基于小波变换技术预测 DNA 序列的编码区

王 玉[1,2], 饶妮妮[1], 匡 斌[1], 袁祚涌[1]
(1. 电子科技大学生命科学与技术学院, 四川成都 610054; 2. 西南科技大学信息工程学院, 四川绵阳 621002)

摘 要: 三周期性是大多数基因组序列的编码区具有的主要特征. 本文提出利用小波变换分析 DNA 序列编码区的三周期性, 形成一种新的基于小波变换的 DNA 序列编码区预测方法, 理论和实验研究证实了新方法的可行性, 探测率和正确率分别达到 81% 和 75%, 特别是探测率较目前常用的其它一些方法有较大改善.

关键词: DNA 序列; 编码区; 小波变换; 预测
中图分类号: Q332 文献标识码: A 文章编号: 0372-2112 (2007) 01-0141-04

Predicting Protein Coding Regions of DNA Sequences Based on Wavelet Translation Technique

WANG Yu[1,2], RAO Ni-ni[1], KUANG Bin[1], YUAN Zuo-yong[1]
(1. *School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China;*
2. *School of Information Engineering, University of Southwest Science and Technology, Mianyang, Sichuan 621002, China*)

Abstract: The major signal in protein coding regions for most of genomic sequences is three-base periodicity. In this paper, we analyze this periodicity using wavelet transformation (WT) and propose a novel prediction approach for the protein coding regions of DNA sequences based on WT. This approach is able to predict and locate the coding regions simultaneously and is independent of training sets or existing database information. The validity of this approach is verified by a great deal of research results from theoretical analysis and experiments. The sensitivity and the specificity of novel approach reach 81% and 75% respectively. So, the prediction effectiveness is good. Especially, the sensitivity of novel approach is greatly improved compared with other techniques currently in use.

Key words: DNA sequence; coding regions (CDS); Wavelet translation; Prediction

1 引言

通过人类基因组计划的实施, 人们得到前所未有的海量生物信息. 解读这些生物信息是生物学领域今后的一项长期工作, 其中预测和定位基因是首先要解决的研究课题之一.

近年来, 大量预测基因的计算方法和软件被提出, 如基于神经网络的方法[1]、相关函数方法[2]、语言学方法[3]等. 原核生物的基因结构较为简单, 在基因组的 DNA 链上表现为一个编码蛋白质的基因对应为一段连续的开放阅读框, 因此, 基因预测问题相对简单, 一些研究小组已经在这方面取得了很好的结果. 而真核生物, 其基因结构很复杂, 许多基因是断裂基因, 间断成外显子和内含子, 且外显子在序列中长度比例极小. 因此, 要从具有较多内含子的真核生物基因组序列中正确识别出编码区是个相当困难的问题, 仍然有大量的工作要做.

本文基于蛋白编码区的显著特征——三周期性来研究编码区的预测技术. 在 DNA 序列的频谱中, 如果在 $f = 1/3$ 处存在一个尖峰, 就可能对应于三联密码子. 这种三周期性在蛋白编码区序列中是普遍存在的, 而在大多数非编码序列中却是不存在的[4]. 据此, 本文提出一种基于小波变换技术预测 DNA 序列编码区的简便方法.

2 方法

2.1 DNA 序列的数值映射

在进行计算分析之前, 首先将 DNA 序列由 A、T、C、G 所组成的符号序列转化为数值序列. 目前, 转化的方法很多, 如 DNA Walk 方法[5]、RY 方法和 SW 方法等[6]. 本文采用分解子序列法. 一个基因组序列在某一位置 j 出现某一种核苷酸 α 这一事件可以被看作是定义在概率空间 (Ω, F, P) 上的随机过程 $U_\alpha(j, \omega)$, 其中, $\Omega = \{A, T, C, G\}$. 因此, 对任意一段 DNA 序列, 都可以把它转化为 4 个子序列 U_A, U_T, U_C, U_G. $U_\alpha = \{U_\alpha(j, \omega); j \in R, \omega \in \Omega\}$, 其映射规则如下:

$$U_\alpha = \begin{cases} 1, & \omega = \alpha \\ 0, & \text{other} \end{cases}$$

例如, 一段 DNA 序列为 ATGCAAGT, 分解为四个子序列 U_A,

收稿日期: 2006-03-22; 修回日期: 2006-10-19
基金项目: 国家自然科学基金(No. 60571047); 四川省学术与技术带头人培养基金(No. 901008); 四川省应用基础项目(No. J13-075); 电子科技大学中青年人才培养计划(No. 601016)

U_T, U_C, U_G 的实例如表 1 所示.

表 1 *DNA* 序列转化为 U_A, U_T, U_C, U_G 4 个子序列实例

DNA 序列	A	T	G	C	A	A	G	T
U_A	1	0	0	0	1	1	0	0
U_T	0	1	0	0	0	0	0	1
U_C	0	0	0	1	0	0	0	0
U_G	0	0	1	0	0	0	1	0

因此, 基于傅立叶变换的 DNA 序列的谱密度可计算如下:

$$S(f)=\sum_{\alpha}S_{\alpha}(f)=\sum_{\alpha}\frac{1}{N}\left|\sum_{j=1}^{N}U_{\alpha}(j,\omega)\exp(-i2\pi fj)\right|^2 \quad (1)$$

其中, 离散频率 $f=j/N$, $j=1,2,3,\cdots,N/2$; N 表示 DNA 序列的长度; $S_{\alpha}(f)$ 表示子序列的谱密度, $\alpha=A$ 或 T 或 C 或 G; $i^2=-1$.

2.2 小波函数的选择

对 DNA 序列的谱密度函数 $S(f)$ 进行连续小波变换为[7]

$$W_f(a,b)=\langle s,\Psi_{a,b}\rangle=|a|^{-1/2}\int_R S(f)\overline{\Psi\left(\frac{x-b}{a}\right)}\,dx \quad (2)$$

其中, $\Psi(X)$ 是 Mexican hat 小波函数, a 为伸缩因子, b 为平移因子.

Mexican hat 函数为:

$$\Psi(x)=\frac{2}{\sqrt{3}}\pi^{-1/4}(1-x^2)e^{-x^2/2} \quad (3)$$

Mexican hat 函数在时域与频域都有很好的局部化, 且满足 $\int_{-\infty}^{\infty}\Psi(x)\,dx=0$. 其傅立叶变换为 $\hat{\Psi}(\omega)=\sqrt{2\pi}(e^{-\omega^2/2}-e^{-2\omega^2})$, 在 $\omega=0$ 处 $\hat{\Psi}(\omega)$ 有二阶零点, 满足容许条件, 而且其小波随参数 ω 衰减得较快. Mexican hat 小波比较接近人眼视觉的空间响应特性, 故选 Mexican hat 小波为本方法的基小波.

2.3 预测方法

以傅立叶变换为基础, 再用小波变换在一定尺度下去除随机涨落引起的高频噪声, 从而实现对 DNA 序列编码区的高精度预测和初步定位. 方法的具体步骤是:

①将生物的 DNA 序列映射为数值序列;

②取分析窗口长度为 M, 用式(1)计算窗口对应序列在 $f=1/3$ 处的值 $S(f)|_{f=1/3}$; 再以 DNA 序列被预测区域的第一个碱基为起点, 沿 DNA 序列以步长 3 滑动窗口, 得到 $S_M(f)|_{f=1/3}$ 相对于核苷酸序列位置 j 的函数 $P_M(j)$ (j 是长度为 M 的窗口的中间位置)[8];

③用式(2)对 $P_M(j)$ 进行小波变换, 得到一定尺度下的小波系数. 如果一个窗口的核苷酸序列在 $f=1/3$ 处有峰值存在, 则这段核苷酸序列就构成编码区的一部分, 否则就是非编码区的一部分.

我们用 MATLAB 实现了上述预测方法.

3 实验结果与分析

从 Genbank 数据库中选取 DNA 序列 ASYRVISP(accession number: M90075)作为实验对象. 已知该 DNA 序列有 6 段编码区, 分别位于: 522~624、745~1041、1166~1334、1419~1584、1676~1915 和 2015~2113. 用上述方法预测该 DNA 序列的编码区, 实验结果如图 1 所示.

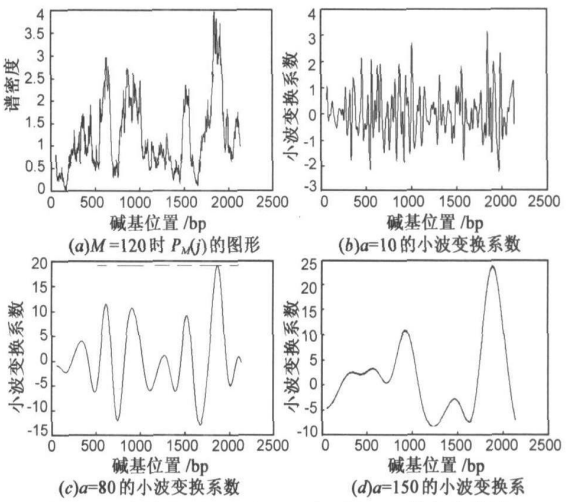


图 1 新方法预测 ASYRVISP 序列编码区的实验结果

图 1(*b*)是小波伸缩因子 a 取 10 的小波变换系数, 可以看出在该尺度下滤波效果明显不够, 未能将噪声信号和有用信号分开; 图 1(*c*)中的短横线代表已知编码区的位置. 由图可见, 曲线的峰值与上面的横线一一对齐, 证实了在该尺度下预测编码区的有效性; 图 1(*d*)是过度滤波的结果, 丢失了较多有用信息. 因此, 采用图 1(*c*)所对应的滤波尺度可实现对 DNA 序列 ASYRVISP 编码区的较高精度预测和初步定位, 预测性能如表 2 所示.

表 2 基因组序列 ASYRVISP 编码区的预测性能($M=120$, $a=80$)

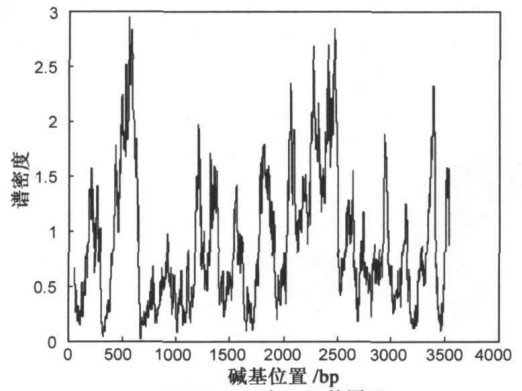
已知 CDS 数	探测到 CDS 数	误探测 CDS 数	缺失的 CDS 数	探测率	正确率
6	6	1	0	1	0.86

其中, 探测率= 探测到 CDS 数/已知 CDS 数.
正确率= 探测到 CDS 数/(探测到 CDS 数+ 误探测 CDS 数).

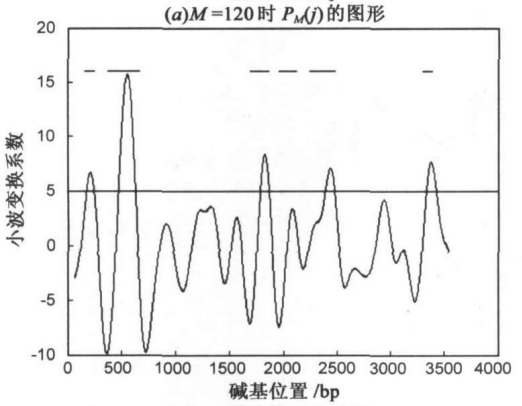
显然, 该方法中分析窗口的长度 M 和滤波尺度 a 是需要选择的. 由于酵母和昆虫病毒的序列只有非常少的内含子, 且开放阅读框的长度不超过 300 bp 是不常见的, 所以对于这些序列取 $M=300$. 实验表明, 窗口长度 M 的范围在 250~400 之间都可以得到相似的结果. M 小于 250, 会增加噪声; 当 M 大于 400 时, 由于数据交叉可能丢失开放阅读框. 较高等的生物存在较多内含子, 所以适宜的窗口长度为 $M\approx 120$. 对于小波伸缩因子 a 的选取, 通过大量实验发现的规律是: 当 M 较大时, a 取 M 的 1/2, 当 M 较短时, a 取 M 的 2/3, 可获得较好的效果. 例如, $M=300$ 时, 取 $a=150$; $M=120$ 时, 取 $a=80$. 研究还发现, 新方法所得到的小波变换系数中, 约 90% 的编码区所对应的波峰值大于 5. 因此, 为提高预测的正确率, 可设定一个判决门限 $C=5$, 即只有大于等于 5 的波峰才判断为基因. 然而, 这是以牺牲探测率为代价的.

用新方法分别在未设门限和设门限的情况下预测序列 AMU12024(accession number: U12024)的编码区, 结果如图 2 所示.

由图 2 知, 该 DNA 序列有 6 段已知编码区, 用设门限的方法可以准确预测出其中 5 段编码区. 位于 2000 bp 附近的编码区的预测受到四个杂波的干扰, 如果不设门限, 可能预测到 10 段编码区. 两种情况下的预测性能如表 3 所示. 由此可见, 设定一个门限值可有效地提高预测方法的正确率.



(a)M=120 时 $P_M(f)$ 的图形



(b)a=80 的小波变换系数

图 2　新方法预测 AMU12024 序列编码区的实验结果

表 3　有无门限两种情况 AMU12024 序列编码区的预测性能(M= 120, a= 80)

门限	已知 CDS 数	探测到 CDS 数	误探测 CDS 数	缺失 CDS 数	探测率	正确率
无门限	6	6	4	0	1	0.60
门限 $C \geqslant 5$	6	5	0	1	0.83	1

为充分证实新方法的有效性, 表 4 给出了 15 组不同生物的 DNA 序列编码区的预测性能.

表 4　15 组 DNA 序列编码区的预测性能(设门限 $C \geqslant 5$)

序号	名称	探测率	正确率	窗口长度 M	小波伸缩因 a
1	ACU08131	0.83	0.63	120	80
2	ALOEGLOBIM	1	1	120	80
3	AUM12024	0.83	1	120	80
4	ASYVISP	0.67	1	120	80
5	ATREGLOBIN	1	1	120	80
6	BCHEGLOBIN	1	1	120	80
7	BOVIAP	0.64	0.78	120	80
8	BTEBGL	1	0.75	120	80
9	BTU02285	1	0.6	120	80
10	ECGCOLIP1	1	1	120	80
11	GGACO1	1	0.71	120	80
12	HSCKBG	1	0.78	120	80
13	OABBGLOB	1	1	120	80
14	CALEGLBIM	1	1	120	80
15	XLBGL3	0.68	1	120	80
平均值		0.91	0.88		

对于上述 15 组 DNA 序列, 平均探测率和正确率分别达到 91% 和 88%.

4　比较与讨论

将新方法应用于序列集 ALLSEQ[9]. 该 DNA 序列集是评价不同基因预测方法或工具的标准序列集. 随机从 ALLSEQ 中选取 50 组序列, 新方法的平均预测性能如表 5 所示.

表 5　ALLSEQ 中随机抽取的 50 组序列的预测性能

已知 CDS 数	探测到的 CDS 数	误探测的 CDS 数	平均探测率	平均正确率
193	157	53	0.81	0.75

该方法与其他方法的预测性能[8]比较如表 6 所示.

表 6　与其他方法的比较

预测性能	基于小波变换的方法	傅立叶分析法	GeneParser2	Genelang	SORFIND
探测率	0.81	0.66	0.65	0.71	0.68
正确率	0.75	0.60	0.78	0.73	0.83

从表 6 知, 基于小波变换的方法在探测率上优于其他方法. 在正确率上该方法略次于 GeneParser2 和 SORFIND. 和基于神经网络的方法相比, 基于小波变换的方法不需要一个训练组来获得某类生物体的先验知识, 因此使用起来更加简便、快速和适用面广. 和基于相关指数的方法相比, 基于小波变换的方法预测较短序列的编码区性能更好. 和基于傅立叶变换的方法相比, 基于小波变换的方法保留了前者的所有优点, 但同时又能有效地消除噪声的影响, 使预测的准确性和正确率大幅度提高.

用本文的方法来预测 DNA 序列的编码区, 其显著的优点之一是可同时实现编码区预测和初步定位, 也可以在不同的尺度下观察事物, 有效地去除统计分析时所引入的噪声, 使预测方法的准确度和正确性都得到有效提高. 由于这种方法不需要基因组序列的任何先验知识, 因此使用简便、快速和适用面广. 当然, 每种方法都有一定的局限性. 该方法是基于蛋白编码区的一个普遍的性质即三周期性来进行预测的, 而有极少数(大约 4% ~ 5%)基因缺乏这种性质[4, 10], 因此, 对于这些基因而言, 该方法就失去效力.

5　结论

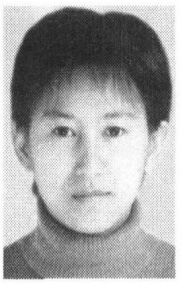
本文提出了一种基于小波变换的 DNA 序列编码区预测方法. 通过对来自于 Genbank 数据库和序列集 ALLSEQ 中的大量 DNA 序列进行的实验证明, 该方法对 DNA 序列编码区进行预测可取得良好效果, 探测率可达 81%, 优于其他方法, 正确率可达 75%, 同时可实现对编码区的初步定位. 正如文献[10]所指出, 通常难以用一种方法将各种生物 DNA 序列编码区预测问题全部解决, 需要多种方法融合, 才能达到准确预测和定位编码区的目的. 将基于小波变换的方法与其他方法融合以实现更高精度的编码区预测和定位, 是我们下一步的研究工作.

参考文献:

[1] R B Farber, A S Lapedes, Sirotkin K M. Determination of eur

karyotic protein coding regions using neural networks and information theory[J]. J Mol Biol, 1992, 226(2): 471- 479.
[2] S V Buldyrev, et al. Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis[J]. Phys Rev E, 1995, 51(5): 5084- 5094.
[3] S Dong, D B Searls. Gene structure prediction by linguistic methods[J]. Genomics, 1994, 23(3): 540- 551.
[4] W Lee, L Luo. Periodicity of base correlation in nucleotide sequence[J]. Phys Rev E, 1997, 56(1): 848- 851.
[5] John A Berger, Sanjit K Mitra, Marco Carli, et al. Visualization and analysis of DNA sequences using DNA walks[J]. Journal of the Franklin Institute, 2004, 341(1- 2): 37- 53.
[6] D Anasstassiou. Frequency-domain analysis of biomolecular sequences[J]. J. Bioinformatics, 2000, 16(12): 1073- 1081.
[7] Stephane Mallat. A Wavelet Tour of Signal Processing. Academic Press[M]. Sept. 15, 1999.
[8] S Tiwari, S Ramachandran, A Bhattacharya, et al. Prediction of probable genes by Fourier analysis of genomic sequences[J]. CABIOS, 1997, 13(3): 263- 270.
[9] M Burset, R Guigó. Evaluation of Gene Structure prediction program[J]. Genomics, 1996, 34(3): 353- 367.
[10] J W Fickett. The Gene identification problem: An overview for developers[J]. Comput Chem, 1996, 20(1): 103- 119.

作者简介:



王 玉 女, 1974 年 2 月生于四川省广元市, 电子科技大学生命科学与技术学院在职研究生, 西南科技大学信息工程学院讲师. 主要研究领域为生物信息学.
E-mail: cliu@uestc.edu.cn

饶妮妮 女, 1963 年生于四川, 1989 年毕业于电子科技大学, 现在是电子科技大学教授、博士生导师, 在国内外发表学术论文 50 余篇. 主要的研究领域包括: 信号与信息处理、生物信息学、远程医疗技术等.