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 **基于联合识别的基因预测**

摘 要：

本文围绕基因序列识别问题，在合理假设的基础上，通过数学推导证明 voss 映射和 Z-curve 映射效果等价，并给出了实数映射的快速算法。而后通过统计分析，模型优化对不同类物种进行了信噪比阈值确定并验证其有效性。最后根据联合识别对未注释 DNA 编码区预测进行了建模，实现和分析。

针对问题 1：推导出了 Voss 映射下功率谱与信噪比的快速算法。其中，功率谱算法仅需统计，无需进行繁杂的 DFT 变换；信噪比算法把加窗处理和平均处理结合起来，大大加快计算速度。同时推导得出：Z-curve 映射下的频谱和信噪比分别为 Voss 映射下的频谱和信噪比的 4 倍和 4/3 倍，并得出实数映射下信噪比的快速计算公式。

针对问题 2：采用 3 种阈值确定方法：经验阈值法、最优化方法、曲线法分别对人、小家鼠、褐家鼠、哺乳动物 4 类基因进行阈值确定；依据敏感性、专一性、总正确率 3 指标分析不同阈值确定方法的基因判别效果，确定了基因的最优的阈值确定方法，并得到其最优信噪比阈值；最后针对误判原因进行了初步探讨。

针对问题 3：单一的固定窗口的功率谱图或是移动序列的信噪比曲线图在识别上不够精确，通过两种曲线的联合识别，能更精确的判断外显子片段（区间和两端点）。且通过 Matlab 里的 Sequence Viewer 对所识别片段进行进一步辨识，提高了端点辨识准确度。

针对问题 4：总结出以下几种能够识别基因编码序列的特征指标，分别为“非 3-碱基周期性”编码序列指标，旋转矢量指标，非均匀指标及干涉指标。并对采用上述指标识别编码序列的原理进行了概括与总结。

本文亮点在于：（1）找到 2 种最优阈值确定方法，通过统计其判别正确率，判别效果均好于以 2 为阈值的经验法；（2）结合固定窗口滑动法与移动序列的信噪比

曲线法对基因识别的特点，提出联合识别理论，提高了辨识效果。

[关键词]：基因识别 功率谱 信噪比 阈值 3-周期特性 联合识别

目录

一：问题的重述

DNA 是脱氧核糖核酸的简称，是绝大部分生物遗传信息的化学载体。一个 DNA 链可分为基因和基因间隔区。基因可分为外显子和内含子。外显子就是基因内的编码区，可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，而内含子并不参与蛋白质的编码。

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。这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。基因预测问题的一类方法是基于统计学的，另一类方法是基于信号处理与分析方法。

根据题目要求、给出的基因预测方法简介及部分基因序列数据解决以下几个问题：

（1） 鉴于很长的 DNA 序列，在计算其功率谱或信噪比时， DFT 的总体计算量仍很大，影响所设计基因识别算法的效率。针对 Voss 映射，探求其功率谱与信噪比的某种快速计算方法。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。同时推导实数映射功率谱与信噪比的快速计算公式。

（2） 找出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。随后，对按照信噪比特征将编码与非编码区间分类的有效性进行统计分析；并对分类识别时所产生的分类错误作适当探讨分析。

（3） 目前，基因识别方面的多数算法结果还不能很充分的探测尚未被批注释的、完整的 DNA 序列的所有外显子，找到好的解决方法对附件中 6 个未被注释的 DNA 序列的编码区进行预测，并对此法的准确率做出适当评估。

（4） 除频谱或信噪比这样单一的判别特征外，总结并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。判断能否利用频谱或信噪比方法发现基因编码序列可能存在的突变，且分析。

二：模型的假设

（1）一个没有错误符号的、长度为 N 的 DNA 序列的总功率的平均值为 N。

（2）题中所给已注释 DNA 序列样本完整、无误，且不考虑基因突变。

三：符号说明

符号	意义
R_I	Voss 映射下的信噪比
R_S	Z-curve 映射下信噪比
R_0	阈值
F	核苷酸域
P	窗口移动量
N	DNA 序列长度
S_1	外显子信噪比组成的集合

S_2	内含子信噪比组成的集合
S_n	敏感性
S_p	专一性
A_c	阈值判别的总正确性

四　问题分析与求解

4.1　问题一

4.1.1　Voss 映射下信噪比快速算法

为简化信噪比的计算，我们进行以下分析：令 x_b, y_b, z_b 分别表示各核苷酸 b 在 3 个密码子位置上的频数，即核苷酸 b 在序列的 0，3，6，…和 1，4，7，…以及 2，5，8，…位置上分别出现的频数。其中 $b \in I$，$I = \{A, T, G, C\}$，有公式如下：

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

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

$$= \sum_{b \in I} \{ (x_b, y_b, z_b) \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} \} = \sum_{b \in I} X_b^T M X_b \tag{4-1}$$

其中，$X_b = (x_b, y_b, z_b)^T$。

文献[1]研究指出：一个没有错误符号的、长度为 N 的 DNA 序列的总功率的表达式如下：

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

从而可得总功率的平均值为 $\overline{E} = E / N = N^2 / N = N$。则信噪比可表示为：

$$R = \frac{P(N/3)}{\overline{E}} = \frac{\sum_{b \in I} X_b^T M X_b}{N} \tag{4-3}$$

由式（4-3）可见：利用此法计算信噪比，不再需要对指示序列进行 DFT 变换，只需统计 4 种核苷酸在 3 种不同位置上出现的频数即可。此外，这样的计算还有累

加功能，可使信噪比计算工作量锐减[2]。

4.1.2 Voss 映射下功率谱快速算法

基于文献[3]的研究，利用 O. Welch 提出的原理，把加窗处理与平均处理结合起来，以达到快速计算功率谱的目的。具体实现方法如下：

（1） 利用 Voss 映射方法将一 DNA 序列映设成四个二进制序列，即生成二进制序列 $\{u_b[n]\}$：$u_b[0], u_b[1], \mathrm{L}$，$u_b[N-1]$（$b \in I$）。$I = \{A, T, G, C\}$。

（2） 将序列分成若干段。分段时，可使各段之间有重叠，进而减小方差。例如假设有一段长为 N 的待分析的信号序列，首先将整个序列划分成独立的 L 段，各段之间重叠 50%，则各段信号由 $m = 2N/(L+1)$ 个采样点组成。

（3） 选择一个适当的窗函数，长度为 m。此窗口的滑动步长是可变的，这种可变的滑动步长能够取得更加高精度的预测结果，并且大大缩短了计算所运行的时间。

（4） 将各分段的数据乘以此窗函数，即进行加窗处理。加窗的优点在于无论什么样的窗函数均可使谱估计非负。接着再利用 FFT 计算加窗处理后的序列的功率谱。

（5） 最后将所有各段序列的功率谱值进行平均，该平均值即为原序列的功率谱密度值。

该方法中对 DNA 序列进行功率谱分析的定义如下：

$$\underset{per}{\hat{S}}(f) = \frac{1}{L}\sum_{i=1}^{L}\left(\sum_{\alpha}\frac{1}{m}\left|\sum_{j=1}^{m}U_{\alpha}(x_j)\exp 2\pi fj\right|^2\right) \tag{4-4}$$

式中，$\alpha = A, T, G, C$，$U_{\alpha}(x_j)$ 为各碱基符号下对应的二进制序列。

4.1.3 Voss 映射频谱和 Z-curve 映射频谱间的关系

（1） Voss 映射下的频谱表示

假设 DNA 序列 $x(n)$ 长度为 N，滑动窗口在上 $x(n)$ 做 M 点的 DFT 变换定义如下：

$$X(m,k) = \sum_{n=0}^{M-1}x(n+m)e^{-j2\pi nk/M} \tag{4-5}$$

窗口起始点 m=0，P，. . . (*N*-1)/*P*，其中 P 是窗口移动量。若果 P=1，即一次移动一个碱基；若 P=3，即一次移动一个密码子。为了得到 $x(n)$ 的三周期组成成分，令 M=3L，其中 L 为正整数，令频率 k 为 L，即 M/3。则式（4-5）变为

$$X(m) \text{@} X(m,L) = \sum_{n=0}^{M-1}x(n+m)e^{-j2\pi n/3} \tag{4-6}$$

当 P=3 时，利用等式 $x(m+n) = x_m(n)$，则 $X(m)$ 可以改写为

$$X(m)=\sum_{r=0}^{2}\sum_{n=r,r+3,\ldots}^{\frac{N-1}{3}}x_m(3n+r)e^{-j2\pi r/3}=\sum_{r=0}^{2}X_{m_r}e^{-j2\pi r/3} \tag{4-7}$$

为了计算 Voss 映射下基因的频谱，根据式（4-7），可以得到下式：

$$X_{1m}\ @X_{1m0}+X_{1m1}e^{-j2\pi/3}+X_{1m2}e^{-j4\pi/\beta}\quad \forall 1\in F \tag{4-8}$$

又因为基因频谱的定义为

$$S_v(m)=\left|X_{Am}\right|^2+\left|X_{Cm}\right|^2+\left|X_{Gm}\right|^2+\left|X_{Tm}\right|^2 \tag{4-9}$$

且$\left|X_{1m}\right|^2=X_{1m}X_{1m}^{*}$。其中*代表复共轭，根据等式（4-9），可得下式：

$$\left|X_{lm}\right|^2=\frac{1}{2}\sum_{r=0}^{2}\left[X_{lm_r}-X_{lm_q}\right]^2 \tag{4-10}$$

所以得下式：

$$S_v(m)=1/2\sum_{1\in F}\sum_{r=0}^{2}\left[X_{1m_r}-X_{1m_q}\right]^2 \tag{4-11}$$

其中，$q=(r+1)$除 3 取余。

（2） Z-curve 映射下的频谱表示

以 DNA 序列$x(n)$为例，可得下式：

$$\begin{aligned}X(m)&=\sum_{n=0}^{M-1}x(n)e^{-j2\pi n/3}\\&=2\sum_{n=0}^{M-1}\left[x_A(n)+x_G(n)\right]e^{-j2\pi n/3}-\sum_{n=0}^{M-1}e^{-j2\pi n/3}\end{aligned} \tag{4-12}$$

因为$M=3L$，（4-12）式中第二部分和为 0。利用多项式，可得：

$$X(m)=2\sum_{r=0}^{2}\left[X_{Am_r}+X_{Gm_r}\right]e^{-j2\pi r/3} \tag{4-13}$$

同 Voss 可得

$$\left|X(m)\right|^2=2\sum_{r=0}^{2}\left[X_{Am_r}+X_{Gm_r}-X_{Am_q}-X_{Gm_q}\right]^2 \tag{4-14}$$

式中，$q=(r+1)$除 3 取余。

同样，计算$y(n)$和$z(n)$，则 z 变换后 DNA 的频谱公式为

$$S_z(m)=2\sum_{1\in F}\sum_{r=0}^{2}\left[X_{Am_r}+X_{1m_r}-X_{Am_q}-X_{1m_q}\right]^2 \tag{4-15}$$

式中，$q=(r+1)$除 3 取余。

令 $\tilde{F}$ 为核苷酸域 F 的一个子集，满足

$$\tilde{F}=\{C,G,T\}\subset F \tag{4-16}$$

所以可得：

$$S_z(m)=4S_v(m)+4\sum_{r=0}^{2}\left(X_{Amr}-X_{Amq}\right)\sum_{1}\left(X_{Amr}-X_{Amq}\right) \tag{4-17}$$

对上式进行化简，[X_{Ar}+X_{Gr}+X_{Cr}+X_{Tr}]等于窗口中出现密码子第 r 位置之和。因为，在每个密码子中，第 r 位置均有核苷酸。上述数量为常数，等于窗口长度的三分之一。

所以可得：

$$S_z(m)=4S_v(m) \tag{4-18}$$

（3） 结论

由公式（4-18）可见，Z-curve 映射下的频谱 $S_z(m)$ 是 Voss 映射下的频谱 $S_v(m)$ 的 4 倍。

4.1.4 Voss 映射信噪比和 Z-curve 映射信噪比间的关系

（1） 参数定义

由文献[4]可知：对于一个长度为 N 的 DNA 序列，其在 Voss 映射下的 4 个指示序列的三维仿射变换可以定义如下：

$$\begin{cases} x[n]=\alpha_{11}u_A[n]+\alpha_{12}u_C[n]+\alpha_{13}u_G[n]+\alpha_{14}u_T[n], \\ y[n]=\alpha_{21}u_A[n]+\alpha_{22}u_C[n]+\alpha_{23}u_G[n]+\alpha_{24}u_T[n], \\ z[n]=\alpha_{31}u_A[n]+\alpha_{32}u_C[n]+\alpha_{33}u_G[n]+\alpha_{34}u_T[n], \end{cases} \tag{4-19}$$

其中 $u_A[n]$， $u_C[n]$， $u_G[n]$， $u_T[n]$， $n=0,1,...N-1$ 是 DNA 序列 Voss 变换的四个指示序列，则（4-19）式可化简为：

$$\begin{pmatrix} x[n] \\ y[n] \\ z[n] \end{pmatrix}=\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \end{pmatrix}\begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix}, n=0,1,...N-1 \tag{4-20}$$

为了方便表达，我们用行矩阵和列矩阵来定义以上的实数系数矩阵

$$F=\left(\alpha_{ij}\right)_{3\times 4}=\begin{pmatrix} \alpha_1^T \\ \alpha_2^T \\ \alpha_3^T \end{pmatrix}=\begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix} \tag{4-21}$$

（2） 原理说明

由文献[4]中的定理三可知：

假设 DNA 序列的长度是 3 的倍数，并且矩阵 F 的列向量满足以下两点

i） $\forall i(1 \le i \le 4), \left\|\beta_i\right\|^2 \equiv c_1$，其中 c_1 为常数。

ii） $\forall i, j(1 \le i, j \le 4, i \ne j), \left\langle \beta_i, \beta_j \right\rangle \equiv c_2$，这里 $\langle \cdot, \cdot \rangle$ 表示两个矢量的内积运算，c_2 为常量。则此特定映射的信噪比 R_S 与 Voss 映射下信噪比 R_I 之间的关系为

$$R_S = \frac{c_1 - c_2}{c_1} R_I = \frac{4}{3} R_I \tag{4-22}$$

又因为 Z-curve 映射的系数矩阵满足以上定理的两个条件，所以可以得出 Z-curve 映射下的信噪比 R_S 与 Voss 映射下信噪比 R_I 之间的关系为 $R_S = \frac{4}{3} R_I$。

（3） 结论

Z-curve 映射下的信噪比 R_S 与 Voss 映射下信噪比 R_I 之间的关系为 $R_S = \frac{4}{3} R_I$。

4.1.5 实数映射下信噪比快速计算

根据文献[4]的定理四可知：一维映射方案即实数数值表达的典型表达式如下：

$$r[n] = 1 \cdot u_A[n] + 2 \cdot u_C[n] + 3 \cdot u_G[n] + 4 \cdot u_T[n], n = 0,1...N-1 \tag{4-23}$$

根据文献[2]的定理四的推论二可知：假设 DNA 序列中四种核苷酸 A、C、G、T 的数量分别为 N_A, N_C, N_G, N_T。对于式（4-23）所示的一维映射，DNA 序列的信噪比（SNR）可以表示为：

$$R_r = \frac{P_r(N/3)}{E_r / N} = \frac{(X_1 + 2X_2 + 3X_3 + 4X_4)^T M (X_1 + 2X_2 + 3X_3 + 4X_4)}{N_A + 4N_C + 9N_G + 16N_T} \tag{4-24}$$

且满足：

$$X_1 + X_2 + X_3 + X_4 = \begin{pmatrix} \frac{N}{3} & \frac{N}{3} & \frac{N}{3} \end{pmatrix}^T, \begin{pmatrix} \frac{N}{3} & \frac{N}{3} & \frac{N}{3} \end{pmatrix} M = 0 \quad and \quad M \begin{pmatrix} N/3 \\ N/3 \\ N/3 \end{pmatrix} = 0 \tag{4-25}$$

其中 M 为系数矩阵，满足

$$M = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \tag{4-26}$$

而 X_1, X_2, X_3, X_4 为出现在三种密码子位置上的四种核苷酸的出现频率矢量。表达式如下：

$$(X_1, X_2, X_3, X_4) = \begin{pmatrix} x_A & x_C & x_G & x_T \\ y_A & y_C & y_G & y_T \\ z_A & z_C & z_G & z_T \end{pmatrix} \tag{4-27}$$

所以可以推断出信噪比的表达式为，此式为信噪比的快速计算公式

$$R_r = \frac{(X_1 + 2X_2 + 3X_4)^T M (X_2 + 2X_3 + 3X_4)}{N_A + 4N_C + 9N_G + 16N_T} \tag{4-28}$$

4.2 问题二

4.2.1 4 类基因外显子、内含子信噪比的统计分析

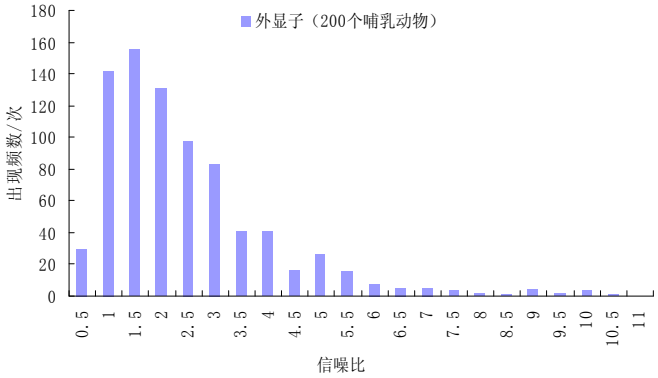
DNA 序列的信噪比值的大小，既表示频谱峰值（*Peak Value*）的相对高度，也反映编码或非编码序列 3-周期性的强弱。信噪比 R 大于某个适当选定的阈值 R_0（比如 R_0=2），是 DNA 序列上编码序列片段（外显子）通常满足的特性，而内含子则一般不具有该性质。现有的对外显子信噪比特性研究的一些文献[2,5]中指出，信噪比 $R \geq 2$ 是外显子一个普遍的特性，从而以此为指标来识别和区分外显子和内含子。然而在实验中发现，这一经验值 R_0=2 虽然可以区分出蛋白编码区和非编码区，但存在很大的预测误差，因为不同生物功率谱峰值高低不同，选用同一个阈值显然未考虑到生物本身的阈值特性。为此我们对上述观点提出疑问。依据上文提出的信噪比快速算法计算题目中给出的带有编码外显子信息的 100 个人和鼠类的，以及 200 个哺乳动物类基因序列的外显子（1264 个）和内含子（962 个）的信噪比，详见附录 A。利用 SPSS 统计分析软件统计、分析，结果详见表 4-1。

表 4-1 四类基因外显子、内含子信噪比均值和标准差统计

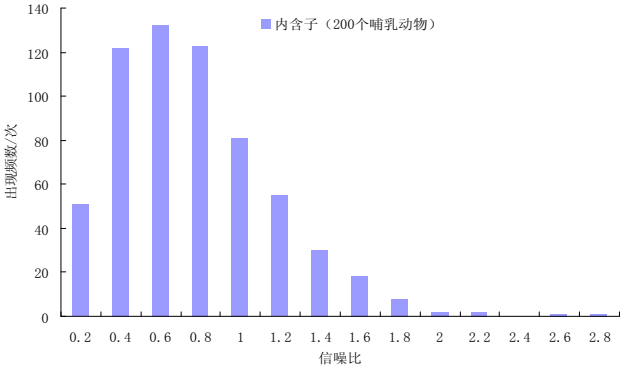
基因类别	外显子			内含子		
	数量	R 均值	R 标准差	数量	R 均值	R 标准差
人	35	3.02	3.071	26	0.82	0.533
小家鼠	357	2.46	2.508	275	0.68	0.414
褐家鼠	45	3	5.233	35	0.83	0.624
哺乳动物	827	2.72	6.243	626	0.67	0.394

注：数据源自著名的生物数据网站：http://www.ncbi.nlm.nih.gov/guide/

同时，对200个哺乳动物类基因序列的外显子、内含子信噪比的区间分布进行统计详见图4-1、4-2，其中外显子、内含子的统计单位区间分别为0.5和0.2。

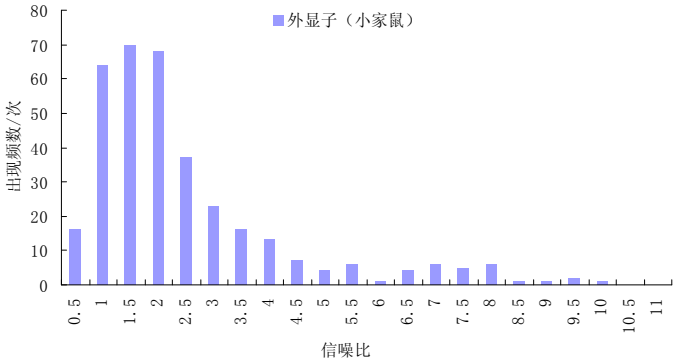


(a) 外显子

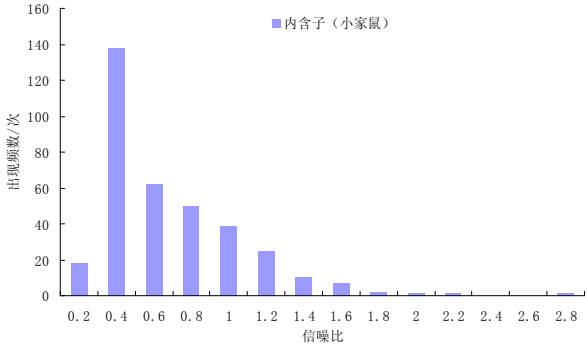


(b) 内含子

图 4-1 哺乳动物类 DNA 序列外显子、内含子的区间分布



(a) 外显子



(b) 内含子

图 4-2 小家鼠 DNA 序列外显子、内含子的区间分布

由表 4-1 可见，尽管上述研究基因的生物种类均为哺乳类生物，但其基因信噪比的均值和标准差均不同，不同生物基因外显子信噪比的标准差差异较大，且外显子的标准差远大于内含子。此外，由图 4-1、4-2 可见，哺乳动物类生物和小家鼠基因的内含子大多小于 2，但其外显子分布于[0,2]区间的频数分别为 458 和 218，占其总数的 55.38%和 61.1%，因此，对基于阈值的外显子判别方法而言，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，因此合理并精确地确定每类甚至每种生物基因的信噪比对基于信噪比阈值识别基因方法有着重要的意义。

文献[3]提出了一种基于靴带抽样算法推断基因预测的最佳阈值的方法。该方法首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，将这些实验观测的阈值作为靴带抽样算法的原始样本集，然后从原始样本集经靴带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。文献[2]提出了三种阈值确定方法。分别为：均值平均法，带标准差的加权平均法和阈值确定的最优化方法。

比较分析上述方法，发现以上方法均要求样本 DNA 序列上部分或全部外显子片段已知。其中，文献[3]中求取最优信噪比阈值的样本数据是基于已标注 DNA 序列实验观测阈值，而文献[2]中最优阈值的求取是基于已标注 DNA 序列内含子、外显子的信噪比值。显然，文献[2]方法对样本数据的要求较低。

在上述研究基础上，本文采用两种最优阈值确定方法对该问题展开研究。

4.2.2 基因阈值确定方法的研究

方法一：最优化方法

（1） 模型的建立：

假设所有外显子、内含子的信噪比组成样本数据集合，分别记为 S_1、S_2, 欲求的最优信噪比阈值为 R_0，且有 $R_i^{(1)} \in S_1, R_j^{(2)} \in S_2$，求解阈值 R_0 的优化模型为

$$\begin{aligned} &\max \sum_i \mathrm{sgn}(R_i^{(1)} - R_0) + \sum_j \mathrm{sgn}(R_0 - R_i^{(2)}) \\ &st: \qquad\qquad a < R_0 < b \end{aligned} \tag{4-29}$$

其中，[a,b]为最优信噪比阈值区间。该模型建立的思想是基于：信噪比 R 大于某个适当选定的阈值 R_0（比如 R_0=2），是 DNA 序列上编码序列片段（外显子）通常满足的特性，而内含子则一般不具有该性质。即在基因外显子、内含子信噪比样本集上，优化模型求得使判别正确率达到最大的阈值解 R_0。

（2） 实验材料：

采用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作为实验材料。因为所给样本序列中都明确标出了外显子和内含子的具体位置，具备该方法样本数据获取的要求。上述方法的实现采用 C#工具实现，具体源代码详见附录 B。

（3） 具体步骤：

① 依据 4.1.1 节中信噪比快速算法，利用公式 $R = \dfrac{P(N/3)}{\overline{E}} = \dfrac{\sum_{b\in I} X_b^T M X_b}{N}$，计算各中

生物 DNA 序列中各外显子和内含子的信噪比

② 建立同一种（类）生物的外显子的信噪比样本，记为集合 S_1，同理建立内含子信噪比样本，记为集合 S_2。

③ 利用 C#工具编程实现上述优化模型。求解得到各种（类）生物的最佳信噪比阈值，详见表 4-2。

表 4-2 四类基因最优信噪比阈值确定结果

基因类别	方法 1 结果	方法 2 结果
人（homo sapiens）	1.01	1.2
小家鼠（mus musculus）	0.937	1.029
褐家鼠（rattus norvegicus）	1.151	1.152
哺乳动物（mammalia）	1.15	1.233

方法二：曲线法

（1） 模型的建立：

首先，假设所选定的信噪比分类阈值为 R_0，即 $R \geq R_0$ 作为外显子的判据，$R < R_0$ 则作为内含子的判据。通过阈值判别外显子与内含子的效果可用式(4-30)、式(4-31)指标表示[2]。

$$\text{敏感性：} \quad S_n = \frac{T_P}{T_P + F_N} \tag{4-30}$$

$$\text{专一性：} \quad S_P = \frac{T_N}{T_N + F_P} \tag{4-31}$$

式中：T_P 为被正确判为外显子的个数；T_N 为被正确判为内含子的个数；F_N 表示被错误判为内含子的个数；F_P 表示为被错误判为外显子的个数。进而定义阈值判别的总正确率 A_c 为：

$$A_c = \frac{S_n + S_p}{2} \tag{4-32}$$

建立总正确率最大化模型即有：

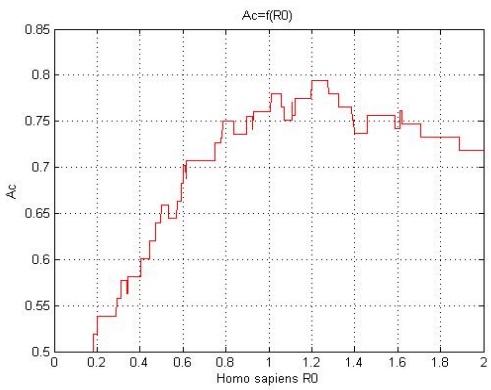
$$\max A_c = f(R_0) \tag{4-33}$$

即首先，建立阈值判别的总正确率 A_c 关于最优阈值 R_0 的函数 $A_c = f(R_0)$，然后针对方法一中得到的外显子、内含子的信噪比组成样本数据集合 S_1、S_2，形成各生物 DNA 序列 $A_c = f(R_0)$ 曲线，找出 $\max A_c$ 所对应的 R_0。优化模型本质也是求得使判别正确率达到最大的阈值解 R_0。

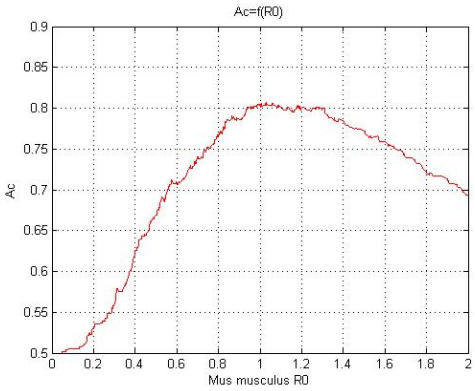
（2） 实验材料：

实验材料同方法一。该方法的实现采用 Matlab/7.0 工具实现，具体源代码详见附录 B。求解得到各种（类）生物的最优信噪比阈值，详见表 4-2，且各种（类）

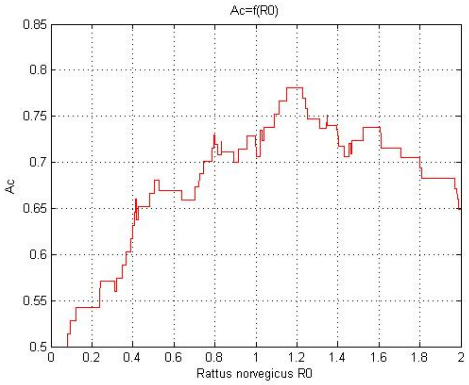
生物 DNA 序列 $A_c = f(R_0)$ 曲线如图 4-3 所示。



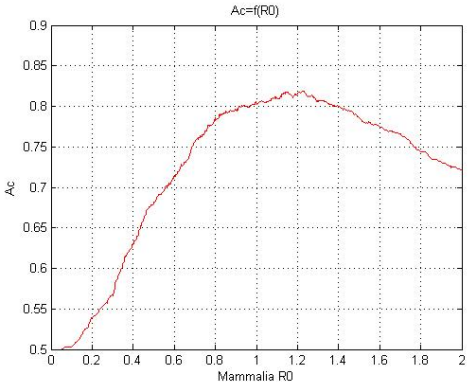
(a) 人



(b) 小家鼠



(c) 褐家鼠



(d) 哺乳动物类

图 4-3 各类生物 DNA 序列 $A_c = f(R_0)$ 曲线

由图 4-3 可见，上述各类生物 DNA 序列的 $A_c = f(R_0)$ 曲线是“凸型”的，即均有一个最大值，该最大值所对应的 R_0 即为最优阈值。且当样本数据较多时，该类 DNA 序列 $A_c = f(R_0)$ 曲线是较为平滑的，如表(b)、(d)所示。

4.2.3 基于信噪比特征的 DNA 识别的效果分析

根据4.2.2节所定义的阈值判别效果评价指标 S_n、S_P、A_c，对按照信噪比特征将编码区和非编码区分类的有效性进行分析，以方法1为例，详见表4-3。

表 4-3 基于方法一最优阈值的判别有效性分析

基因类别	T_P	F_N	S_n	T_N	F_P	S_P	A_c
人	29	6	0.8286	19	7	0.7308	0.7797
小家鼠	289	68	0.8095	218	57	0.7927	0.8011

褐家鼠	33	12	0.7333	29	6	0.8286	0.7809
哺乳动物	624	203	0.7545	605	21	0.9665	0.8605
鼠类	322	80	0.8010	247	63	0.7968	0.7989

注：表中鼠类包括小家鼠和褐家鼠。

基于表 4-3 中对按信噪比将编码与非编码区间分类的有效性分析结果，可以得出以下结论：

(1) 通过表 4-3 可见，样本集合中样本的数目越多，采用阈值确定的最优化方法确定的阈值的判别总正确率 A_c 越高。

(2) 通过比对小家鼠和褐家鼠的分析数据可见：同类生物样本集合中样本数目越多，则专一性 S_P 越大。

(3) 针对某一特定生物（如小家鼠）时，判别结果的敏感性较强，即 S_n 较大，对于褐家鼠较低的原因，可能是由于样本数据较少，导致判别结果中敏感性稍差。

为评价2种阈值确定方法性能，并将其与固定阈值为 2 的方法比较，基于上述4类基因数据的判别正确率 A_c 统计如表4-4、4-5所示。

表 4-4 以 2 信噪比阈值的判别有效性分析

基因类别	T_P	F_N	S_n	T_N	F_P	S_P	A_c
人	18	17	0.5143	24	2	0.923	0.722
小家鼠	142	215	0.3978	272	3	0.9891	0.6934
褐家鼠	16	29	0.3555	33	2	0.9429	0.7187
哺乳动物	372	455	0.4498	622	4	0.9936	0.6492

表 4-5 3 种阈值确定方法的的判别有效性分析

基因类别	最优化方法		经验法(R_0=2)
	方法 1	方法 2	
人	0.7797	0.8192	0.722
小家鼠	0.8011	0.8074	0.6934
褐家鼠	0.7809	0.7945	0.7187
哺乳动物	0.8605	0.781	0.6492

由表4-4可见：基于以R_0=2的经验法进行基因区间分类时，被正确判为外显子的个数较少，即敏感性 S_n 较小，这是由于上述基因类别中外显子信噪比大于2的比重较小，详见图4-1、4-2。尽管如此，以2为信噪比阈值判别的总正确率 A_c 也不低，这是

因为上述4类生物中内含子信噪比较小，几乎全部小于2，进而使得专一性S_P较大。

由表4-5可见：①2种最优化方法的判别正确率均大于以2为信噪比阈值判别的总正确率；②最优化方法所确定的阈值均小于2，这是由上述各类基因内含子、外显子的区间分布决定的；③从数学角度来看，2种最优化方法的本质是一样的，优化模型都是求得使判别正确率达到最大的阈值解R_0，理论上得到的结果也应是相同的，之所以有区别，与样本集合的大小有着直接关系。

4.2.4 基于信噪比特征的 DNA 分类识别错误原因分析

在查阅相关文献的基础上，我们对分类识别时所产生的分类错误进行分析，找出导致错误的可能原因，并对部分原因进行实际验证，得到以下结论：

（1）文献[4,5]指出，当外显子长度较短时，判断正确率较低。我们对 200 个哺乳类动物 DNA 序列中 827 个外显子信噪比进行统计分析，并按照长度将外显子分为 3 类：长度 l≤200bp 为短外显子，200bp<l≤500bp 为中长外显子，l>500bp 为长外显子，用优化方法 1 确定的阈值判别 3 类外显子的正确率统计如表 4-6 所示。

表 4-6 哺乳动物 DNA 序列 3 类外显子阈值判别效果分析

外显子类型	T_P	T_P+F_N	S_n
短外显子	454	645	0.704
中长外显子	140	152	0.921
长外显子	30	30	1

由表 4-6 可见，基于信噪比特征的 DNA 分类识别短外显子（l≤200bp）的正确率是最低的。S_n=0.704<0.8605（表 4-3）。同时也说明：在短编码序列中，3-碱基周期性并不是绝对存在的。

（2）文献[5]研究表明，当编码序列中 A+ T 的含量高于 G+ C 含量时，3-碱基周期性表现不明显，易产生分类错误。

（3）当编码序列中碱基在密码子三个位点上的分布比较均匀时，也易产生分类错误；同时，当编码序列中密码子和氨基酸的使用偏向小时，易产生分类错误。

（4）DNA 序列的碱基组成和分布，所编码蛋白质氨基酸的选用和顺序以及同义密码子的使用都由与 3-碱基周期性有一定的关系，进而也会造成基于信噪比特征的 DNA 分类识别误判。

4.3 问题三

4.3.1 基因识别方法的确定

最常用的方法有固定长度滑动窗口上频谱曲线的基因识别方法。外显子片段具有明显的三周期特性，则计算的 $M/3$ 处的功率谱值较大，利用滑动窗口，当窗口所框碱基片段中外显子占大部分时，功率谱值应较大，因此窗口滑过外显子时会有波

峰出现，则波峰两侧应为外显子的区间。功率谱值上升和下降的过程相对应为窗口滑入和滑出外显子的过程。因此窗口大小的选择，对于外显子的识别也起到很关键的作用。由于时间原因，本组仅根据经验值，及普通外显子片段长度进行了选择。但此法的缺点是由于 DNA 随机噪声的存在，功率谱图两端的识别度很差。

另一种较常用的方法为基于 DNA 序列上“移动序列”信噪比曲线的基因识别方法。由第二问可知，外显子和内含子的信噪比阈值有明显的区别。从 0 开始依次取 3n 个碱基片段，计算片段的信噪比，当所计算片段中，外显子占大部分时，信噪比值会较大，而内含子占大部分时，信噪比值会变小，因此，根据曲线的峰谷变化，可以判别外显子的端点。谷值应对应的外显子的左端点，峰值应对应外显子的右端点。可以用于外显子两端点的判断。

本文将两种方法结合起来，具体步骤如下：

（1） 先由谱功率图判断出外显子的大概区间。

（2） 再由移动信噪比曲线来定两端。对功率谱图确定阈值，做进一步修改。

（3） 确定外显子区间后，查找里面是否有终止子等不应出现的密码子序列。

（4） 由于题中所给出的是完整的基因序列，因此第一段外显子应以启动子开头，而最后一段外显子应由终止子结尾。判断完外显子区段后，可对第一段和最后一段进行检查。精确位置。此步骤可以用 matlab 里自带的生物工具箱的 sequence viewer 实现。

由于 DNA 序列的三周期特性，是由于自然界的蛋白质对氨基酸的使用偏好，造成不同位置上碱基种类的不均衡造成的。是一种统计学的规律。对于较短的基因片段，这种统计学规律表现不明显，本问所探讨的提高基因端点识别的方法，仅适用于碱基数较多的外显子片段（一般大于 200）。

4.3.2 6 组未注释 DNA 序列编码区预测

下面以题目附件中所给的 100 组基因数据中的第 52 号基因为例，对该方法进行验证。图 4-4 和图 4-5 分别为第 52 号基因对应的固定滑动窗口的功率谱图和移动序列的信噪比曲线。其中功率谱图已化为标准图。即均除以谱图中出现的最大幅值。通过步骤的一二部，得到的阈值为 0.1628，确定了外显子位置，如图黑线所示。图中红线表示的为真正的外显子的位置，可以看出，还是有一定偏移的。

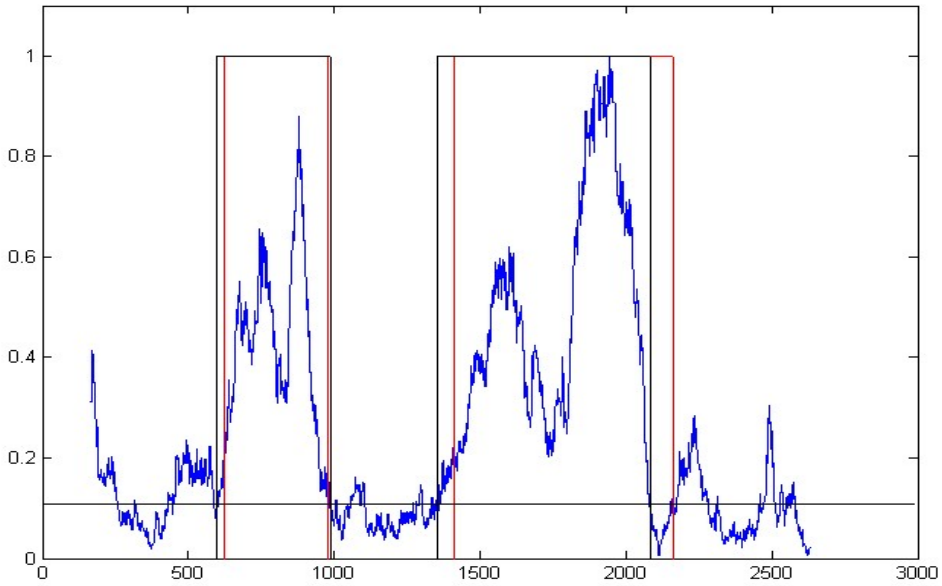


图 4-4　固定长度滑动窗口的频谱 $p = p(n; \frac{M}{3})$ 曲线（mus musculus 基因，AF042783）

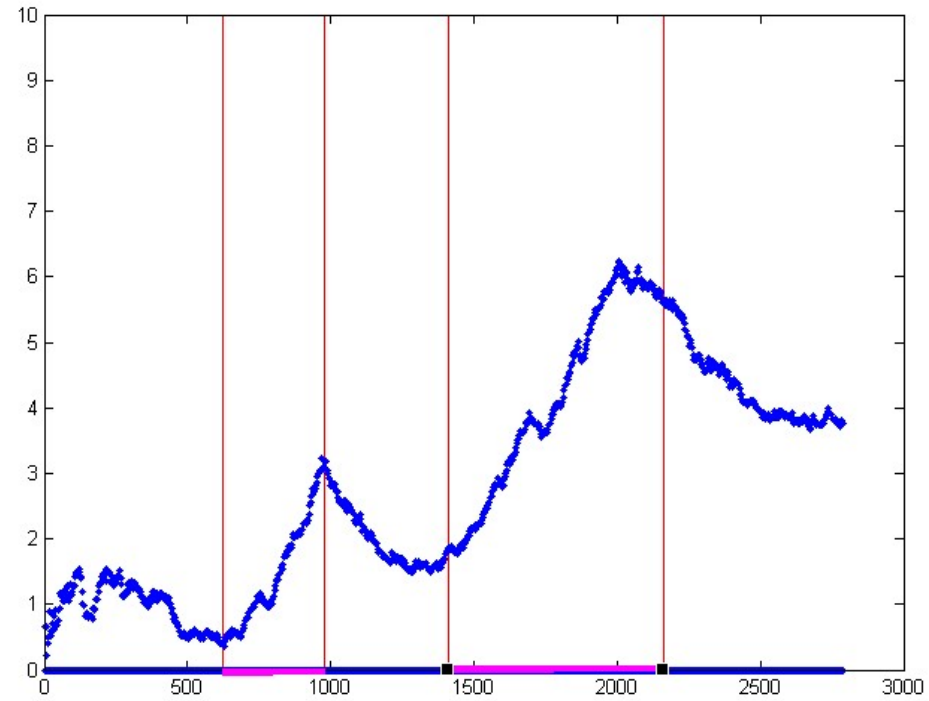


图 4-5 DNA 移动序列其指示序列的信噪比曲线。（mus musculus 基因， AF042783）

由频谱图得出外显子的区段为 596-987，1355-2083，由移动信噪比曲线得出的外显子区段为 627-978，1347-2007。再从 sequence viewer 上看到起始子在 590-627 区间附近，只有 499, 622, 715 三个, 如图 4-6 。

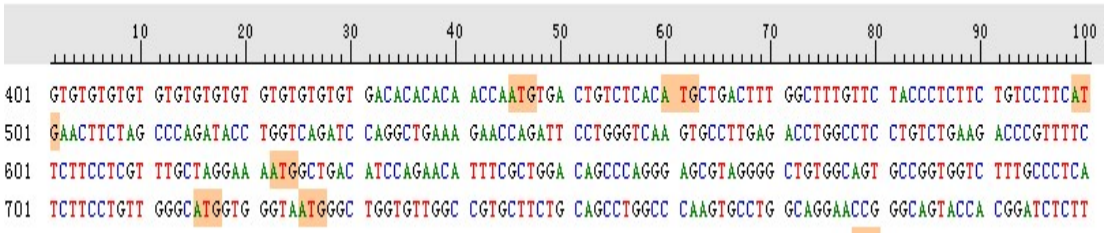
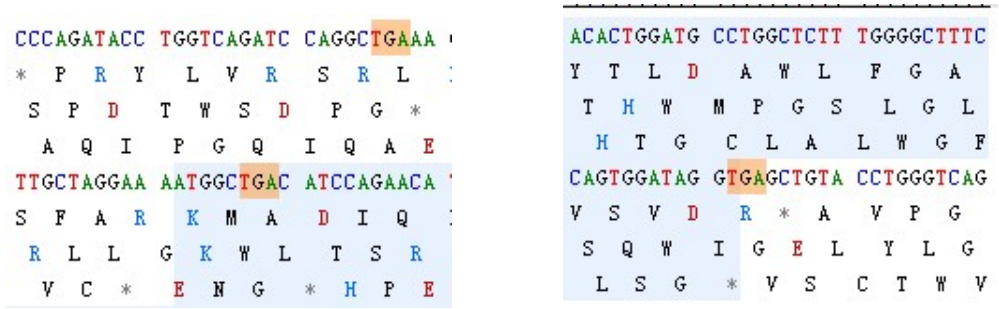


图 4-6 400-701 段启示子出现的位置示意图

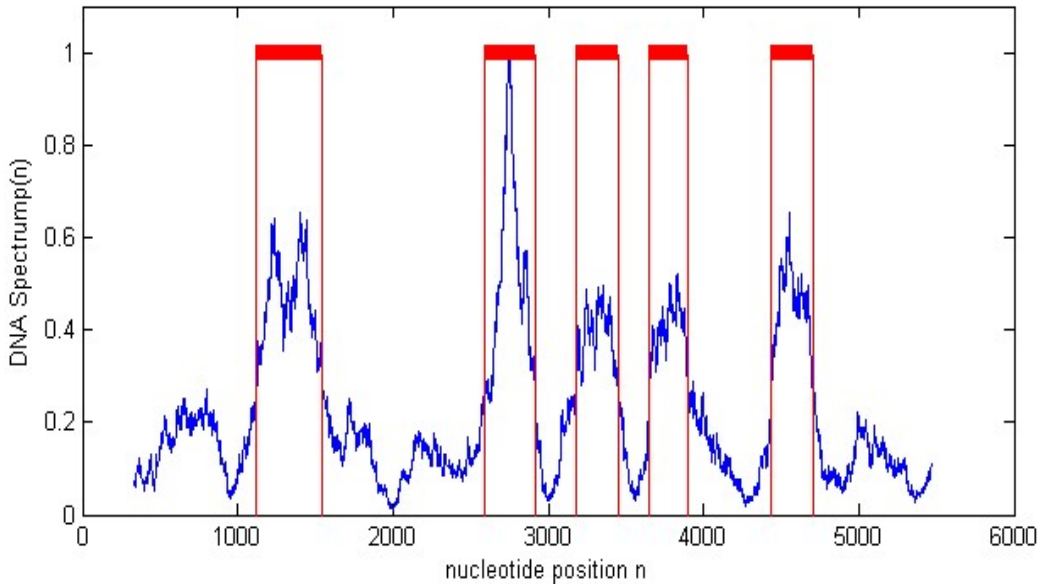
再根据功率谱图和移动信噪比曲线，可知，622 是合适的启动子，则第一段外显子的起始位置应为 622。再从 622 开始查此段区间里的终止子，发现依次编码，在第 982 的位置会出现终止子，如图 4-7 所示。其中（a）中虽出现终止子，但是从起始开始编码，应为 GCT，GAC，并不会表达 TGA。所以第一段外显子的右端点应在 982 之前，则此范围缩减为 978-982。而题目中给出的此段外显子的范围为 622-980.可见在这种情况下，对外显子两端的辨识度还是比较高的。但这仅限于开头和结尾两个外显子，对于中间部分外显子的区分，尽管借助两个图像比对进行，仍会有部分偏移，以及预测失误的地方。特别是长度较短的外显子片段。



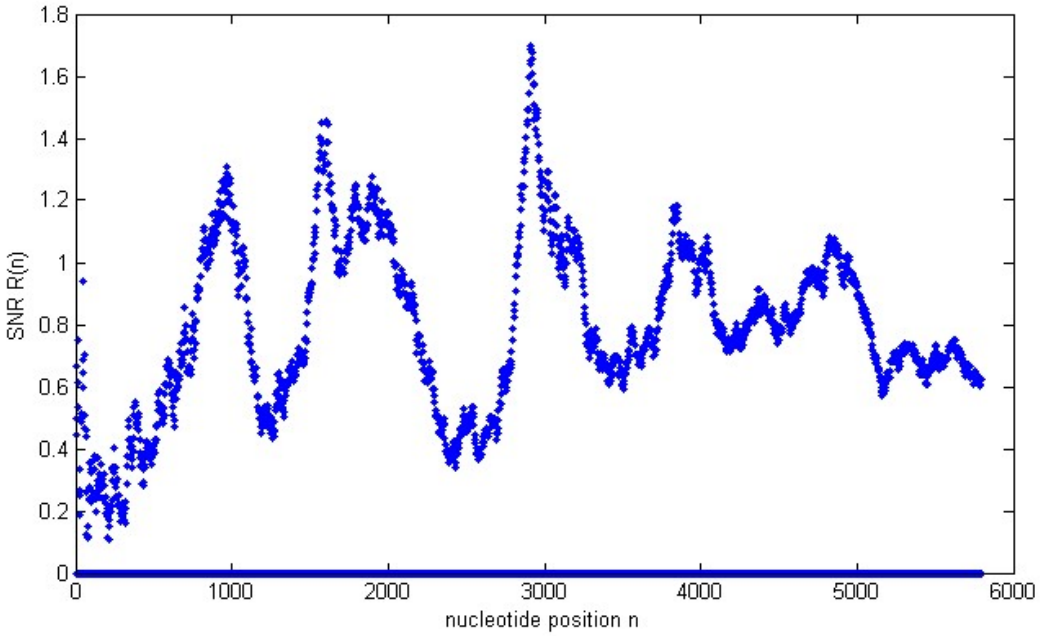
（a）起始端位置示意　　（b）碰到终止子的位置示意图

图 4-7 终止子出现的位置示意图

下面对题中所给的六段基因进行分析，详见图 4-8 至 4-13。

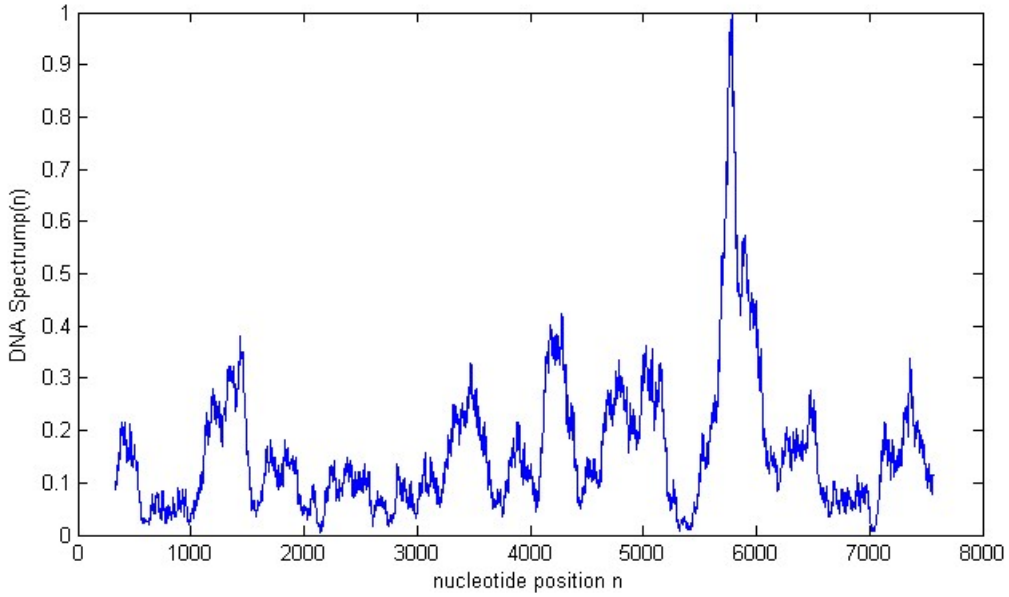


（a） 固定长度滑动窗口的频谱

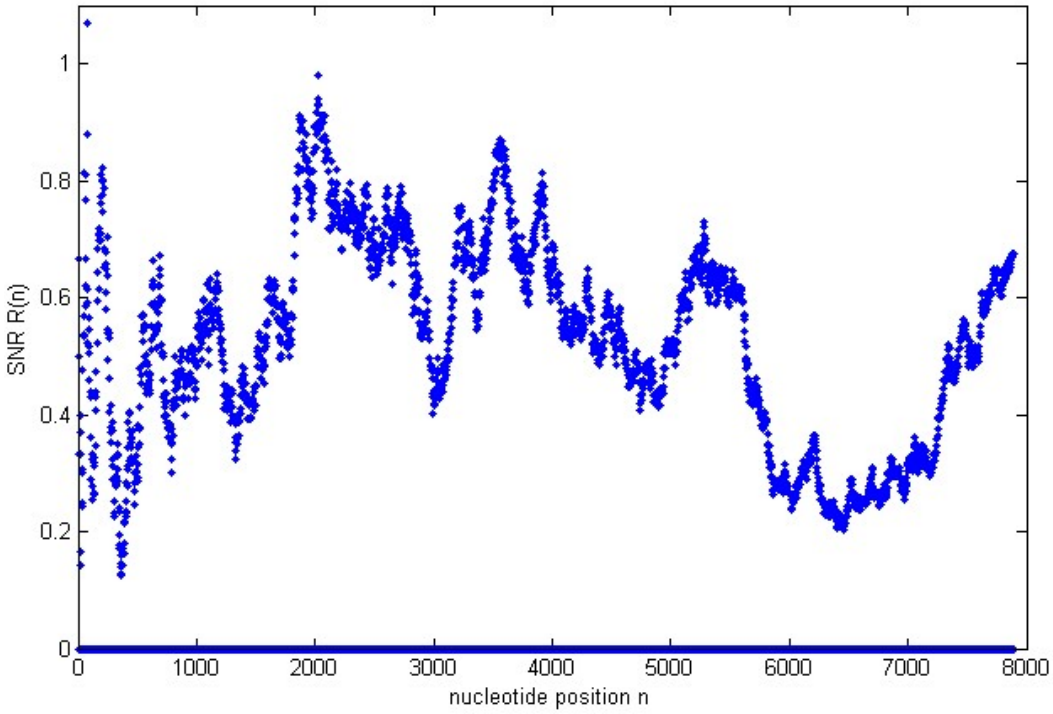


（b）移动序列的信噪比曲线图

图 4-8　Gene1 的预测结果

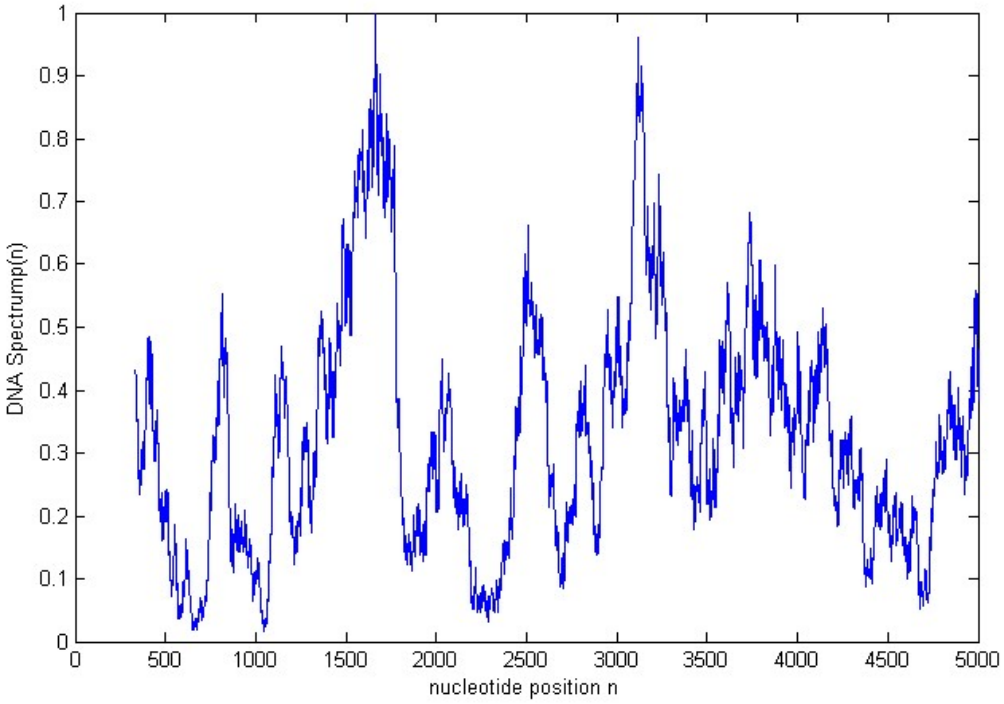


（a） 固定长度滑动窗口的频谱

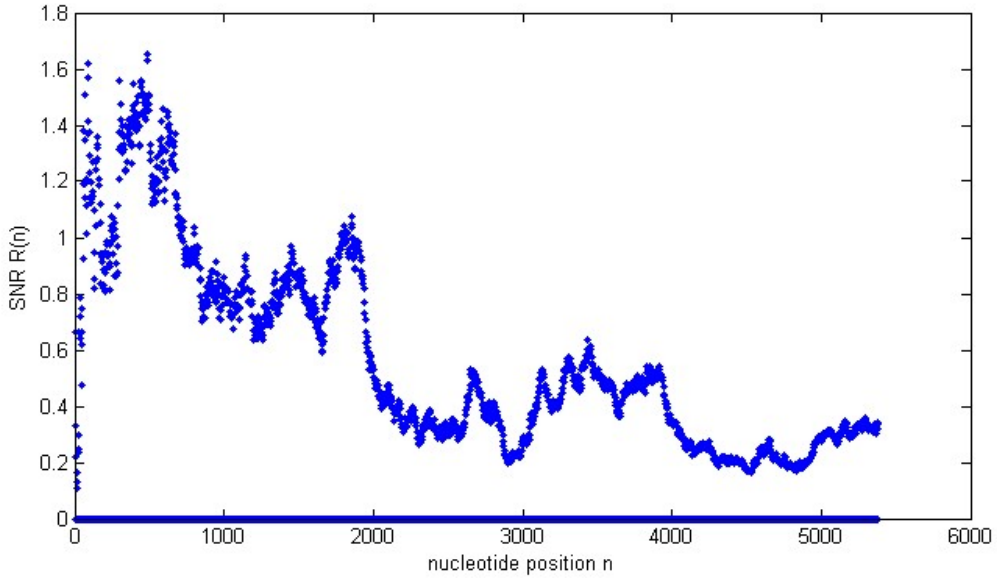


（b）移动序列的信噪比曲线图

图 4-9 Gene2 的预测结果

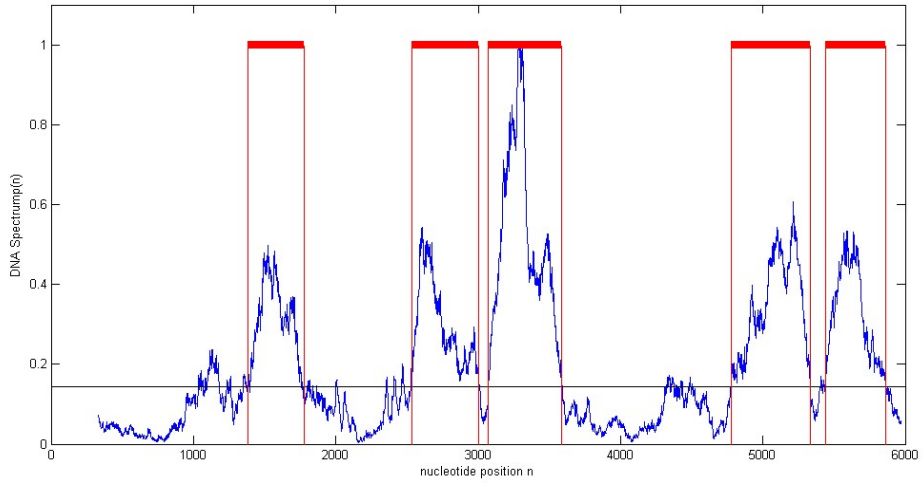


（a） 固定长度滑动窗口的频谱

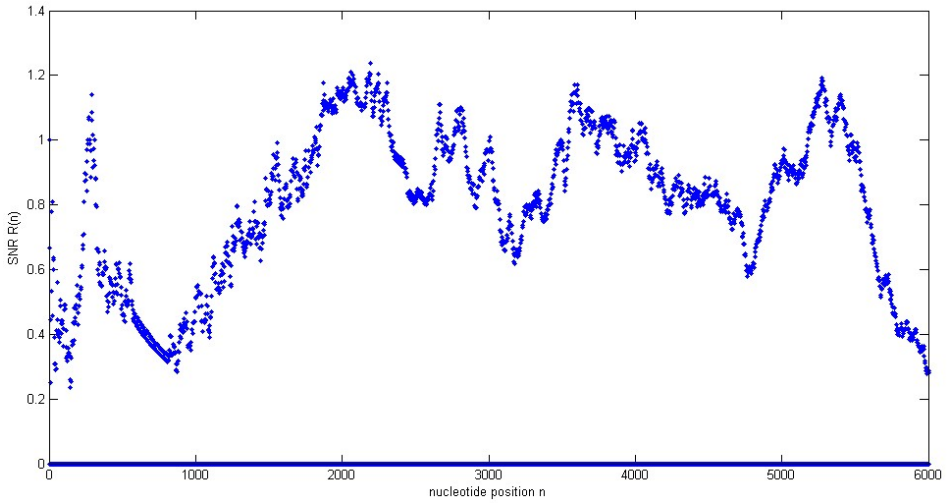


（b）移动序列的信噪比曲线图

图 4-10 Gene3 的预测结果

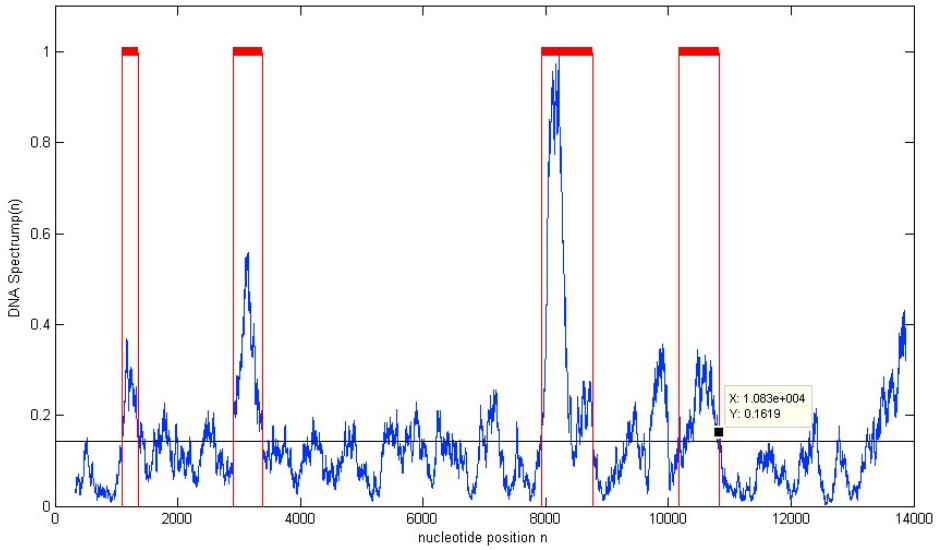


（a） 固定长度滑动窗口的频谱

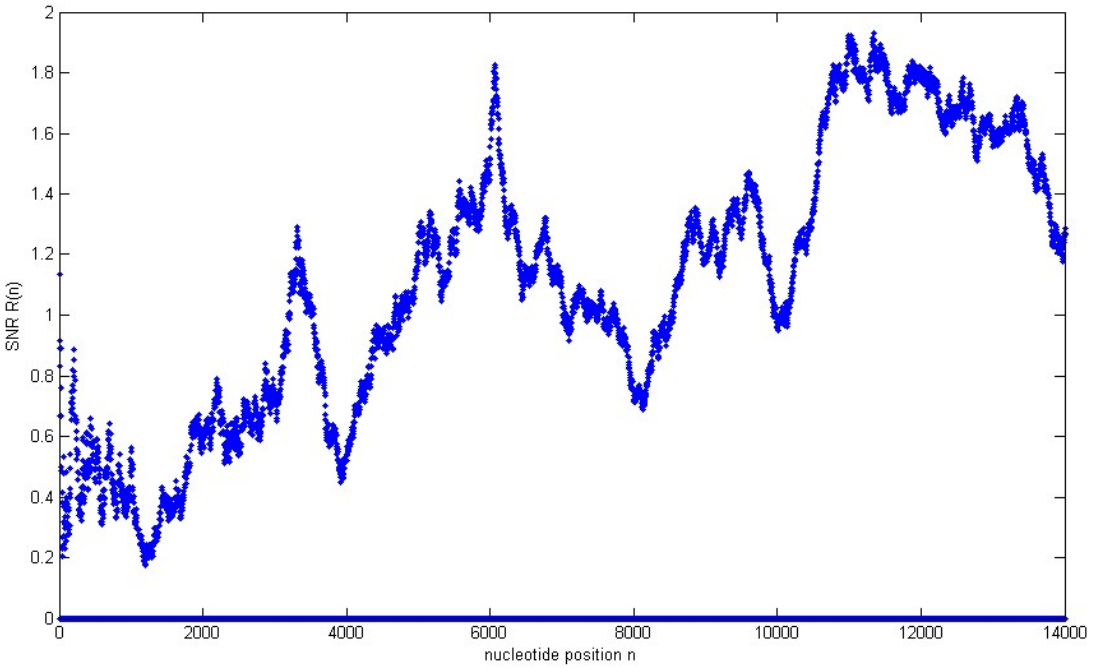


（b）移动序列的信噪比曲线图

图 4-11 Gene4 的预测结果

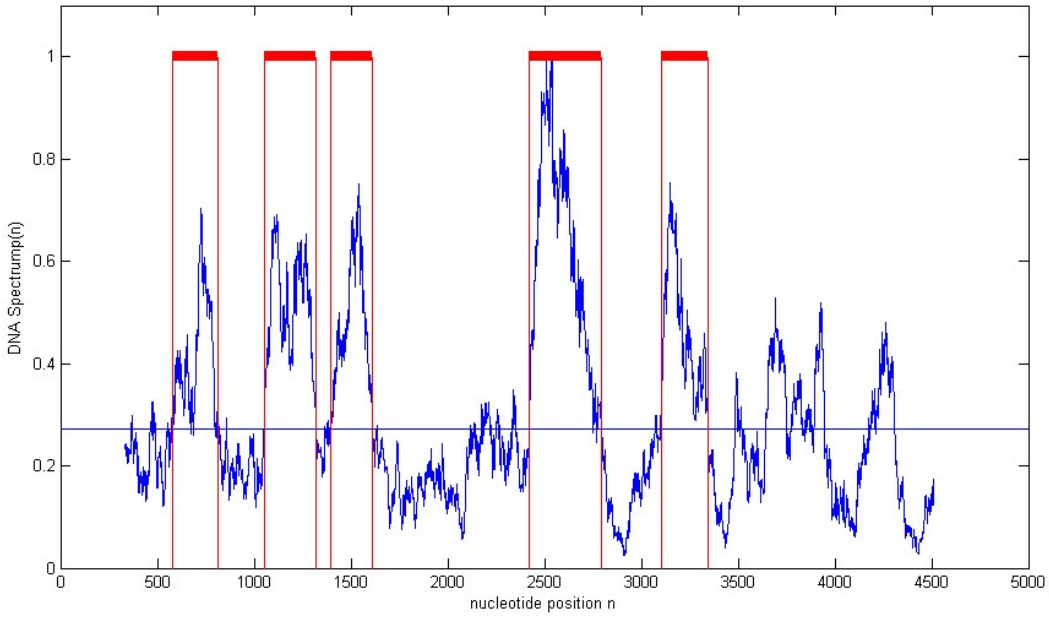


（a） 固定长度滑动窗口的频谱

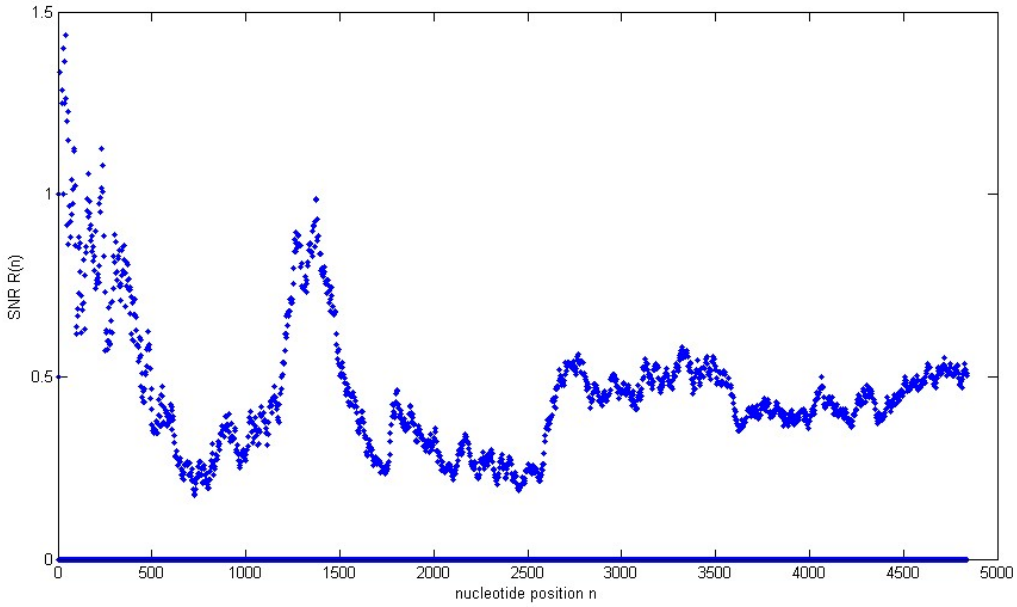


（b）移动序列的信噪比曲线图

图 4-12 Gene5 的预测结果



（a） 固定长度滑动窗口的频谱



（b）移动序列的信噪比曲线图

图 4-13 Gene6 的预测结果

依本文方法，对此六个未知基因进行识别，其中第 1,4,5,6 基因的辨识度较高，且基因 1 的识别外显子片段为 1109-1547, 2595-2914, 3175-3453, 3651-3896, 4439-4706，基因 4 的识别外显子的片段为 1385-1778, 2536-3003, 3071-3584, 4775-5332, 5441-5861.基因 5 识别外显子片段为 1099-1367, 2903-3380, 7931-8767, 10180-10830，基因 6 识别外显子片段为 576-816， 1051-1319, 1395-1610, 2421-2795, 3103-3343。但是也不是所有功率谱确定区段在移动序列额信噪比上均能有正确的对应，比如基因 4 的最后一个外显子片段，有可能是因为 2,3 中的外显子片段相对较短，所以不能有效辨识。在基因一中初始识别第一段基因为 1123，根据 sequence viewer，修正为 1109。由于篇幅问题，其他修正均不一一指出。

4.4 问题四

4.4.1 识别基因编码序列的其它特征指数

采用频谱或信噪比这样单一的判别特征，是影响、限制基因识别正确率的一个重要原因。为了更加准确的预测，尽量避免遗漏，可采用除频谱和信噪比以外的特征指数。通过查阅文献，可将其他指标总结如下：

（1） “非 3-碱基周期性”编码序列指标

根据文献[5]可知，在短编码序列中，3-碱基周期性并不是绝对存在的。这与序列的碱基组成和分布、所编码蛋白质氨基酸的选用和顺序以及同义密码子的使用都有一定的关系。一般的，当序列中 A+T 含量比较高，碱基在密码子三个位点上的分布比较均匀或密码子和氨基酸的使用偏向比较小时，短编码序列的 3-碱基周期性可能都不太明显，容易造成误判。我们已在 4.2.4 节中所得分析结果（表 4-6）是一个很好的验证。

文献[5]收集了 25 个非 3-碱基周期性编码序列，这可以为我们正确判断外显子，减小误判提供一定的依据。

（2） 旋转矢量指标

由文献[6]可知，可通过光谱旋转的方法识别编码序列。文献表明，通过对啤酒酵母细胞所有基因的研究发现：对于四种核氨酸而言，在编码区，各项的分布曲线是一个围绕着中心值的钟形曲线；而在非编码区，各项的分布趋于一致。这个结论同样适用于很多种生物。很多编码区和非编码区的判别方法是基于此相位特性。

文献[6]提出的方法是基于顺时针旋转的矢量，这些矢量是通过 DFT 运算得到。在编码区，所有的矢量在复平面上紧密排列，因此，矢量和的大小就会增大。而在非编码区，此项操作并不能明显增大幅值。可利用此方法，来鉴别内含子与外显子。

（3） 非均匀指标及干涉指标

由文献[7]可知，可根据外显子在密码子三个位点上的分布不均匀性引入非均匀指标 HI，来研究基因外显子与内含子的序列特征。通过推广 HI 参数，定义干涉指标 R，寻找外显子和内含子片段组合的特性，为新基因的预测奠定基础。

HI 的表达式如下：

$$HI=\sum_{l=1}^{3}\sum_{\alpha}^{4}\frac{\left(p_{\alpha}^{(l)}-p_{\alpha}\right)^{2}}{p_{\alpha}}\bullet\frac{N}{3} \tag{4-34}$$

其中，$p_{\alpha}\left(\alpha=1,2,3,4\right)$ 表示序列中 4 种碱基的概率，N 为序列长度，$p_{\alpha}^{(l)}$ 为第 l 个子序列中第 α 种碱基的概率。

定义干涉指标 R 的定义如下：

$$HI=\frac{\sum_{i=1}^{n}\sum_{l=1}^{3}\sum_{\alpha}^{4}\frac{\left(p_{i\alpha}^{(l)}-p_{i\alpha}\right)^{2}}{p_{i\alpha}}\bullet\frac{N_i}{3}}{\sum_{l=1}^{3}\sum_{\alpha}^{4}\frac{\left(p_{\alpha}^{(l)}-p_{\alpha}\right)^{2}}{p_{\alpha}}\bullet\frac{N}{3}} \tag{4-35}$$

这里 N_i 表示第 i 个子序列的长度，$p_{i\alpha}(\alpha=1,2,3,4)$ 表示第 i 个子序列中 4 种碱基的概率。其中 $i=1,2,\cdots,n, N=N_1+N_2+\cdots N_n$。

由文献[7]可以得到：

i）非均匀指数 HI 对于区别外显子和内含子是一个有效而简单的参数

ii）外显子与内含子的干涉指标 R 值分布有明显的区别，且随着 n 的增加，区别越来越明显。同时可以看到，对于相同的片段数 $n(n=n_e+n_o)$，R 值分布图中最概然分布所对应的干涉指标 $R(\max)$ 在 n_o 大时随 n_e 或 n_o 的变化灵敏与 n_e 大时随 n_e 或 n_o 的变化。

iii）无论是由计算方法或理论预测方法得到的异类拼接的组合中 n 值和 $R(\max)$ 的关系都可以看到，外显子片段数 n_e 和内含子片段数 n_o 都对 $R(\max)$ 有影响，后者影响较大，我们可以利用这一性质来区别外显子和内含子，对基因识别做进一步研究。

参考文献

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

[2] Jianfeng， Shao· Xiaohua， Yan · Shuo Shao，SNR of DNA sequences mapped by general affine transformations of the indicator sequences，Mathematical Biology ，DOI 10.1007/s00285-012-0564-3。

[3] 徐尚蕾，Bootstrap 算法在基因预测中的阈值选取研究，电子科技大学硕士学位论文，2011。

[4] 邵建峰，严晓华，DNA 序列信号 3 周期特性， 南京工业大学学报（自然科学版）第 34 卷，第 4 期：133-137，2012。

[5] 张静，石秀凡，不具有 3-碱基周期性的编码序列初探，生物化学与生物物理进展，第 29 卷，第 2 期：267-272，2002。

[6] Daniel Kotlar ，Yizhar Lavner ，Gene Prediction by Spectral Rotation Measure A New Method for Identifying Protein-Coding Regions，Genome Research，Cold Spring Harbor Laboratory Press ISSN：1930-1937，2003。

[7] 张利绒，罗辽复，线虫基因组外显子与内含子序列特征研究， 内蒙古大学学报自然科学版，第 33 卷，第 4 期：401-406，2002。