信息与交流

编者的话 近 20 年来，在科学技术的发展过程中，呈现出各学科及其分支学科之间互相渗透及融合的新潮流和新趋势，新的学术思想，工业产品，乃至新的科系随之应运而生，如数码分类、环境微生物、极端环境微生物、混合发酵、基因工程等。与此相适应，本刊拟开辟此新栏目，以餐本刊的爱好者，并欢迎赐稿，提出批评和建议。让我们共同努力，把本刊办成为学术思想活跃，格式多样，内容丰富，为微生物学工作者喜闻乐见的园地，迎接建国 50 周年和蕴育着新的机遇和激烈挑战的新世纪的到来。

国际互联网上的 NCBI 分子生物学数据库简介

马东晖 李小洁 马辉文*

(武汉大学生命科学学院 武汉 430072)

崔 晓 晖

(武汉测绘科学技术大学网络中心 武汉 430070)

关键词 国际互联网，NCBI 分子生物学数据库，分子生物学

分类号 Q78 **文献识别码** D **文章编号** 0253-0654(1999)-02-0150-53

NCBI 是美国国家生物技术信息中心 (National Center of Biotechnology Information) 的英文缩写，它于 1988 年由美国国会提议创建。其宗旨是开发和研究各种服务于生物医学领域的自动信息存取系统。NCBI 管理着著名的核苷酸和蛋白质序列数据库 GenBank[1]，它是国际电脑互连网 (INTERNET) 上最大的有关分子生物学的数据库系统之一。

NCBI 数据库系统除提供 GenBank 所提供的各种服务外，还与设在美国国立卫生研究院 (NIH) 内的国家医学图书馆联合，提供近 3800 多种有关生物医学方面的文献检索服务。此外，NCBI 数据库还包括有众多生物医学方面的专门数据库和工具软件，如：人类孟德尔遗传学数据库 OMIM，布鲁克海文大分子模型数据库 MMDB，开放阅读框查找软件 ORF 和同源查询工具软件 BLAST 等。另外，NCBI 还与世界上各大著名分子生物学数据库建立有直接联系。因此，从 NCBI 的网络站点 (http://www.ncbi.nlm.nih.gov) 出发去可访问其它各种分子生物学数据库资源。以下简要介绍 NCBI 提供的各项服务。

1 GenBank 数据库

GenBank 是当今全球最大的核苷酸、蛋白质序列数据库之一。GenBank 主要致力于收集全世界各种新发表的核苷酸及蛋白质序列数据，并将收集的数据资料进行分类整理和储存。同时，GenBank 还提供各种实用工具软件以支持 GenBank 的各种服务功能。GenBank 提供的所有数据及工具软件服务均是免费的，用户只需有一台能与 INTERNET 连接的电脑，就可从中取得任何有用的资料。截止 1997 年 7 月 13 日，GenBank 共收录 1525698 个核苷酸序列，碱基 1033207934 个。

1.1 GenBank 数据资源的来源 GenBank 序列资源主要有两个来源：(1) 由序列发现者直接提交。当前几乎所有的国际性生物学刊物均要求作者在文章发表之前将其所测得的序列向 GenBank、EMBL (欧洲分子生物

* 联系人 (责任作者)

1998-01-05 收稿，1998-06-08 修回

学实验室)或 DDBJ(日本 DNS 数据库)提交,以得到上述数据库签发的登录注册号(accession number)。(2)从近 3800 多种有关于生物医学方面的杂志上搜索已发表的序列资料。GenBank 还与设立在欧洲分子生物中心的 EMBL 及日本的 DDBJ 序列数据库合作,每天进行资料的交换互补,以保证数据库的新颖性和代表性。

为方便研究工作者提交自己的序列,GenBank 向作者提供两种主要的序列提交方式。一种是通过 WWW 超文本连接方式。用户可以通过双击 NCBI 主页(Home Page)中的 BankIt 按钮,连接到 BankIt 工具软件上(http://www.ncbi.nlm.nih.gov/Bankit)在线地提交序列资料。由于这种方式是一种启发式的用户界面,操作起来方便、快捷。当前按这种方式提交的序列占 GenBank 序列总数的 80% 以上。另一种方式是作者通过邮寄磁盘或通过电子邮件(E-mail)递交。但在递交前必须用 NCBI 提供的 Sequin 工具软件对所要提交的序列进行处理。Sequin 软件工具可通过 NCBI 的匿名 ftp 文件传输方式获得。其路径为 ftp://ncbi.nlm.nih.gov/pub/sequin。如果作者发现已发表的序列有错时,也可通过上述两种方法向 NCBI 报告。

1.2 GenBank 数据库的管理　由于 DNA 核苷酸序列测定技术的改进,GenBank 搜集的序列数目几乎成几何级数增长。其中表达序列标签数据库(Expressed Sequence Tags database—dbEST)[2]是 GenBank 数据库递增最快的一支。它已占该库序列总数的 70% 以上。而且每天都有近两千个新 EST 序列送入 GenBank.。为方便用户获取 GenBank 序列数据库的资料,GenBank 对所收集的序列进行了分区,如 dbEST 和基因序列标识位点数据库(Sequence tagged site—dbSTS)[3]等数据库就是新划出的分支。此外,GenBank 还与其它序列数据库达成了协议,对所有序列注册登录号进行统一。用户只需通过一个登录号就能顺利进入其它与 GenBank 直接连机的序列数据库,如,EMBL、DDBJ、GSDB、PROSITE、PDB、SWISS-PROT等。

人类基因组计划是当前最大的分子生物学课题,其中转录图谱的构建已成为该课题实施的关键[4],为此 GenBank 专门建立了 dbEST 和 dbSTS 序列数据库。

1.2.1 dbEST 数据库:表达序列标签(EST)是一类仅经过 5′端和 3′端序列测定的 cDNA 片段。根据 Wilcox 等人研究,cDNA 的 3′端非翻译区(3′ UTR)具有分类学特征。他们发现 3′ UTR 相对于 cDNA 的其它区域极为多变,而且来源于同一基因 cDNA 的 3′ UTR 相对保守[5],且均不含内含子[6]。因此,EST 在人类基因组转录图谱的构建、人类新基因的发现和基因序列编码区的鉴定方面具有重要意义[7]。当前这一领域的发展十分迅速。它在很大程度上归功于美国华盛顿大学基因组测序中心和 Merck 公司合作进行的人类 cDNA 克隆的全面 EST 测定计划[8~9]。到 1997 年 7 月,GenBank 已收到人类 EST 序列 768376 个,占 dbEST 总数的 69%。

1.2.2 dbSTS 数据库:dbSTS 数据库是序列标识位点数据库的简称。构建高分辨率的人类基因组物理图谱的主要目的是为人类基因组计划搭建起一个完整的框架和更好地组织已获得的序列资料[10]。为此,NCBI 设计出一种被称之为"电子 PCR"的工具软件[1]。它能允许用户将一段新获得的人类 DNA 序列与 dbSTS 数据库进行对比。一旦发现此新获取序列含有 STS 引物结合位点,此软件就能像"PCR"一样,扩增出一段 DNA 片段。如果此片段大小与预测的片段大小一致,则可将这段新获得的人类 DNA 序列定位于已构建的人类基因组图谱上。

1.3 GenBank 数据库提供的服务功能　GenBank 数据库提供的服务功能可通过多种方式获取,其中最为常用的有以 WWW 超文本连接方式的 Entrez 系统以及 NCBI 提供的各种 E-mail 服务器。

1.3.1 Entrez 系统:Entrez 系统服务器的地址是 http://www.ncbi.nlm.nih.gov/entrez。该系统是一个完整的分子生物学信息获取系统。它将 DNA 和蛋白质序列资料、蛋白质三维结构数据以及 MEDLINE 目录系统有机地统一起来[1]。例如,Entrez 将 MEDLINE 文献目录摘要系统与 GenBank 连接,使用户能方便地由一个序列资料查找 MEDLINE 上与它相关的论文摘要。或者,由 MEDLINE 上一篇论文的摘要,用户可以查找到与之相关的序列资料。同时 Entrez 系统还与世界上其它各大核苷酸及蛋白质序列数据库直接联系,使用户能通过 NCBI 这一结点很方便地获得各大数据库的信息资源。

如果运用 RasMol 或 Kinemege 工具软件,用户还可通过 Entrez 查询系统,直接观察蛋白质分子的三维结构。RasMol 等工具软件可以通过 NCBI 的匿名 ftp 服务器(ftp://ncbi.nlm.nih.gov/entrez/network)直接获得。由于 INTERNET 的广泛应用,现在 NCBI 已不再提供

Entrez 的光盘。

1.3.2 GenBank 的电子邮件服务器：NCBI 向用户主要提供三种形式的 E-mail 信息查询服务器，它们是 Retrieve、Query 和 BLAST E-mail服务器。通过这三种服务器，用户可以获取各种 GenBank 所能提供的信息资源。由于 E-mail 服务在我国已经相当普及，下面拟较为详细地介绍 NCBI 提供的 E-mail 服务器使用方法。

(1) Query E-mail服务器：此服务器的地址是 query@ncbi.nlm.nih.gov。它与 Entrez 查询系统类似，也采用相邻区域法则。用户不但可以获得目的序列信息，而且还可获得位于同一区域或其它相关区域的资料。目前支持这一查询方式的主要数据库资源有：GenBank 核苷酸、蛋白质序列资源，MEDLINE 分子生物学部分及 NCBI 的分子结构数据库、Query E-mail查询服务最短的命令可以仅有两行缩写代码组成。以下为一查询实例：

To: query@ncbi.nlm.nih.gov

From: hwma@whu.edu.cn

Subject:

Cc:

Bcc:

Attachments:

DBn

UID U30150

上面的黑体字是电子邮件软件 EUDORA 提供的信头。第一行为 Query 服务器 E-mail 地址，第二行为用户 E-mail地址(这里是本文作者的地址)，以下几项均为空白。横线以下为邮件主体。其中，第一行规定了查询区域为核苷酸序列，第二行规定了查询方式及内容，其中 u30150 是一 DNA 序列的登录注册号。另外，Query E-mail服务器还容许用户在第二行之后输入一些可变参量，以规定返回数据的形式。同时 Query E-mail服务器也支持布尔逻辑语法规则，可用"或"、"与"、"非"形式查询。要获得有关详细资料，读者只需在邮件主体中(横线以下)输入 help 即可。

(2) Retrieve E-mail 服务器：此服务器的 E-mail 地址是 retrieve@ncbi. nlm. nih. gov。它是最常用的 GenBank 核苷酸、蛋白质序列资源查询服务器。其查询方式灵活多样，可以通过序列登录号、关键词、作者名或杂志名等查询。以下为一查询实例：

To: retrieve@ncbi.nlm.nih.gov

From: hwma@whu.edu.cn

Subject:

Cc:

Bcc:

Attachments:

DATALIB genbank

BEGIN

acetylcholinesterase

横线以下为邮件主体。其中第一行指定了查询数据库的名称(这里是 GenBank 数据库)，第二行为查询服务器的起始识别符，第三行为查询内容。用户也可通过发出 help 指令获取使用此服务器的有关详细说明。

(3) BLAST 同源查找 E mail 服务器：此服务器的 E-mail 地址是 blast@ncbi.nlm.nih.gov。它容许用户能通过 E-mail对某一核酸或蛋白质序列进行同源找寻分析。BLAST 同源查找 E-mail 服务器采用由 NCBI 开发的 BLAST 基本局域配对找寻工具，将用户所输入的序列与 GenBank 中的所有序列进行同源对比查找[11]。BLAST 家族总共有五种形式：BLASTP、BLASTX、TBLASTN、BLASTN 和 TBLASTX，其中 BLASTP 和 BLASTN 分别容许用户输入一段氨基酸或核苷酸序列以对 NCBI 的蛋白质及核苷酸序列数据库进行同源查找分析。BLASTX 则先将用户输入的核苷酸按六个阅读编码框翻译成蛋白质后再分别与 NCBI 蛋白质序列数据库进行同源查找分析。TBLASTN 则将用户输入的一段氨基酸序列与 NCBI 核苷酸序列数据库中的所有序列分别按六个阅读编码框翻译成氨基酸序列后，再进行同源查找分析。而 TBLASTX 则是将用户输入的核苷酸序列和 NCBI 核苷酸序列数据库中的序列同时按六个阅读编码框翻译成氨基酸后再进行对比查找。

BLAST 同源查找 E-mail服务器要求用户严格按其规定的格式输入查找序列。BLAST 同源查找 E-mail服务器一旦发现所接受的查询格式有错误，都会终止查询而返回一帮助信息和用户操作手册。查询实例：

To: blast@ncbi.nlm.nih.gov

From: hwma@whu.edu.cn

Subject:

Cc:

Bcc:

Attachments:

PROGRAM blastn

DATALIB month

BEGIN

> XYZ012 mygene XYZ

taagtctagtcgtgtattttaggctcagtcgtacgtttgttc

tttaaaagccgatcgtactgatttgaaaagggcta

横线以下邮件主体，第一行给出了采用何种 BLAST 查询工具软件(blastn)，第二行则指定了查询数据库为非冗余核苷酸序列数据库，第五行为 FASTA 模式的查询序列。NBCI 的 BLAST 同源查找 E-mail服务器还容许用户根据操作手册进行多项参数的设定。

2 美国医学图书馆 MEDLINE 文献查询指南数据库

美国医学图书馆 MEDLINE 文献查询指南数据库是 INTERNET 上关于生物医学方面最大的虚拟图书馆检索系统之一。它总共收集了近 880 万篇有关生物医学方面的文献。而且每天还在不断地从 3800 多种生物医学杂志上收集新发表的论文。美国医学图书馆于 1997 年 6 月 26 日宣布，它的 MEDLINE 数据库对所有来访者均提供免费的 WWW 查询服务。目前，它主要有两大类服务项目：其一为 PubMed，地址是 http://www.ncbi.nlm.nih.gov/PubMed/；另一为 Internet Grateful Med，地址是 http://igm.nlm.nih.gov。其中 PubMed 查询完全免费，而 Internet Grateful Med 虽也完全免费，但需要有美国医学图书馆的用户帐户。PubMed 主要提供的服务项目有：(1)免费 MEDLINE 生物医学文献检索服务；(2)布尔逻辑方式查询、关键词查询、美国医学图书馆主题词查询(MeSH)；(3)与 24 种杂志直接联机服务，可提供全文；(4)关于临床诊断、治疗及预后的咨询服务；(5)与 NCBI 的 GenBank 序列数据库及三维分子结构数据库的联机服务。它通过 Entrez 查询服务系统与之进行互访连接。此数据库除可用于生物医学基础理论研究目的外，还可用于临床诊断。

3 NCBI 的一些专门性的数据库及工具软件

3.1 OMIM 人类孟德尔遗传数据库[12]　OMIM 数据库是由约翰·霍普金斯大学的 Mckusick 及其同事共同开发的一个遗传学数据库。它收集了各种已发表的人类基因以及由这些基因突变或缺失而导致的各种遗传病。此数据库还同时与其它相关的人类医学遗传学数据库相连接，如，Cardiff 人类基因突变数据库，PAX6 突变数据库等。OMIM 数据库不仅可以帮助医务人员诊断有关的人类遗传疾病，而且医务人员也可以通过这一节点与美国西奈医学中心遗传疾病图象数据库相连接，以获得一些相关的人类基因遗传病的图象资料。

3.2 MMDB 分子模型数据库　它是 Entrez 系统中的一种新型结构数据库。它收集了布鲁克海文蛋白质数据库中所有的大分子三维结构资料[1]。通过 MMDB，分子生物学家可以获得许多有关某一大分子的生物功能及作用机制的信息资料，同时 NCBI 所提供的 Cn3D 软件还可以帮助分子生物学家方便地进行大分子三维结构观察。此软件可通过 NCBI 的匿名 ftp 服务器获得(ftp://ncbi.nlm.nih.gov/entrez/network/)。

3.3 NCBI 分析工具软件　NCBI 不但提供以上众多的分子生物学数据库，而且还提供分子生物学家常用的多种分析软件，例如，ORF 阅读编码框查找软件。用户应用此软件不仅可以快速查找一段核苷酸序列中所有可能的阅读编码框，也可以根据需要决定采用何种遗传密码系统。它特别适合于序列提交前的分析工作。

总的来说，NCBI 提供的分子生物学工具软件不多。欧洲分子生物学数据库(EMBL)提供有大量的蛋白质序列分析工具软件。有兴趣的读者不妨可以登录到它的站点上一观。

参 考 文 献

[1] Benson D A, Boguski M S, Lipman D J *et al.* Nucleic Acids Res, 1997, **25**:1～6.

[2] Aaronson J S, Eckman B, Blevins R A *et al.* Genome Res, 1996, **6**:829～845.

[3] Olson M, Hood L, Cantor D *et al.* Science, 1989, **245**:1434～1435.

[4] Boguski M S, Schuler G D. Nature Genetics, 1995, **10**:369～371.

[5] Wilcox A S, Khan A S, Hopkins J A *et al.* Nucleic Acid Res, 1991, **19**:1837～1843.

[6] Hawkins J D. Nucleic Acid Res, 1988, **16**:9893～9908.

[7] Adams M D, Kelley J M, Gocayne J D *et al.* Science, 1991, **252**:1651～1656.

[8] Boguski M S. Trends Biochem Sci, 1995, **20**:295～296.

[9] Hillier L, Lennon G, Becker M *et al.* Genome Res, 1996, **6**:807～828.

[10] Collins F, Galas D. Science, 1993, **262**:43～46.

[11] Altschul S P, Gish W, Miller W *et al*. J Mol Biol, 1990, **215**:403～410.

[12] Sander D M. J Biotechniques, 1997, **22**:92～94.