

一种用于基因预测的 FIR 数字滤波器

马宝山, 朱义胜

(大连海事大学信息工程学院, 辽宁大连 116026)

摘要: 数字信号处理技术已经用于 DNA(DeoxyriboNucleic Acid) 序列中的基因分析与识别. 本文利用数字滤波器对基因序列进行分析和预测, 根据基因序列周期 3bp(base pair), 设计了一种具有窄带选通特性的 FIR(Finite Impulse Response) 数字滤波器, 对基因序列进行滤波, 与 IIR(Infinite Impulse Response) 数字滤波器计算的结果对比, 获得了较好的仿真结果.

关键词: FIR 数字滤波器; IIR 数字滤波器; 基因预测

中图分类号: TN713 **文献标识码:** A **文章编号:** 0372 2112 (2007) 09 1710 04

FIR Digital Filter for Gene Prediction Application

MA Bao-shan, ZHU Yi-sheng

(College of Information Engineering, Dalian Maritime University, Dalian, Liaoning 116026, China)

Abstract: The techniques based on digital signal processing have been applied to analyze and identify the genes in a DNA (deoxyribonucleic acid) sequence. According to period 3bp (base pair) of gene sequences, a new FIR (finite impulse response) digital filter with narrow band pass response was proposed to predict protein coding regions. The prediction curve of exons was obtained by processing the DNA sequence. Compared with the existing IIR (infinite impulse response) digital filter, the FIR filter can yield better results.

Key words: FIR digital filter; IIR digital filter; gene prediction

1 引言

50 多年前, 人类发现了 DNA 分子的双螺旋结构^[1], 随后对基因组学的研究不断取得显著的成果, 获得了大量的基因组和蛋白质方面的数据, 如何对这些庞大的数据进行处理变得越来越重要. 在这个领域中, 传统和现代的信号处理方法一直发挥着积极的作用^[2], 对 DNA 序列, RNA(Ribonucleic Acid) 序列和蛋白质序列的处理一直是研究的核心. 一个 DNA 序列是由 4 种字符 A (adenine), T (thymine), C (cytosine), G (guanine) 构成的序列. 例如,

.....ATCCCAAGTATAAGAACTA.....

字母 A, T, C, G 分别表示 4 种不同的碱基. 如果把碱基 A 的标量序列表示成二进制序列, 例如: $x_A(n) = \dots 1000011001011011001 \dots$, 其中 1 表示 A 存在, 0 表示 A 不存在. 其它碱基的标量序列也可以类似的定义, 这样由字符组成的序列映射成了数字序列, 如果算法可行, 就可能用数字信号处理的方法来进行基因分析.

2 用数字信号处理的方法研究基因序列

对于真核细胞的 DNA 序列, 如图 1 所示, 大多数的外显子具有周期 3bp(base pair), 而内含子则没有该行为^[6]. 有许多理论可以解释这种现象, 一般认为是由于在密码子转变成氨基酸过程中存在着编码偏好性. 但这种现象也有例外, 某些基因就不表现这种行为. 目前, DNA 中的周期 3bp 已经广泛地用于发现新的 DNA 序列中的外显子(exon).

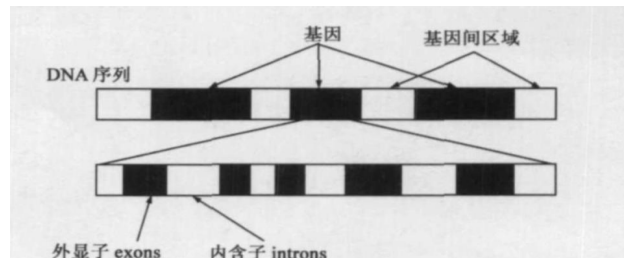


图 1 DNA 序列有基因和基因间区域, 真核细胞的基因具有外显子和内含子

研究函数

$S[k] \triangleq |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2$ (1)
其中, $X_A[k]$ 是长度为 N 的碱基 A 对应的标量序列 $x_A(n)$ 的离散傅立叶变换 DFT (Discrete Fourier Transform), 其它三个符号有对应的意义. 上式包含周期 3bp, 即外显子标量序列的频谱在归一化频率 $\omega = 2\pi/3$ 处有明显的峰值, 如图 2 所示. 周期 3bp 在基因编码区以外不存在, 因此可以用这个特性来预测编码区位置. 如果取 N 为 3 的倍数, 画出序列傅立叶变换后的频谱, 取采样值 $k = f/3$ 处有尖峰(即在 $k = N/3$ 处). 给定一组长碱基序列, 先用一个较短数据窗截取数据并计算出 $S[N/3]$, 然后滑动窗, 这样可以得到 $S[N/3]$ 是如何沿着 DNA 序列变化的.

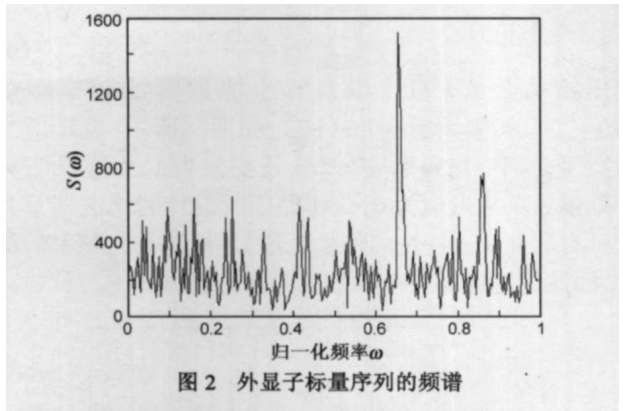


图 2 外显子标量序列的频谱

本文使用的数据是从美国国家生物信息中心网站 www.ncbi.com 下载的基因序列 F56F11.4a(在染色体中排列的碱基号从 7949- 14625, 编号: AF099922), 该序列共有 5 个外显子, 如表 1 所示.

表 1 基因序列 F56F11.4a 的外显子预测结构

Exon #	Relative to Itself		Relative to F56F11	
	Start	End	Start	End
1	1	111	7949	8059
2	1600	1929	9548	9877
3	3186	3449	11134	11397
4	4537	4716	12485	12664
5	6327	6677	14275	14625

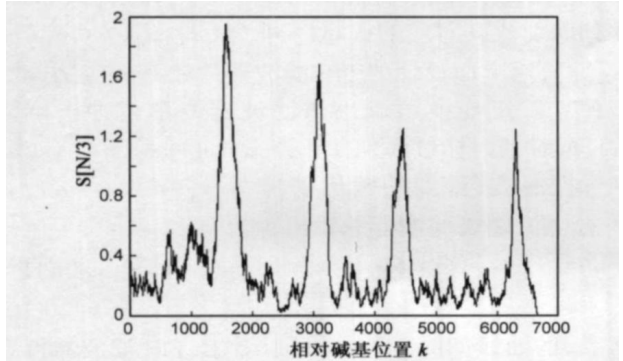


图 3 滑动窗法做离散傅立叶变换的预测曲线

将包含碱基的 DNA 序列表示成由 0 和 1 构成的数字序列, 经过滑动窗法做离散傅立叶变换(DFT)后, 得到曲线如图 3 所示. 窗长度 N 必须足够大, 以保证周期性影响远大于背景噪声 $1/f$ 的影响^[4]. 用这种 DFT 的方法, 不仅需要较大的计算量, 而且从图中可以看出, 对第一个外显子的识别效果也不够理想.

3 反陷波数字滤波器和多级滤波器

滑动窗方法可以看作对基因序列进行数字滤波. 假设滤波器的单位脉冲响应:

$$h(n) = \begin{cases} e^{j\omega_0 n}, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

这是一个带通滤波器, 中心频率 $\omega_0 = 2\pi/3$, 最小阻带衰减约 13dB. 如果把基因序列映射的数字序列当作滤波器的输入信号 $x(n)$, 通过数字滤波器后, 就得到了输出信号 $y(n)$. 由此, 文献[3]中提出了反陷波数字滤波器(notch filter), 对基因标量序列滤波取得一定的预测效果, 如图 4 所示, 但预测出的编码区的峰值不够明显, Vaidyanathan 又提出改进的多级数字滤波器(multi-stage filters), 能够从背景噪声例如 $1/f$ 噪声中更有效地分离周期 3bp, 得到较好的滤波效果, 如图 5 所示.

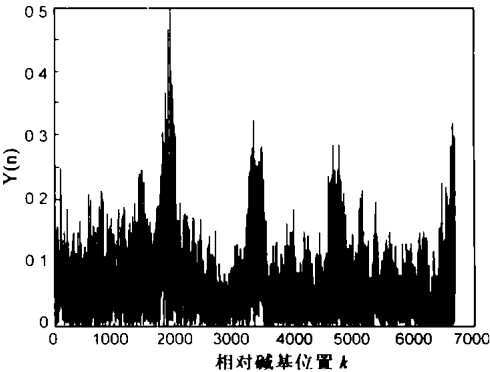


图 4 反陷波滤波器的预测曲线

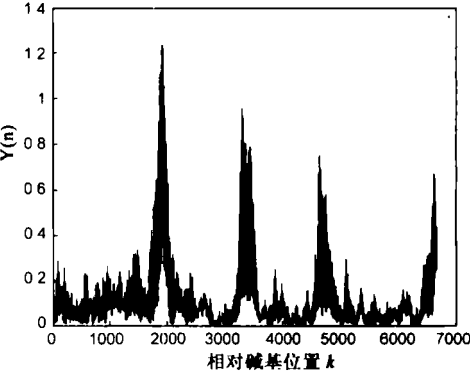


图 5 多级滤波器的预测曲线

由图 4 可以看出, 采用反陷波数字滤波器对基因序列滤波, 可以预测出序列外显子的波峰, 但是背景噪声影响很严重, 第一个峰值不明显. 图 5 所示改进后的多级滤波器效果明显, 背景噪声大大降低, 后 4 个波峰非

常明显.但是,多级滤波器对第一个外显子的识别还不够理想,在第一个外显子处没有明显的波峰.这是因为第一个外显子长度较短,只包含 111 个碱基,并且周期 3bp 特性相对其它 4 个外显子不是很强烈.该外显子的频谱如图 6 所示,而图 2 给出的是第二个外显子的标量序列频谱,比较两图,可知第一个外显子的周期 3bp 不如第二个外显子的明显.因此,影响滤波器预测基因编码区的主要因素是外显子的长度和其周期 3bp 特性的强弱.对于长度较短且周期特性弱的编码区的预测需要进一步设计滤波器的选通特性,以更好地预测第一个外显子.

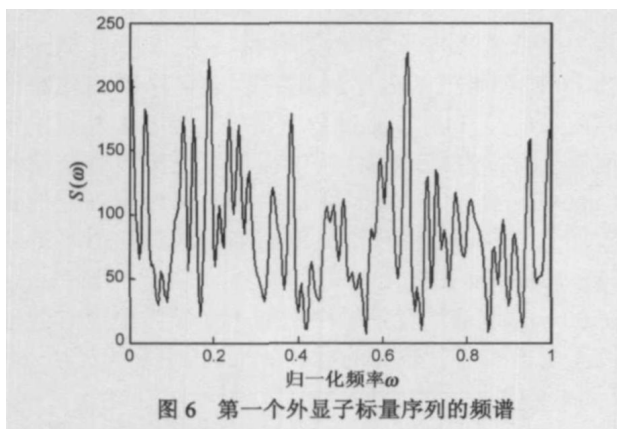


图 6 第一个外显子标量序列的频谱

4 FIR 带通数字滤波器

本文利用频率抽样法^[5]设计 FIR 带通数字滤波器,该方法的特点是可以根据需要设计滤波器的幅频特性.频率抽样法是从频域出发,把给定的理想频率响应 $H_d(e^{j\omega})$ 等间隔抽样,即

$$H_d(k) = H_d(e^{j\omega}) \big|_{\omega = \frac{2\pi k}{N}}, \quad k = 0, 1, \dots, N-1 \quad (3)$$

然后以 $H_d(k)$ 作为实际 FIR 数字滤波器的频率特性的抽样值 $H(k)$, 即

$$H(k) = H_d(k) = H_d(e^{j\omega}) \big|_{\omega = \frac{2\pi k}{N}} \quad (4)$$

得到 $H(k)$ 后,可以用频域的这 N 个抽样值 $H(k)$ 来唯一确定滤波器的单位冲击响应 $h(n)$.

根据基因序列外显子区域的周期 3bp, 设计一个在归一化频率 $\omega = 2/3$ 处有窄带选通特性的数字滤波器,即要设计的 FIR 滤波器的理想频率响应 $H_d(e^{j\omega})$ 在 $\omega = 2/3$ 为中心的很窄范围内幅值为 1, 其它频率处幅值都为 0, 对这个理想频率响应等间距抽取 $N = 2048$ 个采样点,第一个取样点在 $\omega = 0$ 处.由频率抽样理论的内插公式知道,利用这 N 个频域抽样值 $H(k)$ 可以求得 FIR 滤波器的系统函数 $H(z)$ 及频率响应 $H(e^{j\omega})$, 并且 $H(z)$ 或 $H(e^{j\omega})$ 逼近 $H_d(z)$ 或 $H_d(e^{j\omega})$, $H(z)$ 和 $H(e^{j\omega})$ 的内插公式为

$$H(z) = \frac{1-z^{-N}}{N} \sum_{k=0}^{N-1} \frac{H(k)}{1-W_N^k z^{-1}} \quad (5)$$

$$H(e^{j\omega}) = \sum_{k=0}^{N-1} H(k) \Phi\left(\omega - \frac{2\pi k}{N}\right) \quad (6)$$

其中内插函数 $\Phi(\omega)$ 为

$$\Phi(\omega) = \frac{1}{N} \cdot \frac{\sin\left(\frac{\omega N}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} e^{-j\omega\left(\frac{N-1}{2}\right)} \quad (7)$$

把式(7)代入式(6), 得到

$$H(e^{j\omega}) = \frac{1}{N} e^{-j\omega\left(\frac{N-1}{2}\right)} \sum_{k=0}^{N-1} H(k) e^{-j\frac{\pi k}{N}} \cdot \frac{\sin\left(\frac{\omega N}{2} - \frac{\pi k}{N}\right)}{\sin\left(\frac{\omega}{2} - \frac{\pi k}{N}\right)} \quad (8)$$

即

$$H(e^{j\omega}) = e^{-j\omega\left(\frac{N-1}{2}\right)} \sum_{k=0}^{N-1} H(k) \frac{1}{N} e^{j\frac{\pi k}{N}(N-1)} \cdot \frac{\sin\left[N\left(\frac{\omega}{2} - \frac{\pi k}{N}\right)\right]}{\sin\left[\left(\frac{\omega}{2} - \frac{\pi k}{N}\right)\right]} \quad (9)$$

利用插值公式(9)可以计算出滤波器的频率响应 $|H(e^{j\omega})|$, 其幅频响应 $|H(e^{j\omega})|$ 如图 7 所示.设计的滤波器只在归一化频率 $\omega = 2/3$ 处幅值为 1, 而在其它频率处幅频响应近似为 0, 这样保证所设计的滤波器只允许具有周期 3bp 特性的信号通过, 并抑制其它频率的信号, 因此该滤波器具有较好的频率选择性.

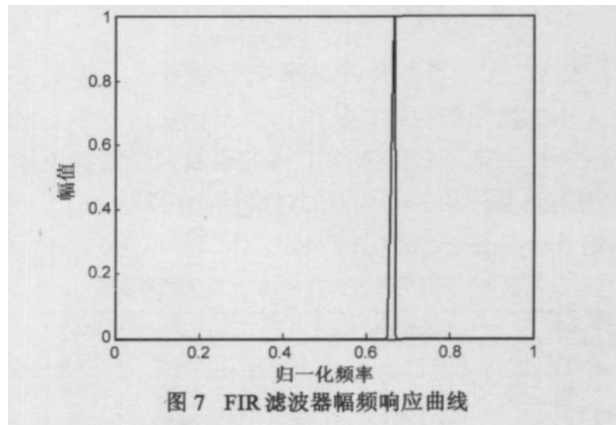


图 7 FIR 滤波器幅频响应曲线

5 仿真实验

本文例子使用从美国国家生物信息中心下载的基因 F56F11.4a 的数据.首先把由四种不同碱基构成的字符序列映射成其对应的标量序列 $x_A(n)$, $x_T(n)$, $x_C(n)$ 和 $x_G(n)$. 对上面设计的 FIR 滤波器的频率响应 $H(e^{j\omega})$ 进行傅立叶反变换可求出该滤波器的单位冲击响应 $h(n)$, 再对基因标量序列 $x_A(n)$, $x_T(n)$, $x_C(n)$, $x_G(n)$ 分别进行滤波, 得到的输出信号为 $y_A(n)$, $y_T(n)$, $y_C(n)$, $y_G(n)$. 定义滤波器总输出^[3]为

$$Y(n) = |y_A(n)|^2 + |y_T(n)|^2 + |y_C(n)|^2 + |y_G(n)|^2 \quad (10)$$

则 $Y(n)$ 的曲线如图 8 所示.可以看出, FIR 滤波器得到了较好的预测结果, 滤波后的曲线比较光滑, 不仅对背

景噪声抑制较好,而且第一个外显子处的波峰也识别得比较清晰.但是,由于第一个外显子的周期 3bp 特性相对较弱,因此它的波峰小于后四个外显子的波峰.

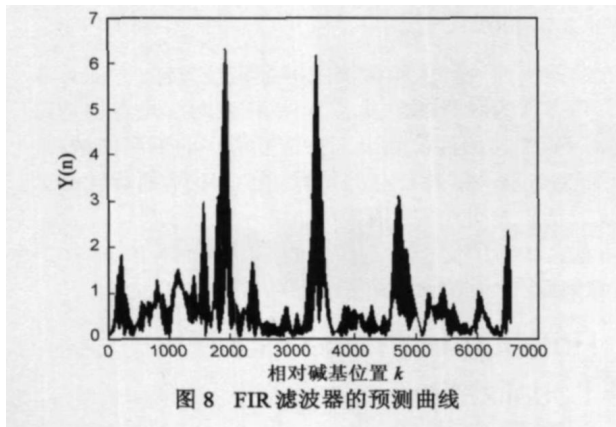


图 8 FIR 滤波器的预测曲线

6 结束语

本文针对现有数字滤波器用于基因预测存在的不足,设计了一种具有窄带选通特性的 FIR 数字滤波器.比较图 5 和图 8,可以看出所设计的滤波器有效抑制了背景噪声,滤波效果较好,预测的 5 个外显子区域清晰.仿真计算说明文中设计的 FIR 数字带通滤波器是有效的.但由于基因序列自身结构的复杂性,使得对它的信号处理和分析也十分复杂,本文利用周期 3bp 性质预测基因只是初步的研究.

参考文献:

[1] J D Watson, F H C Crick. A structure for DNA[J]. Nature,

1953, 171(4): 737.

- [2] D Anastassiou. Genomic signal processing[J]. IEEE Signal Processing Magazine, 2001, 18(7): 8–20.
- [3] P P Vaidyanathan, B J Yoon. Digital filters for gene prediction applications[A]. IEEE Asilomar Conference on Signals, Systems and Computers[C]. Monterey, CA: IEEE Signal Processing Society, 2002. 306–310.
- [4] C K Peng, S V Buldyrev, A L Goldberger, S Havim, F Sciortino, M Simons, H E Stanley. Long range correlations in nucleotide sequences[J]. Nature, 1992, 356(3): 168–170.
- [5] 程佩青. 数字信号处理教程[M]. 北京: 清华大学出版社, 2001.
- [6] E N Trifonov, J L Sussman. The pitch of chromatin DNA is reflected its nucleotide sequence[A]. Proc of the Nat Acad Sci[C]. USA: the National Academy of Sciences of the United States of America, 1980. 3816–3820.

作者简介:



马宝山 男, 1977 年出生于吉林德惠, 2000 年和 2003 年于哈尔滨工程大学分别获得学士学位和硕士学位, 现大连海事大学攻读博士学位, 任大连海事大学讲师, 研究方向: 信号处理.
E-mail: mbs0729@newmail.dlmu.edu.cn

朱义胜 男, 1945 年生, 大连海事大学教授, 电路与系统学科博士生导师, IEEE 高级会员. 主要从事宽带匹配网络、计算机辅助电路设计和信号处理的研究.