安徽农业大学学报, 2014, 41(3): 351-357
Journal of Anhui Agricultural University
[DOI] 10.13610/j.cnki.1672-352x.20140423.032　网络出版时间：2014-4-23　15:08:24
[URL]http://www.cnki.net/kcms/doi/10.13610/j.cnki.1672-352x.20140423.032.html

基于 3-周期性的基因预测

胡连果，李　强，张　瑾，张良云，杨　涛，骈　聪
(南京农业大学理学院，南京 210095)

摘　要：基因预测是 DNA 序列分析的一项重要任务，真核生物基因外显子序列的识别是生物信息学中的难点，本研究提出一种基于 3-周期性小鼠基因的预测方法。基于碱基幅角的功率谱，使小鼠外显子序列的频谱图具有更显著的 3-周期性，且内含子的频谱图更加平稳，从而增大了序列中外显子与内含子的信噪比的区分度，使小鼠基因序列的预测获得更好的效果，在平均值法阈值 R_2 下预测准确率达 94.53%；在分布图法阈值 R_1 下，准确率达 94.50%，敏感性达 99.45%。

关键词：功率谱；3-周期性；幅角；信噪比；阈值
中图分类号：Q811.4　　**文献标识码：**A　　**文章编号：**1672−352X (2014)03−0351−07

An approach to gene prediction based on 3-periodicity

HU Lianguo, LI Qiang, ZHANG Jin, ZHANG Liangyun, YANG Tao, PIAN Cong
(College of Science, Nanjing Agricultural University, Nanjing 210095)

Abstract: Gene prediction is a significant task, in which the recognization of exon of eukaryotic gene is a great challenge. In this study, a gene prediction method which is based on 3-periodicity in gene sequences was performed. A more prominent peak value in exon and more steady values in intron were shaped in the power spectrum curves based on arguments of the four bases. The distinction degree of SNR of exons and introns was enlarged and a better effect in short gene prediction was presented. In the condition of threshold R_2, the accuracy reached to 94.53%; under the threshold R_1, the accuracy reached to 94.50%, and the sensibility reached to 99.45%.

Key words: power spectrum; 3-periodicity; argument; SNR; threshold

自人类基因组计划的展开和各种生物体测序的完成，生命科学研究进入了一个全新的阶段，对于基因组的碱基序列进行测定，破解遗传信息是一项艰巨而又迫切的任务。通过基因识别技术对 DNA 序列进行分析并找到编码蛋白质的区域是破解遗传信息的关键问题。科学研究表明，通过数学模型和算法进行基因识别省时高效，是其他方法所不能比拟的。由于原核生物结构简单，研究已较为深入，而真核生物的编码区是不连续的片段，结构复杂，尚无优良的基因识别算法。

预测基因编码区的方法有很多，大致可以分为两类：①基于统计特征：密码子含量，核苷酸频率，G/C 和 A/T 含量，碱基的位置频率分布，密码子结构因子等等；②基于基因序列碱基的 3-周期性，所谓碱基的 3-周期性，即若对基因序列进行数值化映射和傅里叶变换后得到功率谱序列，基因外显子序列的功率谱序列的频谱图在 1/3 处具有较大的峰值，而内含子序列却没有类似的峰值，把这种统计现象叫做碱基的 3-周期性[1]。这两类基因编码区预测方法不是绝对独立的，在对基因识别过程当中各种因素相互影响。基于统计特征进行基因预测的方法在处理过程中，需要大量数据样本作为训练集，而基于 3-周期性的预测却不需要。基于碱基 3-周期性的基因预测，已有大量研究，一方面对数值化映射进行研究，使得到的数值序列带有更多自身的生物信息，目前已经有较多映射方法，如 Z-Curve 映

收稿日期：2013-12-27
基金项目：国家自然科学基金（11226070）资助。
作者简介：胡连果，硕士研究生。
* **通信作者：**李　强，副教授。E-mail：lq@njau.edu.cn

射[2]、正四面体法映射[3]、复数法映射[4]、实数法[5]等;一方面是对功率谱序列和阈值的选取进行研究，Silverman 和 Linsker[3]提出了最初的功率谱序列，Fickett 和 Tung [6],Tiwari [7],Anastassiou [4]等提出基于不同特征的功率谱序列。

作者是结合外显子序列的碱基幅角分布的统计特征和功率谱序列及阈值的选取进行研究，对小鼠外显子序列进行识别预测，改善了传统方法上对真核生物中较短的基因预测 3-周期性不明显、预测的误差较大的缺点；本文提出的基于碱基幅角的功率谱和信噪比对测试集中平均长度为 150 bp 的小鼠外显子序列的预测取得很好的效果。在基因预测中使较短外显子序列的 3-周期性更显著，进而提高基因的预测的准确率，进一步推动后续工作，如：基因和蛋白质功能分析，基因克隆等。因此有必要研究新的算法来提高准确率。

1 数据和基因预测方法

1.1 数据

数据来源于 http://bpg.utoledo.edu/~afedorov/lab/eid.html，取小鼠基因中 400 个外显子序列(外显子序列的平均长度为 240 pb)和 470 个内含子序列作为训练集，取 2000 个外显子序列(外显子序列的平均长度为 150 pb)和 2000 个内含子序列作为测试集。

1.2 基因序列的映射、功率谱和信噪比

对基因序列进行数值化映射，令 $I=\{A,T,G,C\}$，现对于任意确定的 $b\in I$

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

其中 $S[n]$是所取的长度为 N 的碱基序列，$n=0,1,2,\cdots,N-1$。映射(1)称为 Voss 映射[8][13-14]，生成相应的 4 个二进制序列$\{u_b[n]\}$：$u_b[0]$，$u_b[1]$，…，$u_b[N-1]$，$b\in I$ 称为 DNA 序列的指示序列。如：给定一段 DNA 序列为 S=ATCTCACTGGT，则 u_A={1,0,0,0,0,1,0,0,0,0,0}，u_T={0,1,0,1,0,0,0,1,0,0,1}，u_G={0,0,0,0,0,0,0,0,0,1,1,0}，u_C={0,0,1,0,1,0,1,0,0,0,0}.

对指示序列分别做离散 *Fourier* 变换(*DFT*)[9]

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

得到 4 个长度均为 N 的复数序列$\{U_b[k]\}$，$b\in I$。

复序列$\{U_b[k]\}$的二范数之和，得到整个 DNA 序列 S 的功率谱序列$\{P[k]\}$ [10]：

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

每给定一段 DNA 序列，通过(3)式作出其对应功率谱曲线，图 1 和图 2 分别为小鼠基因中某个外显子序列和内含子序列的功率谱曲线(或频谱图)。

外显子序列的功率谱在三分之一处有比较大的峰值，内含子序列的频谱在三分之一处没有这个特征，故把

$$R=\frac{P(\frac{N}{3})}{\overline{E}} \qquad (4)$$

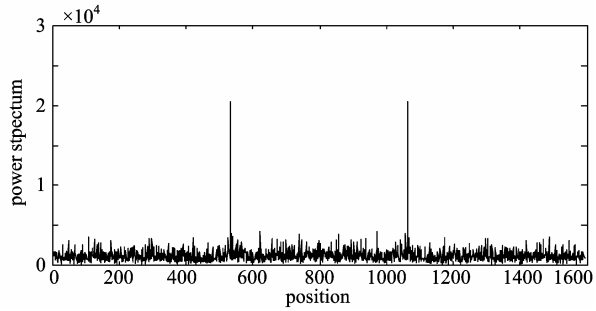

图 1 小鼠编号为 AF019045 中基因序列的一段外显子序列（区间为[634,2226]，长 1593 bp）通过(3)式得到的频谱图

Figure 1 Power spectrum curve of Mus musculus exon numbered AF019045 which is from 634 bp to 2226 bp under equation (3)

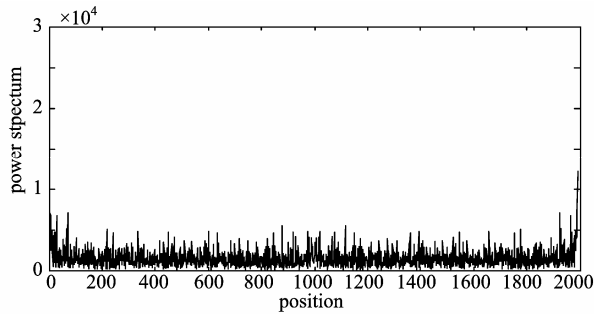

图 2 小鼠编号为 AF019045 中基因序列的一段内含子区序列（区间为[2227,4218]，长 1992 bp）通过(3)式得到的频谱图

Figure 2 Power spectrum curve of Mus musculus intron numbered AF019045 which is from 2227 bp to 4218 bp under equation (3)

称为 DNA 序列信噪比[3]，其中 $\overline{E}=(\sum_{k=0}^{N-1}P[k])/N$ 是给定 DNA 序列的频谱均值。

信噪比值的大小是给定 DNA 序列在 N/3 处的频谱峰值的大小的表征，即外显子或内含子的 3-周期性的强弱，显然外显子序列的信噪比值较大，而内含子的信噪比值相对较小。对于一段 DNA 序列，选取一个最优的值 R_0，使尽可能多的外显子的功率谱或信噪比大于 R_0，而内含子的功率谱或信噪比小于 R_0。根据这个特征选定某个适当的阈值，通过信噪比值和阈值的大小来判断是否是外显子或内含子序列，若信噪比大于阈值则判断为外显子序列，若

信噪比小于阈值则判断为内含子序列。预测的准确率关键在于：①使外显子和内含子的信噪比值的分布区分度尽量大，即使外显子的 3-周期性更明显或内含子的信噪比值更小；②选择一个恰当的阈值也能使判断外显子和内含子的准确率更高。

对 DNA 序列做(1)、(2)和(3)式变换后，得到此 DNA 序列的频谱图，但是很多真核生物的外显子序列都比较短（本研究的测试集中小鼠基因的外显子序列的平均长度为 150 pb），3-周期性不明显，此时得到的外显子和内含子的信噪比值区分度不明显，外显子误判的误差增大，这就导致预测准确率较低。

计算基因序列的信噪比，首先计算，$P[\frac{N}{3}]=\sum_{b\in I}\left|U_b(\frac{N}{3})\right|^2$ 即频谱的峰值关键在于 $U_b\left(\frac{N}{3}\right)$，$b\in I$ 而

$$U_b\left(\frac{N}{3}\right)=\sum_{n=0}^{N-1}u_n e^{-jn\frac{2\pi}{3}}=f(b,1)\times 1+f(b,2)e^{-j\frac{2\pi}{3}}+f(b,3)e^{j\frac{2\pi}{3}} \quad (5)$$

其中 $f(b,i)$是碱基 b 在密码子第 i 个位置上出现的频数。

又因为 $1+e^{-j\frac{2\pi}{3}}+e^{j\frac{2\pi}{3}}=0$，故(5)式可变换为

$$Ub\left(\frac{N}{3}\right)=[f(b,1)-f_{\min}]+[f(b,2)-f_{\min}]e^{-j\frac{2\pi}{3}}+[f(b,3)-f_{\min}]e^{j\frac{2\pi}{3}} \quad (6)$$

其中 $f\min=\min_{i=1}^{3}\{f(b,i)\}$ [11]。

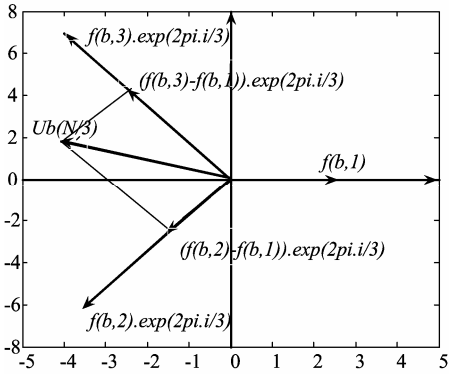

图 3　当 f_{min}=$f(b,1)$时，在复平面上求 $U_b\left(\frac{N}{3}\right)$的示意图

Figure 3　The vector of $U_b\left(\frac{N}{3}\right)$ when $f_{\min}$=$f(b,1)$

取 $f_{\min}$=$f(b,1)$为例，$U_b\left(\frac{N}{3}\right)$，$b\in I$ 在复平面上运算如图 3，可以看到 $U_b\left(\frac{N}{3}\right)$，$b\in I$ 的取值大小和 $f(b,1)$，$f(b,2)$，$f(b,3)$的大小相关，$U_b\left(\frac{N}{3}\right)$所在的复平面的位置即其所对应的幅角也和 $f(b,1)$，$f(b,2)$，$f(b,3)$相关。

当 $f_{\min}$=$f(b,1)$时，$U_b\left(\frac{N}{3}\right)$的幅角

$$\arg\left[Ub(\frac{N}{3})\right]=art\cot\left[\frac{2(f_1-f_{\min})-(f_2-f_{\min})}{\sqrt{3}(f_2-f_{\min})}\right]+\Phi \quad (7)$$

此时 f_1=$f(b,3)$，f_2=$f(b,2)$，$\varphi=\frac{2\pi}{3}$；相应地当 $f_{\min}$=$f(b,2)$时，f_1=$f(b,3)$，f_2=$f(b,1)$，$\varphi=-\frac{2\pi}{3}$；当 $f_{\min}$=$f(b,3)$时，f_1=$f(b,1)$，f_2=$f(b,2)$，ϕ=0。Daniel 和 Yizhar [11]发现酵母菌的外显子和内含子序列的每个碱基对应的幅角 arg[A], arg[T], arg[G] ,arg[C]的值得分布不同，外显子序列的碱基幅角分布呈钟形，本文对小鼠基因研究发现小鼠基因的外显子碱基幅角也有类似分布。这样把

$$P'(s)=\left|\frac{e^{-j\mu A}}{\sigma_A}U_A(s)+\frac{e^{-j\mu T}}{\sigma_T}U_T(s)+\frac{e^{-j\mu C}}{\sigma_C}U_C(s)+\frac{e^{-j\mu G}}{\sigma_G}U_G(s)\right|^2 \quad (8)$$

作为本文小鼠的基因序列的功率谱序列，其中 s=0,1,2,…,N-1，而 μ_b、σ_b（$b\in I$）是所取训练集中外显子对应碱基的幅角的均值和方差。

小鼠的基因序列的信噪比的选取仍然是(4)式的形式

$$R'=\frac{P'(\frac{N}{3})}{\overline{E}'} \quad (9)$$

但是功率谱选取(8)式的计算方式，其中 $\overline{E}'=(\sum_{s=0}^{N-1}P'[s])/N$。

1.3　阈值的选取

对任何物种的基因序列，经过分析计算得到其功率谱频谱图或者信噪比的曲线图后，如何区分外显子和内含子，需要一个评判标准，选取一个最优的值 R_0，我们把这个值作为基因预测中的阈值。要预测基因的外显子和内含子序列，选取恰当的最优的阈值也是非常重要的，这将直接影响基因预测的准确率。若所取的一段 DNA 序列的信噪比大于或等于所取的阈值，则判断这段 DNA 序列为外显子序列；若这段 DNA 序列的信噪比小于所取的阈值，则判断这段 DNA 序列为内含子序列。

3 种阈值的选取方法如下：

①分布图法阈值 R_1，训练集的外显子和内含子序列的信噪比分布直方图中，取两峰之间的谷点作为阈值；

②平均值法阈值 R_2，把外显子的信噪比均值 $m_{外}$和内含子的信噪比均值 $m_{内}$的算术平均值作为阈

值 R_2= ($m_{外}$+$m_{内}$) /2 [12]；

③加权平均值法阈值 R_3，把外显子的信噪比均值 $m_{外}$及其方差 $\sigma_{外}$的积，内含子的信噪比均值 $m_{内}$及其方差 $\sigma_{内}$的积，再除以两方差之和，即为 R_3=($m_{外}\sigma_{外}$+$m_{内}\sigma_{内}$)/($\sigma_{外}$+$\sigma_{内}$) [12]。

2 小鼠基因预测与评价

2.1 小鼠基因的外显子和内含子的幅角分布

统计发现小鼠基因的外显子和内含子序列的每个碱基对应的幅角 arg[A], arg[T], arg[G] ,arg[C]的值得分布不同，外显子序列的碱基幅角分布呈现钟形。本文所取训练集 400 个外显子序列和 470 个内含子序列的 arg[A], arg[T], arg[G] ,arg[C]值的分布情况如图 4 和图 5。

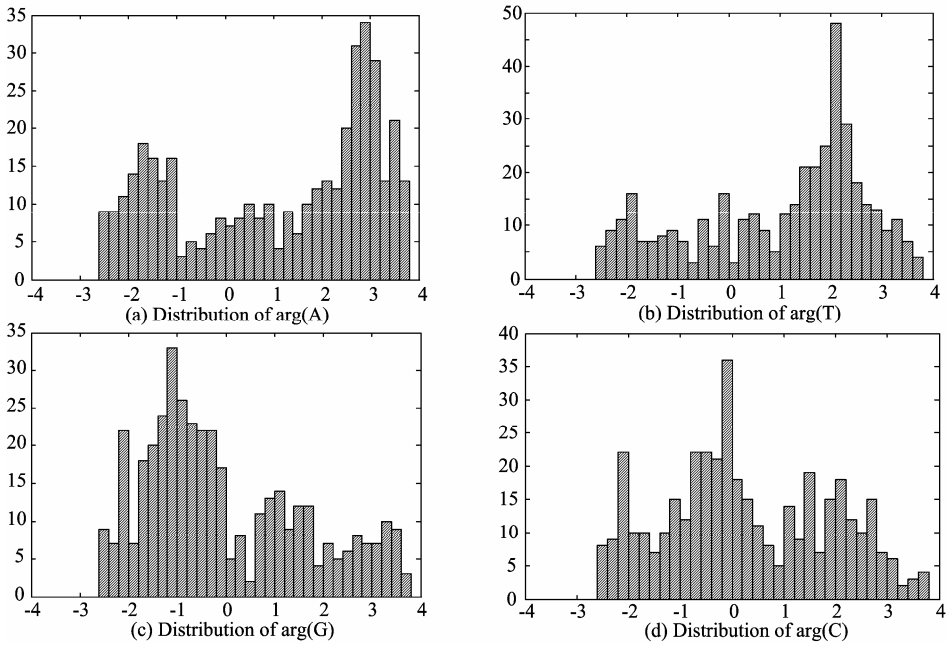

图 4 小鼠的外显子的四个碱基的幅角的分布图

Figure 4 Four bases argument distribution of Mus musculus exon

图 4 中，(a)中 arg[A]的均值为 μ_A=1.0884，方差为 σ_A=1.9943；(b)中 arg[T]的均值为 μ_T=1.0884，方差为 σ_T=1.6764；(c)中 arg[G]的均值为 μ_G=-0.0164，方差为 σ_G=1.6521；(d)中 arg[C]的均值为 μ_C=0.2778，方差为 σ_C=1.5848。

由图 4 可知，4 个碱基的幅角 arg[A], arg[T], arg[G] ,arg[C]分布都有个集中的峰值，图像呈钟形分布，尤其(b)、(d)图像中的钟形分布更明显。由图 5 可知，小鼠的内含子的 4 个碱基的幅角 arg[A], arg[T], arg[G] ,arg[C]分布没有出现相应的钟形，而是均匀分布在每个区间。(8)式中小鼠的基因序列的功率谱序列是在复数序列(2)的基础上再乘以系数 $e^{-j\mu b}/\sigma_b$，使外显子的 3-周期性更加明显，使外显子的频谱峰值更显著，这样增强外显子序列的信噪比。

2.2 小鼠基因的功率谱图

小鼠基因编号为 AF019045（与图 1 同一段外显子序列和图 2 同一段内含子序列）的功率谱图，如图 6 和图 7 所示，外显子序列在频率 k=N/3 处，有更显著的频谱峰值（频谱峰值和平均值的比值更大，图 1 中外显子的均值和频谱峰值比为 1:16.7，而图 6 中外显子均值和频谱峰值比为 1:41.8），3-周期性更显著，而内含子序列的频谱更加平稳。

2.3 小鼠基因的信噪比分布及预测结果

经过优化处理以后的功率谱序列的 3-周期性会更加显著，可以有效改善一些较短的外显子序列的 3-周期性不够显著的问题，使得外显子的 3-周期性更加明显，而内含子的频谱图更加稳定。故外显子的信噪和内含子的信噪比区分度会更大。本研究测试集中，小鼠的 2000 个外显子序列和 2000 个内含子序列的信噪比统计后的分布图如图 8 和图 9 所示。由图 8 和图 9 可知，所取测试集的小鼠外显子序列中，外显子的信噪比的均值为 $m_{外}$=1.2035，方差为 $\sigma_{外}$=1.8715；小鼠内含子序列中，内含子的信噪比的均值为 $m_{内}$=0.2798，方差为 $\sigma_{内}$= 0.5809。3 种阈值分别为：①分布图法阈值 R_1=0.5，其中有 99.45%的外显子信噪比大于 0.5，有 89.55%内含子的信噪比小于 0.5；②平均值法阈值 R_2= ($m_{外}$+$m_{内}$) /2=0.74；③加权平均值法阈值 R_3=($m_{外}\sigma_{外}$+$m_{内}\sigma_{内}$)/($\sigma_{外}$+$\sigma_{内}$) =1.21。

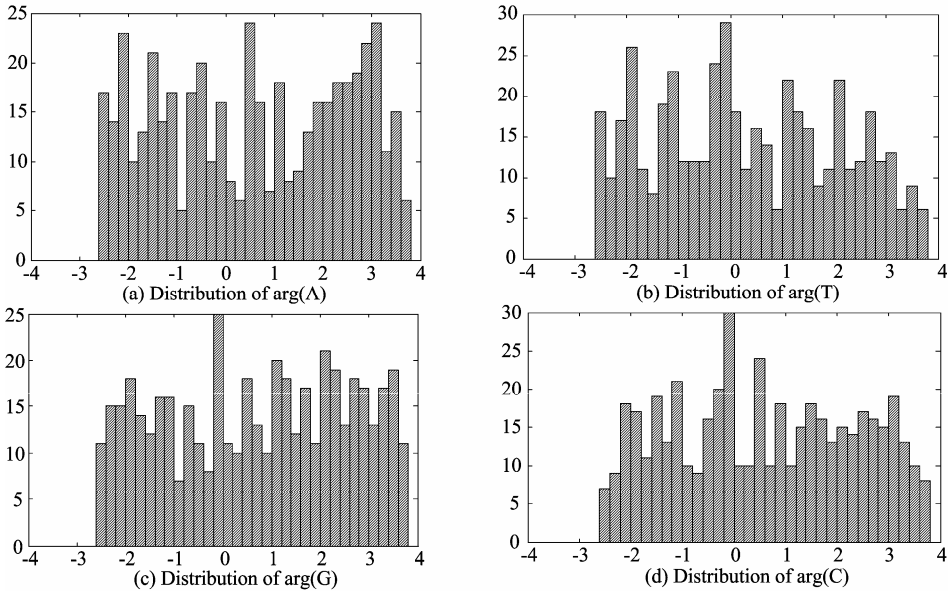

图 5　小鼠的内含子的 4 个碱基的幅角的分布图

Figure 5　Four bases argument distribution of Mus musculus intron

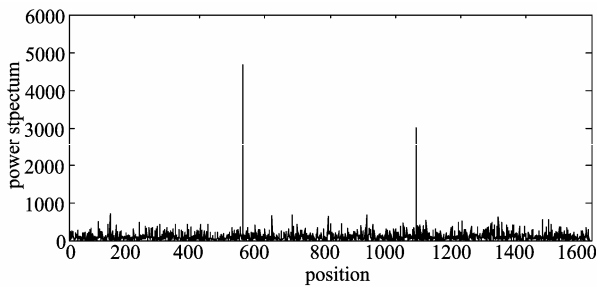

图 6　小白鼠的编号为 AF019045 中基因序列的一段外显子序列（区间为[634,2226]，长 1593 bp）通过(8)式得到的新频谱图

Figure 6　Power spectrum curve of Mus musculus exon numbered AF019045 which is from 634 bp to 2226 bp under equation (8)

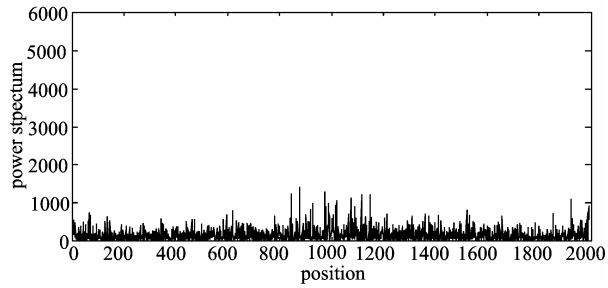

图 7　小白鼠的编号为 AF019045 中基因序列的一段内含子序列（区间为[2227, 4218]，长 1992 bp）通过(8)式得到的新频谱图

Figure 7　Power spectrum curve of Mus musculus intron numbered AF019045 which is from 2227 bp to 4218 bp under equation (8)

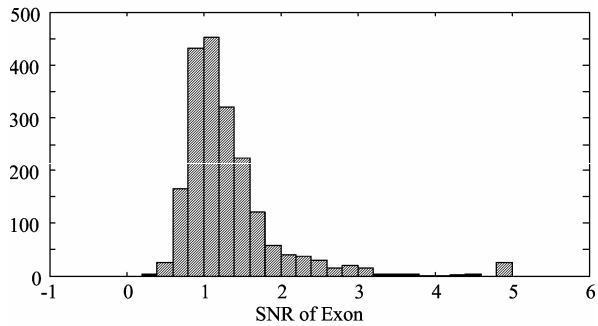

图 8　小鼠的外显子的信噪比分布直方图

Figure 8　SNR distribution of Mus musculus exon

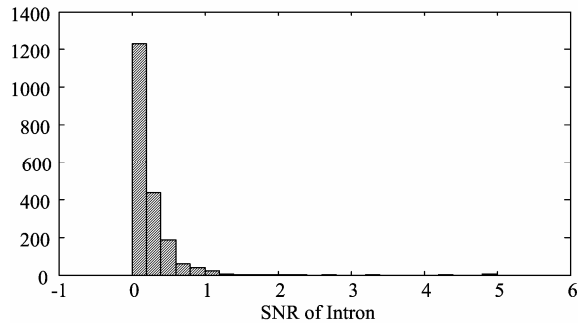

图 9　小鼠的内含子的信噪比分布直方图

Figure 9　SNR distribution of Mus musculus intron

2.4　评价

用敏感性 S_n 和专一性 S_p 分析评价不同阈值的小鼠基因预测，

敏感性　$S_n=T_P/(T_P+F_N)$　　(10)

专一性 $S_p=T_N/(T_N+F_p)$ (11)

其中 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_P 为被错误地判为外显子的个数，F_N 为被错误地判为内含子的个数[1]。

小鼠测试集中，2000 个外显子序列和 2000 个内含子序列用(8)式的功率谱和(9)式的信噪比，对外显子和内含子的信噪比分别用 3 种阈值 R_1, R_2, R_3 判断后，统计 T_P, T_N, F_P, F_N 的个数如表 1 所示。

对小鼠的测试集的基因序列的预测中，3 种阈值对外显子和内含子的判别结果的敏感性(10)和专一性(11)和整体预测的准确率 $A_c=(S_n+ S_p)/2$ 如表 2 所示。

3 种阈值中（表 2），分布图法阈值 R_1 在敏感性 S_n 方面有很好的效果，而专一性 S_p 最好的是加权平均值法阈值 R_3，平均值法阈值 R_2 虽然敏感性 S_n 和专一性 S_p 都不是最好，但其敏感性 S_n=94.10％和专一性 S_P=94.95％都达到了很高的预测值，使在阈值 R_2 的预测下的准确率 A_c=94.53％。加权平均值法阈值 R_3 的专一性虽然达到了 98.70%，但是敏感性 S_n 却只有 44.90%，直接降低了其对应预测的准确率。从基因预测的准确率的角度来看(所取测试集中的外显子序列的平均长度为 150 kp),对于比较短的基因序列，预测的准确率在分布图法阈值 R_1 和平均值法阈值 R_2 都取得了很好的预测效果，当外显子和内含子序列的信噪比分布区分度较大时，分布图法才能取得很好的预测效果；当基因预测其它方面需要较好的专一性时，加权平均值法阈值 R_3 是最适合的阈值。

表 1 3 种阈值判别的结果

Table 1 Results of three thresholds

阈值 Threshold	外显子 Exon		内含子 Intron	
	T_P	F_P	T_N	F_N
R_1=0.50	1989	209	1791	11
R_2=0.74	1882	101	1899	118
R_3=1.21	898	26	1974	1102

表 2 3 种阈值的敏感性、专一性和准确率

Table 2 Sensibility, specificity and accuracy under three thresholds

阈值 Threshold	S_n/%	S_P/%	A_c/%
R_1=0.50	99.45	89.55	94.50
R_2=0.74	94.10	94.95	94.53
R_3=1.21	44.90	98.70	71.80

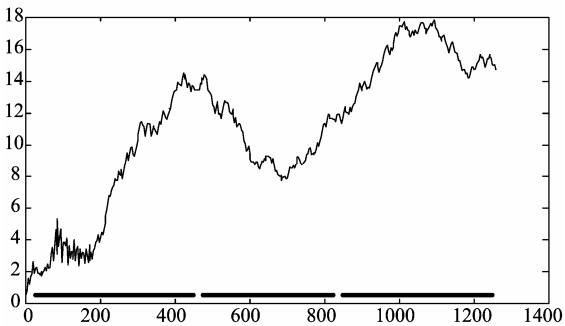

图 10 小鼠编号为'AB010281'的基因移动序列的信噪比曲线（黑色粗线为基因序列实际外显子的位置）

Figure 10 SNR curve of Mus musculus moving sequence numbered AB010281 (the thick black line is the location of Mus musculus exon)

3 小结与讨论

目前对原核生物的基因预测可取得较好的效果，而真核生物基因机构复杂，给精确预测编码区带来困难。用传统功率谱(3)和信噪比(4)预测真核生物的外显子和内含子的效果不佳，特别对于短外显子序列很难被有效预测；且将判别阈值取为经验阈值 R_0=2，带有一定的主观性、经验性，不同的物种基因带有自身的生物性质，所选取的判别阈值也是不同的。

作者统计发现小鼠外显子的碱基幅角分布呈现类似钟形，而内含子碱基幅角趋向于随机分布，在传统功率谱上乘以外显子碱基幅角系数，使小鼠外显子序列的功率谱的 3-周期性更显著，内含子序列的功率谱更平稳，从而使小鼠的外显子序列和内含子序列的信噪比的区分度更大，尤其对较短的 DNA 序列信噪比的区分度更显著，为预测短的 DNA 序列提供了很好的依据，本研究在小鼠的基因预测中取得了很好的效果。取小鼠基因编号为'AB010281'的整段基因序列，用本研究计算信噪比的方法计算该段基因的移动子序列信噪比，得到该段移动序列信噪比的图像如图 10 所示,可看出小鼠基因的移动

序列信噪比曲线的峰、谷区间与外显子区间成明显的对应关系，在小鼠整段基因序列预测中也取得较好的预测效果。

但大部分长度小于 75 bp 的外显子序列没有 3-周期性，利用本研究计算方法仍不能使其有 3-周期性。信噪比是影响、限制基因识别正确率的一个重要原因，但短的外显子序列，可能不具有信噪比显著性，可以试图寻找识别基因外显子序列的其它特征指数。同时本研究提出的直方图法确定阈值，在本文中取得很好的预测效果，但直方图法阈值对外显子和内含子序列的信噪比区分度要求较高，区分度不显著时预测效果不佳。选取不同的阈值对预测不同基因类型的效果不一样，作者只研究了小鼠基因，对于不同物种的阈值确定方法有待研究。目前，虽然已有一系列基因识别软件，如 GenScan，GAZE，NetPlantGene 等，但大多预测算法针对性较强，基因准确率不高。有必要研究模型和算法来提高准确率，同时降低算法的时间复杂度，达到破解基因遗传信息的目的。

参考文献：

[1] Yin C C, Yau S S T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence [J]. J Theor Biol, 2007, 247: 687-694.

[2] Zhang R, Zhang C T. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences [J].Journal of Biomolecular Structure & Dynamics, 1994, 11(4): 767-782.

[3] Silverman B D, Linsker R. A measure of DNA periodicity [J]. Journal of Theoretical Biology, 1986, 118: 295-300.

[4] Anastassiou D. Frequency-domain analysis of biomolecular seqences [J].Bioinformatics, 2000, 16: 1073- 1081.

[5] Chakravarthy N, Spanias A, Lasemidis L D. Autoregressive modeling and feature Analysis of DNA sequences [J]. Eurasip Jasp, 2004, 1: 13-28.

[6] Fickett J W, Tung C S. Assessment of protein coding measures[J]. Nucleic Acids Res, 1992,20: 6441-6450.

[7] Tiwari S, Ramachandran S, Bhattacharya A, et al. Prediction of probable genes by Fourier analysis of genomic sequences[J]. Comput Appl Biosci, 1997, 113: 263-270.

[8] Voss R F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences [J]. Phys Rev Lett, 1992, 68: 3805-3808.

[9] Yan M, Lin Z S and Zhang C T. A new Fourier transform approach for protein coding measure based on the format of the Z-curve [J]. Bioinformatics, 1998,14(8): 685-690.

[10] Eivind C. Equivalence of two Fuirier methods for biological sequences [J].Journal of Mathematical Biology, 1997, 36: 64-70.

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

[12] 邵剑锋, 严晓华, 邵伟, 等. DNA 序列信号 3-周期特性[J]. 南京工业大学学报: 自然科学版, 2012(4): 133-137.

[13] 马玉韬, 张成, 张泽林, 等. DNA 序列映射方法对蛋白质编码区预测准确率的影响[J]. 安徽农业科学, 2012, 40(6): 3234-3238.

[14] Sharma S D, Shakya K, Sharma S N. Evolution of DNA Mapping Schemes for Exon Detection [J]. ICCCET, 2011: 71-74.