蛋白质和 DNA 数据库管理系统

吴　加　金

（军事医学科学院基础医学研究所，北京）

提　　要

本文介绍我们发展的蛋白质和DNA数据库及其计算机管理系统。系统从6个不同途径对数据库检索，也提供多种输出信息的方式，使用方便，并具有丰富的序列数据处理软件，可按不同需要处理所检索的序列数据。

随着基因工程和蛋白质实验研究技术的迅速发展，每年发表基因的 DNA 顺序已超过几十万碱基对。DNA 序列数据的积累达到爆炸性地步。为了合理储存、使用、分析和管理这些数据，生物学家寻找电子计算机的协助。目前世界各国都相继建立了 DNA 或 RNA 和蛋白质数据库。我们也引入了蛋白质数据库[1]和 DNA 序列数据库。并设计相应的检索系统，现简介如下：

一、数据库简介

数据库的数据原为磁带文本，现已转入超

表 1　蛋白质氨基酸序列数据库的数据格式

CODE	DROOPSB4P1 92AA 27/12/85
NAME	OPSIN, EXON 4
SOURCE	D. MELANOGASTER, DNA (CANTON S STRAIN) AND cDNA TO mRNA (OREGON RP2 STRAIN).
ORGANISM	DROSOPHILA MELANOGASTER EUKARYOTA; METAZOA; ARTHROPODA; INSECTA; DIPTERA.
REF	TRANSLATED FROM DROOPSB4 IN GENBANK REV 38.0
COMMENT	BEGINNING OF CODING REGION IS MISSING. END OF CODING REGION IS MISSING.
SEQUENCE	AVSAHEKAMR EQAKKMNVKS LRSSEDAEKS AEG-KLAKVAL VTITLWFMAW TPYLVINCMG LFKFEGLTPL NTIWGACFAK SA/ CYNPIYV GI

表 2　DNA 序列数据库数据的一般格式

ID	LPRS219A unreviewed; DNA 104 BP
××	
AC	KO1647;
××	
DT	18-JUL-1985 (incorporated)
××	
DE	Sea urchin (s, purpuratus) repeat element 2109a.
DE	clone csp21oga
KW	repetitive sequence; repetitive sequence 210ga.
OS	Strongylocentrotus purpuratus
OC	eukaryota; Metazoa; Echinodermata; Echinoidea.
××	
RN	[1] (bases 1-104)
RA	posakony J. W., Flytzanis C. N.,
RA	Britten R. J., Davidson E. H.;
RT	"Interspersed sequence organization and
RT	developmental representation of cloned poly(a)
RT	rnas from sea urchin eggs";
RL	J. Mol. Biol, 167: 361—389(1983).
××	
FH	Key From TO Description
××	
SQ	Sequence 104Bp; 28A; 21C; 26T; 29G.
	AATTCGGGGG GCGTTTCACA AAACTTGTC/ TCAGTGACAC ATGACAGTTG
	TTGTTATAAG CTACTGAGGA AACGTCAAA(CTTTCGCGCT TCGCGCGAAA
	GGGG

级小型计算机 VAX/11-780 的硬盘。蛋白质数据库包含 3450 氨基酸序列，740693 个氨基酸。DNA 数据库包含六千多个序列。这些数据是收集自几十种有关杂志，截止 1985 年初所发表的序列数据。

数据库中每个数据项目的结构大致相同，由两部分组成：1. 原文部分，2. 序列数据部分。

表 1 和表 2 为这两数据库结构的举例。表 2 中每行首的两个字符为该行内部的识别符，是缩写形式。表 1 行首为该行识别符的全称。

二、数据库检索软件的设计

由于数据库容量大，为了便于从几个不同途径迅速检索，在分析数据库结构基础上，作了如下几项工作：

1. 把所有序列数据款目建成单个文件。

2. 建立蛋白质数据文件名称和 DNA 序列数据文件名称的汇总文件。

3. 建立了文件名称和该序列全称的对照文件。

4. 建立期刊名、卷期、年份与该卷期所发表的序列文件名的对照文件。

5. 建立了作者、关键词等与数据文件编号的对照文件。

6. 建立了编号和文件名称的对照文件。

利用上面建立的几个索引文件，用 Pascal 语言设计数据检索程序。此程序可从如下六种途径检索序列数据：

1. 用期刊杂志名称和卷期检索序列数据。

2. 按序列名称缩写码检索

3. 按有关单词检索，可以是关键词或其他有代表性单词。

4. 按序列片段的类似性检索

5. 以作者名检索

6. 以关键词检索。

考虑到可能出现单词拼写差错，按途径 2 和 3 检索时，字符符合 60% 以上就显示出缩写序列名和序列全称，供用户最后取舍。

用途径 4 检索时，可检索属于某关键词所限定的有关序列，也可检索整个数据库中全部序列。检索时凡类似性大于某个预定值的序列都可输出。

检索命中后，可按下面 8 种方式之一输出信息。

1. 打印出序列数据文件名称

2. 列出序列名称

3. 列出微生物种类

4. 列出发表序列的文章的作者

5. 列出文章的题目

6. 列出发表序列的期刊

7. 仅列出序列数据

8. 打印出整个序列数据文件。

此外，检索命中的序列数据还可利用联机传递方式，由 VAX/11-780 终端传送到 BCM-S68K 微机，利用建立在 BCM-S68K 微机上的序列数据处理程序包对序列数据进行处理。

三、序列数据处理的程序包

序列数据处理是利用电子计算机程序，通过对序列数据的各种运算达到对某些实验操作的模拟，如酶切、剪切和重组；研究序列某种特殊功能，如序列重复成分、发夹结构、树叶状结构、氨基酸序列的螺旋结构和 β 折叠结构等；比较不同序列相似性或差异和同源性。我们建立的应用程序包如表 3。表 3 程序都用 Pascal 高级语言编写，很容易转移到别的机种。

四、使用情况

本数据库检索途径较多，适用于不同情况的检索，使用方便。现举两个用例：

例 1　实验测出氨基酸片段

HFAPLSNGSVVDKVTPDMAH

但事前不知此片段和哪种蛋白质序列相类似，为此使用对整个蛋白质数据库进行类似性检索，并预置两序列片段类似性为 85%，检索结果按输出序列名称方式输出，结果如表 4。表 4 说明待检索序列和本数据库中两氨基酸序列存在类似片段，其类似性分别为 100%，类似片段在序列的起始位置分别于第 14 和第 11 个氨基酸。由于对整个数据库检索，加上 VAX/11-780

表 3 序列数据处理程序简表

序号	程序名	功 能
1	EDDNA	1 建立和输入新的 DNA 序列 2 插入、删除、修改 DNA 序列文件 3 剪切基因片段 4 DNA 和 RNA 互相转换 5 DNA 互补链的换算及序列数据的输出
2	UPDENZ	建立限制性内切酶文件及对该文件的编辑
3	RPDNA	检索序列中重复核苷酸片段
4	RPDNAT	检索两序列中相同核苷酸片段
5	RPAM	检索序列中重复氨基酸片段
6	RPAMT	检索两序列中相同氨基酸片段
7	RECSL	检索一序列片段在另序列相同片段的位置
8	SIMILAR	寻找 DNA 序列（称 B 序列）在另一序列［称 A 序列］中类似片段位置
9	DIVATE	统计两序列分散性
10	COMPR2	两序列核苷酸位置最大对准比较
11	COMPR3	三序列核苷酸位置最大对准比较
12	RECS	求限制性内切酶酶切位点
13	MAPL	测定线性 DNA 序列的酶切图谱
14	PAD	估算蛋白抗原决定簇的位置
15	LOCATION	按三相列出 DNA 的氨基酸序列，按氨基酸的名称或密码子印出其位置图谱统计氨基酸的分布，统计起动子和终止子氨基酸分布
16	PROM	检索序列中内含子与外显子交接片段位置或 promoter 片段位置
17	JOIN	联结两序列片段，删除重叠部分
18	PPEDIC	估算蛋白质序列的二级结构（ROBSON 方法）
19	CHOUPD	Choug-Fasman 方法估算蛋白质二级结构
20	MACTHR	用最大碱基配对算法估算 RNA 二级结构
21	SSRP	按配对矩阵和自由能最低算法估算 RNA 二级结构
22	SSRC	用直接求自由能最低的配对方案估算 RNA 二级结构

为多用户机器，检索所用机时与各用户使用 CPU 状态有关，按例 1 情况，命中表 4 两序列

表 4 氨基片段类似性检索的结果

序列名称	类似片段起点	类似性[%]
opsin, exon 1	14	100.0
opsin, exon 2	11	100.0

约用 4 分钟，但整个检索过程使用 CPU 时间为 11 分。

例 2 检索 DNA 序列 INTERLEUKIN-2

以按单词途径检索，打入 IL-2，检索第一次命中后显示如下序列全称，询问是否为待检索序列。检索使用终端约 1 分钟，占 CPU 约 50 秒。

HSILO2 S RNA 801bp Human mRNA encoding interleukin-2

DO you want this sequence?

若此序列属于待检索序列，键入 "y"，此序列被记录。接着又显示另一个命中序列，两次命中间隔仅几秒。

HSILO5 U DNA 6684bp Human interleukin-(IL-2) gene and 5′-flanking

此序列还需检出时，再次打入 "y"，接着又显示。过程重复直至检索完毕。

MMILO4 S RNA 939bp Mouse mRNA for interleukin-2 (IL-2)

最后根据需要，选 8 种输出方式中的一种、输出所有已选中序列的相应内容。

其他应用软件也经常使用，大大节省了实验工作者的劳动，不再举例。

在建立数据库和程序库过程中，得到黄翠芬教授、马贤凯教授、朱伟雄高级工程师和姚志建副教授的大力协助和支持，也得到医科院基础所祝庆林教授的大力支持，在此表示感谢。

参 考 文 献

[1] Claverie, J.M. et al.: *Biochimie*, **61**, 437, 1985.

［本文于 1986 年 11 月 25 日收到］