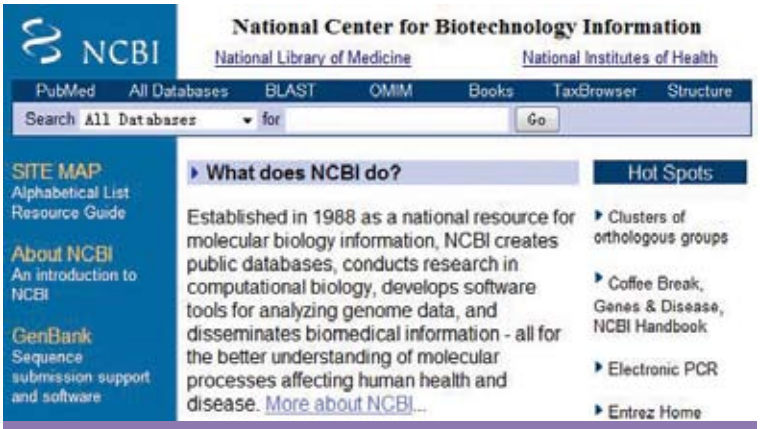
《核酸研究》2009数据库专刊收录数据库1170个，其中详细介绍了179个通用及专用数据库，包括其详尽描述和访问网址，其中95个数据库为首次报道。

《生命奥秘》本月专题将以《核酸研究》所介绍的数据库为基础，按照通用、热点、特色以及应用原则从中选取了部分数据库，如NCBI、UniProt、Ensembl、UCSC GDB等，并详细介绍其用途和所采用的方法，以期能抛砖引玉。

读者可在《核酸研究》网站（http://nar.oxfordjournals.org/content/vol37/suppl_1/index.dtl）获得完整的数据库列表及其摘要。

一 综合数据库

NCBI数据库集



图片来源：NCBI

美国国立生物技术信息中心（National Center for Biotechnology Information），即我们所熟知的NCBI是由美国国立卫生研究院（NIH）于1988年创办。创办NCBI的初衷是为了给分子生物学家提供一个信息储存和处理的系统。除了建有GenBank核酸序列数据库（该数据库的数据资源来自全球几大DNA数据库，其中包括日本DNA数据库DDBJ、欧洲分子生物学实验室数据库EMBL以及其它几个知名科研机构）之外，NCBI还可以提供众多功能强大的数据检索与分析工具。目前，NCBI提供的资源有Entrez、Entrez Programming Utilities、My NCBI、PubMed、PubMed Central、Entrez Gene、NCBI Taxonomy Browser、BLAST、BLAST Link (BLink)、Electronic PCR等共计36种功能，而且都可以在NCBI的主页www.ncbi.nlm.nih.gov上找到相应链接，其中多半是由BLAST功能发展而来的。

1 NCBI最新进展

1.1 PubMed搜索功能的增强

去年，NCBI对PubMed进行了几项改进工作，改动最大的是搜索界面和摘要浏览界面。其中，搜索界面中新增了“Advanced Search”选项（这实际上是对以往“Limits”和“Preview/Index”功能的整合），并

且增加了一个新的窗口，用户可以在此窗口下通过“论文作者名”、“论文所属杂志名称”、“论文出版日期”等限定条件进行搜索。而且，“论文作者名”和“论文所属杂志名称”还设有文本框自动填充功能。现在，在PubMed数据库中进行文本搜索的同时还可以立即通过两个“内容传感器（content sensors）”进行分析。一个“内容传感器”是根据作者姓名、所属杂志名称或杂志名缩写、出版日期、卷号或刊号等信息进行分析，然后将符合条件的搜索结果排列到结果列表的顶端。另一个“内容传感器”是根据文章是否与用户给出的条件，例如是否与某种药物相关，在NCBI的新增数据库PubMed Clinical Q&A中进行搜索，然后给出搜索结果。

1.2 新增Primer-BLAST分析工具

2008年，NCBI新增了设计、分析PCR引物的工具——Primer-BLAST。Primer-BLAST的引物设计功能是基于NCBI现有的Primer3程序发展而来的，Primer3程序可以为一段DNA模板序列设计PCR引物。Primer-BLAST在设计出引物之后还在某些相应数据库中进行BLAST搜索，因此可以得到特异性引物，扩增出目的片段。用户在给出DNA模板的同时还可以限定正向引物或反向引物，这样，NCBI就只会给出另一条引物。如果用户给出了模板DNA和两条引物序列，Primer-BLAST就只会运行BLAST程序，帮助用户对引物进行分析。用户也可以只给出两条引物而不给出模板序列，这时Primer-BLAST会通过BLAST程序分析出与这对引物最匹配的模板序列。Primer-BLAST进行BLAST搜索的数据库包括RefSeq mRNA、BLAST nr和12种模式生物基因组数据库。

1.3 BLAST的改进及更新

NCBI对BLAST进行了全新的改版，推出了最新的web BLAST report。在最新的BLAST比对结果页面中，“图形化概要（Graphic Summary）”、“具体描述（Descriptions）”以及“序列比对（Alignments）”等部分页面都可以展开和收起。此外，网页上还提供了“结果输出格式选项（Formatting）”和“结果下载选项（download）”，在下载选项中还新增了CSV格式下载。这样，读者可以轻松地将BLAST的比对结果输入到表格处理软件中去。另外，BLAST比对结果页面上的“Alignments”部分还提供了每一条命中序列在Entrez Gene中的相关信息，这些信息包括基因名称、来源物种以及在PubMed数据库中与该基因有关条目的数目等。

“BLAST tree”结果输出模式可以测量不同序列间的距离，自动收起亚类信息等。现在，可以以Newick 格式或Nexus格式下载 BLAST tree结果，也可以在进化树图中选择任一节点重新构树。最后还要向读者介绍NCBI BLAST的一个新网址：URL: blast.ncbi.nlm.nih.gov。NCBI建议读者都使用这个网址登陆NCBI BLAST，因为该BLAST使用更多的计算机进行分析，也具有更强的系统容错能力。

1.4 Entrez Gene改进及更新

基因组注释工作当中有一项重要的工作就是定位基因重叠群序列（contig sequences），即在染色体中找出某个基因的定位。实际上基因组测序工作就是将许多基因重叠序列彼此拼接，最后拼出“完整（中间会有一些缝隙）”的基因组图谱。这项工作可以直接将某个基因与某段基因重叠群序列对应起来，但不能直接将该基因与染色体联系起来，而这恰恰是生物学家最感兴趣的地方。因此，为了能让用户在搜索基因的同时，也能了解到一些该基因在染色体中的定位情况，Entrez Gene推出了新的“Limits”服务，用户可以使用该服务在基因组范围内进行基因搜索。用户可以在某个物种染色体的某个区域里进行基因搜索。

Entrez Gene会按以下三种顺序对搜索出的基因进行排序：

1. 按照基因名排序。
2. 按照相关性排序，即按照结果与用户搜索所使用的关键词，例如基因名称等的匹配程度排序。
3. 按照基因重要性排序，即按照该基因在PubMed、Homologene、Protein Clusters、Online Mendelian Inheritance in Man（OMIM）或Bookshelf 中文献数量的多少进行排序。

2 ENTREZ 搜索系统

2.1 Entrez

Entrez数据库是一个整合了多个数据库的综合检索系统，它包含了35个不同数据库的信息，共收录有超过350,000,000条记录（表1）。Entrez数据库支持使用简单的布尔查询（Boolean queries）方式进行文本搜索，可以下载不同格式的数据资料，还可以按照生物学关系提供与其它相关记录的链接。这些链接给出的都是最简要的信息，例如会给出一条序列和报道该序列的论文摘要，或者会给出一条蛋白质序列的编码DNA序列或该蛋白质的3D结构图。这种通过计算机运算，即基于比较序列相似性或PubMed中摘要的相似性，所给出的相关链接信息可以以最快的速度提供给用户大量的相关信息。还有一种叫做“LinkOut”的功能将这种链接功能扩展到了与外部数据库，例如各物种基因组数据库之间的链接。Entrez中搜索到的数据可以以多种格式输出，也可以打包下载或逐个下载。

表1 Entrez数据库（2008年9月30日版）

数据库名称	收录记录数目	在本文中的所属类别
Nucleotide	65,786,674	Genes and associated sequences
EST	56,569,180	Genes and associated sequences
SNP	51,242,511	Genotypes and phenotypes
PubChem Substance	44,576,721	Small molecules and bioassays
GEO Profiles	42,751,725	Gene expression
GSS	24,562,212	Genes and associated sequences
Protein	22,337,204	Genes and associated sequences
PubChem Compound	19,327,825	Small molecules and bioassays
PubMed	18,289,697	Literature resources
Probe	9,650,111	Gene expression
Gene	4,962,281	Genes and associated sequences
UniGene	3,488,940	Genes and associated sequences
PubMed Central	1,683,851	Literature resources
NLM Catalog	1,374,580	Literature resources
UniSTS	514,624	Genes and associated sequences
Taxonomy	460,107	Entrez search and retreival system
Protein Clusters	285,386	Genomes
3D Domains	246,719	Molecular structure and proteomics
Books	229,412	Literature resources
MeSH	205,235	Literature resources
Cancer Chromosomes	131,638	Genomes
Homologene	115,467	Genes and associated sequences
PopSet	85,977	Genes and associated sequences
GENSAT	83,553	Gene expression
Structure	53,266	Molecular structure and proteomics
dbGaP	39,617	Genotypes and phenotypes
CDD	26,660	Molecular structure and proteomics
Journals	22,762	Literature resources
OMIM	19,857	Genotypes and phenotypes
GEO Datasets	16,754	Gene expression
Genome	8792	Genomes
Site Search	4402	Introduction
Genome Project	3900	Genomes
OMIA	2577	Genotypes and phenotypes
PubChem Bioassay	1197	Small molecules and bioassays

2.2 My NCBI

My NCBI功能是为了方便用户储存个人配置信息，例如搜索条件、LinkOut参数或文件出处等而设的。用户登陆自己的My NCBI帐户后，就可以进行保存搜索设置、管理邮件等操作了。My NCBI中有一种称作“Collections”的功能可以让用户储存搜索结果和文献结果。BLAST中也设有类似的功能，这样用户就可以使用同一条件进行多次比对了。

2.3 Entrez programming utilities（E-Utilities）

E-Utilities（Entrez应用程序）由8种服务器程序组成，借助E-Utilities可以设置一套标准参数进行搜索、链接和下载数据（表2）。用户可以到NCBI主页上的‘Entrez Tools’链接中了解更多有关E-Utilities的信息。

表2 E-Utilities部分工具及其功能

名称	功能
Einfo	对指定数据库进行基本的统计学分析（包括最近一次数据更新的时间），列出所有搜索范围和可用链接数等。
Esearch	给出符合文本搜索信息的记录标识符，再结合“Efetch”或“Esummary”程序就可以完成相应数据资料的下载。
Elink	向用户提供大量Entrez数据库中的相关信息链接。 将“Elink”程序与URL或简单对象访问协议（SOAP）结合起来，用户就可以自己编写小程序，并在Entrez数据库中进行自动批量搜索，这在以往的页面模式下是不可能完成的。

2.4 Taxonomy

NCBI Taxonomy（分类）数据库在Entrez生物学数据库中起到了组织中心的作用。该数据库为每一个分类学上的节点，从超界节点（superkingdoms）到亚种节点（subspecies），提供数据链接服务。分类数据库以每月增加2200个新分类单位的速度在增长，共收录有将近300,000种物种信息，这些信息为“属(genus)”级别，或者虽然未达到“属(genus)”级别，但在Entrez至少收录有一条该物种的核酸序列或蛋白质序列信息。使用Taxonomy网页可以了解该物种在分类学上的地位，也可以在某一物种范围内对Entrez数据库进行搜索。

3 BLAST序列相似性搜索程序

BLAST程序是一种进行序列相似性搜索的程序，它可以对核酸序列或蛋白质序列进行分析。经过BLAST程序比对之后会得到各种序列结果，例如转录体序列（UniGene）信息、基因序列（Gene）信息、3D结构信息（MMDB）或芯片信息（GEO）等。用户也可以使用My NCBI功能保留BLAST中设定的搜索题目、近期搜索结果和搜索参数等信息。还有一种BLAST程序——BLAST2Sequences程序，它可以对两条DNA序列或蛋白质序列进行比对，并获得一个点对点的比对结果。BLAST程序也可以作为一个独立的程序下载到本地计算机上使用，用户可以到ftp.ncbi.nih.gov/blast/executables/LATEST/下载（表3）。

表3 NCBI上可供下载的软件列表

软件名称	可运行的操作系统	类别
BLAST (stand alone)	Win、Mac、LINUX、Solaris	BLAST
BLAST (network client)	Win、Mac、LINUX、Solaris	BLAST
BLAST (web server)	Mac、LINUX、Solaris	BLAST
CD-Tree	Win、Mac	Molecular structure and proteomics
Cn3D	Win、Mac、LINUX、Solaris	Molecular structure and proteomics
e-PCR	Win、LINUX	Genes and associated sequences
gene2xml	Win、Mac、LINUX、Solaris	Genes and associated sequences
OMSSA	Win、Mac、LINUX	Molecular structure and proteomics
splign	LINUX、Solaris	Genes and associated sequences
tbl2asn	Win、Mac、LINUX、Solaris	Genomes

3.1 BLAST

BLAST默认的比对信息数据库包括NCBI中的人类基因组数据库和人类RefSeq数据库。比对之后，BLAST会按照评分高低、序列相似度对结果进行排序，另外BLAST还可以对小鼠数据库以及其它数据库进行比对。

蛋白质序列的默认数据库包括GenBank非冗余数据库、RefSeq、Swiss-Prot、PDB、PIR和PRF等。此外，还包括这些数据库下的子数据库以及其它一些专利数据库和诸如核酸数据库等环境样品数据库（environmental samples）。

3.2 BLAST output formats

标准的BLAST输出格式包括默认的配对比对格式（default pairwise alignment）、搜索定位的多序列比对格式（query-anchored multiple sequence alignment formats）、简单的可解析的Hit Table格式以及按照分类学给出的报告格式等。一种叫做“按照同一性进行配对（Pairwise with identities）”的格式能更好地突出目标序列与检索序列之间的差别。而Web BLAST中提供的树状浏览格式则会按照搜索出的目标序列与检索序列之间的距离不同将这些目标序列进行聚类，形成一幅树状图来显示结果。BLAST比对之后给出的每一种格式的比对结果都会有一个分值和E值。用户也可以设定一个E值的阈值来筛选比对结果。

3.3 MegaBLAST

MegaBLAST也是一种BLAST程序，不过它主要是用来在非常相似的序列之间（来自同一物种）比对同源性的。使用者通过网页使用MegaBLAST进行批量比对操作，这比使用标准的BLAST程序要快10倍。MegaBLAST在NCBI基因组BLAST页面下是默认的搜索工具，借助它能对增长迅速的Trace Archives数据库和标准BLAST使用的数据库进行快速检索。NCBI还为跨物种核酸序列快速搜索提供了Discontiguous MegaBLAST，它使用非重叠群字段匹配算法（noncontiguous word match）来进行核酸比对。Discontiguous MegaBLAST比blastx等翻译后比对要快得多，同时它在比较编码区时也具有相当高的敏感度。

3.4 Genomic BLAST

NCBI在Map Viewer中还为100多个物种设有Genomic BLAST。通过默认的Genomic BLAST既能对某个物种的基因组序列进行搜索，也能对其它的数据库进行搜索，比如RefSeqs数据库、EST数据库等。

4 文献资源

4.1 PubMed数据库

目前，PubMed数据库中收录有自1860年以来20,400种生命科学类杂志、刊物刊登过的超过1800万条的文献记录。这些文献中有980万条摘要信息，最早的记录可追溯至19世纪80年代，其中有870万条可以检索到

全文。PubMed数据库与其它Entrez数据库都保持着密切联系，这样可以在不同的数据库之间架起一座连接的桥梁。PubMed数据库还会通过计算机自动检索出包含相近MeSH词汇、文献题目以及摘要的相关文献信息提供给用户。默认的“AbstractPlus”输出格式给出了该文献的摘要信息和五篇与该文献相关信息的简单介绍，这样用户就可以获得更多的有关资讯了。

4.2 PubMed Central

PubMed Central是一个收录生命科学领域同行评审期刊（Peer Reviewed Journals）文献的数据库，现收录超过160万条全文文献，并且仅去年一年就增长了51%。而且，包括《核酸研究》（*Nucleic Acids Research*）在内的480多种期刊会为PubMed Central提供全文文献。

所有参与PubMed Central的出版商也都必须在文献出版后12个月之内免费为PubMed Central提供全文文献。由于NIH于2008年4月7日开始执行向公众免费开放使用的政策，故而PubMed Central也必须免费向公众开放使用。如此一来，用户使用Entrez就可以搜索到PubMed和PubMed Central中的所有文献信息了。

4.3 NCBI Bookshelf、NLM Catalog以及Journals database

NCBI Bookshelf通过与作者和出版商合作，收录了86种在线教科书和生物医药类图书。NCBI Bookshelf作为独立于Entrez数据库的一个单独数据库，它里面的信息也可以通过文本搜索或Entrez数据库，例如PubMed、PubMed Central、Gene和OMIM中的链接搜索到。NCBI Bookshelf中的图书不是象普通图书那样一本一本的存放的，而是按照内容将它们分成了230,000个不同的部分、章节进行储存的。用户浏览其中一个内容的时候也可以跳到该书的其它部分或者直接搜索这本书中的特定内容进行阅读。

NLM Catalog为藏书超过130万册的美国国立卫生图书馆（NLM）记录设立目录信息，包括杂志、图书、手稿、计算机软件、录音文件和其它电子资源。每一条记录都可链接到NLM LocatorPlus和具有相近题目或MeSH词汇的相关文件目录信息。

Journals database（期刊数据库）包含了每一个Entrez数据库中的所有期刊信息。目前共收录有超过22,000条记录，期刊数据库为每一份期刊都建立了ISO刊名缩写索引、出版日期索引和NLM catalog链接索引以及Entrez中引用该期刊中文献的索引。

5 基因序列信息以及相关序列信息

5.1 数据库

5.1.1 Entrez Gene

图片来源：Entrez Gene

Entrez Gene数据库为用户提供基因序列注释和检索服务，还会链接到NCBI的Map Viewer、Evidence Viewer、Model Maker、BLAST Link (Blink)、protein domains from the Conserved Domain Database（CDD）等数据库资源以及其它与基因相关的资源。Entrez Gene数据库收录了来自5300多个物种的430万条基因记录。而且，NCBI除了拥有自己的注释工作人员之外，还在不断从许多其它国际合作组织那里获取新的基因注释记录信息。

Entrez Gene数据库与PubMed中最新引文之间的链接是由基因注释人员负责维护的，这项功能也被称作GeneRIF。完整的Entrez Gene数据集以及物种特异性的数据亚集可以在NCBI FTP站点中的NCBI ASN.1中找到。一种可以将NCBI ASN.1格式转化成XML格式的名为ene2xml的软件也可以到ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml下载。

5.1.2 UniGene和ProtEST

UniGene从属于GenBank的一部分，专门收集转录体序列数据，包括EST序列和非冗余序列，每一条UniGene记录都代表一个潜在的基因。UniGene收录了GenBank中来自所有物种的将近70,000条EST序列，这些物种中包括58种动物、43种植物和真菌以及6种真核生物。现在，在构建基因表达谱芯片时都是参考UniGene中的数据来进行设计的。UniGene数据库每周都会更新EST信息，每两个月会更新序列信息。ProtEST作为UniGene序列的辅助确认工具会预先对序列进行BLAST比对，它所使用的比对方式是将UniGene核酸序列的6种可能翻译蛋白质序列与模式生物蛋白质序列进行比对。

5.1.3 HomoloGene数据库

HomoloGene数据库是一个在20种完全测序的真核生物基因组中自动检索同源基因的系统，包括直系同源与旁系同源。HomoloGene的结果报告包括基因同源性和来自OMIM、小鼠基因组信息学（Mouse Genome Informatics, MGI）、斑马鱼信息网络（Zebrafish Information Network, ZFIN）、酵母基因组数据库（Saccharomyces Genome Database, SGD）、直系同源基因簇（Clusters of Orthologous Groups，COG）和果蝇数据库（FlyBase）的基因表型信息。HomoloGene下载功能能下载HomoloGene中的转录体、蛋白质和基因组序列信息，还能下载基因组中特定基因的上游和下游序列。

5.1.4 Reference Sequences

Reference Sequences（RefSeq）数据库是一个收录注释过的非冗余转录体、蛋白质和基因组序列数据库。2008年，Reference Sequences数据库收录的记录增加了40%。同年7月公布的Reference Sequences数据库30共收录了来自5400种不同物种的300万条核酸序列和560万条蛋白质序列。用户可以通过Entrez核酸和蛋白质数据库搜索到RefSeq序列，也可以通过NCBI FTP站点进入RefSeq数据库。

5.1.5 GenBank和其它数据库来源的序列

用户可以通过三个Entrez数据库——Nucleotide、EST和Genome Survey Sequence（GSS）（这三个数据库在E-Utilities中分别称作nuccore、nucest和nucgss）搜索到GenBank中的序列。Entrez Nucleotide数据库含有除了收录之外的GenBank中所有的序列，它还收录有全基因组鸟枪法测序序列、第三方注释序列（Third Party Annotation sequences）和Entrez结构数据库中的序列。对这些记录中编码序列概念上的翻译信息都收录在了Entrez蛋白质数据库中。EST数据库收录了GenBank EST中的所有数据和没有生物学注释信息的“单分子识别首次通过（first-pass single-read）”的cDNA序列。同样，GenBank中的GSS数据库也收录了没有生物学注释信息的单分子识别首次通过的基因组序列。

5.2 分析工具

5.2.1 ORF Finder、Spidey和Splign

NCBI提供了几种分析工具可以帮助用户在基因组内发现编码序列。Open Reading Frame（ORF）Finder程序可以将一段DNA序列按照6种进行翻译，然后返回某一段DNA序列中可能的ORF。

Spidey工具将一组真核生物的mRNA序列与一个基因组序列进行比对，使用4种物种的RNA剪切模型（脊椎动物、果蝇、秀丽隐杆线虫和植物）来预测RNA剪切位点。

Splign是一种通过比对cDNA和基因组序列来发现剪切位点的工具，它可以在测序出现错误的情况下使用，还可以进行跨物种的比对。Splign使用了一种Needleman-Wunsch算法，与区域化算法（compartmentization algorithm）一起使用能发现可能的基因位点。用户可以在Splign网页上下载单独为大批量分析而专门设计的Splign工具使用。

5.2.2 Electronic PCR（e-PCR）

正向e-PCR能在UniSTS数据库收录的超过510,000条STS标记物中搜索到与STS引物配对的序列。反向e-PCR则通过搜索基因组数据库和转录体数据库来估计基因组结合位点、扩增子大小和引物特异性。用户可以在ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR上找到e-PCR的源代码（source code）。

5.2.3 Conserved CDS database（CCDS）

不同的科研小组使用他们各自的方法研究同一物种基因组时，对于基因组中的基因定位可能会得到相似但不完全相同的结论。这样，就会对其它的科研工作者造成困扰。在所有的模式生物中，目前对人类和小鼠的基因组序列研究得最多也最透彻，因此它们最适合用来作为“标准的（consensus）”基因注释的“实验材料”。

CCDS数据库计划（www.ncbi.nlm.nih.gov/CCDS/）就是由NCBI、欧洲生物信息学研究院（European Bioinformatics Institute）、韦尔科姆基金会桑格研究院（Wellcome Trust Sanger Institute）和加州大学圣克鲁兹分校（UCSC）共同合作建立的标准的有关人类和小鼠基因蛋白质编码区的数据库，该数据库会不断更新来保持其高水准。到目前为止，CCDS数据库共收录了超过20,000条人类基因CDS注释数据和17,500条小鼠基因CDS注释数据。用户可以在CCDS的网页上使用基因名或序列ID进行搜索，还可以链接到Entrez Gene数据库、历史记录信息、转录体和蛋白质序列、Map Viewer、Ensemble Genome Browser、UCSC Genome Browser和桑格研究院的Vega Browser。用户可以到ftp.ncbi.nlm.nih.gov/pub/CCDS/下载CCDS序列数据。

6 基因组信息

6.1 数据库

6.1.1 Entrez Genome

图片来源：Entrez Genome

Entrez Genome数据库收录了850多种微生物、3100多种病毒以及1600多种真核生物细胞器的完整基因组数据以及将近50种动物、绿色植物和真菌的700多条染色体信息，总共收录有6200多条序列，其中有882条是去年新增的序列信息。而对于更高等的真核生物基因组，Entrez Genome数据库会直接链接到NCBI Map Viewer。原核生物、病毒和真核生物细胞器的基因组则可以链接到专门的页面和BLAST页面。另外还专门设有植物基因组页面（Plant Genomes Central Web page），在上面可以查询到完整的植物基因组测序计划、植物基因组BLAST或者Map Viewer等信息。

6.1.2 Entrez Genome Project

Entrez Genome Project数据库（Entrez基因组计划数据库）向用户提供了一个有关正在进行中的大规模植物基因组测序、组装、注释和作图工作的全面概况。目前，该数据库显示，一共对2200种植物进行了测序工作，其中750种已经完成了所有工作，700种正处于草图组装阶段。该数据库的规模还在不断扩大，以至于还囊括了多个单独的测序项目，例如病毒群体计划（viral population projects）、对16S核糖体RNA元基因库（16S ribosomal RNA metagenomic）等靶位点的测序计划（targeted locus sequencing projects）以及转录组计划等。Entrez基因组计划数据库与其它Entrez数据库，例如Entrez核酸数据库和Entrez基因组数据库以及NCBI内部或者外部资源都有广泛的联系。Entrez基因组计划还为原核生物的某些特点，例如表型、活力、致病性和对生存盐浓度、温度、氧气浓度、pH值等环境因素设置了索引，这对于研究原核生物的生物学

家们来说无疑是一项非常有用的功能。NCBI鼓励各个测序中心在开始他们的测序项目之前提前登记自己的项目安排，这样就能更好的统筹安排，共享资源了。

6.1.3 NCBI Trace Archives

Trace Archives数据库储存了由凝胶/毛细血管测序平台（例如Applied Biosystems ABI 3730）测序获得的序列数据。至今，Trace Archives数据库包含有4500个品种的共计超过19亿（12%为人类数据）的序列数据。

6.1.4 Short Read Archive

Short Read Archive（SRA）数据库里收录的数据都是由新一代测序仪（例如Roche-454、Illumina Genome Analyzer、Applied Biosystems SOLiD System platforms）测序产生的基因序列信息。从2007年开始，SRA已经迅速累积到了1.3 Tbp，共180亿条小片段，约占人类基因组序列总长度的85%。

SRA的出现为大家进行数据挖掘提供了更多的机会。出于方便广大用户使用的考虑NCBI还将为SRA数据建立索引，同时更多的辅助工具，例如搜索及比对等功能也将陆续开发出来。

6.2 分析工具及资源

6.2.1 Map Viewer

NCBI的Map Viewer显示了基因组集合、遗传标记及物理标记以及相关注释信息和比对信息等其它分析结果。Map Viewer的主页www.ncbi.nlm.nih.gov/mapview/提供了包括人类、小鼠和大鼠（*Rattus norvegicus*）在内的超过100种物种的基因组数据。用户可以看到的图谱将根据物种的不同可能会有所不同，或许包括细胞遗传图谱（cytogenetic maps）、物理图谱（physical maps）和各种不同的序列图谱。源自同一物种的多个基因组图谱可以在同一个页面中显示。

6.2.2 Model Maker以及Evidence Viewer

Model Maker（MM）是用来构建转录模型的一种工具，它将通过由从头预测法（ab initio predictions）预测出来的外显子以及通过与GenBank中的转录体数据库EST和RefSeq比对之后得来的外显子，与NCBI的人类基因组数据库结合在一起来构建转录模型。

Evidence Viewer（EV）则将所有能支持基因注释信息正确性的序列信息证据进行了归纳总结，它采用的是将RefSeq、EST等GenBank中的转录体信息与基因组重叠群进行比对的方法。EV显示了每一个外显子的详细比对结果，并突出显示了其中不匹配的部分。

6.2.3 Entrez cancer Chromosomes

Entrez cancer Chromosomes （Entrez癌症染色体）数据库包含了与人类癌症有关的人类染色体畸变信息，例如基因缺失或转位等。Entrez癌症染色体数据库由三个部分组成，即NCI/NCBI SKY（Spectral Karyotyping）/M-FISH（Multiplex-FISH）和CGH（Comparative Genomic Hybridization）数据库；美国国立癌症研究院（NCI）为癌症染色体畸变信息设立的Mitelman数据库以及NCI为再发癌症染色体畸变设立的数据库。每一个畸变都以图形的形式表现出来，并附之相关临床病例信息和文献信息。

6.2.4 TaxPlot、GenePlot和gMap

TaxPlot可以同时给出来自两个物种蛋白质之间的相似性以及原核生物或真核生物参考物种的完整基因组信息。与其相关的另一个工具GenePlot则可以给出一对完整微生物基因组内的片段，经可视化的缺失、

转位或倒位操作之后，其编码蛋白质之间的相似性。gMap工具将预先计算过的微生物全基因组比较结果与BLAST比较结果以及核酸序列相似的基因组聚类结果结合在一起进行比对，然后将相似的片段以图形化的方式表现出来。

6.2.5 Influenza Genome Sequencing Project（IGSP）

IGSP（流感基因组测序计划）为研究流感的科研工作者提供了越来越多的序列资料，他们可以借此找出流感病毒致病的遗传性状。到目前为止，该计划已经得到了超过33,000条流感病毒序列。NCBI的流感病毒资源也和IGSP之间设有链接，还可以通过PubMed找到所有最新的有关流感病毒方面的文献和各种在线分析工具及数据库资源。这些数据库包括NCBI的流感病毒序列数据库（Influenza Virus Sequence Database），该数据库收录有GenBank和RefSeq中超过70,000条流感病毒的序列。科研人员借助流感病毒资源提供的各种工具能对超过83,000条流感蛋白质序列进行分析。Entrez的生物学数据库中还收录有超过100条流感病毒蛋白质结构信息和350多条有关流感病毒种群研究的资料。还有一种在线流感病毒基因组注释工具能帮助科研工作者们分析新发现的流感病毒序列并进行注释，然后将结果通过tbl2asn等上传工具递交给NCBI的GenBank数据库。

6.2.6 Entrez Protein Clusters

Entrez Protein Clusters（Entrez蛋白质聚类数据库）收录了由完整的原核生物基因组和叶绿体基因组编码的28万多条已确认的RefSeq蛋白质序列，并将这些序列按照分类学的规则进行了归类（聚类）。NCBI可以将这些蛋白质聚类信息用于基因组范围内的比对，也可以用于简化的BLAST——简单的微生物蛋白BLAST（Concise Microbial Protein BLAST, www.ncbi.nlm.nih.gov/genomes/prokhits.cgi）比对之用。蛋白聚类数据库还包括注释信息、出版信息、结构域和结构信息、相关库外链接和分析工具（例如多序列比对工具和系统发生分析工具）信息等。蛋白质聚类数据库还通过Genome ProtMap（http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi）与其它基因组数据库有链接。

7 基因型和表型信息

7.1 基因型和表型数据库

图片来源：dbGaP

认识遗传和环境因素与人类疾病之间的关系，对于帮助我们提高疾病诊治水平来说具有非常重要的意义。大范围的基因型研究能为基因组相关调查、医疗测序、分子诊断以及发现基因型和非临床特性之间的关系等研究提供数据资料。基因型和表型数据库（dbGaP; www.ncbi.nlm.nih.gov/sites/entrez?db=gap）是Entrez系统的一部分，它负责管理与可见特征（表型）相关的遗传特征（基因型）。该数据库收录的资料来自由NIH资助的全基因组关联分析（genome-wide association study, GWAS）结果。（详见grants.nih.gov/grants/gwas/index.htm。）目前dbGaP数据库收录的数据来自25个研究项目，用户可以通过疾病名称或基因名称进行搜索、浏览。

为了保证研究项目的机密性，dbGaP数据库只接受“去识别（de-identified）”的数据，同时还要求使用个人资料（individual-level）的研究者接受审核。不过，用户可以不受任何限制的浏览研究文件、操作流程和项目问卷调查等资料。

7.2 dbSNP

dbSNP数据库（单核苷酸多态性数据库）收录的是单核苷酸多态性信息，例如单个碱基的替换、缺失或插入信息。共收录有将近1800万条人类SNP信息和3300万条其它各物种的SNP信息。dbSNP数据库还收录确

认信息、种群特异性等位基因频率信息（population-specific allele frequencies）和个体基因型信息。所有这些信息都可以在dbSNP数据库的FTP站点中找到。

7.3 供常规临床应用的数据库

7.3.1 dbMHC

dbMHC数据库是有关主要组织相容性复合体（MHC）的数据库。它收录了各种MHC等位基因的变异信息，这些信息与器官移植以及个体对感染性疾病的易感性有非常重要的关系。dbMHC数据库收录了1000多条MHC等位基因序列以及这些等位基因在人群中出现的频率，还收录了人白细胞抗原（HLA）的基因型与全世界临床造血干细胞移植成功率之间的信息。

7.3.2 dbLRC

dbLRC数据库则是全面收录白细胞受体复合物（LRC）等位基因信息，主要关注LRC中的KIR基因。

7.3.3 dbRBC

dbRBC数据库收录的是与红细胞抗原或血型有关的基因及其序列信息。该数据库是将血型抗原基因突变数据库（Blood Group Antigen Gene Mutation Database）中的资源与NCBI中的相关资源整合到一起组建而成的。dbRBC数据库里收录的每一个基因都有详细的信息，还有国际输血学会（ISBT）通过等位基因命名法对血型等位基因的命名。

上述这三个数据库都带有多序列比对工具、分析纯合型或杂合型序列的工具以及DNA探针比对工具。

7.4 OMIM

NCBI的OMIM是Entrez的一个组成部分，主要收录人类基因和遗传病相关信息，它由约翰霍普金斯大学（Johns Hopkins University）的Victor A. McKusick小组负责维护。OMIM数据库收录了疾病表型与基因的相关信息，包括对该遗传病详细的描述、基因名称、遗传方式、基因定位、基因多态性以及详细的相关参考文献信息。OMIM数据库共有将近20,000条记录，涵盖超过12,500个已知的基因位点数据和表型数据。这些记录还与许多其它重要资源，例如位点特异性数据库（locusspecific databases）和GeneTests（www.genetests.org）之间设有链接。

7.5 OMIA

OMIA（动物在线孟德尔遗传）数据库是一个有关动物（除了人类和小鼠）基因和遗传病的数据库，由澳大利亚悉尼大学（University of Sydney, Australia）的Frank Nicholas教授等人建立。该数据库收录了超过2500条记录，其中包括文本信息、参考资料信息以及与OMIM、PubMed和Entrez Gene这些数据库之间的链接。

8 基因表达

8.1 Gene Expression Omnibus（GEO）

图片来源：GEO

GEO（基因表达精选集）是一个储存高通量功能基因组学数据的数据库，这些高通量功能基因组学数据来自芯片和新一代的测序仪得到的试验数据。GEO除了收录基因表达数据之外还收录其它数据，例如基因组拷贝数变异数据、基因组-蛋白相互作用数据以及基因组甲基化数据等。该数据库既接受原始数据，也接受经过处理的数据，不过这些数据都要符合“有关芯片试验的最小信息（minimum information about a microarray experiment，MIAME）”标准。该数据库能存储好几种

格式的数据，包括web格式、spreadsheets格式、XML格式和纯文本格式。GEO数据库被分为两个部分收录在Entrez中，分别是GEO Profiles数据库（它负责收录一个基因在一次试验中的定量基因表达数据）和GEO数据库（收录整个试验的数据）。目前，GEO数据库共收录了由世界各地5000多家实验室提交的超过10,000条试验数据，以及300,000个样品和对500多个物种进行表达谱测量得到的160亿个基因表达丰度数据。

8.2 GENSAT

GENSAT是有关小鼠中枢神经系统基因表达谱的数据库，这些数据是由美国神经障碍和中风研究院（National Institute of Neurological Disorders and Stroke）提供的。GENSAT储存了小鼠大脑的组织切片图像，这些组织切片中都含有各种标签，例如增强的绿色荧光蛋白标签等，这样可以根据标签的荧光强度来判断基因的表达量。GENSAT共收录了8万多幅图像资料，还提供搜索功能、资料下载功能、缩放功能和比对功能。

8.3 Entrez Probe

NCBI Probe database（探针数据库）是一个公共的核酸试剂数据库，它可以提供试剂信息、销售厂家信息、探针有效性信息，还可以计算序列相似性。该数据库储存了960万条探针序列，这些探针可以分为31大类，包括用于基因分型的探针、发现SNP的探针、基因表达探针、基因沉默探针、基因测序探针等等。

9 分子结构和蛋白质组学

9.1 MMDB

图片来源：MMDB

NCBI的MMDB数据库收录了Protein Data Bank数据库中经试验验证过的数据信息，包括蛋白质结构域注释信息、与相关文献的链接信息、蛋白质和核酸序列信息、PDB异基因（PDB heterogens）信息、CDD中的保守结构域信息和经VAST算法计算出的结构邻域（structural neighbors）信息。用户可以通过在MMDB数据库中进行文本搜索得到相关的简要结构信息图，还能链接到NCBI结构和在比对浏览器Cn3D中查看搜索结果。

9.2 分析工具

Blink工具能显示预先计算（pre-computed）出的BLAST比对结果，即与Entrez数据库中每一条蛋白质序列相似的序列。用户可以限定一些参数，例如物种类别或被比对的数据库等来对结果进行筛选。

9.2.1 开放式质谱搜索算法

开放式质谱搜索算法（Open Mass Spectrometry Search Algorithm, OMSSA）是一种与BLAST类似的算法，利用和BLAST中E值一样的方法在已知的蛋白质序列数据库（非冗余数据库或refseq数据库）中找出与待测序列最相近的已知序列。在OMMSA的网页上可以一次分析2000多个样品。用户还可以到ubchem.ncbi.nlm.nih.gov/omssa/download.htm.站点下载可进行更大量分析的OMSSA软件。

9.2.2 HIV-1/Human Protein Interaction Database

美国国立过敏和传染病研究所艾滋病部（The Division of Acquired Immuno Deficiency Syndrome of The National Institute of Allergy and Infectious）与南方研究院（Southern Research Institute）和NCBI合作，建立了HIV-1/Human Protein Interaction Database（HIV-1/人类蛋白相互作用数据库），用来记录HIV-1病毒蛋白和人类宿主细胞蛋白之间的相互作用。在www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html网页上用户可以找到蛋白质在RefSeq中的检索号、Entrez Gene ID号、相互作用的氨基酸位点、对相互作用的简单描述、关键词和PubMed ID号等信息。

10 PubChem

PubChem是NIH设立分子图书馆以及开展研究小分子化学、结构和生物学特性工作的基础。三个Entrez数据库——PCSubstance、PCCompound和PCBioAssay收录了所有的相关信息。这三个数据库共收录有将近4100万条小分子记录和1900万种结构。其中750,000条记录都在PubChem中收录的1200种生物检测方法中的至少一种生物检测试验中具有活性。PubChem不仅与PubMed、PMC等Entrez数据库有链接，还与Entrez Structure和Entrez Protein有链接，这样就将基因组水平的生物大分子与细胞代谢水平的小分子联系起来了。用户可以使用文本在PubChem数据库中进行搜索，也可以使用各种格式的化学分子式或化学结构进行搜索。

更多信息

用户可以在NCBI Bookshelf中找到NCBI手册，该手册详细的介绍了NCBI中的各种资源。在NCBI的主页上还有“教育（Education）”链接，其中有多个教程可供用户学习。用户还可以通过网站地图了解NCBI中的各种资源。在“About NCBI”网页上还有生物信息学入门和其它补充资源。NCBI还设有面向用户的服务人员回答各种问题，用户可以发邮件到info@ncbi.nlm.nih.gov向他们提问。www.ncbi.nlm.nih.gov/About/newsletter.html网页上还有NCBI最新的更新信息。用户还可以到www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html订阅各种更新信息邮件。NCBI现在还设有RSS服务。

原文检索：*Nucleic Acids Research*, 2009, Vol. 37, Database issue D5–D15

YORK /编译

二 蛋白序列数据库

全球蛋白资源数据库UniProt（2009更新）

图片来源：UniProt

UniProt是一个集中收录蛋白质资源并能与其它资源相互联系的数据库，也是目前为止收录蛋白质序列目录最广泛、功能注释最全面的一个数据库。UniProt是由欧洲生物信息学研究所（European Bioinformatics Institute）、美国蛋白质信息资源（Prontein Information Resource）以及瑞士生物信息研究所（Swiss Institute of Bioinformatics）等机构共同组成的UniProt协会（UniProt Consortium）编辑、制作的一个信息资源，旨在为从事现代生物研究的科研人员提供一个有关蛋白质序列及其相关功能方面的广泛的、高质量的并可免费使用的共享数据库。

UniProt是一个向所有使用者免费开放的数据库，全球科研人员都可以登陆网站www.uniprot.org浏览并下载这些资料。借助它，科研人员可以对目的蛋白进行交互式分析或特定的分析。