第 37 卷 第 9 期
Vol.37 No.9

计 算 机 工 程
Computer Engineering

2011 年 5 月
May 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)09—0181—03 文献标识码：A 中图分类号：TN911.72

基于频谱分析的串联重复序列识别方法

聂俊岚，毛伟伟，王常武，王宝文，刘文远

(燕山大学信息科学与工程学院，河北 秦皇岛 066004)

摘 要：针对现有串联重复序列识别方法存在的计算量大、灵敏度低等问题，提出一种基于频谱分析的串联重复序列识别方法。该方法采用碱基的电子离子相互作用势作为基因序列数字化表示的方法，通过对数字序列作离散傅里叶变换得到序列中串联重复序列出现的频率，并对基因序列做加窗傅里叶变换，找出串联重复序列存在的位置。实验表明，该方法的计算量较已有方法减少了 75%，并能较好地解决已有方法识别灵敏度低的缺点。

关键词：串联重复序列；离散傅里叶变换；电子离子相互作用势；频谱分析；信噪比

Identification Method of Tandem Repeats Based on Spectral Analysis

NIE Jun-lan, MAO Wei-wei, WANG Chang-wu, WANG Bao-wen, LIU Wen-yuan

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

【Abstract】Aiming at the drawbacks of the existing tandem repeats finding methods, such as large number of calculations and feeble sensitivity, this paper presents a tandem repeats identification method which is based on spectral analysis. The technique employs the Electron-Ion Interaction Potential(EIIP) of each nucleotide as the numerical representation for DNA sequence, and obtains the occurrence frequency of the tandem repeats which is buried in the sequence after computing the Discrete Fourier Transform(DFT) of the sequence. The windowed Fourier transform is used, and the tandem repeats location is identified efficiently. Experiment demonstrates that the calculation amount is reduced by 75% compared with the existing methods, and greatly resolves the feeble sensitivity of the existing techniques.

【Key words】tandem repeats; Discrete Fourier Transform(DFT); Electron-Ion Interaction Potential(EIIP); spectral analysis; Signal to Noise Ratio(SNR)

DOI: 10.3969/j.issn.1000-3428.2011.09.063

1 概述

重复基因序列在生物进化过程中起着非常重要的作用。这些重复序列在病毒和原核生物中很少出现，在真核生物中则大量存在。目前科学证实，人类基因组中大约含有 50%以上的重复基因序列。这些重复序列种类繁多，大部分具体功能目前还不十分清楚。然而已有的研究显示，一些特定的重复序列在基因表达、调控和遗传等方面起着十分重要的作用。例如一些三核苷酸重复序列拷贝数的异常增加会导致某些人类遗传病的产生，像脆性 X 染色体综合症、亨廷顿舞蹈症及弗兰德利克氏共济失调症等[1]。因此，对串联重复序列的深入研究有助于揭示某些遗传疾病的发病机制，为诊断和治疗这些遗传疾病提供更有效的方法。

根据所采用方法的不同，现有的重复序列发现方法大体可以分为 2 类：(1)基于字符串匹配的方法；(2)基于数字信号处理的方法。

基于字符串匹配的串联重复序列识别方法主要有 TRF、STRING、STAR、MREPS 以及 ATR 等[2]。然而这些方法都存在一些缺陷，一方面这些方法不能保证发现序列中所有可能的串联重复序列，另一方面，这些方法的计算复杂度会随着基因序列中串联重复序列拷贝的长度呈指数形式增长。

由于基因在长期的进化过程中会在某些位置发生例如插入、删除、替换等突变，因此方法的鲁棒性也成为了研究者必须要考虑的问题。而数字信号处理的方法在鲁棒性方面显然要优于字符串匹配方法。文献[3]提出了一种基于傅里叶变换来识别串联重复序列的方法(SRF 方法)，该方法采用了二进制方法来表示各个碱基。并分别求出各碱基的频谱，最后将 4 个碱基的频谱相加得到基因序列的总频谱。观察频谱图可得到基因序列中串联重复序列拷贝出现的频率。文献[4]提出了一种改进的傅里叶变换方法(傅里叶乘积识别法)识别基因序列中的串联重复序列，将各个碱基的频谱相乘的结果作为总的频谱来观察基因序列中重复序列拷贝出现的频率。

上述 2 种方法都能较好地识别出基因序列中串联重复序列出现的位置。但是在定位重复序列位置时都需要针对每个串联重复拷贝频率分别求其加窗傅里叶变换，才能得到 DNA 序列中所有串联重复序列出现的位置，识别灵敏度低。并且由于这 2 种方法对 DNA 序列均采用二进制表示法，需要对每条基因序列做 4 次离散傅里叶变换才能求出该 DNA 序列的频谱图，计算量大。串联重复序列识别中要处理的数据量一般都比较大，往往是整个基因组，因此计算量是识别方法中应该考虑的重要问题。针对这些问题，本文提出了一种基于电子离子相互作用势(Electron-Ion Interaction Potential, EIIP)[5]的傅里叶频谱分析识别方法。

2 串联重复序列识别方法

频谱分析是指应用傅里叶变换将难以处理的时域信号转换成易于分析的频域信号，得到信号的幅值、相位、能量等与频率的关系，从而方便对信号进行分析。离散傅里叶变换

基金项目：河北省教育厅自然科学研究计划基金资助项目(2009339)

作者简介：聂俊岚(1962－)，女，教授，主研方向：生物信息学，虚拟仿真，图像处理；毛伟伟，硕士研究生；王常武，教授；王宝文，副教授；刘文远，教授

收稿日期：2010-10-25 **E-mail：**maom919@163.com

使数字信号处理可以在频域上以数字运算方式进行，是一种用于描述离散信号时域表示与频域表示关系的数学工具[6]。这里首先将基因序列映射成数字序列，然后对数字序列做离散傅里叶变换得到该基因序列的频谱图，进而对该基因序列中所反映出来的频域信息进行分析。最后对基因序列进行时频分析，便可找出基因序列中串联重复序列的位置信息。由于传统的傅里叶变换在时频分析中没有时间分辨率，因此本文采用加窗傅里叶变换的方法来确定串联重复序列的位置。

加窗傅里叶变换的基本思想是把信号分成很多小的时间间隔，用傅里叶变换分析每一个小时间间隔，以便确定在那个时间间隔内的信号频率。加窗傅里叶变换解决了传统傅里叶变换在时频分析中没有时间分辨率的问题。

2.1 DNA 序列的数字表示

基因序列是由 4 种碱基(A、T、G、C)组成的字符串序列。为了在 DNA 序列分析中使用数字信号处理的方法，首先需要将 DNA 序列中的 4 个碱基分别映射成数字。本文根据各个碱基的 EIIP 值将 DNA 序列映射成为一条数字序列。碱基的 EIIP 描述是其自身价电子的平均能量，是碱基本身的一种物理属性，因此使用这种表示方法更有助于生物学家进一步对串联重复序列功能的研究，并且由于最终得到的数字序列由二进制表示法中的 4 条减少到了 1 条，使得计算量也相应地减少了 3/4。因此，相比传统的二进制表示法，该表示法更有意义。各碱基的 EIIP 如表 1 所示。

表 1 各碱基的 EIIP 值

碱基	EIIP
腺嘌呤 A(Adenine)	0.126 0
胸腺嘧啶 T(Thymine)	0.133 5
鸟嘌呤 G(Guanine)	0.080 6
胞嘧啶 C(Cytosine)	0.134 0

例如伪基因序列 $S[n]$ =ACTGACGATGATGC 的数字表示为： $N_e[n]$ =[0.126 0, 0.134 0, 0.133 5, 0.080 6, 0.126 0, 0.134 0, 0.080 6, 0.126 0, 0.133 5, 0.080 6, 0.126 0, 0.133 5, 0.080 6, 0.134 0]。

2.2 DNA 序列频谱分析

已知一条 DNA 序列 $S[n]$ 及其数字表示 $N_e[n]$，对该序列做频谱分析首先需要对这条序列做离散傅里叶变换。该序列对应的离散傅里叶变换为：

$$X(k)=\sum_{n=0}^{N-1}N_e[n]\mathrm{e}^{-\mathrm{j}2\pi kn/N}$$

其中， $0\leqslant k\leqslant N-1$； $N=length(N_e[n])$。

为了避免基因信号中的直流分量对频谱图的干扰影响到对其频谱的分析，对上式加入了一个参数 c，则该序列对应的离散傅里叶变换变为：

$$X(k)=\sum_{n=0}^{N-1}(N_e[n]-c)\mathrm{e}^{-\mathrm{j}2\pi kn/N}$$

其中， $c=\dfrac{1}{N}\sum_{n=0}^{N-1}N_e[n]$。

由上式可定义该序列的频谱表示：

$$S(k)=|X(k)|^2$$

对基因序列 $S[n]$ 进行频谱分析时，首先观察该序列所对应的频谱图，如果在该基因序列中存在一个长度为 p 的串联重复序列拷贝，那么在频谱图中对应频率 $f=1/p$ 的位置就会有一个波峰存在。因此可以从波峰的位置判断出该序列中存在的串联重复序列拷贝的长度。

2.3 信噪比的设置

在对基因序列进行频谱分析时，在频谱图中可能出现了若干波峰，但是由于噪声信号的存在，并不是所有波峰都是对研究有意义的，因此这里需要设置一个阈值即信噪比(S/N)对频谱图中出现的各个波峰进行评价：

$$\frac{S(k)}{S_m}>\frac{S}{N}$$

其中， S_m 为 $S(k)$ 的平均值。

只有当上式成立时，才认为频谱图中出现的波峰是有研究价值的。通常情况下当信噪比为 $S/N=2$ 时，频谱图中出现的波峰是有意义的。但是经过大量的实验发现当信噪比为 $S/N=4$ 时更有利于串联重复序列的定位。因此这里将信噪比设为 4。

2.4 DNA 序列的加窗傅里叶变换

在分析基因序列 $S[n]$ 的频谱图之后，得到了序列中存在的串联重复序列拷贝的长度 p，再使用加窗傅里叶变换对序列进行时频分析便可找出序列 $S[n]$ 中串联重复序列出现的位置。序列 $N_e[n]$ 的加窗傅里叶变换为：

$$STFT(m,k)=\sum_{n=0}^{M-1}N_e[n]g\cdot(n-mN)\mathrm{e}^{-\mathrm{j}2\pi nk/M}$$

其中， $k=0,1,\mathrm{L}\ ,M-1$。

在对 $N_e[n]$ 求加窗傅里叶变换时需要根据信号的特点选择窗函数的类型以及窗口的宽度。窗函数的时频特性将直接影响信号的频谱，从而对识别性能产生影响。窗口宽度的选取也很重要，适合的窗口宽度能够对原始信号提供较好的分辨率，更加便于观察信号在时域和频域上的分布情况。从加窗傅里叶变换结果可以得到对应频率的串联重复序列在基因序列中出现的位置。

3 实验及结果分析

3.1 频谱分析方法实验

为了验证该方法的高效性，本文从 GenBank 中提取了 2 条人类 DNA 序列进行实验验证，并将该方法与 SRF 方法以及傅里叶乘积识别法从灵敏度以及准确性 2 个方面进行了对比。这 2 条序列分别为 M65145 以及 Y-27H39。

首先对基因序列进行数字化表示，使用碱基的 EIIP 对序列 M65145 中的碱基进行数字映射，然后根据离散傅里叶变换得出该序列相应的频谱图如图 1 所示。

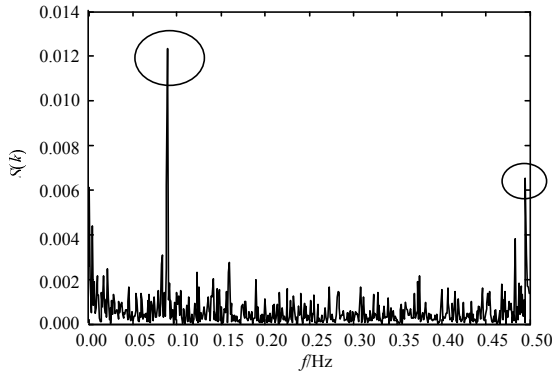

图 1 序列 M65145 的频谱图

从频谱图中可以观察到有若干波峰出现。根据信噪比设置仅选取了图中圆圈标注位置所对应的频率。该频率即为序列中重复序列拷贝出现的频率，对应频率分别为 $f_1\approx 0.09$ Hz， $f_2\approx 0.49$ Hz。由此可知序列 M65145 中存在 2 条不同的串联重复序列，并且对应的拷贝的长度分别为 $p_1=\dfrac{1}{f_1}\approx 11$ bp， $p_2=\dfrac{1}{f_2}\approx 2$ bp。

最后对序列 M65145 的 EIIP 序列做加窗傅里叶变换，此处采用的窗函数为 Hamming 窗，经过反复实验将窗口大小设为 109 时信号在时域上的分辨率最好，从而得到该基因序列中串联重复序列出现的位置信息。如图 2 所示，横坐标表示碱基的位置，纵坐标表示串联重复序列拷贝出现的频率。图中左侧矩形标识的部分表示拷贝长度为 11 bp 的重复序列在整条序列中出现的位置，位于 99 bp~498 bp 之间。右侧矩形标识部分表示拷贝长度为 2 bp 的重复序列在整条序列中出现的位置，位于 830 bp~940 bp 之间。

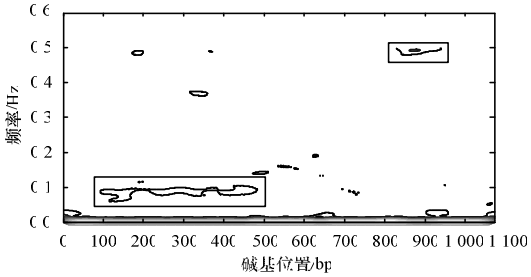

图 2　序列 M65145 的加窗傅里叶变换图

序列 Y-27H39 的频谱图如图 3 所示，从图中可以观察到圆圈标注位置所对应的频率为 $f \approx 0.25$ Hz 。由此可以判断在该序列中存在一个拷贝长度为 4 bp 的串联重复序列。求该序列对应数字序列的加窗傅里叶变换，变换结果如图 4 所示。此处采用的窗函数仍然是 Hamming 窗，多次实验发现窗口大小应设为 45。图中矩形标识部分即为该串联重复序列在基因序列中出现的位置，位于 81 bp~138 bp 之间。

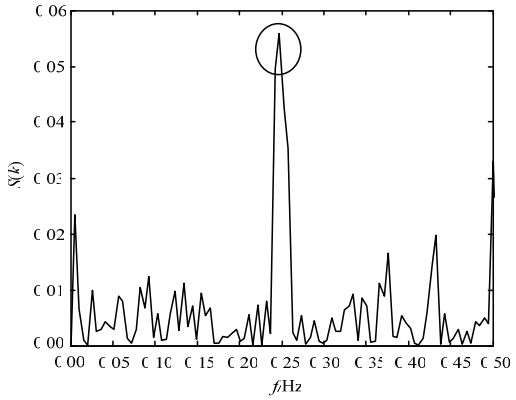

图 3　序列 Y-27H39 的频谱图

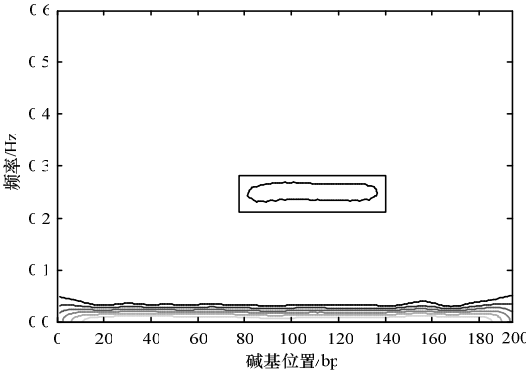

图 4　序列 Y-27H39 的加窗傅里叶变换图

3.2　实验结果分析

SRF 方法以及傅里叶乘积识别法在求基因序列的频谱图时，将 1 条 DNA 序列映射成 4 条数字序列，需要分别求出每条数字序列的离散傅里叶变换才能得到该序列的频谱图，计算量大。本文提出的频谱分析方法只需求一次离散傅里叶变换，便可得到该 DNA 序列的频谱图，因此计算量相对减少了 75%。并且这 2 种方法在求基因序列的加窗傅里叶变换时都需要针对每个串联重复拷贝出现的频率做一次加窗傅里叶变换才能得到所有频率下序列的时频分布，识别灵敏度较低。由图 2 可看出本文提出的频谱分析方法可以一次将基因序列中所包含的所有串联重复序列识别出来。因此，相比上述 2 种方法，本文提出的方法计算量小、灵敏度较好。

本文将 SRF 方法、傅里叶乘积识别法以及频谱分析法分别对序列 M65145 以及 Y-27H39 找出的位置信息与 GenBank 中标识的位置进行了比较，比较结果如表 2 所示。从表 2 中的结果比对可以看出，3 种方法均能很好地标识出基因序列中串联重复序列出现的位置，只是标识精度略有差异。本文提出的频谱分析法在定位精度方面与傅里叶乘积识别法相近，两者的定位精度均要优于 SRF 方法。

表 2　3 种方法的定位精度对比

DNA 序列	方 法	位置信息/bp
M65145	SRF	100~500, 800~900
	傅里叶乘积识别法	98~496, 830~938
	频谱分析法	99~498, 830~940
	GenBank	860~900
Y-27H39	SRF	75~145
	傅里叶乘积识别法	85~140
	频谱分析法	81~138
	GenBank	95~142

4　结束语

为了减小串联重复序列识别方法的计算量，提高识别灵敏度，本文提出了一种使用频谱分析进行串联重复序列识别的方法。实验结果表明，该方法在计算量，识别灵敏度以及定位精度方面均要优于已有方法。下一步要研究的工作是如何能够进一步提高识别精度。

参考文献

[1] Benson G. Tandem Repeats Finder: A Program to Analyze DNA Sequences[J]. Nucleic Acids Research, 1999, 27(2): 573-580.

[2] Du Liping, Zhou Hongxia, Yan Hong. OMWSA: Detection of DNA Repeats Using Moving Window Spectral Analysis[J]. Bioinformatics, 2007, 23(5): 631-633.

[3] Sharma D, Issac B, Raghava G P S, et al. Spectral Repeat Finder(SRF): Identification of Repetitive Sequences Using Fourier Transformation[J]. Bioinformatics, 2004, 20(9): 1405-1412.

[4] Petre G. POP, Tandem Repeats Localization Using Spectral Techniques[C]//Proc. of IEEE International Conference on Intelligent Computer Communication and Processing. Cluj-Napoca, Romania: [s. n.], 2007.

[5] Cosic I. Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules? Theory and Applications[J]. IEEE Transactions on Biomedical Engineering, 1994, 41(12): 1101-1114.

[6] 刘　鹏, 刘定生, 李国庆. 基于矩阵秩估计偏移量的频域超分辨率重建[J]. 计算机工程, 2009, 35(15): 29-31.

编辑　任吉慧