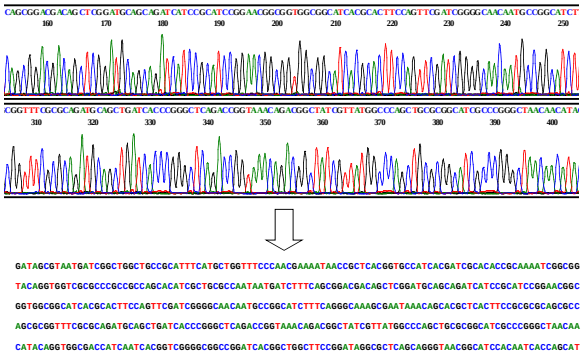
Molekulárně biologické databáze

Pro zajímavost, nebude součástí zkoušky…

Důležité, pravděpodobně bude u zkoušky…

Molekulárně biologická data



Molekulárně biologická data

MALDI-TOF

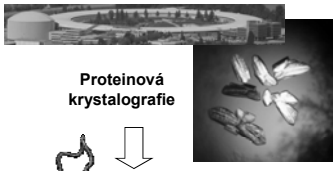
Identifikace proteinů

Sekvenace proteinů

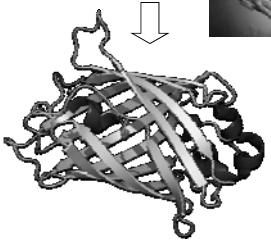
```
MDRNGNFSLPPNTAFKAIFYANAADRQDLK
LFIDDAPEPAATFVGNSEDGVRLFTLNSKG
GKIRIEASANGRQSATDARLAPLSAGDTVW
LGWLGAEDGADADYNDGIVILQWPIT
```

Molekulárně biologická data

NMR spektroskopie



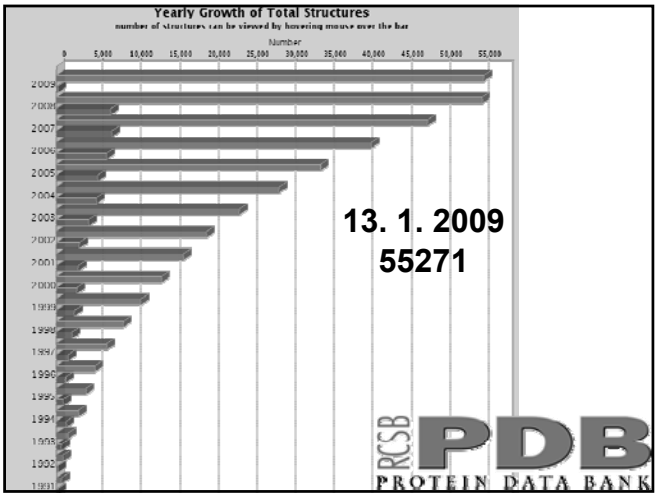
Proteinová krystalografie



Molekulárně biologická data

- **Výkonné technologie:**
  - Automatické sekvencování
  - MALDI-TOF
  - NMR spektroskopie
  - Proteinová krystalografie

Výrazný nárůst množství biologických dat.



Éra reverzní genetiky

Klasická genetika

GATAGCGTAATGATCGGCTGGCTGCCGCATTTC
TACAGGTGGTCGCGCCCGCCGCCAGCACATCGC
GGTGGCGGCATCACGCACTTCCAGTTCGATCGG
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCC

Fenotyp → Genotyp

Reverzní genetika

Automatické DNA sekvencování

Produkce velkého množství dat

GATAGCGTAATGATCGGCTGGCTGCCGCAT
TACAGGTGGTCGCGCCCGCCGCCAGCACAT

Genotyp → Fenotyp

Bi7201 Základy genomiky, podzimní semestr

Molekulárně biologická data

- Nutnost organizovaného ukládání a skladování dat.
- Nutnost prohlížení a analyzování uložených dat.

Databáze je určitá uspořádaná množina informací (dat) uložená na paměťovém médiu.

V širším smyslu jsou součástí databáze i softwarové prostředky, které umožňují manipulaci s uloženými daty a přístup k nim.

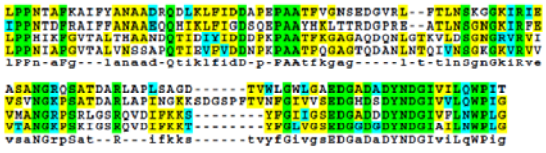
Analytické nástroje

- **Vyhledávácí software**

  Nutnost snadného, rychlého a specifického vyhledání informací.

- **Srovnávání dat (sekvencí)**

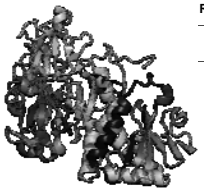
  Sequence alignment – „seřazení“ sekvencí.



Rozdělení molekulárně biologických databází

- Databáze:
  - Primární
  - Sekundární
  - Strukturní

EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK



Ribosome-inactivating protein, subdomain 1

Ribosome-inactivating protein, subdomain 2

Ricin B-like lectins

Rozdělení molekulárně biologických databází

- Databáze:
  - Primární
  - Sekundární
  - Strukturní

EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK

Primární databáze obsahují anotované sekvence NA nebo proteinů.

Rozdělení molekulárně biologických databází

- Databáze:
  - Primární
  - Sekundární
  - Strukturní

EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK

Ribosome-inactivating protein, subdomain 1

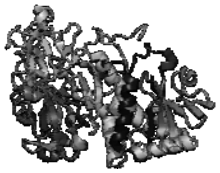
Ribosome-inactivating protein, subdomain 2

Ricin B-like lectins

Sekundární databáze obsahují informace odvozené z primárních databází ve formě charakteristických vzorů sekvencí, tj. funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí).

Rozdělení molekulárně biologických databází

- **Databáze:**
  Primární
  Sekundární
  Strukturní



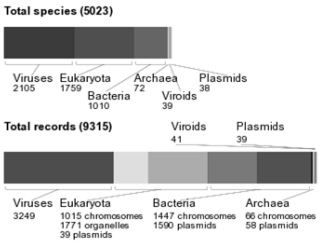
Abrin

Obsahují struktury proteinů (nukleových kyselin) a jejich anotace.

Rozdělení molekulárně biologických databází

- **Databáze:**
  Primární
  Sekundární
  Strukturní

  Genomové zdroje



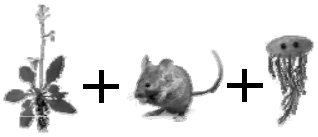
Rozdělení molekulárně biologických databází

- **Databáze:**
  Specializované
  Univerzální

Specializované databáze obsahují informace o určité proteinové rodině nebo skupině proteinů, případně o určitém organismu.

Rozdělení molekulárně biologických databází

- **Databáze:**
  Specializované
  Univerzální



Univerzální databáze obsahují informace o proteinech (NA) ze všech organismů.

Rozdělení univerzálních proteinových databází

- **Univerzální databáze:**

  „Skladiště“ sekvencí – sequence repository

  „Manuálně“ spravovaná – curated database

Rozdělení univerzálních proteinových databází

- **„Skladiště“ sekvencí – sequence repository**

  Kromě sekvencí obsahují málo nebo žádné dodatečné informace.

  Záznamy generovány automaticky.

  Proteiny mohou být zastoupeny několika různými záznamy (sekvencemi) = „nadbytečnost“ (redundance) sekvencí.

Rozdělení univerzálních proteinových databází

- **Manuálně spravované – curated databases**

  Záznamy obsahují dodatečné informace.

  Informace jsou před vložením do databáze validovány experty.

  Všechny záznamy o stejné proteinové sekvenci jsou sdružovány do jediného = non-redundant dataset.

Rozdělení molekulárně biologických databází

- **Databáze:**
  Primární
  Sekundární
  Strukturní

  Genomové zdroje
  Složené databáze

Složené databáze

- **Složené (composite) databáze:**

  Slučují data z několika primárních databází.

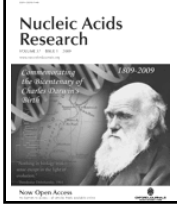
  Eliminace redundantních dat.

  Různá priorita zdrojových databází podle kvality validace a anotace (eliminace redundantních dat z databáze s nižší prioritou).

Molekulárně biologické databáze

Nucleic Acids Research

http://www3.oup.co.uk/nar/database/a/



Nucleotide Sequence Databases
International Nucleotide Sequence Database Collaboration
Coding and non-coding DNA
Gene structure, introns and exons, splice sites
Transcriptional regulator sites and transcription factors
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

1170 databází

EBI/NCBI/CIB

Instituce zabývající se shromažďováním, správou a poskytováním dat a informací a vývojem analytických nástrojů.

EBI
Evropský institut pro bioinformatiku

EMBL-EBI

European Bioinformatics Institute

http://www.ebi.ac.uk/

NCBI
Národní centrum pro biotechnologické informace

NCBI

National Center for Biotechnology Information

http://www.ncbi.nlm.nih.gov/

CIB
Centrum pro informační biologii

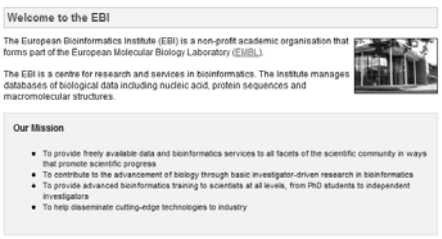
Center for Information Biology

http://www.cib.nig.ac.jp/

EBI – Evropský institut pro bioinformatiku

European Bioinformatics Institute

- **Založen roku 1992 jako součást European Molecular Biology Laboratory - EMBL.**
- **Sídlo v Hinxtonu ve Velké Británii.**



NCBI - Národní centrum pro biotechnologické informace

NCBI — National Center for Biotechnology Information
National Library of Medicine — National Institutes of Health

- Založeno v roce 1988 jako oddělení Národní lékařské knihovny (National Library of Medicine – NLM) v USA.
- Součást National Institutes of Health – NIH.

What does NCBI do?

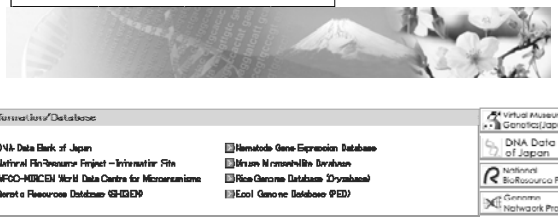
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

CIB – Centrum pro informační biologii

- Založeno jako oddělení Národního genetického institutu (国立遺伝学研究所, NIG) v Japonsku.

Research Organization of Information and Systems
National Institute of Genetics

http://www.nig.ac.jp/



Primární databáze NA

- **EMBL -** Evropský institut pro bioinformatiku

EMBL-EBI

- **GenBank -** Národní centrum pro biotechnologické informace

NCBI

- **DDBJ -** Národní genetický institut (NIG)

大学共同利用機関法人 情報・システム研究機構
国立遺伝学研究所

EMBL

EMBL-EBI

EMBL NUCLEOTIDE SEQUENCE DATABASE

- EMBL Nucleotide Sequence Database (EMBL-Bank) byla založena roku 1980 jako první databáze nukleotidových sekvencí.
- Obsahuje sekvence RNA a DNA.
- Zdroje sekvencí: vloženy přímo autory, genomové projekty, patenty

This morning the EMBL Database contained 244,322,213,780 nucleotides in 153,137,008 entries.

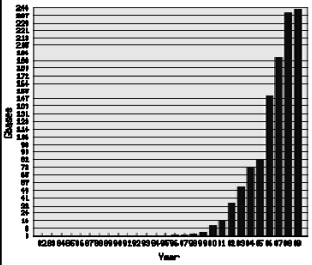
This morning = 21.1.2009

EMBL

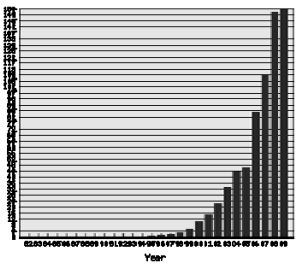
EMBL-EBI

EMBL NUCLEOTIDE SEQUENCE DATABASE

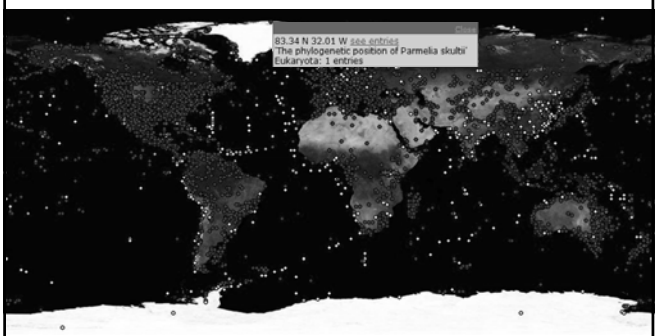
Total nucleotides
(current 244,322,213,780)



Number of entries
(current 153,137,008)



The map shows 16,892,540 entries distributed over 11,882 locations.



The dots on the map have different colours according to the taxonomy of the specimens:
Eukaryota Bacteria Archaea Other Mixed

EMBL „entry“

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifoliese; Trifolium.
XX
RN   [5]
RP   1-1859
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
```

Translation = *proteinová* databáze

```
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE36"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40036.1"
FT                   /translation="MDFIVAILALFVISSFSITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEDGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIM
FT                   KQMGFDSYRFSISWPRILPKGKLSGGINREGIKYYNNLINELLANGIQPFVTLFHWDLP
FT                   YELEEKYGGFLNSGVINDFNDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGSNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKEKFQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRLVKNRLPKFSKFESSLVNGSFDF
FT                   IGIDYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASNWIYVYPKGLQ
FT                   DLKYILKEYTNLPMFITENGMDEFNDATLPVEEALLNDYRIDYYRKHLYYIRSA
FT                   IRAGSNVKGYFAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
aaacaaacca aatatggatt ttattgtagc catatttgct ctgttagtta ttagctcatt     60
cacaattact tccacaaatg cagttgaagc ttctactctt ctttgcaatag gtaacctgag    120
tcggagcagt tttcctcgtg gcttcatctt tgggtctgga tctttcagcat atcaatttga    180
aggtgcagta aacgaaggcg gtagaggacc aagtattttgg gatacctttca ccatcaaata    240
tccagaaaaa atcagggatg gaagccatgc agacatcacg gttgaccaat atcacagcta    300
cagggagat gttgggatca tgaaggatca aaatatggac tcgtatagac tctcaattctc    360
ttggccaata atactccaa aqqqaaqttc aqqcqqaqqc atcaatcacq aqqaatccaa    420
```

Formát EMBL databáze

ID	identification	(begins each entry; 1 per entry)
AC	accession number	(>=1 per entry)
PR	project identifier	(0 or 1 per entry)
DT	date	(2 per entry)
DE	description	(>=1 per entry)
KW	keyword	(>=1 per entry)
OS	organism species	(>=1 per entry)
OC	organism classification	(>=1 per entry)
OG	organelle	(0 or 1 per entry)
RN	reference number	(>=1 per entry)
RC	reference comment	(>=0 per entry)
RP	reference positions	(>=1 per entry)
RX	reference cross-reference	(>=0 per entry)
RG	reference group	(>=0 per entry)
RA	reference author(s)	(>=0 per entry)
RT	reference title	(>=1 per entry)
RL	reference location	(>=1 per entry)
DR	database cross-reference	(>=0 per entry)
CC	comments or notes	(>=0 per entry)
AH	assembly header	(0 or 1 per entry)
AS	assembly information	(0 or >=1 per entry)
FH	feature table header	(2 per entry)
FT	feature table data	(>=2 per entry)
XX	spacer line	(many per entry)
SQ	sequence header	(1 per entry)
CO	contig/construct line	(0 or >=1 per entry)
bb	(blanks) sequence data	(>=1 per entry)
//	termination line	(ends each entry; 1 per entry)

Formát EMBL databáze

```
ID   <1>; SV <2>; <3>; <4>; <5>; <6>; <7> BP.
```

The tokens represent:

1. Primary accession number
2. Sequence version number
3. Topology: 'circular' or 'linear'
4. Molecule type
5. Data class
6. Taxonomic division
7. Sequence length

```
ID   CD789012; SV 4; linear; genomic DNA; HTG; MAM; 500 BP.
```

http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

Formát EMBL databáze

Class	Definition
CON	Entry constructed from segment entry sequences, drawing annotation from segment entries
ANN	Entry constructed from segment entry sequences with its own annotation
PAT	Patent
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTC	High Throughput cDNA sequencing
HTG	High Throughput Genome sequencing
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TSA	Third Party Annotation
STS	Sequence Tagged Site
STD	Standard (all entries not classified as above)

http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

Formát EMBL databáze

Division	Code
Bacteriophage	PHG
Environmental Sample	ENV
Fungal	FUN
Human	HUM
Invertebrate	INV
Other Mammal	MAM
Other Vertebrate	VRT
Mus musculus	MUS
Plant	PLN
Prokaryote	PRO
Other Rodent	ROD
Synthetic	SYN
Transgenic	TGN
Unclassified	UNC
Viral	VRL

http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html

GenBank

- **Založena roku 1982 v rámci institutu NCBI.**

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2008 Jan;36(Database Issue):D25-30). There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.

Sample GenBank Record

http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Sample GenBank Record

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Sample GenBank Record

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevi
  ORGANISM  Saccharomyces cerevi
            Eukaryota; Fungi; As
            Saccharomycetales; S
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P
  TITLE     Cloning and sequence
            DNA damage-induced m
  JOURNAL   Yeast 10 (11), 1503-
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K.
  TITLE     Selection of axial g
            plasma membrane glyc
  JOURNAL   Genes Dev. 10 (7), 7
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-19
            Haven, CT, USA
```

GenBank Division

The GenBank division to which a record belongs is indicated with a three letter abbreviation. In this example, GenBank division is PLN.

The GenBank database is divided into 18 divisions:

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 - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

The DNA Data Bank of Japan

- **Původně zahrnovala data především z japonských výzkumů.**
- **V současnosti úzká spolupráce s ostatními databázemi.**

INSDC International Nucleotide Sequence Database Collaboration

http://www.insdc.org/

DDBJ: DNA Data Bank of Japan
CIB-DDBJ: Center for Information Biology and DNA Data Bank of Japan
NIG: National Institute of Genetics

EBI: European Bioinformatics Institute
EMBL: European Molecular Biology Laboratory

NCBI: National Center for Biotechnology Information
NLM: National Library of Medicine

IAC: International Advisory Committee
ICM: International Collaborative Meeting

Primární databáze proteinů

- **Univerzální databáze:**
  „Skladiště“ sekvencí – sequence repository
  Manuálně spravovaná – curated database

Příklad: **GenBank *versus* RefSeq**

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

Primární databáze proteinů

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

GenPept - GenBank Gene Products Data Bank
RefSeq - Reference Sequence

Primární databáze proteinů

- PIR-PSD - Protein Information Resource Protein Sequence Database.
- Nejstarší univerzální „curated“ databáze proteinů.
- Komplexní, non-redundant data, rozčleněna podle proteinových rodin a nadrodin, doplněna funkčními, strukturními a bibligrafickými daty.

http://pir.georgetown.edu/

Swiss-PROT + TrEMBL

- Swiss-Prot - „Cutared“ databáze založená na Univerzitě v Ženevě v roce 1986. Spravovaná Švýcarským institutem pro bioinformatiku (SIB - Swiss Institute of Bioinformatics).
- Vysoká úroveň anotace ⟹ vkládáno více sekvencí než je možno manuálně anotovat a zařadit do databáze.
- TrEMBL – Počítačově anotovaná data, odvozená z kódujících úseku sekvencí v DDBJ/EMBL/GenBank, která ZATÍM nejsou zařazena v Swiss-Prot.

Swiss-PROT + TrEMBL

- Anotace: Funkce
  Katalytická aktivita
  Podjednotky
  Domény
  Biotechnologické využití
  Sekvenční homologie
  Posttranslační modifikace
  Reference atd.

http://www.expasy.org/sprot/

Složené databáze

- Databáze:
  Primární
  Sekundární
  Strukturní

  Genomové zdroje
  Složené databáze

Složené databáze

- Složené (composite) databáze:

  Slučují data z několika primárních databází.

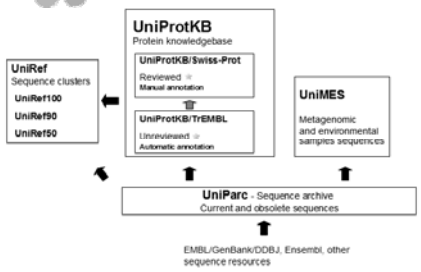
  Eliminace redundantních dat.

  Různá priorita zdrojových databází podle kvality validace a anotace (eliminace redundantních dat z databáze s nižší prioritou).

Swiss-PROT + TrEMBL

OWL (Swiss-PROT + PIR + Genbank + NRL-3D)

UniProt



2002- spolupráce mezi EBI, SIB a PIR

http:/www.uniprot.org

Sekundární databáze NA a proteinů

Sekundární databáze obsahují informace odvozené z primárních databází ve formě charakteristických vzorů sekvencí, tj. funkčních nebo strukturních motivů získaných srovnáním primárních dat (sekvencí).

- **Vyhledávání „vzoru“ charakteristického pro určitou skupinu proteinů.**
- **Možnost predikce funkce proteinů.**

Sekundární databáze NA a proteinů

- **Databáze mohou obsahovat:**

  Proteinové DOMÉNY odvozené ze známých struktur

  Proteinové sekvence seřazené do SEKVENČNÍCH RODIN

  CHARAKTERISTICKÉ MOTIVY odvozené z těchto sekvenčních rodin.

```
LPPNTAFKAIFYANAADRQDLKLFIDD
IPPNTDFRAIFFANAAEQGHIKLFIGD
LPPHIKFGVTALTHAANDQTIDIYIDD
LPPNIAFGVTALVNSSAPQTIEVFVDD
```

[illegible]

This pattern is translated as: [Phe or Cys]-any-Val-any-any-any-any-[any but Glu or Asp]

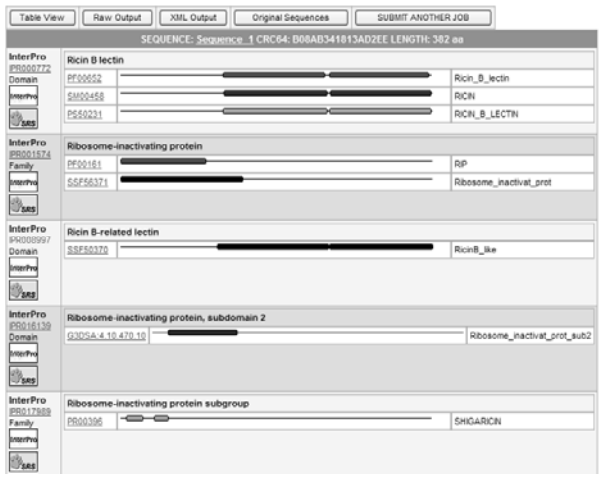
Sekundární databáze NA a proteinů

- **Sekundární proteinové databáze:**

  PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMS

  V současné době sdruženy do integrované klasifikační databáze proteinů InterPro

  http://www.ebi.ac.uk/InterProscan/



Sekundární databáze NA a proteinů

- **Sekundární proteinové databáze:**

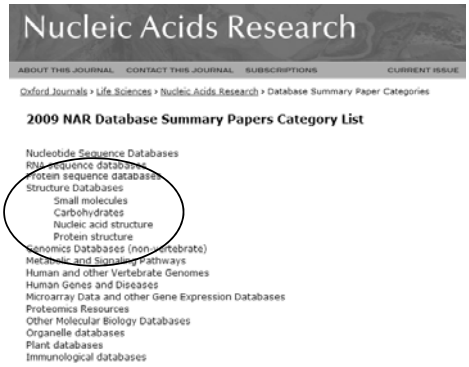
  PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMS

  V současné době sdruženy do integrované klasifikační databáze proteinů InterPro

  http://www.ebi.ac.uk/InterProscan/

- **Sekundární databáze NA**
  TRANSFAC

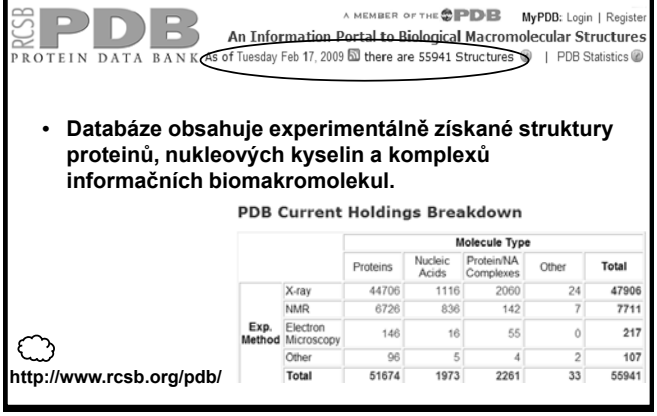
Strukturní databáze



http://www3.oup.co.uk/nar/database/a/

Strukturní databáze proteinů

PDB - Protein Data Bank



- **Databáze obsahuje experimentálně získané struktury proteinů, nukleových kyselin a komplexů informačních biomakromolekul.**

PDB Current Holdings Breakdown

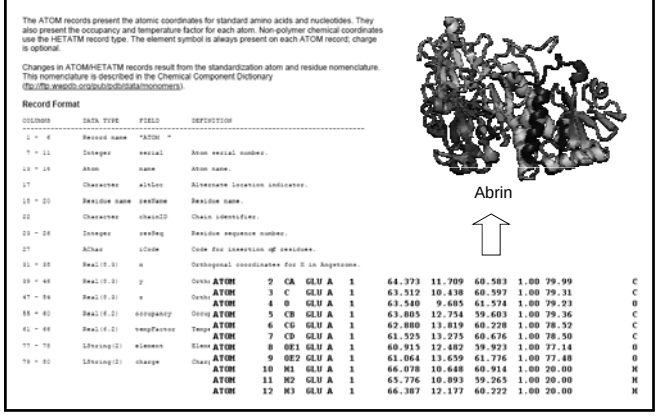
		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	44706	1116	2060	24	47906
	NMR	6726	836	142	7	7711
	Electron Microscopy	146	16	55	0	217
	Other	96	5	4	2	107
	Total	51674	1973	2261	33	55941

http://www.rcsb.org/pdb/

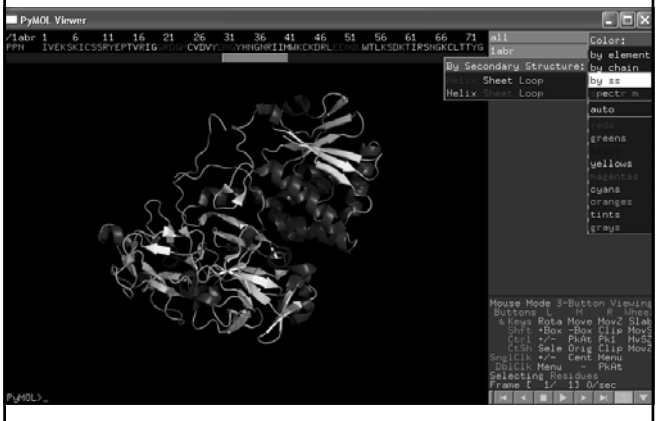
PDB formát

- **PDB formát – původní formát databáze.**
- **1997 – mmCIF (macromolecular Crystallographic Information File).**
- **Záznamy jsou v databázi uloženy v obou formátech a volně stažitelné.**
- **PDB formát – rozeznáván téměř všemi programy pro práci se strukturami.**

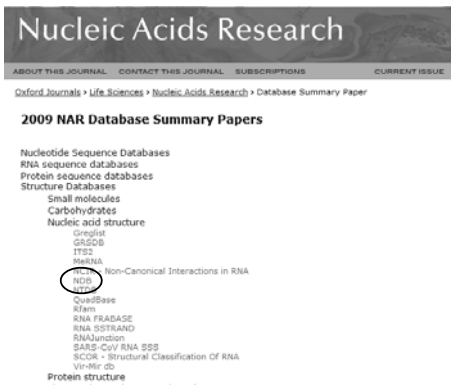
PDB formát



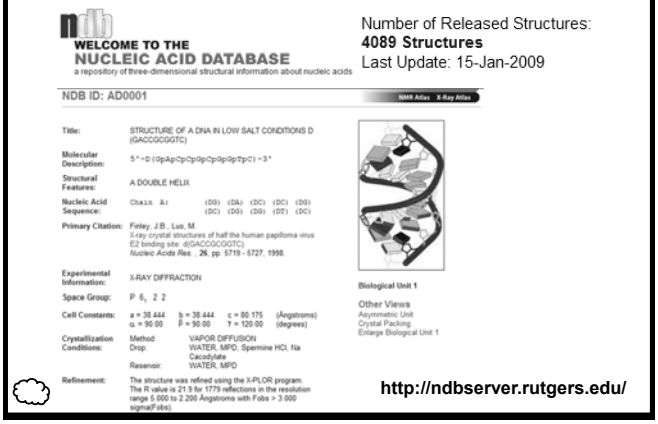
Abrin



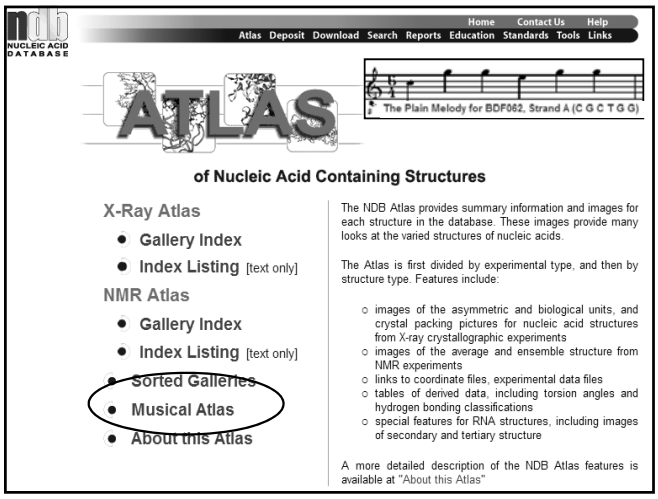
Strukturní databáze NA



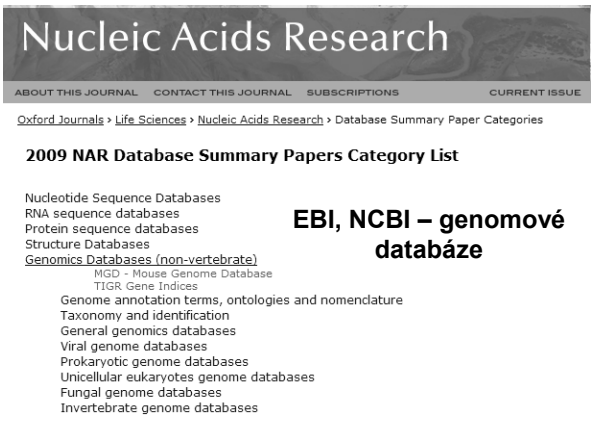
NDB - Nucleic Acid Database



http://ndbserver.rutgers.edu/



Genomové zdroje



EBI, NCBI – genomové databáze

Vyhledávací systémy

- **Nutnost organizovaného ukládání a skladování dat.**
- **Nutnost prohlížení a analyzování uložených dat.**

Databáze je určitá uspořádaná množina informací (dat) uložená na paměťovém médiu.

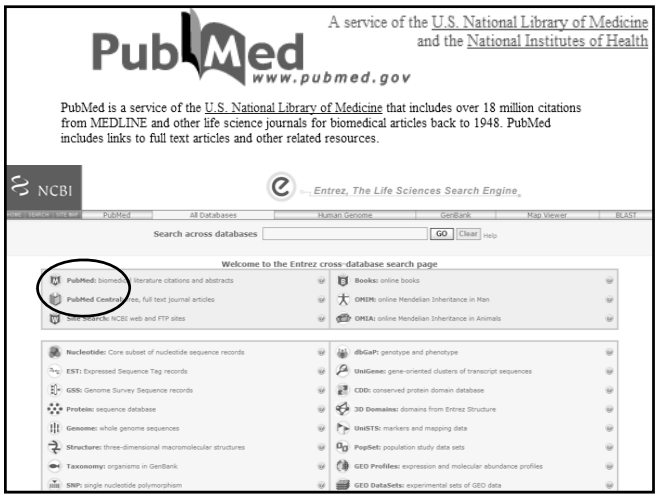
V širším smyslu jsou součástí databáze i softwarové prostředky, které umožňují manipulaci s uloženými daty a přístup k nim.

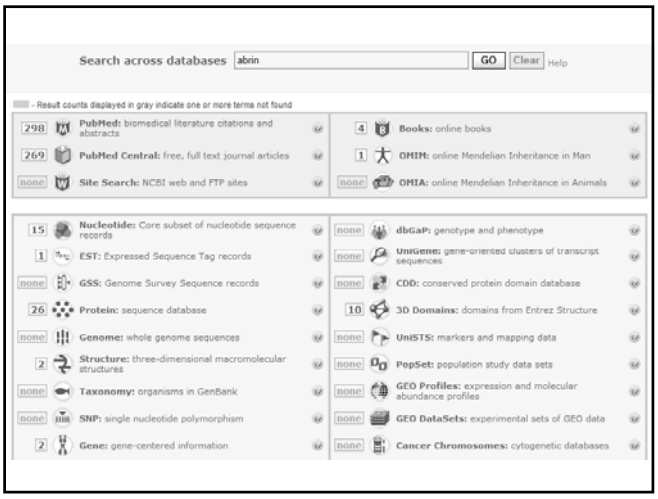
Vyhledávací systémy

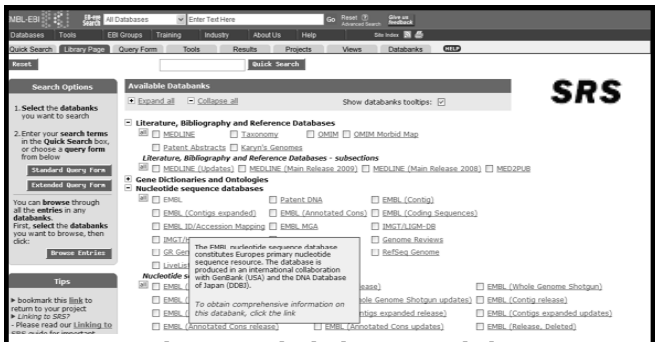
- **Textové vyhledávání v databázích NCBI – Entrez**
  http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others.

http://www.ncbi.nlm.nih.gov/Entrez/tutor.html







- **Textové vyhledávání v databázích**

  EBI– SRS

  Sequence Retrieval System
  http://srs.ebi.ac.uk/

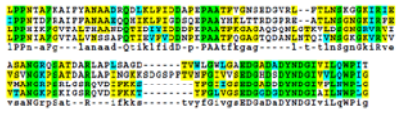
Vyhledávací systémy

- **Vyhledávání podobností sekvencí**

  Textové vyhledávání může selhat (nedostatečná anotace).

  Vyskytuje se shodná nebo podobná sekvence v databázi? (Identifikace možné funkce na základě homologie.)

- **Specializované nástroje (algoritmy) pro „seřazení“ (alignment) sekvencí.**



Shrnutí

- **Výrazný nárůst množství biologických dat vede k nutnosti jejich organizovaného skladování a analyzování (databáze).**
- **Instituce pro správu dat a vývoj nástrojů pro analýzu: EBI/NCBI/CIB**
- **Základní rozdělení databází: primární/sekundární/strukturní databáze**
- **Textové vyhledávací systémy: Entrez/SRS**