IBM PC 机上的核酸顺序数据库及其检索方法

陈农安 夏志清 何东明
(中国科学院上海生物化学研究所)

提 要

我们在微型电脑 IBM PC 机上开发了一个核酸顺序数据库及其管理和分析系统。该库数据丰富且能更新,管理和分析软件系统功能强,结构设计合理、灵活,检索和管理比较方便。本文介绍了该库的作用及其各种检索方法。分析程序系统 NASA (Nucleic Acid Sequence Analysis) 将另文发表[1]。

核酸顺序数据库的用途

当前,分子生物学、基因工程等学科正日新月异地发展,随着 DNA序列测定方法的日益完善和快速,核酸顺序资料迅猛增长,建立核酸顺序的计算机数据库及其管理系统已成为迫切需要。由于被测定的 DNA 序列越来越长,有些杂志已不再刊载它们,而计算机核酸数据库能以磁盘和磁带等为媒介快速传递和交流信息,国际 CODATD(Committee of Data in Science and Technology) 组织为此提供了方便。

在含有几千、甚至几万个碱基的 DNA 序列上,只有使用计算机软件系统进行各种检索、统计和分析,才能充分理解和利用测得的 DNA 序列所含有的信息——限制性内切酶位点,蛋白质基因的位置及启动子和转录起始点等。

猿猴肉瘤病毒携带的癌基因 v-sis 与编码血小板来源的生长因子 (简称 PDGF) 的基因非常相似。这一发现不但第一次揭示了一个病毒癌基因与一个已知生理功能的细胞基因的相似性,而且为癌基因可能异常地产生了控制细胞正常生长的物质使细胞发生癌变的假说提供了富有说服力的证据。这一重要发现是加州大学化学系的 Doolittle 把 PDGF 序列送入电脑,通过与核酸资料库中的所有基因的比较和分析,发现了两基因的相似性[2]。

核酸顺序数据库的检索方法

起初,我们在 TRS-80 型微电脑上编制了一些核酸顺序分析程序[3], 在分子生物学研究中发挥了一定的作用[4]。

本文所要介绍的是我们在 IBM PC 微电脑上开发的一个完整的核酸顺序数据库及其管理和分析软件系统。该数据库拥有 8823 个顺序(sequences), 共 844 万碱基(到 86 年初),以后每年还能通过 CODATA 组织添加和更新数据。所有顺序放在硬盘中,约占 7.5 兆字节。管理和分析软件系统放在一个软盘上。该系统用 BASIC 语言编写,经编译成为可执行文件,以提高检索和处理速度。Menu I 是管理软件系统的主菜单,起动管理系统,用户首先见到的就是它,它共有八个选择项(图 1)

1. 按顺序名 (locus name) 进行检索。pBR322、M13 和 T4PTE114 等都是 locus name, 须注意的是在这里所有输入的 locus name 中的英文字母必须是大写体。当用户键入所要查找顺序的 locus name 后, 该顺序立即从硬盘调入内存并开始在屏幕上显示出来,如显示的内容较多, 向上移动太快以至来不及阅读,则可按 Ctrl 键+Num lock 键,这样显示就暂停,然后按任一键, 显示继续。作为例子, 键入 PBRS2421[5], 屏幕显示如例 1 (见图

```
**************************************
*                                    *
*          GENE DATA BANK            *
*                                    *
**************************************

                OPTIONS:

1------SEARCH SEQ. BASED ON LOCUS NAME
2------SEARCH SEQ. BASED ON DEFINITION
3------SEARCH SEQ. BASED ON AUTHOR'S NAME
4------SEARCH SEQ. BASED ON KEYWORD
5------PRINT SEQ. ON PAPER
6------EDIT SEQ. (TO BE USED BY NASA)
7------NASA PROGRAM
8------QUIT

  YOUR OPTION  1-8?
```

图 1 Menu I

2)。库内每个顺序文件中除贮有整个顺序外，还有该顺序的简短解释（或称定义，Definition）、有关的参考文献和作者名等，为分子生物学等学科的研究提供了方便。

2. 按用户给出的定义词(Definition Word)寻找出顺序名及其整个顺序定义，然后由顺序名查找顺序。不必把整个定义输入，只需几个定义词，键入的定义词越多，寻找的范围就越小，计算机给出的顺序名越少。例 2 (见图 3)是一寻找过程。图 4 是计算机的寻找结果。

3. 从作者名寻找出他所作出的顺序名。如寻找 UUu，E 作出的收集在库的所有顺序名，见例3(图 5)。

4. 按关键字 (Keyword) 寻找顺序名。关键字是把顺序分类后的类名称，由关键字找出的是该类的全部顺序名。如 Promoter，Oncogene 和 Ribosomal Protein S1 等就是关键字。以 Promoter 为例，见例 4 (图 6)。

5. 把顺序在打印机上输出，这里仅输出顺序，顺序前的参考文献和作者名等信息不输出。用户可要求每行打印多少个碱基，行间距离多少等。在选择项 1——按顺序名寻找顺序时，当键入顺序名后，按下回车键↵并立即按 Esc 键，这样就在打印机上输出顺序前的参考文献和作者名等信息，不输出顺序。如要连同顺序一起输出则可采用屏幕打印方法。选择项 5 和选择项 6 都要在选择项 1 后使用，否则被打印或编辑的顺序就成了不确定的了。

6. 用户给出 0—3 个字符或数字作为顺序号 (SEQ. No.)，系统就在 C 盘上建立一个名为 MYSEQXXX. ASC 的文件，×××是用户键入的顺序号。用户在此可进行显示和删除等项工作，把所需要的顺序段贮入上面建立的 MYSEQXXX. ASC 文件，以备 NASA 分析程序使用。这里的编辑功能主要就是显示和删除

```
                YOUR OPTION  1-8? 1
Enter Entry Name (LOCUS name) to search for ? PBRS242I
Group: Synthetic

PBRS2421   8/ 1/ 85          Entered
Definition:
Plasmid vector pBRS242 derived from insertion of

lac UV5 promoter- operator region into pBR322; region 5' to UV5 insertion.
Accession Number(s): K02386
Reference # 1 : Mol Gen Genet 189, 142-147 (1983)
Author(s): Savochkina,L.P., Retchinsky,V.O., Beabealashvilli,R.S.

PBRS2421      Accession # K02386          Number of records:  2
Segment  1  of  2
DS-DNA
CGTCTTCAAGAATTCTCATGTTTGACAGCTTATCATCGAATTC
Enter Entry Name (LOCUS name) to search for ?
```

图 2

```
Sequences are organized in following Groups

      1-----Primate
      2-----Rodent
      3-----Other Mammalian
      4-----Other Vertebrate
      5-----Invertebrate
      6-----Plant
      7-----Organelle
      8-----Bacterial
      9-----Structural RNA
      10----Viral
      11----Bacteriophage
      12----Synthetic
      13----Unannotated

    Group select   (1-13)? 1
Definition Word? Human
Definition Word? 7SL
Definition Word?
```

图 3

```
LOCUS NAME          Definition
HUM7SLR1  ----Human 7SL RNA pseudogene, clone p7L30.1.
HUM7SLR2  ----Human 7SL RNA pseudogene, clone p7L30.2.
HUM7SLRA  ----Human 7SL RNA pseudogene, clone p7L28.
HUM7SLRB  ----Human 7SL RNA pseudogene, clone p7LEM1.
HUM7SLRC  ----Human 7SL RNA pseudogene, clone p7L7.
HUM7SLRD  ----Human 7SL RNA pseudogene, clone p7L23.
HUM7SLRE  ----Human 7SL RNA pseudogene, clone p7L63.
```

图 4

两项，也就是保留用户所需要的顺序片段，然后贮存之。为了丰富编辑功能，我们在 NASA 系统中另行建立了一整套功能齐全的编辑命令，以供用户使用。在本选择项中，显示功能是指系统根据用户的要求显示由选择项 1 所确定的顺序的某一页，每一页以 1000 个碱基为单位，每行为 100 个碱基。显示完毕后系统问用户是否要进行编辑；如要，则进入编辑状态，否则将顺

```
YOUR OPTION 1-8? 3

Enter Author Name: ULLU,E
LAST NAME: ULLU          FIRST INITIAL: E

DROSG7SL   HUM7SLR1   HUM7SLR2   HUM7SLRA   HUM7SLRB   HUM7SLRC
HUM7SLRD   HUM7SLRE   HUMSR7SL   XENSR7SL
```

图 5

```
YOUR OPTION 1-8? 4

Enter Keyword Phrase: PROMOTER
AD2VAIPRO  AHYPS2CAT  BKVECR501  BKVECR522  BKVECR530  BKVECR531
BKVECR532  BSUPR      BSUSPRE    BSUVEGPRO  CELVIT1    CELVIT2    CELVIT4
CELVIT5    CELVIT6    CHKCON5E   CLODF13B   COLE1PRM   CRETBA1G   CRETBA21
CRETBA22   CRETBB1G   CRETBB2G   DROACT5C   DROHIS3P   DROHSPCAT  ECOAMPCF1
ECOAMPCF2  ECOBLIPR   ECOC625    ECODEOAB   ECODNAKA   ECOGLNA    ECOGLNAL
ECOGLNALG  ECOGROE    ECOHISCB   ECOHISSP   ECOIFNAWA  ECOILVBPR  ECOILVIHP
ECOLAC     ECOMELOP   ECOMETA    ECOMETLB1  ECONARPR   ECOOMPA    ECOPYRBI
ECOPYRBIB  ECORGNA    ECORGNABP  ECORGNB    ECORGNBP   ECORGNC    ECORGNDS1
ECORGNE    ECORGNG    ECORGNX1   ECOSPO1PR  ECOTACPRM  ECOTETPRO  ECOTGL6KP
ECOTGY1    ECOTGYPR   ECOTRPPO   ECOTRPPRO  HBVECOH80  HBVECOI10  HBVSV40PR
HHARGD     HUM1FN6S   KPNN1FL    MUSC1A2    MUSCRY62D  PBRS188V   PBRS240
PBRS2421   PBRS2422   PBRT7PRA   PBRT7PRB   PBRT7PRC   PFR10V     PSKS104V
PSKS105V   PSKS106V   PTRS3      RABHBBPR   RATPEC     RBTN1FH    RJAN1FDK
RJAN1FH    RMEN1FP4   RMEN1FPR1  RMEN1FPR2  RMEN1FPR3  SLMEB4     SPB2PREG
SPO1EP     SPO1EP230  SV40PBRA   SV40PBRB   SV40PBRC   SYNRNAPRA  SYNRNAPRB
T4PTE114   T4PTE122   T4PTE123   T5L1       T5L2       TOLXPRM    TOLXYLD
VACMLG     YSCCYC1    YSCGAL     YSCHIS3    YSCHSVIK   YSCHSVTKY  YSCODCD
YSCODCF    YSCT6L     YSGGALS1

Enter Keyword Phrase:
```

图 6

(下转第56页)

冲频率 50kHz，电压 5kV，样品浓度 10^{-4}mol/L，发射波长 435nm，测量表明 9，10-Diphenyl Anthracene 是单指数衰减，用矩方法去卷积得到的寿命是 6.8ns，这一结果与以前工作者得到的结果是一致的[4]。

图 3 是 9，10-Diphenyl Anthracene 的荧光衰减曲线。$L(t)$ 是仪器响应曲线，$R(t)$ 是荧光衰减实验曲线，为了检验拟合的质量，由计算机解出的 $F(t)$ 通过 (1) 式与 $L(t)$ 卷积计算出拟合曲线 $C(t)$，用点线与实验曲线同时画在半对数坐标系上，以便比较。

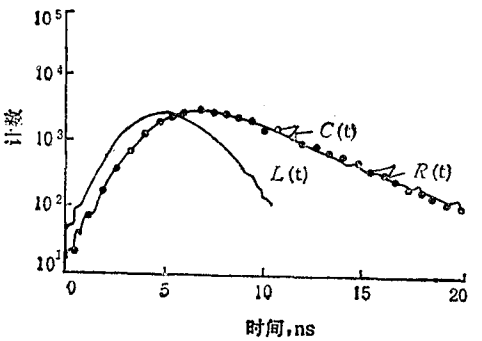

图 3 9,10-Diphenyl Anthracene 的荧光衰减曲线

本工作通过 9,10-Diphenyl Anthracene 荧光寿命的测量为例子，说明我们的系统对 ns 量级荧光寿命的测量与解算是可行的。由于配备了计算机，所以一旦样品的毫微秒荧光谱记录完成，就可以从计算机上求得样品的荧光寿命参数，具有快速、准确的优点。这对生物样品间的比较测量是很方便的。

本系统研制中得到中国科学院生物物理所计算机组的帮助，在此表示感谢。

参 考 文 献

1 彭程航等. 生物化学与生物物理进展，1987;(2): 49
2 Isenberg I *et al. Biophys J*, 1969;9:1337
3 Ygurabide J. *Methods in Enzymology*, 1972;26:496
4 Porter G. *Progress in Reaction Kinetics*, 1967;4:289

[本文于1988年12月16日收到]

(上接第59页)

序贮入 MYSEQXXX. ASC 文件。在编辑状态下，如操作 Alt + D 则进入删除状态，只要告诉系统用户所需顺序片段的第一个碱基的序号及最后一个碱基的序号；如操作 Alt + R 则退出编辑状态。

7. NASA (Nucleic Acid Sequence Analysis) 程序将另文发表[1]。

8. 退回到操作系统。

总的来说，当用户试图使用本系统时，应首先根据用户所掌握的有关顺序的信息——如顺序的作者名、顺序的生物来源或生物分类名等等，从选择项 2、3、4 入手，找出所要顺序的顺序名 (即 locus name)，然后由选择项 1. 找出该顺序，再经选择项 6 编辑后贮入 C 盘中的 MYSEQXXX. ASC 文件中，也可使用一下选择项 5。把该顺序打印出来，这样，余下的工作也许需要 NASA 系统帮忙了。

参 考 文 献

1 何东明，陈农安. 生物物理学报，1989; 5(1) (待发表)
2 桂建芳，张奇亚. 生命的化学，1984; 4(6),9
3 夏志清，陈农安. 生物化学与生物物理进展，1983;5:78
4 夏志清. 生物化学与生物物理进展，1984; 5: 83
5 Savochkina L P *et al. Mol Gen Genet*, 1983; 189: 142

[本文于 1988 年 10 月 4 日收到]