在 DNA 序列上寻找蛋白质基因的计算机程序

何东明　陈农安
（中国科学院上海生物化学研究所）

测定 DNA 分子的核苷酸排列顺序的快速方法已经出现，在一个 DNA 分子的序列被阐明后，如何尽快地搞清分子上哪些区域是蛋白质的编码区将是一个重要的问题。近年内，Staden 在这方面做了一些工作[1,2]。

程 序 介 绍

我们在 TRS-80 I 型微计算机上实现了一组从 DNA 分子的核苷酸顺序出发来预测蛋白质基因区的计算机程序。程序是用 Basic 语言写的，它能够把计算结果在 WX 4671 型绘图仪上输出，绘制出使人一目了然的基因分布图。

下面简单地介绍一下程序的原理及其作用：

程序 GENE 1　在组成蛋白质的 20 种氨基酸中，18 种氨基酸具有简并密码子，其简并密码子中的各个密码子在不同生物种类的 DNA 分子中出现的频率是不一样的[3]；并已发现，属于同一生物种类的 DNA 分子中的蛋白质基因区在使用简并密码子时具有一致的偏爱性，而在其它区域却未发现这种现象。

GENE 1 就是基于这种偏爱性，并根据用户提出的一个标准基因的位置，采用 Bayes 方法进行统计分析来确定其它蛋白质基因的位置。

GENE 1 首先统计出 64 个密码子中的每一个在标准基因中出现的频率 f_{abc}，然后取其自然对数 $F_{abc} = \ln f_{abc}$，当 $f_{abc} = 0$ 时，取 $F_{abc} = \ln(1/25)$。一个 DNA 序列有三个三联体链（frame），我们取 25 个三联体作为一个窗口，每一次窗口向前移动一个三联体，对窗口所在的每一位置，对三个三联体链求值：

$$(1)\ H_1 = \sum_{abc} F_{abc},\quad H_2 = \sum_{bca} F_{bca},$$

和D均为 0.05—0.5 mg，维生素E为 0.5—5.00 mg，维生素K是 0.10—1.00 mg，在此范围内用微量注射器吸取不同量注入柱内，所得四种维生素峰强度与浓度间均呈线性关系，试样的测定按外标法计算（见表 1）。

表 1 说明人胎盘制剂中维生素A的含量为 5.5 微克/克，高于牛胎盘制剂（2.5 微克/克），两者相差约一倍。维生素D在人胎盘制剂中是 9.9 微克/克，在牛胎盘制剂中含 9.6 微克/克，两者差别不大，而维生素E和维生素K与上述情况相反，牛胎盘制剂中E比人胎盘制剂的E几乎高 16 倍，K高约 1 倍。

回收率实验。维生素A取量 4.8 微克，回收率(五次实验平均值；下同)为 97.1% ±4.7%；维生素D取量 3.2 微克，回收率为 100.2% ± 2.2%；维生素E取量 23.2 微克，回收率为 96.5% ±2.5%；维生素K取量 24.0 微克，回收率为 95.4 ± 6.8%。

天津杏林生物药厂傅一心教授对本工作大力支持谨此致谢。

参 考 文 献

[1] DoLan, J. W. et al.: *J. Chromatogr. Sci.*, **16**, 616 (1976).
[2] Huguette Cohen et al.: *J. Agric. Chem.*, **26** (5), 1210, 1978.
[3] T. Van De Weerdhof et al.: *J. Chromatogr.*, **83**, 455, 1973.
[4] Pellerin, F. et D. Dumitrescu.: *Talanta*, **27** (3), 243, 1980.

[本文于 1984 年 5 月 29 日收到]

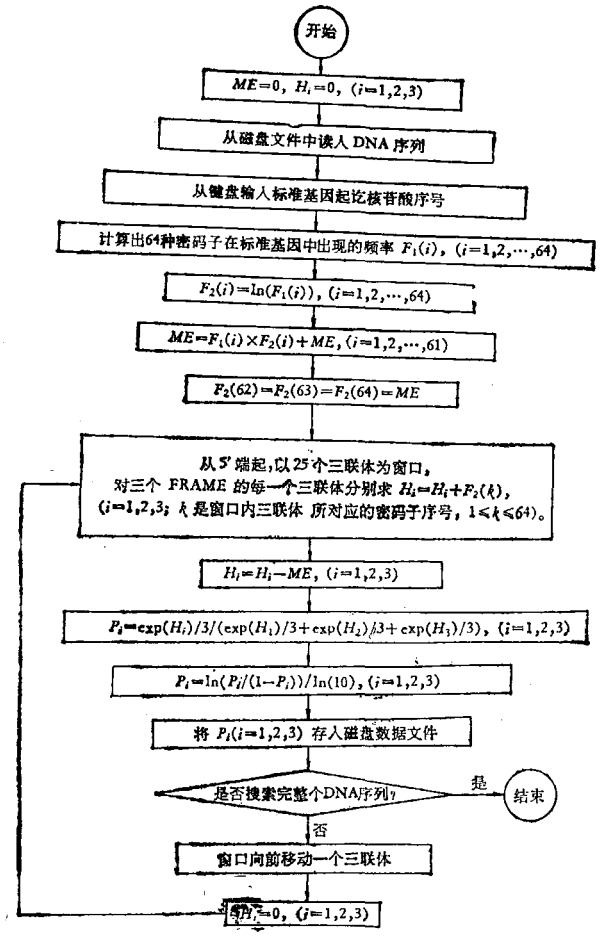

图 1　GENE 1 的框图

$$H_3 = \sum_{\text{cab}} F_{\text{cab}};$$

其中：abc、bca 和 cab 分别表示在窗口中第一、第二和第三三联体链中出现的密码子。

(2) $K_1 = H_1 - \text{MEAN}$，$K_2 = H_2 - \text{MEAN}$，$K_3 = H_3 - \text{MEAN}$；

其中：MEAN 是标准基因中 64 个密码子的 f_{abc} 与 F_{abc} 的乘积之和。

$$(3)\ P_1 = \frac{1}{3} \cdot e^{K_1} \Big/ \left(\frac{1}{3} \cdot e^{K_1} + \frac{1}{3} \cdot e^{K_2} + \frac{1}{3} \cdot e^{K_3}\right),$$

$$P_2 = \frac{1}{3} \cdot e^{K_2} \Big/ \left(\frac{1}{3} \cdot e^{K_1} + \frac{1}{3} \cdot e^{K_2} + \frac{1}{3} \cdot e^{K_3}\right),$$

$$P_3 = \frac{1}{3} \cdot e^{K_3} \Big/ \left(\frac{1}{3} \cdot e^{K_1} + \frac{1}{3} \cdot e^{K_2} + \frac{1}{3} \cdot e^{K_3}\right)。$$

(4) $y_i = \ln[P_i/(1 - P_i)]/\ln(10)$，$(i = 1, 2, 3)$，

当 $P_i > 0.99999995$ 时取 $P_i = 0.99999995$，以免运算溢出

y_i $(i = 1, 2, 3)$ 就是 GENE 1 给出的运算结果，我们把这些数据储存在软盘上的数据文件中，以备重复使用。

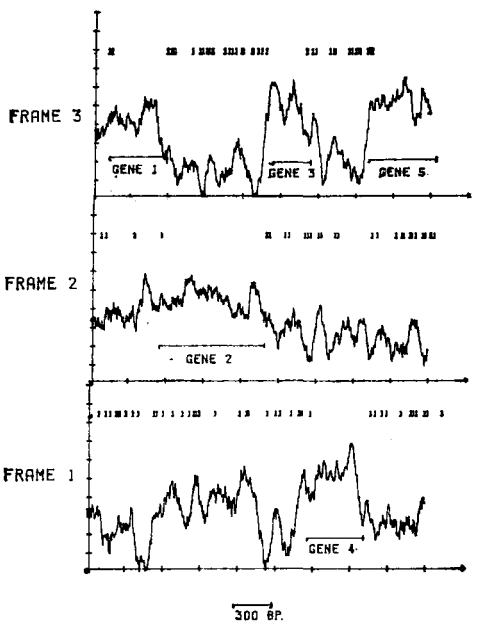

图 2　由 PLOT 1 绘制的大肠杆菌 unc 操纵子靠近启动子的部分蛋白质基因图

每一框上端的一排小 1 表示该三联体链(FRAME)中终止密码子出现的位置，基因就出现在无终止密码的高峰处

程序 PLOT 1　是一个绘图程序。为了在 WX 4671 型绘图仪上画出合适的基因图，PLOT 1 从 GENE 1 接受数据 y_i $(i = 1, 2, 3)$，然后进行数据转换：$Z_i = (y_i + 15) \times 25$。PLOT 1 以 Z_i 为纵坐标，窗口内第一个三联体在整个 DNA 序列中的三联体序号为横坐标。图 2 所示是由 GENE 1 计算，经 PLOT 1 绘制的大肠杆菌 unc 操纵子靠近启动子的部分蛋白质基因图[4]，采用的标准基因是基因 5。

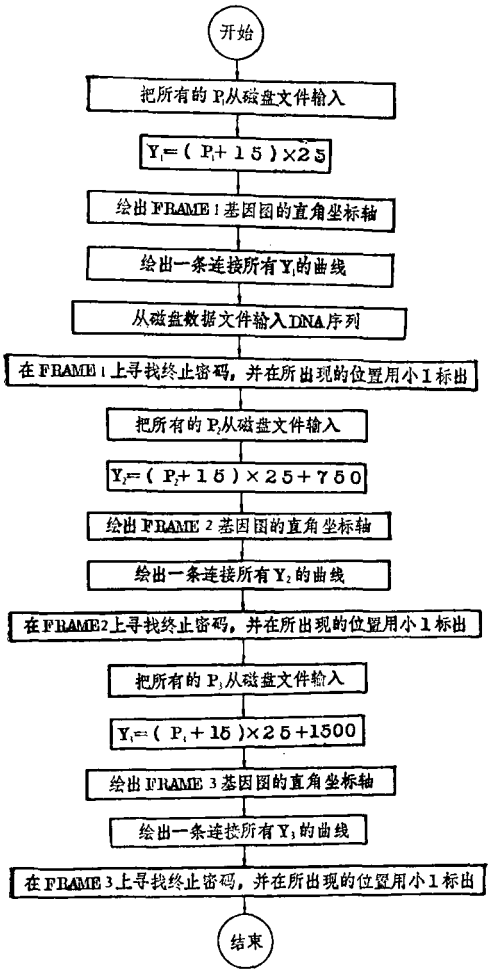

图 3 PLOT 1 的框图

P_1、P_2 和 P_3 的含义见图 1

程序 GENE 2 表 1 是从 EMBL 核酸顺序库中所有的蛋白质基因区统计出的在密码子三个位置上 A、T、C、G 出现的概率[2]。

表 1

	T	C	A	G
位置 1	0.1733	0.2104	0.2845	0.3317
位置 2	0.2748	0.2378	0.3094	0.1780
位置 3	0.2622	0.2782	0.2066	0.2530

由表 1 可知在密码子的三个位置上，A、T、C、G 出现的概率不是均等地为四分之一，GENE 2 就是利用这一特性来预测 DNA 序列上的蛋白质基因区的。

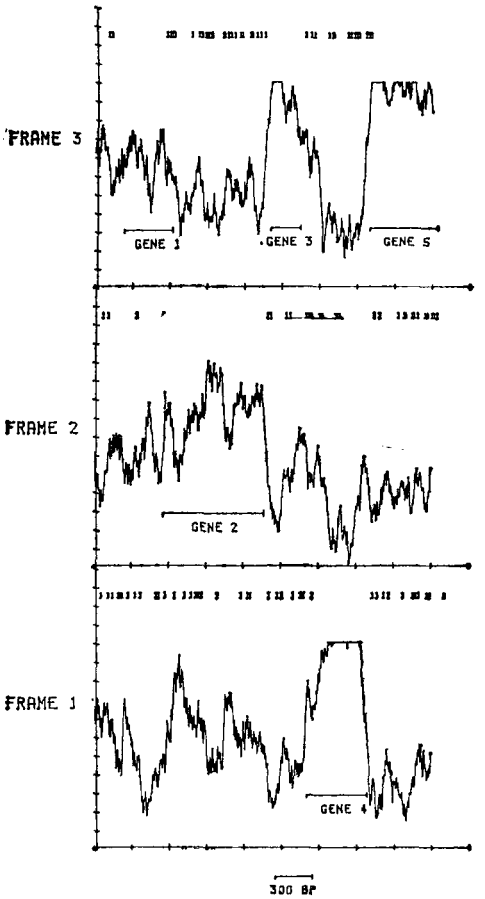

图 4 由 PLOT2 绘制的大肠杆菌 unc 操纵子靠近启动子的部分蛋白质基因图

说明详见图 1

GENE 2 的计算过程是这样的：把表 1 看作一个二维矩阵 E (3, 4)，行代表密码子中的三个位置，列代表四个核苷酸。同 GENE 1 的处理方法相似，GENE 2 也以 25 个三联体为窗口，每次向前移动一个三联体。对三联体链 1 (FRAME 1) 的每一窗口求出四个核苷酸分别在密码子的三个位置中每一位置出现的数目，以此值作为元素组成一个 3×4 矩阵 $P(3, 4)$，其行和列的次序及意义与 $E(3, 4)$ 相同，进而求值 $S_1=\sum_{i=1}^{3}\sum_{j=1}^{4}E(i,j)\times P(i,j)$。对三联体链 2 和链 3 (FRAME2 和 FRAME3)，同样可求出 S_2 和 S_3。这样，我们就可求值：$y_1=S_1/(S_1+S_2+S_3)$，$y_2=S_2/(S_1+S_2+S_3)$，$y_3=S_3/$

(下转第 34 页)

重视。我们的正常人末梢血淋巴细胞电泳组织图呈双峰形，与多数作者的结果一致[8-12]。由于我们的 T、B 淋巴细胞电泳组织图分别与总淋巴细胞的 HMC 及 LMC 峰重叠，这一事实证明了淋巴细胞电泳的 HMC 和 LMC 分别与 T、B 淋巴细胞密切相关。Bona 亦曾用细胞化学技术证实了它们之间的关系[3]。Hanjan 曾进一步证明了淋巴细胞电泳不仅可以准确地定量末梢血的 T、B 淋巴细胞，而且也能定量 T 细胞亚类[19]。他认为 HMC 中又可分为电泳率不同的两个组分，其中电泳率较高的部分与早期玫瑰花相关，电泳率较低部分与晚期玫瑰花相关。Uzgiris 等用 Laser Doppler Spectrometer 测量淋巴细胞电泳率，证明了 T 细胞两个电泳组分与 E-RFC 的亲合力相关[13]。

关于第三个群体 (IMC) 的意义，目前看法不一。Brown 等证明了它与 HMC 同属于 T 细胞群体[14]。而 Chollet 等证明这第三个群体既无 T 细胞标志，又无 B 细胞标志。

目前许多作者已将淋巴细胞电泳技术用于多种疾病的研究，如传染性单核细胞增多症[16]。类风湿性关节炎[14,17]、先天性免疫缺陷[18]及淋巴细胞性白血病的免疫学分型[11,15]等。这些研究表明了这一技术对于某些疾病的诊断、治疗监视以及发病机理的研究具有一定的意义。

参 考 文 献

[1] Kaplan, M. E. et al.: *J. Immunol. Methods*, **5**: 131, 1974.
[2] Saxon, A.: *J. Immunol. Methods*, **12**: 285, 1976.
[3] Bona, C.: *Biomedecine*, **22**: 97, 1975.
[4] Zeiller, K. et al.: *Physiol. Chem.*, **352**: 1168, 1971.
[5] Andersson, L. C. et al.: *J. Immunology*, **114**: 1226, 1975.
[6] Dumout, F. et al.: *J. Immunol. Methods*, **53**: 233, 1982.
[7] Bubenik, J. et al.: *J. Neoplas ma*, **28**: 517, 1981.
[8] Sabolovic, D. et al.: *Lancet*, **28**: 927, 1972.
[9] Wiig, J. N. et al.: *Clin. Exp. Immunol.*, **15**: 497, 1973.
[10] Donald, D.: *J. Immunol. Methods*, **6**: 151, 1974.
[11] Goldstone, A. H. et al.: *Clin. Exp. Immunol.*, **17**: 113, 1974.
[12] Chollet, ph. et al.: *J. Immunol. Methods*, **11**: 25, 1976.
[13] Uzgiris, E. E. et al.: *Eur. J. Cancer*, **15**: 1275, 1979.
[14] Brown, K. A. et al.: *Clin. Exp. Immunol.*, **36**: 272, 1979.
[15] Müller, M. et al. (eds): *Modern Trends in Cell Electrophoresis*, Dresden. GDR, 1978.
[16] Donald, D. C. et al.: *Clin. Exp. Immunol.*, **40**: 197, 1980.
[17] Brown, K. A. et al.: *Lancet*, **15**: 114, 1977.
[18] Durandy, A. et al.: *Clin. Exp. Immunol. Immunopathol.*, **4**: 440, 1975.
[19] Hanjan, S. N. S. et al.: *J. Immunol.*, **118**: 253, 1977.

[本文于 1985 年 3 月 25 日收到]

(上接第 77 页)

$(S_1+S_2+S_3)$。最后，把所有的 y_i $(i=1, 2, 3)$ 存入软盘。GENE 2 的框图与 GENE 1 的框图类似，就不再给出。

程序 PLOT 2 它的功能与 PLOT 1 类似，不同之处仅是：PLOT 2 是从 GENE 2 得到数据 y_i $(i=1, 2, 3)$，数据转换公式不同于 PLOT 1，这里是：$Z_i = y_i \times 5000 - 1500$。图 4 是由 GENE 2 计算，经 PLOT 2 绘制的大肠杆菌 unc 操纵子靠近启动子的部分蛋白质基因分布图。PLOT 2 的框图与 PLOT 1 的框图大同小异，也不再给出了。

用 PLOT 1 或 PLOT 2 来绘制基因图各有其优缺点。如用 PLOT 1，所绘制的图案较清晰准确，但需用户已掌握一个标准基因的位置，而 PLOT 2 则不需。

本工作得到了洪国藩教授的指导，在此表示感谢。

参 考 文 献

[1] Staden, R. and McLachan, A. D.: *Nuc. Acids Res.*, **10**, 141, 1982.
[2] Staden, R.: *Nuc. Acids Res.*, **12**, 551, 1984.
[3] Grantham, R., Gautier, C. and Gouy, M.: *Nuc. Acids Res.*, 8, 1893, 1980.
[4] Gay, N. J. and Walker, J. E.: *Nuc. Acids Res.*, **9**, 3919, 1981.

[本文于 1984 年 6 月 30 日收到]