IDEAS 的引进和开发应用

吴加金
(军事医学科院基础医学研究所,北京)
朱伟雄
(军事医学科院计算中心,北京)
曹德贤
(中国医学科学院肿瘤研究所,北京)

提 要

本文介绍从美国 NIH 肿瘤研究所引进的核酸和蛋白质序列数据库的开发应用情况。此数据库包括国际上几个主要的 DNA 数据库和蛋白质数据库，是86年底的版本。该数据库管理程序提供了方便的数据检索途径，具有一般数据处理能力。该系统还配有一套适于结构-功能分析的序列同源性比较和二级结构分析的程序。此系统现已开发成功，可提供使用。

表1 IDEAS 所集合的数据库

1 GENBANK	-GenBank (R) Genetiec Sequence data Bank Release 44.0,30 AUG 1986 Los Alamos National Laboratory 8823 entries, 8, 442, 357 bases
2 EMBL	-EMBL Nucleotide Sequence Data Library Release 8.0 April 1986 European Molecular Biology Laboratory 6395 entries, 6353040 bases
3 NBRFNUC	-Nucleic Acid Sequence Database of the Release 28.0, July 1986 Protein Identification Resource at the National Biomedical Research Foundation 1917 entries, 3502068 bases
4 NBRF	-Protein Sequence Database of the Release 10.0, AUG 1986 Protein Identification Resource at the National Biomedical Research Foundation 3800 entries, 890703 residues,
5 PDBDIR	-Protein Data Bank Directory April 1986 Brookhaven National Laboratory 292 entries
6 PDBSTR	-Protein Data Bank Structures April 1986 Reoreanized at Institute for Chemical Research Kyoto University 391 entries, 69308 residures
7 KABAT	-Sequences of Proteins of Immunological Interest 1983 Bolt Beranek and Newman, Inc.

自70年代以来，由于分子克隆和 DNA 序列分析技术的迅速进展，被搞清的基因 DNA 序列和蛋白质序列急剧增加，为了充分利用已获得的信息，各国相继建立了 DNA 序列和蛋白质序列的数据库及相应的电子计算机管理系统，大量开展了电子计算机在 DNA 序列数据和蛋白质序列数据处理中的应用，IDEAS 系统就是其中之一。

IDEAS 是核酸和蛋白质数据库及其管理分析系统的简称，是 Integrated Database and Extended Analysis for Nucleic Acids and Protein 的缩写，它由美国 NIH 的国家肿瘤研究所发展并在 VAX-11/780 计算机上运行，其数据库集合了目前国际上几个主要的核酸序列数据库及蛋白质的氨基酸序列数据库，是 86 年 8 月份发表的版本，还集合了蛋白质结构参数数据库 PDBSTR，免疫学所关心的蛋白质序列数据库 KABAT 等。这些数据库详列于表 1。

IDEAS 通过其数据库管理程序 SEQMAN 的各种命令对上述数据库进行文件系统的管理，可从各数据库中方便地获取有关信息，如核酸或蛋白质序列数据，发表的文章题目，作者、期刊、序列的某些特征数据，序列的来源等。

IDEAS 除了提供文件系统管理程序 SEQMAN 外，还配备一套适用于序列数据处理的应用程序，这些程序的名称及其功能列于表 2

表 2 序列数据处理程序的简表

序列名称	功　　能
SEQA	两 DNA 序列对准比较
SEQAP	两蛋白质序列对准比较
SEQOP	统计蛋白质序列间同源性的可信限
SEQF	检索某序列片段在 DNA 数据库中的同源性
SEQFN	快速检索某片段在 DNA 数据库中的同源性
SEQH	全面检索某片段在 DNA 数据中的同源性
SEQFP	检索蛋白片段在蛋白数据库中的同源性
SEQHP	全面检索蛋白质序列间的局部同源性
STRALI	用相似片段对准方法预计二级结构
CHOFAS	用 Chou-Fasmans' 方法预计二级结构
DELPHI	用 Garnier's 方法预计二级结构
DSSP	标注蛋白质二级结构
ANNOT	蛋白质数据库中蛋白二级结构的标注
HPLOT	计算和绘制疏水性值和带电氨基酸的分布曲线。
HCOMP	比较两蛋白质间疏水性值的分布图
WUKAB	显示蛋白质序列每个氨基酸的可变性
ALOM	膜蛋白的定位

由于计算机的配置和计算机操作系统版本的差异，储存于磁带上的 IDEAS 版本送入我们的 VAX-11/780 计算机时，必须根据具体条件适当修改。

IDEAS 软件使用方法可参阅 IDEAS 的使用手册或借助于 IDEAS 软件的 HELP 的帮助，我们也已写出关于 IDEAS 使用方法的讲义，其中包括一些用例。

下面仅举几个实例说明 IDEAS 运行情况：

(1) IDEAS 运行选择

运行 IDEAS 时，显示屏显示如表 3 操作选择

表 3 前一部分是 IDEAS 的自我提示，后一部分为 IDEAS 的任务选择，可打入相应名称，选择运行相应程序。

(2) IDEAS 数据库管理程序 SEQMAN 的运行

键入 SEQMAN 回答 IDEAS 提问后，SEQMAN 程序开始运行，文件系统管理程序 SEQMAN 的几个主要命令如表 4。

表 4 各命令的用法可详阅已编写成的

表3 IDEAS 的初始操作选择

```
I  D  E  A  S
Inlegrated Database and Extended Analysis System
for Nucleic Acids and Proteins ***
Last database updates ***
08/30/86 GENBANK Release 44.0(AUG 86)
08/05/86 EMBL    Release 8.0 (APr 86)
08/13/86 NBRFNUC Release 28.0 (JUl 86)
08/05/86 NBRF    Release 10.0 (May 86)
Options available:
SEQMAN FRAMIS SEQF SEQFN SEQH SEQFP SEQHP SELECT GRAPH
SEQDP SEQFT SEQA SEQAP SEQL STRALI CHOFAS DELPHI DSSP
ANNOT HPLOT HCOMP WUKAB ALIGN ALOM
DCL    Symbol Manual Help Menu More Exit
Option? SEQMAN
```

表4 文件系统管理程序 SEQMAN 主要命令简表

命令符号	简要功能
find	检索指定数据库中匹配字符串的款目
get	列出指定数据库中指定款目的原文
Scan	检查指定的数据库的内容
SRCH	报告序列文件中和指定字符串相匹配的序列位置
Pr	按一定格式列出序列数据
tr	把核苷酸序列翻译成氨基酸序列
rc	列出指定序列的互补序列
freq	统计序列中碱基或氨基酸频率
diff	标出两序列的差别
def	定义和检查数据库个数
type	打印指定款目的内容

IDEAS 的使用讲义。现举一个用例说明:

如检索 GENBANK 中与 T-CELL RECEPTOR 的有关序列,可打入命令:

FIND GENBANK T-CELL RECEPTOR

```
HUMTCAXA Human T-cell receptor active alpha-chain
         mRNA from Jurkat cell lien. 728bp
HUMTCAXA Human T-cell receptor germline beta-chain
         partial C-beta-2 gene. 278bp
   :        :        :
   :        :        :
```

如果要检索 EMBL 数据库中的 T-CELL RECEPTOR 的序列,打入命令:

FIND EMBL T-CELL RECEPTOR

命令执行后,输出和上例类似的结果。需要打印出从 GENBANK 中检索出某个序列文件的详细内容时,如打印出 HUMTCAXA, 可用命令:

于是在 GENBANK 数据库中,属于 T-CELL RECEPTOR 的序列名称及其缩写名将按下面方式列出:

GET GENBANK HUMTCAXA

将把识别符号为 HUMTCAXA 的内容全文输出。SEQMAN 的其它命令的用法不再例举。

(3) IDEAS 其它序列处理程序的运行

我们编写的 IDEAS 的使用讲义,详细地介绍了这些序列分析处理程序的运行方法,必要时还可参阅每个程序中所用算法的原始文

表5 CHOFAS 程序运行会话简表

```
CHOFAS---Protein Secondary Structure Prediction
Input file [NBRF database] = NBRF
Output file [Your terminal] =
Output line width [80] = 60
Plot the Profiles? (Y/N) [N] = N
Sequence or LIBRARY or END: RWHUVY
Start, End ([CR] for entire sequence)
```

表6 用CHOFAS 程序预测 RWHUVY 蛋白质的结果

```
      Prediction of Secondary Structures in RWHUVY (Length 135)
            10        20        30        40        50
 MDSWTFCCVSLCILVAKHTDAGVIQSPRHEVTEMGQEVTLRCKPISGHNS
  -                    +.  -         +. - - -      ++
      <--------------->      <------------>
 EEEEEEEEEEEEEE          EEE          EEEEEEEEEEEE
  T                         TT             T
       60        70        80        90        100
 LFWYRQTMMRGLELLIYFNNVPIDDSGMPEDRFSAKMPNASFSTLKIQPS
       +   +   -                  --       --+  +
<--------------->       <------><---------->
EEEEEEEEEEEEEEEEEEEEE                          EEEEEEE
       I              T       TT   TT             T       TTTTT
       110       120       130
EPRDSAVYFCASSFSTCSANYGYTFGSGTRLTVV
  +       -    +-                 +
    <---------->
        EEEEEEEEEEEEEEEEEEEEEEE
        I     T     I         T
```

献，这些文献资料也可从 IDEAS 系统中的 HELP 命令查出。

IDEAS 的程序使用都有类似的格式，如程序运行时先提问待输入文件名称；回答可为 IDEAS 已具备的数据库名称或者用户自己的某子数据库名；然后询问输出设备号；是否需要修改程序中的有关参数；指定序列名称及序列起止编号等等；最后输出计算结果。凡需要绘图部分，还提问绘图终端的型号。现用 CHOU FASMAN'S 方法预测蛋白质序列 RWHUVY 的二级结构为例：

在 IDEAS 的问话 OPTION? 时，回答 CHOFAS，开始运行二级结构预测程序，运行会话过程如表5。

程序预测蛋白质二级结构的结果如表6

从 IDEAS 的数据库及其所提供的应用程序看，IDEAS 除了提供检索 DNA 序列的有关信息外，它对于进行序列间同源性比较和蛋白质序列二级结构预测及二级结构图形显示功能较强，因此 IDEAS 系统适合进行结构-功能方面的研究工作，但它缺乏用于从事基因工程实验研究所需要的某些有关序列数据处理的程序，如酶切图谱的构建；PROMOTOR 位置的测算；内含子与外显子交接片段的定位；RNA 二级结构的预测；蛋白质抗原决定簇预测等等方面的程序。我们实验室已建立了上述列举的用于基因工程实验研究所需要序列数据分析程序，从而可弥补 IDEAS 的不足之处。

[本文于 1987 年 7 月 28 日收到]