗口功率谱的基因识别算法仍具有较高的识别效果，但移动序列功率谱的基因识别算法则出现了一些误判。

4.3 功率谱与信噪比快速计算公式的理论推导

设 DNA 序列 S 的四个指示序列 $\{u_b[n]\}$，$b \in I = \{A,C,G,T\}$，设 DNA 序列中核苷酸个数为 N，令 S 中核苷酸 A、C、G 和 T 在 DNA 序列中出现的次数分别为 N_A、N_C、N_G 和 N_T，则有如下定理成立。

定理 4.1 整个 DNA 序列 S 的功率谱序列 $\{P[k]\}$ 满足

$$P[k] = \sum_{b \in I} |U_b[k]|^2 = N \cdot N_b \tag{4.23}$$

则 DNA 序列的总功率谱 E 满足

$$E = \sum_{k=0}^{N-1} p[k] = N \cdot N_A + N \cdot N_C + N \cdot N_G + N \cdot N_T = N^2 \tag{4.24}$$

证明： 对指示序列 $\{u_b[n]\}$ 分别做离散 Fourier 变换，可得

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-j\frac{2\pi nk}{N}}, \quad k = 0,1,\cdots,N-1 \tag{4.25}$$

以此可得到四个长度均为 N 的复数序列 $\{U_b[k]\}$，$b \in I$。

对复数序列 $\{U_b[k]\}$ 分别作离散 Fourier 变换的逆变换，可得

$$u_b[n] = \frac{1}{N} \sum_{n=0}^{N-1} U_b[k] e^{j\frac{2\pi nk}{N}}, \quad n = 0,1,\cdots,N-1 \tag{4.26}$$

利用 Parseval 定理[3]，有

$$\sum_{n=0}^{N-1}\left|u_b[n]\right|^2=\frac{1}{N}\sum_{n=0}^{N-1}\left|U_b[k]\right|^2 \tag{4.27}$$

易知，上式的左边满足$\sum_{n=0}^{N-1}\left|u_b[n]\right|^2=N_b$，进而有

$$\left|U_b\right|^2=\sum_{n=0}^{N-1}\left|U_b[k]\right|^2=N\cdot N_b \tag{4.28}$$

因此，利用$N_A+N_C+N_G+N_T=N$，易推得 DNA 序列的总功率谱E满足

$$E=\sum_{b\in I}|U_b|^2=N\cdot N_A+N\cdot N_C+N\cdot N_G+N\cdot N_T=N^2 \tag{4.29}$$

证毕。

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[4][5]

$$P[\frac{N}{3}]=\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b) \tag{4.30}$$

根据定理 4.1 和式（4.30），易得如下推论。

推论 4.1 指示序列$\{u_b[n]\}$的信噪比R_b满足

$$R_b=\frac{x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b}{N_b} \tag{4.31}$$

证明：根据信噪比的定义，有

$$\begin{aligned}R_b&=\frac{P_b\left[\frac{N}{3}\right]}{\overline{E_b}}\\&=\frac{|U_b|^2}{E_b/N}\\&=\frac{x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b}{(N\cdot N_b)/N}\\&=\frac{x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b}{N_b}\end{aligned} \tag{4.32}$$

证毕。

根据定理 4.1 和推论 4.1，易得如下推论。

推论 4.2 DNA 序列的信噪比R满足

$$R=\frac{N_A}{N}\cdot R_A+\frac{N_C}{N}\cdot R_C+\frac{N_G}{N}\cdot R_G+\frac{N_T}{N}\cdot R_T \tag{4.33}$$

证明：根据信噪比的定义，有

$$
\begin{aligned}
R &= \frac{P[\frac{N}{3}]}{\bar{E}} \\
&= \frac{\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)}{E/N} \\
&= \frac{\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)}{N} \\
&= \frac{N_A}{N}\cdot R_A+\frac{N_C}{N}\cdot R_C+\frac{N_G}{N}\cdot R_G+\frac{N_T}{N}\cdot R_T
\end{aligned}
\tag{4.34}
$$

证毕。

鉴于式(4.33)在简化信噪比计算方面的重要意义以及其体现了信噪比与DNA 序列中核苷酸出现频次之间的关系，在此将式(4.33)称为 SNR-F 公式，并将其应用于后文中的基因识别模型的建立与求解中。

通过以上的理论推导过程，针对 Voss 映射，本文建立了功率谱、信噪比与DNA 序列中核苷酸出现的频次之间的关系。利用上述关系，计算功率谱与信噪比将不再需要离散 Fourier 变换等计算量较大的运算，只需要对 DNA 序列中核苷酸出现的频次进行统计，然后进行简单的数值运算即可，这将极大简化功率谱与信噪比的计算。文中以下针对功率谱与信噪比的计算方法如不加特殊说明均采用上述算法进行。

4.4 Z-curve 映射和 Voss 映射的关系

Z-curve方法[6]是上个世纪90年代中期，天津大学张春霆院士首先提出的一种表示DNA序列新方法。它从几何学的角度，阐述了基因识别的新方法。该方法从根本上区别于以往基因识别方法：动态规划[7]和隐马尔可夫模型方法[8]等。并在基因起始位点识别、Isochore结构识别上都有很好的应用。Z-curve方法已经成为了一个研究DNA序列的比较完整的、系统化的方法。由于Z-curve映射具有非常深刻的生物学意义，若仅仅探讨Z-curve映射的功率谱与信噪比和Voss映射下的功率谱与信噪比之间的关系，不足以全面刻画出两种映射之间的关系。

因此，本文将先重点分析Z-curve映射的功率谱与信噪比和Voss映射下的功率谱与信噪比之间的关系，然后从理论基础、生物学意义和特征三个方面将Z-curve映射和Voss映射进行对比，从而刻画出两种映射之间更深层次、更全面的关系。

4.4.1 两种映射下的功率谱与信噪比之间的关系

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 $b_n=\sum_{i=0}^{n}u_b[i]$。

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

$$\begin{pmatrix}\Delta x[n]\\ \Delta y[n]\\ \Delta z[n]\end{pmatrix}=\begin{pmatrix}1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix}\begin{pmatrix}u_A[n]\\ u_C[n]\\ u_G[n]\\ u_T[n]\end{pmatrix} \tag{4.35}$$

则有

$$\begin{pmatrix}x[n]\\ y[n]\\ z[n]\end{pmatrix}=2\begin{pmatrix}1 & 0 & 1 & 0\\ 1 & 1 & 0 & 0\\ 1 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}u_A[n]\\ u_C[n]\\ u_G[n]\\ u_T[n]\end{pmatrix}-\begin{pmatrix}1\\ 1\\ 1\end{pmatrix} \tag{4.36}$$

由此，可以看出 Z-curve 映射是 Voss 映射的仿射变换，Z-curve 映射与 Voss 映射相比，它具有更强的生物学背景意义，且在方便运算上也有较大的优势。

类似于 Voss 映射，定义 Z-curve 映射的总功率谱

$$P_Z[k]=\left|\Delta X[k]\right|^2+\left|\Delta Y[k]\right|^2+\left|\Delta Z[k]\right|^2 \tag{4.37}$$

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 $\Delta z[n]$ 的离散傅变换。

同样，我们也可以定义 Z-curve 映射的信噪比为

$$R_Z=\frac{P_Z[\frac{N}{3}]}{\overline{E}}=\frac{\left|\Delta X[\frac{N}{3}]\right|^2+\left|\Delta Y[\frac{N}{3}]\right|^2+\left|\Delta Z[\frac{N}{3}]\right|^2}{\overline{E}} \tag{4.38}$$

其中， $\overline{E}=\dfrac{\sum_{k=0}^{N-1}P_Z[k]}{N}$ 是 Z-curve 映射的平均功率谱。

Ahmad Rushdi 和 Jamal Tuqan 在文献[9]中，已经给出了 $P_z[k]$ 和 $P[k]$ 之间的数值关系，描述为如下引理。

引理 4.1[9] 设 $\{P_z[k]\}$ 和 $\{P[k]\}$ 分别为 Z-curve 映射与 Voss 映射下 *DNA* 序列 S 的功率谱序列，则

$$P_z[k]=P[k],\quad k=0,1,\cdots,N-1 \tag{4.39}$$

根据引理 4.1，有如下定理。

定理 4.2 E_z 和 E 分别为 Z-curve 映射与 Voss 映射下 *DNA* 序列 S 的总功率谱，则

$$E_z=4E \tag{4.40}$$

证明： $E_z=\sum_{k=0}^{N-1}P_z[k]=\sum_{k=0}^{N-1}4P[k]=4\sum_{k=0}^{N-1}P[k]=4E$ 。证毕。

定理 4.3 R_z 和 R 分别为 Z-curve 映射与 Voss 映射下 *DNA* 序列 S 的信噪比，则

$$R_z=R \tag{4.41}$$

证明： $R_z=\dfrac{P_z\left[\dfrac{N}{3}\right]}{\overline{E_z}}=\dfrac{P_z\left[\dfrac{N}{3}\right]}{E_z/N}=\dfrac{4P\left[\dfrac{N}{3}\right]}{4E/N}=\dfrac{P\left[\dfrac{N}{3}\right]}{E/N}=\dfrac{P\left[\dfrac{N}{3}\right]}{\overline{E}}=R$。证毕。

通过以上的推导过程，从数值上，本文给出了 Z-curve 映射的功率谱与信噪比和 Voss 映射下的功率谱与信噪比之间的关系。

4.4.2 理论基础及其生物学意义对比

Z-curve 方法研究起始于对组成 DNA 序列的 A，G，C，T 这四种碱基的对称性的观察，按双环或单环是否存在氨基或酮基、碱基对形成氢键的数目或强弱，将这四种碱基进行划分，并用正六面体来表示它们的对称性，从而导出了 Z 变换，然后利用 Z 变换将 DNA 序列转换为三维空间中的点，得到 Z 曲线。

从以上 Z-curve 映射的理论基础描述中可以看出，Z-curve 映射和 Voss 映射两者的理论基础具有较为明显的区别，描述如表 4.1 所示。

表 4.1 Z-curve 映射和 Voss 映射的理论基础及其生物学意义对比

映射	理论基础	DNA 序列表达形式	生物学意义	评价
Voss	信号处理与分析	用A、C、G和T四种字母代表四个碱基，由不同数量的这四个字母按不同顺序排列成一维链就构成了DNA序列，进而形成DNA双螺旋结构。	离散Fourier变换所固有的“栅栏效应”在这里体现十分明显，在某些非整数的周期点上，并不具有生物学意义，而一些比较重要的频率点却可能被漏掉。	在发表、存储及提供计算机进行快速分析等方面有其不容争议的优点，但却又有它不可克服的严重缺陷。其一是忽视了人脑在模式识别方面的强大能力；其二是它的“解像力”不可调节。
Z-curve	信号处理与分析、几何学	给定的DNA 序列唯一决定了$x[n]$、$y[n]$和$z[n]$的分布；三种分布也唯一决定了DNA序列。$x[n]$、$y[n]$和$z[n]$三种分布是相互独立的，表现在以下事实上：任何一种分布不能由其他两种分布的线形叠加表示出来。	$x[n]$表示嘌呤碱基和嘧啶碱基沿DNA序列的分布；当嘌呤碱基总数过半时，$x[n]>0$，否则$x[n]<0$。$y[n]$表示氨基和酮基沿DNA序列的分布；当氨基碱基占优时，$y[n]>0$，否则$y[n]<0$。$z[n]$表示氢键碱基和弱氢键碱基沿DNA序列的分布；当弱氢键碱基占优时，$z[n]>0$，否则$z[n]<0$。从方法学的角度讲，这是DNA序列的一种几何学研究途径。	利用几何学分析和研究DNA序列的直观工具，是表达碱基排列顺序的一种崭新形式，不仅具有解像力连续可调的优点，还可充分调动人脑来参与模式识别，从而克服了传统的表达DNA序列形式的缺陷。

4.4.3 特征对比

Z-curve 是三维空间的曲线，它是显示和分析 DNA 序列的基本工具，能以直观的形式体现出 DNA 序列的整体特性和局部细节，而且不论基因组的序列有多长，总可以较方便通过计算机显示出其 Z-curve，并可进行任意的旋转，以便从不同角度观察 Z-curve 的三维折叠结构。与 Voss 映射相比，Z-curve 映射有以下特征：

1) 等步性：两个相邻节点的坐标之差，即 Δx 和 Δy，或者等于l，或者等于-l，没有其他可能性。
2) 回路性：即在任意一个节点处，Z-curve在往前走若干步之后，有可能返回此点，形成所谓回路。
3) 渐进性：当 $N \geq 1$时，Z-curve要么趋于0，要么是一个被慢变化函数调制的线性增长的函数，Z-curve的终点只与序列的碱基组成有关而与其排列顺序无关。
4) 对称性：Z-curve是基于几何学中正四面体的性质得到的，理论研究表明DNA序列中的映象点(D格点)具有对应关系，由此可得到Z-curve束具有对称性，而且Z-curve均在正四面体的内切球内问。
5) 手性：Z-curve有左旋和右旋之分，分别对应着左旋和右旋DNA序列；而传统的字母DNA序列的表达形式，则无手性问题。

4.5 实数映射下功率谱和信噪比的快速计算公式

从以上对 Voss 映射与 Z-curve 映射的分析可以看出，Voss 映射是 DNA 序列的一种 4-D 表示，Z-curve 映射是 DNA 序列的一种 3-D 表示。事实上，除了这些表示方法，还存在 2-D 和 1-D 的数值映射，具体可参见文献[10]。根据文献[10]的分析，DNA 序列的任意一种数值映射都可以看作是 Voss 映射的仿射变换。因此，对实数映射的研究就可以转化对 Voss 映射的仿射变换的研究。

下面我们考虑一般情况，在 Voss 映射下，定义 DNA 序列 S 的四个指示序列 $\{u_b[n]\}$ 的任意一个 3-D 实数仿射变换为

$$\begin{pmatrix} x[n] \\ y[n] \\ z[n] \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix}$$

令 R_S 和 R 分别为上述实数映射和 Voss 映射下的信噪比。为方便描述，作如下定义。

$$H = \left(r_{ij}\right)_{3\times 4} = \begin{pmatrix} r_1^T \\ r_2^T \\ r_3^T \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \end{pmatrix} \text{且} U = \begin{pmatrix} u_A(n) \\ u_C(n) \\ u_G(n) \\ u_T(n) \end{pmatrix} = \begin{pmatrix} u_1(n) \\ u_2(n) \\ u_3(n) \\ u_4(n) \end{pmatrix}$$

令核苷酸符号 $b \in I = \{A,T,G,C\}$ 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，具体表示如下。

$$X = \begin{pmatrix} x_A & x_C & x_G & x_T \\ y_A & y_C & y_G & y_T \\ z_A & z_C & z_G & z_T \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{pmatrix} = \left(X_1, X_2, X_3, X_4\right)$$

假设 DNA 序列 S 的长度为 3 的倍数，有如下定理成立。

定理 4.4 当矩阵 H 的列向量满足如下两个条件时：

1) 对任意的 $i\ (1 \leq i \leq 4)$，$\| \alpha_i \| \equiv c_1$，其中 c_1 是个常数；

2) 对任意的 $i, j\ (1 \leq i, j \leq 4, i \neq j)$，$< \alpha_i, \beta_i > \equiv c_2$，其中 $<\cdot,\cdot>$ 表示两个向量

的内积运算，c_2是个常数。

则

$$R_S = \frac{c_1 - c_2}{c_1} \cdot R = \frac{4}{3}R$$

证明：由于$x[n] = r_1^T \cdot U$，可得$\{x[n]\}$进过离散 *Fourier* 变换在$N/3$处的功率谱为

$$\begin{aligned}
\left| X\left[\frac{N}{3}\right] \right|^2 &= \left| \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi n \cdot \frac{N}{3}}{N}} \right|^2 \\
&= \left| \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{3}n} \right|^2 \\
&= \left| \sum_{j=1}^{4} r_{1j} \left(\sum_{n=0}^{N-1} u_j[n] \cdot e^{-j\frac{2\pi n}{3}} \right) \right|^2 \\
&= \left| \sum_{j=1}^{4} r_{1j} \left(x_j + y_j \cdot e^{-j\frac{2\pi}{3}} + z_j \cdot e^{j\frac{2\pi}{3}} \right) \right|^2 \\
&= \left| \left(\sum_{j=1}^{4} r_{1j} x_j \right) + \left(\sum_{j=1}^{4} r_{1j} y_j \right) e^{-j\frac{2\pi}{3}} + \left(\sum_{j=1}^{4} r_{1j} z_j \right) e^{j\frac{2\pi}{3}} \right|^2 \\
&= r_1^T X^T M X r_1
\end{aligned} \tag{4.42}$$

其中，$M = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}$。

同理，可得

$$\left| Y\left[\frac{N}{3}\right] \right|^2 = r_2^T X^T M X r_2 \tag{4.43}$$

$$\left| Z\left[\frac{N}{3}\right] \right|^2 = r_2^T X^T M X r_2 \tag{4.44}$$

仿射变换的总功率谱为

$$P_S\left[\frac{N}{3}\right] = \left| X\left[\frac{N}{3}\right] \right|^2 + \left| Y\left[\frac{N}{3}\right] \right|^2 + \left| Z\left[\frac{N}{3}\right] \right|^2 = \sum_{i=1}^{3} r_i^T X^T M X r_i \tag{4.45}$$

令$\beta = (1,1,1,1)^T$且

$$X^T MX = \begin{pmatrix} X_1^T \\ X_2^T \\ X_3^T \\ X_4^T \end{pmatrix} M\left(X_1, X_2, X_3, X_4\right) = \begin{pmatrix} X_1^T MX_1 & \cdots & X_1^T MX_4 \\ \vdots & \ddots & \vdots \\ X_4^T MX_1 & \cdots & X_4^T MX_4 \end{pmatrix} \triangleq F \qquad (4.46)$$

由于

$$X_1 + X_2 + X_3 + X_4 = \begin{pmatrix} x_A + x_C + x_G + x_T \\ y_A + y_C + y_G + y_T \\ z_A + z_C + z_G + z_T \end{pmatrix} = \begin{pmatrix} \frac{N}{3} & \frac{N}{3} & \frac{N}{3} \end{pmatrix}^T \qquad (4.47)$$

$$\begin{aligned} MXr &= M\left(X_1, X_2, X_3, X_4\right)(1,1,1,1)^T = M\left(X_1 + X_2 + X_3 + X_4\right) \\ &= \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} \frac{N}{3} & \frac{N}{3} & \frac{N}{3} \end{pmatrix}^T = 0 \end{aligned} \qquad (4.48)$$

易得

$$\begin{aligned} r^T Fr = r^T X^T MXr &= (1,1,1,1)F(1,1,1,1)^T = \sum_{i,j=1}^{4} X_i^T MX_j \\ &= \sum_{i=1}^{4} X_i^T MX_i + 2\sum_{i<j} X_i^T MX_j \\ &= 0 \end{aligned} \qquad (4.49)$$

结合（4.45）式，有

$$P_S\left[\frac{N}{3}\right] = r_1 Fr_1^T + r_2 Fr_2^T + r_3 Fr_3^T = c_1\left(\sum_{i=1}^{4} X_i MX_i^T\right) + 2c_2\left(\sum_{i<j} X_i MX_j^T\right) \qquad (4.50)$$

事实上，在上式中 $X_i^T MX_i$ 的系数可以利用如下方式计算得到。

$$r_{1i}^2 + r_{2i}^2 + r_{3i}^2 = \| \alpha_i \|^2 = c_1, \quad i = 1,2,3,4$$

同理，有

$$r_{1i}r_{1j} + r_{2i}r_{2j} + r_{3i}r_{3j} = < \alpha_i, \alpha_j > = c_2, \quad i, j = 1,2,3,4, i \neq j$$

利用（4.49）和（4.50），有

$$\begin{aligned} P_S\left[\frac{N}{3}\right] &= (c_1 - c_2)\left(\sum_{i=1}^{4} X_i^T MX_i\right) + \left[c_2\left(\sum_{i=1}^{4} X_i^T MX_i\right) + 2c_2\sum_{i<j} X_i^T MX_j\right] \\ &= (c_1 - c_2)P\left[\frac{N}{3}\right] + 0 \\ &= (c_1 - c_2)P\left[\frac{N}{3}\right] \end{aligned} \qquad (4.51)$$

其对应的总功率谱为

$$E_S=\sum_{k=0}^{N-1}\left(|X(k)|^2+|Y(k)|^2+|Z(k)|^2\right)$$

根据 Parseval 定理，可得

$$\sum_{n=0}^{N-1}|x(n)|^2=\frac{1}{N}\sum_{k=0}^{N-1}\left(|X(k)|^2\right)$$

由此可得

$$\begin{aligned}\sum_{k=0}^{N-1}|X(k)|^2&=N\cdot\sum_{n=0}^{N-1}|x[n]|^2=N\cdot\sum_{n=0}^{N-1}|r_1^T\cdot(u_A[n],u_C[n],u_G[n],u_T[n])^T|^2\\&=N\cdot(r_{11}^2N_A+r_{12}^2N_C+r_{13}^2N_G+r_{14}^2N_T)\end{aligned}\quad(4.52)$$

其中 S 中核苷酸 A、C、G 和 T 在 DNA 序列中出现的次数分别为 N_A、N_C、N_G 和 N_T。

同时，有

$$\sum_{k=0}^{N-1}|Y(k)|^2=N\cdot(r_{21}^2N_A+r_{22}^2N_C+r_{23}^2N_G+r_{24}^2N_T)$$

$$\sum_{k=0}^{N-1}|Z(k)|^2=N\cdot(r_{31}^2N_A+r_{32}^2N_C+r_{33}^2N_G+r_{34}^2N_T)$$

利用上述三式，有

$$E_S=N\left(c_1N_A+c_1N_C+c_1N_G+c_1N_T\right)=c_1N^2$$

因此，有

$$R_S=\frac{P_S\left[\frac{N}{3}\right]}{E_S/N}=\frac{(c_1-c_2)P\left[\frac{N}{3}\right]}{c_1N}=\frac{(c_1-c_2)}{c_1}\cdot\frac{P\left[\frac{N}{3}\right]}{N}=\frac{(c_1-c_2)}{c_1}\cdot R$$

同时，c_1 与 c_2 之间存在着线性关系：

$$\left\|\sum_{i=1}^{4}\alpha_i\right\|^2=\sum_{i=1}^{4}<\alpha_i,\alpha_i>+2\sum_{i<j}<\alpha_i,\alpha_j>=4c_1+12c_2=0$$

则 $c_1=-3c_2$。

故，可得

$$R_S=\frac{4}{3}R$$

证毕。

通过对上述定理的进一步研究，以下给出定理的更一般形式。

定理 4.5 假设 DNA 序列 S 的长度为 3 的倍数，设矩阵 $D=\left(d_{ij}\right)_{4\times4}$ 满足：对任意 $i,j\left(1\le i,j\le4\right)$，$d_{ij}=<\alpha_i,a_j>$。特别，当 $i=j$ 时，有 $<\alpha_i,a_j>=\|\alpha_i\|^2=d_{ii}$，则有

$$R_S=\frac{P_S\left[\frac{N}{3}\right]}{E_S/N}=\frac{\sum_{i,j=1}^{4}d_{ij}(X_i^TMX_j)}{d_{11}N_A+d_{22}N_C+d_{33}N_G+d_{44}N_T} \tag{4.53}$$

证明：根据等式（4.45），有

$$P_S\left[\frac{N}{3}\right]=\left|X\left[\frac{N}{3}\right]\right|^2+\left|Y\left[\frac{N}{3}\right]\right|^2+\left|Z\left[\frac{N}{3}\right]\right|^2=\sum_{i=1}^{3}r_i^TX^TMXr_i$$

因此，可得

$$P_S\left[\frac{N}{3}\right]=r_1Fr_1^T+r_2Fr_2^T+r_3Fr_3^T=\sum_{i=1,j=1}^{4}d_{ij}(X_i^TMX_j) \tag{4.54}$$

事实上，有

$$r_{1i}r_{1j}+r_{2i}r_{2j}+r_{3i}r_{3j}=<\alpha_i,\alpha_j>=d_{ij},\quad i,j=1,2,3,4$$

对比等式（4.52）和（4.53），易得

$$E_S=\sum_{k=0}^{N-1}\left(|X(k)|^2+|Y(k)|^2+|Z(k)|^2\right)=N\left(d_{11}N_A+d_{22}N_C+d_{33}N_G+d_{44}N_T\right)$$

故，可得

$$R_S=\frac{P_S\left[\frac{N}{3}\right]}{E_S/N}=\frac{\sum_{i,j=1}^{4}d_{ij}(X_i^TMX_j)}{d_{11}N_A+d_{22}N_C+d_{33}N_G+d_{44}N_T}$$

证毕。

上述定理给出了 Voss 映射的仿射变换下功率谱与信噪比的快速计算公式，对于更加一般的实数变换，上述定理也是适用的。

在此，以 $A\to 0,C\to 1,G\to 2,T\to 3$ 为例，给出其信噪比的快速计算公式。在此实数映射下，根据定理，信噪比的计算公式为

$$R_S=\frac{P_S\left[\frac{N}{3}\right]}{E_S/N}=\frac{(X_2+2X_3+3X_4)^TM(X_2+2X_3+3X_4)}{N_C+4N_G+9N_T}$$

4.6 小结

针对问题一，本节完成了以下工作：

针对 Voss 映射，给出了计算基因序列功率谱或信噪比的快速 Fourier 变换和 AR 模型，仿真实验结果表明，计算效率有所提升。经过理论推导，建立了功率谱、信噪比与 DNA 序列中核苷酸出现的频次之间的关系，计算功率谱与信噪比将不再需要离散 Fourier 变换等计算量较大的运算，只需要对 DNA 序列中核苷酸出现的频次进行统计，然后进行简单的数值运算即可，有效提升了功率谱与信噪比的计算效率。

推导出了 Z-curve 映射的功率谱与信噪比和 Voss 映射下的功率谱与信噪比之间的数值关系，并从理论基础、生物学意义和特征三个方面对 Z-curve 映射

和 Voss 映射进行了对比分析，刻画出了两种映射之间更深层次、更全面的关系。

经过理论推导，给出了一般的实数映射下功率谱、信噪比的快速计算公式，将其功率谱、信噪比的计算简化为核苷酸出现频次的统计和简单数值运算，极大简化了实数映射下功率谱与信噪比的计算。

五、问题二模型的建立与求解

问题二：

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：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，对按照功率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。

问题分析：

功率谱的分析方法，存在影响基因预测精度的一个重要因素——阈值。阈值的设定是用来区分一段 DNA 序列为蛋白编码区还是非编码区的重要指标。一个固定的阈值并不适用于所有生物，经验阈值缺乏通用性，而且事实表明，为各类不同生物选用同一阈值缺乏合理性，因为不同生物有着不同的基因结构特性，对某些生物而言，某个阈值会过高，使得预测结果虽然有较高的正确率，但探测率势必下降，从而使预测精度降低。反之，对有些生物而言，某个阈值会过低，使得预测结果有较高的探测率但正确率很低，同样预测精度将降低。除了不同生物其基因结构特性的因素之外，传统经验阈值不具有通用性的其它因素在于：功率谱分析方法中不同的窗口大小、不同的功率谱计算方法等都会产生功率谱预测曲线的幅值变化，由此看来，经验阈值会带来明显的预测误差。因此，解决阈值的选取问题在功率谱分析方法中是十分关键的。

在本节中，我们结合重采样技术，提出了最佳阈值推断算法。本文提出的方法基于这样一个事实：对近 2000 多种生物，它们仅有部分基因已被标注。我们希望利用这些已标注的基因信息，结合重采样算法，为一个特定的生物推测其最佳阈值。实验仿真结果表明，推断的最佳阈值能显著提高基于功率谱分析方法的基因预测精度，该最佳阈值可用来预测该生物目前尚未标注确认的其它基因。

5.1 固定长度窗口滑动功率谱分析方法

如题目所述，在真核生物的基因组中，基因是断裂的，外显子在序列中的长度很短且被大量内含子隔开。因此，要从真核生物的基因序列中识别蛋白编码区并定位出起始点是一项重要且具有挑战性的工作。现在，已经有多种生物的测序工作已经完成，这些核酸序列数据隐含有大量的生命信息，值得进行系

统挖掘。

目前用于基因预测的方法有很多，但它们有各自的使用范围和适用对象，并会产生不同的结果。主要存在以下问题：只能检测蛋白质编码基因，即外显子部分，预测非编码区的方法比较少；假阳性，即在非编码区预测出编码区；只能检测典型基因；假阴性，即将编码区预测为非编码区；预测结果保守，很难发现新的基因；融合化，即将 DNA 序列上距离过近的基因被预测成一个大的基因；过界预测，即预测超出实际的边界；片段化，即对于一个完整的基因，由于其内含子过大，会被预测成多个基因[11]。

在诸多从头计算的预测方法中，基于功率谱分析的基因预测方法是利用蛋白编码区的 3-周期性的一种计算方法，该方法易于实现并且所需的待分析生物的先验信息少，因此面对世界范围内急剧增长的基因组序列，功率谱分析方法可以快速有效的识别 DNA 序列中可能的基因。学者们认为编码区的 3-周期性特征形成的的部分原因，是由于以三个核苷酸为密码子的位置分布不平衡造成的，构建蛋白质的二十种氨基酸的出现概率不同。而在一个内含子序列，核苷酸均匀分布在这三个密码子位置。该分布的不平衡的原因是蛋白质对氨基酸组成有偏好，同种氨基酸的同义密码子使用频率不同，以及每种氨基酸的密码子兼并度不同等原因也是形成 3-周期性的潜在因素。

目前常用的一个基因预测法为固定长度窗口滑动功率谱分析方法，因此，本文选择以固定长度窗口滑动功率谱分析方法为基础。固定长度窗口滑动功率谱分析方法的预测过程简要叙述如下：

（1）将 DNA 符号序列映设成为四个二进制的数值序列，并利用先验的知识选取合适的滑动窗宽 M。序列的映射规则为 Voss 映射。

（2）对第 k 个滑动窗，计算该窗下序列对应的局部信噪比 $p(n;\frac{M}{3})$，其中 M 为窗口宽度。

（3）设置局部信噪比函数的阈值 R。一般情况下，那些局部信噪比 $p(n;\frac{M}{3})> R$的序列被标识为蛋白编码区域，而局部信噪比 $p(n;\frac{M}{3})\le R$的序列被标识为非编码区域。

然而，这类方法需要设定阈值来区分蛋白编码区和非编码区，且预测精度严重受到阈值大小的影响。传统经验阈值能区分绝大多数生物的蛋白编码区和非编码区，但并不能达到最好的预测精度。由于不同真核生物的基因结构之间存在差异，所以，对于某一特定的真核生物，仅仅依赖于有限的先验生物知识很难为其确定一个合适的基因预测阈值。显然，为所有生物选取统一的预测阈值更难以取得理想的预测结果。如何利用某一特定的生物根据其自身基因结构特征，寻找一个最佳阈值，本节将对基因识别的阈值确定方法进行研究。

5.2 最佳阈值确定算法的设计

在本节中，我们主要运用了重采样技术来确定基因预测算法的最佳阈值，下面先简要介绍重采样算法原理。

美国 stanford 大学统计系教授 Efron 在总结、归纳前人研究成果的基础上提出一种新的统计方法—Bootstrap 方法，重采样算法是一种基于计算机的统计推断方法，无需知道待估计参数的先验信息，通过少量的样本则可估计参数的分布，即只依赖于给定的观测信息，不需要其它假设的一种新统计推断方法[12]。

目前被广泛应用于雷达信号处理、生物医学工程、图像处理、模式识别及控制领域等，如基于 Bootstrap 方法的基因表达谱数据互信息估计研究，在医学统计分析中也发挥了很好的作用。Bootstrap 的逻辑基础是某个统计量的所有准确度估计指标都来源于该统计量的抽样分布。如果这个统计量是用来源于某一个总体的含量为 n 的随机样本估计而得到的，那么它的抽样分布就可以显示该统计量的各种值的相对频数。抽样分布是由总体分布和估计统计量所用的计算公式所决定的。目前对重采样算法的理论研究和应用表明，该方法明显优于传统的参数估计方法，传统的参数估计方法都是基于对原始样本的大量采样。因此，重采样算法方法被应用到许多现实工程领域问题中，如生物医学工程、雷达信号处理以及人工神经网络等。

重采样算法的基本思想是：令 $X=\{x_1,x_2,\cdots,x_n\}$ 是从分布未知的总体 F 中随机抽取的 n 个样本，因此，$x_i(i=1,2,\cdots n)$ 是独立同分布的随机变量。θ 表示来自未知总体 F 的一个未知参数，我们要解决的问题是寻找参数 $\hat{\theta}$ 的分布特性来作为对未知参数 θ 的一个估计。X 称之为原始样本，从样本 X 中计算要估计的参数 $\hat{\theta}$，则要进行重采样过程，即从原始样本 X 中进行 n 次有放回的独立抽样，产生一个新的样本 $X^*=\{x_1^*,x_2^*,\cdots,x_n^*\}$，称为重采样样本。通常重采样次数达到 1000 以上时便可获得相当好的估计。

重采样算法能为待估计的参数给出高质量的准确度估计，如置信区间、标准差。本文利用重采样算法来为估计的最佳阈值求得一个置信区间，即进行参数的区间估计（interval estimation）。区间估计是指按一定的可信度（概率）估计未知的总体参数可能所在的范围的估计方法，该范围也称置信区间。区间估计是在考虑抽样误差存在的情况下，经随机抽取样本后进行的估计，是一种较为准确可靠的方法。一般情况下，统计学上通常采用 95%的置信区间，表示总体参数在某一范围的概率为 95%，或将置信区间选为 99%，可根据不同情况的条件要求选用不同的概率。反映置信区间的两个重要要素有：一是准确度，即所选的可信度的大小，也是一个区间包含总体均数的概率的大小，越接近 100%越好，说明估计的参数在该区间内取值的概率就越大。二是精密度，反映在区间的长度上，该长度越小越好，能给估计的结果带来更现实的意义。显然，在样本含量确定的情况下，准确度和精密度是矛盾的，如果将可信度提的很高，那么会把区间变得很长，因此不能认为 99%置信区间比 95%置信区间好，准确度和精密度二者要兼顾，才能获得更好的统计结果。一般情况下，95%置信区间比 99%的置信区间更为常用，因为在可信度一定的情况下，若能增加样本的含量，那么区间长度可以明显减少，从而提高精密度。

5.3 最佳阈值确定算法的基本步骤

对于某一特定的生物，基于重采样算法的最佳阈值推断算法的基本思路为：先从该生物已标注的核苷酸序列中截取若干序列，从各段序列上实验获得最优预测阈值，利用这些实验观测的阈值作为重采样算法的原始样本集。然后从原始样本集经重采样获得最优阈值的置信区间。最后由该置信区间计算获得该生物的最佳阈值。置信区间本身代表了最优阈值的分布特征，该特征是由生物的基因结构特征所决定的。本文提出的最佳阈值推断方法的详细步骤描述如下：

（1）从一个待预测其基因的生物体中的线粒体或染色体上，随机截取 n

段已标注基因的 DNA 序列，本文给出的算法要求 $n \geq 10$。

（2）对于每一段已标注基因的 DNA 序列，运用基于 AR 模型法的的功率谱分析方法预测其基因，通过与已标注的基因进行比对，获得预测该段 DNA 序列基因的一个最优阈值 $R_i(i=1,2,\cdots,n)$。由 n 段 DNA 序列获得的最优阈值组成原始样本 $\chi=\{R_1,R_2,\cdots,R_n\}$，且该样本含有待估计的最佳阈值的分布特性，最优阈值表示在该阈值下相应序列有着最高的预测精度。

（3）对于原始样本 $\chi=\{R_1,R_2,\cdots,R_n\}$，设置重采样次数 C≥1000。

（4）令 $j=1,2,\cdots,C$，从 χ 中做 n 次有放回的随机抽取，获得一个无序采样集，称其为一次重采样 $\chi_j^*=\{R_1^*,R_2^*,\cdots,R_n^*\}$。计算 χ_j^* 的均值，则得到一个阈值的估计参数 $\hat{\theta}_j^*$。

（5）如果 $j \leq C$，则令 $j=j+1$，转到（4）。否则，转到（6）。

（6）将 C 个阈值的估计值从低到高排列，组成待估计参数的分布特征 $\hat{\theta}^*=\{\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_C^*\}$。

（7）为待估计的参数 $\hat{\theta}^*$ 计算 $100\times(1-\alpha)\%$ 的置信区间 $(\hat{\theta}_p,\hat{\theta}_q)$，其中 α 为置信水平，且 $p=\left\lfloor \frac{\alpha C}{2} \right\rfloor$，$q=C-p+1$。

（8）通过计算以上置信区间的均值，为选定的生物获得一个预测其蛋白编码区的最佳阈值。

由于对绝大部分生物而言，已有部分基因已被标注，那么提出的方法中所需的先验信息，即重采样的原始阈值样本集是可以获得的。该方法首先利用一个生物体中少量已标注的基因先验信息来构建原始的基因预测阈值样本，再运用重采样算法推断阈值的置信区间，最后结合该生物体的基因结构分布特征，由置信区间内求得一个最佳阈值。迄今为止，已经有大量生物体的部分基因被标注，如人类、鼠类等生物体，这些实验数据为本方法提供了可靠的先验信息。

5.4 重采样阈值确定算法的仿真实验

我们利用 MATLAB 2011a 软件进行编程，对带有编码外显子信息的 100 个人和鼠类及 200 个哺乳动物类的基因序列样本数据进行了仿真实验，源代码见附录三。

首先我们需要确定 n 段 DNA 序列中每一段的一个最优阈值。最优阈值指的是在该阈值下，相应序列有着最高的预测精度，故我们采用每段 DNA 序列对应内含子的固定长度滑动窗口功率谱的最大值作为该段 DNA 序列获得的最优阈值，在这种情况下的阈值，能使外显子尽可能被准确标识出来。我们对 100 个人和鼠类的数据进行仿真实验，由于篇幅所限，我们仅列出 1 到第 10 个的最优阈值，结果如表 5.1 所示。

表 5.1 人和鼠类第 1 到 10 个样本的最优阈值

样本序号	最优阈值	该段预测精度
1	3.879055	91%
2	2.859477	91%
3	2.392866	72%
4	4.158346	82%

5	4.587965	87%
6	2.657533	96%
7	1.723214	93%
8	2.805396	75%
9	4.359883	90%
10	3.16985	86%

我们设置信水平为 95%，利用 MATLAB 2011a 软件编程得到 95%的置信区间为（1.6167，1.9379）。

最后，执行阈值推断方法中的第 8 步，计算置信区间下的均值，获得人和鼠类生物基因预测的最佳阈值 P=1.7773。为了验证提出的方法推断的最佳阈值的有效性，我们从给定的生物序列上选取另外一段 DNA 序列，利用固定长度窗口滑动功率谱分析方法，并设定预测的阈值为 P=1.7773，窗口长度为 99，对人和鼠类的基因序列进行了检测。图 5.1 和 5.2 为人和鼠类的数据第 19 和 23 个基因数据外显子的预测图，图中的粗线为设定的阈值，阈值以上的部分即为算法预测出的外显子区域。

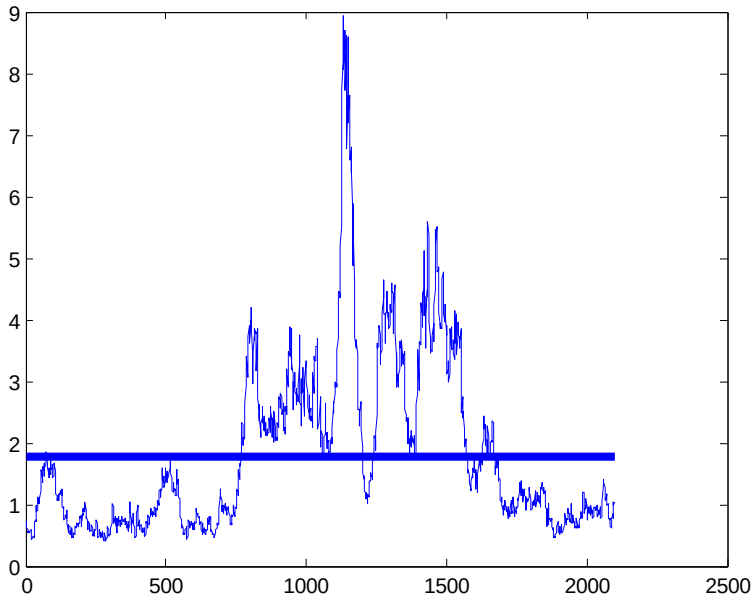

图 5.1 窗口长度为 99 时人和鼠类的数据第 19 个数据外显子的固定窗口法预测图

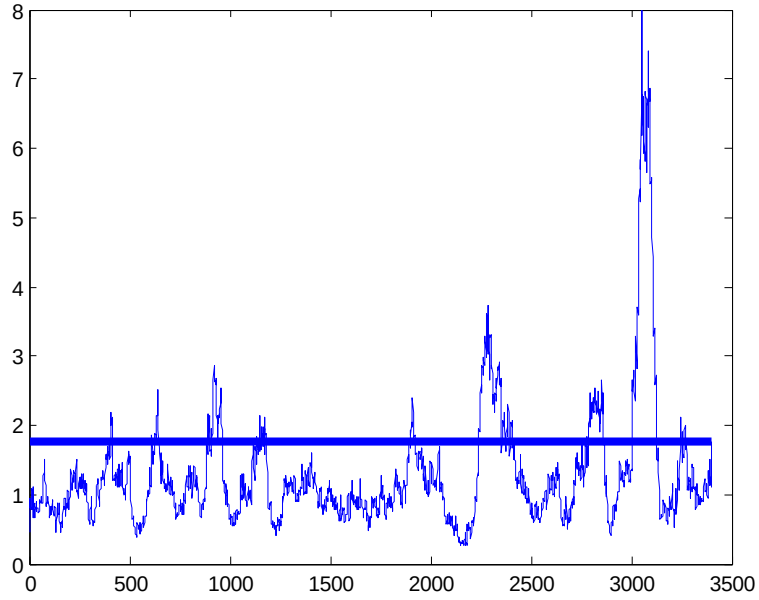

图 5.2 窗口长度为 99 时人和鼠类的数据第 23 个数据外显子的固定窗口法预测图

从上图可知，第 19 个数据预测的外显子区间为（664，1700），而第 19 个基因数据的实际外显子区间（675，1847）。第 23 个数据预测的外显子区间为（610，650）、（1214，1290）、（701，810）、（2256，2452）和（3014，3017），而 23 个数据实际的外显子区间（2264，2432）和（2950，3062），其中的区间（610，650）、（1214，1290）和（701，810）属于误判。因此我们考虑加大窗口长度的大小，调整窗口大小为 255，预测阈值为仍 P=1.7773，再利用固定长度窗口滑动功率谱分析方法对第 23 个基因序列进行预测，结果如图 5.3 所示。从图 5.3 可知，此时，不再出现窗口为 99 时产生的误判，这说明我们在设置算法参数时，还需要考虑窗口的大小。

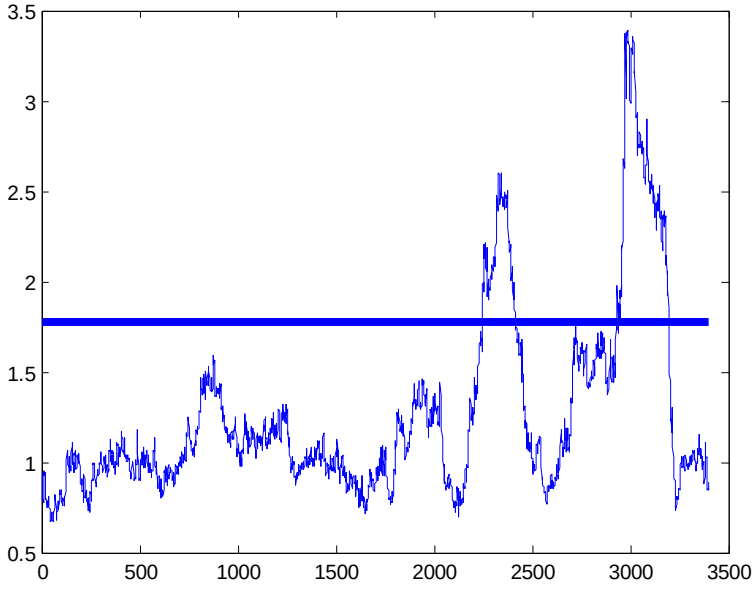

图 5.3 窗口长度为 255 时人和鼠类的数据第 23 个数据外显子的固定窗口法预测图

我们对 200 个哺乳动物类的数据进行实验，我们仅列出 1 到第 10 个的最优阈值，结果如表 5.2 所示。

表 5.2 哺乳动物类第 1 到 10 个样本的最优阈值

样本序号	最优阈值	该段预测精度
1	2.644552	87%
2	4.8764	91%
3	2.840304	82%
4	3.505751	79%
5	5.149509	92%
6	3.003013	85%
7	1.714498	88%
8	2.743955	85%
9	3.435222	96%
10	2.241927	91%

我们设置信水平为 95%，利用 MATLAB 2011a 软件编程得到 95%的置信区间为（2.1765，2.2164）。

同样执行阈值推断方法中的第 8 步，计算该置信区间下的均值，获得给定生物基因预测的最佳阈值 P=2.18。为了验证提出的方法推断的最佳阈值的有效性，我们从给定的生物序列上选取另外一段 DNA 序列，利用固定长度窗口滑动功率谱分析方法，并设定预测的阈值为 P=2.18，窗口长度为 255。图 5.4 为哺乳动物类的数据第 1 个基因数据外显子的预测图，图 5.4 为哺乳动物类的数据第 1 个数据外显子的预测图，图中的粗线为阈值，阈值以上的部分为预测的外显子区域。

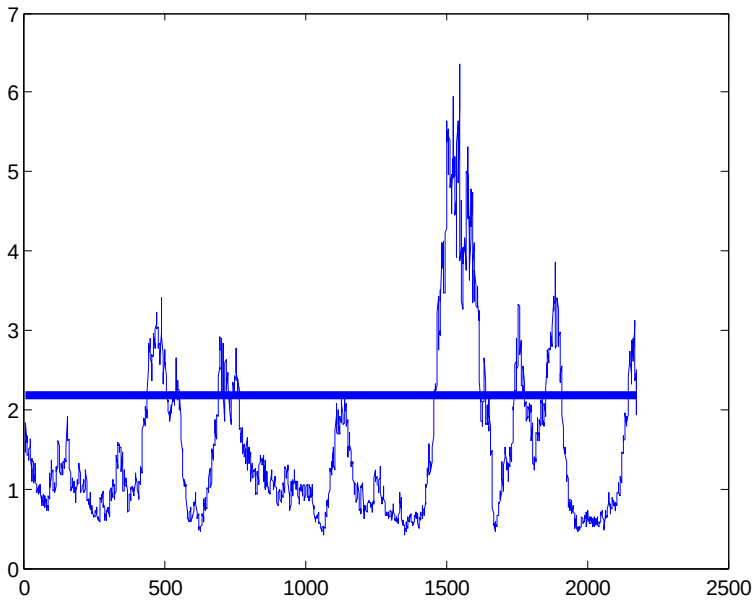

图 5.4 窗口长度为 99 时哺乳动物类的数据第 1 个数据外显子的固定窗口法预测图

从上图可知，第 1 个数据预测的外显子区间为（452，536），（665，705），（1441，1929），而 1 个数据实际的外显子区间（493，535），（652，805），（1341，1938）。这说明本文提出的重采样阈值确定算法预测的阈值，对固定长度窗口滑动功率谱分析方法具有较高的精确性。

5.5 小结

针对问题二，本文完成了以下工作：

结合重采样技术，提出了最佳阈值确定算法，能为每一个特定种类的生物推测其最佳阈值。其具体思路为：先从该生物已标注的核苷酸序列中截取若干序列，从各段序列上实验获得最优预测阈值；然后，从原始样本集经重抽样获得最优阈值的置信区间；最后，由该置信区间计算获得该生物的最佳阈值。仿真实验结果表明，在合理确定窗口大小的基础上，利用该最佳阈值能显著提高基于功率谱分析方法的基因预测精度，同时还可用来预测该生物目前尚未标注确认的其它基因。

六、问题三模型的建立与求解

问题三：

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”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。

问题分析：

全世界有超过25种生物的测序工作已经完成，随着各种生物序列信息的指数增长，大大增加了用于基因预测的参考序列集，势必会提高估计的参数的准确性。虽然目前很多基因预测方法有很高的准确性，一些功能强大的基因识别软件对蛋白编码区的识别率很高，但仍然需要开发新一代的基因识别程序，并结合各种方法提高预测的精度。在今后的基因预测中，通过多种预测方法的合理组合，相互取长补短，将显著增加预测精度，即通过不同方法的结合来提高预测精度是一大研究趋势。学者们相继发表了多种提高基因预测精度的概念和方法，大多是将多种预测方法联合使用，也有对某种现有预测方法的改进等。因为基于单一模型的预测方法存在模型限制条件和适用范围等问题，很难获得理想的结果，多种基因预测方法综合在一起可以改善预测结果。因此，本节我们利用两种预测方法相结合，提高现有基因预测方法的预测精度。

6.1 基于 AR 模型重采样的基因预测方法

在第四节中，我们详细介绍了基于AR模型重采样的基因预测方法，在这里不再赘述，我们利用该模型仅对附件中给出的6个未被注释的DNA序列（gene6）的编码区域进行预测。

由于在该算法中，需要先收集已知DNA序列的一些特征对阈值进行估计，但附件中给出的6个未被注释的DNA序列并没有说明属于哪类基因，故我们只能简化阈值的估计。在第五节中，我们分别对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进行了阈值的估计。进一步，我们另外下载了50个基因序列数据，并对这50个基因序列的阈值进行了估计。最后对这三类基因序列的阈值取其平均值作为6个未被注释的DNA序列的新阈值。我们利用MALAB 2011a软件进行编程，得到新的阈值为1.965。利用该阈值，选择窗口长度为255，对6个未被注释的DNA序列进行了预测，结果如下。

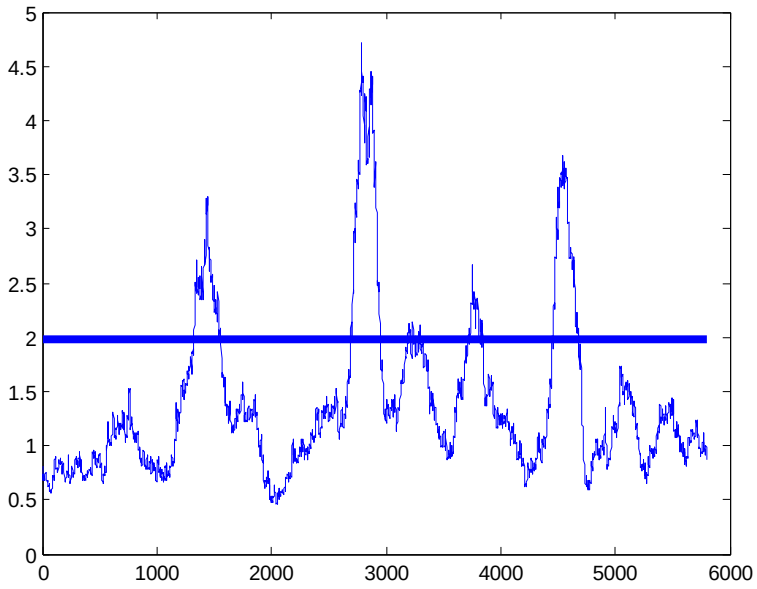

图 6.1 窗口长度为 255 时第 1 个未被注释的 DNA 序列预测图

从上图可看出第1个未被注释的DNA序列的外显子区域为（1290，1531），（2714，2981），（3179，3213），（3701，3857）和（4564，4715）。

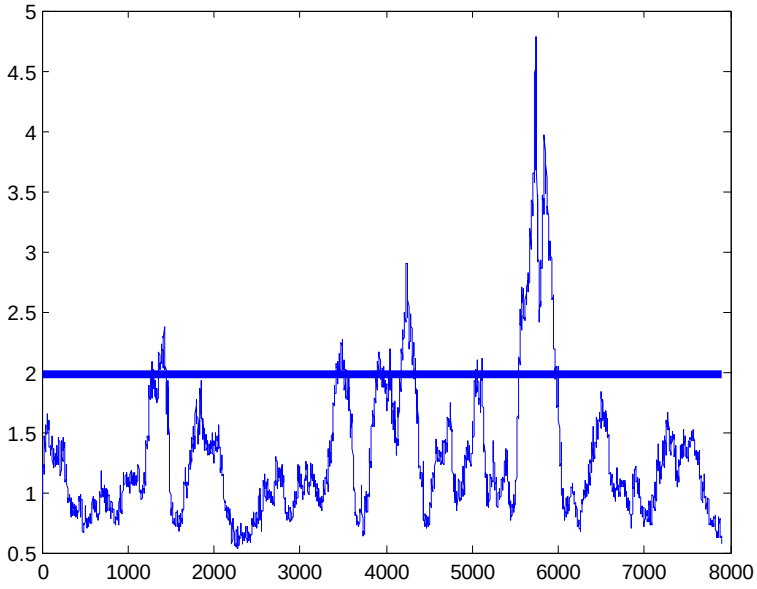

图 6.2 窗口长度为 255 时第 2 个未被注释的 DNA 序列预测图

从上图可看出第2个未被注释的DNA序列的外显子区域为（1180，1438），（3375，3512），（3894，4278），（5078，5112）和（5512，6023）。

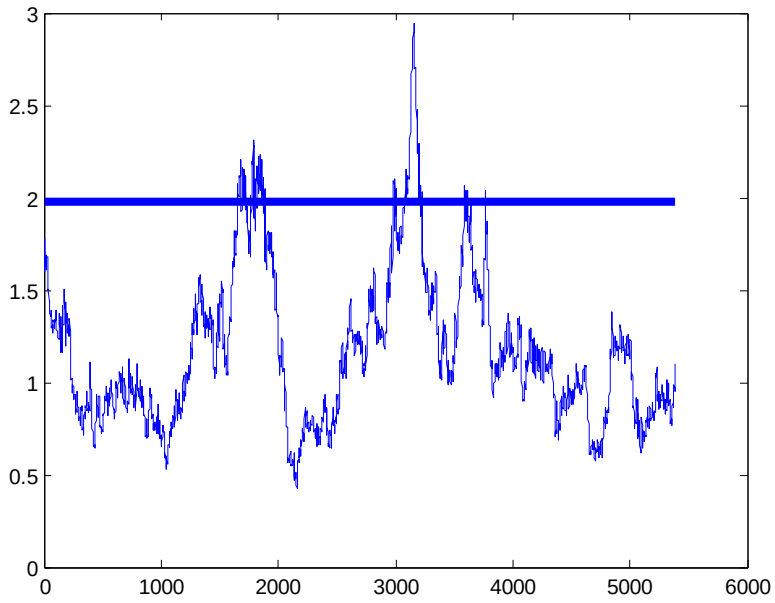

图 6.3 窗口长度为 255 时第 3 个未被注释的 DNA 序列预测图

从上图可看出第3个未被注释的DNA序列的外显子区域为（1728，1932），（3005，3114）和（3672，3873）。

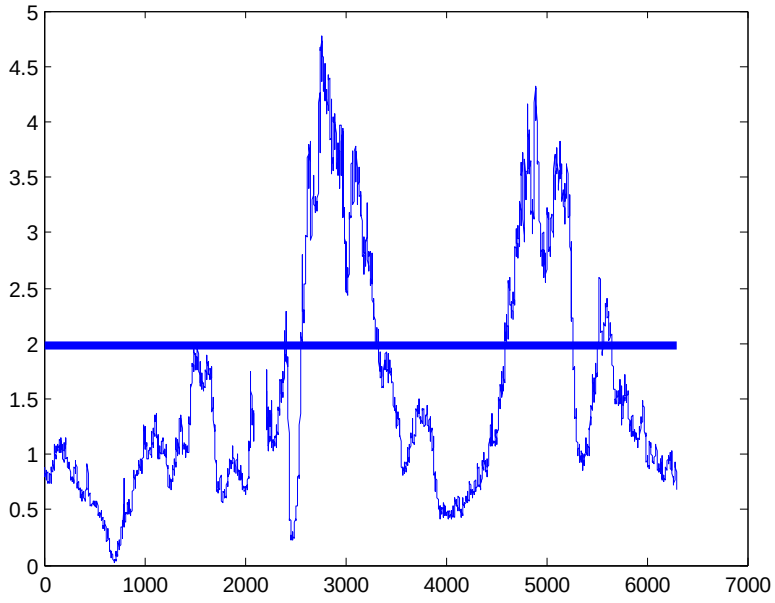

图 6.4 窗口长度为 255 时第 4 个未被注释的 DNA 序列预测图

从上图可看出第4个未被注释的DNA序列的外显子区域为（2615，3331），（4017，5213）和（5578，5670）。

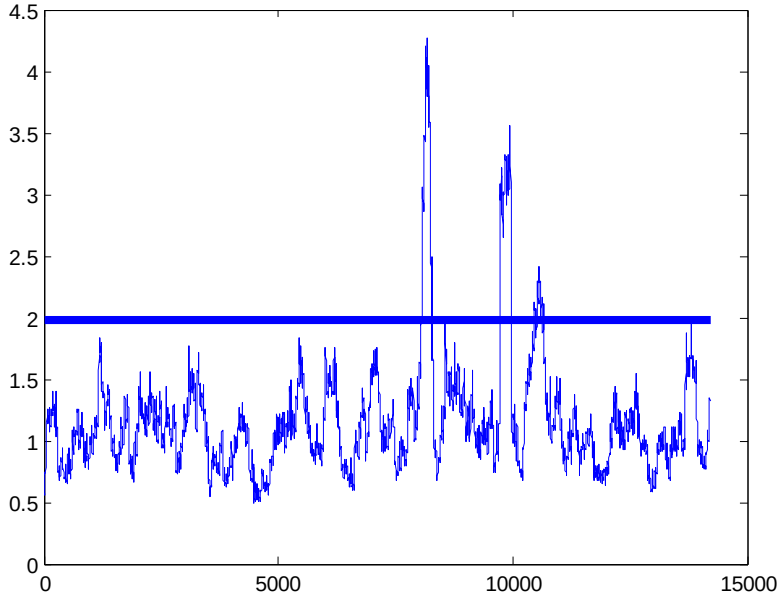

图 6.5 窗口长度为 255 时第 5 个未被注释的 DNA 序列预测图

从上图可看出第5个未被注释的DNA序列的外显子区域为（6259，6547），（9923，10034）和（10243，10357）。

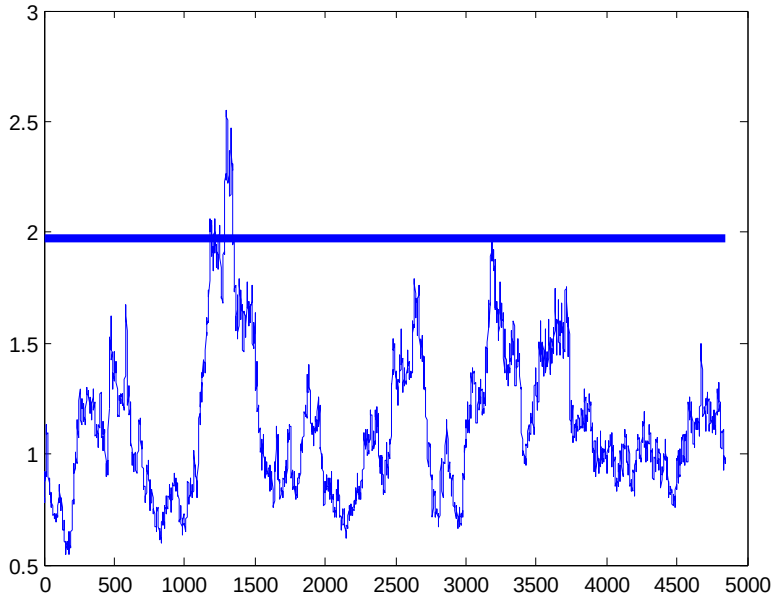

图 6.6 窗口长度为 255 时第 6 个未被注释的 DNA 序列预测图

从上图可看出第5个未被注释的DNA序列的外显子区域为（1287，1424）。

6.2 基于 SNR-F 的基因识别模型

本节利用基因预测的数字滤波器与第四节中提出的功率谱与信噪比快速计

算公式相结合，提出了一种基于SNR-F的基因识别模型。该模型克服了现有Fourier方法对序列长度的要求，而且易于实现。

6.2.1 数字滤波器设计

从信号处理的角度看，滑动窗方法可以看作对基因序列进行数字滤波。假设滤波器的单位脉冲响应：

$$h(n)=\begin{cases} e^{jw_0 n} & 0\le n\le N-1 \\ 0 & otherwise \end{cases} \tag{6.1}$$

这是一个带通滤波器，中心频率 $\omega_0=2\pi/3$，最小阻带衰减约13 dB。如果把基因序列映射的数字序列当作滤波器的输入信号 $x(n)$，通过数字滤波器后，就得到了输出信号 $y(n)$。由于基因编码区具有频谱3-周期性，因此输入碱基的标量序列中处于非编码区的部分会被滤波器衰减掉，保留编码区位置的信号，故输出信号在编码区位置会出现波峰。

通过以上分析可知，设计的滤波器是以 $\omega=2\pi/3$ 为中心频率的窄带滤波器，而且带宽尽量小，这种性能的滤波器可以通过一个二阶全通滤波器实现。本节利用频率抽样法[13]设计带通数字滤波器，该方法的特点是可以根据需要设计滤波器的幅频特性。频率抽样法是从频域出发，把给定的理想频率响应 X 等间隔抽样，即

$$H_d(k)=H_d(e^{jw})\big|_{\omega=\frac{2\pi}{N}k} \quad k=0,1,\cdots N-1 \tag{6.2}$$

然后，以 $H_d(k)$ 作为实际数字滤波器的频率特性的抽样值 $H(k)$，可以用频域的这N个抽样值 $H(k)$ 来唯一确定滤波器的单位冲击响应 $h(n)$。根据基因序列外显子区域的频谱3-周期性性质，设计一个在归一化频率 $\omega=2\pi/3$ 处有窄带选通特性的数字滤波器，即要设计的滤波器的理想频率响应 $H(e^{jw})$ 在 $\omega=2\pi/3$ 为中心的很窄范围内幅值为1，其它频率处幅值都为0。对这个理想频率响应等间距抽取N个采样点，第一个取样点在 $\omega=0$ 处。由频率抽样理论的内插公式知道，利用这N个频域抽样值 $H(k)$ 可以求得数字滤波器的系统函数 $H(z)$ 及频率响应 $H(e^{jw})$。此方法设计的滤波器只允许具有频谱3-周期特性的信号通过，并抑制其它频率的信号，因此该滤波器具有较好的频率选择性。

6.2.2 改进的 SNR-F 算法

对于较长的基因编码区，比如300 bp或者更长的序列，利用传统的Fourier变换就能够较好探测到频谱3-周期性，但对较短的基因编码区，比如长度100 bp左右或更短时，利用传统的Fourier分析很难探测到频谱3-周期性。由此，文献[14]中提出了延长打乱Fourier方法，该算法能够有效放大3-周期性信号，对于较短基因序列的预测精度有所提高，同时抑制背景噪声。利用延长打乱Fourier方法计算出基因序列功率谱在 $N/3$ 处的幅值 $P_x(N/3)$、$P_y(N/3)$、$P_z(N/3)$ 作为算法中的统计特征量。

延长打乱Fourier方法中采用延长和打乱的目的是消除噪声并且放大频谱3-周期性信号。但这种方法可以改进，一方面至少随机打乱序列10000次以上实现起来比较消耗时间，另一方面该方法要求处理的基因序列长度必须是3的整数倍，

这些缺点降低了预测算法的效率，并且限制了所处理序列的长度范围。在信号处理领域，使用数字滤波器来消除信号中的噪声，保留有用信号是方便可行的。结合延长打乱Fourier方法的思想，利用SNR-F公式，本文提出改进算法如下：

（1）延长序列。对于一段较短的DNA序列$S(i), i \in (1,\mathrm{M})$，M是任意整数，重复序列$S(i)$得到延长后的序列$X(n)$，长度大于1024 bp。

$$X(n) = \{\underbrace{S\ S\ S \cdots S}_{L}\}, n \in (1, L), L > 1024$$

取$X(n)$前面长度是1024 bp的序列作为待处理序列$S(n)$。

（2）序列滤波。通过Z-curve映射将DNA序列$S(n)$转换为数字序列$S_x(n), S_y(n)$和$S_z(n)$。采用上节中所设计的数字滤波器，滤除背景噪声，用滤波器对数字序列$S_x(n), S_y(n)$和$S_z(n)$进行滤波，获得滤波后的序列$x(n), y(n)$和$z(n)$。

（3）功率谱及信噪比计算。利用公式（4.23）计算序列$x(n), y(n)$和$z(n)$的功率谱$P_x(N/3)$、$P_y(N/3)$、$P_z(N/3)$，利用SNR-F公式得出DNA序列的信噪比。

改进的方法处理的基因序列长度不需要是3的整数倍，这样扩大了识别算法的适用范围；另一方面，对序列滤波处理易于实现，省略了随机打乱序列次序10000次以上的操作。除此之外，改进的方法能够更好提取基因编码区的频谱3-周期性信号，在基因识别的准确性方面也有一定的提高。

6.2.3 基因识别方法实验分析

以上算法在一定程度上克服了传统Fourier方法对于基因识别存在的问题，对于较短基因序列的预测精度有所提高，本节提出的基于改进SNR-F算法的基因识别方法在此基础上又有所提高。但这几种方法都是对基因序列的分类，即给定一条未知序列，通过算法判定它属于编码序列还是非编码序列，而基因预测中给定的DNA序列有时很长，既包括编码序列又包括非编码序列，由此，本节给出一种加窗的方法，可以对一条完整的DNA序列进行预测，区分出序列中的编码区部分和非编码区部分。

算法的思路是：预先设定好窗口长度L，用窗从头开始截取DNA序列，对于截取的DNA序列使用改进SNR-F算法进行识别，确定其属于编码序列还是非编码序列，然后沿着DNA序列滑动这个窗，由此可以判断出整个序列中哪部分是编码区，哪部分是非编码区。

算法的评价采用的指标是(1-伪正率)和(1-伪负率)，探测率(sensitivity，*Sn*)和准确率(specificity，*Sp*)。真阳性(*TP*)是正确识别的外显子，假阳性(*FP*)是错误识别的内含子，真阴性(*TN*)是正确识别的内含子，假阴性(*FN*)是错误识别的外显子。伪正率(false positive rate，*FPR*)是非编码区预测的错误率，伪负率(false negative rate，*FNR*)是编码区预测的错误率，定义分别如下：

$$FNR = FN / (TP + FN) \tag{6.3}$$

$$FPR = FP/(TN+FP) \tag{6.4}$$

令预测精度(*Ap*)为：

$$Ap = [(1-FPR)+(1-FNR)]/2 \tag{6.5}$$

探测率(*Sn*)和准确率(*Sp*)的定义分别如下：

$$Sn = TP/(TP+FN) \tag{6.6}$$

$$Sp = TP/(TP+FP) \tag{6.7}$$

令探测率(*Sn*)和准确率(*Sp*)的平均值为预测精度*Ac*：

$$Ac = (Sn+Sp)/2 \tag{6.8}$$

此外，还可应用相关系数 *CC* 和近似相关 *AC* 衡量总体的预测正确率。其中，

$$CC = \frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \tag{6.9}$$

近似相关 *AC* 定义为

$$AC = (ACP-0.5)\times 2 \tag{6.10}$$

其中，

$$ACP = \frac{1}{4}\left[\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right] \tag{6.11}$$

从NCBI下载了一些已标注的DNA序列，使用该方法对这些DNA序列进行预测，其中的识别算法分别使用了传统Fourier方法和基于改进SNR-F算法的方法，获得实验结果如表6.1所示。评价的指标选用探测率(*Sn*)、准确率(*Sp*)和预测精度(*Ac*)。

表 6.1 基因识别方法对比

基因名称	外显子数	方法	*Sn*	*Sp*	*Ac*
AMU12024	6	传统 Fourier	67%	75%	71%
		改进 SNR-F	83%	83%	83%
ACU08131	6	传统 Fourier	67%	80%	74%
		改进 SNR-F	83%	83%	83%
GGCALB	17	传统 Fourier	65%	79%	72%
		改进 SNR-F	71%	86%	79%
GGCRYDI	16	传统 Fourier	63%	77%	70%
		改进 SNR-F	75%	86%	81%

BTHOR	3	传统 Fourier	67%	67%	67%
		改进 SNR-F	100%	75%	88%

由表6.1可以看出，该方法具有可行性，能够从基因序列中探测出编码区和非编码区，而且使用改进SNR-F算法预测的精度要高于传统Fourier方法的预测精度，平均精度高10%左右，再次说明了改进SNR-F算法具有一定优越性。

6.2.4 对 genes6 中基因序列的预测

利用上述方法，对题目附件中给出的6个未被注释的DNA序列（genes6）的编码区域的预测。与6.1小节相同，选取阈值为1.965，选择窗口长度为255，对6个未被注释的DNA序列进行了预测，结果如图6.7所示。

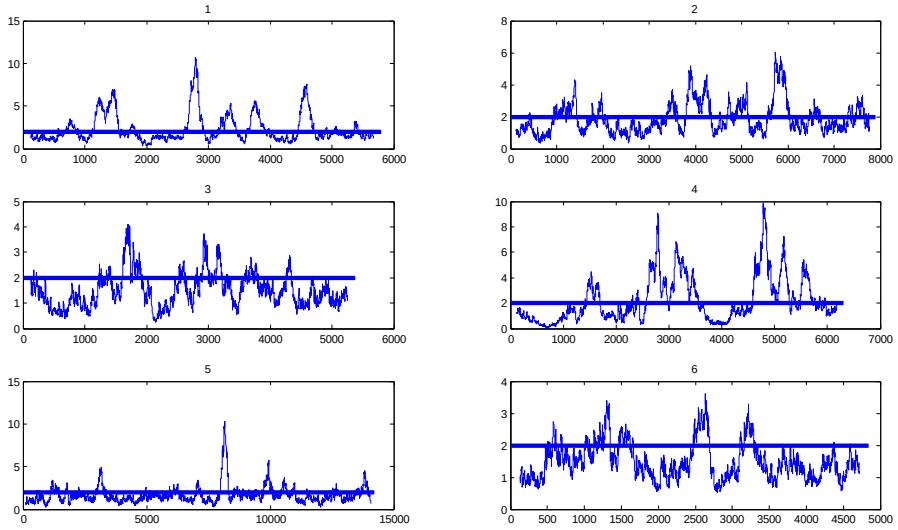

图 6.7 窗口长度为 255 时 genes6 中未被注释的 DNA 序列预测图

根据图6.7所示，在genes6中：

第1个未被注释的DNA序列的外显子区域应为（720，912），（1113，1455），（2646，2903），（3167，3383），（3743，3937），（4453，4735）和（5320，5429）；

第2个未被注释的DNA序列的外显子区域应为（987，1463），（1694，2021），（3395，4296），（4663，5151），（5503，6008）和（7370，7685）；

第3个未被注释的DNA序列的外显子区域应为（1238，1946），（2490，3398），（3679，3826）和（4230，4357）；

第4个未被注释的DNA序列的外显子区域应为（1439，1705），（2691，3703），（4708，5237）和（5409，5684）；

第5个未被注释的DNA序列的外显子区域应为（1225，1447），（2578，2694），（7008，7453），（7608，8309）、（9624，10082）和（13884，14891）；

第6个未被注释的DNA序列的外显子区域应为（498，702），（1002，1627），（2485，2741）和（3125，3308）。

6.3 小结

在对问题三的解决过程中，我们采用了两种不同的方法，预测的genes6中未

被注释的DNA序列的外显子区域也不相同。但是，从预测的结果不难看出，基于SNR-F的基因识别方法比基于AR模型重采样的基因预测方法预测出了更多的外显子区域。其中，前一种方法的预测能够基本覆盖未被注释的DNA序列的外显子区域，而后一种方法预测出的外显子区域则几乎可以确定是存在的。因此，两种预测方法重合的区域以及前一种方法独立预测出的部分区域应当是实际genes6中未被注释的DNA序列的外显子区域。两种预测方法相结合有助于提高基因预测的精度，同时能够使后期进行基因识别更具有针对性。

七、问题四模型的建立与求解

问题四：

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。你们团队能否总结，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呢？

上面提出的基于频谱 3-周期性的基因预测四个方面问题中，“快速算法”与“阈值确定”是为设计基因预测算法做准备的。此外，在最后的延展性研究中，各队也可以对你们自己认为有价值的其它相关问题展开探讨。

问题分析：

（1）频谱或信噪比只是基因识别的众多判断特征之一。随着基因序列研究的不断发展，新的基因识别的判断特征不断被发现；同时，在众多的判断特征中，哪些是显著的，哪些是不显著的？这两个问题的解决就要求在应用基因识别算法时，判断特征的选取必须是动态的，允许新的判断特征的加入，而且选取需要按一定的指标进行有效筛选。

（2）检测基因突变就是检测出 DNA 序列的“错误”（如替换，删除或者插入等），这类似于信息论中的纠错码理论，频谱或信噪比方法在检测 3-周期性方面具有优势，但在检测 DNA 序列的单点突变方面并没有好的效果，Z-曲线的提出表明利用几何工具可以有效分析 DNA 序列，受此启发，本文基于改进的基于 DNA 序列的“四线”图，提出了基于改进“四线”图的 DNA 序列突变分析模型，为检测基因突变提供一种新的思路。

7.1 基因识别特征的动态筛选模型

基因识别研究中常用的方法有傅立叶[15]、Wang[16]、 Zhang[17]、频谱[18]等方法，这些方法均利用编码区与非编码区一种或多种特征差异进行识别，这些特征一般通过经验或观察选取，不一定是基因组的主特征。但如果将所有的特征进行简单相加，特征向量的维数会变得很大；并且，某些特征会干扰其它特征的判别，最终影响判别效率。

在此，我们列出了一系列基因的典型特征，提出了一种特征筛选方法，它能动态的筛选 DNA 序列编码区与非编码区主要的特征差异。将基因的组成特征、结构特征以及信号特征作为备选特征，为它们建立相应的特征模型；再通过模型计算特征得分，根据得分对特征集进行筛选和优化；然后将筛选过的特征得分组合成特征向量，继而在特征空间中利用判别分析进行训练和判别。对实际数据集的测试结果表明，我们提出的方法可以对编码区进行有效的识别。

7.1.1 动态特征筛选

1. 问题描述

将基因组序列视为一段由字母{A，C，G，T}组成的字符串 S，由开放阅读框架方法可提取出 ORF（open reading frame，开放阅读框）片断，这些片段中有的是基因 ORF，有的是非基因 ORF。假设对提取出的 ORF，该片段包含足够的信息来判别该 ORF 是不是一个基因 ORF，这样的片段称为待判别的序列。给定一组训练数据集，包括正数据集（经实验证实的基因 ORF）和负数据集（非基因 ORF），本节的基因识别就是通过训练，对任意给定的待判别 ORF，判断其是编码 ORF 还是非编码 ORF 序列。

2. 组成特征的选取

核苷酸的组成特征是基因的基本特征。如果把长度为 k 的核苷酸看作是一种“字”，那么，碱基使用偏好可看作是 1 字使用偏好，氨基酸使用偏好可看作是 3 字使用偏好，六核苷酸选用频率可看作是 6 字选用频率。我们首先考虑序列的 1-核苷酸，2-核苷酸，3-核苷酸和 6-核苷酸组成。长度为 k 的字具有 4^k 种可能组合。将 1-核苷酸使用偏好用 4 维向量 (f_A, f_T, f_G, f_C)，其中 f_A, f_T, f_G, f_C 分别表示 A、T、G 和 C 的使用频率。类似的，2-核苷酸和 3-核苷酸的使用频率分别用 16 维和 64 维向量描述。

6-核苷酸的使用频率具有 4096(4^6)种可能。所以我们使用香农熵进行降维，如下式所示。

$$H = -\sum_{w=1}^{4096} p(w)\log_2 p(w) \tag{7.1}$$

其中 $p(w)$ 表示 6-核苷酸的每种可能。

G+C 含量评估编码区序列的显著性。我们定义 G+C 含量与 A+T 含量的比值来描述这一特征。

因此，对 1-核苷酸，2-核苷酸，3 核苷酸，6-核苷酸和 G+C 含量，我们分别得到 4 维、16 维、64 维和 1 维向量。

3. 结构特征的选取

将位置在 1，4，7，…的核苷酸序列命名为子序列-1，位置在 2，5，8，…的核苷酸组成的序列命名为子序列-2，位置在 3，6，9，…的核苷酸命名为子序列-3。子序列-1 的嘌呤、氨基、强氢基分别由 r (1)、m(1)和 s(1)表示，子序列-2 中的嘌呤、氨基、强氢基由 r(2)、m(2)和 s(2)表示，子序列-3 中的嘌呤、氨基、强氢基由 r(3)、m(3)和 s(3)表示。

这样，我们用一个 9 维向量来描述该特征。

4. 信号特征的选取

终止密码子[19]是 DNA 序列中非常强的信号。终止密码子对编码区碱基的使用起重要限制作用，在编码区 DNA 序列中，TAA、TAG、TGA 的含量较低；在非编码区序列中，终止密码子的含量较高。我们用 TAA，TAG 或 TGA 的使用频率来描述这一特征。这样，对信号特征我们得到一个 1 维向量。

5. 特征的筛选和组合

至此，我们选取了序列的 7 个候选特征，分别是 1-核苷酸，2-核苷酸，3-核苷酸，6-核苷酸，G+C 含量，密码子频率和终止密码子频率。但这些特征对判别的贡献是不一样的，并不是每个特征都能提供显著的信息，判别力不强的特征还会干扰和影响判别效果。因此需要根据对判别实际贡献的大小对备选特征进行筛选，选取“最优”特征集。

为实现这个算法，我们需要构造两个数据集，一个是正数据集，它是编码序列的集合；一个是负数据集，它是非编码序列的集合。每个序列都可由向量 u 来表示，特征判别力的大小可通过计算正负数据集间 Mahalonobis[20]平方距离 D^2 来估计， D^2 越大，判别力也越大。

D^2 的计算方式如下：设编码序列为第一类，非编码序列为第二类，用下式计算每个类的 $\overline{U}_k^g$

$$\overline{U}_k^g = (\overline{u_1}^g, \overline{u_2}^g, \cdots, \overline{u_m}^g),\ g = 1,2 \tag{7.2}$$

其中 m 表示 k 个特征向量的维数，且

$$\overline{U}_j^g = \frac{1}{n_g}\sum_{l=1}^{n_g} u_{jl}^g,\ g = 1,2,\ j = 1,2,\cdots,m \tag{7.3}$$

用 $S^k = (s_{ij}^k)$ 两类的协方差矩阵，则

$$s_{ij}^k = \sum_{g=1}^{2}\sum_{l=1}^{n_g}(u_{ij}^g - \overline{u_i}^g)(u_{ij}^g - \overline{u_j}^g),\ i,j = 1,2,\cdots,m \tag{7.4}$$

k 个特征的 D_k^2 用下式计算：

$$D_k^2 = \left(\overline{U}_k^1 - \overline{U}_k^2\right)^T \left(S^k\right)^{-1}\left(\overline{U}_k^1 - \overline{U}_k^2\right) \tag{7.5}$$

特征筛选可以提高计算效率，减小干扰，提高判别精度，在特征较多时尤为必要。

使用我们的新方法来筛选最优特征，具体过程如下：

(1) 计算每个候选特征的 Mahalonobis 平方距离 D_1^2，将 D_1^2 按降序排列组成候选特征队列 Q。选择 D_1^2 值最大的特征作为特征集中的初始特征，放入 A 集合。

(2) 添加特征：假设我们已选取了 k 个特征到集合 A，对队列 Q 中剩下的候选特征 α，计算 $A+a$ 的 Mahalonobis 平方距离 D_{k+1}^2。如果 $D_{k+1}^2 - D_k^2 > d_0$（d_0 为计算过程中确定的阈值），则说明此特征能够提供显著的附加信息，则将 α 移入集合 A，并进入步骤(3)；否则，选择集合 Q 中的下一个特征并进入步骤(2)，直到队列 Q 为空，此时算法终止。

(3) 删除特征：假设步骤(2)中我们向集合 A 加入了新的特征，对于集合 A 中的任何其它特征 x，依次计算 $A-a$ 的 Mahalonobis 平方距离。如果

$D_{k+1}^2 - D_k^2 < d_0$，则说明随着新特征的加入，此特征提供的信息已经不再显著，则从集合 A 中删除 x，并进入步骤(2)。

经过以上过程的筛选，我们在集合 A 中得到了 3 个特征：密码子使用，6-核苷酸和终止密码子使用频率，分别用 9 个元素、1 个元素和 1 个元素的向量表示。把它们合成一个向量 $u_1 : u = (u_1, u_2, \cdots, u_{11})^T$。此时的特征集即为“最优”特征集。

6. 特征筛选算法的伪代码描述

以下为特征筛选的算法的伪代码描述：

输入: 候选特征向量
输出: 主要特征向量 A
开始

Step1: 对每个候选特征，计算 D_1^2。

Step2: 将 D_1^2 按降序排列组成候选特征队列 Q。

Step3: 选择 D_1^2 值最大的特征作为特征集中的初始特征，放入 A 集合。

Step4: 假设我们已选取了 k 个特征到集合 A，对队列 Q 中剩下的候选特征 α，计算 $A+a$ 的 Mahalonobis 平方距离 D_{k+1}^2。

若 $D_{k+1}^2 - D_k^2 > d_0$，则将 α 移入集合 A，并进入 Step5；

否则，选择集合 Q 中的下一个特征并进入 Step4，直到队列 Q 为空，此时算法终止；

Step5: 假设 Step4 中我们向集合 A 加入了新的特征，对于集合 A 中的任何其它特征 x，依次计算 $A-a$ 的 Mahalonobis 平方距离。

若 $D_{k+1}^2 - D_k^2 < d_0$，从集合 A 中删除 x，进入 Step4；

结束

7.1.2 数据库

为测试基于动态特征筛选算法的基因识别的性能，我们需要构建一个数据库。S.cerevisiae 是第一个被测序的真核单细胞生物，为了方便与已有结果的对比，本文使用该基因组来构建数据库。S.cerevisiae 基因组中共有 16 条染色体，全长 12.16Mbp。

S.cerevisiae 基因组共包含 6449 个 ORF，共分为 6 类，分别是已知的蛋白质(known proteins)、不相似类(no similarity)、有问题的 ORFs(questionable ORFs)、相似于或弱相似于已知的蛋白质(similarity or weak similarity to known proteins)、相似于未知的蛋白质(similarity to unknown proteins)和强相似于已知的蛋白质(strong similarity to known proteins)，每类分别包含中分别包含了 3410，516，471，820，1003 和 229 个序列。我们选择第一类的 3410 个 ORF 作为正数据集。

而负数据集用以下方式构建：

（1）记下每个 ORF 区的起始位置；

（2）从 16 条染色体中选出长度不小于 300bp 的基因间序列；

（3）在选出的长度大 300bp 的 DNA 序列中，从第一个碱基开始寻找密码子‘ATG’；然后寻找{TAA，TGA，TAG}直到找到第一个终止密码子。从所有的选出的序列中随机挑出 3410 条作为负数据库。

7.1.3 Fisher 判别

经过特性筛选之后，每条序列都由一个 11 维向量表示。我们使用 Fisher 判别来判别编码区与非编码区序列，Fisher 线性判别分析(FLDA)是由 Fisher 于 1936 年提出的用于两类问题特征提取的一种有效方法，其基本思想是寻找一投影方向，使训练样本投影到该方向时尽可能具有最大类间距离和最小类内距离，详细见文献[21]。应用 Fisher 判别法到以上构造的训练库，得到参数 c 和阈值 c_0，检验库中编码区与非编码区的判别则可简单由 $c\cdot u > c_0$ / $c\cdot u < c_0$ 得到。

参数 c 的计算方法如下：首先定义方差 S，令 u_{ij}^g 表示 g 组中第 k 条序列中的第 j 个分量，其中 $g=1,2;\ j=1,2,\cdots,11;\ k=1,2,\cdots,n_g(n_1=n_2)$。令 $\overline{u_j^g}$ 为 u_{ij}^g 中对所有 k 的均值，$\overline{U_g}$ 表示向量 $\overline{u_j^g}$，即 $\overline{U_g}=\left(\overline{u_1^g},\overline{u_2^g},\cdots,\overline{u_{11}^g}\right)^T$，其中 $g=1,2$。这样方差矩阵 S 中的第 i 行第 j 列的元素为

$$s_{ij}=\sum_{g=1}^{2}\sum_{k=1}^{n_g}(u_{ik}^g-\overline{u_i^g})(u_{ij}^g-\overline{u_j^g}),\ i,j=1,2,\cdots,11 \tag{7.6}$$

这样，参数可简单由下式确定：

$$c=S^{-1}\left(\overline{U_1}-\overline{U_2}\right) \tag{7.7}$$

其中，S^{-1} 为矩阵 S 的逆阵，具体细节可以参考文献[22]。

在训练库中，调整 c_0 值使编码区的错误率与非编码区的错误率相等，将此时的 c_0 值作为阈值。将得到的 Fisher 参数 $c_1,c_2,\cdots,c_{11}$ 及阈值 c_0 运用到检验库中，这样，如果 $c\cdot u > c_0$ 则序列被判别为编码序列，否则被判别为非编码区序列。其中 $c=\left(c_1,c_2,\cdots,c_{11}\right)$，$u=(u_1,u_2,\cdots,u_{11})^T$。

7.1.4 评价指标

对某一算法的优劣进行衡量的标准通常是：灵敏度和特异度，这两个测度经常用来描述一个算法或一个识别函数的正确度。要评估一个算法，重新替换(re-substitution)测试和交叉确认(cross-validation)测试通常被认为是一种行之有效的方法。重新替换测试反映了算法的自身的一致性而交叉确认测试反映了算法判断的有效性。在本模型中，我们使用 3 路交叉测试，即将正负数据集分别分为 3 部分，依次把其中 2 个部分作为训练集而余下的 1 部分作为测试集。

7.1.5 实验结果

利用训练集中的序列，我们就可以求出 Fisher 参数 $c_1,c_2,\cdots,c_{11}$ 及阈值 c_0（参见方程），利用这些值我们就可以得到这个算法应用于测试集中的正确度，以此来评价这个算法的优劣。这在这一节里，由于我们使用了 3 路交叉测试，因此有三个不同的训练集和测试集。实验的平台是基于 MATLAB 2011a，对于在每一个数据集上测试的结果进行性能评估，三个数据集测试的结果和它们的均值被列在表 7.1。

表 7.1 三个数据集测试结果

测试集	S_n(%)	S_p(%)	AC(%)
1	98.0	98.1	98.1

2	98.1	98.3	98.2
3	97.9	98.1	98.0
均值	98.0	98.2	98.1

作为比较，我们列出了 Zhang[16]和 Wang[17]的结果，Zhang 使用了编码序列的 1-核苷酸和密码子特征，Wang 整合了编码序列的多种特征。为使这样结果具有可比性，我们重新编写了上述两种方法的程序，并使用同一个数据库。

表 7.2 显示了这些方法比较的结果。从上面的结果可以看出，与其它算法相比较，本文的算法有比较高的正确度。

表 7.2 不同方法的识别结果

方法	S_n(%)	S_p(%)	AC(%)
Zhang 的方法	95.2	96.3	95.7
Wang 的方法	97.2	97.3	97.3
动态特征筛选方法	98.0	98.2	98.1

7.2 基于改进“四线”图的 DNA 序列突变分析模型

7.2.1 改进的 DNA 序列“四线”图

Milan Randic[23]基于 DNA 序列的“四线”图表示分析了序列间的比对。该“四线”图表示方法是把构成 DNA 序列的四种碱基 A、T、G 及 C 按照它在序列中的次序被依次分配到一个平面内等距的四条水平线上。每个碱基对应于这四条线上的一个点，用线段连接相邻的点，就可以得到一条“之”字形曲线。假设序列 $a=a_1a_2\cdots a_n\in\{A,C,G,T,-\}$，$1\le i\le n$，定义了一个如下的映射：

$$\rho(a_i)=\begin{cases}(i,h_1) & a_i=A\\(i,h_2) & a_i=T\\(i,h_3) & a_i=G\\(i,h_4) & a_i=C\\(i,\infty) & a_i='-'\end{cases}$$

其中 n 表示所研究序列的长度，$h_i(i=1,2,3,4)$ 是整数且它们相互之间不相等。例如，我们可以取 h_i 分别为 1，2，4 和 8，也可以取为-1，0，2 及 6 和 0，1，3 及 7 等数组。为了便于表达序列比对的结果，使用坐标 (i,∞) 来表示序列中的空位“—”。这种做法不影响序列的图形表示，但方便于后面的应用。

通过所定义的映射，很显然，DNA 序列中的每一个碱基都对应笛卡尔坐标平面内的一个点，一条序列对应于 2 维有序整数数组。例如，令 $h_1=1,h_1=2,h_1=4,h_1=8$，那么 DNA 序列 *g=ATGGCATTGACAAACTCG* 被映射为一个有序的整数数组{(1，1)，(2，2)，(3，4)，(4，4)，(5，8)，(6，1)，(7，2)，(8，2)，(9，4)，(10，1)，(11，8)，(12，1)，(13，1)，(14，1)，(15，8)，(16，2)，(17，8)，(18，4)}。这个有序整数数组对应于坐标平面内一个有序的点序列，用线段连接相邻的点，就得到一条如图 7.1 的图形曲线。很显然，通过该方法得到的 DNA 序列的图形有的如下特点：直观、简单。除此之外，我们可以根据图形曲线来重构一条唯一的 DNA 序列。也就是说，在 DNA 序列转化为图形曲线的过程中，序列中的信息没有丢失。因此，DNA 序列的图形曲线可以看作 DNA 序列的“特征”。

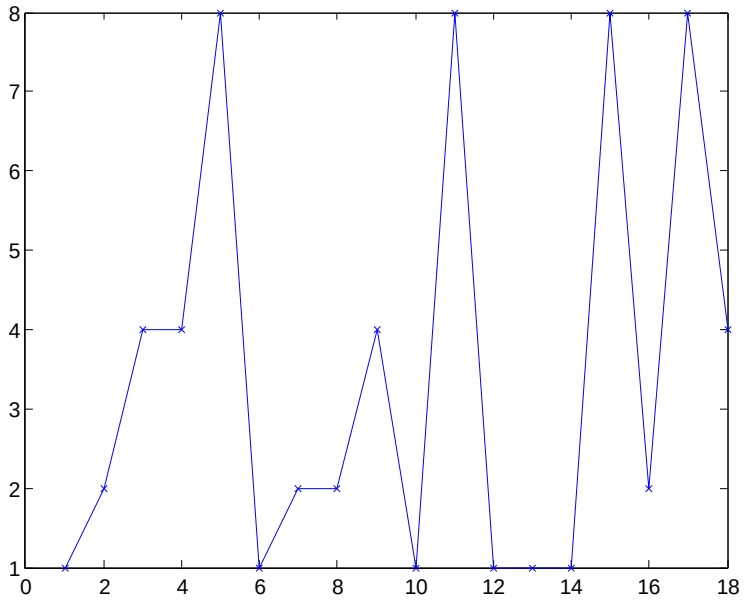

图 7.1 序列 g 改进的“四线”图

假设任意 2 条序列 $a=a_1a_2\cdots a_n$ 和 $b=b_1b_2\cdots b_m$，它们的长度分别为 n 和 m。我们可以证明如下的结论：

结论 7.1 如果 $\rho(a_i)-\rho(b_i)=(0,0)$，其中 $1\le i\le \min(n,m)$，那么 a_i 与 b_i 匹配。

证明：反证法。令 $\rho(a_i)=(i,\lambda),\rho(b_i)=(i,\mu)$，$\lambda,\mu\in\{h_1,h_2,h_3,h_4\}$。如果 a_i 与 b_i 不匹配，那么 λ 与 β 不相等，这与等式 $\rho(a_i)-\rho(b_i)=(0,0)$ 相矛盾。因此，a_i 与 b_i 一定匹配。

结论 7.2 如果 DNA 序列 $a=a_1a_2\cdots a_n$ 的图形曲线平行于 x-坐标轴平移 s 个单位（s 是一个整数），那么我们会得到一条新的 DNA 序列 $c=c_{1+s}c_{2+s}\cdots c_{n+s}$，且等式 $\rho(c_{i+s})-\rho(a_i)=(s,0)$，或 $\rho(c_{i+s})=(s,0)+\rho(a_i)$，对于满足 $1\le i\le n$ 的任意的 i 都成立。

证明：不妨假设序列中任意一个碱基 a_i 对应于点坐标 (i,λ)，其中 $1\le i\le n$。如果点 (i,λ) 平行 x-坐标轴平移 s 个单位，然后它将到达新序列的第 $i+s$ 个碱基的坐标位置 $(i+s,\lambda)$。因此等式 $\rho(c_{i+s})=(s,0)+\rho(a_i)$ 必然成立。

推论 7.1 对于满足 $1\le i\le \min(n,m)$ 的任意的 i，其中 m 和 n 分别为序列 a 和 b 的长度，如果 $\rho(b_i)-\rho(a_i)=(0,0)$，那么序列 a 是序列 b 的子序列或者序列 b 是序列 a 的子序列。

推论 7.2 序列 b 的图形曲线平行 x-坐标轴移动 s 个单位，然后形成一条新的图形曲线（我们不妨把它假设为新序列 $c=c_{1+s}c_{2+s}\cdots c_{n+s}$，(i) 如果 $\rho(c_{i+s})-\rho(a_i)=(0,0)$，其中 $\max(1,s+1)\le i\le \min(n,s+m)$，那么碱基 a_i 与新序列中的碱基 c_{i+s} 匹配。(ii) 如果 $\rho(c_i)-\rho(a_i)=(0,0)$，$i=d_1,d_1+1,\cdots,d_2$，其中 $\max(1,s+1)\le d_1\le d_2\le \min(n,s+m)$，那么序列 a 的子序列 $a_{d_1}a_{d_1+1}\cdots a_{d_2}$ 将与序列 b 中的子序列 $c_{d_1-s}c_{d_1+1-s}\cdots c_{d_2-s}$ 匹配。

推论 7.3 序列 a 中的子序列 $a_{i+1}a_{i+2}\cdots a_{i+d}$ 与序列 b 中子序列 $b_{j+1}b_{j+2}\cdots b_{j+d}$ 匹配，如果存在一个整数 s 使得平移 s 个单位后新序列 $c=c_{1+s}c_{2+s}\cdots c_{n+s}$ 满足 $\rho(c_{k+s})-\rho(a_k)=(0,0)$，其中 $\max(1,s+1)\le d_1\le d_2\le\min(n,s+m)$，而 $d_2-d_1\succ d$。那么序列 a 和 b 拥有最长的公共子串 $a_{d_1}a_{d_1+1}\cdots a_{d_2}$。

7.2.2 基于改进“四线”图的 DNA 序列突变模型

DNA 突变通常可以分成下列 4 类：替代，转移，插入和删除。前两种突变类型产生编码错误，而后两者突变导致 DNA 序列的长度发生改变。Liao[69]提出了一种判断 DNA 突变的方法，它是首先将 DNA 序列分别基于{A，C}，{A，T}和{A，G}分类标准作出 3 条特征曲线，然后根据这 3 条特征曲线来进行 DNA 突变分析。这种方法很显然给突变分析带来了可视化的优点，但在文献[24]中的方法存在着计算复杂性高的问题。

下面我们将基于改进 DNA 序列的“四线”图介绍判断 DNA 序列间的碱基突变的方法。在本文中，我们只考虑两序列对应位置上的碱基的替代、插入和删除的情况。令序列 $a=a_1a_2\cdots a_n$ 和 $b=b_1b_2\cdots b_m$，为了便于序列比对，假定 Δ_{ij} 对不同的 i 和 j 都是不相等的，其中 $\Delta_{ij}=h_i-h_j$，$i\neq j$ 且 $i,j\in\{1,2,3,4\}$，我们可以得到如下的定理。

结论 7.3 如果 $\rho(b_i)-\rho(a_i)=(0,h_2-h_1)$，那么序列 a 中的第 i 个碱基 A 被碱基 T 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_1-h_2)$，那么序列 a 中的第 i 个碱基 T 被碱基 A 替代。

结论 7.4 如果 $\rho(b_i)-\rho(a_i)=(0,h_3-h_2)$，那么序列 a 中的第 i 个碱基 T 被碱基 G 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_2-h_3)$，那么序列 a 中的第 i 个碱基 G 被碱基 T 替代。

结论 7.5 如果 $\rho(b_i)-\rho(a_i)=(0,h_3-h_1)$，那么序列 a 中的第 i 个碱基 A 被碱基 G 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_1-h_3)$，那么序列 a 中的第 i 个碱基 G 被碱基 A 替代。

结论 7.6 如果 $\rho(b_i)-\rho(a_i)=(0,h_4-h_3)$，那么序列 a 中的第 i 个碱基 G 被碱基 C 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_3-h_4)$，那么序列 a 中的第 i 个碱基 C 被碱基 G 替代。

结论 7.7 如果 $\rho(b_i)-\rho(a_i)=(0,h_4-h_2)$，那么序列 a 中的第 i 个碱基 T 被碱基 C 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_2-h_4)$，那么序列 a 中的第 i 个碱基 C 被碱基 T 替代。

结论 7.8 如果 $\rho(b_i)-\rho(a_i)=(0,h_4-h_1)$，那么序列 a 中的第 i 个碱基 A 被碱基 C 替代；如果 $\rho(b_i)-\rho(a_i)=(0,h_1-h_4)$，那么序列 a 中的第 i 个碱基 C 被碱基 A 替代。

结论 7.9 如果 $\rho(b_i)-\rho(a_i)=(0,\infty)$，那么序列 a 中的第 i 个碱基被删除或者插入；如果 $\rho(a_i)=(i,h_i)\big((i,h_2),(i,h_3)or(i,h_4)\big)$，那么序列 a 中的第 i 个碱基 A（T，G 或 C）被删除；如果 $\rho(b_i)=(i,h_i)\big((i,h_2),(i,h_3)or(i,h_4)\big)$，那么序列 a 中的第 i 个碱基 A（T，G 或 C）被插入。

结论 7.3 证明:不妨假设 $a_i \neq A,T, a_i \in \{A,C,G,T\}, b_i \neq A,T, b_i \in \{A,C,G,T\}$，那么 $a_i = G$ 或 C，$a_i = C$ 或 G，$\rho(b_i) - \rho(a_i) \neq (0, h_2 - h_1)$ 且 $\rho(b_i) - \rho(a_i) \neq (0, h_1 - h_2)$。很显然，这与条件 $\rho(b_i) - \rho(a_i) = (0, h_4 - h_2)$ 或 $\rho(b_i) - \rho(a_i) = (0, h_1 - h_2)$ 相矛盾。因此结论成立。

同理，我们能够证明其它结论也是成立的。基于以上的结论，通过平行 *y*-坐标轴的方向在不同序列的图形曲线间平移不同的向量（例如 $(0, h_2 - h_1)$ 和 $(0, h_1 - h_2)$），我们可以判断 DNA 序列间碱基突变的类型和突变的位置。例如，判断序列 *c=GTTCGACGGT* 和序列 *d=GCTCAATAT* 间的突变。为了便于叙述，不妨假设 $h_1 = 1, h_1 = 2, h_1 = 4, h_1 = 8$。序列的图形曲线如图 7.2 所示。在两序列的图形曲线间移动向量（0，6），我们发现在序列 *c* 中第 2 个位置上的碱基 *T* 被碱基 *C* 替代；移动向量（0，-2），发现在序列 *c* 中第 9 个位置上的碱基 *G* 被碱基 *T* 替代；移动向量（0，-5），发现第个和第 8 个位置上的碱基被碱基 *A* 替代；移动向量（0，-6），发现第 7 个位置上的碱基 *C* 被碱基 *T* 替代。

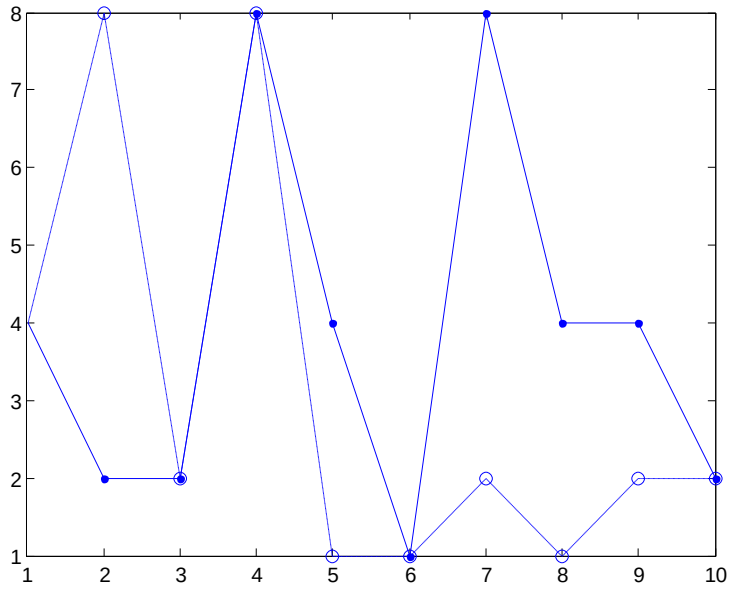

图 7.2 序列 c 和 d 基于改进的“四线”图的突变分析效果图

7.3 小结

本节针对目前常用的基因识别算法对特征选取的主观性，提出了基因识别特征的动态筛选模型，该模型根据特征之间的相关性来获取基因的主特征，利用提取的主特征进行编码区与非编码区的识别。针对基因突变问题，使用改进的 DNA 序列“四线”图，提出了一种图形化的分析 DNA 序列间碱基突变的方法，为检测基因突变提供一种新的思路。

八、结束语

生命科学技术的发展积累了大量的生物数据，这些数据里面隐藏着生命的规律，为人类探索生命的奥秘提供了大量的素材，为生物信息学的研究带来前

所未有的良机。同时，面对海量的数据，人类如何有效刻画、分析这些数据，从中提炼出生命的规律并加深对生命的认识，是广大科研人员面临的巨大挑战。本文在前人相关研究的基础上，综合运用信号处理理论等方法研究了 DNA 序列表示及基因识别的有关问题。尽管本文在 DNA 序列功率谱及信噪比计算和基因识别算法等方面做了一些努力，但由于生命科学本身的特殊性，还有很多内容诸如如何精确确定外显子区间端点等问题需要进一步研究。随着社会的进步和科技的发展，基因识别技术必将越来越成熟，越来越完善，基因识别技术的应用也会越来越广泛，我们期待基因识别早日为人类带来福音。

参考文献

[1]. 傅广操, 樊明捷. MATLAB 在现代功率谱估计中的应用[J]. 电脑学习. 2003 年 12 月, 第 6 期. p6.

[2]. 姚文俊, 自相关法和 Burg 法在 AR 模型功率谱估计中的仿真研究[J]. 计算机与数字工程, 2006, 6(35): 32-33.

[3]. 刘智, 游中胜, 邹枝玲. 基于小波变换的多元时间序列相似性研究[J]. 《西南师范大学学报(自然科学版)》2009 年 8 月，第 34 卷 04 期. 73-76.

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

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

[6]. Zhang C T, Zhang R. Z curves an intuitive tool for visualizating and analyzing the DNA sequences [J]. J Biomolec Struct Dyn, 1994, 11: 767-782.

[7]. Needleman S B，Wunsch C D．A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]．Journal of Molecular Biology，1970，48(3):443-453.

[8]. 谢惠民. 生物序列分析中的若干数学方法[J]. 高等应用数学学报 A 辑, 2005, 20(4):379-392.

[9]. Rushdi A, Tuqan J (2006) Gene identification using the Z-curve representation[J]. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, vol 2, pp 1024-1027.

[10]. Sharma SD, Shakya K, Sharma SN (2011) Evaluation of DNA mapping schemes for exon detection[J]. In: International conference on computer, communication and electrical technology, ICCCET 2011.

[11]. A.A.Tsonis, J.B.Elsner, P.A.Tsonis. Periodicity in DNA coding sequences: implications in gene evolution[J]. J Theor Biol, 1991, 151(3): 323-331.

[12]. A.M.Zoubir, D.R.Iskander. Bootstrap Techniques for Signal Processing[M]. Cambridge University Press, 2004: 1-15.

[13]. 程佩青. 数字信号处理教程.北京:清华大学出版社[M], 2001:359-368.

[14]. Yan M, Lin Z S, Zhang C T. A new fouirer transform approach for protein coding measure based on the format Z-curve[J]. Bioinformatics, 1998,14(8):

685-690.
[15].Tsonis A A, Elsner J B, Tsonis P A. Periodicity in DNA coding sequences: Implications in gene evolution[J]. J Theor Biol, 1991, 151 (3) : 323-331.
[16].Zhang C T, Wang J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve[J]. Nucl Acid Res, 2000, 28(14): 2804-2814.
[17].Wang Y H, Zhang C T, Dong P X. Recognizing shorter coding regions of human genes based on the statistics of stop codons[J]. Biopolymers, 2002, 63(3): 207-216.
[18].Kotlar D, Lavner Y. Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein-Coding Regions[J]. Genome Res, 2003, 13 (8): 1930-1937.
[19].Fickett J W, Tung C S. Assessment of protein coding measures[J]. Nucl Acid Res. 1992, 20 (24) : 6441-6450.
[20].王学仁. 实用多元统计分析[M]. 上海科学技术出版社, 1990.
[21].Goffeau A, Barrell B G, Bussey H, et al. Life with 6000 genes[J]. Science, 1996, 274(5287) : 546–567.
[22].Mardia K V, Kent J T, Bibby J M. Multivariate Analysis[J]. Academic Press, London, UK, 1979.
[23].Randi M, Zupan J, Vikic-Topic D, Plav i D. A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences[J]. Chemical Physics Letters, 2006, 431(4-6):375-379.
[24].Liao B, Ding K. Graphical approach to analyzing DNA sequences[J]. Journal of Computational Chemistry, 2005, 26(14):1519-1523.

附录

附录一

基于Lenvinson-Durbin递推算法求功率谱及信噪比

```
clear;
clc;
Ax=load('A.txt');
Tx=load('T.txt');
Cx=load('C.txt');
Gx=load('G.txt');
order1=100;
nfft=256;
fs=1638;
range='onesided';
t0=clock;
[PAxx,fA]=pyulear(Ax,order1,nfft,fs,range);
[PTxx,fT]=pyulear(Tx,order1,nfft,fs,range);
[PCxx,fC]=pyulear(Cx,order1,nfft,fs,range);
[PGxx,fG]=pyulear(Gx,order1,nfft,fs,range);
Pxx=PAxx+PTxx+PCxx+PGxx;
Pxx(1)=mean(Pxx(2:length(Pxx)));
t1=clock;
plot(fA,Pxx);
xlabel('Frequency(hz)');
ylabel('Power Spectral Density');
title('Lenvinson-Durbin Psd Estimate');
SNR=Pxx( floor(nfft/3)+1)/mean(Pxx)
etime(t1,t0)
```

基于 Burg 递推算法求功率谱及信噪比

```
clear;
clc;
Ax=load('A.txt');
Tx=load('T.txt');
Cx=load('C.txt');
Gx=load('G.txt');
order1=50;
nfft=256;
fs=1638;
range='onesided';
t0=clock;
[PAxx,fA]=pburg(Ax,order1,nfft,fs,range);
```

```
[PTxx,fT]=pburg(Tx,order1,nfft,fs,range);
[PCxx,fC]=pburg(Cx,order1,nfft,fs,range);
[PGxx,fG]=pburg(Gx,order1,nfft,fs,range);
Pxx=PAxx+PTxx+PCxx+PGxx;
Pxx(1)=mean(Pxx(2:length(Pxx)));
t1=clock;
plot(fA,Pxx);
xlabel('Frequency(hz)');
ylabel('Power Spectral Density');
title('Burg Psd Estimate');
SNR=Pxx( floor(nfft/3)+1)/mean(Pxx)
etime(t1,t0)
```

附录二

基于 AR 模型的固定长度滑动窗口功率谱的基因识别算法

```
clear;
clc;
right=0;
all=0;
M=99;
load('Genes100.mat');
startpos=2;
endpos=2;
for j=startpos:endpos
      genes(j).SNRplot=zeros(1,genes(j).Length);
      for k=1:genes(j).Length
   fs=min(genes(j).Length,k+(M-1)/2)-max(1,k-(M-1)/2);
   Ax=genes(j).ua(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Cx=genes(j).uc(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Gx=genes(j).ug(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Tx=genes(j).ut(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
%fs=genes(j).ExonPos(k*2)-genes(j).ExonPos(k*2-1);
order1=min(30,fs);
nfft=256;
range='onesided';
[PAxx,fA]=pyulear(Ax,order1,nfft,fs,range);
[PTxx,fT]=pyulear(Tx,order1,nfft,fs,range);
[PCxx,fC]=pyulear(Cx,order1,nfft,fs,range);
[PGxx,fG]=pyulear(Gx,order1,nfft,fs,range);
Pxx=PAxx+PTxx+PCxx+PGxx;
Pxx(1)=mean(Pxx(2:length(Pxx)));
SNR1=max(Pxx( ceil(nfft/3):ceil(nfft/3)))/mean(Pxx(2:length(Pxx)));
SNRplot(k)=SNR1;
      end
```

```
        plot(1:genes(j).Length,SNRplot);
        for l=1:length(genes(j).ExonPos)/2
line([genes(j).ExonPos(l*2-1),genes(j).ExonPos(l*2)],[min(SNRplot),min(SNRplot)],'LineWidth',4);
        hold on
        end
end
```

基于 AR 模型移动序列功率谱的基因识别算法

```
clear;
clc;
right=0;
all=0;
M=99;
load('Genes100.mat');
startpos=2;
endpos=2;
for j=startpos:endpos
        genes(j).SNRplot=zeros(1,genes(j).Length);
        for k=1:genes(j).Length
    fs=k;
    Ax=genes(j).ua(1:k);
    Cx=genes(j).uc(1:k);
    Gx=genes(j).ug(1:k);
    Tx=genes(j).ut(1:k);
order1=min(30,fs);
nfft=256;
range='onesided';
[PAxx,fA]=pyulear(Ax,order1,nfft,fs,range);
[PTxx,fT]=pyulear(Tx,order1,nfft,fs,range);
[PCxx,fC]=pyulear(Cx,order1,nfft,fs,range);
[PGxx,fG]=pyulear(Gx,order1,nfft,fs,range);
Pxx=PAxx+PTxx+PCxx+PGxx;
Pxx(1)=mean(Pxx(2:length(Pxx)));
SNR1=max(Pxx( ceil(nfft/3):ceil(nfft/3)))/mean(Pxx(2:length(Pxx)));
SNRplot(k)=SNR1;
         end
        maxsnr=max(SNRplot);
        for k=1:genes(j).Length
                SNRplot(k)=SNRplot(k)/maxsnr;
        end
        plot(1:genes(j).Length,SNRplot);
end
```

附录三

重采样阈值确定算法

```
clear;
clc;
all=0;
M=99;
load('Genes100.mat');
startpos=1;
endpos=10;
best=zeros(1,endpos-startpos+1);
for j=startpos:endpos
      genes(j).SNRplot=zeros(1,genes(j).Length);
      for k=1:genes(j).Length
   fs=min(genes(j).Length,k+(M-1)/2)-max(1,k-(M-1)/2);
   Ax=genes(j).ua(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Cx=genes(j).uc(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Gx=genes(j).ug(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
   Tx=genes(j).ut(max(1,k-(M-1)/2):min(genes(j).Length,k+(M-1)/2));
order1=min(30,fs);
nfft=256;
range='onesided';
[PAxx,fA]=pyulear(Ax,order1,nfft,fs,range);
[PTxx,fT]=pyulear(Tx,order1,nfft,fs,range);
[PCxx,fC]=pyulear(Cx,order1,nfft,fs,range);
[PGxx,fG]=pyulear(Gx,order1,nfft,fs,range);
Pxx=PAxx+PTxx+PCxx+PGxx;
Pxx(1)=mean(Pxx(2:length(Pxx)));
SNR1=max(Pxx( ceil(nfft/3):ceil(nfft/3)))/mean(Pxx(2:length(Pxx)));
SNRplot(k)=SNR1;
     each_length=length(genes(j).ExonPos)/2-1;
   if each_length~=0
   eachSNR=zeros(1,each_length);
   for k=1:each_length
   eachSNR(k)=max(SNRplot(genes(j).ExonPos(k*2):genes(j).ExonPos(k*2+1)));
   end
   genes(j).bestSNR=max(eachSNR);
   best(j)=genes(j).bestSNR;
   else
best(j)=max(max(SNRplot(1:genes(j).ExonPos(1))),max(SNRplot((genes(j).ExonPos(2)):genes(j).Leng
th)));
   end
end
save('Genes100_Exon_SNR.mat',    'genes');
save('best.mat',    'best');
```