遗传 HEREDITAS (Beijing) 12 (6): 15—18 1990

一组 DNA 序列分析的微机程序

杨 子 恒[1)]

(甘肃农业大学畜牧系，兰州，730070)

本文介绍了作者编制的一组用于分析 DNA 序列资料的计算机程序。程序用 BASIC 语言写成，在 IBM 微型机上调试运行，包括序列打入、核苷酸频率统计、转译及限制酶切位点查找等几部分。

关键词：DNA 序列，计算机分析，BASIC 程序

随着确定的 DNA 序列的数量的日益增加，计算机已经成为整理、分析这些资料必不可少的工具。过去十年来国外不少人编制了对核酸与蛋白质资料进行管理与分析的软件[3-6]，而在国内这方面的报道尚不多见。为此笔者运用 BASIC 语言编制了一组核酸序列的分析程序，在 IBM 兼容机上调试运行，希望有助于国内分子生物学和分子进化领域的研究工作。本文报道单个序列分析的部分。

本组程序主要包括序列文件输入，序列打印，单、双或三核苷酸频率统计，一般字符串查找和限制酶切作图以及转译等几个部分。

（一）序列文件的输入

DNA 序列的输入可用任何标准字处理软件，也可以按用户给定的核苷酸频率由一个子程序随机产生。由于 BASIC 语言中要求字符串长度不超过 254，所以长的 DNA 序列就以 240bp 的长度分段贮存，以便于将来的分析处理，当然最后一个片段可能会短于 240bp。这样可以分析的序列长度就仅受内存的限制。输入序列时可以按较短的片段，如 40 或 60bp 输入，便于查对，然后再将它们连接成 240bp 的标准长度贮存。

（二）打印序列

这个程序以单链或双链的形式打印一个序列或其片段，并给出位置标号(图 1. 横线上为键盘输入，"↘"表示回车，下同。)

```
要打印的序列 ? test,↘
每行打印的核苷酸数? 60↘
开始位点? 0↘
结束位点? 0↘
单链(1)还是双链(2)? 2↘
打印序列 test, 从 1 到 100
         .        10        .        20        .        30
CCGTAGATTG AGCTATACCA TTCGAATTAG
GGCATCTAAC TCGATATGGT AAGCTTAATC
         .        40        .        50        .        60
AAGCTGAGGT CAGCTGGTCC CAGCTAATCT
TTCGACTCCA GTCGACCAGG GTCGATTAGA
         .        70        .        80        .        90
TAATCGTCAA ATCCACGAAA GCTTAGATCT
ATTAGCAGTT TAGGTGCTTT CGAATCTAGA
         .       100
CGCGCCAATA
GCGCGGTTAT
```

图 1 序列打印程序的输出结果

（三）单碱基、双碱基及三碱基频率统计

这个程序计算并打印一个序列或其片段中的碱基组成、双碱基频率及三碱基频率。双碱基频率用于考察碱基分布的独立性，是将所有相邻碱基组合计数来计算的，即长为 l 的片段中有$(l-1)$个双碱基组合。三碱基频率，即可能的密码子使用表可以用于判断一个序列中

Yang Ziheng: A Collection of Microcomputer Programs for the Analysis of DNA Sequences

1) 现在北京农业大学畜牧系攻读博士学位，邮政编码：100094。

本文于 1989 年 5 月 10 日收到。

要考查的序列? test↘

限制酶切图

```
     .    10        .    20        .    30        .    40        .    50        .    60
CCGTAGATTGAGCTATACCATTCGAATTAGAAGCTGAGGTCAGCTGGTCCCAGCTAATCT
     |                   ‖              |              |              |
   AluI              Asu II          AluI          Alu I          Alu I
                     Taq I
     .    70        .    80        .    90        .   100
TAATCGTCAAATCCACGAAAGCTTAGATCTCGCGCCAATA
                     ‖              |         |
                   Alu I          Mbo I     Fnu DII
                   Hind III
```

切割位置表

酶	识认的片段	切割位点
Alu I	AGCT	11,32,42,52,80
Asu II	TTCGAA	21
Eco PI	GGTCT	未找到
Eco RI	GAATTC	未找到
Fnu DII	CGCG	91
Hind III	AAGCTT	79
Hap I	GTTAAC	未找到
Kpn I	GGTACC	未找到
Mbo I	GATC	86
Taq I	TCGA	22
Xba I	TCTAGA	未找到
Xho I	CTCGAT	未找到

图 2　字符串查找程序的输出结果

可能的编码区域，或者用于确定不同外显子中的密码子使用情况。

（四）一般字符串查找与限制酶切作图

本程序用于在一 DNA 序列上寻找限制酶切位点、TATA 框或者其他任何由用户给定的碱基序列片段，以模拟基因工程中的限制酶消化实验。输出结果包括一个限制酶切图，其中指出酶切位点并在序列下面标注限制酶的名字，或者标记查找到的序列片段的开始位置并在下面标记片段的名字。此外还输出一个酶切位置表，其中列出酶(序列片段)的名称、所识别的序列片段以及切割位点(参见图 2)。

（五）转译

该程序将 DNA 序列（向前）或其互补链(向后)转译成氨基酸序列。可以使用通用密码词典(参见图 4 中的密码使用表)，也可使用哺乳动物线粒体密码词典[2]。氨基酸可用单字母符号，并跟三联体中间一个碱基对齐，终止密码用星号表示，也可使用三字母符号或汉字单字符号，其中蛋氨酸用大写 MET 表示。输出结果中每行打印的字符个数可任意给定。

这个程序有两种选择形式。第一种选择用于分析尚未确定编码区域的序列，程序给出所有三种可能的阅读格式 (open reading frames) 下的转译结果，通过考察其中起始密码和终止密码的位置、数目，并结合密码子使用的非随机性[1,7]，可以预测该序列或其互补链上的蛋白质编码区域。

另一种选择主要用于转译序列或其互补链上已知的编码区域。要求输入起始位点和终止位点(默认值为全长)。如有必要，还可以打印出转译区段内三个密码位点上的碱基组成，并给出根据四种碱基在三个位点上独立分布的假定计算的 χ^2 值（其自由度为 6，显著值为 $\chi^2_{0.05}=12.6$，$\chi^2_{0.01}=16.8$）以及密码子使用表。

图 3 和图 4 分别是这两种选择下运行的示

```
要转译的序列？test ↘
转译方向（0=向前，1=向后，& 2=两者）？2 ↘

转译 test

    .  10    .  20    .  30    .  40    .  50    .  60
CCGTAGATTGAGCTATACCATTCGAATTAGAAGCTGAGGTCAGCTGGTCCCAGCTAATCT
 P  *  I  E  L  Y  H  S  N  *  K  L  R  S  A  G  P  S  *  S
  R  R  L  S  Y  T  I  R  I  R  S  *  G  Q  L  V  P  A  N  L
   V  D  *  A  I  P  F  E  L  E  A  E  V  S  W  S  Q  L  I

    .  70    .  80    .  90    . 100
TAATCGTCAAATCCACGAAAGCTTAGATCTCGCGCCAATA
 *  S  S  N  P  R  K  L  R  S  R  A  N
  N  R  Q  I  H  E  S  L  D  L  A  P  I
L  I  V  K  S  T  K  A  *  I  S  R  Q

转译 test 的互补链

    .  10    .  20    .  30    .  40    .  50    .  60
TATTGGCGCGAGATCTAAGCTTTCGTGGATTTGACGATTAAGATTAGCTGGGACCAGCTG
 Y  W  R  E  I  *  A  F  V  D  L  T  I  K  I  S  W  D  Q  L
  I  G  A  R  S  K  L  S  W  I  *  R  L  R  L  A  G  T  S  *
   L  A  R  D  L  S  F  R  G  F  D  D  *  D  *  L  G  P  A

    .  70    .  80    .  90    . 100
ACCTCAGCTTCTAATTCGAATGGTATAGCTCAATCTACGG
 T  S  A  S  N  S  N  G  I  A  Q  S  T
  P  Q  L  L  I  R  M  V  *  L  N  L  R
D  L  S  F  *  F  E  W  Y  S  S  I  Y

是否还要转译 test (Y/N) ↘
```

图 3　转译程序 I 的输出结果

本图及图 4 中所用氨基酸中文单字缩写、英文单字母、三字母缩写意义如下：

苯，F，Phe；亮，L，Leu；异，I，Ile（异亮）；蛋，M，Met；
缬，V，Val；丝，S，Ser；脯，P，Pro；苏，T，Thr；
丙，A，Ala；酪，Y，Tyr；组，H，His；官，Q，Gln(谷酰)；
天，N，Asn（天酰）；赖，K，Lys；冬，D，Asp(门冬)；谷，E，Glu；
胱，C，Cys（半胱）；色，W，Trp；精，R，Arg；甘，G，Gly。
* 或 Ter 表示终止密码。

意结果。

可以看出，编制程序时充分利用了 BASIC 语言人机对话的特点，并采用中英文两种提示方式，即使没有计算机使用经验的用户亦可方便地使用。进一步的工作将包括多序列分析，即寻找同源性、序列间替代数估计等。

参 考 文 献

[1] Almagor, H.: 1985. *J. theor. Biol.*, 117: 127—136.

[2] Anderson, S.: 1981. *Nature*, 290: 457—464.

[3] Brutlag, D. L. et al.: 1982. *Nucl. Acids Res.*, 10: 279—294.

[4] Conrad, B. and D. W. Mount: 1982. *Nucl. Acids Res.*, **10**: 31—38.

[5] Staden, R.: 1977: *Nucl. Acids Res.*, 4: 4037—4051.

[6] Staden, R.: 1986. *Nucl. Acids Res.*, 14: 217—231.

[7] Staden, R. and A. D. MacLachlan: 1982. *Nucl. Acids Res.*, 10: 141—156.

要转译的序列? test↘

转译方向(0=向前,1=向后)? 1↘

起始位点? 6↘

终止位点? ↘

要打印碱基频率和密码使用表吗 (Y/N)? y↘

转译 test，从 6 到 98

```
     .    10        .    20        .    30        .    40        .    50        .    60
TATTGGCGCGAGATCTAAGCTTTCGTGGATTTGACGATTAAGATTAGCTGGGACCAGCTG
      A   R   D   L   S   F   R   G   F   D   D   *   D   *   L   G   P   A

          .    70        .    80        .    90        .   100
ACCTCAGCTTCTAATTCGAATGGTATAGCTCAATCTACGG
D   L   S   F   *   F   E   W   X   S   S   I   Y
```

碱基频率(卡方=9.262122)

密码位点 1: T = 11 C = 6 A = 4 G = 10

密码位点 2: T = 8 C = 4 A = 11 G = 8

密码位点 3: T = 8 C = 10 A = 9 G = 4

总和: T = 27 C = 20 A = 24 G = 22

密码子使用表

苯 TTT 2	丝 TCT 0	酪 TAT 1	胱 TGT 0
TTC 2	TCC 0	TAC 1	TGC 0
亮 TTA 0	TCA 1	* TAA 2	* TGA 0
TTG 0	TCG 0	TAG 1	色 TGG 1
亮 CTT 0	脯 CCT 0	组 CAT 0	精 CGT 1
CTC 1	CCC 0	CAC 0	CGC 0
CTA 1	CCA 1	官 CAA 0	CGA 1
CTG 1	CCG 0	CAG 0	CGG 0
异 ATT 0	苏 ACT 0	天 AAT 0	丝 AGT 0
ATC 1	ACC 0	AAC 0	AGC 3
ATA 0	ACA 0	赖 AAA 0	精 AGA 0
蛋 ATG 0	ACG 0	AAG 0	AGG 0
缬 GTT 0	丙 GCT 1	冬 GAT 3	甘 GGT 0
GTC 0	GCC 0	GAC 2	GGC 0
GTA 0	GCA 0	谷 GAA 1	GGA 2
GTG 0	GCG 1	GAG 0	GGG 0

是否还要转译序列 test(Y/N)? ↘

图 4 转译程序 II 的输出结果