

基于全相位频谱分析的基因识别算法研究

王飞宇, 郑紫微*, 杨任尔

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 为对未知基因序列的编码区进行预测, 利用基因编码区存在的频谱 3-周期性质, 在传统的固定窗口长度滑动窗算法的基础上, 提出了一种改进型的基因识别算法. 新算法将全相位谱分析技术与多采样率数字信号处理技术相结合, 有效地降低了传统滑动窗算法中存在的截断效应, 减少了计算量并且可实现流水线操作. 最后通过计算机仿真将该算法所得结果与其他识别算法所得结果相比较, 结果表明该算法在核苷酸水平上有较高的预测准确性.

关键词: 基因识别; 3-周期性; 全相位 FFT 谱分析; 截断效应

中图分类号: TN911.6

文献标志码: A

文章编号: 1001-5132 (2016) 02-0037-06

DNA 是生物遗传信息的载体, 随着人类基因组计划的完成, 得到了海量的基因组信息, 对这些庞大的数据进行处理变得越来越重要. 真核生物的基因序列通常由编码区和非编码区构成, 编码区中蕴含的生物信息控制了生物蛋白质的合成. 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, 也是生物信息学当前的一个研究热点^[1].

由于基因频谱 3-周期性(3-Base Periodicity)^[2]的存在, 已有部分学者利用此性质提出了相应的基因预测算法, Tiwari^[3]和 Anastassiou 等^[4-5]先后提出了频谱特征识别算法(Spectral Content, SC)与优化的频谱特征识别算法(Optimized Spectral Content, OSC), 这 2 种方法都基于固定长度滑动窗来提取频谱特征. 马宝山等^[6]利用自适应滤波器(Recursive Least Square method, RLS)来实现基因预测. Mena-Chalco 等^[7]构造了修正的 Gabor 小波基(Modified Gabor-Wavelet Transform, MGWT)来提取频率特性.

傅里叶谱分析是最为常用且较为成熟的频谱分析工具. 笔者根据 DNA 序列映射与其频谱存在的 3-周期性特点, 然后在传统的固定窗口长度滑

动窗算法^[3-5]上进行改进, 考虑到当其在做加窗傅里叶变换时, 对数据截断操作带来的频谱泄露效应会影响识别精度, 由此引入全相位谱分析技术(All-phase FFT spectrum analysis)^[8-9]可以大幅度减少截断效应, 并且将其中的 FFT 分析模块处做了适当改变, 设计多采样率数字信号处理模型将其替代, 减少计算量的同时易于实现流水线操作. 最后通过仿真得到实验结果, 引入评价指标并与其他基因识别算法进行比较.

1 DNA 序列映射与频谱 3-周期性

1.1 Voss 映射介绍

在 DNA 序列研究中, 首先需要把 A、T、G、C 这 4 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, 以便于进行数字处理. 令:

$$I = \{A, T, G, C\}. \quad (1)$$

长度(即核苷酸符号个数, 又称碱基对长度, 单位记为 bp)为 N 的任意 DNA 序列可表达为:

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

现对于任意确定的 $b \in I$, 令:

$$u_b(n) = \begin{cases} 1, & S(n) = b, \\ 0, & S(n) \neq b, \end{cases} \quad n = 0, 1, 2, \dots, N-1, \quad (3)$$

收稿日期: 2014-10-26.

宁波大学学报(理工版)网址: <http://journal.nbu.edu.cn/>

基金项目: 国家自然科学基金(60972063); 浙江省杰出青年科学基金(R111041); 宁波市科技创新团队(2011B81002); 教育部高等学校博士学科点专项科研基金(20113305110002).

第一作者: 王飞宇(1988-), 男, 浙江宁波人, 在读硕士研究生, 主要研究方向: 数字信号处理. E-mail: 406120368@qq.com

*通信作者: 郑紫微(1975-), 男, 福建永春人, 教授, 主要研究方向: 数字移动通信. E-mail: zhengziwei@nbu.edu.cn

称之为 Voss 映射^[10]。例如, 假设给定的一段 DNA 序列片段为 $S = ATCGTACTG$, 则所生成的 4 个 0~1 序列分别如下:

$$\begin{aligned} \{u_A(n)\} &: \{1, 0, 0, 0, 0, 1, 0, 0, 0\}, \\ \{u_G(n)\} &: \{0, 0, 0, 1, 0, 0, 0, 0, 1\}, \\ \{u_C(n)\} &: \{0, 0, 1, 0, 0, 0, 1, 0, 0\}, \\ \{u_T(n)\} &: \{0, 1, 0, 0, 1, 0, 0, 1, 0\}. \end{aligned} \quad (4)$$

由此, 这样产生的 4 个数字序列又称为 DNA 序列的指示序列(Indicator Sequence)。

1.2 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傅里叶变换(DFT):

$$U_b(k) = \sum_{n=0}^{N-1} u_b(n) e^{-j \frac{2\pi nk}{N}}, k = 0, 1, \dots, N-1, \quad (5)$$

以此可得到 4 个长度均为 N 的复数序列。计算每个复数序列的功率谱, 并相加则得到整个 DNA 序列的功率谱序列:

$$P(k) = \sum_{b \in I} |U_b(k)|^2. \quad (6)$$

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 1 为酵母基因中的一段编码区序列(BK006948.2)在 Voss 映射下的功率谱。由图可见, 外显子序列的功率谱曲线在频率 $N/3$ 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-周期性^[11]。

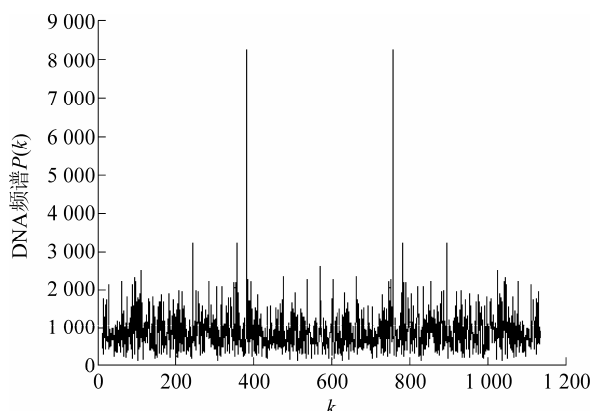


图 1 Voss 映射下外显子功率谱

2 算法分析

2.1 传统滑动窗算法分析

频谱 3-周期性的发现, 或者频谱概念的引入,

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。

对 1 个 DNA 序列和它的指示序列, 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。在以 n 为终点的长度为 M 的序列片段 $[n, n+M-1]$ 上, 作 4 个指示序列的离散傅里叶变换:

$$U_b(n, k) = \sum_{i=n}^{n+M-1} u_b(i) e^{-j \frac{2\pi ik}{M}}, k = 0, 1, \dots, M-1, \quad (7)$$

并求出它在 n 处的 3 周期频谱, 即将 4 个碱基的 3 周期功率谱谱值相加得到:

$$P(n, \frac{M}{3}) = \sum_{b \in I} \left| U_b(n, \frac{M}{3}) \right|^2. \quad (8)$$

按照这种方法得到频谱值, 并画出其关于碱基相对位置 n 的频谱曲线(图 2), 图中水平线条即为 DNA 序列实际的基因外显子区间。由此可以看到, 滑动窗频谱曲线的峰值与基因外显子区间具有“对应”关系。在确定外显子区间时, 通常的方法是设定 1 个相应的阈值, 取高于该阈值的部分即可得到外显子区间。但是由于基因序列内部随机噪声的存在, 以及传统滑动窗算法对序列直接截断产生的频谱泄露, 都会对识别效率产生不利的影响。

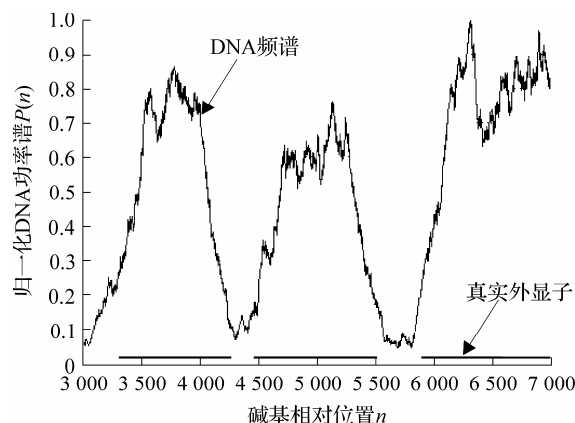


图 2 传统滑动窗算法得到的频谱
(人类线粒体基因, NC_012920_1)

2.2 全相位 FFT 谱分析技术

DFT 是最经典的谱分析手段, 然而其谱泄露是一个很突出的缺陷。谱泄露能严重影响谱分析的性能, 它源于在做 DFT 运算时, 隐含了对输入数据的截断操作。而截断会破坏波形的连续性, 因此在谱分析效果上以谱泄露的方式反映出来。

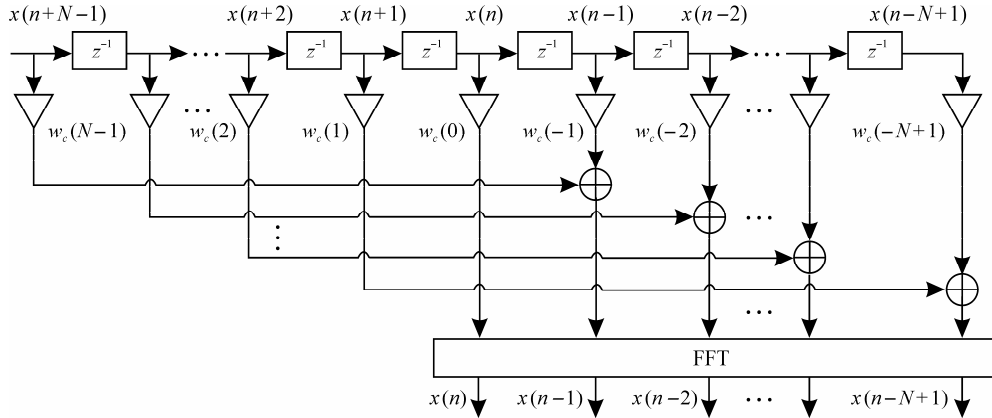


图3 全相位 FFT 谱分析基本框图

然而, 全相位 FFT 谱分析技术则可以减少截断效应. 数据在经过全相位数据预处理后, 再进行 FFT 技术可得到如图 3 所示的全相位 FFT 谱分析框图结构, 其中的 w_c 为卷积窗. 首先经过对数据的全相位预处理, 然后在进行常规的 FFT 运算, 这种数据截断方式的改进使振幅谱性能和相位谱性能大大改善.

考虑单频复指数 FFT 信号如下:

$$x(n) = e^{j(\omega_0 n + \varphi_0)} = e^{j(\frac{2\pi n\beta}{N} + \varphi_0)}, \quad (9)$$

可导出其不加窗的传统 FFT 谱为:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} e^{j(\frac{2\pi n\beta}{N} + \varphi_0)} \cdot e^{-j\frac{2\pi kn}{N}} = \frac{1}{N} \cdot \frac{\sin[\pi(\beta - k)]}{\sin[\pi(\beta - k)/N]} e^{j[\frac{N-1}{N}(\beta - k)\pi + \varphi_0]}. \quad (10)$$

再考虑由图 3 所示的框图, 先通过全相位预处理模块, 然后再做 FFT, 即可得到全相位 FFT 的输出为:

$$\tilde{X}(k) = \frac{1}{N^2} \cdot \frac{\sin^2[\pi(\beta - k)]}{\sin^2[\pi(\beta - k)/N]} e^{j\varphi_0}. \quad (11)$$

将式(10)与式(11)比较可以看出, 全相位 FFT 谱幅值为传统 FFT 谱幅值的平方, 并且它的理论相位值与频率偏移值无关. 这意味着旁谱线相对于主谱线的比值将以这种平方关系衰减下去, 从而使主谱线更加突出, 因而全相位 FFT 具有很好的抑制频谱泄露的性能.

由于 DNA 信号本身存在许多背景噪声, 并且直接对序列进行截断会引起很强的截断效应, 这时会影响对分界点的判断. 因此, 笔者应用全相位谱分析技术以减少谱分析时的截断效应, 以提高

基因预测的准确率.

2.3 多采样率模型

在全相位 FFT 谱分析结构图中, 由于我们关心的是 DNA 的 3-周期性质, 所以对于其中的 FFT 处理模块可替换为多采样率信号处理模型来简化计算, 并且易于实现流水线操作.

首先分析单个碱基的 3-周期特性:

$$U_b(n, \frac{M}{3}) = \sum_{i=n}^{n+M-1} u_b(i) e^{-j\frac{2\pi ik}{M} \cdot \frac{M}{3}} = u_b(n) + u_b(n+1) e^{-j\frac{2\pi}{3}} + u_b(n+2) e^{j\frac{2\pi}{3}} + u_b(n+3) + u_b(n+4) e^{-j\frac{2\pi}{3}} + u_b(n+5) e^{j\frac{2\pi}{3}} + \dots + u_b(n+M-3) + u_b(n+M-2) e^{-j\frac{2\pi}{3}} + u_b(n+M-1) e^{j\frac{2\pi}{3}},$$

可以得到:

$$U_b(n, \frac{M}{3}) = \sum_{k=0}^2 \left(\sum_{i=0}^{M/3-1} u_b(3i+n+k) \right) e^{-j\frac{2\pi k}{3} n}. \quad (12)$$

定义下采样后的子序列, 其中, k 为偏移量:

$$u_{bk}(n) = u_b(3n+k), \quad k=0, 1, 2. \quad (13)$$

则(12)式相应地可改写为:

$$U_b(n, \frac{M}{3}) = \sum_{k=0}^2 U_{bk}(n) e^{-j\frac{2\pi k}{3} n}, \quad (14)$$

其中, $U_{bk}(n)$ 表示如下:

$$U_{bk}(n) = \sum_{i=0}^{M/3-1} u_b(3i+n+k) = u_{bk}(n) \times h(n). \quad (15)$$

它们反映了在第 0、第 1、第 2 号阅读框位置上所对应碱基出现的频数, 具有生物学上的意义, 相当于将子序列 $u_{bk}(n)$ 通过传递函数为(16)式的滤波器:

$$H(z) = 1 + z^{-1} + \cdots + z^{-(L-1)}, L = M/3, \quad (16)$$

其实质为将下采样序列加矩形窗的运算, 可视相应情况将其替换为其他窗函数对应的 FIR 滤波器系统函数(如三角窗、汉明窗、布莱克曼窗等), 可进一步减少频谱泄露。具体的实现手段只需改变 FIR 滤波器的抽头系数即可, 如下式:

$$H(z) = \alpha_0 + \alpha_1 z^{-1} + \cdots + \alpha_{L-1} z^{-(L-1)}. \quad (17)$$

由(14)式可得到对应图4所示的多采样率信号处理模型。在图3所示的全相位FFT谱分析框图中, 保留全相位数据预处理模块, 将最后的FFT处理模块替换为图4中的多采样率信号处理模块。

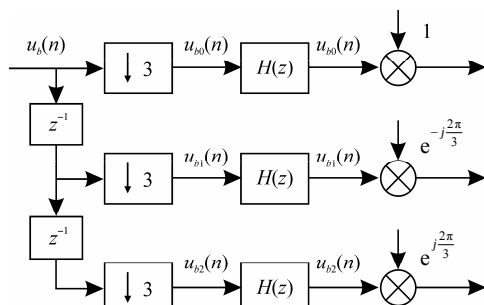


图4 多采样率信号处理模型

对(14)式做进一步推导, 最终可推导出:

$$U_b(n, \frac{M}{3}) = \sum_{k=0}^2 [U_{bk}(n) - U_{bl}(n)]^2, \quad (18)$$

$$l = (k+1) \bmod 3.$$

将直接计算频谱((7)式)与通过多采样率模型((18)式)计算频谱的方法相比较, 可以看出此设计避免了计算序列所有频段的谱值, 减少了计算量, 同时可实现流水线操作。表1则为两者计算复杂度的比较情况。

表1 计算复杂度比较

算法	实数乘法	实数加法
直接计算——(7)式	$2M$	$2M-2$
多采样率模型——(18)式	$M+3$	$M+3$

3 总体框图与仿真结果

综上所述, 基于全相位FFT谱分析的基因预测算法总体框图如图5所示。

对于给定基因序列(人类线粒体基因, NC_012920_1)分别做传统滑动窗算法与全相位FFT谱分析算法, 得到各自的基因频谱(图6与图7)。与传统滑动窗算法得到的结果相比较, 首先可见其归

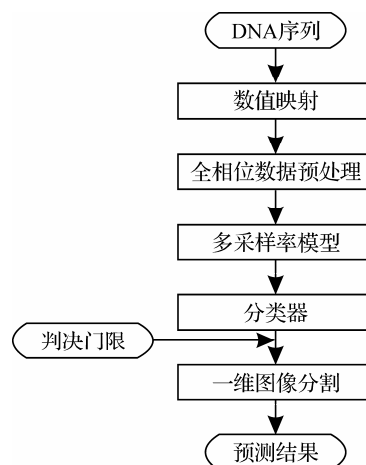


图5 基因预测算法总体框图设计

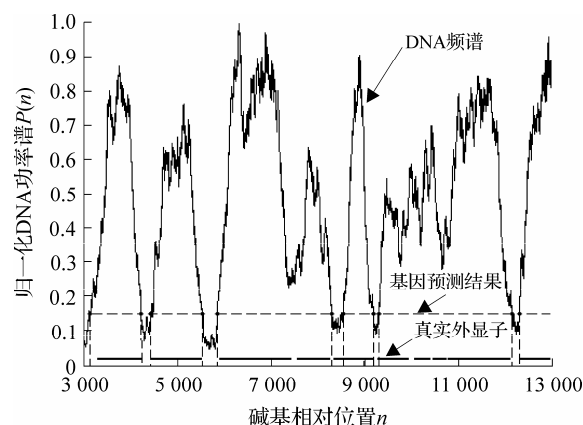


图6 传统滑动窗算法得到的频谱及预测结果

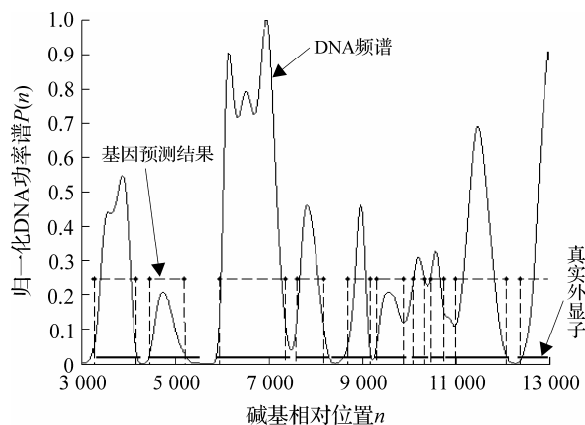


图7 全相位FFT谱分析法得到的频谱及预测结果

一化频谱幅值在非编码区衰减较大, 接近于零, 可见利用阈值选取法即可简单地剔除部分较短的非编码区片段; 而对于传统方法而言, 非编码区的频谱衰减较弱。其次, 全相位FFT谱分析算法得到的频谱曲线的上升沿与下降沿较为清晰, 便于利用一维图像分割来提取编码区间, 并进一步剔除那

些无法靠单一阈值筛选的非编码区区间; 但是对于传统的滑动窗算法而言, 由于在一整段编码区内就有较多的起伏, 因此难以直接用一维图像分割处理。

基于全相位谱分析方法的优势, 对频谱做阈值处理后进行一维图像分割^[12]。由图7可见, 第1、2、3、4、5段编码区可以直接靠单一阈值直接筛选, 随后进行的一维图像分割处理又可将第6、7、8段编码区较好地分离出来, 以此得到最终的预测区间, 具体预测数据见表2。

表2 预测结果与真实区间比较

预测区间 ¹⁾	预测区间 ²⁾	真实区间
	3260...4159	3307...4262
	4462...5206	4470...5511
3146...4241	5948...7357	5904...7445
4444...5543	7598...8176	7586...8269
5862...8327	8704...9184	8366...9207
8567...9214	9328...9900	9307...9990
9338...12185	10124...10351	10059...10404
12342...13812	10474...10755	10470...10760
14740...15912	11001...12083	10766...12137
	12393...14234	12337...14673
	14799...15854	14747...15887

注: 1)为传统滑动窗算法; 2)为全相位谱分析法。

现以核苷酸水平下^[13-14]对笔者提出的基因识别算法的准确度作出评估。由图8可见, 通过比较序列上预测的外显子与真实的外显子是否一致来度量预测的准确性。

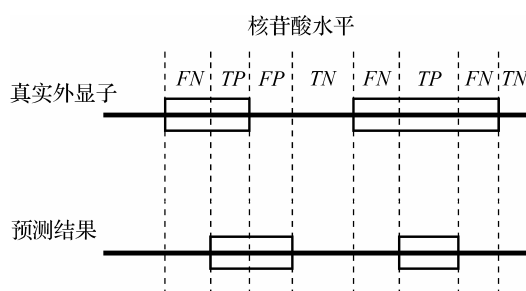


图8 核苷酸水平上的比较

在核苷酸水平上衡量预测准确性时, 通常采用 R_{CC} (Correlation Coefficient) 和 R_{AC} (Approximate Correlation) 指标来度量。计算公式如下:

$$R_{CC} = (TP \times TN - FP \times FN) / ((TP + FN) \times (TN + FP)(TP + FP)(TN + FN))^{1/2}, \quad (19)$$

$$R_{ACP} = \frac{1}{4} \left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN} \right), \quad (20)$$

$$R_{AC} = (R_{ACP} - 0.5) \times 2. \quad (21)$$

从美国国家生物信息中心(NCBI)下载相关的200组人类基因数据, 使用各种不同的算法进行识别, 并计算相应的评价指标。由表3可见, 全相位FFT谱分析的 R_{AC} 与 R_{CC} 值达到了0.85, 较其他方法效果较好, 比MGWT方法略低, 说明其在核苷酸水平上具有较高的预测准确性。考虑到MGWT需要进行小波变换, 全相位算法则在算法实现上面较为简单。

表3 核苷酸水平上各种预测算法比较

预测算法	R_{AC}	R_{CC}
All-phase FFT	0.85	0.85
Modified Gabor-Wavelet	0.87	0.87
Recursive Least Square method	0.71	0.71
Optimized Spectral Content Method	0.72	0.72
Fourier Spectral Content Method	0.66	0.66

4 结语

主要研究了基于全相位FFT谱分析技术的基因识别算法。在传统滑动窗算法的基础上, 对传统的数据截断方式进行改进, 通过全相位数据预处理的方式有效地减少了频谱泄露, 并且在其后的FFT分析模块上设计了多采样率信号处理模型, 简化了计算, 并且易于实现流水线操作。通过仿真得到频谱曲线及预测结果, 并基于核苷酸水平上的评价指标来证明此方法的有效性。

参考文献:

- [1] WANG Z, CHEN Y Z, LI Y X. A brief review of computational gene prediction methods[J]. *Geno Prot Bioinfo*, 2004, 2(4):216-221.
- [2] VOSS R F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences[J]. *Phy Rev Lett*, 1992, 68(25):3805-3808.
- [3] TIWARI S, RAMACHANDRAN S, BHATTACHARYA A, et al. Prediction of probable genes by Fourier analysis of genomic sequences[J]. *Comput Appl Biosci*, 1997, 13(3):263-270.

- [4] ANASTASSIOU D. Frequency-domain analysis of biomolecular sequences[J]. *Bioinformatics*, 2000, 16(12): 1073-1081.
- [5] ANASTASSIOU D. Genomic signal processing[J]. *IEEE Signal Processing Magazine*, 2001, 18(4):8-20.
- [6] 马宝山, 朱义胜. 用于基因预测的自适应滤波器的仿真研究[J]. *系统仿真学报*, 2007, 19(24):5620-5623.
- [7] MENA-CHALCO J P, CARRER H, ZANA Y, et al. Identification of protein coding regions using the modified gabor-wavelet transform[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008, 5(2):198-207.
- [8] 黄翔东, 王兆华. 基于全相位幅频特性补偿的 FIR 滤波器设计[J]. *电路与系统学报*, 2008, 13(2):1-5.
- [9] 黄翔东, 王兆华. 基于全相位频谱分析的相位差频谱校正法[J]. *电子与信息学报*, 2008, 30(2):293-297.
- [10] SHARMA S D, SHAKYA K, SHARMA S N. Evaluation of DNA mapping schemes for Exon detection[C]. 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), 2011:71-74.
- [11] YIN C C, STEPHEN S T Y. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence[J]. *Journal of Theoretical Biology*, 2007, 247: 687-694.
- [12] GONZALEZ R C, WOODS R E. Digital image processing[M]. 3rd ed. New York: Prentice Hall Presses, 2008:460-494.
- [13] RAMACHANDRAN P, LU WUSHENG, ANTONIOU A. Filter-based methodology for the location of hot spots in proteins and Exons in DNA[J]. *IEEE Transactions on Biomedical Engineering*, 2012, 59(6):1598-1609.
- [14] 孙红卫, 翁洋, 朱允民. 基因预测准确性的度量标准分析[J]. *四川大学学报(自然科学版)*, 2006, 43(3):649-654.

Gene Prediction Based on All-Phase Spectrum Analysis

WANG Fei-yu, ZHENG Zi-wei*, YANG Ren-er

(Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China)

Abstract: The 3-base periodicity, which exists in the coding regions, is adopted to predict the location of coding regions in an unknown DNA sequence. Based on the fixed-length sliding window approach, an improved algorithm for gene prediction is proposed, which combines all-phase FFT spectrum analysis and multi-rate digital signal processing. The method possesses some merits including reducing the truncation effect, improving the operation efficiency and providing a scheme for pipeline operation. Computer simulation results are given to compare the presented approach with other method for gene prediction, indicating that the proposed method achieves a satisfactory prediction accuracy at the nucleotide level.

Key words: gene prediction; 3-base periodicity; all-phase spectrum analysis; truncation effect

(责任编辑 章践立)