DOI:10.13657/j.cnki.gxkxyxb.2013.01.022

基因预测中的信噪比计算新模型
Calculation of Signal-to-Noise Ratio in Gene Prediction

万芷君,明 媚,王 婷,栗丽兵
WAN Zhi-jun, MING Mei, WANG Ting, LI Li-bing

(桂林电子科技大学计算机科学与工程学院,广西桂林 541004)
(School of Computer Science and Engineering of Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China)

摘要:以快速傅立叶变换(FFT)求出DNA序列功率谱后,用重复计算实验的办法,对大量DNA序列中核苷酸的分布情况进行统计,得出综合考察变量,然后以核苷酸分布频数和DNA序列长度为自变量,建立Z-curve映射规则得到DNA序列信噪比的新模型,并进行实例分析。新模型不需要对序列进行离散Fourier变换(DFT),而且不要求DNA指示序列长度必须为3的倍数,应用范围更大。

关键词:信噪比 基因预测 数字映射 功率谱
中图法分类号:TP274 **文献标识码**:A **文章编号**:1002-7378(2013)01-0011-03

Abstract: Fast Fourier Transform (FFT) was used to get the power spectrum. By means of repeating computational experiments, the distribution frequency of nucleotides distribution was statistically analyzed in a large number of DNA sequences to get a aggregate variable. Then let the distribution frequency and the length of the DNA sequence be the independent variable, the SNR under Z-curve was calculated to establish the new model, which avoids the constraints of DNA sequence length and Discrete Fourier Transform (DFT) in traditional methods, so it's easy to popularize.

Key words: signal-to-noise ratio, gene prediction, digital mapping, power spectrum

DNA化学名称为脱氧核糖核酸(Deoxyribonucleic acid,缩写为DNA),承载着生物体的遗传信息。DNA序列由腺嘌呤(Adenine, A),鸟嘌呤(Guanine, G),胞嘧啶(Cytosine, C),胸腺嘧啶(Thymine, T)这四种核苷酸(nucleotide)符号按一定的顺序连接而成。真核生物的DNA序列被划分为许多间隔的片段,其中参与编码蛋白质的部分称为外显子(Exon),不参与编码的部分称为内含子(Intron)[1]。基因预测即对给定的DNA序列,识别出其中的外显子。这是一个尚未完全解决的问题,也是当前生物信息学的一个最基础、最首要的问题。目前大多数研究者都采用信号处理与分析的方法(即频谱分析)来进行基因预测。其主要思想是利用离散Fourier变换(DFT),对数值化映射后的基因序列进行频谱分析,以其信噪比的值来区分编码区和非编码区[2]。但是对于很长的DNA序列,在计算信噪比时,离散Fourier变换(DFT)的总体计算量太大,有必要寻找更佳的方法来计算信噪比。本文主要探索计算信噪比的新模型,并通过实例检验新模型的有效性。

收稿日期:2012-12-20
修回日期:2013-01-08
作者简介:万芷君(1987-),女,硕士研究生,主要从事网络工程,非线性规划理论及算法研究。

1 信噪比计算的思路

对基因进行预测之前,需根据一定的规则将DNA序列的四种核苷酸排列结构映射成相应的数值序列,得到DNA序列的数字映射之后,再考察两个非常重要的指标:功率谱和信噪比。常用的数字映射主要有Voss映射和Z-curve映射,由于Z-curve映射的维度比Voss映射低,运算时间少,因此我们以Z-curve映射序列为标准对功率谱和信噪比进行研究[3]。

Z-curve 映射的功率谱定义[4]为

$$P_z[k]=|\Delta X[k]|^2+|\Delta Y[k]|^2+|\Delta Z[k]|^2, \quad (1)$$

其中 $\Delta X[k]$,$\Delta Y[k]$ 和 $\Delta Z[k]$ 分别表示数字序列 $\Delta x[n]$，$\Delta y[n]$ 和 $\Delta z[n]$ 的离散傅立叶变换。

Z-curve 映射的信噪比[4]为

$$R_Z=\frac{P_Z[\frac{N}{3}]}{\bar{E}}=\frac{\left|\Delta X[\frac{N}{3}]\right|^2+\left|\Delta Y[\frac{N}{3}]\right|^2+\left|\Delta Z[\frac{N}{3}]\right|^2}{\bar{E}}, \quad (2)$$

其中 $\bar{E}=\frac{\sum_{k=0}^{N-1}P_Z[k]}{N}$ 是 Z-curve 映射的平均功率谱。

利用离散 Fourier 变换(DFT)[4]求序列的功率谱时，在 DNA 序列长度 N 很大的情况下，求其指示序列的 DFT 变换要完成 $N\times N$ 次复数乘法和 G_p 次复数加法，其计算量相当大，需要消耗大量的计算时间。而且(2)式要求 DNA 序列长度必须为 3 的倍数，否则所求峰值失准。所以考虑用快速傅立叶变换(FFT)[5]求功率谱序列，以缩短计算时间。

计算信噪比时，考虑到碱基的 3-周期性[6]是用于基因识别的一个重要特征信息，以及密码子使用的偏向性现象，即说明核苷酸符号 $b\in I=\{A,T,G,C\}$ 或其组合序列出现在该序列的 $0,3,6,\cdots,N-3$ 与 $1,4,7,\cdots,N-2$ 以及 $2,5,8,\cdots,N-1$ 等位置上的频数与信噪比之间有着密切联系[7]，而核苷酸在某个位置上出现的频数是比较容易统计的。再考虑到 Z-curve 映射的三维序列分别表示嘌呤、氨基类、存在弱氢键(以 1 为准)的核苷酸的分布，即表示的是三类核苷酸的分布。基于此，我们统计出这三类核苷酸在序列的上述第一、第二、第三个子序列上出现的频数，考察它们与 $k=\frac{N}{3}$ 处的功率谱值以及整个序列 S 的总功率谱的平均值之间的关系，建立新的信噪比计算模型，再利用 Matlab 软件进行数值拟合求出模型参数。

2 信噪比计算的新模型及求解

2.1 模型建立

建立 Z-curve 映射三维序列表示的三类核苷酸的分布与其对应 DNA 序列信噪比之间的联系，考虑创建能反映这三类核苷酸分布的综合变量。设 AG、AC、AT 分别表示 Z-curve 映射三维序列表示的三类核苷酸。设 F_{x1}、F_{x2}、F_{x3} 分别表示第一、第二和第三密码子位置上核苷酸 $x\in(AG,AC,AT)$ 的发生频数(在 Z-curve 映射三维序列中，即为每一行序列中 1 的个数)。又由于核苷酸在三个密码子位置上分布的非均衡性导致 DNA 序列的 3-周期性，即导致 $k=\frac{N}{3}$ 处的功率谱峰值，因此考虑每类核苷酸在密码子位置上发生频数的方差作为变量，测量密码子位置上核苷酸分布变化。

x 类核苷酸在三个密码子位置上发生频数的方差表示为

$$\sigma_x=\sum_{i=1,2,3}(F_{xi}-\frac{1}{3}\sum_{j=1,2,3}F_{xj})^2。$$

则三类核苷酸在所有密码子位置上发生频数的方差表示为

$$F_3=\sum_{x=AG,AC,AT}\sigma_x=\sum_{x=AG,AC,AT}\sum_{i=1,2,3}(F_{xi}-\frac{1}{3}\sum_{j=1,2,3}F_{xj})^2。$$

设 F_3 与 $P_z[\frac{N}{3}]$ 的关系表达式为

$$P_z[\frac{N}{3}]=(A)F_3,$$

其中 (A) 有待模型来求解

考察平均功率谱 $\bar{E}$ 和每个密码子位置上核苷酸分布之间的关系。因为 F_{xi} 表示第 i 个密码子位置上核苷酸 x 的发生频数，那么明显有

$$F_{AGi}+F_{ACi}+F_{ATi}=\frac{N}{3},i=1,2,3。$$

第 i 个密码子位置上三类核苷酸分布频数的方差表示为

$$D_i=\sum_{x=AG,AC,AT}(F_{xi}-\frac{N}{3})^2,$$

则 3 个密码子位置上三类核苷酸分布频数的方差表示为

$$F_c=\sum_{i=1,2,3}\sum_{x=AG,AC,AT}(F_{xi}-\frac{N}{3})^2。$$

设 F_c 与 $\bar{E}$ 的关系表达式为

$$\bar{E}=(B)F_c,$$

其中 (B) 有待模型求解。则信噪比模型表示为

$$R_Z=\frac{P_Z[\frac{N}{3}]}{\bar{E}}=\frac{(A)F_3}{(B)F_c}。$$

2.2 模型求解

为了求出模型中的参数 A 和 B，考虑用计算实

验的方法,并对多组数据进行数值拟合。

取 AB304259.1(酿酒酵母 ATP1a,ATP1b,为 F1F0-ATP 酶复杂,完整的 CD ATP1c 基因)中的 7 个 DNA 序列样本,AF100306.1(线虫粘粒 T24C4,完整序列)中 17 个 DNA 序列样本和 NC_012920_1(人线粒体全基因组)中 23 个 DNA 序列样本,一共 47 个样本为实验样本(基因序列样本均采集于生物数据网站:http://www.ncbi.nlm.nih.gov/guide/)。

首先分别求得其 Voss 映射序列,再变换得到 Z-curve 映射序列。然后用 Matlab 软件编写读频程序,分别得到每组 Z-curve 映射三维序列中,三类核苷酸在每个密码子位置的分布频数矩阵,即

$$\begin{pmatrix} F_{AG1} & F_{AG2} & F_{AG3} \\ F_{AC1} & F_{AC2} & F_{AC3} \\ F_{AT1} & F_{AT2} & F_{AT3} \end{pmatrix}。$$

得到 47 个频数矩阵之后,即可分别求其对应的 DNA 序列的 F_3 和 F_c。根据模型由 Matlab 得到 47 组 Z-curve 映射序列中三类核苷酸在所有密码子位置上发生频数的方差 F_3 和三个密码子位置上三类核苷酸分布频数的方差 F_c。

再在 Matlab 中由快速傅立叶变换(FFT)对 47 组 Z-curve 映射序列求功率谱序列,由此得到 $P_z[\frac{N}{3}]$ 和 $\bar{E}$ 的统计值。

以 F_3 为 x 轴,以 $P_z[\frac{N}{3}]$ 为 y 轴,由 47 组数据在 Matlab 中进行非线性拟合,得到曲线如图 1 所示。

由此得到以 F_3 为自变量,$P_z[\frac{N}{3}]$ 为因变量的关系式为

$$P_z[\frac{N}{3}] = -2.434272362842894e-05 * F_3^2 + 1.292877133985642 * F_3 + 7.775670761551705e+02,$$

即解得(A)。

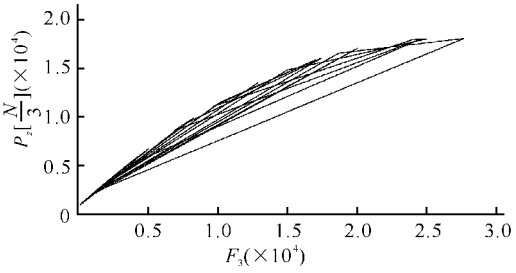

图 1 $F_3 \sim P_z[\frac{N}{3}]$ 曲线

再以 F_c 为 x 轴,以 $\bar{E}$ 为 y 轴,由 47 组数据在 Matlab 中进行非线性拟合,得到曲线如图 2 所示

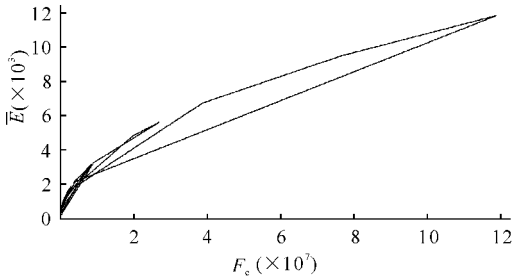

图 2 $F_c \sim \bar{E}$ 曲线

由此得到以 F_c 为自变量,$\bar{E}$ 为因变量的关系式为

$$\bar{E} = [(1.196073803300283 * Fc - 2.268927020220067e+04)]^{\frac{1}{2}},$$

即解得(B)。

则信噪比公式为

$$R_Z = \frac{P_Z[\frac{N}{3}]}{\bar{E}} = (-2.434272362842894e-05 * F_3^2 + 1.292877133985642 * F_3 + 7.775670761551705e+02) / [(1.196073803300283 * Fc - 2.268927020220067e+04)]^{\frac{1}{2}}。 \quad (3)$$

3 模型检验

选取一段序列长度为 5800 的 DNA 序列为检验样本,按照定义对信噪比求解结果为 $R = 0.9843$。利用 FFT 变换求得 DNA 序列的功率谱序列,在功率谱序列中取 $k = 1933$($N = 5800$, $5800/3 \approx 1933$)处的功率谱值,得 $P[\frac{N}{3}] = 5.7090e+003$,在 Matlab 软件中用 mean 函数对功率谱序列进行求平均,得 $\bar{E} = 5.8000e+003$,两者相除即得 R。根据模型中所给算法求出其 F_3、F_c、$P_z[\frac{N}{3}]$ 和 $\bar{E}$ 值,代入公式(3)得到信噪比 $R_z = 0.6669$。

显然按照定义求解与模型求解得出的 R 值是有差异的,这主要是因为检验样本和数值拟合实验所取样本序列长度大多(比如 5800)不是 3 的倍数,所求 $P_z[\frac{N}{3}]$ 不是准确峰值,影响其信噪比求解的准确性。但是两者差距不是很大,且都小于 1(一般取 2 为阈值),所以不影响外显子的判断,因此新模型是有效的。

(下转第 16 页)

/＊＊＊＊＊＊＊＊＊＊网页显示部分＊＊＊＊＊＊＊＊＊＊＊＊/
}catch(Exception e){……}。

(4)启动 Tomcat 服务器。在地址栏里输入 http://localhost:8080/axis/RequestTest.jsp,即可以看到运行结果。

3 结束语

本文研究 Web 服务下.NET 和 J2EE 的应用集成,给出了两者互操作的模型,为现有的企业实现不同开发框架下异构资源的整合提供了一定的参考。如今,Web 服务已基本上覆盖了传统分布式计算技术的应用领域,但是更加复杂的应用领域要求将现有不同的应用程序无缝的组合起来,形成功能更加强大,更完善的应用程序,这还需进一步的研究。

参考文献:

[1] 杨德华.利用 J2EE 实现 Web 服务模型与应用[J].计算机工程与应用,2004,9:123-125.
[2] Champ ionM,Ferris C,et al. Web Services Architecture [R]. http://www.w3.org/TR/2002/WD2ws2arch220021114, 2002.
[3] 郑小平. NET 精髓-Web 服务原理与开发[M].北京:人民邮电出版社,2002.
[4] 刘宏.. NET 与 J2EE 在 Web Service 领域之比较[J].电脑学习,2004,4:2-3.

(责任编辑:尹 闯)

(上接第 13 页)

本文方法在计算功率谱时引进 FFT 算法,减少了传统 DFT 变换的计算次数;在计算信噪比时,通过重复计算实验,考察 Z-curve 映射序列中反映出的三类核苷酸的分布与信噪比的关系。结果显示,与外显子判别密切相关的信噪比的值可以通过该 DNA 序列相应的数字映射到序列中,由核苷酸或其组合的分布情况求得。该方法避开了傅里叶变换的繁冗工作,而且此方法不要求 DNA 序列长度必须为 3 的倍数,更利于推广和应用。

参考文献:

[1] Yin Changchuan,Stephen S-T YAU. A Fourier characteristic of coding sequences:origins and a non-Fourier approximation[J]. Journal of Computational Biology, 2005,12(9):1153-1165.
[2] 田元新,陈超,邹小勇,等.外显子周期三行为特征的研究[J].化学学报,2005,63:1215-1219.
[3] 邵建峰,严晓华,邵伟,等. DNA 序列信号 3-周期特性[J].南京工业大学学报,2012,34(4):133-137.
[4] 马玉韬,张成,杨泽林,等. DNA 映射方法对蛋白质编码区预测准确率的影响[J].安徽农业科学,2012,40(6):3234-3238.
[5] 刘小群,周云波.基于 Matlab 的 DFT 及 FFT 频谱分析[J].山西电子技术,2012,4:48-49.
[6] 马玉韬,车进,刘大铭.基于傅里叶分析的蛋白质编码区预测中功率谱密度计算方法研究[J].宁夏大学学报,2011,32(2):134-138.
[7] 张志涌,杨祖. MATLAB 教程[M].北京:北京航空航天大学出版社,2012.

(责任编辑:尹 闯)