基于傅里叶技术快速预测DNA序列编码区

王 玉，饶妮妮

(电子科技大学生命科学与技术学院 成都 610054)

【摘要】利用功率谱分析探测DNA序列编码区的主要特征信号三周期性，需要计算1/3频率点的傅里叶频谱。针对该问题，提出了只计算1/3频率点处的傅里叶频谱快速预测DNA序列编码区的方法。理论分析和实验证明，该方法的计算速度比使用傅里叶变换或快速傅里叶变换的方法快，计算准确性保持不变，不需要一个训练组或现有数据库的信息。

关 键 词 傅里叶变换; 功率谱分析; 基因组序列; 编码区
中图分类号 Q-332 文献标识码 A

An Efficient Algorithm for Prediction Genes of Genomic Sequences Based on Fourier Analysis

WANG Yu，RAO Ni-ni

(School of Life Science and Technology, Univ. of Electron. Sci. & Tech. of China Chengdu 610054)

Abstract The major signal in protein coding regions of genomic sequence is three-base periodicity. We use Fourier transform as a spectral analysis tool for genes detection, all that is required is a spot Fourier coefficient at $M/3$, and the complete Fourier spectrum is not required. An algorithm for computing spot Fourier coefficients is presented. Thereby, a method is developed to recognize the protein coding region of genomic sequence quickly. An important feature of the method is that its computational speed is very fast. Furthermore, this method is independent of training sets or existing datebase information and thus can find general applications.

Key words Fourier transformation; power spectrum analysis; genomic sequence; protein coding region

随着人类基因组计划的发展，近年来GenBank里的碱基数目呈指数增长，如何从大量的数据中挖掘出有用的生物信息，是生物信息学领域今后几十年都需要致力解决的问题，用计算方法识别DNA序列中蛋白编码区更是迫切需要解决的研究课题之一。近年来，国际上已经有一些研究小组进行基因识别的研究工作，并取得了一定的进展，提出了基于神经网络的方法、自相关函数方法、密码学方法等。但是，这些方法都存在一定的局限。因此面对世界范围内急剧增长的基因组序列，需要进一步完善这些方法或发展新的快速有效的方法，识别DNA序列中可能的基因。本文提出了基于傅里叶变换的一种快速预测DNA序列编码区的方法。

1 方 法

1.1 DNA序列的数值映射

一个基因组序列可以看作是由A,T,C,G四种碱基所构成的符号序列，在对基因组序列进行计算分析之前，先将其转化为数值序列。转化的方法有：(1) 自相关函数方法[1]；(2) DNA Walk方法[2]；(3) RY方法和SW方法等[3]。为了对DNA序列应用功率谱进行分析，可采用下面所述的方法将DNA序列转换为数值序列。一个基因组序列在某一个位置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$(\Omega, \square, P)$上的随机过程$X_\alpha(j, \omega)$，其中，$\Omega=\{A, T, C, G\}$。因此，对任意一段DNA序列，都可以把它转化为四个子序列X_A, X_T, X_C, X_G。$X_a=\{X_a(j,\omega); j\in R, \omega\in\Omega\}$，其映射规则为：$X_a=\begin{cases}1 & \omega=a\\ 0 & \text{其他}\end{cases}$。

例如，对任意一段DNA序列$AGCAGTACAGTGTACGGAT$，本文把它转化为X_A, X_T, X_C, X_G四个子序列，

收稿日期：2005 - 10 - 19
作者简介：王 玉(1974 -)，女，硕士，主要从事生物信息学方面的研究.

如表1所示。

表1 DNA序列转化为X_A, X_T, X_C, X_G四个子序列

DNA序列	*A*	*G*	*C*	*A*	*G*	*T*	*A*	*C*	*A*	*G*	*T*	*G*	*T*	*A*	*C*	*G*	*G*	*A*	*T*
转化为X_A	1	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0
转化为X_T	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1
转化为X_C	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
转化为X_G	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0

一个长度为N的子序列 X_a 的功率谱为：

$$P_a(f)=1/N\left|\sum_{j=1}^{N}X_a(j,\omega)\exp 2\mathrm{i}\pi fj\right|^2 \tag{1}$$

式中 离散频率$f=k/N$，k=1,2,⋯, N/2；$\mathrm{i}^2=-1$。长度为N的DNA序列总的功率谱为：

$$P(f)=\sum_a P_a(f)=\sum_a\frac{1}{N}\left|\sum_{j=1}^{N}X_a(j,\omega)\exp 2\mathrm{i}\pi fj\right|^2 \tag{2}$$

1.2 预测方法

对基因组序列进行功率谱分析发现，大多数蛋白编码区序列在f=1/3处出现一个峰值，而大多数非蛋白编码区序列的傅里叶频谱却没有峰值出现。根据这一特征，本文提出一个简便的基于功率谱分析的DNA序列编码区预测方法，具体步骤是：(1) 将任意一种生物的DNA序列映射为X_A, X_T, X_C, X_G 四个数值子序列；(2) 对每一个子序列，取分析窗口长度为M，用式(1)计算窗口在f=1/3(k=M/3)处的值，再用式(2)计算对应窗口的DNA序列在f=1/3(k = M/3)处的总的功率谱$P_M(f)|_{f=1/3}$；(3) 沿着DNA序列以步长3滑动窗口，可得到$P_M(f)|_{f=1/3}$相对于DNA序列位置j的函数$S_M(j)$ (j是长度为M的窗口的中间位置)[4]。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。

1.3 快速计算方法[5]

用上述预测方法对基因组序列的编码区进行预测和定位时，只需要计算每个窗口在f=1/3处的频谱。但是，傅里叶变换是同时计算M个点的频谱(M为窗口长度)。如果窗口滑动P次，则要重复计算M个点的频谱P次，计算量非常大。即使是采用快速傅里叶变换，完成一个DNA序列编码区的预测也需要相当长的时间。实际上只需要f=1/3 (k = M/3)这一点的频谱，就能实现预测。因此，重复计算不需要的M–1个点的频谱是一种浪费。为此，本文提出了一种快速计算方法。对离散信号 $x(n)$ 进行离散傅里叶变换的定义为：

$$\boldsymbol{Y}(k)=D_{\mathrm{DFT}}[x(n)]=\sum_{n=0}^{N-1}x(n)\exp(-2\pi\mathrm{i}nk/N) \tag{3}$$

式中 $m=0,1,\cdots,N-1$；$\mathrm{i}^2=-1$。也可将式(3)写成如下的矩阵形式：

$$\boldsymbol{Y}=\boldsymbol{W}\boldsymbol{x}=\begin{pmatrix}1 & 1 & 1 & 1 & \cdots & 1 & 1\\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{(N-2)} & \omega^{(N-1)}\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 1 & \omega^{(N-1)} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-2)(N-1)} & \omega^{(N-1)(N-1)}\end{pmatrix}\begin{pmatrix}x(0)\\ x(1)\\ \vdots\\ x(N-1)\end{pmatrix}$$

式中 $\omega=\exp(-2\mathrm{i}\pi/N)$。因此，$N$/3点处的傅里叶频谱计算如下：

$$Y(N/3)=(1\quad \omega^{N/3}\quad \omega^{2N/3}\quad \cdots\quad \omega^{(N-2)N/3}\quad \omega^{(N-1)/3})\begin{pmatrix}x(0)\\ x(1)\\ \vdots\\ x(N-1)\end{pmatrix}=\sum_{n=0}^{N-1}x(n)\omega^{nN/3} \tag{4}$$

根据式(4)，可以很容易地实现求N/3频率点处的傅里叶频谱，而不需要对整个窗口求傅里叶变换，从而节省运算时间，提高运算速度。用MATLAB语言实现求任意一点的傅里叶频谱的程序如下：

input ***X*** %待分析的数据向量

N %向量的长度

k%待求傅里叶频谱对应的频率点

$\omega = \exp(-\mathrm{i} * 2 * \pi * k) / N$

$Y = x(N-1)$

For n=N–2 downto 0

$Y = \omega * Y + x(n)$

End for

output Y

2　计算机实验

本文用计算机实验验证了所提出的快速预测方法的正确性和有效性。上面所叙述的快速预测方法用MATLAB语言实现。

2.1　实验结果

首先从Genbank基因数据库中选取Triticum Aestivum(bread wheat)的DNA序列(Accession Number: AB166873)作为实验对象。已知该DNA序列有7段编码区序列，分别位于186～258、336～351、477～607、717～837、952～1 055、1 172～1 293和1 384～1 440。用上述方法预测该DNA序列的编码区，实验结果如图1所示。再从Genbank基因数据库中选取Oryza Sativa (Japonica Cultivar-Group)的DNA序列(Accession Number: AB093593)作为实验对象。该DNA序列有7段编码区序列473～800、911～1 173、1 281～1 766、1 883～2 308、3 307～3 653、3 761～4 023和4 134～4 461。用该基因序列做实验，其结果如图2所示。此外，还对多种生物的大量基因组序列做了实验，都取得了相似的结果。

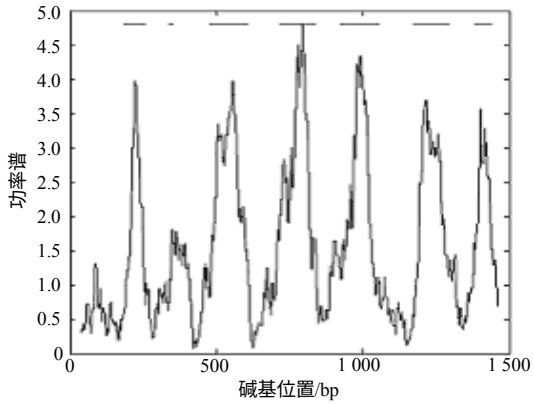

图1　预测Triticum Aestivum DNA序列

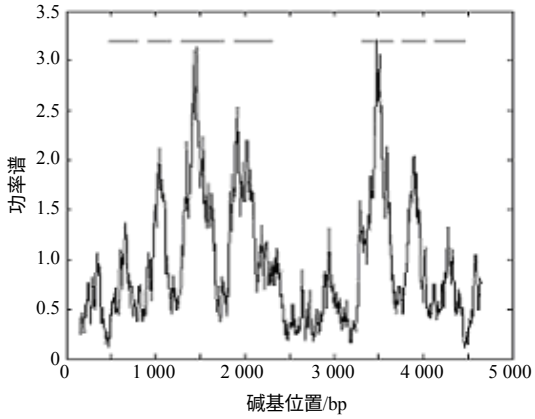

图2　预测Oryza Sativa DNA序列

2.2　实验结果分析与讨论

图1、2中所示的横线表示已知编码区序列的位置。由图1、2可看出，曲线的峰值位置刚好和编码区序列相对应。由此可见，利用本文所提出的方法预测这些序列的编码区取得了良好的效果，达到了准确预测和定位的目的。利用本文方法来预测基因组序列的编码区的优点有：(1) 运算快速：和传统的傅里叶变换的预测方法相比，由于不需要计算所有点的傅里叶频谱，而只计算f=1/3点的频谱，因此运算速度大大提高。特别是基因组序列越长，节省的运算时间就越多，更能显示出它的优越性；(2) 应用简单方便、适用面广：与神经网络的方法相比，由于该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。当分析一个新的或是不常见的序列时，由于没有该DNA序列的任何先验知识，利用神经网络的方法就无法对这类序列进行分析，但利用谱分析的方法仍然能够以较高精度预测这些序列的编码区。与基于相关指数的方法相比，用谱分析的方法预测较短序列的编码区性能更好。

当然，每种方法都不是完美无缺的，由于本文的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，但是有极少数(大约4%～5%)基因缺乏这种性质[6]，如S. cerevisiae中的Mating-Type基因和E.histolytica中的Amoebapore基因，对于这些基因而言，该方法就失去了效力。

3 结 束 语

通过理论分析和实验证实，利用本文提出的基于傅立叶技术的快速预测方法对基因组序列的编码区进行预测可取得良好的效果。该方法的显著优点是运算速度比利用FFT的方法快，容易应用，不需要基因组序列的任何先验知识；并且可同时实现基因的预测和定位。预测出编码区的大概位置，为进一步用实验方法精确定位编码区打下基础。正如文献[7]所指出的，通常难以用一种方法将各种生物DNA序列的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。面对世界范围内急剧增长的生物序列信息，相信对简便、快速、准确和适应性强的编码区预测方法的需求将会越来越大。

参 考 文 献

[1] Dodin G, Vandergheynst P, Levoir P, et al. Fourier wavelet transform analysis a tool for visualizing regular patterns in DNA sequences[J]. J Theo. Biol., 2000, 206: 323-326.
[2] Berger J A, Mitra S K, Carli M, et al. Visualization and analysis of DNA sequences using DNA walks[J]. Journal of the Franklin Institute, 2004, 341: 37-53.
[3] Buldyrev S V, Goldberger A L, Havlin S, et al. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis[J]. Physical Review, 1995, 51(5): 5084-5091.
[4] Tiwari S, Ramachandran S, Bhattacharya A, et al. Prediction of probable genes by Fourier analysis of genomic sequences[J]. CABIOS, 1997, 13(3): 263-270.
[5] Shepherd S J, Kay N D, Van Eetvelt P. An efficient algorithm for computing genetic spectra[C]// Oxford Bioinformatics Forum, Oxford, UK, 2003: 11-15.
[6] Chechetkin V R, Turygin A Y J. Study of correlation in DNA sequences[J]. Theo. Biol., 1996, 178: 205-217.
[7] Fickett J W. The gene identification problem: An overview for developers[J]. Comp. Chem., 1996, 20: 103-118.

编 辑 孙晓丹

(上接第806页)

由表1可以得出，自体库选过小，会造成单抗体的高扰动率，频繁更新抗体群，缺乏抗体的多样性，覆盖范围减小；而自体库选过大，会造成训练网络的时间急剧增多。本文采用200条为自体库大小，然后通过新型网络模型对这30万条数据记录进行检测，并与单免疫算法模型和传统的抗体网络模型进行对比。虽然此网络模型在时间上略逊于其他两种已知算法模型，但在准确率上却有明显的提高，如表2所示。

3 结 束 语

本文构建了基于免疫算法和神经网络的新型抗体网络，针对传统BP神经网络在入侵检测应用中学习性能的不足，引入免疫算法原理，对已有的抗体网络进行改造。通过对网络模拟数据集的测试，相对于单免疫网络和传统的抗体网络，检测效率和学习性能有明显的提高。

参 考 文 献

[1] 赵俊忠, 黄厚宽. 免疫机制在计算机网络入侵检测中的应用[J]. 计算机研究与发展, 2003, 40(9): 1293-1299.
[2] 吴 知, 许家玷. 免疫原理在多Agent入侵检测系统中的应用[J]. 电子科技大学学报, 2005, 6(3): 381-384.
[3] D Castro L N, Von Zuben F J, de Deus J G A. The construction of a boolean competitive neural network using Ideas from Immunology[J]. Neurocomputing, 2003, (50c): 51-85.
[4] Kim J, Peter J B. Towards an artificial immune system for network intrusion detection:An investigation of dynamic clone selection [J/OL]. IEEE2002, 0-7803-7282-4/02, 2005-10-21.
[5] D`haeseleer P, Forrest S. An immunological approach to change detection: algorithms, analysis and implications[C]//IEEE Symposium on Research in Security and Privacy, Oakland, 1996.

编 辑 熊思亮