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 基于谱分析的 DNA 序列识别算法研究

摘 要：

本文对基于谱分析的 DNA 序列识别中相关问题建立了数学模型，并进行了计算及结果分析。

针对问题一 Voss 映射下 DFT 算法效率低的问题，提出了一种基于频数二次型的功率谱快速算法和基于帕斯瓦尔定理的信噪比快速算法，该算法与 DFT 及其快速算法 FFT 的计算量对比见表 1；对 Voss 映射和 Z-curve 映射下的谱关系进行了研究，结果表明 Z-curve 映射是 Voss 映射的线性变换和降维，Z-curve 映射的功率谱值和信噪比值分别是 Voss 映射下功率谱值和信噪比值的 4 倍和 4/3 倍；研究了实数映射 $A\to 0, C\to 1, G\to 2, T\to 3$，并给出了功率谱和信噪比的快速计算公式，该公式适用于任意实数映射。

表1 计算量分析

计算量	DFT	FFT	频数二次型
实数乘	$4N^2$	$2N\log_2 N$	48
实数加	$4N^2-2N$	$3N\log_2 N$	$N+5$

针对问题二阈值确定的问题，提出了一种能同时反映敏感性和特异性的基因预测准确率指标，并基于该指标建立了最优固定阈值分析法的预备模型，用于样本训练时确定已注释基因的阈值；建立了基于模糊逻辑的自适应阈值模型，该模型综合考虑基因序列信噪比的均值、方差和峰值等特征，确定了模糊集合和模糊逻辑规则；根据自适应阈值模型计算出了 100 个人和鼠基因的固定阈值和自适应阈值，对预测结果的评估分析表明自适应阈值法优于固定阈值法，表 2 列举了 3 个基因的固定阈值和自适应阈值以及不同阈值下的预测正确率。

表 2　固定阈值与自适应阈值的预测正确率

基因序号	固定阈值	正确率	自适应阈值	正确率	改进度
30	1.6	0.6783	1.4446,1.4345,2.3696,0.80887,1.8158,1.2983	0.8488	25.14%
35	1.8	0.7685	1.7832,0.44012,1.75,1.4659	0.8899	15.80%
97	1.9	0.6496	1.7464,1.6108,1.9251	0.9408	44.83%

针对问题三，提出了两种确定外显子端点的方法，一是基于小波变换的梯度边缘检测法，通过对端点模糊区的信噪比曲线进行多项式拟合实现了曲线的平滑，再用梯度算法检测端点位置；通过分析和仿真表明了序列的多次重复可放大序列的 3-周期特性，利用这一现象提出了基于重复序列的边界搜索算法，进一步提高了端点定位的准确性。对 27 号基因上一段起始端点在 802 位的外显子进行仿真，初步预测的端点在 771 位，边缘检测法将端点确定为 817 位，边界搜索算法将端点确定为 794 位，定位误差缩小至 8 位。综合前三问的解决方法，对 6 个未注释的基因进行了预测（编码区位置见正文 27 页）。

在延展性研究中，引入信号增强技术应用于短编码序列的谱分析，对 3-周期性较强的部分进一步增强，而抑制 3-周期性较弱的部分，提高了预测效果；将谱分析应用于检测基因突变，仿真和分析表明，当处于外显子内部的碱基发生删除或插入的突变时，该处的 3-周期特性被破坏，其信噪比曲线在突变处会出现波谷，称之为“鞍部效应”。针对这一现象提出了伪鞍部识别算法，首先检测信噪比曲线中的鞍部，在鞍部中心附近尝试插入/删除碱基，通过观察鞍部是否消失来判断是否存在碱基删除/插入形式的突变。

目　录

基于谱分析的 DNA 序列识别算法研究

1 问题引入

基因是指携带有遗传信息的DNA序列，由腺嘌呤（A），鸟嘌呤（G），胞嘧啶（C），胸腺嘧啶（T）这四种核苷酸（碱基）符号按一定的顺序连接而成。基因通过指导蛋白质的合成来表达自己所携带的遗传信息，从而控制生物个体的性状表现。基因预测，一般是指识别出DNA序列中编码区（即外显子）。

随着人类基因组计划的实施和顺利完成，基因预测成为生物信息学中最基础，也是最首要的问题。真核生物的DNA结构较为复杂，在目前基因预测研究中，采用数字信号处理（Digital Signal Processing，DSP）方法来发现基因编码序列受到广泛重视。

长为N的DNA序列在利用傅里叶变换对数值化映射后的基因序列进行频谱分析中，外显子序列的功率谱曲线在频率k=N/3处，具有较大的频谱峰值，称为3-周期特性。这种特性在基因编码区以外是不存在的。产生这种现象的原因是由于编码区碱基分布的不平衡性。本文基于DNA序列信号频谱的3-周期特性，对以下四个问题进行深入的探讨：

（1）功率谱与信噪比的快速算法

探求Voss映射下功率谱与信噪比的快速计算方法；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；对实数映射的可行性进行分析。

（2）对不同物种类型基因的阈值确定

对不同的基因类型研究其阈值确定方法和阈值结果；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进行分析。

（3）基因识别算法的实现

针对未被注释的、完整的DNA序列，设计基因识别算法并评估算法的准确率，并将算法用于对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

在基因识别研究中，对其它有价值的相关问题展开探讨。

2 DNA 序列的数值映射及谱分析

本章主要解决问题（1）。由于DNA序列是由A、T、G、C 四种碱基构成的字符序列，为了能够让计算机处理，必须先将这些字符序列转换成二进制数值序列，然后才能采用数字信号处理的相关方法对其进行进一步地处理，因此将DNA字符序列转换成数值序列是基于DSP方法对序列进行分析的基础。目前有很多种映射方法，并且每种方法都有自身的优势。本章重点研究了Voss映射、Z-curve映射与实数映射，并提出了不同映射条件下功率谱与信噪比的快速计算方法。

2.1 研究现状

研究表明，DNA序列数值映射方法的优劣会直接影响到最终分析结果以及生物学意义的解释[1]。目前，DNA序列数值映射方法可分为数值表示法和图形表示法两大类。表2-1列举了部分数值映射方法。数值表示法是将序列中的四个字符按照某种规则映射成相应的数值序列，以达到对DNA序列进行数值分析的目

的。其中按照被分析的碱基个数，数值表示法又可分为以单碱基为映射对象的数值表示和以多碱基为对象的数值表示。映射后的数值可为实数也可为复数。如果数值选用实数，会将原来变化的信号（即交流分量）变为不变的信号（即直流分量），容易使得DNA序列频谱中的直流分量增大，而交流分量削弱；如果数值选用复数，直流分量被抑制得相对较小，而在3-周期处的分量被增大，易于观察基因的编码区。图形表示法是将序列表示成几何空间中的曲线。其最大的优势是使得基因数据的可视化。目前，国内外提出了大量的DNA序列的图像表示法，具有代表性的有五维空间上的G曲线法[2]、三维空间上的H曲线法[3]、基于笛卡尔坐标的二维行走曲线[4]。数值表示法中最具有代表性的方法是Voss映射法，而图形表示法中最具代表性的Z-curve映射法。

表2-1 DNA序列数值映射方法举例

映射方法	碱基			
	A	C	G	T
整数法	0	1	2	3
电子势能法	0.1260	0.1340	0.0806	0.1335
复数法	1+j	-1-j	-1+j	1-j

基于统计学特征的方法分为基于传统数字信号处理技术的方法和基于智能信号处理的方法。在基于传统信号处理技术中，傅里叶分析和数字滤波器是被用于识别有3-周期特性的编码区中最常用的方法[5]。这两类方法类似的结果可以通过平均幅值差分函数（AMDF）和时域周期图（TDP）获得[6]。文献[7]采用改进后的小波方法提高了识别精度。基于智能信号处理技术的方法采用的主要智能技术包括：神经网络、隐形马尔科夫模型，支持向量机，动态规划等[8]。

在目前对生物信息领域的研究热潮下，各种DNA序列的数值映射和基因预测方法层出不穷，预测精确度提高的同时算法日趋复杂。因此，本章节将以经典的、基础的映射方法为基础，以降低算法复杂度、加快计算速度为出发点，力求有所突破。

2.2 符号说明与假设

表 2-2　符号约定与含义

符号	含义
I	A、T、G、C 四种碱基集合
S	长度为 N 的 DNA 序列
$u_b(n)$	基因序列数值映射数列
$U_b(k)$	DFT 频谱序列
$P(k)$	功率谱序列
$\bar{E}$	功率谱的平均值
R	信噪比
x_b, y_b, z_b	$b \in I$ 出现在序列的 0,3,... N-3 与 1,4,…N-2 及 2,5,…N-1 等位置上的频数

2.3 Voss 映射与快速谱分析

2.3.1 Voss 映射与谱分析原理

Voss映射是最早出现且最简单的将DNA字符序列转化成数值序列的方法[9]，因此在基因序列的研究中被广泛地使用。它的基本思想是将DNA序列按照四种碱基存在与否映射成四维的数值序列，在某一个位置，如果某种碱基存在就将其映射为1，否则映射为0。

令 $I=\{A,T,G,C\}$，长度（即碱基符号的个数，单位记为 bp）为 N 的任意 DNA 序列，可表达为 A、T、G、C 的字符序列：

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

现对于任意确定的 $b\in I$，令：

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

通过将字符序列映射为数值序列，把它看作一条离散的时间信号，就可以使用数字信号领域的序列分析方法来对它进行分析。傅立叶变换是一种常见的从变换域的角度处理信号的方法，这种方法将 DNA 序列从时域映射到频域，并且运用谱分析的方法查找 DNA 序列中的外显子。对于长度为 N 的序列，其傅立叶变换为：

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

计算每个复序列 $\{U_b(k)\}$ 的平方功率谱，并相加则得到整个 DNA 序列 S 的功率谱序列 $\{P(k)\}$：

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

记 DNA 序列 S 的总功率谱的平均值为：

$$\bar{E}=\sum_{k=0}^{N-1}P(k)\Big/N \tag{2-4}$$

而将 DNA 序列在特定位置，即 $k=N/3$ 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：

$$R=\frac{P(N/3)}{\bar{E}} \tag{2-5}$$

外显子序列的功率谱曲线在频率 $k=N/3$ 处，具有较大的频谱峰值，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-周期特性。根据3-周期性，信噪比 R 大于某个适当选定的阈值 R_0（比如 $R_0=2$）的DNA片段，可作为DNA序列上可能的编码序列（外显子）。

2.3.2 Voss 映射下的快速谱算法

对于很长的DNA序列，在计算其功率谱或信噪比时，DFT变换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。根据DFT公式(2-1)，求出N点$U_b(k)$需要N^2次复数乘法和$N(N-1)$次复数加法。众所周知，实现一次复数乘法需要四次实数乘法和两次实数加法，实现一次复数加法需要两次实数加法。尽管DFT可用快速傅里叶变换FFT计算，仍然难以快速处理长度数以千记的DNA序列。为此，本文提出一种快速功率谱与信噪比算法。

2.3.2.1 基于频数二次型的功率谱快速算法

在长为N的DNA序列中，若N为3的倍数，将碱基符号$b \in I = \{A,T,G,C\}$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$k = N/3$处的总功率谱值即为：

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

$$= \sum_{b\in I}\left|x_b + y_b\cdot e^{-j\frac{2\pi}{3}} + z_b\cdot e^{j\frac{2\pi}{3}}\right|^2 = \sum_{b\in I}(x_b^2 + y_b^2 + z_b^2 - x_b y_b - x_b z_b - y_b z_b) \qquad (2\text{-}6)$$

公式(2-6)根据Voss映射的特点将DFT计算转为碱基出现频次的运算，说明DNA序列在$k = N/3$处的功率谱可以通过计算碱基出现的频次而获得。

线性代数中，设F是一个数域，F上n元二次齐次多项式

$$q(x_1, x_2, \cdots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{n-1,n}x_{n-1}x_n$$

叫做F上的一个n元二次型。若$a_{ij} = a_{ji}(1 \le i, j \le n)$，令$\mathbf{A} = (a_{ij})$，二次型可以写为：

$$q(x_1, x_2, \cdots, x_n) = (x_1, x_2, \cdots, x_n)\mathbf{A}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

根据二次型的定义，公式(2-6)可以写为：

$$P(\frac{N}{3}) = \sum_{b\in I}(x_b, y_b, z_b)\begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{pmatrix}\begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} = \sum_{b\in I}\mathbf{BAB}^T \qquad (2\text{-}7)$$

其中，$\mathbf{B} = (x_b, y_b, z_b)$。

2.3.2.2 基于帕斯瓦尔定理的平均功率和信噪比快速算法

帕斯瓦尔定理表明，信号在时域和频率变换域中能量是守恒的，其公式为：

$$\sum_{n=0}^{N-1}|u_b(n)|^2=\frac{1}{N}\sum_{k=0}^{N-1}\left|U_b(k)\right|^2 \tag{2-8}$$

公式（2-8）表明，序列功率谱的平均值等于序列时域值的平方和。DNA序列总功率的平均值是四个分序列功率谱均值之和：

$$\bar{E}=\sum_{b\in I}\sum_{n=0}^{N-1}|u_b(n)|^2=N \tag{2-9}$$

将公式(2-7)和公式(2-9)代入2.3.1节信噪比的表达式，可以得到信噪比的快速计算表达式：

$$R=\frac{\sum_{b\in I}\mathbf{BAB}^T}{N} \tag{2-10}$$

2.3.3 计算结果与分析

2.3.2.1节给出了基于频数二次型的功率谱快速算法，以此为基础，2.3.2.2节给出了基于帕斯瓦尔的信噪比快速算法。与DFT分析法相比，本文的快速算法复杂度降低，计算量锐减。表2-3分析了DFT、FFT和频数二次型三种方法求解 $k=N/3$ 处的功率谱值的计算量。由该表可看出，随着N的增大，快速算法的优势越来越明显。当DNA序列长度N为5000时，FFT计算的实数乘法可达快速算法的2560倍。

表2-3 计算量分析

计算量	DFT	FFT	频数二次型
实数乘	$4N^2$	$2N\log_2 N$	48
实数加	$4N^2-2N$	$3N\log_2(N)$	$N+5$

根据本文提出的功率谱和信噪比对人类线粒体基因NC_012920_1进行分析，图2-1显示了基于频数二次型求解功率谱的结果，图2-2给出了基于帕斯瓦尔定理求解信噪比的结果。与题目给出的曲线相比可以看出，该快速算法与FFT算法求解的结果十分吻合，表明了快速算法的正确性。

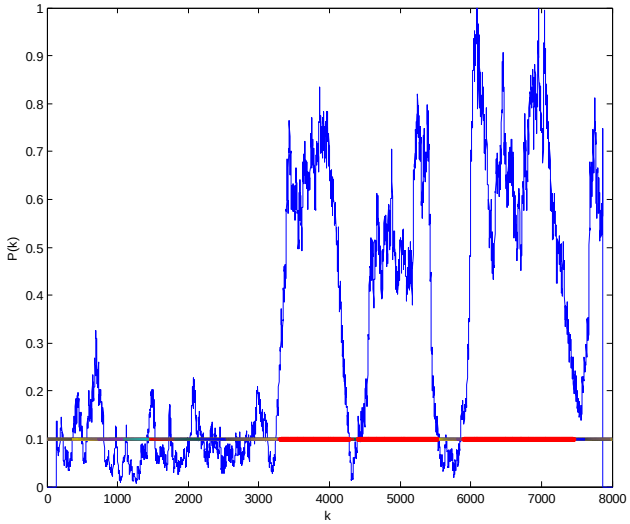


图2-1　Voss映射下频次二次型法求解功率谱（人类线粒体基因NC_012920_1）

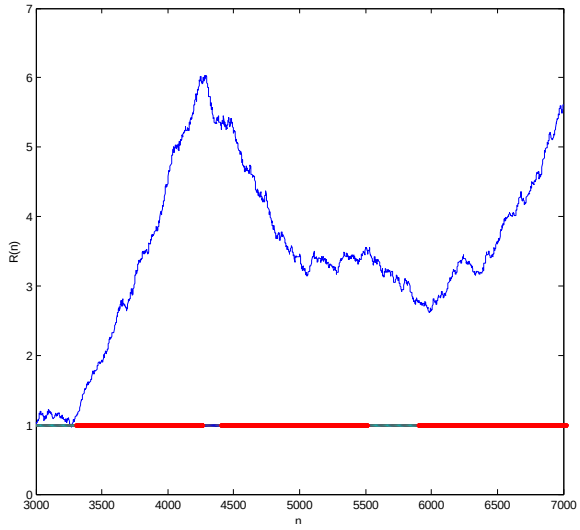


图2-2　Voss映射下帕斯瓦尔定理求解信噪比（人类线粒体基因NC_012920_1）

2.4 Z-curve 和 Voss 映射下的谱分析关系

2.4.1 Z-curve 映射与谱分析原理

1994 年，张春霆院士提出了基因序列的 Z-curve 定义[9]。设 DNA 序列 S 的四个指示序列 $\{u_b(n)\}$ 的累积序列 $b_n(n=0,1,...,N-1)$ 为 $b_n=\sum_{i=0}^{n-1}u_b(i)$。则定义三个序列：

$$\begin{cases} x(n)=2(A_n+G_n)-n \\ y(n)=2(A_n+C_n)-n \\ z(n)=2(A_n+T_n)-n \end{cases} \tag{2-11}$$

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 $\Delta z(n)=z(n)-z(n)$，于是得到 Z-curve 映射：

$$\begin{pmatrix} \Delta x(n) \\ \Delta y(n) \\ \Delta z(n) \end{pmatrix}=\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} u_A(n) \\ u_C(n) \\ u_G(n) \\ u_T(n) \end{pmatrix} \tag{2-12}$$

Z-curve 映射的总功率谱：

$$P_Z(k)=|\Delta X(k)|^2+|\Delta Y(k)|^2+|\Delta Z(k)|^2 \tag{2-13}$$

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

同样，定义 Z-curve 映射的信噪比为：

$$R_Z=\frac{P_Z\left(\frac{N}{3}\right)}{\overline{E}}=\frac{\left|\Delta X\left(\frac{N}{3}\right)\right|^2+\left|\Delta Y\left(\frac{N}{3}\right)\right|^2+\left|\Delta Z\left(\frac{N}{3}\right)\right|^2}{\overline{E}} \tag{2-14}$$

其中 $\bar{E}=\sum_{k=0}^{N-1}P_Z(k)\Big/N$ 是 Z-curve 映射的平均功率谱。

从公式(2-12)可以看出，对于序列 $x(n)$，A 和 G 被表示成 1，而 C 和 G 被表示为-1；同样，在序列 $y(n)$ 中，A 和 C 被表示为 1，而 G 和 T 被表示为-1；序列 $z(n)$ 中，A 和 T 被表示为 1，而 C 和 G 被表示为-1。从某中程度上来说，这种表示方法具有一定的生物特性，比如 A 和 G 都是嘌呤而 C 和 G 均为嘧啶，A 和 C 都属于氨基类而 G 和 T 为酮类，A 和 T 之间存在弱氢键，而 C 和 G 之间存在着强氢键。根据 Z-curve 的既能够表现生物特性，又能够节省计算时间的特征，这种表示方法得到了广泛地应用。

2.4.2 Z-curve 映射与 Voss 映射的关系

四种碱基的数目符合 $A_n+T_n+G_n+C_n=n$，对特定的 n， A_n、T_n、G_n、C_n 中只有三个是独立的，Z-curve 就是将三个独立的数用三维空间坐标系中的一点 P_n 唯一表示出来了。对于长为 N 的 DNA 序列，当 n 从 1 到 N 增加时，三维空间中表述 A_n、T_n、G_n、C_n 关系的点 $P_1,P_2,...,P_N$ 的连线就构成了 Z 曲线。因此，Z-curve 映射本质上是 DNA 序列的几何表达形式。Voss 映射是四维映射，而 Z-curve 映射是三维映射，可以说 Z-curve 映射是对 Voss 映射所得到的序列进行线性变换得到的，下面定量分析 Z-curve 映射与 Voss 映射下功率谱和信噪比的关系。

2.4.2.1 功率谱的关系

在 Z-curve 映射下，总功率谱值为 $\Delta x(n),\Delta y(n),\Delta z(n)$ 三个序列功率谱值之和，而三个序列的功率谱又可由同样的方法得到。下面以 $\Delta x(n)$ 序列为例，给出求解功率谱的公式。

由公式(2-12)不难发现， $\Delta x(n)$ 只能取到两个值：-1 或 $1(n=0,1,...,N-1)$。这是由于 Voss 映射中 $u_A(n),u_C(n),u_G(n),u_T(n)$ 在处 n 只有一个为 1，其余三个都为 0 的缘故。在 DNA 序列的第 n 位，$\Delta x(n)=1$ 就表示该位上碱基 A 或 G 的出现，而 $\Delta x(n)=-1$ 则表示该位上碱基 C 或 T 的出现。$\Delta x(n)$ 序列中 1 的总个数表示碱基 A 和 G 出现的频次，-1 的总个数表示碱基 C 和 T 出现的频次。基于这一关系， $\Delta x(n)$ 序列在 $k=N/3$ 的功率谱值为：

$$
\begin{aligned}
P_Z^X\left(\frac{N}{3}\right)&=\left|\Delta X\left(\frac{N}{3}\right)\right|^2=\left|\sum_{n=0}^{N-1}\Delta x(n)\cdot e^{-j\frac{2\pi}{3}n}\right|^2\\
&=\left|\sum_{m=0}^{N/3-1}\left(\Delta x(3m)+\Delta x(3m+1)e^{-j\frac{2\pi}{3}}+\Delta x(3m+2)e^{j\frac{2\pi}{3}}\right)\right|^2 \qquad (2\text{-}15)\\
&=\left|x_b'+y_b'e^{-j\frac{2\pi}{3}}+z_b'e^{j\frac{2\pi}{3}}\right|^2=x_b'^2+y_b'^2+z_b'^2-x_b'y_b'-x_b'z_b'-y_b'z_b'
\end{aligned}
$$

其中 $x_b'=x_A+x_G-x_C-x_T$， $y_b'=y_A+y_G-y_C-y_T$， $z_b'=z_A+z_G-z_C-z_T$。

$P_Z^Y(N/3)$ 和 $P_Z^Z(N/3)$ 同理可得。由公式(2-15)可发现， $P_Z^X(N/3)$、$P_Z^Y(N/3)$

和 $P_Z^Z(N/3)$ 同样只与 Voss 映射下不同碱基在不同位置上出现的频次有关，这样就将 Z-curve 映射和碱基出现的频次联系起来了。那么，在 $k=N/3$ 处 Z-curve 映射下的总功率谱值为：

$$\begin{aligned} P_Z\left(\frac{N}{3}\right) &= P_Z^X\left(\frac{N}{3}\right)+P_Z^Y\left(\frac{N}{3}\right)+P_Z^Z\left(\frac{N}{3}\right) \\ &= 4\sum_{b\in I}(x_b^2+y_b^2+z_b^2-x_by_b-x_bz_b-y_bz_b)=4\sum_{b\in I}\mathbf{BAB}^T \end{aligned} \tag{2-16}$$

公式(2-16)表明：在 $k=N/3$ 处，Z-curve 映射下总功率谱值是 Voss 映射下总功率谱值的 4 倍。

2.4.2.2 信噪比的关系

长度为 N 的 DNA 序列，$\Delta x(n),\Delta y(n),\Delta z(n)$ 序列仅有-1 和 1 两个值，因此根据帕斯瓦尔定理 $\Delta x(n),\Delta y(n),\Delta z(n)$ 三个序列的功率平均值是其序列长度 N。在 Z-curve 映射下，总功率谱的平均值如公式(2-17)所示。

$$\bar{E}_Z=\bar{E}_{\Delta x}+\bar{E}_{\Delta y}+\bar{E}_{\Delta z}=3N \tag{2-17}$$

公式(2-17)表明：在 $k=N/3$ 处，Z-curve 映射下的总功率谱均值是 Voss 映射下总功率谱均值的 3 倍。

根据公式(2-16)和(2-17)，Z-curve 映射下的信噪比公式为：

$$R=\frac{4\sum_{b\in I}\mathbf{BAB}^T}{3N} \tag{2-18}$$

公式(2-18)表明：在 $k=N/3$ 处，Z-curve 映射下信噪比值是 Voss 映射下信噪比值的 4/3 倍。

2.4.3 计算结果与分析

图 2-3 和图 2-4 给出了 Z-curve 映射下快速算法求解功率谱和信噪比的结果，与题目给出的曲线十分吻合，表明了 Z-curve 映射下快速算法的正确性。

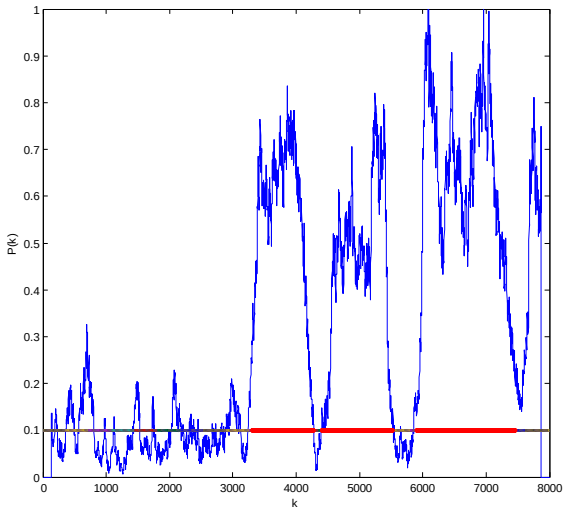


图2-3 Z-curve映射下频次二次型法求解 $k=N/3$ 处功率谱（人类线粒体基因

NC_012920_1）

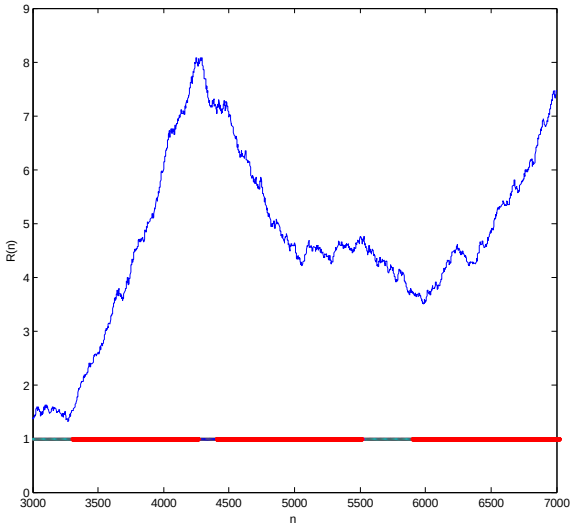


图2-4 Z-curve映射下帕斯瓦尔定理求解信噪比（人类线粒体基因NC_012920_1）

结论：

对于长度为 N 的 DNA 序列，Z-curve 映射和 Voss 映射存在以下关系：

1. Z-curve 映射是对 Voss 映射所得到的序列进行线性变换和降维处理后得到的；

2. 在 $k = N/3$ 处，Z-curve 映射的功率谱值是 Voss 映射功率谱值的 4 倍，Z-curve 映射总功率谱的平均值是 Voss 映射总功率谱均值的 3 倍，进而 Z-curve 映射下的信噪比是 Voss 映射信噪比的 4/3 倍；

3. 对 DFT 算法，Z-curve 映射和 Voss 映射相比，降低了数据的维数，使得数据处理的复杂度减小。

2.5 实数映射下的通用快速谱算法

本小节讨论一维映射的情况。将 A、C、G、T 四个碱基映射为实数值 0, 1, 2, 3。假设给定的一段 *DNA* 序列片段为 S = ATCGTACTG，则映射后的一维序列是 {031230132}。下面在该映射下探求功率谱与信噪比的关系。

长为 N 的 DNA 序列映射为一维序列 $u(n)$，若 N 为 3 的倍数，$u(n)$ 序列在 $k = N/3$ 处的功率谱值为：

$$
\begin{aligned}
P\left(\frac{N}{3}\right) &= \left|U\left(\frac{N}{3}\right)\right|^2 = \left|\sum_{n=0}^{N-1} u(n)\cdot e^{-j\frac{2\pi}{3}n}\right|^2 \\
&= \left|\sum_{m=0}^{N/3-1}\left(u(3m)+u(3m+1)e^{-j\frac{2\pi}{3}}+u(3m+2)e^{-j\frac{4\pi}{3}}\right)\right|^2 \\
&= \left|\left(x_C+2x_G+3x_T\right)+\left(y_C+2y_G+3y_T\right)e^{-j\frac{2\pi}{3}}+\left(z_C+2z_G+3z_T\right)e^{j\frac{2\pi}{3}}\right|^2 \\
&= \mathbf{XAX}^T
\end{aligned}
\qquad (2\text{-}19)
$$

其中 $X=(x_C+2x_G+3x_T, y_C+2y_G+3y_T, z_C+2z_G+3z_T)$。

根据帕斯瓦尔定理可得功率谱的平均值，则序列的信噪比为：

$$\begin{aligned} R&=\frac{P(N/3)}{\bar{E}}=\frac{\mathbf{XAX}^T}{\sum_{n=0}^{N-1}|u(n)|^2} \\ &=\frac{\mathbf{XAX}^T}{(x_C+y_C+z_C)+4(x_G+y_G+z_G)+9(x_T+y_T+z_T)} \end{aligned} \tag{2-20}$$

公式(2-19)除以公式(2-20)便可得 DNA 序列的信噪比。该映射下对人类线粒体基因 NC_012920_1 信噪比分析的结果见图 2-5，与题目所给曲线十分吻合，证明了实数映射下快速算法的正确性。

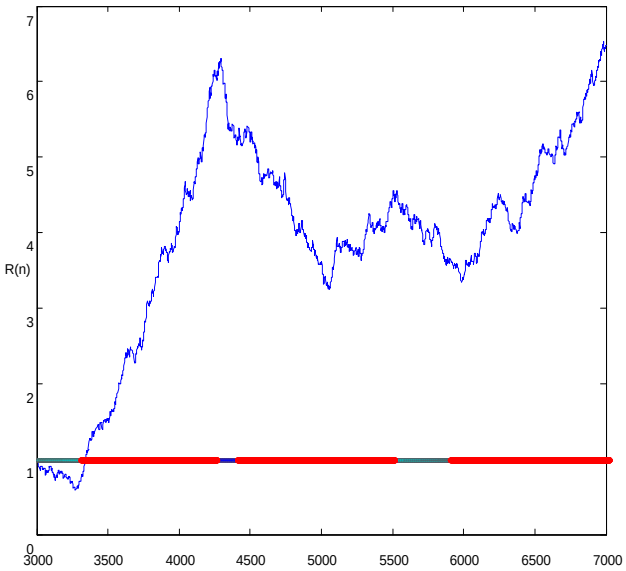


图2-5　实数映射下信噪比（人类线粒体基因NC_012920_1）

从这两个公式可以看出，实数映射的功率谱和信噪比求解最终转化为四种碱基出现频次的关系，即虽然其映射为一维映射，实际求解中仍是个四维问题。

下面给出实数映射下求解功率谱和信噪比的一般公式。假设将 A、C、G、T 四个碱基映射为实数值 a,c,g,t，碱基 $b\in I=\{A,T,G,C\}$映射后对应的数值为 $r_b, r_b\in\{a,c,g,t\}$。则长为 N 的 DNA 序列（N 为 3 的倍数）在 $k=N/3$ 处的功率谱值和信噪比值分别见公式(2-21)和公式(2-22)：

$$P\left(\frac{N}{3}\right)=\mathbf{XAX}^T \tag{2-21}$$

其中 $X=(ax_A+cx_C+gx_G+tx_T, ay_A+cy_C+gy_G+ty_T, az_A+cz_C+gz_G+tz_T)$。

$$R=\frac{\mathbf{XAX}^T}{\sum_{b\in I}(x_b+y_b+z_b)\cdot r_b^{\,2}} \tag{2-22}$$

3 基于模糊逻辑的自适应基因阈值确定

本章主要解决问题（2）。对基因识别来说，阈值是一个非常重要的参数。基因预测的最终目的是从 DNA 序列检测出外显子部分，因此对前一章节对 DNA 序列进行数值映射和谱分析都是为了能够更准确地从 DNA 序列中将外显子分离出来。进行外显子预测时，需要设定一定的阈值。对于不同的基因类型，所选取的判别阈值也许应该是不同的;对同一基因的不同片段,阈值或许也不完全相等。

有些 DNA 序列 3-周期特性不明显[9]，在仿真的过程中，甚至发现了部分具有“假 3-周期”性的 DNA 序列，即内含子对应的频谱也会出现谱峰。因此，本章将选取具有 3-周期特性的基因，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。

3.1 问题分析与符号说明

根据论文和题目的连续性，基因阈值确定方法将继续基于频谱 3-周期性展开。

不同的基因类型阈值不同，而对于同一基因的不同外显子片段的确定，其阈值或许也不尽相同,因此对基因不同的片段考虑设置不同的阈值,即自适应阈值。将模糊逻辑的方法引入自适应阈值，本文将建立基于模糊逻辑的自适应阈值模型。

表 3-1 符号约定与含义

符号	含义	符号	含义
R_0	阈值	A_c	基因预测的正确率
R^{opt}	最优阈值	F_s	模糊集合
$R_{\min}$	阈值搜索范围下界	F_r	模糊逻辑规则
$R_{\max}$	阈值搜索范围上界		

3.2 预备模型：已注释基因的最优固定阈值分析

无论采取何种阈值确定方法，得到的结果都是经验性的结论，都需要使用已知样本（即已知外显子位置的基因）进行训练。在进行样本训练时，目标是在一个阈值范围内找出最适合训练样本的阈值,本文采取一种最优阈值分析法确定样本训练终止条件。

根据定义可知，当阈值很低时，大量内含子容易被虚判成外显子，而当阈值过高时，大量外显子又会被误判为内含子，这两种情况下基因预测的正确率都不会太高。定义样本训练中**最优阈值**为使基因检测达到最高正确率的阈值，并用符号 R^{opt} 表示。

这时，如何合理定义正确率的问题出现了。从最优阈值的定义来看，不同正确率的定义会影响到最优阈值的确定。已有研究中用到的正确率定义多用于分析基因预测的结果，并不适用于样本训练。例如使用敏感性和特异性两个指标作为评估标准时，在最佳的阈值下基因预测的上述两指标值肯定是相对较高的，但仅其中某一个指标达到最高值并不能表示对应的阈值是最佳的。

令 T_{ex} 表示被正确预测为外显子的碱基个数，L_{ex} 表示真实序列外显子上的碱基个数，$\tilde{L}_{ex}$ 表示预测序列外显子上的碱基个数。本文给出了一种简便但十分合理的正确率 A_c 的定义方法，如公式(3-1)所示：

$$A_c = \frac{T_{ex}}{\max\left(L_{ex}, \tilde{L}_{ex}\right)} \tag{3-1}$$

对题目给出的 300 个已知基因，其中任意一个都可用最优阈值分析法确定一个最优阈值，算法总体结构采用一次循环即可，步骤如下：

（1）计算样本基因序列的 SNR；

（2）选择阈值范围 $[R_{\min}, R_{\max}]$，确定搜索步长 $STEP$；

（3）for i= $R_{\min} : STEP : R_{\max}$

计算 $A_c(i)$

end

（4）选取 $\max(A_c)$ 的阈值，定为 R^{opt}。

需要注意的是，采用最优固定阈值分析法可为每个已注释序列确定一个最优的固定阈值，但此阈值并不具有普适性，对于一个未注释基因，还是无法确定应采用的阈值。本小节的最优固定阈值分析法是为样本训练提出来的，它是未注释基因阈值确定的准备工作，对未注释基因阈值确定方法将在 3.3 节具体分析。

根据阈值分析法，图 3-1 给出了题目数据 100 个人和鼠基因的第 14 号基因 AF019045 的正确率随阈值变化的曲线，曲线的峰值就是我们为这一基因选定的最优阈值。

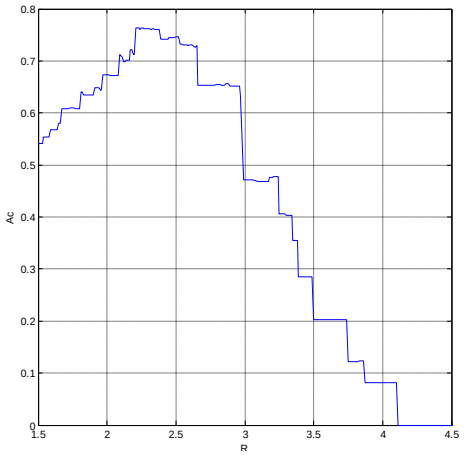


图 3-1 基因 AF019045 正确率随阈值变化曲线

3.3 基于模糊逻辑的自适应阈值模型

3.3.1 模糊逻辑概述

模糊逻辑[10]指模仿人脑的不确定性概念判断、推理思维方式，对于模型未知或不能确定的描述系统，以及强非线性、大滞后的控制对象，应用模糊集合和

模糊规则进行推理，表达过渡性界限或定性知识经验，模拟人脑方式，实行模糊综合判断，推理解决常规方法难于对付的规则型模糊信息问题。模糊逻辑善于表达界限不清晰的定性知识与经验，它借助于隶属度函数概念，区分模糊集合，处理模糊关系，模拟人脑实施规则型推理，解决因“排中律”的逻辑破缺产生的种种不确定问题，因此非常适用于自适应阈值的求解。

3.3.2 模型参数优化

图 3-2 给出了基于频谱 3-周期性的的基因预测方法流程图，图中的每个步骤的处理都会影响基因序列的信噪比，而阈值就是根据信噪比的特征确定的，因此本章节逐一对每个步骤中的参数进行优化，致力于减小每个步骤对阈值确定造成的误差。

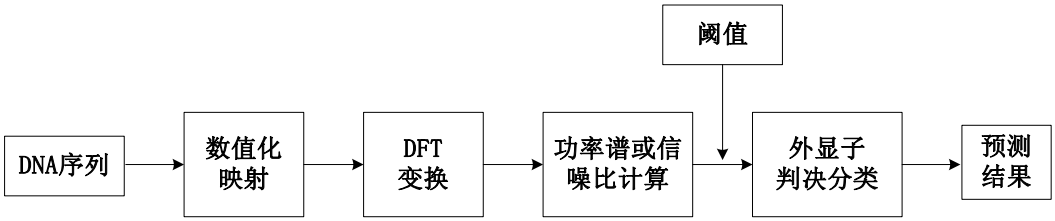


图 3-2 基于序列频谱 3-周期性的的基因预测方法流程图

基于模糊逻辑的自适应阈值搜索算法需要首先将基因序列划分为不同的段，利用模糊集合序列信噪比的均值、方差、峰值等特征分为低、中、高三档，并制定模糊逻辑规则。采用“训练——测试”的模式确定最优自适应门限，首先选定参数，然后对训练样本（已注释的 DNA 序列）进行训练，形成训练集，最后对测试集进行测试。

“训练——测试”模式的参数包括：

（1）段长：每段序列不能太短，必须保证能包含足够的有效信息来确定阈值的范围，又不能太大，以便在序列的不同位置上确定不同的阈值。

（2）窗尺寸：将基因序列分段后，信噪比序列长度相应减小，信噪比曲线中的毛刺现象更加突出，窗尺寸对阈值的判断影响增大，需要根据准确率的变化选择适当的窗尺寸。

（3）窗类型：文献[11]采用了矩形窗对 DNA 序列的功率谱进行处理，在数字信号处理中，矩形窗和巴特利特窗的比较表明后者可使信噪比曲线更加平滑并可去除添加矩形窗后带来的额外波峰。

实际进行基因序列的谱分析时，无论是基于固定窗口还是滑动窗口，都等同于把 DNA 的数值序列限制在一定的长度内。这样，取用有限个数据，即将数据截断的过程，就等于将信号进行加窗函数操作$u_b(n)\cdot w(n), n=0,1,...,M-1$。而这样操作以后，常常会发生频谱分量从其正常频谱扩展开来的现象，即所谓的“频谱泄漏”，表现在 DNA 序列上就是编码区的频谱泄漏到了非编码区，降低了基因预测的准确性。当进行离散傅立叶变换时，时域中的截断是必需的，因此泄漏效应也是离散傅立叶变换所固有的，必须进行抑制。由此可见，窗函数选取在谱分析中占有的重要地位。

目前窗口内的数据都是直接取出的，相当于对 DNA 数值序列加了矩形窗，而矩形窗导致很严重的频谱泄露，泄露的频谱衰减很慢，因此本文采用巴特利特窗进行修正。

矩形窗函数为：

$$w_{\text{rectangular}}=\begin{cases}1, 0\le n\le M-1\\ 0, \text{其他}\end{cases} \tag{3-2}$$

巴特利特窗函数为：

当 n 为偶数时

$$w_{\text{Bartlett}}=\begin{cases}\dfrac{2n}{M}, 0\le n\le \dfrac{M}{2}\\ 2-\dfrac{2n}{M}, \dfrac{M}{2}<n\le M-1\end{cases} \tag{3-3}$$

当 n 为奇数时

$$w_{\text{Bartlett}}=\begin{cases}\dfrac{2n}{N}, 0\le n\le \dfrac{M-1}{2}\\ 2-\dfrac{2(M-n)}{M}, \dfrac{M+1}{2}<n\le M-1\end{cases} \tag{4-3}$$

采用矩形窗和巴特利特窗的 DNA 数值序的信噪比曲线如图 3-3 和 3-4 所示，可以看出引入巴特利特窗可以有效去除矩形窗引起的泄漏的频谱。

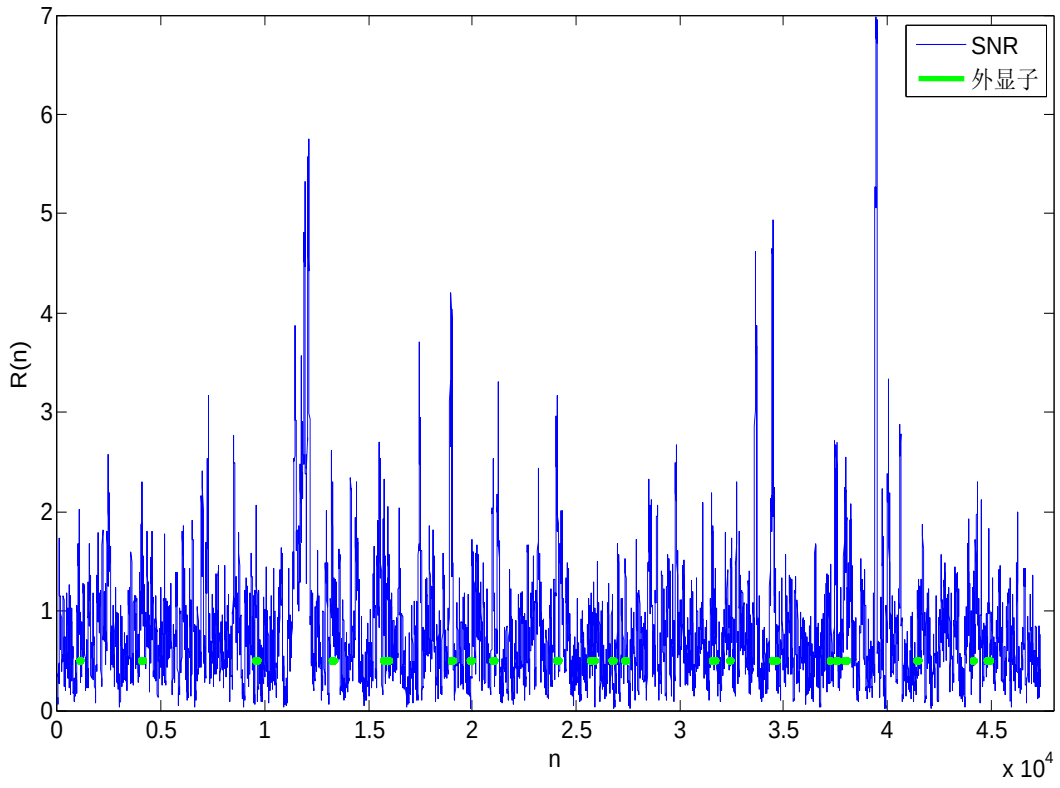


图 3-3 加矩形窗信噪比曲线（39 号基因）

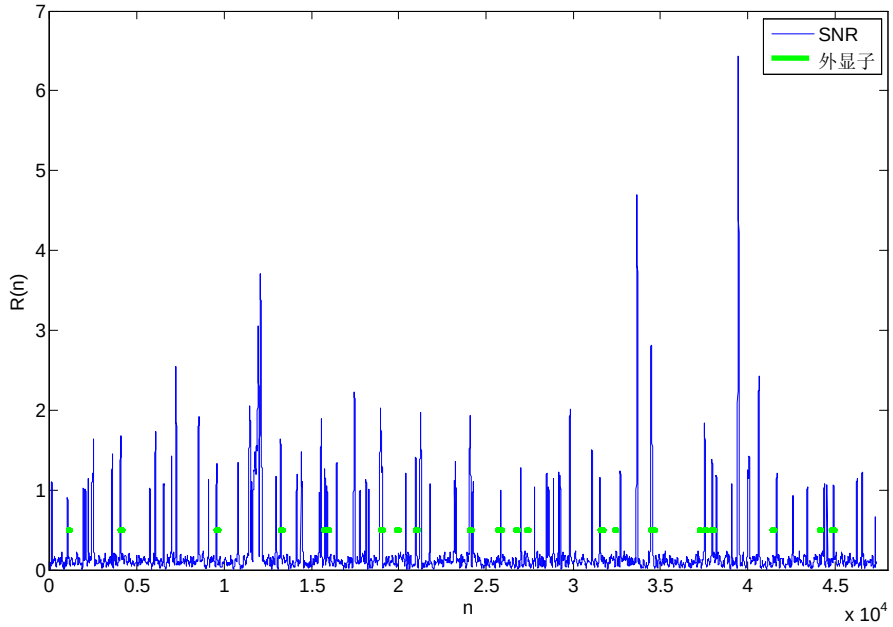


图 3-4 加巴特利特窗的信噪比曲线（39 号基因）

（4）外显子窗口扩展：经过阈值法选出的外显子实际为原滑动窗口的中点，若此点的 $P(N/3)$ 值大于阈值，则表征此时窗口内的基因段为外显子段，因此需要将阈值选出的端点向两边扩展，扩展窗口的大小就是之前加窗的尺寸。

（5）模糊集合：不同特征（均值、方差、峰值）的模糊集合 F_s 分为低、中、高三类，隶属度函数曲线如图 3-5 所示。

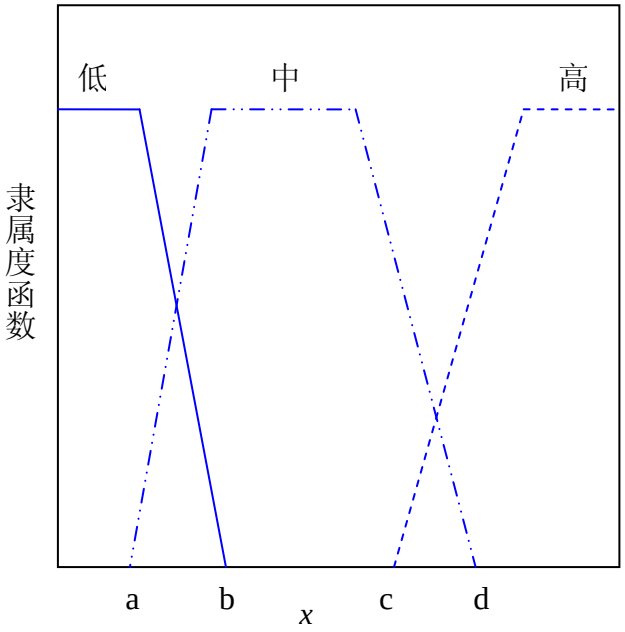


图 3-5　隶属度函数曲线

特征函数属于低、中、高三个模糊集合的隶属度函数为：

$$l_f(x)=\begin{cases}1, x\le a\\ \dfrac{b-x}{b-a}, a<x\le b\\ 0, x>b\end{cases}$$

$$m_f(x)=\begin{cases}0, x\le a\\ \dfrac{x-a}{b-a}, a<x\le b\\ 1, b<x\le c\\ \dfrac{d-x}{d-c}, c<x\le d\\ 0, x>d\end{cases}$$

$$h_f(x)=\begin{cases}0, x\le c\\ \dfrac{x-d}{d-c}, c<x\le d\\ 1, x>d\end{cases}$$

隶属度函数表示样本的某个特征属于低、中、高三档的概率，根据模糊判定规则，具有相同特征值的样本不一定隶属于同一档，判定时按照特征值确定的概率随机归档。

（6）模糊逻辑规则：在阈值范围$[R_{\min}, R_{\max}]$上的模糊逻辑规则F_r如表 3-2 所示。

表 3-2 确定自适应阈值范围的模糊逻辑规则

序号	均值	方差	峰值	$\mathbf{R_{min}}$	$\mathbf{R_{max}}$
1	低	低	低	$\bar{r}$	$\bar{r}+4\sigma$
2	低	低	中	$\bar{r}-0.5\sigma$	$\bar{r}+3.5\sigma$
3	低	低	高	$\bar{r}-\sigma$	$\bar{r}+3\sigma$
4	低	中	低	$\bar{r}$	$\bar{r}+2\sigma$
5	低	中	中	$\bar{r}-0.25\sigma$	$\bar{r}+1.75\sigma$
6	低	中	高	$\bar{r}-0.5\sigma$	$\bar{r}+1.5\sigma$
7	低	高	低	$\bar{r}$	$\bar{r}+\sigma$
8	低	高	中	$\bar{r}-0.05\sigma$	$\bar{r}+0.95\sigma$
9	低	高	高	$\bar{r}-0.1\sigma$	$\bar{r}+0.9\sigma$
序号	均值	方差	峰值	$\mathbf{R_{min}}$	$\mathbf{R_{max}}$
10	中	低	低	$\bar{r}-1.5\sigma$	$\bar{r}+2.5\sigma$
11	中	低	中	$\bar{r}-2\sigma$	$\bar{r}+2\sigma$
12	中	低	高	$\bar{r}-2.5\sigma$	$\bar{r}+1.5\sigma$
13	中	中	低	$\bar{r}-0.75\sigma$	$\bar{r}+1.25\sigma$
14	中	中	中	$\bar{r}-\sigma$	$\bar{r}+\sigma$
15	中	中	高	$\bar{r}-1.25\sigma$	$\bar{r}+0.75\sigma$
16	中	高	低	$\bar{r}-0.45\sigma$	$\bar{r}+0.55\sigma$

17	中	高	中	$\bar{r}-0.5\sigma$	$\bar{r}+0.5\sigma$
18	中	高	高	$\bar{r}-0.55\sigma$	$\bar{r}+0.45\sigma$

序号	均值	方差	峰值	$\mathbf{R_{min}}$	$\mathbf{R_{max}}$
19	高	低	低	$\bar{r}-3\sigma$	$\bar{r}+\sigma$
20	高	低	中	$\bar{r}-3.5\sigma$	$\bar{r}+0.5\sigma$
21	高	低	高	$\bar{r}-4\sigma$	$\bar{r}$
22	高	中	低	$\bar{r}-1.5\sigma$	$\bar{r}+0.5\sigma$
23	高	中	中	$\bar{r}-1.75\sigma$	$\bar{r}+0.25\sigma$
24	高	中	高	$\bar{r}-2\sigma$	$\bar{r}$
25	高	高	低	$\bar{r}-0.9\sigma$	$\bar{r}+0.1\sigma$
26	高	高	中	$\bar{r}-0.95\sigma$	$\bar{r}+0.05\sigma$
27	高	高	高	$\bar{r}-\sigma$	$\bar{r}$

3.3.3 模型建立

上一小节确定模型参数后开始训练：

输入：DNA 序列，分段长度 C，窗口尺寸 N，窗的类型，外显子窗口扩展，阈值步长，模糊集合 F_s，模糊逻辑规则 F_r。

输出：

（1）将 DNA 序列按长度 C 分段。

（2）对每个分段 C_R

1)根据第 2 章的方法计算分组序列在 $k=N/3$ 处的信噪比；

2)计算信噪比的均值、方差及峰值等特征值；

3)依据 F_s 将信噪比的特征值分为低、中、高三个模糊集合；

4)通过模糊逻辑 F_r 规则确定阈值范围 $[R_{\min},R_{\max}]$；

5)根据上一章节的最优阈值分析法在 $[R_{\min},R_{\max}]$ 上求解最优阈值 R^{opt}；

6)形成 C_R 与 R^{opt} 对；

7)将 $\langle C_R,R^{opt}\rangle$ 加入训练集中。

（3）将训练集内按照 R^{opt} 的大小升序排列。

训练结束后形成训练集，接下来进行测试：

输入：未注释的基因序列，训练集 DB，分段长度 C，窗口尺寸 N，窗的类型，外显子窗口扩展，阈值步长，模糊集合 F_s，模糊逻辑规则 F_r。

输出：将基因序列识别为外显子和内含子。

（1）将 DNA 序列按长度 C 分段。

（2）对每个分段 C_R

1)根据第 2 章的方法计算分组序列在 $k=N/3$ 处的信噪比；

2)计算信噪比的均值、方差及峰值等特征值；

3)依据 F_s 将信噪比的特征值分为低、中、高三个模糊集合；

4)通过模糊逻辑 F_r 规则确定阈值范围 $[R_{\min}, R_{\max}]$；

5)根据阈值范围 $[R_{\min}, R_{\max}]$ 在波形样本集中搜索最相似的波形及相应的阈值 $\langle C_R, R^{opt} \rangle$；

6)将 R^{opt} 定为该段基因序列的最优阈值。

（3）综合所有分段的最优阈值形成自适应阈值，将 SNR 大于阈值的位置将其判为外显子，并进行端点扩展。

3.4 评价指标

在信号检测理论中，ROC 曲线是一种对于灵敏度进行描述的功能图像[12]。ROC 曲线可以通过描述击中率和虚惊率来实现。ROC 曲线首先是由二战中的电子工程师和雷达工程师发明的，他们用来检测战场中的敌军，也就是信号检测理论。之后很快就被引入了心理学来进行信号的知觉检测。ROC 分析现在已经在相关的领域得到了很好的发展，特别是在医学，无线电领域中，而且最近在机器学习和数据挖掘领域也得到了很好的发展。

ROC 曲线方法简单、直观，通过图示可观察分析方法的准确性，并可用肉眼作出判断。ROC 曲线将灵敏度与特异性以图示方法结合在一起，可准确反映某分析方法特异性和敏感性的关系，是试验准确性的综合代表。ROC 曲线不固定分类界值，允许中间状态存在，利于使用者结合专业知识，权衡漏报与虚报的影响，选择一更佳截断点作为诊断参考值。提供不同试验之间在共同标尺下的直观的比较，ROC 曲线越凸越近左上角表明其评价价值越大，利于不同指标间的比较。曲线下面积可用于评价诊断准确性。

因此，本文选取 ROC 曲线在碱基水平上对预测结果进行评估。如图 3-6 所示，定义 T_{ex} 表示被正确判为外显子的碱基个数，即正确的肯定；T_{in} 表示被正确判为内含子的碱基个数，即正确的否定；F_{ex} 表示被错误地判为外显子的碱基个数，即错误的肯定，表示假报警，是第一类错误；F_{in} 表示被错误地判为内含子的碱基个数，即错误的否定，表示未命中，是第二类错误。

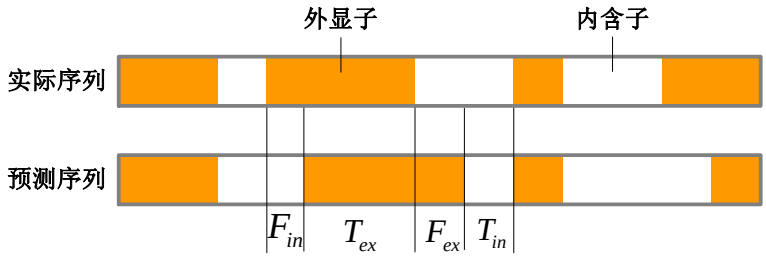


图 3-6 碱基水平上的参数表示

ROC 曲线引入敏感性和特异性两个指标分析外显子的预测准确率：

敏感性 S_n：表示实际编码区的碱基个数被正确预测为外显子碱基的比例：

$$S_n = \frac{T_{ex}}{T_{ex} + F_{in}}$$

特异性 S_p：表示被预测为外显子的碱基中真实来自外显子的比例：

$$S_p = \frac{T_{ex}}{T_{ex} + F_{ex}}$$

用敏感性和特异性两个指标评价基因预测准确性，这种方法充分考虑到 DNA 序列中的每一个碱基，在碱基水平上将预测结果和序列中外显子的实际位置一一进行比较，既要使全部外显子碱基都判出来，又要使判出来的外显子尽可能正确，因此可以精确地评估预测结果。

在文章前几节求解最优阈值时采用本文自定义的准确率 A_c 作为评价指标，而在本节评价预测时又选取一套新的敏感性和特异性评价指标，对此作如下解释。

说明

敏感性 S_n、特异性 S_p 和准确率 A_c 的关系：

分析三者的定义不难发现，准确率 A_c 其实就是敏感性和特异性二者的较小值，用其作为确定最优阈值的条件，既保证了使敏感性和特异性同时达到较高水平，又便于程序实现。而为了全面评价基因预测的结果，采用敏感性和特异性两个指标又更加直观。

3.5 结果分析与小结

本文从美国国家生物信息中心网站 NCBI 数据库[14]中下载了 1000 个人类和鼠类的基因样本，对每个基因按每段长度 900 进行分段，通过模糊逻辑自适应阈值确定方法进行训练从而建立训练集，对题目给出的 100 个人和鼠类的样本作为测试集进行测试分析，自适应阈值结构见附录 1，固定阈值与自适应阈值的对比见附录 2。

结论

1. 表 3-3 列出了 10 个基因的自适应阈值，与最优固定阈值分析法得到的固定阈值进行了对比，同时也给出了两种阈值的敏感性和特异性指标值，可以看出自适应阈值的两个指标普遍高于固定阈值。最优固定阈值分析法得到的是使预测正确率最高的阈值，其他任何一个固定阈值预测出来的正确率一定低于该正确率，表 3-3 表明最好的固定阈值预测效果大部分都比本文求解的自适应阈值效果差，说明自适应阈值效果好于固定阈值。

2. DNA 序列频谱的 3-周期特性是碱基分布不平衡造成的，是一个统计学特征，并不是所有基因编码碱基分布都不平衡，也并不是所有非编码碱基分布都平滑，进而并不是只有编码区表现出 3-周期特性，文献[13]证实了这一猜想，本文的分析也验证了这一点，分析表 3-3 可以看出有部分基因无论是固定阈值分析还是自适应阈值分析，敏感性和特异性指标都不好，例如 41 号，图 3-4 给出了 41 号基因样本的信噪比曲线及已知的编码区位置，可以看出编码区具有 3-周期

特性，而在非编码区也有较高的 SNR 值，就扩大了编码区的选取从而造成预测正确率的下降。

3. 无论是固定阈值还是自适应阈值，其值都在 2 左右波动，这验证了题目中将人和鼠阈值设为 2 的合理性，也说明了本文阈值确定方法的正确性，但要提高预测准确性，必须进行自适应阈值分析。

表 3-3 测试基因的阈值

基因序号	固定阈值	固定阈值 S_n	固定阈值 S_p	自适应阈值	自适应阈值 S_n	自适应阈值 S_p
1	2.3	0.3732	0.3649	1.2494,1.6052,2.3285,2.0873,1.7768, 2.4342,1.4426,1.7833,2.1772	0.6342	0.6025
6	1.8	0.9127	08473	1.7218,1.1399,0.73111,1.7833,1.3969	0.9554	0.8965
10	1.9	0.7116	0.1871	2.0115,1.7681,2.1731,2.065,1.5509, 1.8046,1.182	0.3889	0.142
12	2.4	0.4486	0.2775	2.4849,1.8,1.9961,1.769	0.4486	0.4034
14	2.3	0.8638	0.7615	2.1703,0.15021,1.88,2.0947,1.8158, 2.5491,1.4207	0.9962	0.8469
35	1.8	0.8062	0.7685	1.7832,0.44012,1.75,1.4659	0.8899	0.8945
51	1.5	0.6111	0.8800	1.8029,1.2959	0.8577	0.8255
65	1.9	0.4482	0.4307	2.4478,1.9022,1.6923	0.3176	0.6912
74	2	0.7756	0.6334	1.9551,1.7666,1.96,1.4019,1.4951, 1.9101	0.7854	0.5493
92	1.6	0.5692	0.3593	1.342,2.0468,2.1108,1.421,1.199, 2.1958,1.2555,	0.9304	0.7188

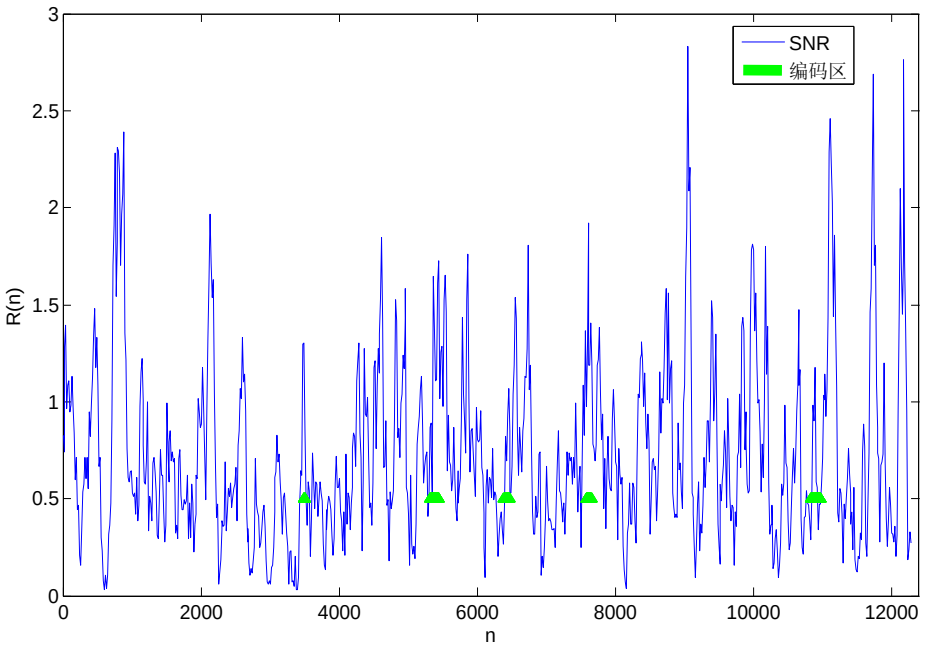


图 3-7 41 号基因的信噪比及编码区位置

图 3-8 至 3-11 给出了测试集 100 个基因中的第 84 号基因和 97 号基因固定阈值和自适应阈值的基因预测结果，其中 84 号基因的预测准确率由 0.4532 提高到了 0.7925，97 号基因的预测准确率由 0.6494 提高到了 0.9408，从图中也可以直观地看出与固定阈值比，自适应阈值确定的外显子位置与实际位置更相符，可见，对于某些基因，划定同一的阈值会严重影响其预测准确率，只有根据其波形特征，自适应地选择阈值，才能准确预测基因中的外显子。

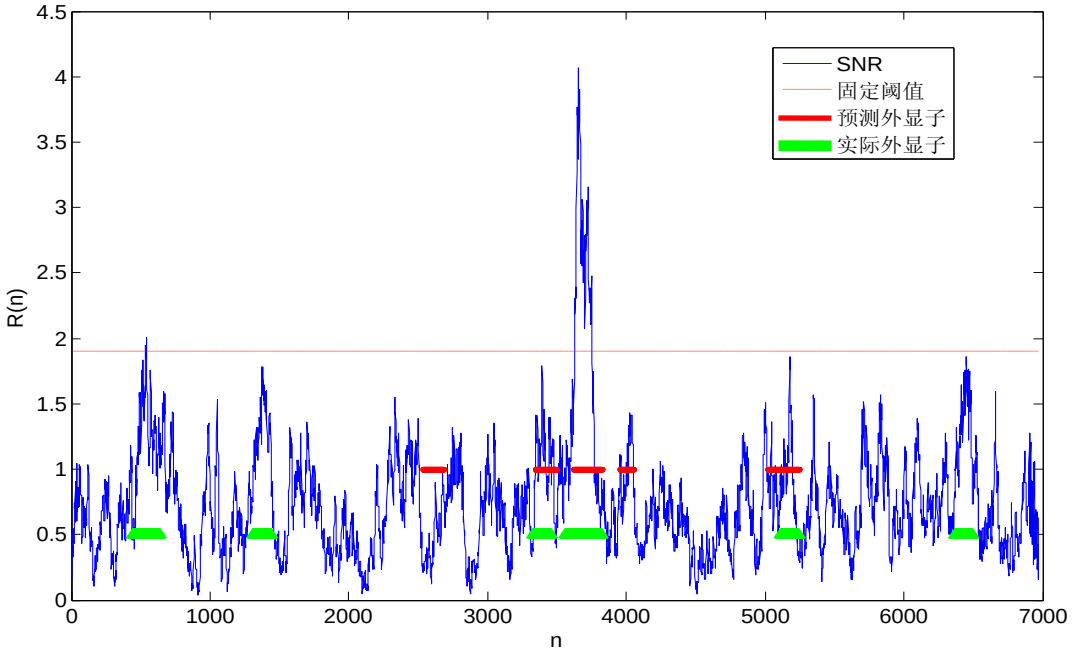


图 3-8 84 号基因固定阈值下的信噪比

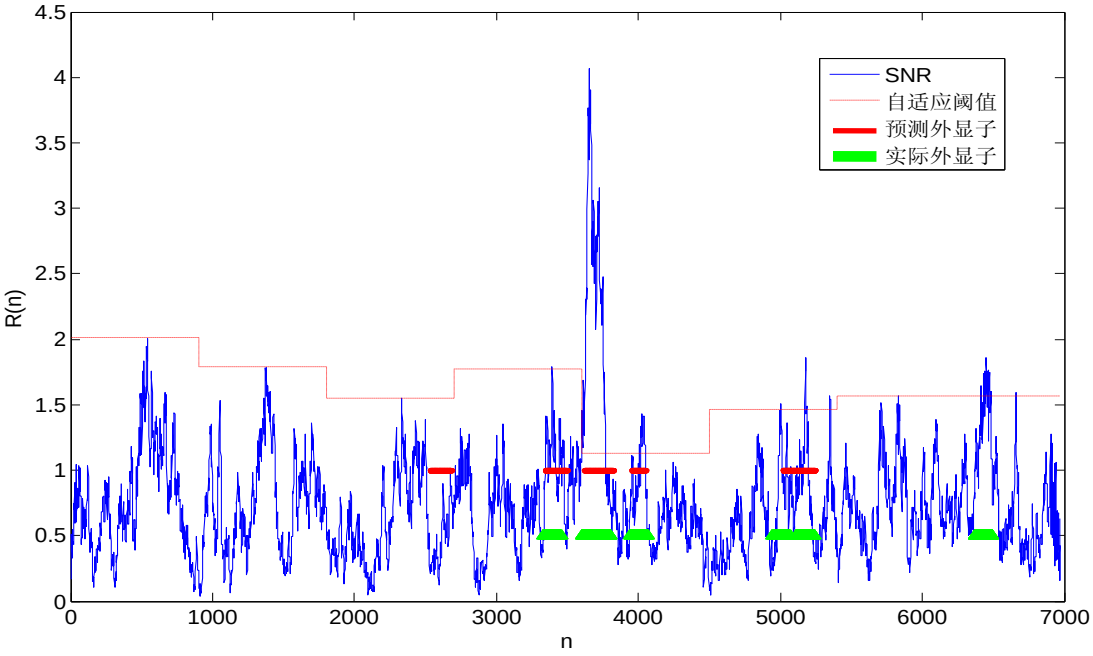


图 3-9 84 号基因自适应阈值下的信噪比

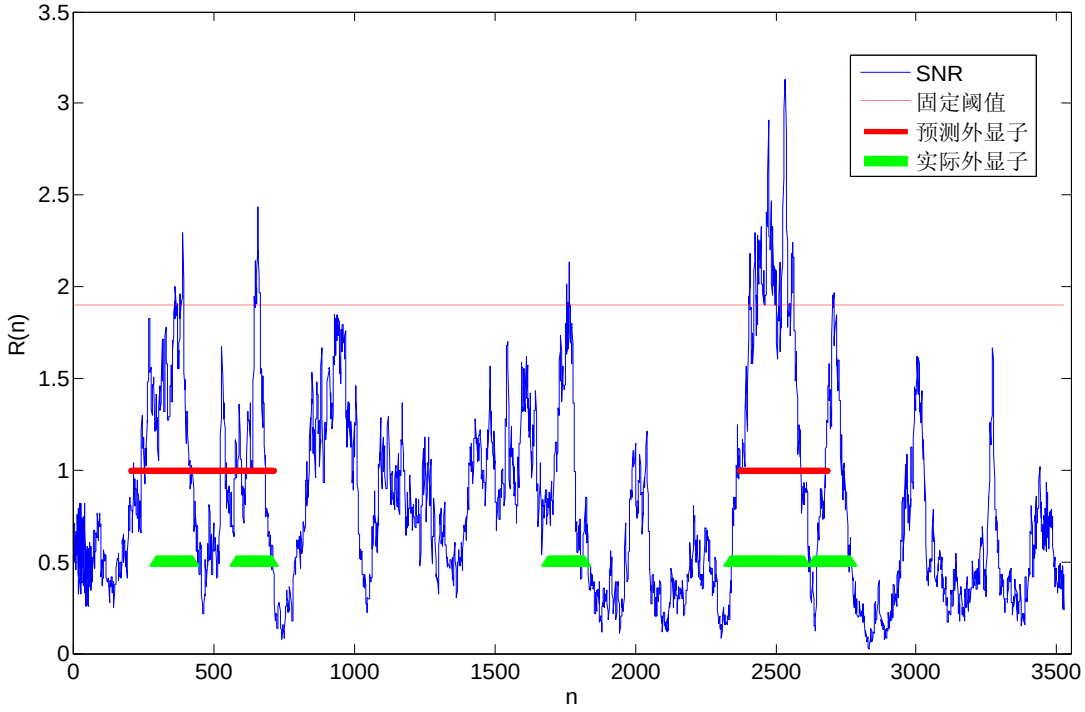


图 3-10 97 号基因固定阈值下的信噪比

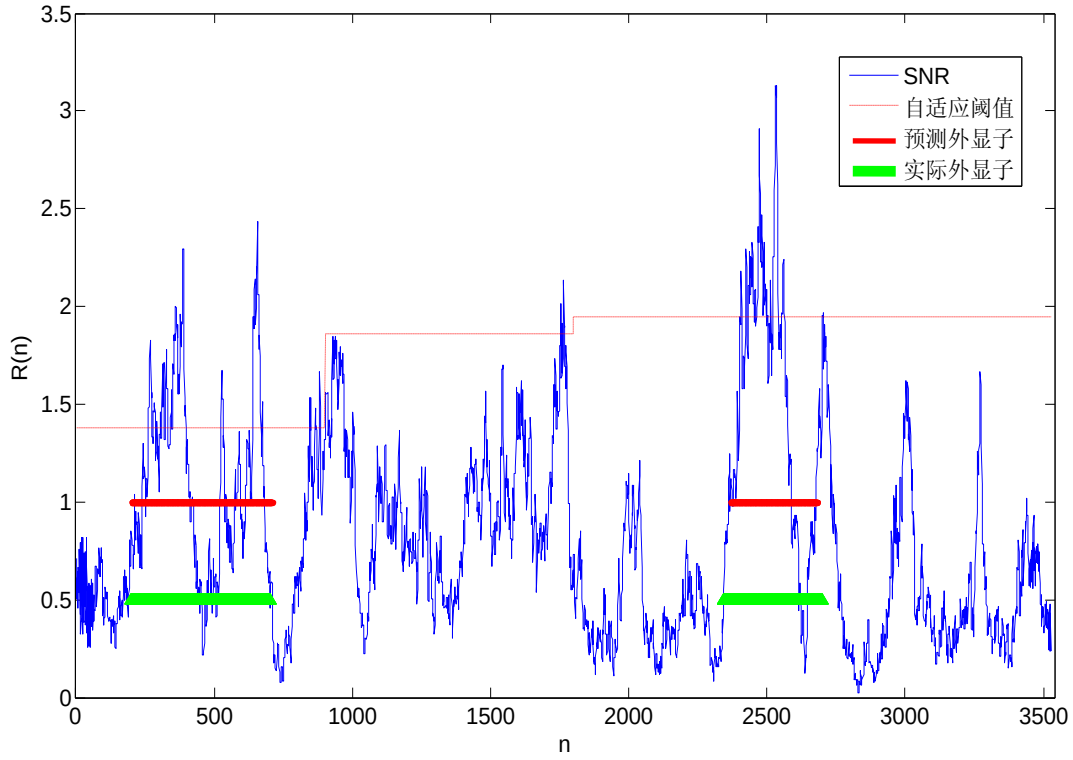


图 3-11 97 号基因自适应阈值下的信噪比

4 基于边界搜索的基因识别算法

本章主要解决问题（3）。频谱峰值特征的发现，或者频谱与信噪比概念的引

入，以及阈值的确定方法，其最终目的是要探测、预报一个尚未被注释的完整的 DNA 序列的所有基因编码序列（外显子）片段，即尽可能“精确地”确定基因外显子区间的两个端点。

4.1 基于小波变换的梯度边缘检测

基因外显子区间的精确确定，体现在功率谱上，即功率谱幅度的突变点或不连续点，即“边缘”，它往往是信号分析的信息所在。但由于噪声的影响，使得其边缘点的判定难度较大。小波变换是近年得到广泛应用的数学工具。与傅立叶变换和加窗傅立叶变换相比，小波变换是时间和频率的局域变换，因而能有效地从信号中提取信息，它通过伸缩和平移等运算功能对函数或信号进行多尺度细化分析，解决了傅立叶变换不能解决的很多困难问题，因而被誉为“数学显微镜”。信号边缘点检测及由边缘点重建原始信号是小波变换应用的一个很重要的方面。但考虑到小波分析方法复杂度较高，本文没有选用该方法，而选用复杂度较低的梯度法，并在理论上证明了函数的小波变换和其梯度方法在一定条件下等价[15]。

设 $\phi(x)$ 是一个在不同尺度下能对信号适当平滑的函数，并满足条件：

$$\int_{-\infty}^{\infty}\phi(x)dx=1 \quad 及 \lim_{t\to\infty}\phi(x)=0 \tag{5-1}$$

在应用中 $\phi(x)$ 常选取 Gaussian 函数。

定义 $\varphi(x)=\dfrac{d\phi(x)}{dx}$，由于 $\int_{-\infty}^{\infty}\varphi(x)dx=0$，函数 $\varphi(x)$ 是一个小波。

小波函数基定义为：

$$\varphi_j(x)=2^{-j/2}\varphi(2^{-j}x) \tag{5-2}$$

对于任何函数，小波变换表示为 $T_j=f(x)*\varphi_j(x)$，则：

$$T_j=f*\left[2^{-j/2}\frac{d\phi_j}{dx}\right](x)=2^{-j/2}\frac{d}{dx}\left[f*\phi_j\right](x) \tag{5-3}$$

由公式（5-3）可以得出，用 $\phi_j(x)$ 对 $f(x)$ 平滑后的函数的微分对应于 $f(x)$ 的小波变换。所以，用梯度分析的方法可以有效的分析功率谱的边缘点，且复杂度较低。图像处理领域中，往往可以利用边缘曲线变化的一阶或二阶导数特点，将边缘点检测出来。梯度法是在边缘模糊区域对每个微元处的梯度进行分析计算来检测边缘点，梯度的大小代表边缘的强度，通过选取最大梯度值来判断边缘点。

但当功率谱受到随机噪声干扰时，由于差分运算对噪声非常敏感的固有特性，使得常常把噪声也当作边缘点检测出来，而真正的边缘由于受到干扰却没有检测到，所以其应用受到某些限制。因此，可以先对功率谱进行平滑，然后再用梯度算法进行检测。

对曲线的平滑方法一类是采用 Gaussian 型滤波器对功率谱图进行卷积运算 $P(x)=P(x)*\phi(x)$，其中 $\phi(x)=\dfrac{1}{2\pi\sigma^2}e^{-\frac{x^2}{2\sigma^2}}$，此方法可在一定程度上抑制噪声的影响。另一类是用一个平滑的曲线对原图像数据点进行拟合，有效的去除了噪声的

影响。在边缘模糊区间内（L 为区间长度）对曲线进行多项式拟合 $P_n(x)=\sum_{k=0}^{n}a_k x^k$，求 a_k 的估计值，使得

$$\min\ E\{a_0,a_1,\ldots a_n\}=\sum[f(x)-P_n(x)]^2 \qquad (5\text{-}4)$$

经过平滑滤波抑制噪声的影响而使信号的边缘效应得到凸显，此时再利用梯度法进行边缘检测，可以得到很高的边缘定位精度。

对 27 号基因外显子端点的模糊区域进行边缘检测，结果如图 4-1 所示。在该段上预测的外显子起始点在 771 位，而实际外显子起始点在 802 位，预测的早于实际的 31 位，采用边缘检测算法后最终确定的边界为 817 位，提高了端点的定位精度。

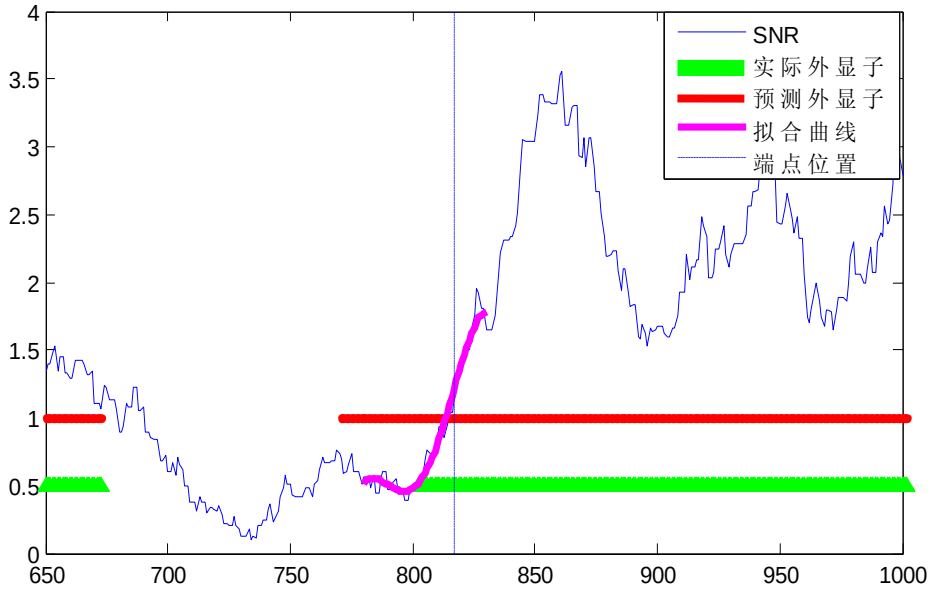
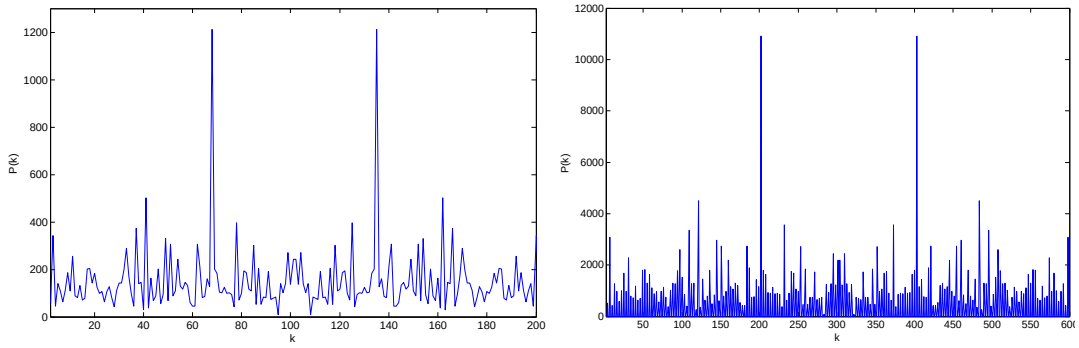


图 4-1 基于噪声抑制的边缘检测

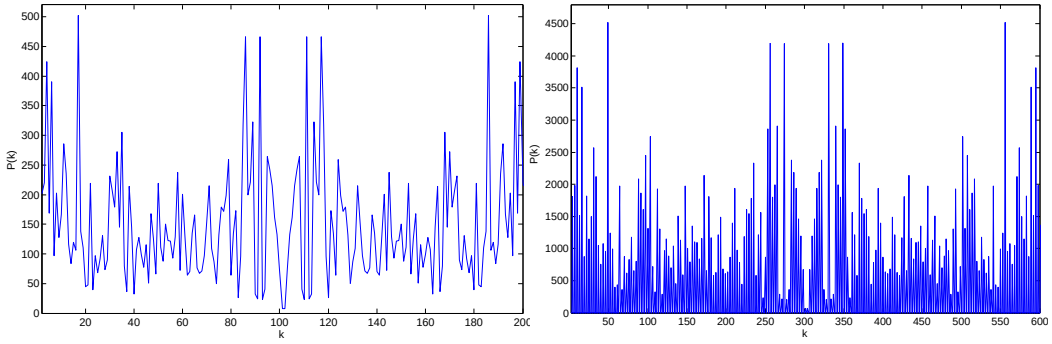
4.2 基于序列重复的边界搜索算法

在 100 个人和鼠的基因样本中取 3 号基因，选取基因上的一个外显子序列和一个内含子序列，将序列分别重复 3 遍后的功率谱与重复前的功率谱对比如图 4-2 和图 4-3 所示。可以看出，经过重复外显子 3-周期特性进一步增强，而内含子仍然不具有这一特性。受这一现象启发，将外显子端点处的小段序列多次重复后进行谱分析，可使外显子的 3-周期特性更加显著，从而更精确地确定外显子端点的位置。以此为基础，提出了一种基于序列重复的边界搜索算法。



(a)重复前　　(b)重复后

图 4-2 外显子重复前后功率谱对比



(a)重复前　　(b)重复后

图 4-3 内含子重复前后功率谱对比

边界搜索算法的基本思想是通过序列重复放大外显子的 3-周期特性，算法以初步预测的外显子端点为中心形成搜索序列，对搜索序列中的每个值都确立一个求 R 序列，对该求 R 序列重复扩展后作谱分析。

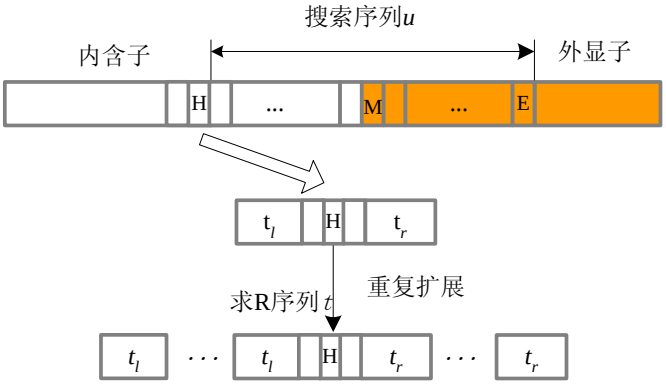


图 4-4 边界搜索算法的序列

边界搜索算法流程描述如下：

（1）确定搜索序列。以外显子的边界碱基 M 为中点分别向两边扩展 $(N-1)/2$ 位，形成长为 N 的搜索序列 $u(n), n=0,1,...,N-1$。序列起始碱基称为 H，对应序列中的 $u(0)$，末尾碱基称为 E，对应序列中的 $u(N-1)$。

（2）确定求 R 序列。对搜索序列中的每个 $n, n=0,1,...,N-1$ 加窗，以 n 为中心向两边扩展 $3l+1$ 位，形成长为 $L, L=6l+3$ 的求 R 序列 $t(n), n=0,1,...,L$

（3）重复扩展。将求 R 序列左右各重复扩展 k 次，在扩展后序列上求解 $m/3$ 处的信噪比 $r(n)$。

（4）对 $n=0,1,...,N-1$， $r(n)$ 形成搜索序列的信噪比曲线。

（5）对搜索序列的信噪比曲线作阈值分析，形成新的外显子端点。

在边界搜索算法中，选取 N 为 100，L 为 33，重复扩展次数 k 为 3 进行仿真分析。27 号基因的信噪比曲线如图 4-5 所示，在该段上预测的外显子起始点在 771 位，而实际外显子起始点在 802 位，预测的早于实际的 31 位，采用边界搜索算法对进一步精确预测结果，搜索序列的信噪比曲线如图 4-6 所示，经阈值判断后最终确定的边界为 794 位，大大提高了端点的定位精度。

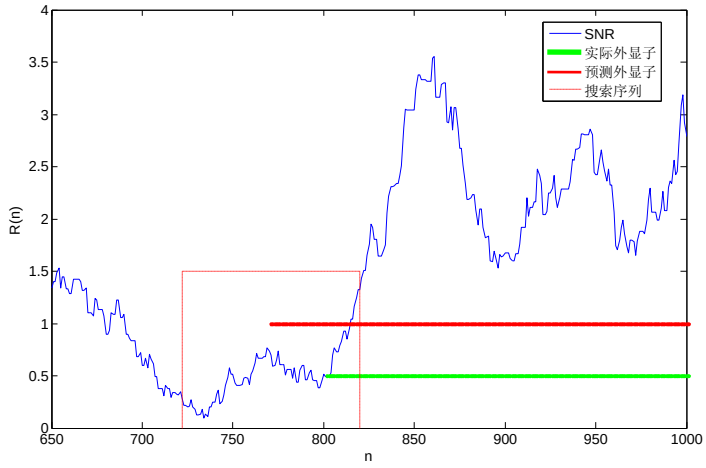


图 4-5 初步预测的基因片段

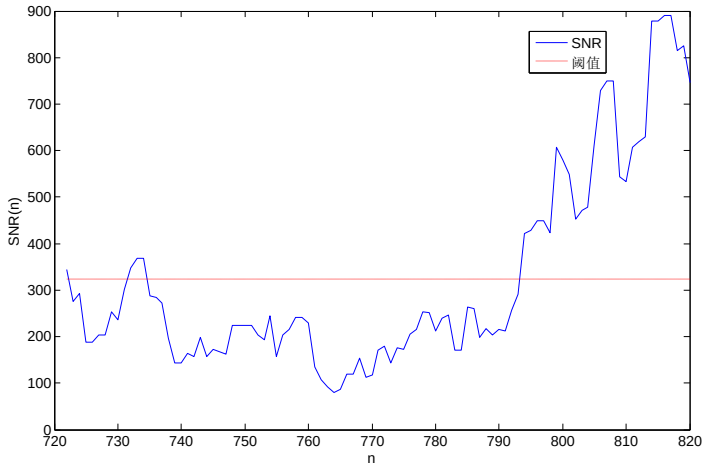


图 4-6 搜索序列的信噪比曲线

4.3 对未注释基因的预测

结合第三章建立的自适应阈值算法和本章的外显子端点确定方法，对题目给出的 6 个未注释基因进行预测，预测出的 6 个基因外显子位置如表 4-1 所示，图

4-7~图 4-12 给出了 6 个基因的信噪比曲线及自适应阈值。

表 4-1　6 个未注释基因的预测编码区

基因序号	编码区位置
1 号	1087…1570,　1704…1749,2401…2492,2716…2881,3114…3159,　3208…3391, 3678…3832, 4372…4417, 4501…4676, 4821…4866, 5001…5175
2 号	1276…1393,3412…3560,3901…4022,4098…4333,5591…5685, 5766…5944,7526…7571
3 号	1310…1403,1542…1801,1956…2001,2515…2686,2781…2960, 3025…3290,3615…3723
4 号	1528…1645,2608…2769,2824…2989,3050…3511,4405…4500, 4658…4913,5015…5269,5477…5667
5 号	725…770,2931…3377,5737…5782,7001…7118,7901…8322, 9773…9890,10348…10393,13548…13975
6 号	145…190,322…394,1201…1283,1367…1462,2414…2751,4626…4695

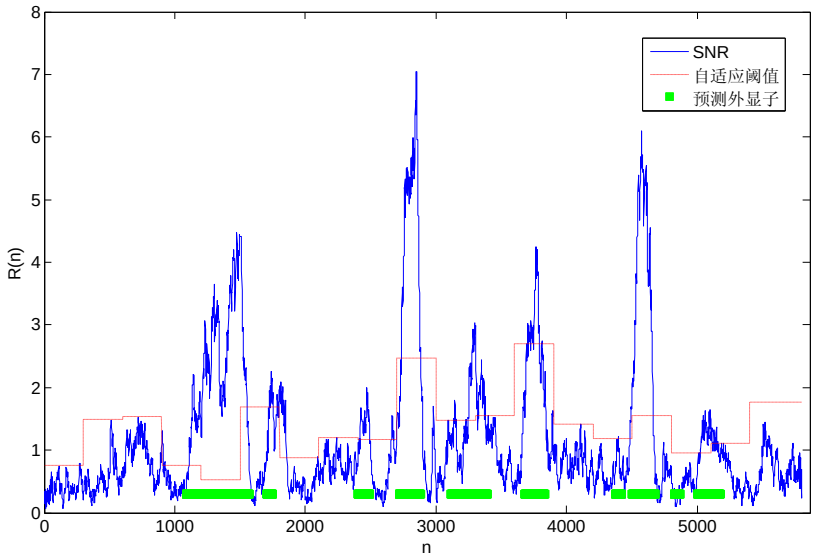


图 4-7　1 号基因的信噪比

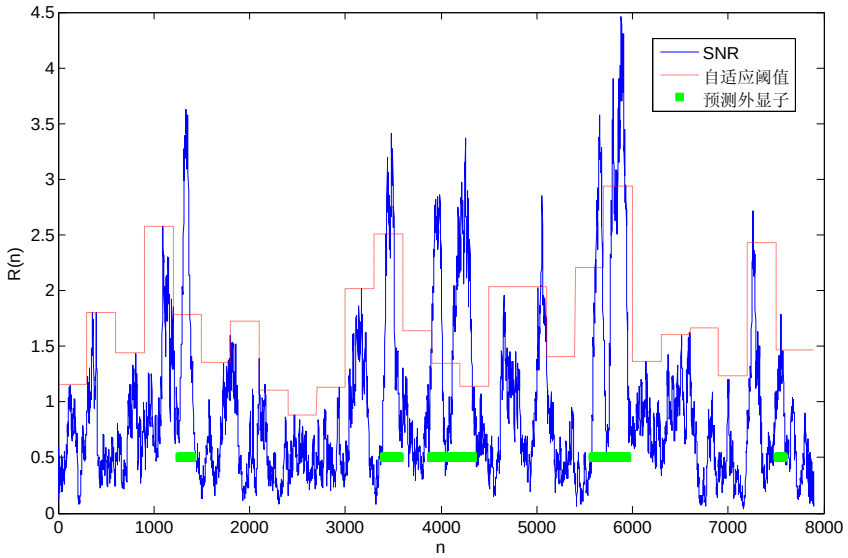


图 4-8　2 号基因的信噪比

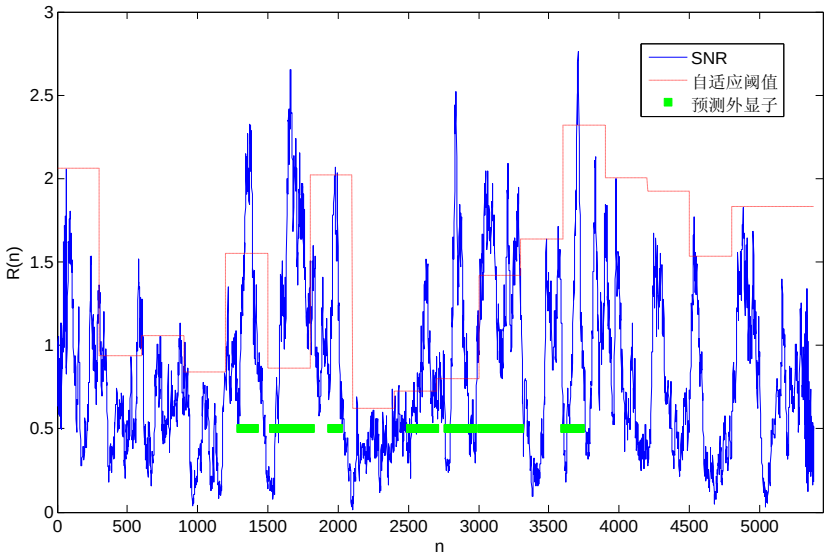


图 4-9　3 号基因的信噪比

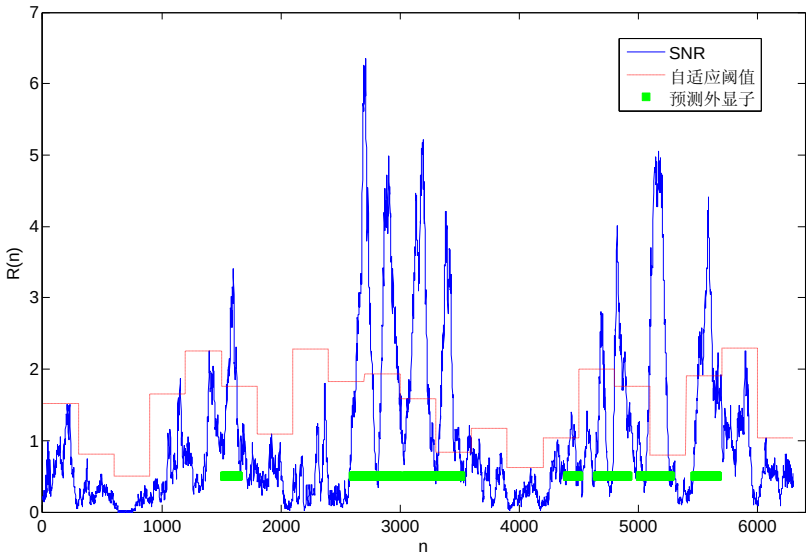


图 4-10　4 号基因的信噪比

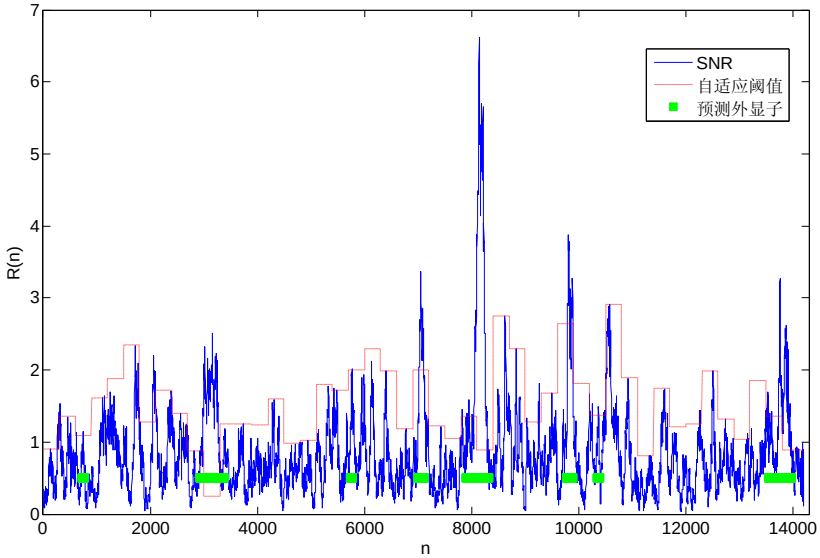


图 4-11　5 号基因的信噪比

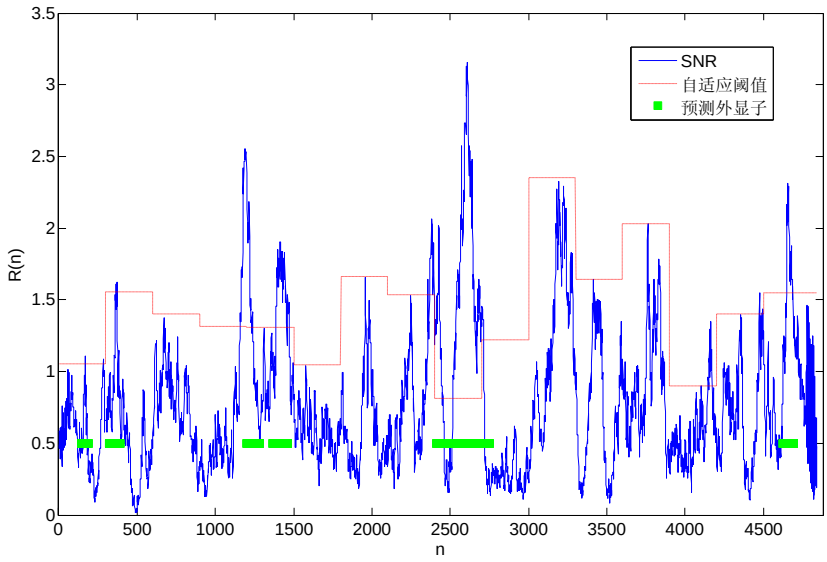


图 4-12　6 号基因的信噪比

5 延展性研究

5.1 基于信号增强的谱分析技术

由于部分基因序列编码区（比如，过短的外显子序列）的 3-周期性不强，甚至体现不出 3-周期性，经过 DFT 变换后，编码区与非编码区的频谱区分不明显，幅度差别不大。考虑到此种情况，本文设计了一种信号增强技术，对 DFT 变换后 $M/3$ 处峰值得到的功率谱 $R(m)$ 进行平滑滤波和信号增强，得到优化的功率谱，再进行谱分析和基因检测，显著提高了检测性能。该方法的本质，是将编码区的谱看作“信号”，而将非编码区的谱看作“噪声”，采用有效的技术对“信号”进行加强，对“噪声”进行抑制，提高对于编码区检测的准确度和精度。模型的核心思想是对弱 3-周期性的有效放大，实现准确判决。

5.1.1 算法设计

算法的关键是基于编码区“信号”与非编码区“噪声”的比例，设计一个有效的增益函数 $\Gamma(m)$。其中，信噪比可以通过短时平均信号能量与估计的噪声门限的比值来计算。该二者都是随时间变化的，并且自适应于局部的信号特征。短时平均信号能量计算公式为：

$$P(m)=\sum_i \alpha_i P(m-i)+\beta R(m) \tag{5-1}$$

其中 i 为相关器的阶数；α_i，β 是一个正的常数因子，控制算法随信号能量变化的敏感度，并起到平滑因子的作用，$\sum_i \alpha_i+\beta=1$。

慢变的噪声门限估计值计算公式为：

$$Q(m)=\begin{cases}(1+\gamma)Q(m-1), Q(m-1)\le P(m)\\ P(m), Q(m-1)>P(m)\end{cases} \tag{5-2}$$

其中，γ 是一个较小的正常数，控制噪声门限只适应变化的速度。

增强的信号 $\hat{R}(m)$ 通过下式进行计算：

$$\hat{R}(m)=\Gamma(m)\cdot R(m)=\frac{P(m)}{Q(m)}R(m) \tag{5-3}$$

5.1.2 结果分析

初始化 $P(1)=0.5R(1)$，$Q(1)=0.1$，相关器的阶数设为 $i=2$，$\alpha_1=\alpha_2=0.4$，$\beta=0.2$，$\gamma=2\times10^{-4}$。同时，若 $\Gamma(m)$ 过大，可能导致信号过大的放大，将比较模糊的信号滤掉，因此限制 $\Gamma(m)$ 不超过 3dB。

对 33 号基因的仿真结果如图 5-1 和图 5-2 所示。由仿真图可以清晰的看出，功率谱的曲线平滑了很多，且干扰噪声的幅度大大降低了，而信号的幅度得到了明显的提高。另外，通过数值计算得出，在不作处理的情况下，基因检测的正确率为 0.6354，在采用信号增强技术的情况下，基因检测的正确率为 0.9733。由此说明，采用信号增强技术可以对“噪声”有效的滤除，减小噪声对于阈值选取和基因检测的影响，提升检测性能。

对于短序列，它的 3-周期性特性表现不明显（即信噪比特性不显著），通过信号增强技术来提高其信噪比，其检测性能一定能得到有效的提高。

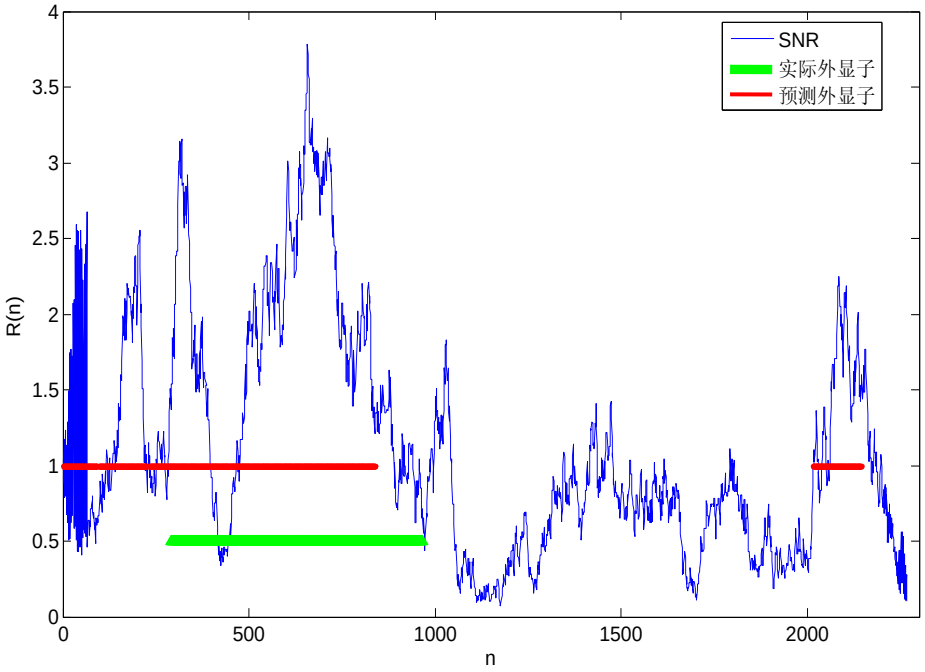


图 5-1　信号增强前信噪比及预测结果

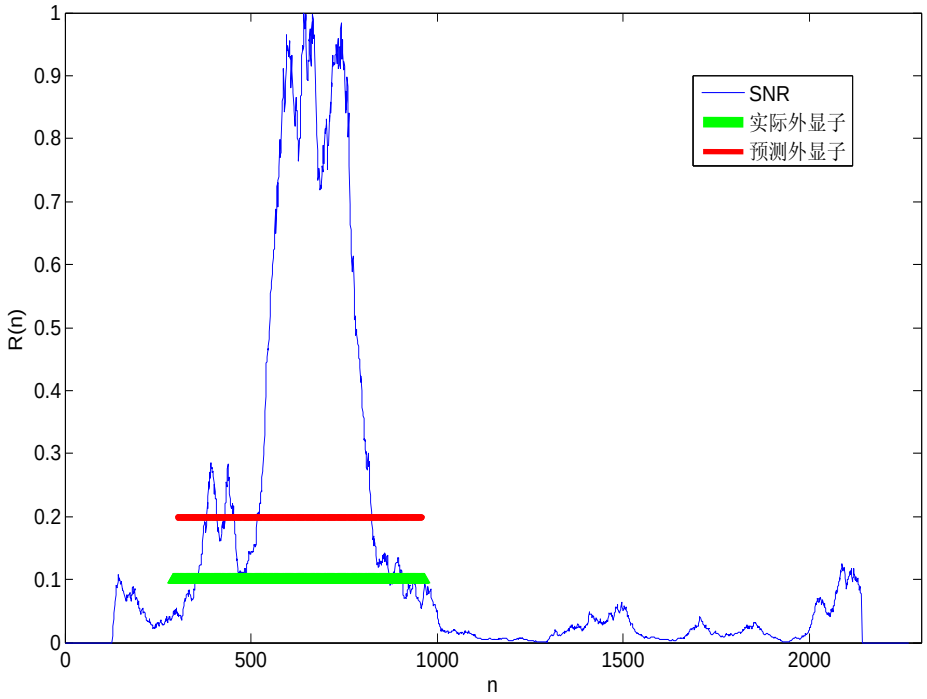
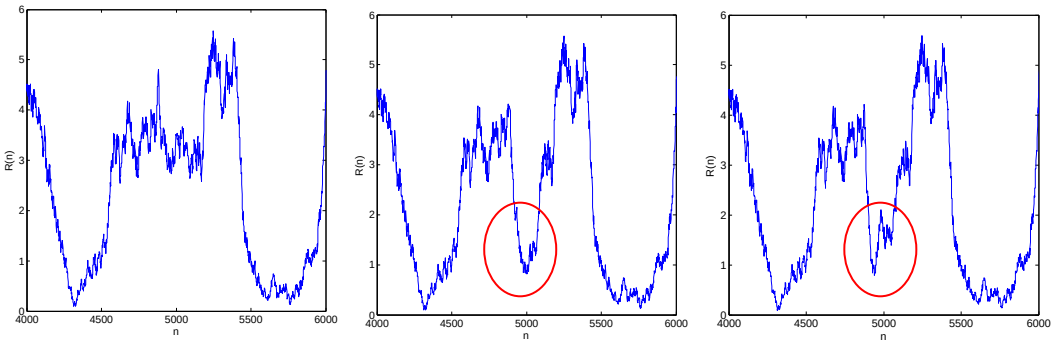


图 5-2　信号增强后的信噪比及预测结果

5.2 基因突变中的伪鞍部识别

由于内含子不具有 3-周期特性，发生在内含子上的基因突变是很难通过谱分析检测出来的。而对具有较好的 3-周期特性的外显子来讲，可以通过谱分析检测部分情况下的基因突变。由于外显子的 3-周期特性，当处于信噪比峰值处的碱基删除或插入时，该处的 3-周期特性将会由于碱基的删除或插入而消失，表现在信噪比曲线上是，原本正常曲线上的峰将变在突变处出现谷值，形成鞍部，因此称之为鞍部效应。形成鞍部效应的根本原因是在外显子中间及附近部位发生插入或删除突变时，在计算碱基频次时发生了较大变化。

以人类线粒体基因（NC_012920_1）为例，分别作插入和删除形式的基因突变，在 n=5000 处插入一个碱基 C，将 n=5000 处的碱基删除，对突变前后的信噪比曲线如图 5-3 所示，可以看出插入和删除确实造成了鞍部效应。



(a)突变前　　(b)插入　　(c)删除

图 5-3　基因突变前后信噪比曲线对比

通过上述分析表明，发生插入和删除形式的基因突变会使信噪比曲线产生鞍部效应，形成伪鞍部。对该类基因突变的识别方法就是对基因的信噪比，逐一分

析每个鞍部，判断其是否属于伪鞍部。在鞍部周围分别插入/删除一个碱基，判断信噪比曲线的变化，若鞍部发生上移甚至有变为波峰的趋势，说明基因发生了删除/插入的突变。

以在人类线粒体基因上插入碱基 C 的突变为例，在 n=5000 的鞍部周围，删除一个碱基，判断信噪比曲线的变化。如图 5-4 所示，在鞍部周围删除碱基，可以发现鞍部上升为波峰的趋势，说明该方法可以有效检测到碱基插入或删除形式的基因突变。仿真结果表明，若在某位置插入/删除一个基因，使伪鞍部恢复为波峰，说明与其相邻的左右 80 个基因范围内有一个发生了删除/插入的基因突变，但基因突变的位置暂时无法精确的检测出来。该方法对于外显子中间部位发生插入或删除突变时，鞍部效应比较显著，但是，对于外显子两端发生突变时检测效果不佳。

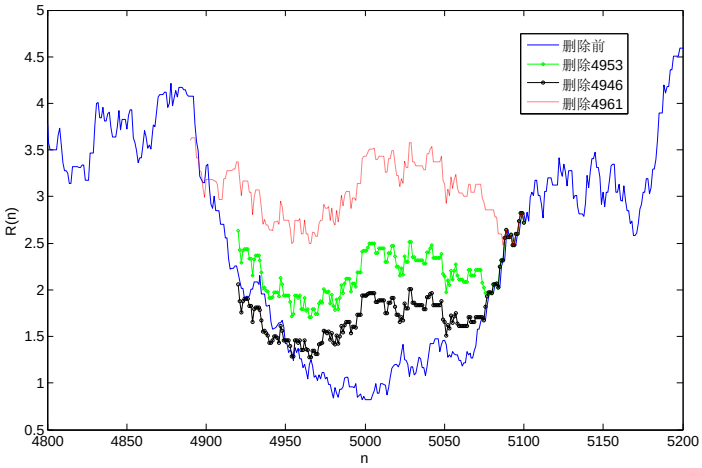


图 5-4 伪鞍部随基因删除位置的变化

6 总结

DNA 是遗传物质的载体，但并不是所有的 DNA 都能够反映生物的性状，只有编码区能够编码蛋白质，因此如何从 DNA 序列中找到编码区部分成为近年来的一个热门话题。DNA 序列是由一系列字符构成的，怎样将这些字符转换为数字信号，与数字信号处理的方法结合起来，并通过这些方法来挖掘 DNA 序列的特性是本文的重点。

由于 DNA 序列本身是由 A、C、G 和 T 四种字符构成的字符串序列，因此需要用一定的方法将其转换成计算机能够处理的数值序列。第二章重点研究了常用的数值映射方法包括 Voss 映射、Z-curve 映射和实数映射，以及用 DFT 进行谱分析的方法。针对 DFT 计算量庞大的问题，提出了不同映射下的功率谱和信噪比的快速算法。

计算出 DNA 序列的信噪比谱后，需要设定一定的决策阈值。第三章重点研究了阈值确定方法，建立了基于模糊逻辑的自适应阈值模型，并设定了具体的评估标准来评估基因预测，获得了较好的预测效果。

为进一步提高基因预测的准确性，又研究了外显子端点的精确定位问题。分别采用图像分割中的技术和基于边界搜索的算法对端点的确定展开了研究，提高了端点定位的精度。

最后将信号增强引入谱分析中，将编码区的谱视为信号，而将非编码区的谱

视为噪声，形成了基于信号增强的谱分析技术；对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问题进行了分析，结论是利用信噪比方法可发现外显子峰值处碱基删除及插入形式的突变。

参考文献

[1] 王震, 基于信号处理方法的基因识别算法研究, 博士论文, 天津大学, 2011.

[2] E. Hamori, H curve, a novel method of representation of nucleotide series especially suited for long DNA sequence. Journal of Biological Chemistry, 258(2): 1318-1327,1983.

[3] E. Hamori, Graphic representation of long DNA by method of H curves-current results and further aspects. Biotechniques,7(7)：710-720，1989.

[4] M. A. Gates. A simple way to look at DNA. Journal of theoretical Biology, 119(2):319-328, 1986.

[5]P. P. Vaidyan athan. Genomics and proteomics: a signal processor’s tour. Circuits and Systems Magazine, 4(4):6-29, 2004.

[6]EAmbikairajah. Gene and exon prediction using time-domain algorithms. Proceeding of ISSP, 2005.

[7] J.P. Mena-chalco. Identification of protein coding regions using the modified gabor-wavelet transforms. IEEE Transaction on Computational Biology and Bioinformation, 2007.

[8]马宝山，基于信号处理理论和方法的基因预测研究, 博士论文，大连海事大学，2008.

[9]S. D. Sharma, K. Shakya, S. N. Sharma. Evaluation of DNA Mapping Schemes for Exon Detection. IEEE ICCCET. 2011: 71-74.

[10]百度百科, 模糊逻辑, http://baike.baidu.com/view/338152.htm，2012-9-23.

[11]A. Agrawal, PCRPA: Protein-coding region predictor and analyzer-exon prediction using power spectrum estimation method, in proceedings of the national conference on IT and its application, 2006.

[12]百度百科, ROC 曲线, http://baike.baidu.com/view/42249.htm，2012-9-24.

[13]田元新，陈超，邹小勇等．外显子周期三行为特征的研究，化学学报，63(13): 1215-1219,2005.

[14]美国国家生物信息中心, www.ncbi.com, 2012-9-22

[15]张晔, 信号时频分析及应用, 哈尔滨, 哈尔滨工业大学出版社, 2006.

附录

1、100 个人和鼠基因的自适应阈值

1.2494,1.6052,2.3285,2.0873,1.7768,2.4342,1.4426,1.7833,2.1772
1.8799
2.2778,1.6708,2.1636,1.7733,2.3357,1.9872,2.6399,1.7151,1.319,2.0664
1.8459,1.5744,1.7288,3.2271,1.5763,2.32,2.0625,1.8441,2.1129,1.7922,2.1097,2.1564,1.771,1.87
22,1.9251,1.5758,1.8799,2.2606,1.1238,2.0229,1.7277,1.4332,1.2908,1.8945,2.0451,2.1926,3.416
6,3.0745,3.0723,3.1479,1.801,1.8575
1.7218,1.1399,0.73111,1.7833,1.3969
0.36703
1.4126,2.1196,1.547,1.3341,1.6723,1.5481,1.5989,1.3969,2.78,1.801
1.6062,1.8358,1.0085
2.0115,1.7681,2.1731,2.065,1.5509,1.8046,1.182
1.4583,1.32
2.4849,1.8,1.9961,1.769
1.5953,1.0048
2.1703,0.15021,1.88,2.0947,1.8158,2.5491,1.4207
3.0764,1.1785,2.4075
1.6527,1.4809,2.0279,1.2983,1.5185
1.5718,3.2254
1.5678,2.6998,2.1559,1.5671,1.2018,1.69
1.3614,0.79003
1.4449,1.8309,1.9741,1.5835,1.7418,2.6531,1.648,1.8336,1.6325,1.8064,2.9839,1.8137,2.3448,1.
4236,1.9453,1.35,1.833,0.21178,1.5953,1.8945
1.9719,1.7464,2.1719,1.5198,0.2268,0.72218,1.9991
0.48442
1.6983,1.8146,2.4724
1.7714,1.8798,2.28
1.5066,1.9288,1.1629,2.5337,2.2077,0.2268
1.8367,1.6088,1.0702
2.1161,1.7151,3.2167,1.7026,1.432,1.6155
0
2.2501,2.0939,1.1672,2.4265,2.5114,2.6998,2.0229
1.4446,1.4345,2.3696,0.80887,1.8158,1.2983
1.5677,2.0131,2.83,1.5505,1.9269,1.0615,1.4332,1.5429
2.0698,1.2461,2.2843,1.5988,1.7073,1.9827,2.3251,1.7714,1.7633,2.3415,1.6456,2.2081
1.7134,1.8325
1.7891,1.3839,2.0098,1.8529,2.15,1.8099,1.7666,1.6169,2.2516,1.5472
1.7832,0.44012,1.75,1.4659
1.9083,1.4632,1.7541,1.8382,1.4428,1.97,1.847,1.8316,1.3152,1.7831,1.76,1.5859
1.7924,1.6971,1.1786,2.1643,3.012,2.3487,1.3095
0.21178
1.7408,1.8997,2.5748,1.4037,2.2061,1.8078,1.8645,2.4195,1.46,1.7911,2.008,1.4463,2.2098,3.87

8,2.3412,1.6304,1.6158,2.2746,2.0395,2.463,1.8203,1.98,1.7922,2.1927,1.4261,2.4343,2.2955,1.4
983,1.1672,1.3041,1.4583,2.3276,2.0623,2.6739,2.0866,1.9083,2.3088,1.2674,2.14,1.3715,1.678,
2.5588,2.2489,1.59,2.3938,2.2843,1.6062,1.5706,1.8722,2.3251,1.4576,2.0017
1.28
2.5988,1.3181,2.1559,1.5678,1.6346,1.8575,1.8529,2.0395,1.7877,2.0176,2.1298,2.059,1.5454
1.4866
1.9489,0.81503
1.0894,1.7904,1.7459,2.4623,2.0972,2.1135,1.0782,1.595,1.35,1.4993,1.7405
2.2178,1.6795,1.3712,1.6608,1.3041,1.5116,0.29952,1.0701,1.7265
0.56044
2.5904,1.2966,1.758,1.2733,1.6897,1.4671,1.66,1.699,1.83,1.0615,2.0647
1.2966,1.6139,1.3758,1.7134
1.3842,1.7151,2.7213,2.4503
2.0162,2.2546,1.0311,2.0873,2.2061
1.8029,1.2959
1.869,1.4108,2.49
1.8722,1.7049,2.49,1.9908,1.6284,0.19119,1.7408,1.5587,1.8239,2.52,1.6811,2.14,1.6818,2.1455,
1.5797
1.8064,1.3545,1.3842,1.7514,1.342,2.34,1.6773,1.3566,1.8854,1.8325,2.13,2.49,1.5463,1.4886,2.
2081,2.5963
1.6443,1.3095,1.8772,1.5099,1.8099,0.92487,2.1736,2.0219,1.5127,1.6595,1.3819,2.1107,2.059
1.3895,1.708,1.9743
1.0997,1.9354,2.1957
1.7833,1.5044,1.44,2.3339,1.6336,1.6923,1.3935,1.5799,1.6457,0.83191,2.2269,1.8538
1.5066,1.3772,2.3036,2.3189,2.26,2.1793,2.0872,2.4833,2.3088,1.9311,2.1736,2.5175,1.8858,1.4
367
1.9551,2.2269,1.7892,1.4195,1.5217,1.3567
2.04,2.1282,1.5671,2.0105,2.2826,2.6998,1.7904,0.054587
1.6466,2.3183,2.0682,1.5099,1.7457
1.3531,2.2078
2.3647,1.3818
2.4478,1.9022,1.6923
1.9627,2.065,1.3077,1.5553,1.2394
1.4993,2.3285,2.0395,1.9719,1.4291,2.2489,1.4583,2.3189,2.7026
2.4618,1.7207,1.9581,2.2993,2.7229,1.5412,1.9927,1.9245
1.8664,1.6481,1.474
1.5345
1.6681,1.1795,1.5553,1.7892,0.4963
2.2949,1.7151,0.92487,1.4144,2.3189,1.8772
1.6341,1.2173,2.11,0.27034,1.432
1.9551,1.7666,1.96,1.4019,1.4951,1.9101
1.5392,0.81372,0.66483
1.365,1.6543,1.7633,2.14,1.1367,1.7049,2.2843,1.9,1.6719,1.2757,2.1957,1.5588,1.6526,1.6336,1
.4291

1.6945,1.7805,1.2465,1.5012,1.9006
0.96912
1.4623,2.23,2.1559,1.5454,1.7414,1.4025,2.2108,2.4276,1.5512,1.924,1.9298,1.82,1.795,2.1484,1
.3341,1.3918,2.4849,1.4386,1.6445,2.4478,2.21,2.4538,2.4265,2.2369,2.925,1.6766,2.0191,1.650
9,1.3971,1.7278,1.0304,1.6527,2.7381,1.3683,1.7684,1.8931,2.9839,1.2489,1.7459,2.1958,1.6687
,2.8269
2.196,2.5439,2.0244,1.4654,1.0048
1.7119
2.28,1.5859,0.96979,1.4019,1.1629,2.2751,2.6998,2.6212,2.5114,1.6233,2.4284,1.4162
2.0131, 1.7916,1.5597,1.7694,1.1253,1.4615,1.5683
1.3972, 1.0395,1.2465,1.4977,1.2877,1.6416,1.4615,1.1119,1.0894,1.6867
0.2216
1.6983,0.78407,0.2314
1.7539,2.9472
1.9593,1.4152
1.8352,1.9699,2.0395,2.4344,1.6644,1.2173
1.6112,1.841
1.342,2.0468,2.1108,1.421,1.199,2.1958,1.2555
1.769,1.6019,2.1097,1.0789,2.192,2.0603,1.8575
0.24625,0.83558
2.4891,2.0537,2.6994,1.5995,1.44,2.3015,1.6139,1.8081,2.0017
1.7464,1.6108,1.9251
1.3803,1.8598,1.95
2.2076,2.1958,1.3964,1.9083,1.4152,1.6062,1.659,1.7515,1.5099
1.7541,1.6711,1.2966,1.8616,1.7063,1.9255,1.6495,1.7637,2.4501,1.8726,1.7651,1.2912,2.0647,1
.3758,2.3339,2.0105,2.5116,1.6213,2.0478,2.83,2.227
2.1861

2、100 个人和鼠基因的固定阈值与自适应阈值对比

基因序号	固定阈值	固定阈值 Sn	固定阈值 Sp	自适应阈值 Sn	自适应阈值 Sp
1	2.3	0.3732	0.3649	0.6342	0.6025
2	3.4	0.8364	1	0.9907	0.9772
3	2	0.7264	0.8014	0.865	0.4758
4	2.2	0.551	0.3768	0.8762	0.7288
5	2	0.5054	0.1152	0.7917	0.2479
6	1.8	0.9127	0.8473	0.9554	0.8965
7	1.9	0.9241	0.9636	1	0.94
8	3	0.7627	0.5448	0.5721	0.3087
9	1.5	0.6263	0.5475	0.9099	0.69
10	1.9	0.7116	0.1871	0.3889	0.142
11	1.6	0.9747	0.7612	1	0.7754
12	2.4	0.4486	0.2775	0.4486	0.4034
13	1.6	0.1364	0.1552	0.5682	0.6429
14	2.3	0.8638	0.7615	0.9962	0.8469
15	2.2	0.9155	0.8471	1	0.9726
16	1.9	0.3285	0.2689	0.5996	0.4843
17	3.3	0.8593	0.956	0.8878	0.956

18	1.9	0.4906	0.4994	0.5813	0.6096
19	1.5	0.8738	1	0.9352	1
20	2.3	0.3635	0.3611	0.6316	0.2571
21	1.7	0.6128	0.6079	0.8922	0.8663
22	1.9	0.8691	1	1	0.9933
23	2.4	0.8333	0.7253	0.7574	0.9922
24	2.1	0.8939	0.5175	0.8902	0.7532
25	2.1	0.3802	0.2829	0.8733	0.264
26	1.9	0.8027	0.9917	0.9866	0.8626
27	2.2	0.8385	0.8798	0.8895	0.9021
28	1.5	0.4005	0.5833	0	0
29	2.1	0.7658	0.7207	0.7048	0.665
30	1.6	0.7215	0.6783	0.8488	0.9756
31	1.6	0.5829	0.4568	0.6974	0.7961
32	2.4	0.1949	0.3224	0.4379	0.2545
33	1.9	0.8133	0.5743	0.8874	0.6877
34	2.5	0.7278	0.7815	0.9647	0.6698
35	1.8	0.8062	0.7685	0.8899	0.8945
36	2.1	0.7659	0.8075	0.9683	0.8309
37	2.1	0.3883	0.3961	0.8447	0.836
38	1.5	0.1584	0.7791	1	0.9311
39	2.5	0.297	0.2523	0.5846	0.3566
40	1.9	0.5094	0.5564	0.9531	0.5971
41	1.9	0.3197	0.0726	0.2063	0.0984
42	1.5	0.7333	1	0.9233	0.9395
43	1.9	0.5882	0.7397	0.6895	0.7468
44	2.2	0.2759	0.186	0.623	0.1617
45	1.7	0.6799	0.6552	0.9393	0.9712
46	1.6	0.931	0.9865	1	0.9867
47	2	0.5071	0.5071	0.8525	0.5724
48	1.6	0.5519	0.2769	0.5445	0.4391
49	2.7	0.6197	0.6626	0.5851	0.6918
50	2.1	0.8273	0.9889	0.8282	0.9889
51	1.5	0.6111	0.88	0.8577	0.8255
52	2	0.8886	0.8593	0.8562	0.8294
53	1.8	0.7306	0.5285	0.8864	0.8668
54	2.6	0.6797	0.6003	0.5923	0.3291
55	1.8	0.4553	0.2477	0.4493	0.2656
56	1.9	0.7	0.5929	0.8172	0.7011
57	2.1	0.4202	0.411	0.6197	0.4084
58	2.1	0.1729	0.3076	0.5874	0.4795
59	4	0.4644	0.3827	0.7416	0.2371
60	2	0.7745	0.6581	0.9655	0.8107
61	2.1	0.6559	0.6794	0.8037	0.7375
62	2.4	0.6822	0.7556	0.7198	0.5268
63	2.3	0.7139	0.6642	0.8703	0.574
64	1.9	0.8883	0.7433	0.9451	0.8408
65	1.9	0.4482	0.4307	0.3176	0.6912
66	2.2	0.8408	0.8342	0.8955	0.841

67	2.2	0.4471	0.3779	0.319	0.2196
68	2.8	0.4261	0.436	0.7955	0.4878
69	1.8	0.7447	0.5481	0.789	0.6703
70	1.6	0.7626	1	0.6028	1
71	1.9	0.9194	1	0.9372	0.9056
72	1.5	0.416	0.3382	0.3993	0.4254
73	2.2	0.7465	0.8097	0.9426	0.7079
74	2	0.7756	0.6334	0.7854	0.5493
75	1.6	0.9682	0.9024	0.9845	0.9967
76	2.2	0.5333	0.5166	0.7741	0.435
77	2	0.7447	0.757	0.8634	0.5611
78	1.6	0.8347	0.8733	0.967	1
79	1.5	0.7754	0.8093	0.9956	0.8019
80	3.6	0.6784	0.692	0.7895	0.2329
81	2.2	0.7635	0.7503	0.8603	0.641
82	1.7	0.8078	0.9162	0.8975	0.9165
83	2	0.5515	0.5284	0.6513	0.6612
84	1.7	0.6562	0.4532	0.8527	0.7925
85	1.7	0.5949	0.5648	0.8568	0.5861
86	1.5	0.861	0.9343	1	0.9144
87	1.5	0.7173	0.8614	0.9112	0.8016
88	3.3	0.9214	0.9293	0.8784	0.7015
89	2.5	0.9018	0.9671	0.7157	0.8906
90	2	0.7914	0.7649	0.8306	0.8193
91	2.1	0.7586	0.7306	1	0.6561
92	1.6	0.5692	0.3593	0.9304	0.7188
93	2	0.5831	0.5262	0.8317	0.8226
94	1.6	0.5425	0.9268	0.8908	0.8691
95	2.2	0.4701	0.5036	0.7484	0.6047
96	1.8	0.4805	0.5103	0.3519	0.4903
97	1.9	0.6496	0.6847	0.9902	0.9408
98	1.9	0.455	0.263	0.5827	0.3108
99	2.6	0.6734	0.7746	0.6516	0.3009
100	2.2	0.697	0.6987	0.697	0.6987

3、6 个未注释基因的自适应阈值

1 号基因 0.7552 1.4855 1.5306 0.7644 0.5256 1.6844 0.8761 1.2032
1.1629 2.4732 1.4785 1.5543 2.7021 1.4154 1.1858 1.5579
0.9615 1.1164 1.763

2 号基因 1.1535 1.803 1.438 2.5777 1.7832 1.3587 1.7292 1.109
0.8862 1.1326 2.0219 2.5095 1.6371 1.35 1.14 2.0319 2.0318
1.4055 2.2055 2.9389 1.3678 1.6058 1.6661 1.2355 2.433
1.4655

3 号基因 2.06 0.9379 1.0571 0.8424 1.549 0.8633 2.0219 0.6218 0.724
0.799 1.4186 1.6396 2.319 2.0054 1.9223 1.5362 1.8326

4 号基因 1.5183 0.8158 0.5045 1.6513 2.2597 1.7649 1.0893 2.2823
1.8315 1.9396 1.5865 0.837 1.1704 0.6168 1.043 2.0072
1.7649 0.7949 1.9029 2.3011 1.0406

5 号基因.9065	1.3581	1.0893	1.6177	1.8751	2.348	1.2845	1.7241	1.4041
	0.8813	0.2448	1.2564	1.2564	1.2376	1.5986	0.9785	1.024
	1.8043	1.724	2.0054	2.3011	1.9926	1.1858	2.0072	1.2206
	1.0551	1.3531	0.8889	2.7455	2.3011	1.2801	1.6801	2.6461
	1.8086	1.3742	2.9095	1.8891	0.8158	1.7401	1.2069	1.2471
	1.9937	1.3211	1.0407	1.8507	1.3531	0.8902		
6 号基因 1.0551	1.5531	1.4041	1.3128	1.3054	1.0466	1.6615	1.5351	
	0.8119	1.2206	2.355	1.6439	2.0319	0.9008	1.4021	1.549