BIOTECHNOLOGY

21世纪生物技术系列

生物信息学理论与技术

Shengwu Xinxixue
Lilun Yu Jishu

第3版

主编
王廷华
王廷勇
张　晓

科学出版社

21世纪生物技术系列

生物信息学理论与技术

主 编 王廷华 王廷勇 张 晓

科学出版社
北京

内 容 简 介

本书是《21 世纪生物技术系列》的一个分册，书中全面介绍了现有生物信息数据库、网络构建、基因表达谱分析、LncRNA 数据生物信息学分析、miRNA 相关研究、DNA 甲基化检测、生物信息学预测蛋白质相互作用，以及利用生物信息学设计 PCR 引物和探针等。本书图文并茂，既有一定的理论基础，又有很强的实用性，同时总结了编者多年来的实验经验。

本书可供生物医学专业研究生、本科生及从事生物信息学研究的科研人员阅读和实验时参考。

图书在版编目(CIP)数据

生物信息学理论与技术 / 王廷华，王廷勇，张晓主编 . —北京：科学出版社，2015. 3

(21 世纪生物技术系列)

ISBN 978-7-03-043842-3

Ⅰ. 生… Ⅱ. ①王… ②王… ③张… Ⅲ. 生物信息论 Ⅳ. Q811. 4

中国版本图书馆 CIP 数据核字(2015)第 054716 号

责任编辑：丁慧颖 沈红芬 / 责任校对：刘亚琦

责任印制：李 利 / 封面设计：范璧合

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www. sciencep. com

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2015 年 3 月第 一 版 开本：787×1092 1/16

2015 年 3 月第一次印刷 印张：12 1/4 插页：20

字数：338 000

定价：54. 00 元

(如有印装质量问题，我社负责调换)

《21世纪生物技术系列》第3版编审委员会

主　审　李云庆

委　员　（按姓氏笔画排序）

《生物信息学理论与技术》编写人员

主　编　王廷华　王廷勇　张　晓

副主编　王昆华　谢露霜　李翠英

编　者　（按姓氏笔画排序）

习杨彦彬　王乃超　王廷华　王廷勇

王昆华　刘素娟　李劲涛　李翠英

邹　宇　张　砚　张　晓　张连峰

陆炳团　林　娜　尚飞飞　贺琴琴

夏庆杰　黄　强　葛文平　谢露霜

《21世纪生物技术系列》前言

21世纪是生命科学飞速发展的时代。如果说20世纪后半叶是信息时代，那么21世纪上半叶，生命科学将成为主宰。我国加入WTO后与世界科技日益接轨，技术的竞争已呈现出其核心地位和作用。正是在此背景下，为适应我国21世纪生物技术的发展和需求，科学出版社于2005年组织编写了一套融基础理论和实践技术为一体、独具特色、主要面向一线科技人员的学术著作——《21世纪生物技术丛书》，包括《组织细胞化学理论与技术》、《神经细胞培养理论与技术》、《蛋白质理论与技术》、《分子杂交理论与技术》、《PCR理论与技术》、《基因克隆理论与技术》、《抗体理论与技术》和《干细胞理论与技术》共8个分册。本丛书自2005年3月问世以来，即受到了广大生物技术科技工作者的喜爱，2006年1月进行了重印；2009年出版了第2版。本丛书对满足我国日益扩大的科研人员及研究生实践需求，以及推动我国21世纪生物技术的普及和发展起到了积极的作用。

生物技术发展迅速，为了满足广大科技工作者的需求，本丛书于2013年推出第3版。在第2版的基础上，第3版主要对实验技术中的经验体会部分进行了全面增补，同时补充了新的理论技术，包括免疫荧光染色、诱导型干细胞理论与培养、基于病毒载体的转基因及RNA干扰技术、免疫共沉淀与蛋白质相互作用、蛋白芯片等实用技术，并对各技术的相关实践经验进行了更全面的总结。重要的是，为了应对和满足前沿技术的发展需要，推出第3版的同时还增补了4个分册，即《基因沉默理论与技术》、《电生理理论与技术》、《生物信息学理论与技术》和《神经疾病动物模型制备理论与技术》，并将丛书名更改为《21世纪生物技术系列》。至此，本丛书已达12个分册，从行为、形态、细胞、分子生物学、电生理和生物信息等多个层面介绍了目前常用生物技术的基本理论、进展及其相关技术与应用，是我国21世纪生物技术著作中覆盖面最广、影响最大的一套著作。本丛书从培养科学思维能力和科研工作能力的目标出发，以实用性和可操作性为目的，面向我国日益增多的研究生和广大一线科研人员。在编写方式和风格方面，力求强调对基本概念和理论进行简明扼要的阐述，注重基本技术实践，认真总结了编者的实验经验和体会，并提供了大量原版彩图，使丛书在兼顾理论的同时更具实用价值。

本丛书由王廷华教授牵头，邀请国内外一批知名专家教授参加编写和审阅。本丛书是全体参编人员实践经验的总结，对从事科研的研究生和一线研究人员有很好的参考价值。

由于编写时间有限，加之科学技术发展迅速，书中的错误和不足之处在所

难免，恳请各位读者批评指正。

值本丛书出版之际，感谢为我国生物技术及科学发展孜孜不倦、奉献一生的老一辈科学家，他们的杰出工作为我国中青年一代的发展奠定了基础；感谢国内外一批知名专家教授对丛书的指导和审阅；感谢编者们所付出的辛勤劳动；感谢中国解剖学会长期以来对本丛书组织工作的支持；感谢各位同道给予的鼓励和关心！

《21 世纪生物技术系列》编审委员会

2013 年 4 月 8 日

目　录

第一章　生物信息数据库

数据库是指以一定方式储存在一起、能为多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。近年来，随着生物学的发展特别是分子生物学实验数据的积累，国内外建立了各类生物信息学数据库，涵盖了生命科学的各个领域。

生物信息数据库根据数据的来源和数据整理的不同分为一级数据库和二级数据库：一级数据库的数据直接来源于实验原始数据，只经过简单的归类、整理和注释；二级数据库是在一级数据库、实验数据和理论分析的基础上，针对不同的研究内容和需要，对生物学知识和信息进行综合整理构建的数据库。根据数据库存储的内容不同，数据库又分为核酸序列数据库、基因组数据库、核酸/蛋白质结构数据库等。其中核酸序列数据库、基因组数据库等属于一级数据库。

目前生物信息数据库主要有以下几个明显的特征：数据库的更新速度不断加快；数据量呈指数增长趋势；数据库使用频率增长迅速；数据库复杂程度不断增加；数据库网络化；数据库功能面向应用；拥有先进的软硬件配置等。

本章节将对几个重要的核酸序列数据库、蛋白质结构数据库、基因组数据库等进行介绍，主要阐述其数据库的特征、功能及简明使用方法等。

第一节　核酸序列数据库

序列数据库是生物信息最基本的数据库，包括核酸和蛋白质两类，分别以核苷酸碱基序列或氨基酸残基序列为基本内容，并附有注释信息。序列数据库的内容主要包括两部分：原始序列数据（sequence data）和描述这些数据生物学信息的注释。

目前，国际上最权威、最主要的3大核酸序列数据库是美国国家生物技术信息中心（National Center for Biotechnology Informatin，NCBI）所维护的GenBank数据库、欧洲生物信息学研究所（European Bioinformatics Institute，EBI）的EMBL数据库和日本国立遗传学研究所（Japan National Institute of Genetics Center for Information Biology）的DDBJ数据库。这些数据库成立初期，数据相对独立，至1998年这些机构共同成立了国际核苷酸序列数据库协会（International Nucleotide Sequence Database Collaboratin，INSDC），每天进行数据交换，同步更新，因此目前用户在任一数据库得到的信息都是完整和全面的。

一、GenBank数据库

（一）GenBank数据库简介

GenBank数据库是1982年4月由Los Alamos National Lab创建，属一级序列数据库，包

含了目前所有已知的核苷酸序列和蛋白质序列及其相关的文献著作和生物学注释。其数据来源于约 10 万个物种,其中 56%是人类的基因组序列。GenBank 的数据来源有 3 种:直接来源于测序工作者提交的序列;INSDC 交换和共享数据;美国专利局提供的专利数据。目前世界上的权威期刊在作者发表论文时都要求提供 GenBank 数据库的序列号。用户通过 NCBI 数据库中的 Entrez 检索查询系统可以方便地检索 GenBank 数据库的核苷酸数据,还可以检索 GenBank 数据库和其他数据库的蛋白质序列数据、基因组图谱数据、来自分子模型数据库(MMDB)的蛋白质三维结构数据、种群序列数据集以及 Pubmed 和 MEDLINE 中的文献数据等。

(二) GenBank 数据库的使用

进入 NCBI 主页(http://www.ncbi.nlm.nih.gov/),主页上方的基本检索输入框打开下拉式菜单,选择查询的数据库,在查询栏中输入检索内容,单击“Search”按钮获取检索结果。可是在 NCBI 主页的中没有“GenBank”选项,选择菜单内的“Nucleotide”、“EST”或“GSS”数据库可以获取 GenBank 中核苷酸序列和其他信息。本部分将以“Nucleotide”数据库为例,查询细胞因子“IL-1β”的信息,展示 GenBank 的使用。

(1) 进入 NCBI 主页,选择“Nucleotide”,查询栏中输入“IL-1beta”,如图 1-1。

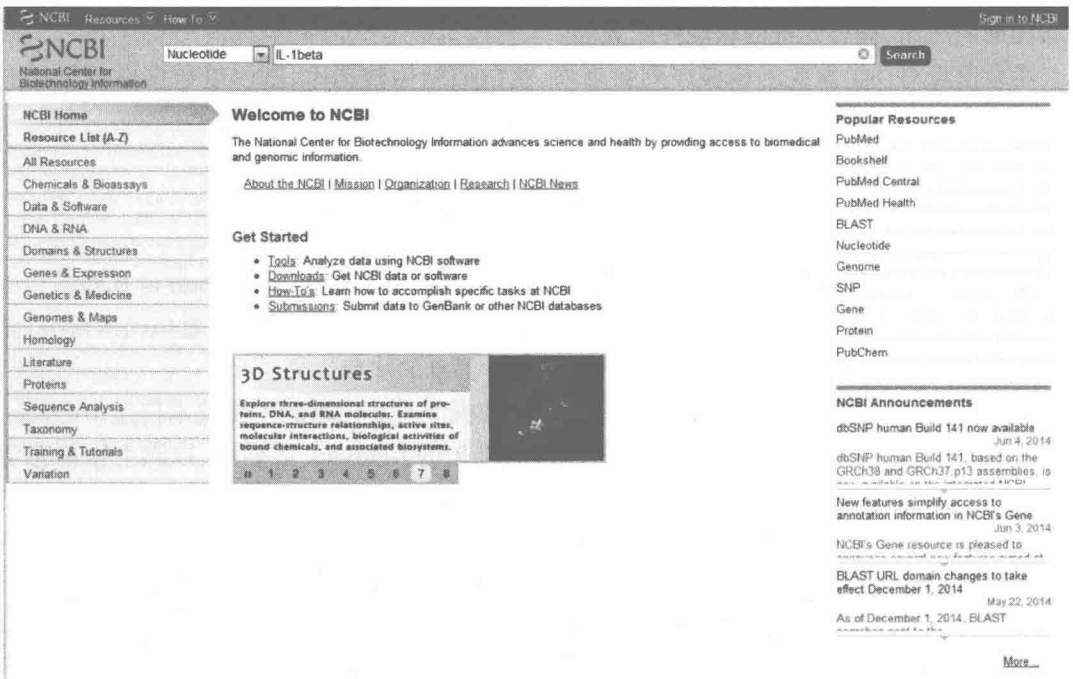

图 1-1 NCBI 页面

(2) 点击 Search 按钮获取相关信息共 647 条(2014.6.6),在右上角的“Filter your results”中可见 INSDC(GenBank)共 268 条,如图 1-2。

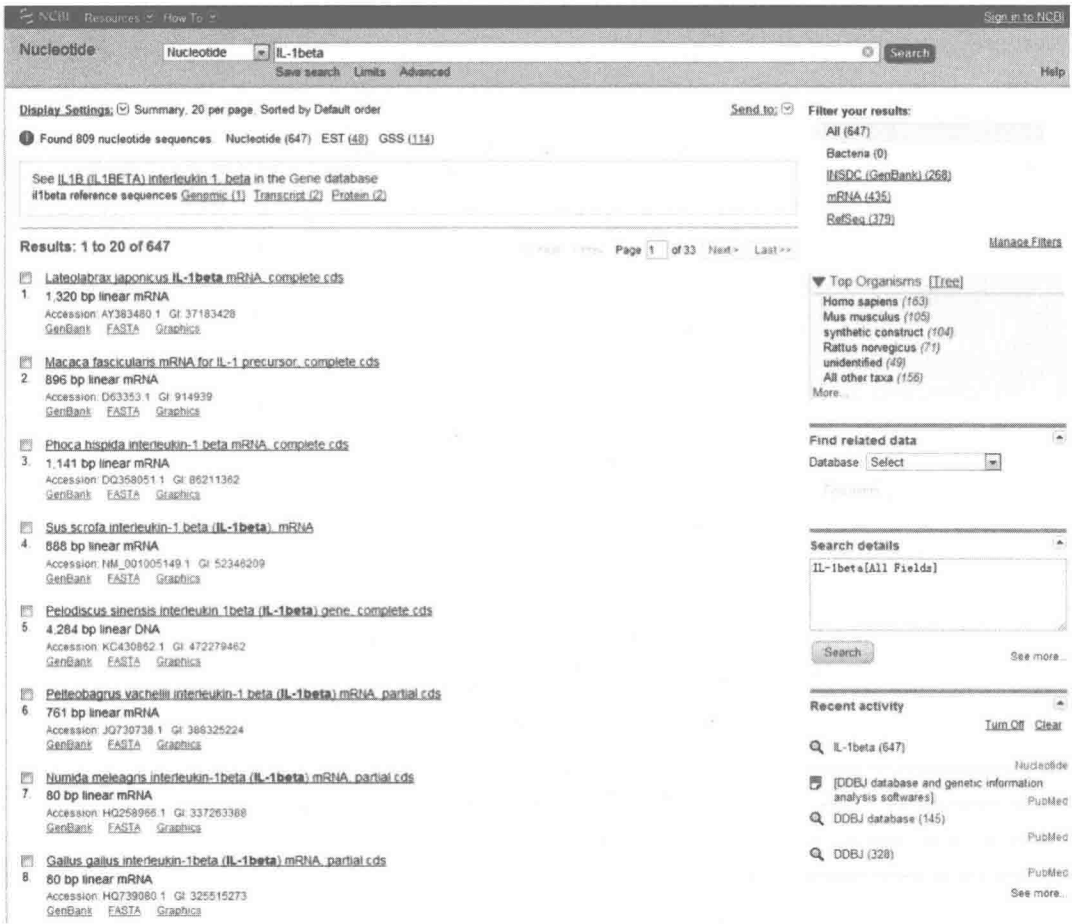

图 1-2　INSDC 检索结果

(3) 根据研究需要选择相关条目,研究者也可以在初选时进行物种等的限制,本例提取种系小鼠(Mus)的 IL-1β 完整 mRNA 序列。点击进入,如图 1-3。

(4) 研究者可以根据目的查询与使用核苷酸信息,如点击“Gene”可获取 IL-1β mRNA 的全部序列,点击“CDS”可获取 IL-1β 的蛋白质翻译序列等。

二、EMBL 数据库

(一) EMBL 数据库简介

EMBL 数据库 1980 年由德国科隆大学收集整理,是世界上第一个核酸序列数据库。数据主要来源于基因组计划、序列中心、科研工作者提交的序列和专利局提供的专利数据等,随着 INSDC 项目的实施,EMBL 数据库数据增长更快,数据信息更完善。EMBL 数据库管理

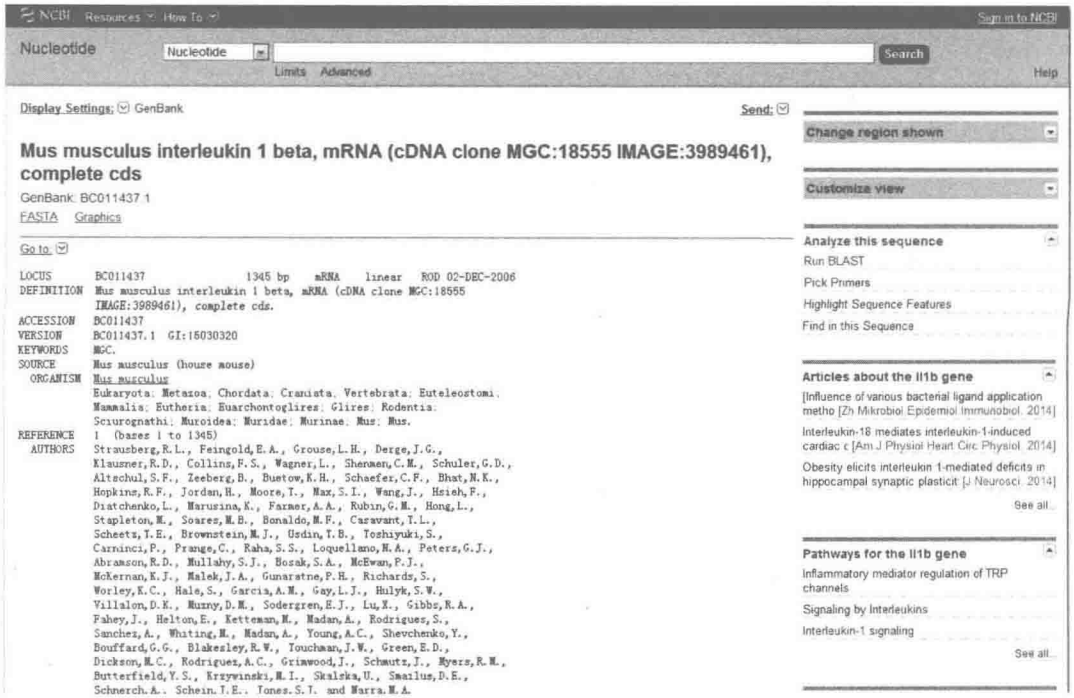

图 1-3 种系小鼠的 IL-1β 完整 mRNA 序列的检索结果

团队负责检查审核新提交的数据,他们确保数据的基本信息完整,序列的生物功能已经过充分描述,任何编码区的注释遵循通用的转换规则。EMBL 数据库条目对序列包含足够的生物信息描述,三种数据类型包括 EST(表达序列标签)、HTG(高通量序列)和 GSS(基因组序列)。EMBL 的一个重要特征是核苷酸的蛋白质编码区(CDS),其 CSD 条目可自动添加到 TrEMBL 数据库,并可以被选择列入 SWISS-PROT 蛋白质序列数据库,因此 EMBL 数据库可以同 TrEMBL、SWISS-PROT 进行信息的共享。

(二) EMBL 数据库的使用

EMBL 数据库的信息存储和维护在 Oracle 数据管理系统中,在互联网上通过序列检索系统 SRS 服务可查询数据信息。网址为 http://www.ebi.ac.uk/ena/。进入主界面后,ENA(Europen Nucleotide Archive)提供两类查询形式,一种是 Text Search(文本查询),可输入已知基因的编号、名字等,另一种为 Sequence Search(序列查询),可输入待查询的核苷酸序列或序列编号,如图 1-4。本部分以小鼠种系的细胞因子 IL-6 为例,展示 ENA 的使用。

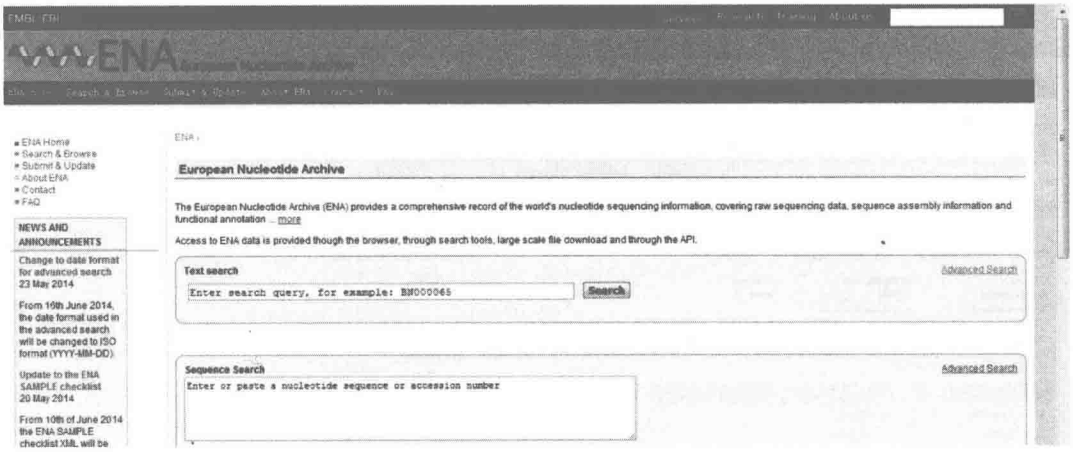

图 1-4　EMBL 数据库主界面

（1）进入 ENA 主页，在 Text search 搜索栏中输入已知 IL-6 的编号：AAI32459，如图1-5。点击 Search 按钮。

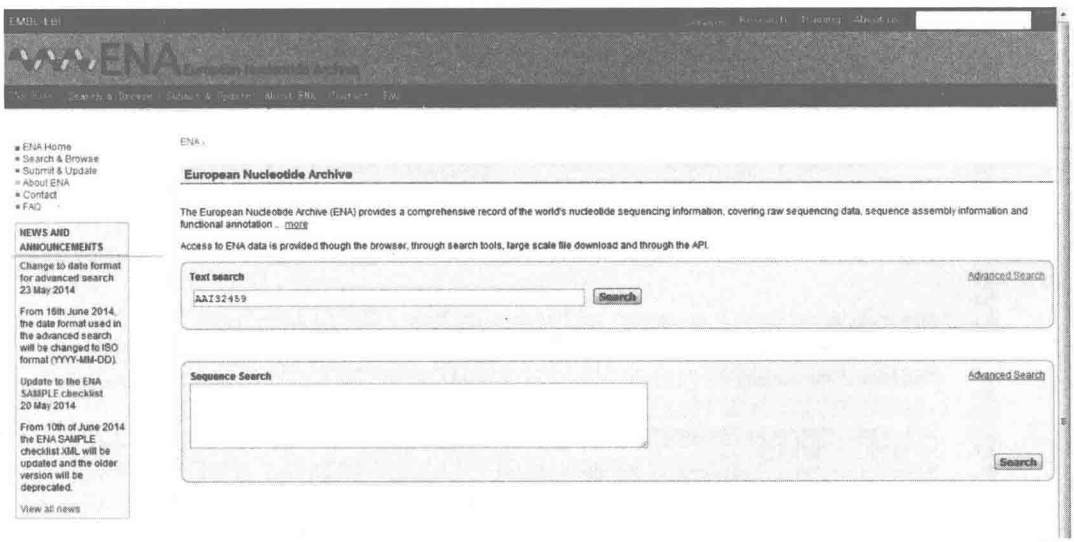

图 1-5　IL-6 检索结果

（2）获取 IL-6mRNA 信息，如图 1-6，在页面的最下端是 IL-6mRNA 的全序列信息，在页面的上端提供了序列信息可供选择的格式，包括 TEXT，FASTA 和 XML。点击相应的链接可获取不同的文件形式，如图 1-7～图 1-9。

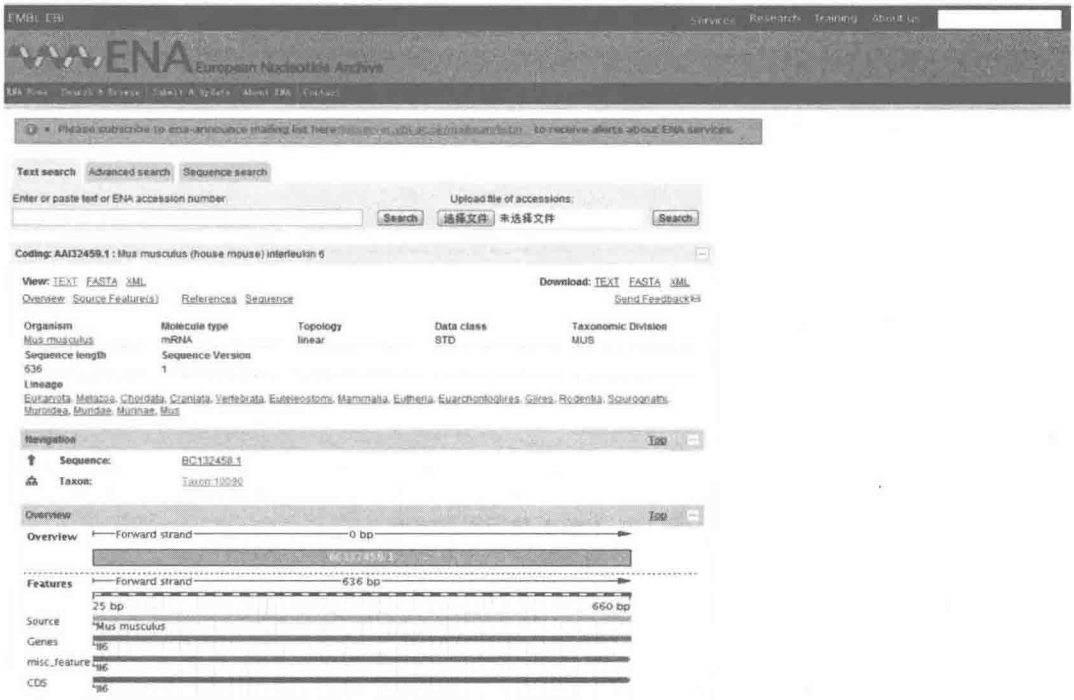

图 1-6 IL-6mRAN 检索结果

```
ID   AAI32459; SV 1; linear; mRNA; STD; MUS; 636 BP.
XX
PA   BC132458.1
XX
DT   02-FEB-2007 (Rel. 90, Created)
DT   24-SEP-2008 (Rel. 97, Last updated, Version 3)
XX
DE   Mus musculus (house mouse) interleukin 6
XX
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Mus; Mus.
XX
RN   [1]
RX   DOI; 10.1073/pnas.242603899.
RX   PUBMED; 12477932.
RG   Mammalian Gene Collection Program Team
RA   Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G., Klausner R.D.,
RA   Collins F.S., Wagner L., Shenmen C.M., Schuler G.D., Altschul S.F.,
RA   Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K., Hopkins R.F., Jordan H.,
RA   Moore T., Max S.I., Wang J., Hsieh F., Diatchenko L., Marusina K.,
RA   Farmer A.A., Rubin G.M., Hong L., Stapleton M., Soares M.B., Bonaldo M.F.,
RA   Casavant T.L., Scheetz T.E., Brownstein M.J., Usdin T.B., Toshiyuki S.,
RA   Carninci P., Prange C., Raha S.S., Loquellano N.A., Peters G.J.,
RA   Abramson R.D., Mullahy S.J., Bosak S.A., McEwan P.J., McKernan K.J.,
RA   Malek J.A., Gunaratne P.H., Richards S., Worley K.C., Hale S., Garcia A.M.,
RA   Gay L.J., Hulyk S.W., Villalon D.K., Muzny D.M., Sodergren E.J., Lu X.,
RA   Gibbs R.A., Fahey J., Helton E., Ketteman M., Madan A., Rodrigues S.,
RA   Sanchez A., Whiting M., Madan A., Young A.C., Shevchenko Y., Bouffard G.G.,
RA   Blakesley R.W., Touchman J.W., Green E.D., Dickson M.C., Rodriguez A.C.,
RA   Grimwood J., Schmutz J., Myers R.M., Butterfield Y.S., Krzywinski M.I.,
RA   Skalska U., Smailus D.E., Schnerch A., Schein J.E., Jones S.J., Marra M.A.;
RT   "Generation and initial analysis of more than 15,000 full-length human and
RT   mouse cDNA sequences";
RL   Proc. Natl. Acad. Sci. U.S.A. 99(26):16899-16903(2002).
XX
```

图 1-7

```
RN   [2]
RC   NIH-MGC Project URL: http://mgc.nci.nih.gov
RG   NIH MGC Project
RA   ;
RT   ;
RL   Submitted (31-JAN-2007) to the INSDC.
RL   National Institutes of Health, Mammalian Gene Collection (MGC), Bethesda,
RL   MD 20892-2590, USA
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..636
FT                   /organism="Mus musculus"
FT                   /lab_host="DH10B"
FT                   /mol_type="mRNA"
FT                   /clone_lib="NIH_MGC_415"
FT                   /clone="MGC:164089 IMAGE:40130735"
FT                   /tissue_type="Lung, mouse"
FT                   /db_xref="taxon:10090"
FT   CDS             BC132458.1:25..660
FT                   /codon_start=1
FT                   /gene="Il6"
FT                   /product="interleukin 6"
FT                   /db_xref="GOA:A2RTD1"
FT                   /db_xref="InterPro:IPR003573"
FT                   /db_xref="InterPro:IPR003574"
FT                   /db_xref="InterPro:IPR009079"
FT                   /db_xref="InterPro:IPR012351"
FT                   /db_xref="MGI:MGI:96559"
FT                   /db_xref="UniProtKB/TrEMBL:A2RTD1"
FT                   /protein_id="AAI32459.1"
FT                   /translation="MKFLSARDFHPVAFLGLMLVTTTAFPTSQVRRGDFTEDTTPNRPV
FT                   YTTSQVGGLITHVLWEIVEMRKELCNGNSDCMNNDDALAENNLKLPEIQRNDGCYQTGY
FT                   NQEICLLKISSGLLEYHSYLEYMKNNLKDNKKDKARVLQRDTETLIHIFNQEVKDLHKI
FT                   VLPTPISNALLTDKLESQKEWLRTKTIQFILKSLEEFLKVTLRSTRQT"
XX
SQ   Sequence 636 BP; 218 A; 142 C; 121 G; 155 T; 0 other;
     atgaagttcc tctctgcaag agacttccat ccagttgcct tcttgggact gatgctggtg        60
     acaaccacgg ccttccctac ttcacaagtc cggagaggag acttcacaga ggataccact       120
     cccaacagac ctgtctatac cacttcacaa gtcggaggct taattacaca tgttctctgg       180
     gaaatcgtgg aaatgagaaa agagttgtgc aatggcaatt ctgattgtat gaacaacgat       240
     gatgcacttg cagaaaacaa tctgaaactt ccagagatac aaagaaatga tggatgctac       300
     caaactggat ataatcagga aatttgccta ttgaaaattt cctctggtct tctggagtac       360
     catagctacc tggagtacat gaagaacaac ttaaaagata acaagaaaga caaagccaga       420
     gtccttcaga gagatacaga aactctaatt catatcttca accaagaggt aaaagattta       480
     cataaaatag tccttcctac cccaatttcc aatgctctcc taacagataa gctggagtca       540
     cagaaggagt ggctaaggac caagaccatc caattcatct tgaaatcact tgaagaattt       600
     ctaaaagtca ctttgagatc tactcggcaa acctag                                 636
```

图 1-7(续)　IL-6mRNA 的 TEXT 格式序列信息

```
>ENA|AAI32459|AAI32459.1 Mus musculus (house mouse) interleukin 6
ATGAAGTTCCTCTCTGCAAGAGACTTCCATCCAGTTGCCTTCTTGGGACTGATGCTGGTG
ACAACCACGGCCTTCCCTACTTCACAAGTCCGGAGAGGAGACTTCACAGAGGATACCACT
CCCAACAGACCTGTCTATACCACTTCACAAGTCGGAGGCTTAATTACACATGTTCTCTGG
GAAATCGTGGAAATGAGAAAAGAGTTGTGCAATGGCAATTCTGATTGTATGAACAACGAT
GATGCACTTGCAGAAAACAATCTGAAACTTCCAGAGATACAAAGAAATGATGGATGCTAC
CAAACTGGATATAATCAGGAAATTTGCCTATTGAAAATTTCCTCTGGTCTTCTGGAGTAC
CATAGCTACCTGGAGTACATGAAGAACAACTTAAAAGATAACAAGAAAGACAAAGCCAGA
GTCCTTCAGAGAGATACAGAAACTCTAATTCATATCTTCAACCAAGAGGTAAAAGATTTA
CATAAAATAGTCCTTCCTACCCCAATTTCCAATGCTCTCCTAACAGATAAGCTGGAGTCA
CAGAAGGAGTGGCTAAGGACCAAGACCATCCAATTCATCTTGAAATCACTTGAAGAATTT
CTAAAAGTCACTTTGAGATCTACTCGGCAAACCTAG
```

图 1-8　IL-6mRNA 的 FASTA 格式序列信息

```
<?xml version="1.0" encoding="UTF-8"?>
<ROOT request="AAI32459&amp;display=xml">
<entry accession="AAI32459" version="1" entryVersion="3" dataClass="STD" taxonomicDivision="MUS" moleculeType="mRNA" sequenceLength="636" topology="linear" firstPublic="2007-02-02"
firstPublicRelease="90" lastUpdated="2008-09-24" lastUpdatedRelease="97">
    <description>Mus musculus (house mouse) interleukin 6</description>
    <reference type="article" number="1">
        <title>Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences</title>
        <author>Strausberg R.L.</author>
        <author>Feingold E.A.</author>
        <author>Grouse L.H.</author>
        <author>Derge J.G.</author>
        <author>Klausner R.D.</author>
        <author>Collins F.S.</author>
        <author>Wagner L.</author>
        <author>Shenmen C.M.</author>
        <author>Schuler G.D.</author>
        <author>Altschul S.F.</author>
        <author>Zeeberg B.</author>
        <author>Buetow K.H.</author>
        <author>Schaefer C.F.</author>
        <author>Bhat N.K.</author>
        <author>Hopkins R.F.</author>
        <author>Jordan H.</author>
        <author>Moore T.</author>
        <author>Max S.I.</author>
        <author>Wang J.</author>
        <author>Hsieh F.</author>
        <author>Diatchenko L.</author>
        <author>Marusina K.</author>
        <author>Farmer A.A.</author>
        <author>Rubin G.M.</author>
        <author>Hong L.</author>
        <author>Stapleton M.</author>
        <author>Soares M.B.</author>
        <author>Bonaldo M.F.</author>
        <author>Casavant T.L.</author>
        <author>Scheetz T.E.</author>
        <author>Brownstein M.J.</author>
        <author>Usdin T.B.</author>
        <author>Toshiyuki S.</author>
        <author>Carninci P.</author>
        <author>Prange C.</author>
        <author>Raha S.S.</author>
        <author>Loquellano N.A.</author>
        <author>Peters G.J.</author>
        <author>Abramson R.D.</author>
        <author>Mullahy S.J.</author>
        <author>Bosak S.A.</author>
        <author>McEwan P.J.</author>
        <author>McKernan K.J.</author>
        <author>Malek J.A.</author>
        <author>Gunaratne P.H.</author>
        <author>Richards S.</author>
        <author>Worley K.C.</author>
        <author>Hale S.</author>
        <author>Garcia A.M.</author>
        <author>Gay L.J.</author>
        <author>Hulyk S.W.</author>
        <author>Villalon D.K.</author>
        <author>Muzny D.M.</author>
        <author>Sodergren E.J.</author>
```


图 1-9 IL-6mRNA 的 XML 格式序列信息

三、DDBJ 数据库

日本 DNA 数据库 DDBJ(DNA Data Bank of Japan),于 1984 年建立,DDBJ 主要向研究者收集 DNA 序列信息并赋予其数据存取号,信息来源主要是日本的研究机构,亦接受其他国家呈递的序列。他们开发了 SQmateh 工具,用来搜索基因或蛋白质中短的碱基或氨基酸序列区域,并建立了简便且易操作的 SOAP(simple object access protocol)服务器。它的数据主要通过 Sakura 和 MST 工具来完成,主页见图 1-10。