

欧美 DNA 序列资料库近况

随着分子生物学的飞速发展,电子计算机已经成为其不可缺少的工具。建立电脑贮存的 DNA 序列资料库,是其中一项重要应用。目前建立的此种资料库主要的有两个,一在美国,一在欧洲。

生物学中许多重要现象,如基因的不连续性,重叠基因的发现等等,都是通过 DNA 的序列测定而得到揭示。快速 DNA 序列测定法和克隆技术的出现,使得阐明 DNA 结构的工作成为一些实验室的常规操作,同时可以测定越来越大的基因组。近年来 DNA 序列资料似洪水般大量涌来,迫使生物学家寻求一种新的手段,去贮存和处理这些信息,同时因为测定的序列

越来越长,有些杂志已不再刊载它们,于是建立电脑贮存的 DNA 序列资料库,被提上了议事日程。

表 1 EMBL Library 资料发布概况

发表次数	1	2	3	4
日期	82.4	83.4	83.12	84.8
引用文献数	387	610	943	1275
收集序列数	568	811	1481	1698
总长度(碱基)	583,433	1,114,447	1,654,863	2,147,205

表 2 EMBL Library 收录的 DNA 序列格式

```

ID      MMIG20    MUS. MUSCUL. IG. MOPC41; DNA; 350 BP.
XX
DT      82.01.01 (first entry)
XX
DE      First two exons in immunoglobulin light chain genes from
DE      cell line MOPC41.
XX
KW      differentiated gene; immunoglobulin.
XX
OS      Mus musculus (house mouse, souris domestique, Hausmaus)
OC      Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda;
OC      Mammalia; Eutheria; Rodentia.
XX
RN      [1] (bases 1-350)
RA      Altenburger W., Steinmetz M., Zachau H. G.;
RT      "Functional and non-functional joining in immunoglobulin light
RT      chain genes of a mouse myeloma";
RL      Nature 287:603-607(1980).
XX
FT      Key      From  To      Description
FT
FT      CDS      126   176      first exon (leader peptide)
FT      CDS      303 > 350  second exon (variable part)
XX
SQ      Sequence 350 BP; 80A; 82 C; 122 T; 66 G.
  
```

```

CGTGACCAAT  CCTAACTGCT  TCTTAATAAT  TTGCATACCC  TCACTGCATC  GCCTTGGGGA
CTTCTTTATA  TAACAGTCAA  ACATATCCTG  TGCCATTGTC  ATTGCAGTCA  GGACTCAGCA
TGGACATGAG  GGCTCCTGCA  CAGATTTTTG  GCTTCTTGTT  GCTCTTGTTT  CAAGGTAA
ATGAACTTA  AAATGGGAA  TTTTCCAACG  TTTCCAACG  TGGTTAGTGT  TGACTGGCAT
TTGGGGGATG  TCCTCTTTA  TCATGCTTAT  CTATGTGGAT  ATTCATTATG  TCTCCACTCC
TAGGTACCAG  ATGTGACATC  CAGATGACCC  AGTCTCCATC  CTCCTTATCT
  
```

在含有几千、甚至上万个碱基的 DNA 序列中,各种重要的区域,诸如限制性内切酶的切点,不连续基因的断裂点,转录起始位点等等,很难用肉眼识别;同时,也只有使用计算机软件系统,才能进行各种统计和分析。例如从被测序列中找出与已知的基因的同类区域。换言之,使用电子计算机,才可以完全理解和充分利用测得的 DNA 序列所提供的信息。

早在 1979 年,分子生物学界开始迫切感到有此需要。1980 年美国国立卫生研究所 (NIH) 和欧洲分子生物学实验室 (EMBL) 商定合作建立一个 DNA 序列资料库。计划的第一阶段,是收集并贮藏所有的 DNA 序列资料,第二阶段,是发展软件系统,以便进行各种分析。双方还商定,各自系统能互相兼容,这样各方只需收集一半的资料。美国方面由于种种原因,未能按计划进行,于是 EMBL 独自行动,在 1982 年 5 月建立了自己的核苷酸序列数据库 (Data Library),并开始发布第一期资料。五个月后,美国 NIH 也与 Bolt, Beranek & Newman 公司签订了五年合同,投资三百万美元,建立美国的资料库 Gen Bank。目前世界上主要的 DNA 资料库,就是这两个。在它们建立前后,还有些实验室建立一些较小规模的资料库。据说,日本也打算建立一个。

EMBL 资料库已发布四期资料。资料以印刷品和

磁带两种形式散发,概况见表 1:

EMBL 资料库收的序列,都给出其名称、长度、贮存日期、对序列的描述、生物体的详细分类地位、文献出处、作者、序列中各基因或功能区域位点,以及序列本身。同时还列出关键词,以便电脑检索。举例见表 2:

为执行计划的第二阶段,1984 年 NIH 与 Intelli-Genetics 公司签订,投资 560 万美元,建立一个全美范围电脑网络,它不仅能直接取用 GenBank 中 DNA 和蛋白质序列,还提供一整套软件,可以对资料作比较分析。EMBL Library 目前则建立了一个软件交换“窗口”,定期发布消息,便于软件编写者之间互相沟通,共享软件,避免重复。

DNA 序列资料库的最终远景,是建立一个计算机网络,将分散在各地的研究人员连结起来,改变目前各自为战和重复劳动的局面,并可以很方便地利用各序列资料库的资料,进行分析。

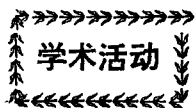
[Lewin, R.: *Sciencs*, **217**, 817—818, 1982.

——: *ibid.*, **223**, 1379—1380, 1984.

EMBL Nucleotide Sequence Data Library

Release Notes, Aug. 1984.

上海交通大学生物科学与技术研究所 陈民伟 范宝玲 编译]



第三届全国酶的结构和功能学术讨论会征文通知

为促进我国酶学研究的发展,按学会工作计划,中国生物化学学会酶学专业组将于今年 10 月下旬在成都召开全国第三届酶的结构和功能学术讨论会,现开始征集论文:

一、征文内容

近几年来,主要在国内完成的较有系统的关于酶的结构和功能方面的研究成果,包括酶工程和临床酶学中的基础研究。

二、征文要求

1. 论文应为有创新成就的科研成果,摘要文字精

练,不超过一千字。

2. 应征论文摘要需经本单位或地方学会推荐,北方地区请寄北京中国科学院生物物理研究所酶室练永宁收,南方地区请寄上海中国科学院生物化学研究所酶室许根俊收,征文截止日期 1985 年 6 月 15 日。

3. 征文上写明主要作者详细通讯地址,以便联系。

4. 经生化学会组织有关专家审查后,入选论文将在会议文集上刊登。

5. 会议文集将刊登少量与专业有关的广告。有关单位如欲刊登请与上海生化所许根俊同志联系。