

连续小波变换法分析核酸序列的长程相关性^{*}

陈晓燕¹, 鲍伦军², 莫金垣¹

(1. 中山大学化学与化学工程学院, 广东 广州 510275;

2. 广州出入境检验检疫局, 广东 广州 510623)

摘要: 提出了基于连续小波变换的检测核酸序列长程相关性的新方法。利用小波变换来观察序列, 再用分数布朗运动建立数学模型, 计算 Hurst 指数, 并以此来衡量序列的长程相关性。最后发现在同一个基因内, 外显子序列的 Hurst 指数明显小于内含子序列, 此特征是这两种序列的显著区别。

关键词: 连续小波变换; 分数布朗运动; Hurst 指数; 外显子序列; 内含子序列

中图分类号: O652 7 **文献标识码:** A **文章编号:** 0529-6579 (2003) 03-0111-03

在宇宙天地万物之中, 最奇妙的现象莫过于生命。长久以来, 众多的科学家都在为揭示生命的秘密, 探索相关的信息而努力。但是, 由于生理学过程的非平稳和非线性特性, 给研究工作带来了许多困难, 然而借助工程数学研究的新成果, 则常常可以使问题豁然开朗。根据尺度不变性的分形概念来研究复杂基因的工作早在十几年前就引起科学家的广泛关注^[1]。核酸序列中的长程相关性与物种的进化动力学有关。对于核酸的研究方法, 以前运用了多种不同的技术包括均方根涨落^[2]、熵近似方法^[3]、平滑序列绝对值 (AVAS) 方法^[4]、Higuchi 方法^[4]和傅立叶分析^[5]等等, 用各种统计方法来比较不同种生物以及核酸编码区、非编码区, 外显子、内含子之间长程相关性质的差别。尽管如此, 仍然有许多问题尚未解决。对长程相关性分析, 一个主要的障碍就是核酸序列中的镶嵌结构, 它会形成链偏差 (strand bias), 这些链偏差会表现在全局图上^[6]。许多用于长程相关性现象的研究方法并不适用研究核酸序列。有人引入关于序列复杂度的计值方法来区分外显子和内含子^[7], 但是反映外显子和内含子复杂度的值的分布有很大的一部分是重叠的, 往往不利于从基因片段中找出外显子部分。然而, 小波基础上的多重分形是一个研究核酸序列非常有用的工具。本文采用了具有伸缩、平移自由的小波变换, 它是一架变焦镜头, 既可作为望远镜, 又可作显微镜^[8], 再结合分形理论来揭示多尺度下核酸序列的特征, 这使得对于核酸序列可以有更深入的认识。

1 核酸序列的映射表达

核酸序列中的 DNA 序列是由 A、G、T、C 4 个碱基按一定顺序排列而成的, 也就是 4 个符号的组合。由于标准的信号和时间序列分析需要输入实数或复数值, 所以, 首先要将核酸序列表示为一个数字序列。方法有很多种, 经过比较, 选择核酸游走 (DNA walk) 的表达方式。此名称是 Peng 等^[3]首次提出的, 它来源于统计学上随机游走 (random walk) 的概念。定义为:

$$X_i = \begin{cases} 1 & i \text{ 位置上为任意两种核苷} \\ -1 & i \text{ 位置上为另外两种核苷} \end{cases} \quad (1)$$

i 是整数 1, 2, 3, ..., L , 其中 L 是核酸链的碱基对数目; 将 $X = X_1, X_2, X_3, \dots, X_L$, 进行累加, 就得到一条一维曲线:

$$f(x) = \sum_{i=1}^L X(i) \quad (2)$$

根据 A 与其它 G、T、C 3 种碱基的组合, 可以得到 3 种映射方式, 即 AG walk 曲线、AT walk 曲线及 AC walk 曲线。图 1 显示了对小鼠一个基因的外显子序列和内含子序列进行数字化映射的结果曲线。

2 核酸序列长程相关性的表述

将核酸序列表达成信号后, 就可以用创造出自然界许多现象, 简单而研究较深入的自相似模型——分数布朗运动 (fractional brown movement, FBM) 来建立数学模型。作为普通布朗运动的推广, FBM 是 Mandelbrot 在 1968 年提出来的。

为了描述物体的分形特征, 需要指定一个判定标准, 所以引入 Hurst 指数, 即:

^{*} 收稿日期: 2002-11-14

基金项目: 国家自然科学基金资助项目 (20205003, 29975033)

作者简介: 陈晓燕 (1976 年生), 女, 博士研究生; 通讯联系人: 莫金垣, E-mail: cesmjv@zsu.edu.cn

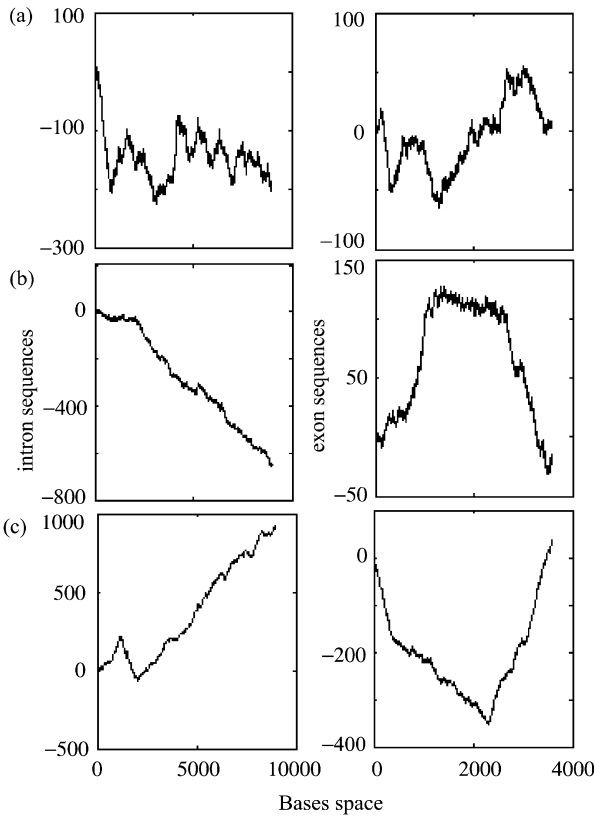


图 1 一个小鼠基因的 DNA 游走

Fig 1 DNA walk for a gene of mouse

(a) AG walk; (b) AC walk; (c) AT walk

$$B_H(t + \tau) - B_H(t) \stackrel{(d)}{=} a^{-H} [B_H(t + a\tau) - B_H(t)] \quad (3)$$

式中, B 是信号, a 是尺度, H 是 Hurst 指数, 或称为局部尺度指数, $0 < H < 1$, 当 $H = 1/2$ 时, FBM 为布朗运动。FBM 曲线的 Hausdorff 维数及记盒维数为 $D = 2 - H$, (d) 代表统计意义。若 $B_H(0) = 0$, 且取 $t = 0$, 则上式可写成:

$$B_H(\tau) \stackrel{(d)}{=} a^{-H} B_H(a\tau) \quad (4)$$

用基本小波 $\Psi(t)$ 对 $B_H(t)$ 作小波变换, 得:

$$WT_{B_H}(a, \tau) = \frac{1}{a} \int B_H(t) \Psi\left(\frac{t - \tau}{a}\right) dt \quad (5)$$

Mexican Hat 函数在时间域与频率域都具有很好的局部化, 并且满足:

$$\int_{-\infty}^{\infty} \Psi(x) dx = 0 \quad (6)$$

它在 $\omega = 0$ 处 $\Psi(\omega)$ 有二阶零点, 所以满足容许条件, 而且其小波随频率 ω 衰减得较快, 将它选

为本方法的基本小波。它是 Gauss 函数的二阶导数, 因为它像墨西哥帽的截面, 所以也称为墨西哥帽函数。Mexican Hat 函数定义为:

$$\Psi(x) = \frac{2}{\sqrt{3}} \pi^{-1/4} (1 - x^2) \exp\left[-\frac{x^2}{2}\right] \quad (7)$$

由于小波变换有一个尺度因子, 对小波函数具有伸展和收缩的作用, 也叫伸缩参数, 所以可以从不同长度来分析研究对象。就像用不同大小的尺子去量一个固定的长度, 量得的次数必然不同, 小的尺子得出的结果往往比大的尺子得到的结果更接近于事物的真实长度。只要找到了该事物测量的规律, 就算用大的尺子去测量, 经过运算, 也能得到较确切的结果。小波变换的这个特征, 在数学上被形象地称为“显微镜”的作用, 即它能够进行变焦, 伸缩参数 a 不同, 对应信号的变换结果也不同。于是, 利用这个特点, 可以分别从不同的特征长度来考察 DNA walk 的波动, 譬如小波变换尺度为 $2^3, 2^5, 2^7$ 时, 分别对应核酸序列的特征长度为 32, 128, 512 个核苷酸。

每一个尺度和小波变换系数就对应一个值, 即二维图中的一个点, 根据分形理论自相似的特性和分维数的定义, 不同尺度和自相似指数的双对数图谱中的点在一定范围内应该成一条直线, 类似于分析检测中经常使用的标准曲线, 而直线部分的斜率近似等于 Hurst 指数。

从欧洲分子生物学实验室 EMBL 的核酸序列数据库中提取出 300 个基因, 其中人类、植物、动物各 100 个。从这些基因中, 分别取出外显子和内含子, 并随机取出其中各物种的 80 个序列作为样本, 进行计算。计算的结果限于篇幅不予全部列出, 表 1 显示了各物种选出一个基因的计算结果。

3 小 结

从结果中发现同一基因的外显子维数比内含子大, 不同的核酸序列表达方式计算得到的维数也不相同, AC walk 的维数最高, AG walk 次之, AT walk 最小。以上是一般情况, 有个别基因例外。必须指出, 分形维数是分形的重要特征, 但同一分形在各种不同的定义与计算方法下得到的分形维数可能不尽相同, 而选择的标准就是计算得到的分形维数应与豪斯道夫 (Hausdorff) 维数呈单调递增或递减的关系, 且对研究对象计算得到的维数差别应充分大, 利于区分。

表 1 不同物种外显子和内含子序列的分形维数¹⁾

Tah 1 The fractal dimension of different species between exon and intron sequences

项目		Exon sequences			Intron sequences		
		Ag walk	Ac walk	At walk	Ag walk	Ac walk	At walk
Plant	<i>H</i>	0.524 0	0.458 9	0.573 1	0.678 9	0.622 5	0.734 1
	<i>D</i>	1.476 0	1.541 1	1.426 9	1.321 1	1.377 5	1.265 9
Human	<i>H</i>	0.550 8	0.469 4	0.623 1	0.586 4	0.585 3	0.731 7
	<i>D</i>	1.449 2	1.530 6	1.376 9	1.413 6	1.414 7	1.268 3
Animal	<i>H</i>	0.513 2	0.438 2	0.540 4	0.633 8	0.603 3	0.705 0
	<i>D</i>	1.486 8	1.561 8	1.459 6	1.366 2	1.396 7	1.295 0

1) *H* 是 Hurst 指数, *D* 是分形维数

参考文献:

[1] MANDELBROT B B. The fractal geometry of nature[M] . San Francisco: Freeman, 1982; 166—205.

[2] PENG C K , BULDYREV S V, GOLDBERGER A L, et al. Long-range correlations in nucleotide sequences[J] . Nature, 1992, 356: 168—170.

[3] ZHENG J, YU Z G. Correlation dimension and Kolmogorov entropy of DNA sequences[J] . Natural Science J Xiangtan University, 2000, 22(1): 115—119.

[4] LU X, SUN Z R, CHEN H M, LI Y D. Characterizing self-similarity in bacteria DNA sequences[J] . Physical Review E, 1998, 58: 3578—3584.

[5] DODIN G, VANDERGHEYNST P, LEVOIR P, et al. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences[J] . J Theor Biol, 2000, 206: 323—326.

[6] ARNEODO A, AUDIT B, BACRY E, et al. Thermodynamics of fractal signals based on wavelet analysis: application to fully developed turbulence data and DNA sequences[J] . Physica A, 1998, 254: 24—45.

[7] 沈如群, 凌伦奖, 陈润生, 等. 外显子和内含子的序列复杂性[J] . 生物物理学报, 1995, 11(3): 429—432.

[8] BAO L J, MO J Y, TANG Z Y. The application in processing analytical chemistry signal of a cardinal spline approach wavelets [J] . Analytical Chemistry. 1997, 69 (15) : 3053—3057.

Using Continuous Wavelet Transform to Analysis
the Long range Correlation of Nucleotide Sequences

CHEN Xiao yan¹, BAO Lun jun², MO Jin yuan¹

(1. School of Chemistry and Chemical Engineering, Sun Yat sen University, Guangzhou 510275, China;
2. Guangzhou Entry Exit Inspection and Quarantine Bureau, Guangzhou 510623, China)

Abstract: A new method to measure the long range correlation proprieties in nucleotide sequences was proposed, in which the sequences were investigated by wavelet transform, the mathematics model was established by fractional brown movement, and thus the Hurst exponent was calculated. The results were used to evaluate the long range correlation of the sequences.

Key words: continuous wavelet transforms; fractional brown movement; Hurst exponent; intron; exon