遗传 HEREDITAS (Beijing) 13 (2): 9—11 1991

多个核酸序列的计算机比较分析

杨 子 恒

(北京农业大学畜牧系，100094)

本文介绍了一组应用于多个 DNA 序列比较分析的 BASIC 程序。程序 K1246P 利用 1-*P*、2-*P*、4-*P* 和 6-*P* 替代模型估计同源 DNA 序列分化后的核苷酸替代数 *K*；程序 DOT 利用点阵方法比较两个 DNA 或蛋白质序列，寻找序列间的相似性以及寡聚核苷酸、回文序列等初级结构特征。

关键词：DNA 序列；微机程序

DNA 序列分析技术的发展提供了大量的核酸资料，使得在分子水平上研究生物的进化过程成为可能。按照分子钟假说，同源序列分化后每位点平均核苷酸替代数 *K* 是与时间成线性关系的[14]，从而一直被用于推断物种间的进化关系，确定重大进化事件的年代等。常用的估计 *K* 值的方法假定核苷酸的替代过程是一时齐的马氏过程，并且根据所假定的突变率矩阵中未知参数的个数有 1-*P*、2-*P*、4-*P* 和 6-*P* 模型(分别为一、二、四、六参数模型)[2,4,5,6,12]。本文第一部分介绍用这四种方法估计替代数 *K* 的程序。

其次，众所周知基因组中存在着大量的重复序列，初级结构的比较发现蛋白质中亦存在着大量重复或近重复序列[1]。寻找这种相似性对于研究 DNA 和蛋白质的结构、功能和演化都具有重要的意义。Heiter 等[3]和 Steinmetz 等[11]分别于 1980 年和 1981 年提出的点阵方法 (dot matrix method) 可以有效地揭示两序列间可能存在的相似性，还可以找出寡聚核苷酸、回文序列等初级结构特征。本文第二部分介绍实现上述方法的 DOT 程序。

这组程序是在单序列分析程序 (包括序列打印、限制酶切位点查找、转译等)的基础上编制的，全部用 BASIC 写成，在 IBM 微型机上调试运行。

一、同源 DNA 序列的比较——分化后每位点替代数 *K* 的估计

估计替代数 *K* 的资料一般是经过匹配(alignment) 的同源序列[13]，其形式如图 1a 所示：一般第一个序列是基准序列 (consensus sequence)，其他序列中的碱基若与基准序列中的相同时，可用"·"来表示；"—"表示未确定碱基，而空格则表示序列间丢失或插入的关系。

这种资料首先经过一个复杂的"加工"程序 (PRO CESS)，将同源序列中的"·"变为相应的碱基；如在同源序列中发现有空格或"-"等异常字符时，则将所有序列中这个位点上的碱基删除掉；若 DNA 序列是编码蛋白质的结构基因，且三个位点分别考虑计算 *K* 值时，则将该碱基位置所在的整个密码三联体全部删除。

图 1b 是 a 中的序列经该程序加工后的结果。图 1a 上面的参数表明只要求考虑从第 4 碱基到第 45 碱基的部分，并且按编码序列 (N31 = 3，否则 N31 = 1) 加工；原来的序列以 3 (NSUB0) 个长为 16 (LSUB0) 的亚片段贮存，加工后则要求以长为 10 (LSUB1) 的亚片段贮存。结果得到的同源序列长为 21(L1)。实

Yang Ziheng: Computer Comparison Analysis of Multiple Nucleic Acid Sequences
本文于 1989 年 11 月 8 日收到。

```
N31 = 3
NSUB0 = 3    LSUB0 = 16    ICUT1 = 4    ICUT2 = 48
LSUB1 = 10
ACTATCATCATATATC  TAGATAGAGCTAGCGA  ATAGTAGTGT
....CT-  CATG....  ..AC.....A...A..  AC........
...C...C.........  ....-.......-....  ....
ACGT.....    .....  .....-.....-.....  ACT.......
CTG.............  .TG.....-.....  ....-.......
                        a
ACTTATCTAA  GAGATAGAGA  A
CTCTATCTAA  GAGCTAGCGA  A
TTCTATCTAA  GAGCTAGCGA  A
ATCTATCTTA  GAGCTAGCGA  A
L1 = 21 NSUB1 = 3
                        b
```

图 1　加工程序 PROCESS 的输入参数和待加工序列 (a) 以及加工后的同源序列 (b)
详细说明见正文。

```
N = 5    L = 345    LSUB = 240    NSUB = 2
K-1p Values for Codon Position 1
              1              2              3              4
2 0.063±0.024
3 0.390±0.072  0.436±0.078
4 0.165±0.041  0.133±0.036  0.362±0.068
5 0.092±0.030  0.063±0.024  0.436±0.078  0.143±0.038
K-1p Values for Codon Position 2
              1              2              3              4
2 0.000±0.000
3 0.054±0.022  0.054±0.022
4 0.045±0.020  0.045±0.020  0.102±0.031
5 0.009±0.009  0.009±0.009  0.063±0.024  0.054±0.022
K-1p Values for Codon Position 3
              1              2              3              4
2 0.693±0.117
3 1.528±0.341  -1.000±0.000
4 1.165±0.217   1.051±0.188  1.165±0.217
5 0.952±0.165   0.671±0.113  1.969±0.592  0.921±0.159
```

图 2　人、鸡、酵母、果蝇和鳟鱼的组蛋白 H2A 基因的替代数估计值与标准误
示程序 K1246P 的运行结果，估计值为−1 时表明估计方法不适用。

际中，建议都以 240bp 的亚片段贮存，此处仅为说明程序功能而已。

经过加工的同源序列可用于两两比较计算替代数 K。编码序列可将三个位点分别计算。其方法是首先对两个序列比较计数得到碱基对频率矩阵 $X = \{x_{ij}\}$，其中 x_{ij} 表示序列 I、II 中分别为碱基 i, j 的位点比例。K 值的估计可任选下述四种替代模型，即 Jukes & Cantor 的 1-P 方法[4]、Kimura 的 2-P 方法[5]、Takahata & Kimura 的 4-P 方法[12]以及 Kimura 和 Gojobori 等的 6-P 方法[6,2]。对于 1-P 和 2-P 模型还计算了 K 值估计的标准误差[7,5]（参见图 2）。此处顺便指出上面提到的文献 [6] 和 [12] 中所给的公式以及 [8] 中所给的 4-P、6-P 公式均有误。

二、一对核酸或蛋白质序列间相似性的比较：点阵方法

所谓点阵方法[3,11]是将两个长为 l_1、l_2 的序列(DNA 或蛋白质)在两个坐标轴上铺开，构成一个 $l_1 \times l_2$ 的矩阵 M。序列 I 的每个残基跟序列 II 的每个残基相比较，相同时 M 的元素取“·”，不相同时为空白[10]，如图 3 所示。如果两序列相同，则会形成一个从 $M(1,1)$ 到 $M(l_1, l_2)$ 的对角线，缺失或插入则造成一系列与主对角线平行的斜线；寡聚核苷酸(如聚 A、聚 T 等)则表现为与某个轴相平行的直线，并常常形成直方块(图 3a)；反向重复序列则构成垂直于主

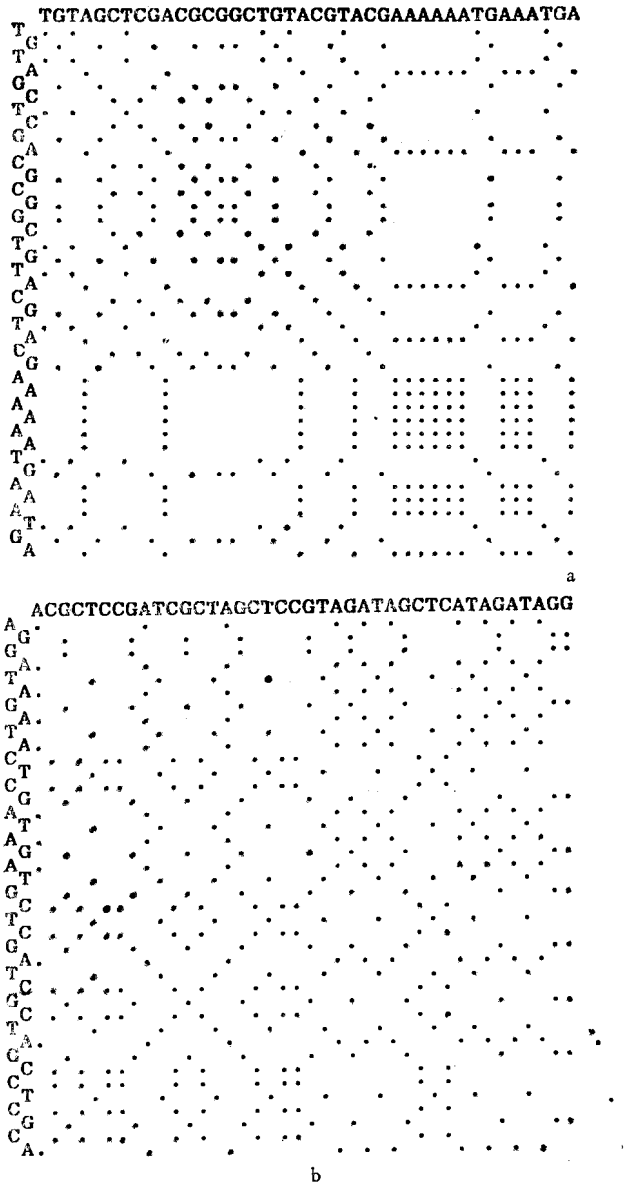

图 3　程序 DOT 的运行结果

图中所示是用字处理软件缩小行距后得到的。

对角线的直线(图 3b)。

由于核酸中只有四种碱基，所以随机的"噪声"很大，若每次对两个或三个核苷酸一起比较，则可以消除部分噪声，这种方法称之为滤波(filtering)。

其次，还可以假定对某些残基(核苷酸或氨基酸)不加区别，这种方法称之为 fudging，在分析远缘蛋白质序列时甚为有用。本程序中采用的 DNA 中模糊碱基字符是国际生物化学协会命名委员会提供的"标准字符"[9]，它们是：

Y:T 或 C; R:A 或 G; W:T 或 A;

M:C 或 A; K:T 或 G; S:C 或 G;

H:T 或 C 或 A(非 G); B:T 或 C 或 G(非 A);

D:T 或 A 或 G(非 C); V:C 或 A 或 G(非 T);

N:T 或 C 或 A 或 G。

当程序要求输入 Fudging word 时，可将上述适当的模糊碱基字符输入，如 RY 表示对两种嘧啶、两种嘌呤分别不加区别。这些特殊字符也应用于限制酶识认的序列片段的表示中，并且此处的算法也跟限制酶切位点查找的算法相似，都是采用一种模糊字符对残基编码的算法而实现的。

程序的编制是可以同时考虑滤波和 Fudging 的，不过作者建议滤波主要用于 DNA 序列的比较，而 fudging 则主要应用于蛋白质序列的比较。

M矩阵的输出是将横轴分为几个区段进行的，然后可将结果拼对起来查看，这样可以比较的序列长度可以很大。

参 考 文 献

[1] Doolittle, R. F.: 1987. *Biol. Bull.*, **172**: 269—283.

[2] Gojobori, T., Ishii, K. & M. Nei: 1982. *J. Mol. Evol.*, 18: 414—423.

[3] Heiter, P. A. et al.: 1980. *Cell*, **22**: 197—207.

[4] Jukes, T. H. & Cantor, C. R.: 1969. *In*: *Mammalian Protein Metabolism* (ed. H. N. Munro), Academic Press New York. pp. 21—123.

[5] Kimura M. &: T. Ohta, 1972. *J. Mol. Evol.*, 2: 87—90.

[6] Kimura, M.: 1980. *J. Mol. Evol.*, 16: 111—120.

[7] Kimura, M.: 1981. *Proc. Natl. Acad. Sci. USA*, 78: 454—458.

[8] Li, W. -H., Luo, C. -C. & C.-I. Wu: 1985. *In*: *Molecular Evolutionary Genetics* (ed. J. MacIntyre), Plenum Press, New York, pp. 1—94.

[9] NC-IUB: 1985. *Eur. J. Biochem.*, 150: 1—5.

[10] Novotny, J.: 1982. *Nucl. Acids Res.*, 10: 127—131.

[11] Steinmetz, M. et al.: 1981. *Cell*, 24: 125—134.

[12] Takahata, N. & Kimura, M.: 1981. *Genetics*, 98: 641—657.

[13] Wells, D. E.: 1986. *Nucl. Acids Res.*, 14: r119—r149.

[14] Zuckerkandl, E. & Pauling, L.: 1965. *In*: Evolving Genes and Proteins (eds. V. Bryson & H. J. Vogel), Academic Press, New York, pp. 97—116.