

# 生物信息学

## 数据库

张礼斌：华中科技大学生命科学与技术学院

Email: [libinzhang@hust.edu.cn](mailto:libinzhang@hust.edu.cn)

# 生物信息学

- **说文解字：** 生物 + 信息 + 学 (bioinformatics)  
biology + information + theory
- **广义：** 应用信息科学的方法和技术，研究生物体系和生物过程中信息的存贮、信息的内涵和信息的传递，研究和分析生物体细胞、组织、器官的生理、病理、药理过程中的各种生物信息，或者说也可以说成是生命科学中的信息科学。
- **狭义：** 应用信息科学的理论、方法和技术，管理、分析和利用生物分子数据。

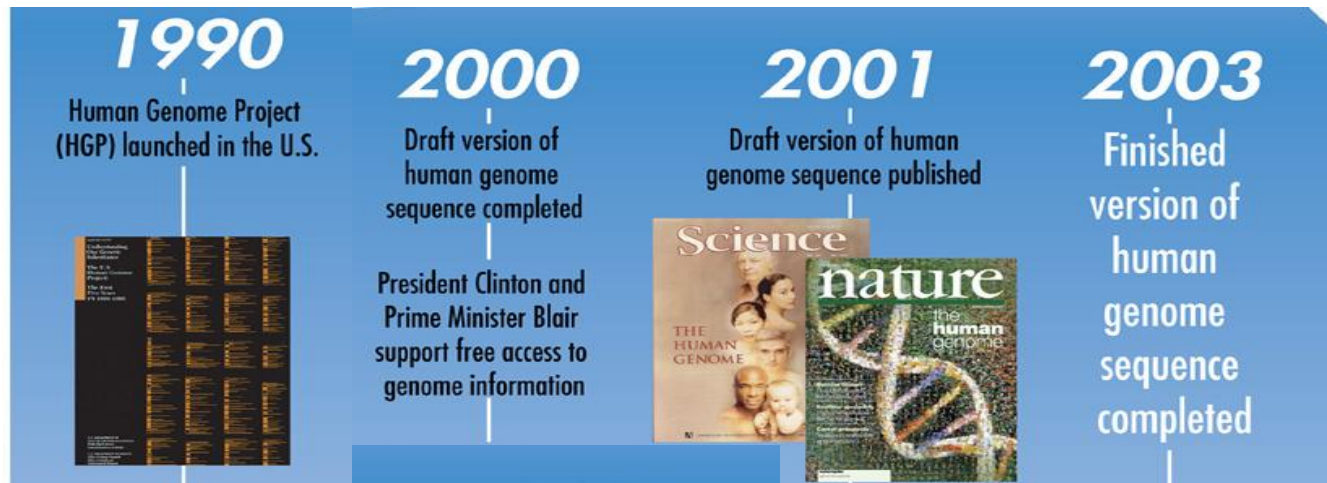
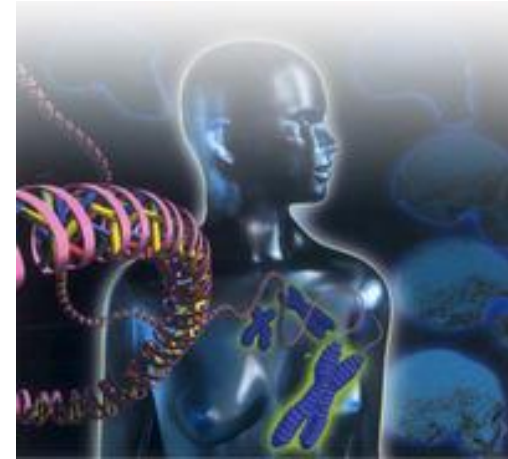
# 人类基因组计划

为什么要开展人类基因组计划？

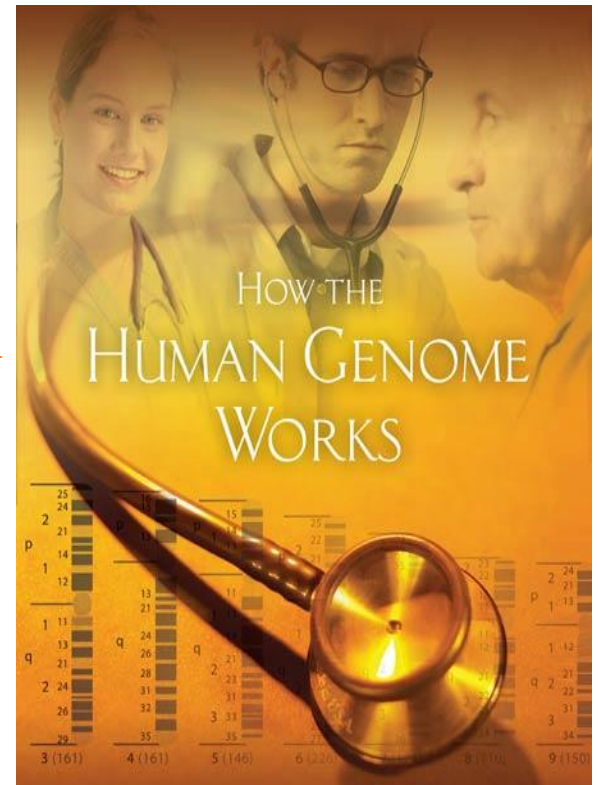
1984.12 犹他州阿尔塔组织会议，初步研讨测定人类整个基因组DNA序列的意义

1985 Dulbecco在《Science》撰文“肿瘤研究的转折点：人类基因组的测序”

有助于认识自身、掌握生老病死规律、疾病的诊断和治疗、了解生命的起源。



# 人类基因组计划目标



Human Genome = three billion ( $3 \times 10^9$ ) base pairs

人类基因组计划  
(HGP, Human Genome Project)

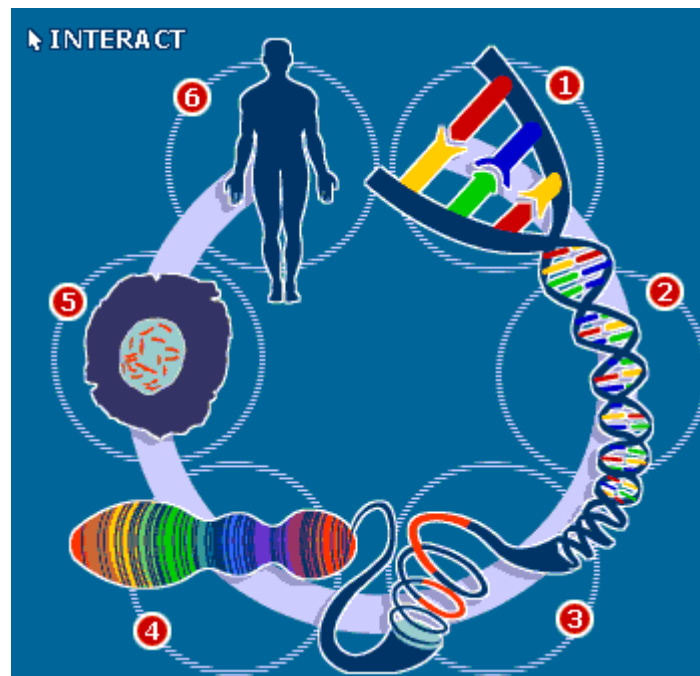
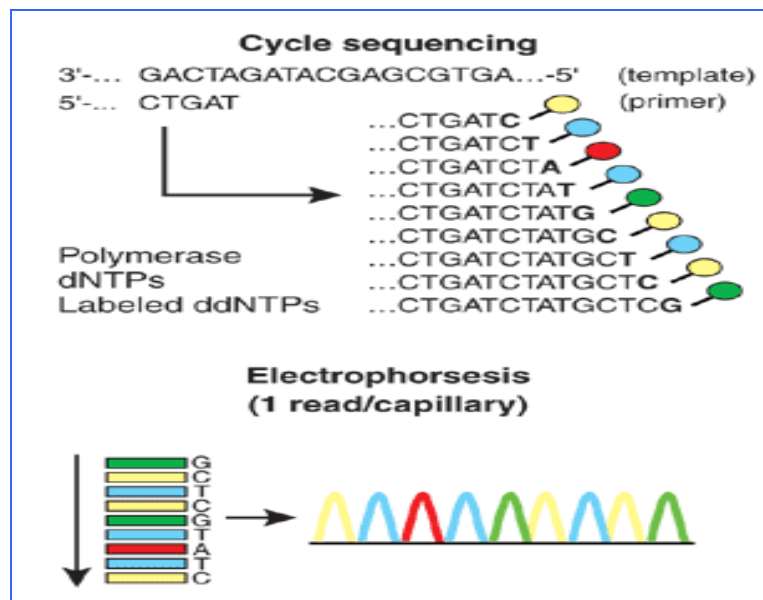
**目标：整体上破解人类遗传信息的奥秘**

# 人类基因组计划-DNA 测序技术

## ■ Sanger测序法

双脱氧链终止法

Sanger测序法



## 新的测序技术

- 焦磷酸测序法（454, Solexa, Solid），单分子测序
- 新的整合技术



生物分子数据  
高速增长

分子生物学  
及相关领域研究人员  
迅速获得最新实验数据



建立生物分子数据库

# 什么是数据库 (Database)

---

**用于收集、整理、储存、加工、发布和检索数据的系统。**

- ◆ **生物类的数据库种类很多 (序列、结构、生物分子互作、其它)**
- ◆ **投稿文章首先要将核苷酸序列或蛋白质序列提交到相应的数据库中**



# 什么是数据库（Database）

---

- ◆ **数据库记录通常包括两部分**
  - ❖ **原始数据**
  - ❖ **对这些数据进行的生物学意义的注释**
- ◆ **一个数据库通常链接了多个相关数据库**

## 生物分子数据库应满足**5**个方面的主要需求：

- ❖ **时间性：**对于新发表的数据，应该能够在很短的时间内（几个小时至几天）通过国际互连网访问
- ❖ **注释：**对于每一个基本数据（如序列），应附加一致的、深层次的辅助说明信息
- ❖ **支撑数据：**在有些情况下，数据库使用者需要得到原始的实验数据，因而要提供访问原始数据的方法。数据库中应包含原始数据，或者能够通过交叉索引访问实验数据库中的原始数据

❖ **数据质量：** 必须保证数据库中数据的质量，数据库管理机构应对数据来源进行检查，并且关注数据库用户和专家提出的意见

❖ **集成性：** 三种基本生物分子数据库（核酸序列、蛋白质序列、蛋白质结构）的集成对于用户来说是非常重要的。对于数据库中的每一个数据对象，必须与其它数据库中的相关数据联系起来，这样就可以从某些分子数据出发得到一系列的相关信息。例如，从某个核酸序列出发，通过交叉索引，可进一步得到对应的基因、蛋白质序列、蛋白质结构，甚至得到蛋白质功能的信息

# ❖ 生物分子数据库

## 一级数据库

- ❖ 数据库中的数据直接来源于实验获得的原始数据，只经过简单的归类整理和注释

## 二级数据库

- ❖ 对原始生物分子数据进行整理、分类的结果，是在一级数据库、实验数据和理论分析的基础上针对特定的应用目标而建立的。

# (一) 数据库工具

## ◆ 建立纯文本数据库

- ❖ GenBank 数据库、EMBL 核苷酸数据库

## ◆ 数据库工具

Access

SQL

Oracle

- ❖ SQL (结构化查询语言) 是世界上流行的和标准化的数据库语言
- ❖ 能够快速灵活存储记录文件和图像
- ❖ MySQL 下载网址

<https://www.mysql.com/cn/>

## （一）数据库工具

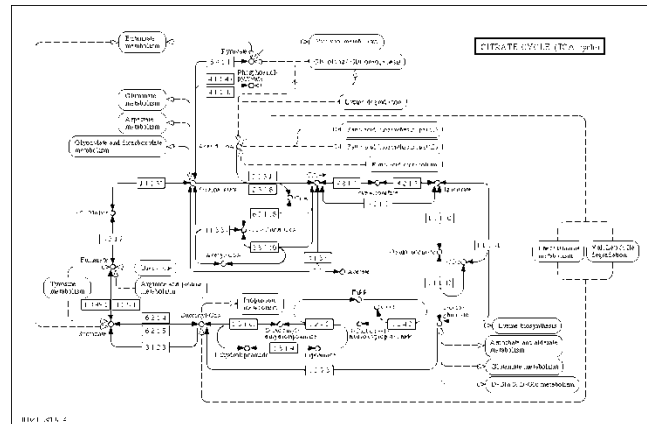
# AceDB 数据库工具

- ❖ **AceDB: A C. elegans DataBase (线虫数据库)**
- ❖ **被广泛应用的管理和提供基因组数据的工具**
- ❖ **数据形式丰富**

## ✓ 遗传图谱

## ✓ 物理图谱

## ✓ 新陈代谢途径



1	gggctccacc	actagtaccc	ctcactacag	gtagccataa	aaaaaatcga	tcacccaaaac
61	ccattatttag	gttgtgtact	gatacagaaa	gttggaacc	aatctcccag	cacagaaaac
121	ggtacggttc	attagcgcgt	gattaattaa	atattttacta	tttttttaaaa	aaaatagatc
181	aatatgattt	ttaagcaact	ttcgtataaa	tacttttttca	aaaaaacaca	ccgtttttcta
241	gtttgaaaag	cgtacacgcg	tgaaatgagg	gagaaagggtt	ggaaacgtgg	gattgcaaac

## (二) 各种生物数据库

### 1、核苷酸数据库

- ◆ DNA、mRNA、tRNA、rRNA序列
- ◆ RNA序列以cDNA序列的形式收集
- ◆ 核苷酸序列直接来源于实验数据
- ◆ 大量氨基酸序列
  - ❖ 主要是非实验来源数据
  - ❖ coding sequence (CDS)

# 核苷酸数据库

## ◆ 三大核苷酸数据库

❖ GenBank、EMBL核苷酸数据库、DDBJ



**信息资源共享：以天为基础进行数据库之间的序列数据交换**

## ❖ 专利核苷酸序列

- ✓ United States Patent and Trademark Office (USPTO)
- ✓ European Patent Office (EPO)
- ✓ Japan Patent Office (JPO)



# 核苷酸数据库

---

**(1) GenBank**      <http://www.ncbi.nlm.nih.gov/genbank>

- ◆ **美国NCBI的核苷酸数据库，包括部分蛋白质序列**
- ◆ **数据每天更新**
  - ❖ 132,015,054 sequences
  - ❖ 124,277,818,310 bases
- ◆ **来源于380,000多个物种**
- ◆ **大约12%的序列来源于人 (*Homo sapiens*)**

# (1) GenBank

NCBI Entrez Protein

My NCBI  
[Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display GenP Accession number

Range: from [ ] CDD Refresh

☐ 1: [NP\\_477382](#). Reports Bub1 CG7838-PA [D...[gi:17137586

BLink, Conserved Domains, Links

Comment Features Sequence

LOCUS NP\_477382 1460 aa linear INV 12-OCT-2006

DEFINITION Bub1 CG7838-PA [Drosophila melanogaster].

ACCESSION NP\_477382

VERSION NP\_477382.1 GI:17137586

DBSOURCE REFSEQ: accession [NM\\_058034.3](#)

KEYWORDS .

SOURCE Drosophila melanogaster (fruit fly)

ORGANISM [Drosophila melanogaster](#)  
Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;  
Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;  
Ephydroidea; Drosophilidae; Drosophila.

REFERENCE 1 (residues 1 to 1460)

AUTHORS Qian, H., Bergman, C.M., and Anxolabehere, D.,

TITLE C. elegans genome annotation of

JOURNAL Proc. Comput. Biol. 1 (2), e22 (2005)

PUBMED [16110336](#)

REFERENCE 2 (residues 1 to 1460)

AUTHORS Hoshino, D., Smith, C.D., Gendron, J.W., de Camargo, A.P.

序列长度

数据类型

Definition: 标题

版本号

GI number

# GenBank

---

## The divisions of GenBank

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. STS - STS sequences (sequence tagged sites)
14. GSS - GSS sequences (genome survey sequences)
15. HTG - HTG sequences (high-throughput genomic sequences)
16. HTC - unfinished high-throughput cDNA sequencing
17. ENV - environmental sampling sequences
18. TSA-Transcriptome Shotgun Assembly
19. PAT - patent sequences
20. WGS-whole genome shotgun

## (2) EST数据库

---

**dbEST (Database of Expressed Sequence Tags)**

**<http://www.ncbi.nlm.nih.gov/dbEST/index.html>**

- ◆ **GenBank的二级数据库**
- ◆ **5' 端或3' 端的cDNA 序列 (EST)**
- ◆ **200-500 bp**

**GenBank 中60%以上的序列是 EST**

### (3) TSA数据库



## Transcriptome Shotgun Assembly Sequence Database

<http://www.ncbi.nlm.nih.gov/genbank/TSA.html>

**TSA** is an archive of computationally assembled sequences from primary data submitted to dbEST, the Short Read Archive (SRA), or the Trace Archive. The overlapping sequence reads from a complete transcriptome are assembled into transcripts by computational methods instead of by traditional cloning and sequencing of cloned cDNAs.

The primary sequence data used in the assemblies and the assemblies must be submitted by the same submitter.

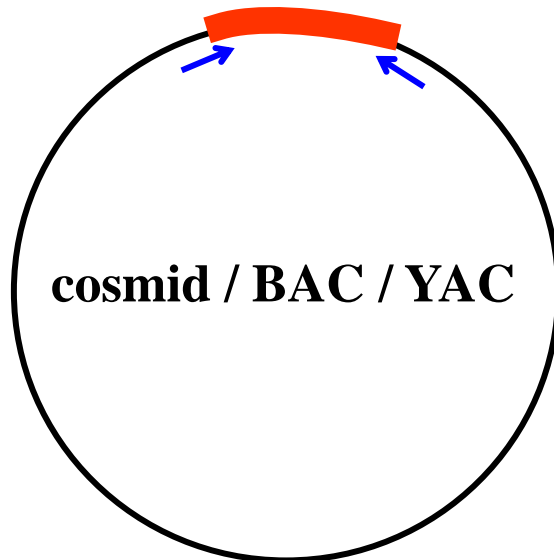
TSA sequence records differ from EST and GenBank records because there are no physical counterparts to the assemblies inserted in the TSA record.

## (4) GSS数据库

dbGSS (Database of Genome Survey Sequences)

<http://www.ncbi.nlm.nih.gov/dbGSS/index.html>

- ◆ GenBank的二级数据库
- ◆ 基因组短序列
- ◆ cosmid / BAC / YAC 外源插入片段的末端序列
- ◆ Alu PCR 序列



## (5) HTGS数据库

**HTGS (High-Throughput Genomic Sequences)**

<http://www.ncbi.nlm.nih.gov/HTGS/>

- ◆ **GenBank 的二级数据库**
- ◆ **尚未完成测序的重叠群 (> 2 kb) 的序列**
- ◆ **新序列的增加速度很快**

## (6) 基因组数据库

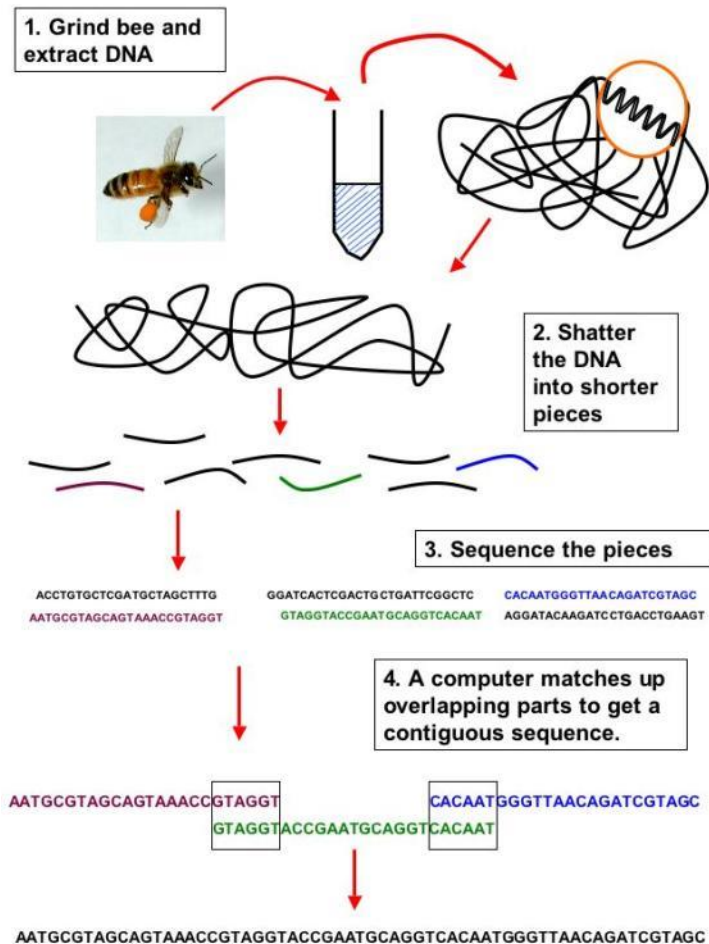
### Genome

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>

- ◆ NCBI 的另一个数据库
- ◆ 测序完成和正在测序物种基因组序列、遗传图、物理图等
- ◆ 序列收集在GenBank
- ◆ 已经完成测序的基因组

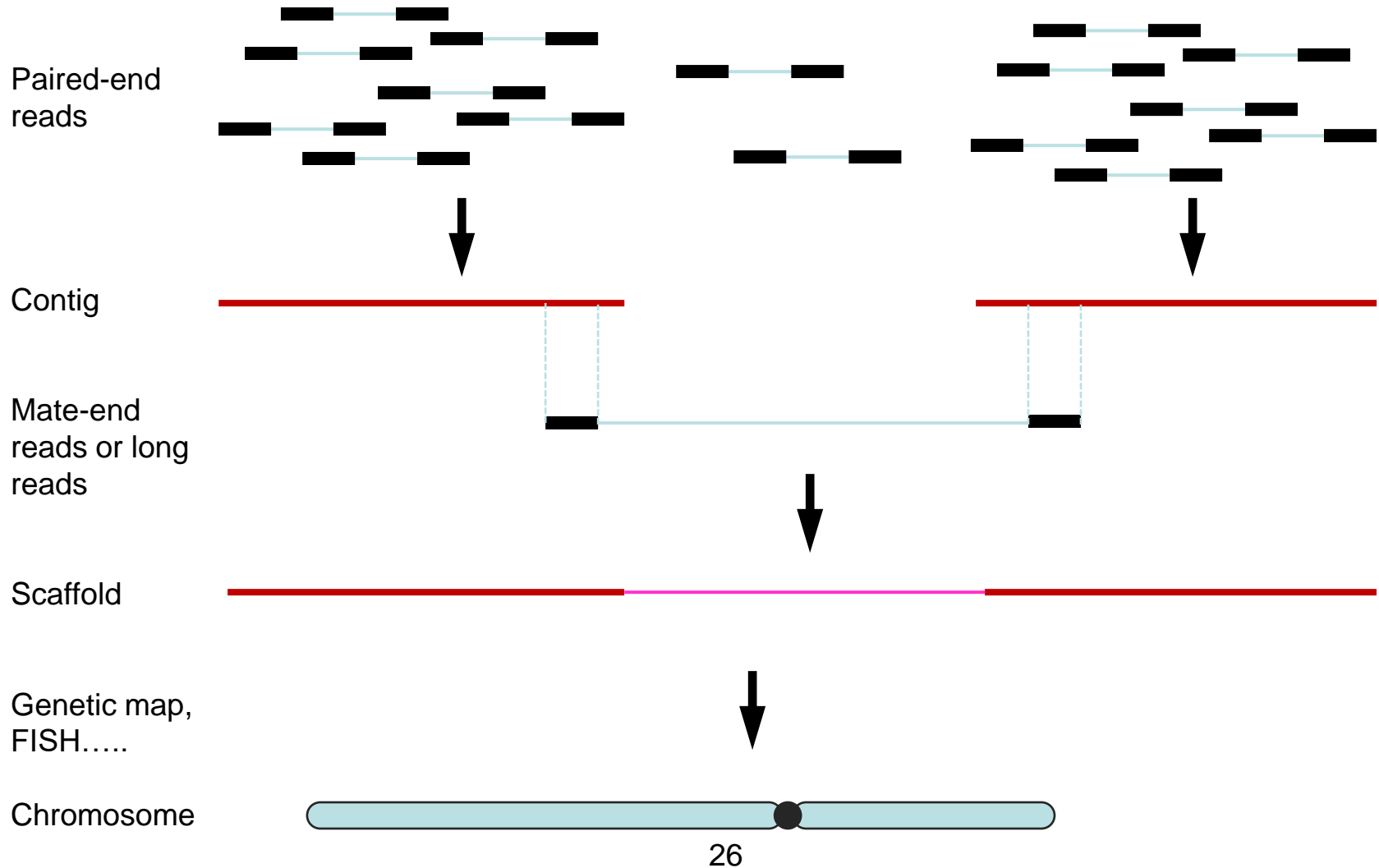


# WGS (whole genome shotgun)



- **Coverage depth** (覆盖深度or测序深度) :  
每个碱基被测序的平均次数，是用来衡量测序数据量的首要参数。  
测序总数据量/基因组大小
- **Coverage ratio** (覆盖率) :  
被测序到的碱基占全基因组大小的比率。  
覆盖比率随覆盖深度升高而提高，亦受测序bias的影响，如illumina测序会受到GC bias的影响，而导致测序不均匀。
- 理论上（完全随机打断）测序深度达到20x即可覆盖整个基因组。实际工作中一般需要50x以上（100 bp读长）。
- Reads长度越长越好。

# *De novo* assembly



## (7) 单核苷酸多态性数据库

dbSNP (Database of Single Nucleotide Polymorphisms)

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp>

◆ NCBI的数据库，创建于1998.9

◆ 约每300 bp 有一个SNP

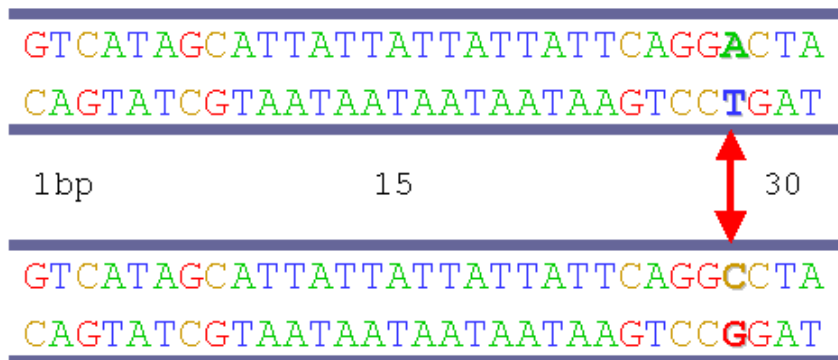
◆ 数据种类

SNP

Insertion/deletion (Indel)

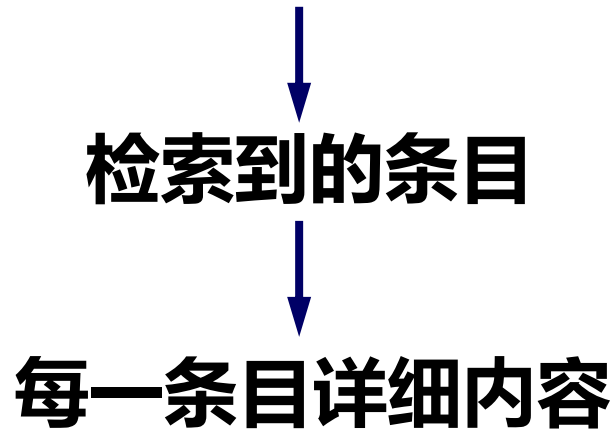
Deletion/insertion/substitution (DIS)

◆ 发现致病基因、进化分析...



# 单核苷酸多态性数据库

## ◆ dbSNP主页输入关键词



## 标准碱基多意代码

代码	碱基
M	A或C
R	A或G
W	A或T
S	C或G
Y	C或T
K	G或T
V	A、C或G
H	A、C或T
D	A、G或T
B	C、G或T
N	G、A、T或C

## **(8) EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Database**

**EBI (European Bioinformatics Institute) 管理  
与GenBank收集的数据相同**

**序列数据文档格式与 GenBank 不同**

**数据库主页<https://www.ebi.ac.uk/>输入关键词**



**检索到的条目**



**每一条目详细内容**

## (9) DDBJ (DNA Data Bank of Japan)

### ◆ 与GenBank收集的序列数据相同

数据库主页 <https://www.ddbj.nig.ac.jp/index-e.html> 输入关键词



检索到的条目



每一条目详细内容

发表文章要提供 Accession number  
(在三大核苷酸数据库中通用)

## **(10) 启动子数据库**

**EPD (Eukaryotic Promoter Database)**

**<https://epd.epfl.ch//index.php>**

- ◆ **由Weizmann Institute of Science in Rehovot (Israel) 开创**
- ◆ **4806条真核生物启动子序列**
- ◆ **人类基因组中的启动子大约19万个**
- ◆ **同一个基因具有多个启动子**

## (11) miRNA数据库

**MicroRNA (miRNA)** 是一类由内源基因编码的长度约为22个核苷酸的非编码单链RNA分子，它们在动植物中参与转录后基因表达调控。



# miRBase

**<http://www.mirbase.org/>**

- ◆ 收集了>15000条 hairpin precursor miRNA 序列
- ◆ 来源于>100个物种
- ◆ 可以通过miRNA名称、关键词、染色体位置等信息检索数据库
- ◆ 分析一条DNA序列中是否可能包含miRNA





## (12) 非编码RNA数据库

<http://bigdata.ibp.ac.cn/npinter4>

➤ 不能翻译成蛋白的功能性RNA分子

➤ Housekeeping non-coding RNA

- tRNAs、 rRNAs、 snRNAs etc.

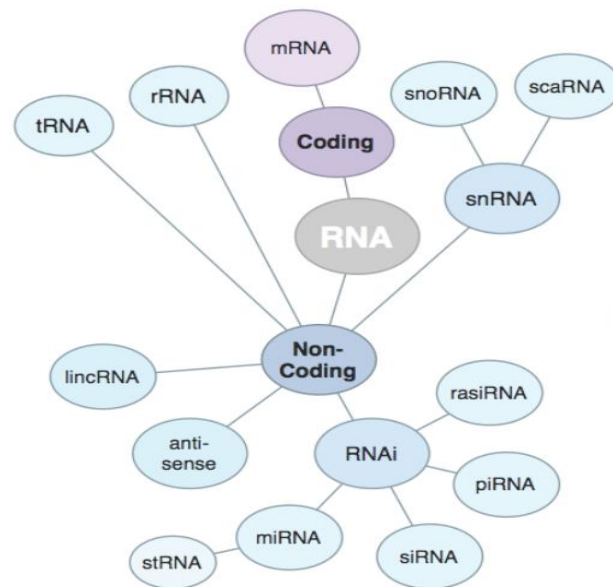
➤ Regulatory non-coding RNA

- small non-coding RNA

✓ siRNA、 miRNA、 piRNA(Piwi-interactiing RNA,

piRNA (Piwi-interactingRNA是从哺乳动物生殖细胞中分离得到的一类长度约为30nt的小RNA, 并且这种小RNA与PIWI蛋白家族成员相结合才能发挥它的调控作用), CircRNA etc.

- Long non-coding RNA (lncRNA, >200 nt )



## 2、蛋白质数据库

以 蛋白质氨基酸顺序及注释信息为基本内容的数据库称为蛋白质序列数据库。

蛋白质序列测定技术的发明早于**DNA**序列测定技术，但是蛋白质序列数据库的建立比核酸序列数据库晚一些。

# 蛋白质序列数据库的发展

- 1984年，“蛋白质信息资源”（**protein information resource PIR**）计划正式启动，由美国国家医学研究基金会管理的蛋白质序列数据库**PIR-PSD**也因此而诞生。
- 1986年，欧洲瑞士日内瓦大学的研究人员设计了一个蛋白质序列分析工具（**COMPSEQ-PC/Gene**）并建立了**swiss-prot**数据库。该数据库中所有序列条目都经过有经验的分子生物学家和蛋白质化学家通过计算机工具并查阅文献资料仔细核实，因此，又称为蛋白质专家库。
- 随着核酸序列的快速增加，由**DNA**序列翻译而来的蛋白质序列也日益增多，相应的**DNA**序列翻译而来的蛋白质序列数据库也开始建立。第一个这样的数据库是**TrEMBL**数据库。（**Translation of EMBL**），它是从**EMBL**中的**cDNA**序列翻译得到的数据库。

## (1) UniProt

<http://www.uniprot.org/>



- ◆ 由PIR、EBI 和SIB创办
- ◆ 分为两个部分：来源于实验的有详细注释的序列（SwissProt）和自动注释序列（TrEMBL）
- ◆ 与100多个数据库相互参照（cross-reference）
- ◆ 可用关键词（Text search）和序列比对（BLAST similarity search）进行检索

## **(2) PIR (Protein Information Resource)**

**<http://pir.georgetown.edu>**

- ◆ **由National Biomedical Research Foundation 创办**
- ◆ **可将蛋白质序列分类**
- ◆ **结构域**

### **(3) PRF (Protein Research Foundation)**

**<http://www.prf.or.jp/>**

- ◆ 由日本的 **Protein Research Foundation** 创办
- ◆ 已发表在杂志上的蛋白质序列
- ◆ 修饰位点、S—S键等
- ◆ 两月更新一次

## **(4) PDBSTR (Re-Organized Protein Data Bank)**

**<http://www.genome.ad.jp>**

◆ 蛋白质序列和二级结构

◆  $\alpha$  螺旋结构

## **(5) Prosite**

**<http://www.expasy.org/prosite>**

- ◆ 蛋白质家族
- ◆ 结构域



- 目前现有的蛋白质序列数据资源

General sequence databases

EXProt

MIPS

NCBI Protein database

PA-GOSUB

PIR-NREF

PIR-PSD

PRF

RefSeq

Swiss-Prot

TCDB

UniParc

UniProt

UniProtKB/TrEMBL

UniRef

### 3、结构数据库

#### (1) PDB (Protein Data Bank)

<http://www.rcsb.org>



◆ 由 Brookhaven National Laboratories 创办

◆ 71,415个结构图

❖ 蛋白质

❖ 核酸

❖ 其它

◆ 可通过 BLAST 系统检索

## (1) PDB (Protein Data Bank)

◆ X 射线衍射图、核磁共振 (NMR) 光谱图和电镜图 (文字和三维结构图)

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	58016	1261	2801	17	62095
NMR	7666	940	168	7	8781
ELECTRON MICROSCOPY	245	22	86	0	353
HYBRID	28	2	1	1	32
other	132	4	5	13	154
Total	66087	2229	3061	38	71415

## (2) SWISS-3D IMAGE

<http://www.expasy.ch/sw3d/>

蛋白质的平面和立体图

◆ 来源于实验结果

◆ 理论模型

**(3) NDB (Nucleic Acid Database)**

**<http://ndbserver.rutgers.edu/>**

核酸的结构

**(4) DNA-Binding Protein Database**

**<http://ndbserver.rutgers.edu/>**

**DNA 结合蛋白质的 X 射线衍射结构图**

## 4、酶和代谢数据库

**KEGG (Kyoto Encyclopedia of Genes and Genomes)**

- ◆ **各种代谢、遗传等路径图**
- ◆ **可检索参与各种路径的基因**

**KEGG主页**<http://www.genome.ad.jp/kegg/>  
**点击 “PATHWAY”**



**“PATHWAY”网页点击任一代谢路径，如糖酵解/糖原异生途径 (Glycolysis/Gluconeogenesis)**

# KEGG数据库

## ◆ 检索Genetic Information Processing

KEGG主页点击 “PATHWAY”



“PATHWAY”网页点击任一遗传信息路径，  
如 Protein export 路径



可以查看参加这一路径蛋白质的信息

# KEGG数据库

## ◆ 检索Environmental Information Processing

KEGG主页点击 “PATHWAY”



“PATHWAY”网页点击任何Environmental Information Processing 路径，如 MAPK signaling pathway 路径



可以查看与这一路径相连的其它信号路径  
或参加这一路径的蛋白质信息

# KEGG数据库

## ◆ 检索Cellular Processes

KEGG主页点击 “PATHWAY”



“PATHWAY”网页点击任何Cellular Processes  
路径，如 Cell cycle 路径



可以查看与这一路径相连的其它信号路径  
或参加这一路径的蛋白质信息



## 5、物种分类数据库

### ◆ 物种分类

界 (Kingdom)

门 (Phylum)

纲 (Class)

目 (Order)

科 (Family)

属 (Genus)

种 (Species)

Mouse: *Mus musculus*

动物界 (Animal)

脊索动物门 (Chordata)

脊椎动物亚门 (Vertebrata)

哺乳纲 (Mammalia)

啮齿目 (Rodentia)

鼠科 (Muridae)

小家鼠属 (Mus)

小家鼠种 (musculus)

每一分类等级下可加设亚级 (Sub-) , 如亚门、亚纲、亚科等。  
每一分类等级上可加设总级 (Super-) , 如总纲、总目、总科等。

# 物种分类数据库

<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>

## ◆ 查找某一物种的系谱树

在NCBI Taxonomy 主页输入物种名称 “pig”



lineage

## 6、文献数据库

### (1) PubMed

<http://www.ncbi.nlm.nih.gov/PubMed/>

- ◆ 美国国家医学图书馆的数据库
- ◆ 医学、分子生物学、基础生物学
- ◆ 5400多种刊物，来源于80多个国家
- ◆ 文献年限：1947年至今
- ◆ 提供摘要，全文链接
- ◆ 免费全文收集在



## (2) 其它类型的文献数据库

### **OMIM (Online Mendelian Inheritance in Man)**

**<http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM>**

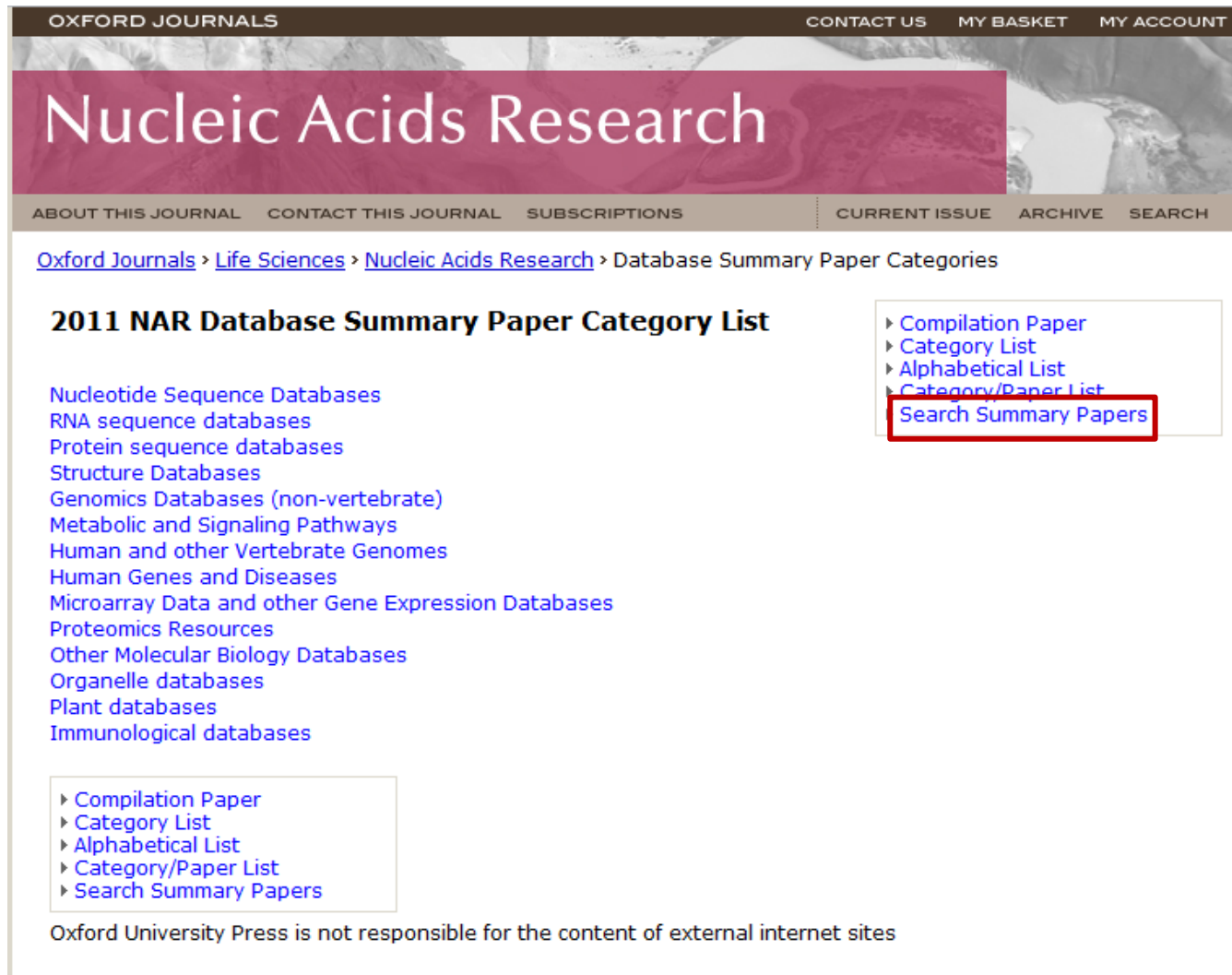
- ◆ **NCBI 的数据库，每天更新数据**
- ◆ **人类基因、遗传疾病**
- ◆ **输入疾病、基因名称**

### **Agricola**

**<http://agricola.nal.usda.gov/>**

- ◆ **美国农业部农业图书馆的数据库**
- ◆ **农业类刊物**

# 7、更多的数据库



The screenshot shows the Oxford Journals website for Nucleic Acids Research. The header includes 'OXFORD JOURNALS' and navigation links: 'CONTACT US', 'MY BASKET', and 'MY ACCOUNT'. The main title 'Nucleic Acids Research' is prominently displayed. Below the title, there are links for 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', 'SUBSCRIPTIONS', 'CURRENT ISSUE', 'ARCHIVE', and 'SEARCH'. The breadcrumb trail indicates the current location: 'Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories'. The main heading is '2011 NAR Database Summary Paper Category List'. A list of database categories is provided, including Nucleotide Sequence Databases, RNA sequence databases, Protein sequence databases, Structure Databases, Genomics Databases (non-vertebrate), Metabolic and Signaling Pathways, Human and other Vertebrate Genomes, Human Genes and Diseases, Microarray Data and other Gene Expression Databases, Proteomics Resources, Other Molecular Biology Databases, Organelle databases, Plant databases, and Immunological databases. Two sidebars offer navigation options: 'Compilation Paper', 'Category List', 'Alphabetical List', 'Category/Paper List', and 'Search Summary Papers'. The 'Search Summary Papers' option in both sidebars is highlighted with a red box. A disclaimer at the bottom states: 'Oxford University Press is not responsible for the content of external internet sites'.

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

## Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories

### 2011 NAR Database Summary Paper Category List

Nucleotide Sequence Databases  
RNA sequence databases  
Protein sequence databases  
Structure Databases  
Genomics Databases (non-vertebrate)  
Metabolic and Signaling Pathways  
Human and other Vertebrate Genomes  
Human Genes and Diseases  
Microarray Data and other Gene Expression Databases  
Proteomics Resources  
Other Molecular Biology Databases  
Organelle databases  
Plant databases  
Immunological databases

- ▶ Compilation Paper
- ▶ Category List
- ▶ Alphabetical List
- ▶ Category/Paper List
- ▶ Search Summary Papers

Oxford University Press is not responsible for the content of external internet sites

<http://www.oxfordjournals.org/nar/database/c/>

## 8、向数据库提交和修改核苷酸和蛋白质序列

**提交: Submission**

**修改: Update**

**数据库中的数据由大家  
无偿提供, 共同享用**

# **(1) 向 GenBank提交或修改核苷酸序列**

## **◆ 在GenBank主页用 BankIt 功能提交序列**

- ❖ 网上直接提交，简单方便**
- ❖ 提交后立刻得到临时编号**
- ❖ 二天内得到 Accession number**

## **◆ 用 Sequin 方法提交序列**

- ❖ 可下载的电子表格**
- ❖ 自动确定 CDS、ORF 和查找重复序列**

## **◆ 用Update 功能修改 GenBank 中的序列和相关信息**

- ❖ 修改一次，version 的编号就进一位**
- ❖ Accession number不变**

## (2) 向 UniPROT 提交或修改蛋白质序列



- ◆ 网上直接操作
- ◆ 只接收用蛋白质直接测序的序列
- ◆ 由核苷酸序列翻译得到的蛋白质序列  
将进入TrEMBL



## 9、常用序列格式

- Fasta

> *Xa26, mRNA*

```
ATGGCCATGGGGTCCACACGCAGTGAGATG
AATGCTAGATCTCACGAGAAAAAAGAAAT
ACATCTCAGGGGGTTGTGATGTACTGGATAA
TTTGCTCGTCATATTAACCATTAGCTTACTC
TAGTTGATGTGGGCGATGGATGGAGCCGGC
AGCCGGGCGATCCTATTAA
```

# 上机操作



1. **熟悉各种数据库**
2. **重点了解 GenBank 和 UniPROT的各种功能和适用范围**

## ***Xa26* nucleic acid sequence (DQ426646,6000 bp):**

### ***> Xa26, mRNA***

ATGGCCATGGGGTCCACACGCAGTGAGATGAATGCTA  
GATCTCACGAGAAAAAAGAAATACATCTCA  
GGGGTTGTGATGTACTGGATAATTTGCTCGTCATATT  
AACCATTAGCTTACTCTAGTTGATGTGGGCATG  
GATGGAGCCGGCAGCCGGCGATCCTATTAA ...

## ***Xa26* amino acid sequence (ABD84047,1103 aa):**

### ***> Xa26, protein***

MALVRLPVWIFVAALLIASSSTVPCASSLGPIASKSNSS  
DTDLAALLAFKAQLSDPNNILAGNWTGTPF  
CRWVGVSCESSHRRRRQRVTALELPNVPLQGELSS...

