

生物医学工程导论

第十八章 生物信息学导论

背景

- 人类基因组计划(Human Genome Project, HGP): 1990年正式启动, 旨在完成人类基因组约30亿个碱基的全序列测定。
- 海量生物数据的迅速膨胀: DNA、RNA和蛋白质序列, 蛋白质二级结构和三维结构数据, 蛋白质相互作用数据等。
- 对大量生物数据的管理、分析和信息化需求促进了生物信息学的迅速发展。

人类基因组计划HGP (Human Genome Project)

- 由美国**NIH**和能源部提出和带头，美、英、德、法、日、中共同参与的国际合作项目。
- 重大国际研究项目：测定人类基因组全部**DNA**序列，构建人类基因组遗传图谱和物理图谱。
- **1990年**：正式启动，**30亿美元**。
- **2001年**：人类基因组草图公开发表。
- **2003年**：美国宣布该项目完成。

人类基因组计划

- 基因组图谱：遗传图谱，物理图谱
- 遗传图谱(**genetic map**): 连锁图谱，显示所知的基因和/或遗传标记的相对距离位置与次序。
- 物理图谱(**physical map**):表示某些基因和/或遗传标记之间在基因组上的精确位置和距离（如间隔的bp数目）的图谱。

生物信息学定义的历史演变

- 定义一：生物信息学是一门收集、分析遗传数据以及分发给研究机构的新学科。(Dr. Hwa A. Lim, 1987)
- 定义二：生物信息学特指数据库类的工作，包括持久稳固的在一个稳定的地方提供对数据的支持。(Dr. Hwa A. Lim, 1994)
- 定义三：生物信息学是在大分子方面的概念型的生物学，并且使用了信息学的技术，这包括了从应用数学、计算机科学以及统计学等学科衍生而来各种方法，并以此在大尺度上来理解和组织与生物大分子相关的信息。(Luscombe,2001)

什么是生物信息学？

生物学研究可以被看成是研究信息的传递：从DNA经转录翻译到蛋白质，从细胞质中到细胞核内，从母细胞到子细胞，从一个细胞或一个组织到另一个细胞或另一个组织，从一代到下一代，从一个物种到另一个物种的进化演变。这种信息论的观点即可称为生物信息学。

(Bioinformatic challenges for the next decade(s),
David Eisenberg et al., 2006)

生物信息学的主要研究内容

- 生物信息的存储与查询；
- 序列比对；
- 基因预测及基因组分析；
- 分子进化与系统发育分析；
- **RNA**结构预测；
- 蛋白质结构预测；
- 分子设计与药物设计；
- 生物网络；
- 生物芯片.

生物信息学的发展历程

- 1952年，**Sanger**根据胰岛素蛋白质的测序结果，推断蛋白质是排列完美的分子。——最早的信息论观点。
- 之前的观点认为蛋白质可能是由相似氨基酸堆积在一起的混合物，而不具有特定的结构。而**Sanger**发现胰岛素蛋白质作为一种纯净物（单一的分子），并且具有特定的三级结构。因此推断蛋白质是排列完美的分子，而这种排列的完美性，其中应当蕴含着一些未知的机理。
- 1955年，**Sanger**与合作者分别对牛、猪和羊的胰岛素蛋白质进行了测序并做了序列上的比较。——最早的序列比对。

最早的序列分析：胰岛素蛋白质

Insulin Chain A: 8-10位存在着不同

(牛, ASV; 猪, TSI; 羊, AGV)

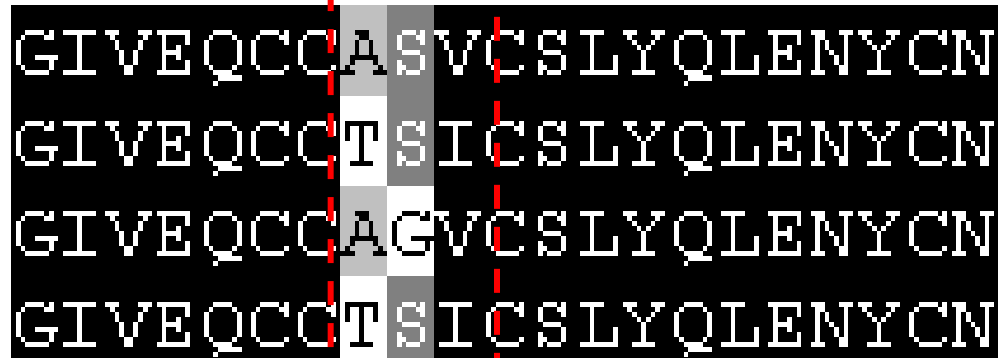
(Brown *et al.*, 1955)

Bovine

Pig

Sheep

Human



G	I	V	E	Q	C	C	A	S	V	C	S	L	Y	Q	L	E	N	Y	C	N
G	I	V	E	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N
G	I	V	E	Q	C	C	A	G	V	C	S	L	Y	Q	L	E	N	Y	C	N
G	I	V	E	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N

Made by GeneDoc

80年代：DNA序列数据库

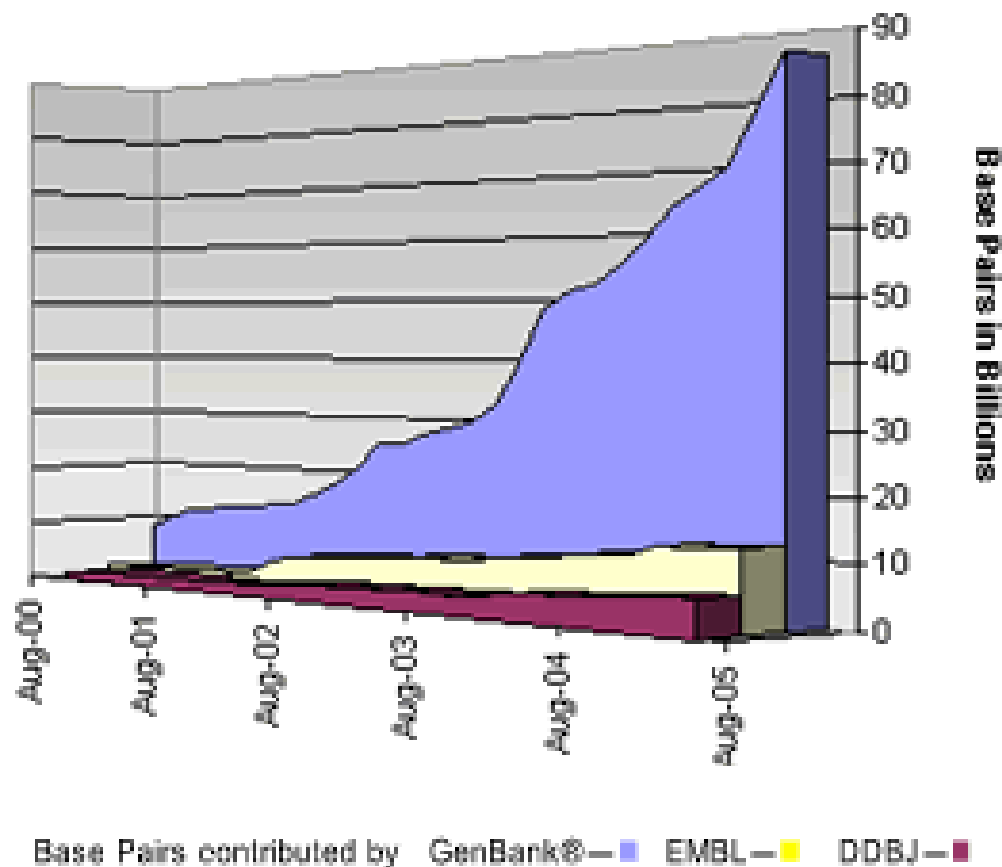
- 1974年，George I.Bell等人收集DNA序列，构建GenBank数据库。1982~1992开发第一个版本。
- 1980年，EMBL数据库成立。
- 1984年，日本DDBJ数据库成立。
- 核酸序列数据的去冗余：Refseq数据库，对于相同的序列只列一条目录。

序列数据的存储

- 核酸序列数据库
 - 国际三大核酸序列数据库：
GenBank, EBML, DDBJ
 - RefSeq: The Reference Sequence Database
 - dbEST: Expressed Sequences Tags数据库
 - UniGene等

核酸数据库数据的增长

**Growth of the
International Nucleotide Sequence Database Collaboration**



较早的基因组数据库- GDB

- 为人类基因组计划(HGP)保存和处理基因组图谱数据。
- **GDB**的目标是构建关于人类基因组的百科全书，除了构建基因组图谱之外，还开发了描述序列水平的基因组内容的方法，包括序列变异和其它对功能和表型的描述。

GenBank

- 由美国国立卫生研究院NIH下属国立生物技术信息中心NCBI建立。
- 汇集并注释了所有公开的核酸以及蛋白质序列。每个记录代表了一个单独的、连续的、带有注释的DNA或RNA片段。

GenBank中测序最多的20个物种

Entries	Bases	Species
11148092	12700084970	Homo sapiens (人)
7200432	8291244632	Mus musculus (小鼠)
1288005	5766221181	Rattus norvegicus (大鼠)
2026919	3808273911	Bos taurus (牛)
2841072	3564487204	Zea mays (玉米)
1559584	2764828000	Danio rerio (斑马鱼)
2058320	1863733664	Sus scrofa (猪)
1179148	1517127691	Oryza sativa (水稻)
227831	1352463179	Strongylocentrotus purpuratus (海胆)
1417622	1135336003	Xenopus tropicalis (爪蟾)
212172	943046501	Pan troglodytes (黑猩猩)
734569	897464279	Drosophila melanogaster (果蝇)
1949707	879897433	Arabidopsis thaliana (拟南芥)
802461	857512666	Gallus gallus (鸡)
497579	810324848	Vitis vinifera (葡萄)
75850	708598911	Macaca mulatta (恒河猴)
1220300	694494794	Canis lupus familiaris (狗)
1006209	657603350	Sorghum bicolor (高粱)
1102504	655077659	Triticum aestivum (小麦)
409757	520072874	Medicago truncatula (蒺藜状苜蓿)

EMBL核酸序列数据库

- EMBL-EBI (European Bioinformatics Institute)维护;
- <http://www.ebi.ac.uk/embl/>

EMBL-EBI

EB-eye Search

All Databases

Enter Text Here

Go

Reset ?

Advanced Search

Give us feedback

Databases

Tools

EBI Groups

Training

Industry

About Us

Help

Site Index

■ EMBL-Bank Home

■ Access

■ Documentation

■ News

■ Submission

■ Publications

■ People

■ Contact

EMBL Fetch

Fetch an EMBL record by id

Go

Hands-on Training

30th April - 1st May 2009:
Short Read Bioinformatics
hands-on EBI training
course...[more](#)

Collaborations

■ INSDC - International

EBI > Databases > EMBL-Bank

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are [direct submissions](#) from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international [collaboration](#) with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The [current database release](#) (Release 101, Sept 2009), with according [Release notes](#) and [user manual](#) are available from the EBI servers. A sample database entry is shown [here](#).

A publication in [Nucleic Acids Research 2009 37: D19-D25](#), provides further information and details.

The EMBL nucleotide sequence database forms part of the [European Nucleotide Archive](#), an EBI project led by [Guy Cochrane](#) as part of the [The Protein and Nucleotide Database Group \(PANDA\)](#) under [Ewan Birney](#).



Link	Explanation
Access	Database queries , Completed genomes webserver , FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography .
Submission	Primary sequence submissions, third party annotation, updates.
Documentation	Release notes user manual , Information for Submitters , FAQ , Release information , Forthcoming Changes .

DDBJ

NIG (National Institute of Genetics)

CIB (Center for Information Biology)

<http://www.ddbj.nig.ac.jp/index-e.html>

The screenshot shows the DDBJ website with a blue header and a navigation menu. The main content area features a large image of Mount Fuji and a description of the database. A sidebar on the left contains links for About DDBJ, How to Use, Q and A, Sequence Submission, Search, Phylogenetics, and Genome Analysis. The main content area includes sections for Hot Topics and Maintenance, each with a list of recent updates and links to more information.

DDBJ
DNA Data Bank of Japan

Accession [DNA](#) [Protein](#) [AlIBDs](#) [Taxonomy](#) [Site Search](#)

Accession numbers

☒ DDBJ ☐ UniProt ☐ PDB ☐ OAD ☐ PRF ☐ Patent [>>more](#)

HOME [Submission](#) [How to Use](#) [Search/Analysis](#) [FTP/WebAPI](#) [Report/Statistics](#) [Contact Us](#) [Japanese](#)

About DDBJ
How to Use
Q and A

Sequence Submission
[SAKURA](#)
[Mass Submission](#)
[Data Updates](#)

Search
[getentry](#)
[ARSA](#)
[TXSearch](#)
[BLAST](#)
[PSI-BLAST](#)
[FASTA](#)
[SSEARCH](#)

Phylogenetics
[ClustalW](#)

Genome Analysis
[GIB](#)

DDBJ : DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, through close collaboration with EBI in Europe and NCBI in USA.

Hot Topics [More](#)

- Aug. 28, 2009 [Change of directory structure of anonymous FTP site: Made a new directory "tpa"](#)
- Aug. 13, 2009 [A new high speed BLAST API was opened to the public from WABI](#)
- Aug. 11, 2009 [Redistribution of genomic sequence \(build 4\) of the cultivar Nipponbare of Japanese rice \(Oryza sativa Japonica Group\) assigned with RAP annotation](#)

Maintenance [More](#)

- Sep. 11, 2009 [Suspension of some parts of DDBJ activities during national holidays \(Sep.19-23\)](#)
- Aug. 04, 2009 [Request for re-submission of the queries \(Re: Homology Search and ClustalW temporary down \(Aug.3\)\)](#)
- Jun. 23, 2009 [Completion of a bug fixation: error in displaying suppressed entries by getentry](#)

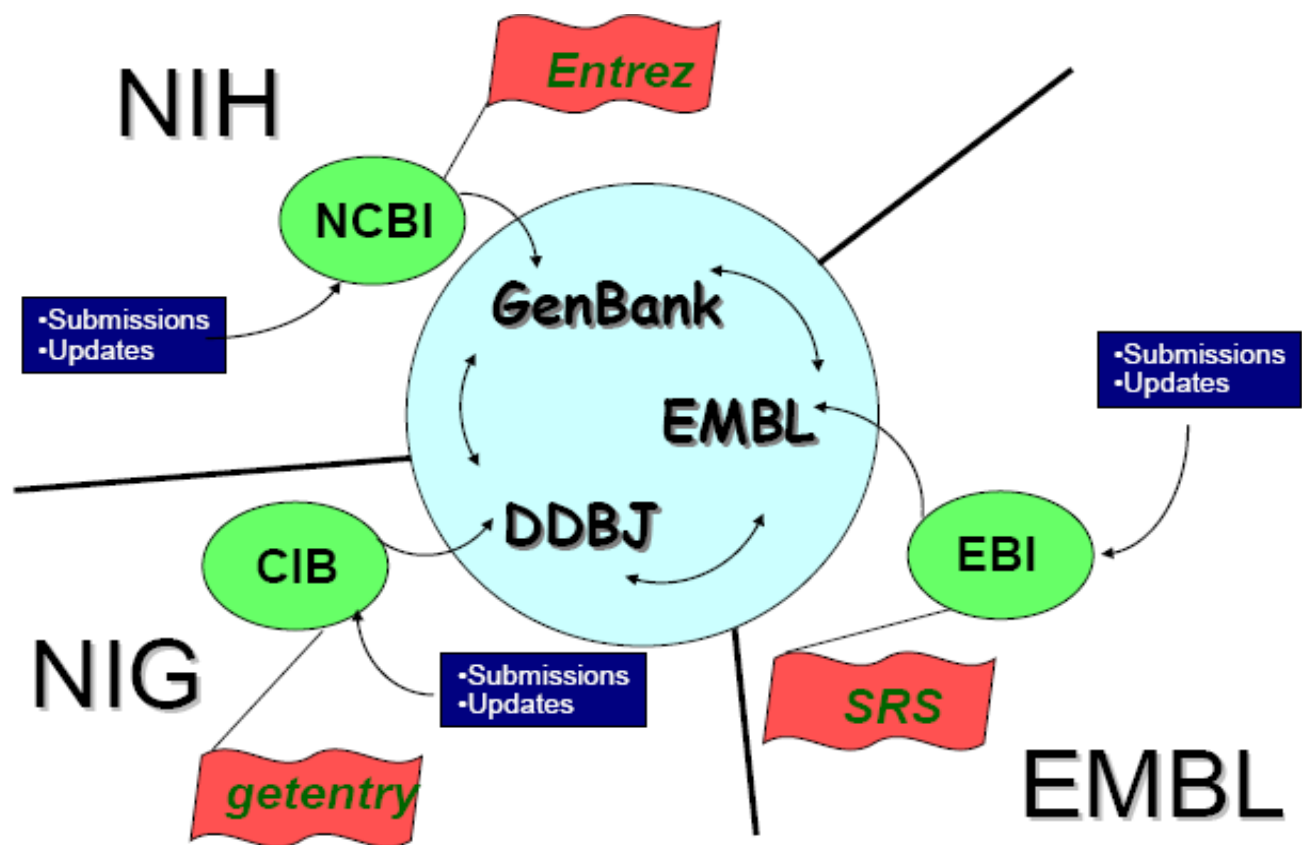
Sequence Data Submission

- Submit my sequences**
Orientation for the data submission

FTP/Web API

- FTP (ftp.ddbj.nig.ac.jp)**
Download data files

三大数据库之间的联系



RefSeq数据库

- 提供非冗余的，高质量的，经检验校正的序列信息；
- 包括染色体、基因组（细胞器、病毒、质粒）、蛋白质、**RNA**等；
- 序列文件的标识符：
mRNA序列：NM_123456
非编码RNA：NR_123456
蛋白质序列：NP_123456

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

dbEST: 表达序列标签数据库

最多的20个物种:

Homo sapiens (human)	8,134,045
Mus musculus + domesticus (mouse)	4,850,243
Bos taurus (cattle)	1,497,461
Sus scrofa (pig)	1,470,315
Danio rerio (zebrafish)	1,358,222
Arabidopsis thaliana (thale cress)	1,276,692
Xenopus tropicalis	1,271,375
Oryza sativa (rice)	1,211,418
Zea mays (maize)	1,159,264
Triticum aestivum (wheat)	1,050,926
Rattus norvegicus + sp. (rat)	889,896
Ciona intestinalis	686,396
Xenopus laevis (African clawed frog)	677,784
Gallus gallus (chicken)	599,330
Brassica napus (oilseed rape)	567,177
Drosophila melanogaster (fruit fly)	542,180
Hordeum vulgare + subsp. vulgare (barley)	437,713
Salmo salar (Atlantic salmon)	432,630
Glycine max (soybean)	392,321
Canis familiaris (dog)	365,909

<http://www.ncbi.nlm.nih.gov/dbEST/> 2007.08, 总序列45,660,524条

UniGene:

An Organized View of the Transcriptome

为每一个基因创造一个唯一的条目，收集这个基因所有的**ESTs**

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus (cow)	43,490
Canis lupus familiaris (dog)	24,459
Equus caballus (horse)	5,992
Homo sapiens (human)	123,173
Macaca fascicularis (crab-eating macaque)	12,659
Macaca mulatta (rhesus monkey)	6,489
Monodelphis domestica (gray short-tailed opossum)	693
Mus musculus (mouse)	79,222
Ornithorhynchus anatinus (platypus)	749
Oryctolagus cuniculus (rabbit)	6,796
Ovis aries (sheep)	19,084
Papio anubis (olive baboon)	11,904
Peromyscus maniculatus (deer mouse)	10,429
Pongo abelii (Sumatran orangutan)	7,286

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

DNA测序

- DNA一次连续测序的长度约为500bp;
- EST (Expressed sequence tag) 测序: 细胞中mRNA反转录成cDNA, 方向不定测序;
- GSS (Genome Survey Sequences, 基因组勘测序列): 类似于ESTs, 来源基因组;
- HTG (High-throughput genome sequences, 高通量基因组序列): 高通量、尚未完工的DNA序列;

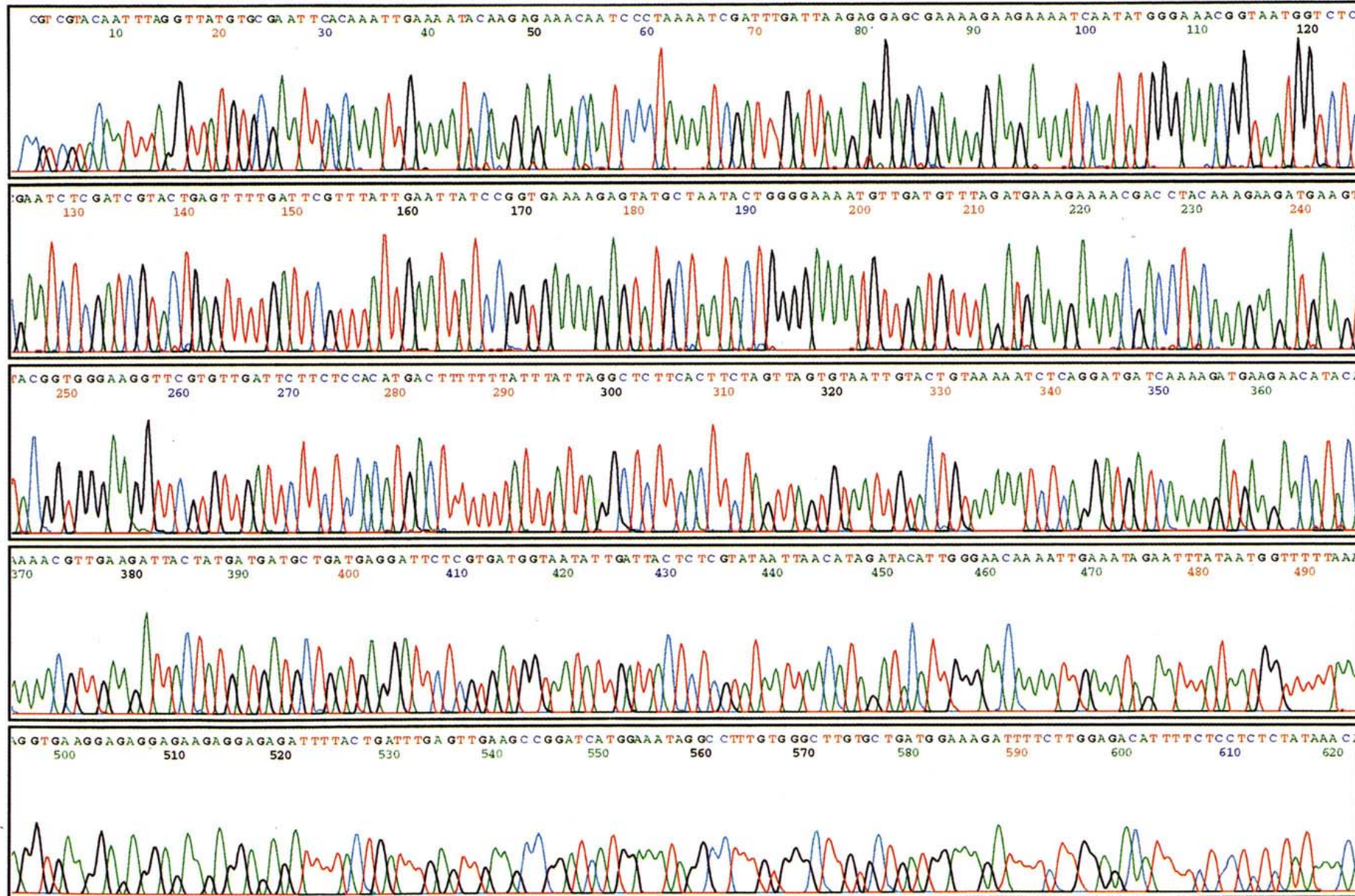
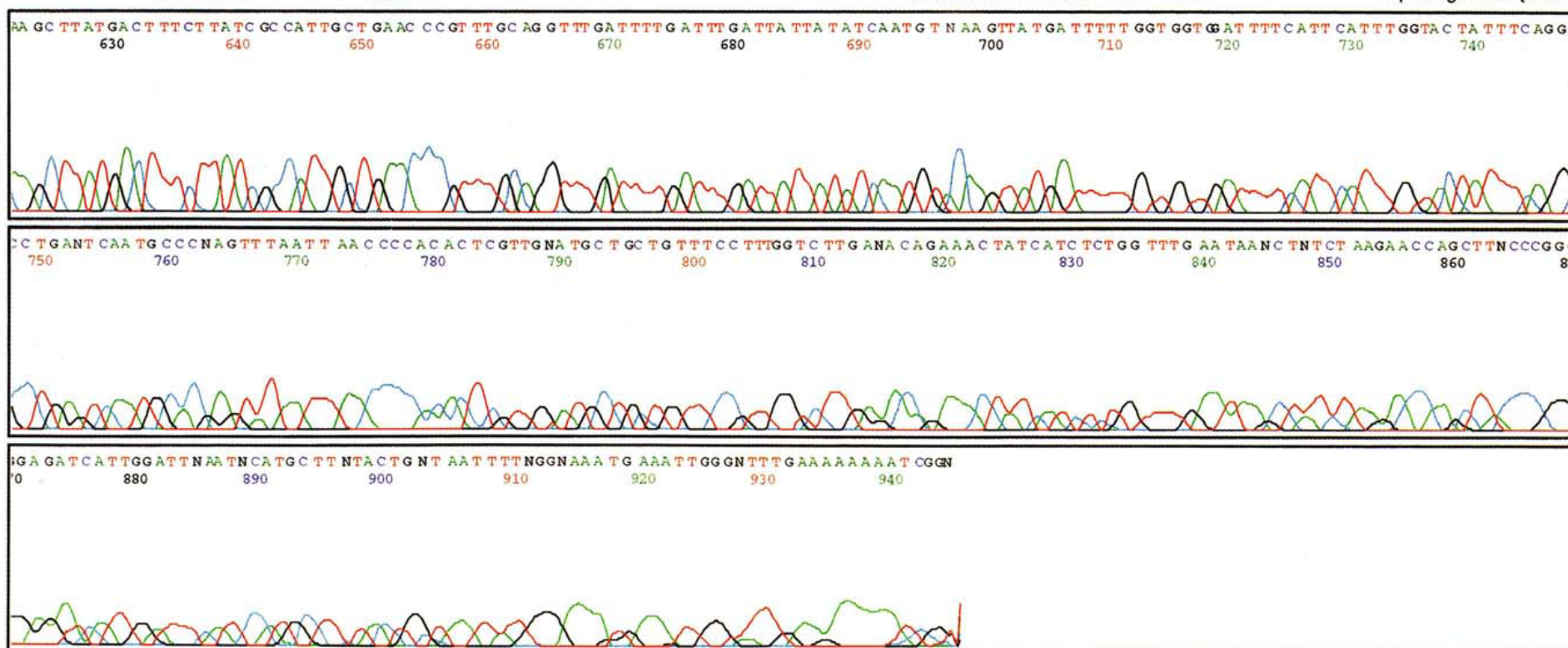


FIGURE 2.2. See legend on p. 37.



1	CGTCGTACAA	TTTAGGTTAT	GTGCGAATTC	ACAAATTGAA	AATACAAGAG	AAACAATCCC	TAAAATCGAT	TTGATTAAGA	GGAGCGAAAA	90
91	GAAGAAAATC	AATATGGGAA	ACGGTAATGG	TCTCGAATCT	CGATCGTACT	GAGTTTGTAT	TCGTTTATTG	AATTATCCGG	TGAAAAGAGT	180
181	ATGCTAATAC	TGGGGAAAAT	GTTGATGTTT	AGATGAAAGA	AAACGACCTA	CAAAGAAGAT	GAAGTACGGT	GGGAAGGTTT	GTGTTGATTG	270
271	TTCTCCACAT	GACTTTTTTT	ATTTATTAGG	CTCTTCACTT	CTAGTTAGTG	TAATTGTACT	GTAAAAATCT	CAGGATGATC	AAAAGATGAA	360
361	GAACATACAA	AACGTTGAAG	ATTACTATGA	TGATGCTGAT	GAGGATTCTC	GTGATGGTAA	TATTGATTAC	TCTCGTATAA	TTAACATAGA	450
451	TACATTGGGA	ACAAAATTGA	AATAGAATTT	ATAATGGTTT	TTAAAGGTGA	AGGAGAGGAG	AAGAGGAGAG	ATTTTACTGA	TTTGAGTTGA	540
541	AGCCGGATCA	TGGAAATAGG	CCTTTGTGGG	CTTGTGCTGA	TGGAAAGATT	TTCTTGAGAG	CATTTTCTCC	TCTCTATAAA	CAAGCTTATG	630
631	ACTTTCTTAT	CGCCATTGCT	GAACCCGTTT	GCAGGTTTGA	TTTTGATTGT	ATTATATATAT	CAATGTNAAG	TTATGATTTT	TGGTGGTGGG	720
721	TTTTTCATTCA	TTTGGTACTA	TTTCAGGCCT	GANTCAATGC	CCNAGTTTAA	TTAACCCAC	ACTCGTTGNA	TGCTGCTGTT	TCCTTTGGTC	810
811	TTGANACAGA	AACTATCATC	TCTGGTTTGA	ATAANCTNTC	TAAGAACCAG	CTTNNCCCGG	GAGATCATTG	GATTNAATNC	ATGCTTNTAC	900
901	TGNTAATTTT	NGNNAATGA	AATTGGGNTT	TGAAAAAAA	TCGGN					990

基因组测序：两种方案策略

- 基因图谱法：DNA片段在染色体上的位置、方向已知。首先染色体被打断成150~200kbp左右的大片段，然后克隆到BACs (Bacterial Artificial Chromosome)中，再进一步随机打断，克隆，测序，依靠计算机组装成长的序列(contig)。
- “鸟枪法” (shotgun)：DNA片段在染色体上的位置和方向未知。全基因组随机打断成小片段，克隆，双向测序，计算机组装成长的序列。

序列数据的存储

- 蛋白质序列数据库
 - UniProt
 - Swiss-prot & TrEMBL, PIR
- 基因组数据库：Ensembl

UniProt

- **Universal Protein Resource:**
Swiss-prot(TrEMBL), PIR两大蛋白数据库的整合体;
- 收录蛋白质序列目录最广泛、功能注释最全面的数据库;
- 包含三个子库:
 - UniProtKB (UniProt Knowledgebase)
 - UniRef (UniProt Reference Clusters)
 - UniParc (Uniprot Archive)
- <http://www.uniprot.org>

Search in

Query

Protein Knowledgebase (UniProtKB) ▾

I

Search

Clear

Fields »

Search

Blast

Align

Retrieve

ID Mapping

WELCOME

The mission of [UniProt](#) is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ★ TrEMBL, which is automatically annotated and is not reviewed.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations , taxonomy , keywords and more .



NEWS



UniProt release 15.7 - Sep 1, 2009

Formyl peptide receptors: the missing link between olfaction and immune system • [Cross-references to STRING](#)

- › [Statistics for UniProtKB:](#)
[Swiss-Prot](#) • [TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

SITE TOUR



Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

paint my thoughts

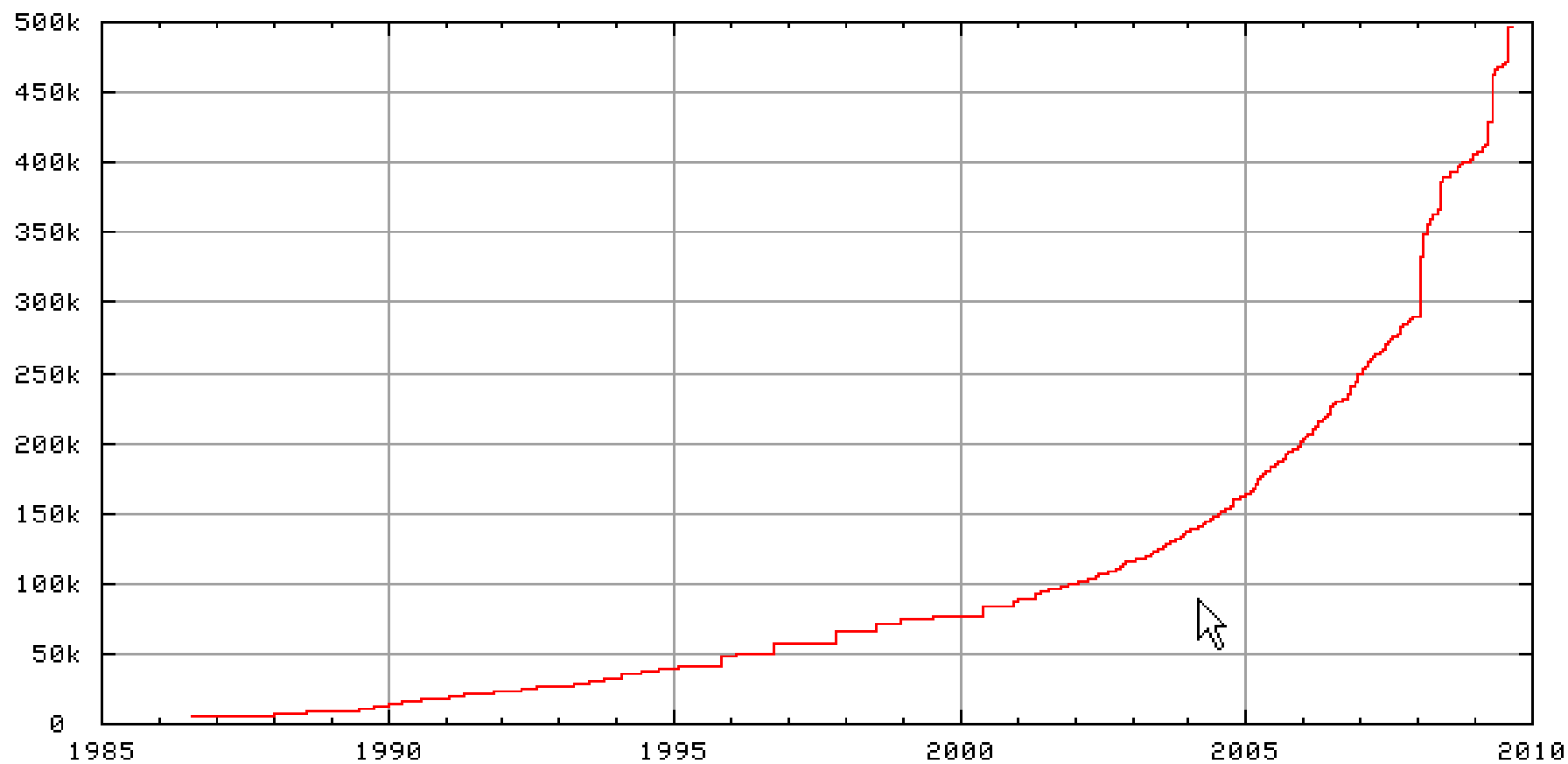
August 2009

Drawing is probably not a talent the layman would normally associate with Science...

UniProtKB

- UniProt Knowledgebase:
Release 15.4 , 16-Jun-2009 , 包括:
 - Swiss-Prot Release 57.7 : 497293 entries
 - TrEMBL Release 40.4 : 9145906 entries
- 包含蛋白质序列全面的信息, 提供准确、丰富的序列与功能注释。
- 记录以6位字母和数字组成, 例: [Q5K8D3](#)

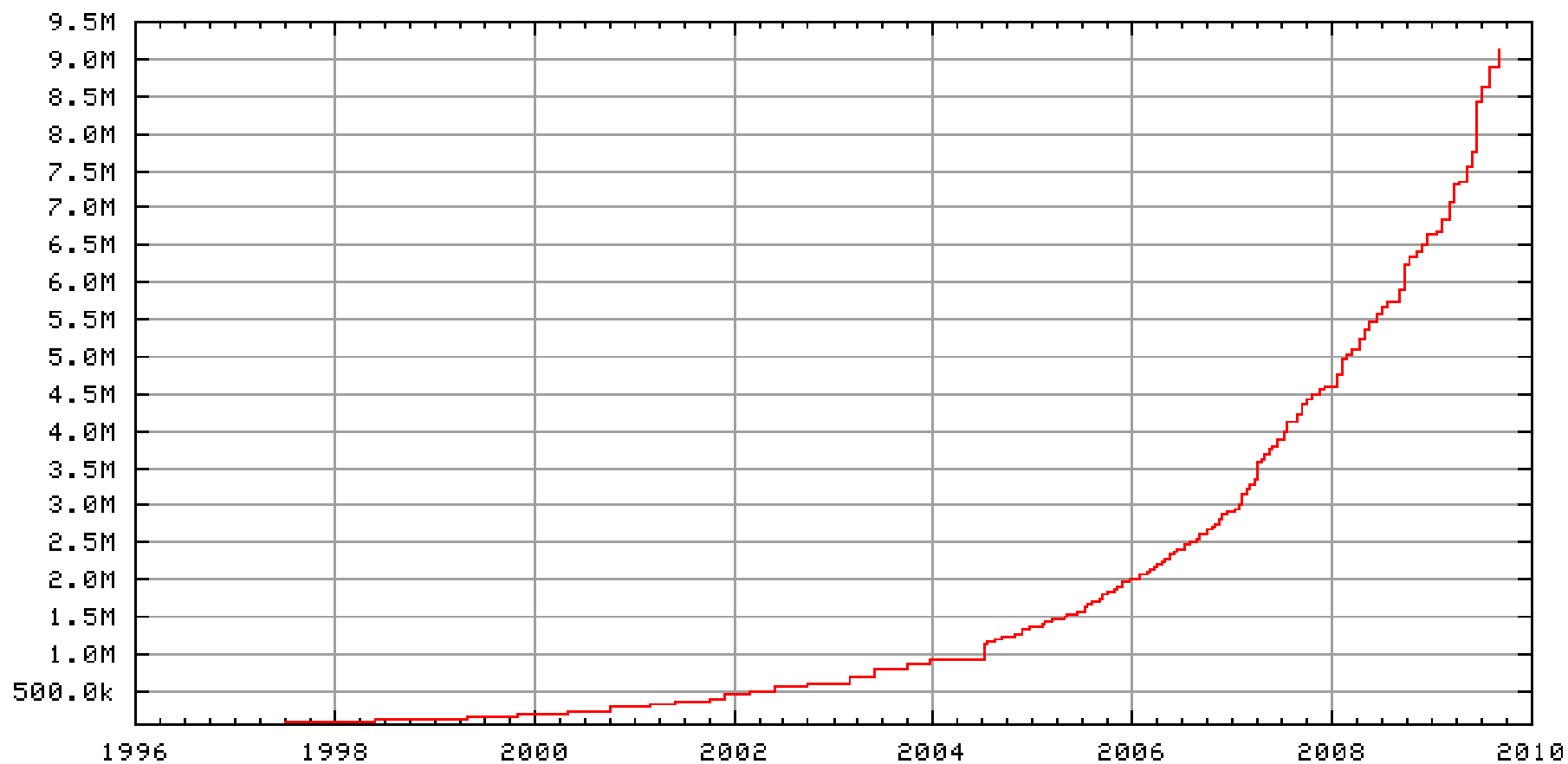
Number of entries in UniProtKB/Swiss-Prot



Swiss-Prot Release 57.7

-----	-----	-----
Number	Frequency	Species
-----	-----	-----
1	20332	Homo sapiens (Human)
2	16177	Mus musculus (Mouse)
3	8601	Arabidopsis thaliana (Mouse-ear cress)
4	7419	Rattus norvegicus (Rat)
5	6552	Saccharomyces cerevisiae (Baker's yeast)
6	5711	Bos taurus (Bovine)
7	4957	Schizosaccharomyces pombe (Fission yeast)
8	4358	Escherichia coli (strain K12)
9	3965	Bacillus subtilis
10	3829	Dictyostelium discoideum (Slime mold)
11	3250	Caenorhabditis elegans
12	3118	Xenopus laevis (African clawed frog)
13	3012	Drosophila melanogaster (Fruit fly)
14	2545	Danio rerio (Zebrafish) (Brachydanio rerio)
15	2263	Oryza sativa subsp. japonica (Rice)
16	2215	Pongo abelii (Sumatran orangutan)
17	2138	Gallus gallus (Chicken)
18	1988	Escherichia coli O157:H7

Number of entries in UniProtKB/TrEMBL



TrEMBL Release 40.4

-----	-----	-----
Number	Frequency	Species
-----	-----	-----
1	189924	Human immunodeficiency virus 1
2	95761	Oryza sativa subsp. japonica (Rice)
3	54942	Homo sapiens (Human)
4	50189	Trichomonas vaginalis G3
5	49246	Mus musculus (Mouse)
6	43743	Arabidopsis thaliana (Mouse-ear cress)
7	39845	Paramecium tetraurelia
8	39335	Oryza sativa subsp. indica (Rice)
9	38075	Hepatitis C virus
10	28042	Tetraodon nigroviridis (Green puffer)
11	26735	Drosophila melanogaster (Fruit fly)
12	24813	Vitis vinifera (Grape)
13	22276	Medicago truncatula (Barrel medic)
14	20542	Danio rerio (Zebrafish) (Brachydanio rerio)
15	20440	Trypanosoma cruzi
16	20409	Caenorhabditis elegans
17	19235	uncultured bacterium
18	16851	Aedes aegypti (Yellowfever mosquito)
19	16685	Tetrahymena thermophila SB210
20	16420	Phaeosphaeria nodorum (Septoria nodorum)

Swiss-Prot & TrEMBL

- 最早广泛使用的蛋白数据库；欧洲最主要的蛋白序列数据库；
- <http://www.expasy.ch/sprot/>
- SIB（Swiss Institute of Bioinformatics）
- 可由ExPASy（Expert Protein Analysis System）系统访问；
- 所有序列条目均经过有经验的分子生物学家和蛋白质化学家审核，因此又称为蛋白质专家库。

PIR

- 1984年，美国国家医学研究基金会（NREF）正式启动蛋白质信息资源（Protein Information Resource, PIR）计划；
- 美国最主要的蛋白序列数据库；
- 非冗余、高质量注释、全面分类；
- PIR数据库按照数据的性质和注释层次分为PIR1、PIR2、PIR3和PIR4。PIR1中的序列已经验证，注释最为详尽。
- <http://pir.georgetown.edu/>

基因组数据库-Ensembl

EMBL-EBI和Sanger研究所共同开发。

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

**Home**



[/](#) | [BLAST/BLAT](#) | [BioMart](#) | [Docs & FAQs](#) 

Search for:

e.g. gene **BRCA2** or **AL032821.2.1.143563** or **muscular dystrophy**

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

 **Human**
GRCh37

 **Mouse**
NCBIM37

New to Ensembl?

Did you know you can:

-  [Learn how to use Ensembl](#)
with our video tutorials and walk-throughs
-  [Add custom tracks](#)
using our new Control Panel
-  [Upload your own data](#)
and save it to your Ensembl account
-  [Search for a DNA or protein sequence](#)
using BLAST or BLAT
-  [Fetch only the data you want](#)
from our public database, using the Ensembl Perl API
-  [Download our databases via FTP](#)
in FASTA, MySQL and other formats

获取序列及检索公共数据库

- **NCBI: Entrez**的开发, **D.Lipman**等人。
- 提供关键字的搜索的方法。
- “硬搜索”:包含关键字的, 完全匹配的结果。
- “软搜索”:与查询内容相关的信息。
- 查询内容: 基因/蛋白质的名称、标识符, 文献、蛋白质结构, 等等。



Search across databases

GO

Clear

Help

Welcome to the Entrez cross-database search page



PubMed: biomedical literature citations and abstracts



PubMed Central: free, full text journal articles



Site Search: NCBI web and FTP sites



Books: online books



OMIM: online Mendelian Inheritance in Man



OMIA: online Mendelian Inheritance in Animals



CoreNucleotide: Core subset of nucleotide sequence records



EST: Expressed Sequence Tag records



GSS: Genome Survey Sequence records



Protein: sequence database



Genome: whole genome sequences



Structure: three-dimensional macromolecular structures



Taxonomy: organisms in GenBank



SNP: single nucleotide polymorphism



dbGaP: genotype and phenotype



UniGene: gene-oriented clusters of transcript sequences



CDD: conserved protein domain database



3D Domains: domains from Entrez Structure



UniSTS: markers and mapping data



PopSet: population study data sets



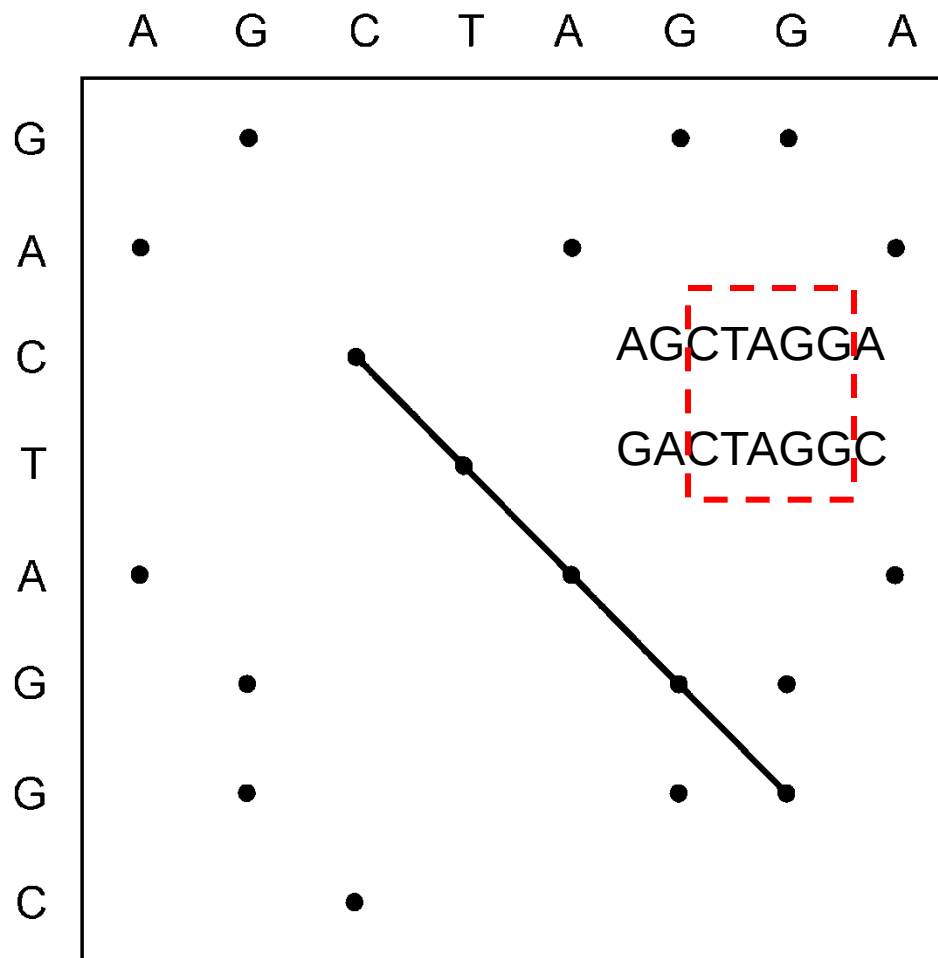
GEO Profiles: expression and molecular abundance profiles



GEO DataSets: experimental sets of GEO data



两条DNA序列的点阵法比较



Needleman-Wunsch算法

	G	A	T	C	T	A	
G	1						
A		2				1	GATCTA
T			3		1		GATCA
C				4			
A		1				5	(minus gap penalty) 空位罚分

Deduced alignment with gap Δ

G	A	T	C	T	A
G	A	T	C	Δ	A

全局优化 vs. 局部优化

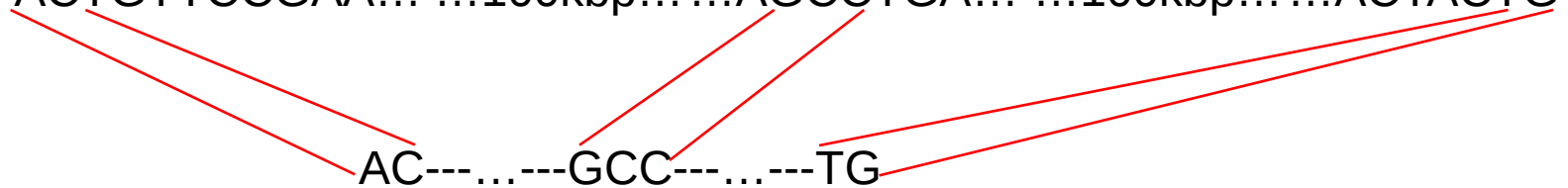
ACTGTTCCGAA... ..100kbp... ..AGCCTGA... ..100kbp.....ACTACTG

ACGCCTG

全局优化

ACTGTTCCGAA... ..100kbp... ..AGCCTGA... ..100kbp.....ACTACTG

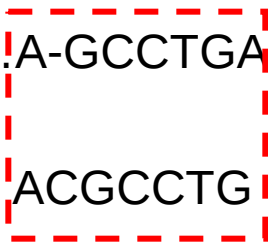
AC---...---GCC---...---TG

A diagram illustrating global sequence alignment. Red lines connect the sequence 'ACTGTTCCGAA... ..100kbp... ..AGCCTGA... ..100kbp.....ACTACTG' to the sequence 'AC---...---GCC---...---TG'. The connections are: 'A' to 'A', 'C' to 'C', 'T' to 'T', 'G' to 'G', 'T' to 'T', 'C' to 'C', 'C' to 'C', 'G' to 'G', 'A' to 'A', and 'A' to 'A'.

局部优化

ACTGTTCCGAA... ..100kbp... ..A-GCCTGA.....100kbp.....ACTACTG

ACGCCTG

A diagram illustrating local sequence alignment. A red dashed box highlights the alignment between the subsequence 'A-GCCTGA' of the top sequence and the full sequence 'ACGCCTG' of the bottom sequence.

最佳局部比对的得分要大于或等于最佳全局比对的得分

数据库中搜索相似序列

- 通过搜索数据库中相似序列发现基因功能
 - 例如反转录病毒编码的致癌基因 **v-sis**和**v-src**
 - 通过和模式生物已知遗传或生化信息的基因序列进行相似性搜索，预测新基因功能。
- **FASTA和BLAST**
 - **FASTA**:以几个残基长度的' word' 为单元进行检索；W. Pearson和D. Lipman开发。
 - **BLAST**:应用最广泛的序列相似性搜索工具，相比**FASTA**有更多改进，速度更快。
 - **PSI-BLAST** (Position-Specific Iterated BLAST)：位点特异性迭代BLAST
 - **PHI-BLAST** (Pattern Hit Initiated BLAST))：模式发现迭代BLAST

FastA和BLAST

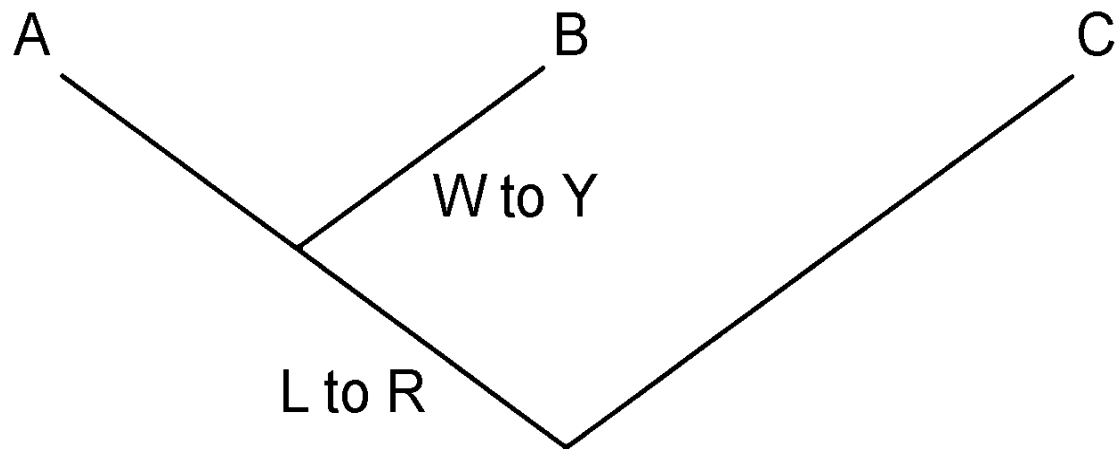
- **FastA和BLAST**程序是目前最常用的基于局部相似性的数据库搜索程序，它们都基于查找完全匹配的短小序列片段，并将它们延伸得到较长的相似性匹配。它们的优势在于可以在普通的计算机系统上运行，而不必依赖计算机硬件系统而解决运行速度问题。
- **BLAST**是目前常用的数据库搜索程序，它是**Basic Local Alignment Search Tool**的缩写，意为“基本局部相似性比对搜索工具”。
- 国际著名生物信息中心都提供基于**Web**的**BLAST**服务器。**BLAST**程序之所以使用广泛，主要因为其运行速度比**FastA**等其它数据库搜索程序快，而改进后的**BLAST**程序允许空位的插入。我们可以访问**NCBI**的网站在线进行**BLAST**和**FastA**的搜索

基于序列信息研究分子进化

- 构建进化树，分析蛋白质的超家族及亚家族分类。
- 寻找Ortholog (直系同源物)或者Paralog (旁系同源物)。
- 分子进化树的构建方法：
 - ✓ 邻接法(Neighbor-Joining),
 - ✓ 最大简约法(Maximum Parsimony),
 - ✓ 最大似然性法(Maximum Likelihood),
 - ✓ 以及贝叶斯类算法(MCMC)。
- 构建进化树的第一步：可靠的多序列比对。

不同物种的系统发育分析

ORGANISM A	A	W	T	V	A	S	A	V	R	T	S	I
ORGANISM B	A	Y	T	V	A	A	A	V	R	T	S	I
ORGANISM C	A	W	T	V	A	A	A	V	L	T	S	I



序列数据的文件格式

- DNA/RNA/氨基酸代码的标识
- GenBank数据格式
- EMBL & UniProt数据格式
- FASTA 数据格式

DNA代码

Symbol	Meaning
G	G
A	A
T	T
C	C
R	A or G
Y	C or T
M	A or C
K	G or T
S	C or G
W	A or T
H	A, G or T not C
B	C, G or T not A
V	A, C or G not T/U
D	A, G or T not C
N	A, C, G, or T

氨基酸代码

1-letter code	3-letter code	Amin Acid
A	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
O	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xxx	Any
Y	Tyr	tyrosine
B	Asx	Asp or Asn
Z	Glx	Glu, or Gln

GenBank数据文件格式

Format: **GenBank** [FASTA](#) [Graphics](#) [More Formats](#) ▼

GenBank: AJ310483.1

Thermoproteus tenax fba gene for fructose-bisphosphate aldolase (Class I)

[Comment](#) [Features](#) [Sequence](#)

LOCUS AJ310483 792 bp DNA linear BCT 02-AUG-2001
DEFINITION Thermoproteus tenax fba gene for fructose-bisphosphate aldolase (Class I).
ACCESSION AJ310483
VERSION AJ310483.1 GI:15130893
KEYWORDS class I; fba gene; fba-pfp operon; fructose-bisphosphate aldolase.
SOURCE Thermoproteus tenax
ORGANISM [Thermoproteus tenax](#)
Archaea; Crenarchaeota; Thermoprotei; Thermoproteales; Thermoproteaceae; Thermoproteus.
REFERENCE 1
AUTHORS Siebers,B., Brinkmann,H., Dorr,C., Tjaden,B., Lilie,H., van der Oost,J. and Verhees,C.H.
TITLE Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase
JOURNAL J. Biol. Chem. 276 (31), 28710-28718 (2001)
PUBMED [11387336](#)
REFERENCE 2 (bases 1 to 792)

GenBank数据文件格式

LOCUS	AJ310483	792 bp	DNA	linear	BCT 02-AUG-2001
DEFINITION	Thermoproteus tenax fba gene for fructose-bisphosphate aldolase (Class I).				
ACCESSION	AJ310483				
VERSION	AJ310483.1 GI:15130893				
KEYWORDS	class I; fba gene; fba-pfp operon; fructose-bisphosphate aldolase.				
SOURCE	Thermoproteus tenax Kra 1				
ORGANISM	Thermoproteus tenax Kra 1 Archaea; Crenarchaeota; Thermoprotei; Thermoproteales; Thermoproteaceae; Thermoproteus.				
REFERENCE	1				
AUTHORS	Siebers,B., Brinkmann,H., Dorr,C., Tjaden,B., Lilie,H., van der Oost,J. and Verhees,C.H.				
TITLE	Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase				
JOURNAL	J. Biol. Chem. 276 (31), 28710-28718 (2001)				
PUBMED	11387336				
REFERENCE	2 (bases 1 to 792)				
AUTHORS	Siebers,B.				
TITLE	Direct Submission				
JOURNAL	Submitted (13-FEB-2001) Siebers B., Microbiology, Universitaet Essen, Universitaetsstr. 5, Essen 45117, GERMANY				
COMMENT	related entry Y14655 fba gene forms an operon with the pfp gene coding for PPI-dependent phosphofructokinase fba-pfp operon.				
FEATURES	Location/Qualifiers				
source	1..792 /organism="Thermoproteus tenax Kra 1"				

GenBank数据文件格式

Locus名字

序列长度

序列类型

mRNA (= cDNA)
rRNA
snRNA
DNA

序列形状

修改日期

子库

```
LOCUS      AJ310483                792 bp      DNA      linear      BCT 02-AUG-2001
DEFINITION Thermoproteus tenax fba gene for fructose-bisphosphate aldolase
            (Class I).
ACCESSION   AJ310483
VERSION     AJ310483.1  GI:15130893
```

检索号

Genbank号

定义 (标题)

GenBank的数据类型

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. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTGS sequences (high throughput genomic sequences)
17. HTC - HTC sequences (high throughput cDNA sequences)
18. ENV - Environmental sampling sequences
19. CON - Constructed sequences

GenBank数据文件格式

FEATURES	Location/Qualifiers
source	1..792 /organism="Thermoproteus tenax Kra 1" /mol_type="genomic DNA" /strain="Kra 1 (DSM 2078)" /db_xref="taxon:768679" /country="Iceland"
gene	1..792 /gene="fba"
CDS	1..792 /gene="fba" /EC_number="4.1.2.13" /function="carbohydrate metabolism" /codon_start=1 /transl_table=11 /product="fructose-bisphosphate aldolase (Class I)" /protein_id="CAC48235.1" /db_xref="GI:15130894" /db_xref="GOA:P58315" /db_xref="InterPro:IPR002915" /db_xref="PDB:1OJX" /db_xref="UniProtKB/Swiss-Prot:P58315" /translation="MANLTEKFLRIFARRGKSIILAYDHGIEHGPADFMDNPDSADPE YILRLARDAGFDGVVFQRGIAEKYYDGSVPLILKLNGKTTLYNGEPVSVANCSVEEAV SLGASAVGYTIYPGSGFEWKMFEEELARIKRDAVKFDLPLVVWSYPRGGKVVNETAPEI VAYAARIALELGADAMKIKYTGDPKTFWAVKVAGKVPVLMSSGGPKTKTEEDFLKQVE GVLEAGALGIAVGRNVWQRRDALKFARALAEVYGGKKLAEPLNV"

GenBank数据文件格式

ORIGIN

```
1 atggcaaacc tcaccgagaa attcttaagg atattcgcgga ggaggggggaa gtccataata
61 ctggcctacg accacggcat tgagcacggg cccgcggact tcatggacaa cccggattca
121 gccgaccgag agtatatact gaggctcgcg agggacgccg gcttcgacgg agttgtgttc
181 cagaggggca tcgccgagaa gtactacgac gggagcgtgc cgctgatatt gaagctcaac
241 ggcaagacga cgttgtacaa cggcgagcct gtgtcgggtg ccaactgtag cgtcgaggag
301 gccgtaagcc tcggcgcaag cgccgtgggc tacacaatat acccaggcag cggctttgag
361 tggaagatgt ttgaggagct ggccagaatt aagagggacg ccgttaaatt cgacctgcc
421 ctggtggtct ggtcgtaccc gaggggcggg aaagtcgtca acgagacggc gcctgagatc
481 gtcgcctacg cggcgagaaat agccctggag ctcggcgcag atgctatgaa gataaagtac
541 acgggagatc ccaagacctt ctctggggcc gtcaaagtgg ccggcaaagt gcctgtgctt
601 atgtccggag gcccgaagac gaagactgag gaggacttcc tcaaacaagt ggagggagtc
661 cttgaggcgg gggccttggg cattgccgtc ggcagaaacg tctggcagag gagggacgcc
721 ctcaagtctg ccagagcgct tgcagagttg gtgtacggcg gaaagaagct ggccgagcct
781 ctgaacgtat ga
```

//

EMBL (UniProt) 数据格式

ID ALF1_THETE Reviewed; 263 AA.
AC P58315;
DT 18-OCT-2001, integrated into UniProtKB/Swiss-Prot.
DT 18-OCT-2001, sequence version 1.
DT 16-JUN-2009, entry version 41.
DE RecName: Full=Fructose-bisphosphate aldolase class 1;
DE EC=4.1.2.13;
DE AltName: Full=Fructose-bisphosphate aldolase class I;
DE Short=FBP aldolase;
GN Name=fba;
OS Thermoproteus tenax.
OC Archaea; Crenarchaeota; Thermoprotei; Thermoproteales;
OC Thermoproteaceae; Thermoproteus.
OX NCBI_TaxID=2271;
RN [1]
RP NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND CHARACTERIZATION.
RC STRAIN=Kra 1 / DSM 2078;
RX PubMed=11387336; DOI=10.1074/jbc.M103447200;
RA Siebers B., Brinkmann H., Doerr C., Tjaden B., Lilie H.,
RA van der Oost J., Verhees C.H. ;
RT "Archaeal fructose-1,6-bisphosphate aldolases constitute a new family
RT of archaeal type class I aldolases.";
RL J. Biol. Chem. 276:28710-28718(2001).
CC -!- CATALYTIC ACTIVITY: D-fructose 1,6-bisphosphate = glycerone
CC phosphate + D-glyceraldehyde 3-phosphate.

EMBL和GenBank数据格式的对比

EMBL	GenBank	含义
ID	LOCUS	序列名称
DE	DEFINITION	序列简单说明
AC	ACCESSION	序列编号
SV	VERSION	序列版本号
KW	KEYWORDS	与序列相关的关键词
OS	SOURCE	序列来源的物种名
OC	ORGANISM	序列来源的物种学名和分类学位置
RN	REFERENCE	相关文献编号，或递交序列的注册信息
RA	AUTHORS	相关文献作者，或递交序列的作者
RT	TITLE	相关文献题目
RL	JOURNAL	相关文献刊物杂志名，或递交序列的作者单位
RX	MEDLINE	相关文献 Medline引文代码
RC	REMARK	相关文献注释
RP		相关文献其它注释
CC	COMMENT	关于序列的注释信息
DR		相关数据库交叉引用号
FH	FEATURES	序列特征表起始
FT		序列特征表子项
SQ	BASE COUNT	碱基种类统计数
空格	ORIGIN	序列

FASTA格式

The screenshot shows a FASTA file viewer interface. The top bar includes a 'Display' dropdown set to 'FASTA', a 'Show' dropdown set to '20', a 'Send to' button, a 'File' dropdown, and a 'Get Subsequences' button. The main window displays a FASTA sequence starting with a definition line: `>gi|460243|gb|U07163.1|SCU07163`. This line is highlighted in a light blue box and labeled 'FASTA Definition Line'. Below the definition line, the sequence is shown in blocks of 60 characters. Several yellow callout boxes point to specific parts of the definition line: 'gi number' points to '460243', 'Database Identifiers' points to 'gb', 'Accession number' points to 'U07163.1', and 'Locus Name' points to 'SCU07163'. A table of database identifiers is also shown in a yellow box.

Database Identifiers	Database
gb	GenBank
emb	EMBL
dbj	DDBJ
sp	SWISS-PROT
pdb	Protein Databank
pir	PIR
prf	PRF
ref	RefSeq

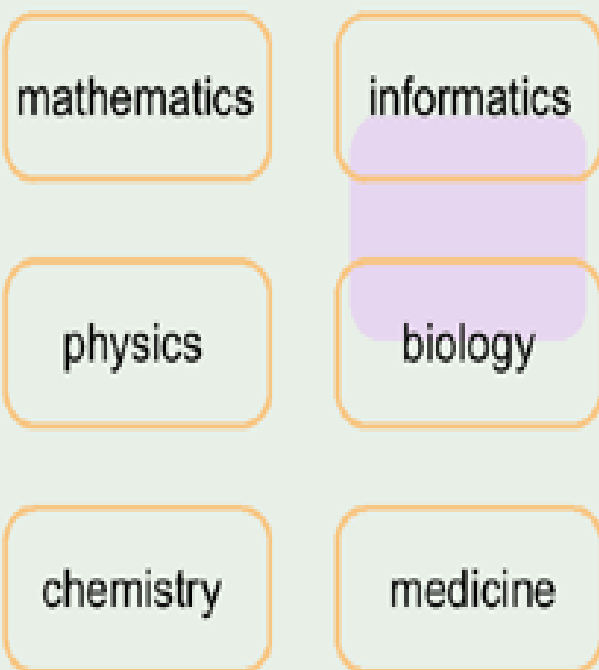
FASTA格式

>1I60:A|PDBID|CHAIN|SEQUENCE

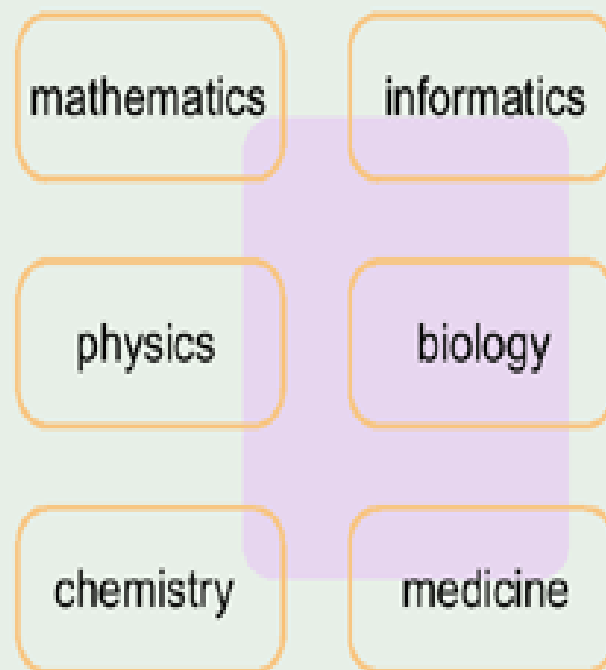
MKLCFNEATTLENSNLKLDLELCEKHGYDYIEIRTMDKLPEYL
KDHSLDDLAEYFQTHHIKPLALNALVFFNNRDEKGHNEIITE
FKGMMETCKTLGVKYVVAVPLVTEQKIVKEEIKKSSVDVLTEL
SDIAEPYGVKIALEFVGHPQCTVNTFEQAYEIVNTVNRDNVG
LVLDSFHFHAMGSNIESLKQADGKKIFIYHIDDTEDFPIGFLTD
EDRVWPGQGAILDAHLSALKEIGFSDVVSVELFRPEYYKLT
AEEAIQTAKKTTVDVVSKYFSM

生物信息学：学科交叉

bioinformatics now



bioinformatics in the future



生物信息学的相关知识储备

- 生物学背景：e.g. 分子生物学、细胞生物学、发育生物学、生物化学，...
- 数学知识：概率论与统计学等
- 算法及编程能力：JAVA, Perl/Python, PHP+MySQL, ...

生物信息学的常用算法与方法

- 动态规划算法(Dynamic programming);
- 贝叶斯统计(bayesian statistic);
- 人工神经网络(ANNs);
- 马尔可夫模型和隐马尔科夫模型(HMM);
- 遗传算法(Genetic Algorithm);
- 蒙特卡洛方法(Monte Carlo);
- 模拟退火算法(Simulated Annealing);
- 支持向量机(SVM);

...

科研机构及网络资源中心

- **NCBI:** 美国国立卫生研究院NIH下属国立生物技术信息中心NCBI。 <http://www.ncbi.nlm.nih.gov>
- **EMBnet:** 欧洲分子生物学网络
<http://www.embnet.org/>
- **EMBL-EBI:** 欧洲分子生物学实验室下属欧洲生物信息学研究所。 <http://www.ebi.ac.uk/>
- **ExPASy:** (Expert Protein Analysis System)瑞士生物信息研究所SIB下属的蛋白质分析专家系统;
<http://www.expasy.org/>

科研机构及网络资源中心

- Bioinformatics Links Directory:
http://bioinformatics.ca/links_directory/
- 各种数据库等
如 PDB (Protein Data Bank)
UniProt 数据库
- 软件资源:
<http://www.expasy.org/tools/>
<http://www.ebi.ac.uk/Tools/>
<http://www.ncbi.nlm.nih.gov/Tools/>

国内生物信息中心举例

□ **CBIPKU:** 北京大学生物信息中心

www.cbi.pku.edu.cn/chinese/

□ **BioSino:** 中国生物信息

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

中国科学院上海生命科学院生物信息中心

□ 上海生物信息技术研究中心

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