基因识别问题及其算法实现

一、背景介绍

DNA是生物遗传信息的载体，其化学名称为脱氧核糖核酸（Deoxyribonucleic acid，缩写为DNA）。DNA分子是一种长链聚合物，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*）（见图1第一行）。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（见图1第二行），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*）。外显子在*DNA*序列剪接（*Splicing*）后仍然会被保存下来，并可在

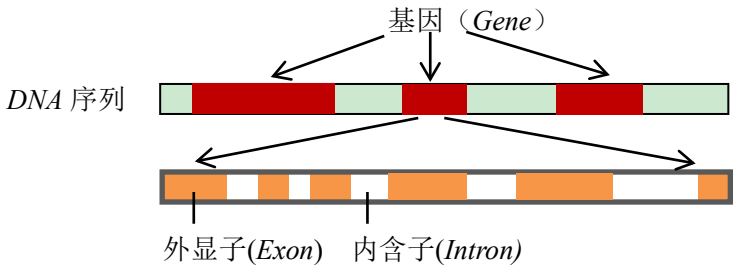

图 1 真核生物 *DNA* 序列（基因序列）结构示意图

蛋白质合成过程中被转录（*transcription*）、复制（*replication*）而合成为蛋白质（见图 2）。*DNA* 序列通过遗传编码来储存信息，指导蛋白质的合成，把遗传信息准确无误地传递到蛋白质（*protein*）上去并实现各种生命功能。

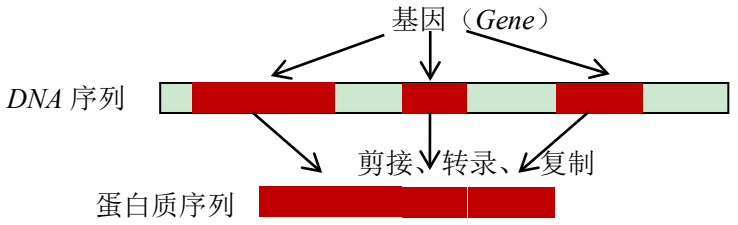

图 2 蛋白质结构示意图

对大量、复杂的基因序列的分析，传统生物学解决问题的方式是基于分子实验的方法，其代价高昂。诺贝尔奖获得者 W.吉尔伯特（Walter Gilbert，1932—；【美】，第一个制备出混合脱氧核糖核酸的科学家）1991 年曾经指出：“现在，基于全部基因序列都将知晓，并以电子可操作的方式驻留在数据库中，新的生物学研究模式的出发点应是理论的。一个科学家将从理论推测出发，然后再回到实验中去，追踪或验证这些理论假设。”随着世界人类基因组工程计划的顺利完成，通过物理或数学的方法从大量的 *DNA* 序列中获取丰富的生物信

息，对生物学、医学、药学等诸多方面都具有重要的理论意义和实际价值，也是目前生物信息学领域的一个研究热点。

二、数字序列映射与频谱 3-周期性：

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。

基因预测问题的一类方法是基于统计学的[1]。很多国际生物数据网站上也有“基因识别”的算法。比如知名的数据网站 http://genes.mit.edu/GENSCAN.html 提供的基因识别软件 GENSCAN（由斯坦福大学研究人员研发的、可免费使用的基因预测软件），主要就是基于隐马尔科夫链（HMM）方法。但是，它预测人的基因组中有 45000 个基因，相当于现在普遍认可数目的两倍。另外，统计预测方法通常需要将编码序列信息已知的 *DNA* 序列作为训练数据集来确定模型中的参数，从而提高模型的预测水平。但在对基因信息了解不多的情况下，基因识别的准确率会明显下降。

因此在目前基因预测研究中，采用信号处理与分析方法来发现基因编码序列也受到广泛重视 [4]。

1. 数字序列映射

在 *DNA* 序列研究中，首先需要把 *A*、*T*、*G*、*C* 四种核苷酸的符号序列，根据一定的规则映射成相应的数值序列，以便于对其作数字处理。

令 $I=\{A,T,G,C\}$，长度（即核苷酸符号个数，又称碱基对（*Base Pair*）长度，单位记为 *bp*）为 *N* 的任意 *DNA* 序列，可表达为

$$S=\{\,S[n]\mid S[n]\in I,\ n=0,1,2,\cdots N-1\}$$

即 *A*、*T*、*G*、*C* 的符号序列 S：$S[0],S[1],\cdots,S[N-1]$。现对于任意确定的 $b\in I$，令

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

称之为 *Voss* 映射[5]，于是生成相应的 0-1 序列（即二进制序列）$\{u_b[n]\}$：$u_b[0],u_b[1],\cdots,$，$u_b[N-1]$ $(b\in I)$。

例如，假设给定的一段 *DNA* 序列片段为 *S = ATCGTACTG*，则所生成的四个 0-1 序列分别为：

$\{u_A[n]\}$：$\{1,0,0,0,0,1,0,0,0\}$；　　$\{u_G[n]\}$：$\{0,0,0,1,0,0,0,0,1\}$；

$\{u_C[n]\}$：$\{0,0,1,0,0,0,1,0,0\}$；　　$\{u_T[n]\}$：$\{0,1,0,0,1,0,0,1,0\}$。

这样产生的四个数字序列又称为 *DNA* 序列的指示序列(*indicator Sequence*)。

2. 频谱 3-周期性

为研究 *DNA* 编码序列（外显子）的特性，对指示序列分别做离散 *Fourier* 变换(*DFT*)

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

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

$$P[k]=\left|U_A[k]\right|^2+\left|U_T[k]\right|^2+\left|U_G[k]\right|^2+\left|U_C[k]\right|^2,\ k=0,1,\cdots N-1 \tag{2}$$

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

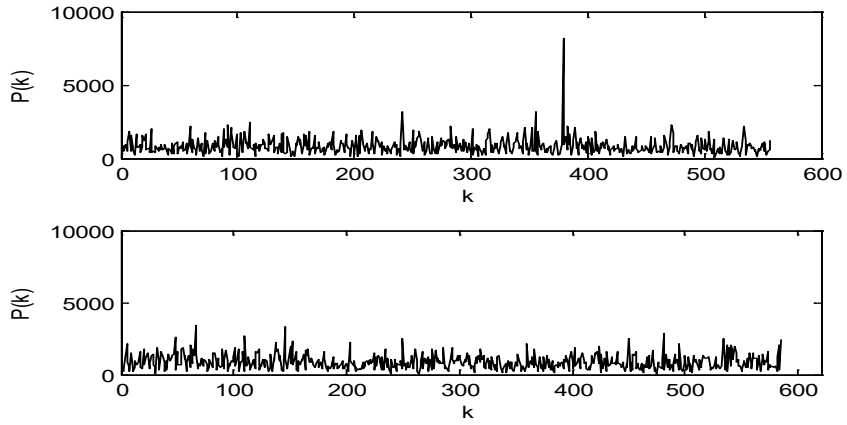

图 3 编号为 *BK006948.2* 的酵母基因 *DNA* 序列的功率谱（**因为对称性，实际这里只给出了功率谱图的一半**）。 (*a*) 上图是基因上一段外显子(区间为[81787，82920]，长 1134*bp*) 对应的指示序列映射的功率谱，它具有 3-周期性；(*b*) 下图是基因上一段内含子（区间为[96361,97551]，长 1191*bp*）的指示序列的功率谱，它不具有 3-周期性。

可以看到：外显子序列的功率谱曲线在频率 $k=\frac{N}{3}$ 处，具有较大的频谱峰值(*Peak Value*)，而内含子则没有类似的峰值。这种统计现象被称为碱基的 3-周期(*3-base Periodicity*)[2][3]。

记 *DNA* 序列 S 的总功率谱的平均值为

$$\overline{E}=\frac{\sum_{k=0}^{N-1}P[k]}{N} \tag{3}$$

而将 *DNA* 序列在特定位置，即 $k=\frac{N}{3}$ 处的功率谱值，与整个序列 S 的总功率谱的平均值的比率称为 *DNA* 序列的“信噪比”（*Signal Noise Ratio*，*SNR*），即

$$R=\frac{P[\frac{N}{3}]}{\bar{E}} \tag{4}$$

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[6]。

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[3][6]

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

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

易见，当四种核苷酸符号 b（$b\in I$）在序列的上述第一、第二、第三个子序列上出现的频数 x_b,y_b,z_b 越接近相等时，$\frac{N}{3}$ 处的谱值也就越接近于零。所以，基因外显子序列的功率谱曲线，在 $\frac{N}{3}$ 频率处具有较大的频谱峰值(*Peak Value*)，反映了在基因外显子片段上，四种核苷酸符号在序列的三个子序列上分布的“非均衡性”。通常认为这种现象源于编码基因序列“密码子”（*coden*）使用的偏向性（*bias*）。虽然目前对此现象产生的“机理”还不是十分地清楚，但是频谱的3-周期性被普遍认为是可用于识别基因编码序列（外显子）的一个重要的特征信息。

3. 基因识别

频谱峰值特征的发现，或者频谱与信噪比概念的引入，其最终目的是要探测、预报一个尚未被注释的完整的 *DNA* 序列的所有基因编码序列（外显子）片段。

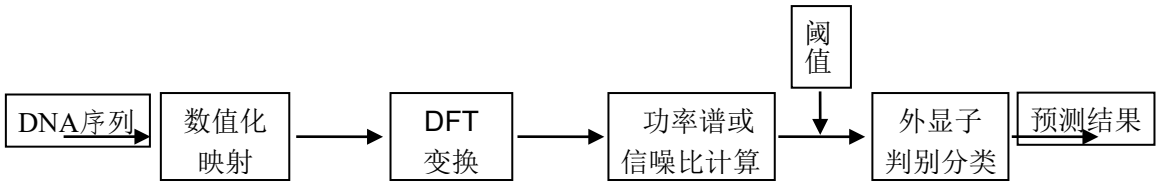

图 4 基于序列频谱 3—周期性的的基因预测方法流程图

已经有一些研究者提出了识别基因的算法（如参见[6]及其后面的文献）。目前利用信噪比的基因识别算法通常有两种：一是固定长度窗口滑动法[2] [3]；另一是移动信噪比曲线识别法[6]。

基于固定长度滑动窗口上频谱曲线的基因识别方法:

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

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

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

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

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

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

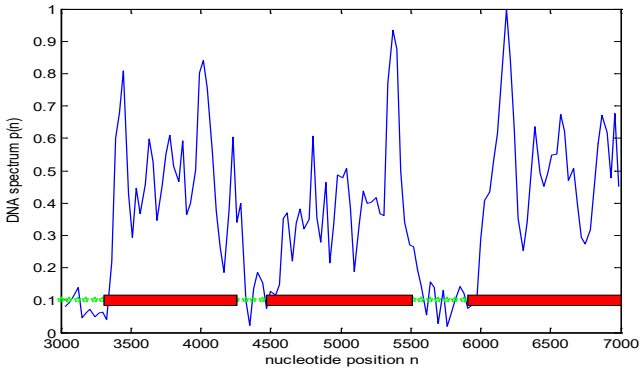

图 5 固定长度滑动窗口的频谱 $p = p(n;\frac{M}{3})$ 曲线（人类线粒体基因，NC_012920_1.fasta）

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

基于 DNA 序列上“移动序列”信噪比曲线的基因识别方法：

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

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

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

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

其中 $\overline{E}[n]$ 为移动子序列 $S_{0\sim n-1}$ 的功率谱的平均值 $\overline{E}[n] = \frac{\sum_{k=0}^{n-1} P[k]}{n}$。在坐标系中画出移动序列 $S_{0\sim n-1}$ 的信噪比曲线 $R[n]$（称为信噪比移动曲线（SNR walk curve），见图 6）

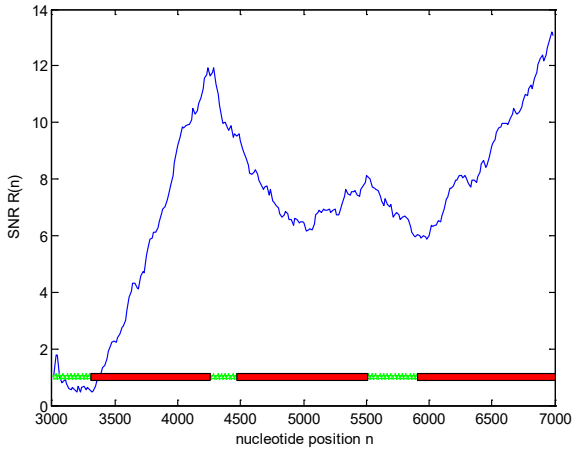

图 6　DNA 移动序列其指示序列的信噪比曲线。（人类线粒体基因，NC_012920_1.fasta）

图中红色水平细线条是 DNA 序列实际的基因外显子的区间。DNA 序列的信噪比移动曲线

的峰、谷与基因外显子区间的端点也具有较“明显的”的对应关系。

三、请研究的几个问题：

1. 功率谱与信噪比的快速算法

对于很长的 DNA 序列，在计算其功率谱或信噪比时，离散 *Fourier* 变换(*DFT*)的总体计算量仍然很大，会影响到所设计的基因识别算法的效率。大家能否对 *Voss* 映射，探求功率谱与信噪比的某种快速计算方法？

在基因识别研究中，为了通过引入更好的数值映射而获取 DNA 序列更多的信息，除了上面介绍的 *Voss* 映射外，实际上人们还研究过许多不同的数值映射方法。例如，著名的 Z-curve 映射（参见[5]或者附件 1）。试探讨 Z-curve 映射的频谱与信噪比和 Voss 映射下的频谱与信噪比之间的关系；

此外，能否对实数映射，如：$A \to 0, C \to 1, G \to 2, T \to 3$，也给出功率谱与信噪比的快速计算公式？

2.对不同物种类型基因的阈值确定

对特定的基因类型的 *DNA* 序列，将其信噪比 R 的判别阈值取为 $R_0 = 2$，带有一定的主观性、经验性。对不同的基因类型，所选取的判别阈值也许应该是不同的。附件中给出了来自于著名的生物数据网站：http://www.ncbi.nlm.nih.gov/guide/ 的几个基因序列数据，另外也给出了带有编码外显子信息的 100 个人和鼠类的，以及 200 个哺乳动物类的基因序列的样本数据集合。大家还可以从生物数据库下载更多的数据，找你们认为具有代表性的基因序列，并对每类基因研究其阈值确定方法和阈值结果。此外，对按照频谱或信噪比特征将编码与非编码区间分类的有效性，以及分类识别时所产生的分类错误作适当分析。

3. 基因识别算法的实现

我们的目的是要探测、预报尚未被注释的、完整的 DNA 序列的所有基因编码序列（外显子）。目前基因识别方面的多数算法结果还不是很充分。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“精确地”确定基因外显子区间的两个端点。

对此，你的建模团队有没有更好的解决方法？请对你们所设计的基因识别算法的准确率做出适当评估，并将算法用于对附件中给出的 6 个未被注释的 DNA 序列（gene6）的编码区域的预测。

4. **延展性研究**

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 100*bp*）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。

参考文献：

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

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

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

【4】Berryman, M. J., Allison, A., 2005. Review of signal processing in genetics. Fluctuation and Noise Letters. 5(4), 13-35.

【5】Sharma, S. D., Shakya，K., Sharma，S. N., 2011. Evaluation of DNA Mapping Schemes for Exon Detection. International Conference on Computer, Communication and Electrical Technology– ICCCET. 2011, 18th & 19th

【6】Yin, C., Yau, S.S.-T. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of Theoretical Biology. 247, 687–694.