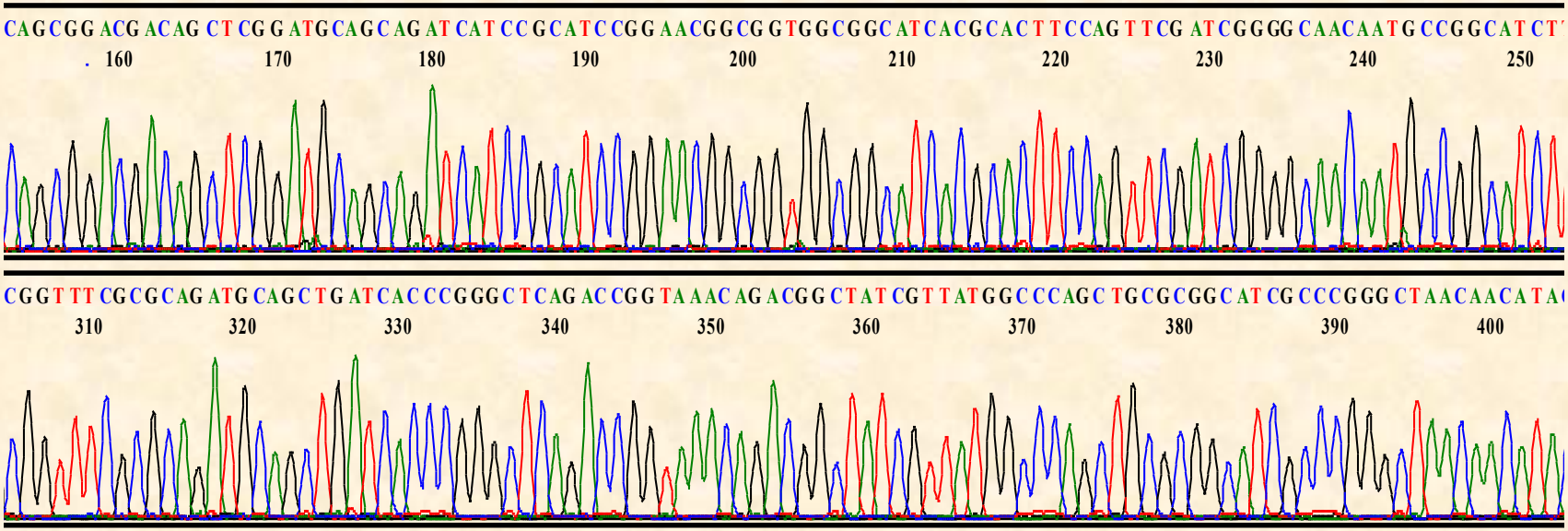
Molekulárně biologické databáze

Pro zajímavost, nebude součástí zkoušky…

Důležité, pravděpodobně bude u zkoušky…

Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

Molekulárně biologická data

MALDI-TOF

Identifikace proteinů

Sekvenace proteinů

MDRNGNFSLPPNTAFKAIFYANAADRQDLK

LFIDDAPEPAATFVGNSEDGVRLFTLNSKG

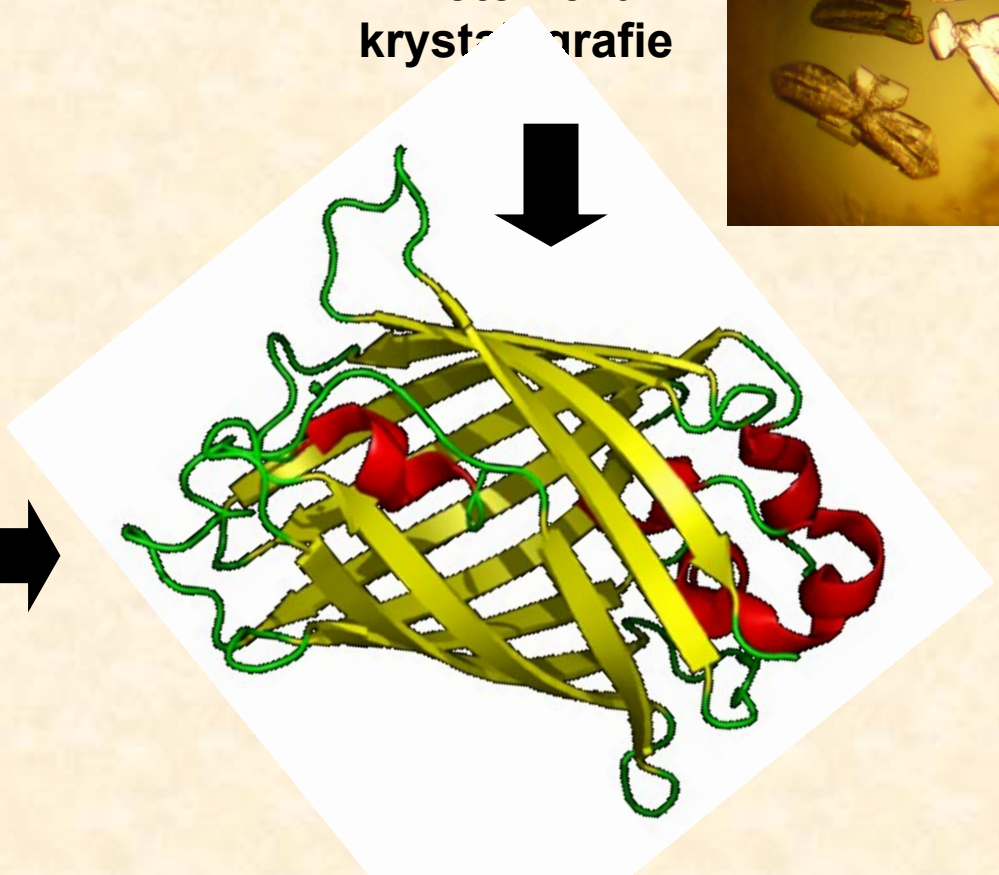
GKIRIEASANGRQSATDARLAPLSAGDTVW

LGWLGAEDGADADYNDGIVILQWPIT

Molekulárně biologická data

NMR spektroskopie

Proteinová krystalografie



Molekulárně biologická data

- **Výkonné technologie:**

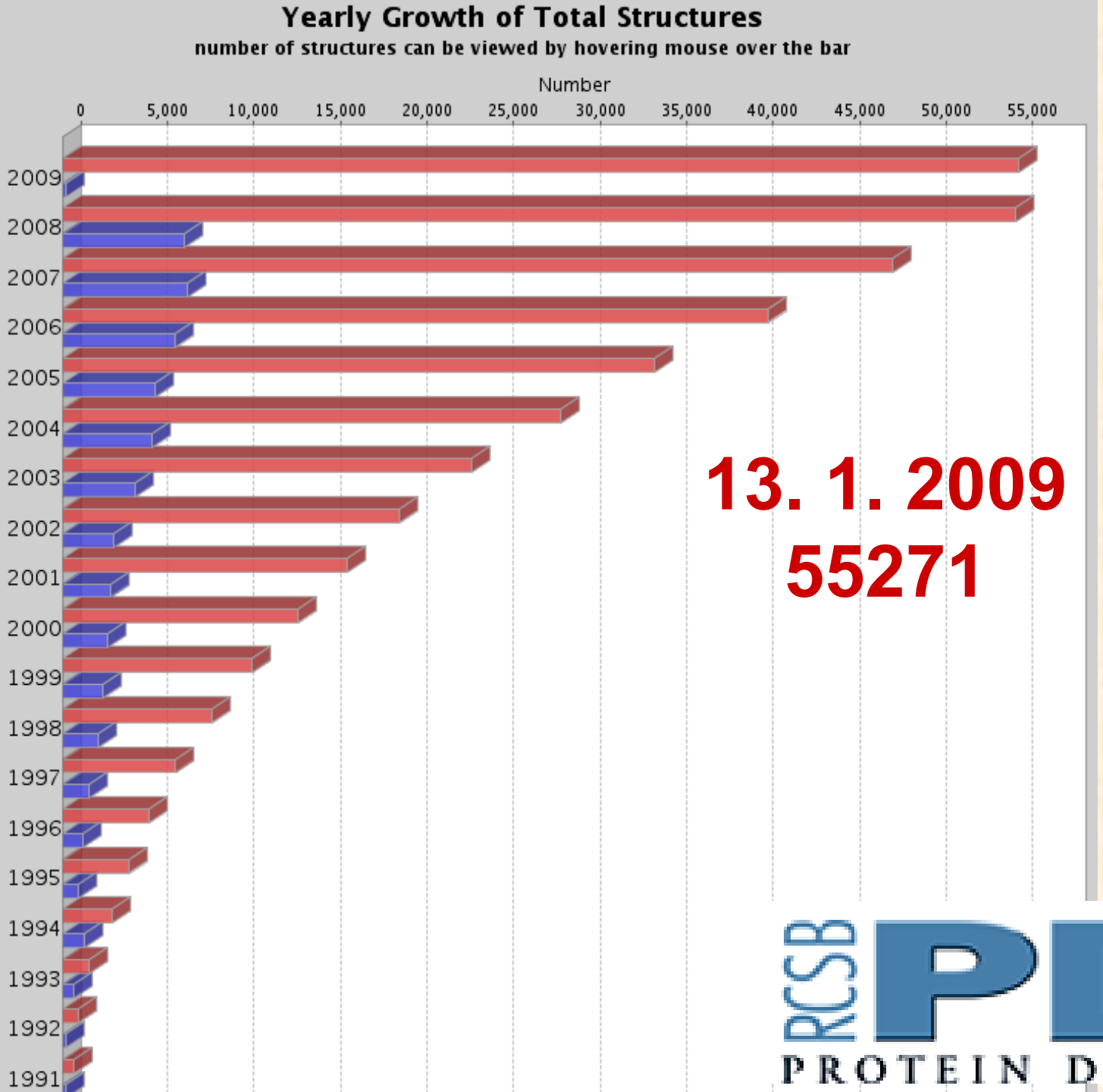
  Automatické sekvencování

  MALDI-TOF

  NMR spektroskopie

  Proteinová krystalografie

Výrazný nárůst množství biologických dat.



13. 1. 2009

55271

RCSB PDB

PROTEIN DATA BANK

Éra reverzní genetiky

Klasická genetika

GATAGCGTAATGATCGGCTGGCTGCCGCATTTC

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGG

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCC

Fenotyp → Genotyp

Reverzní genetika

Automatické DNA sekvencování

Produkce velkého množství dat

GATAGCGTAATGATCGGCTGGCTGCCGCAT

TACAGGTGGTCGCGCCCGCCGCCAGCACAT

Genotyp → Fenotyp

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.

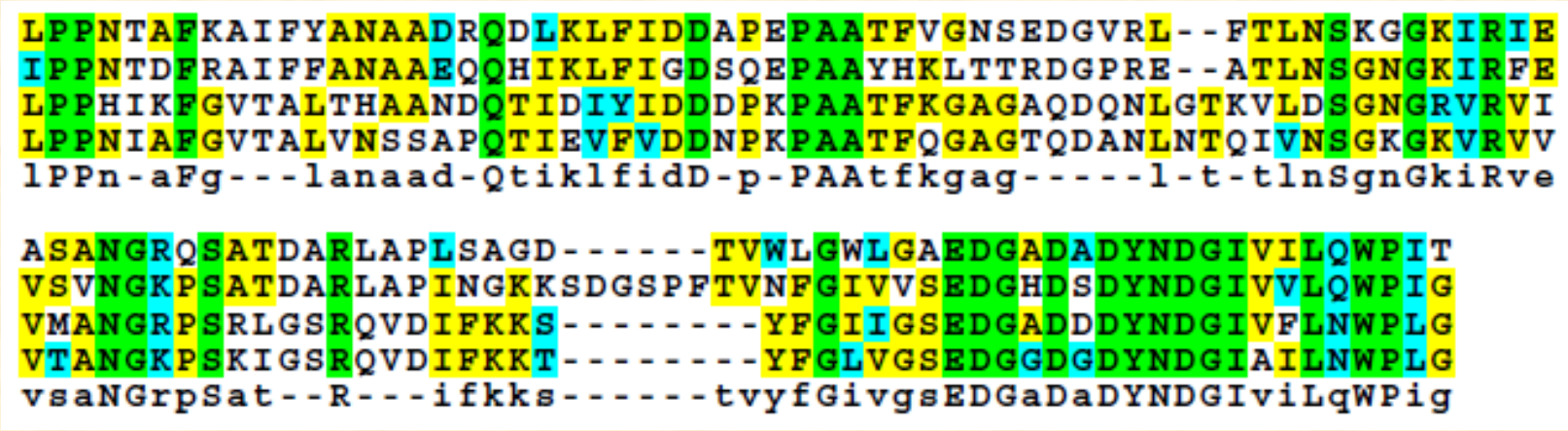
Analytické nástroje

- **Vyhledávácí software**

  Nutnost snadného, rychlého a specifického vyhledání informací.

- **Srovnávání dat (sekvencí)**

  Sequence alignment – „seřazení“ sekvencí.



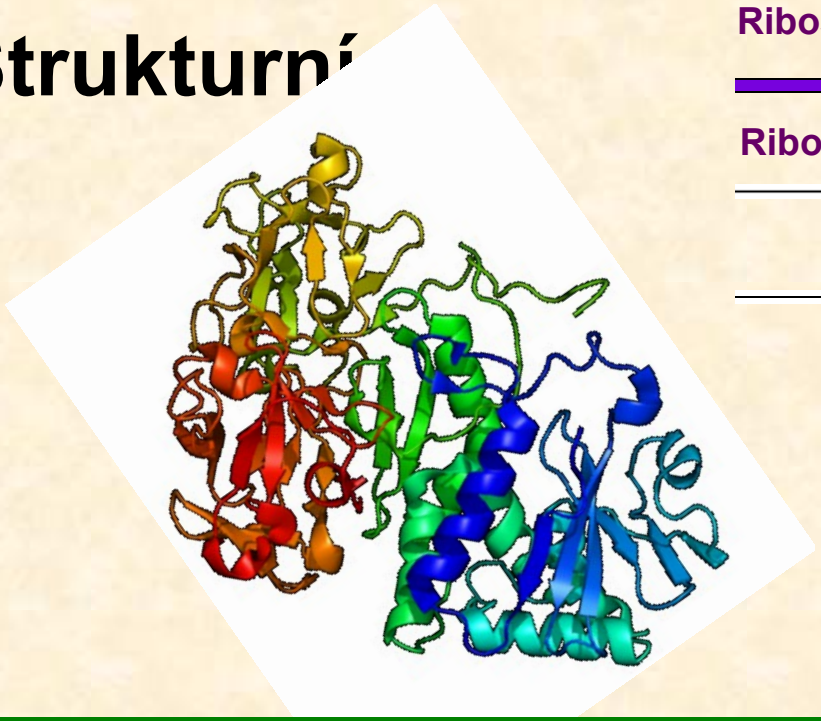
Rozdělení molekulárně biologických databází

- **Databáze:**

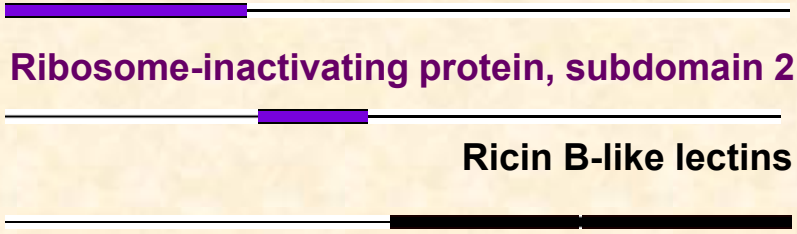
  Primární

  Sekundární

  Strukturní

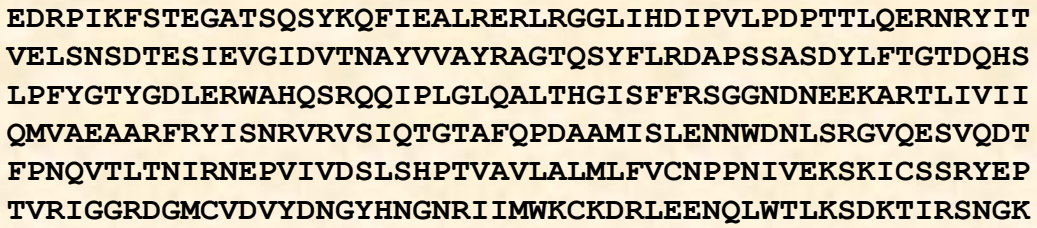


```
EDRPIKFSTEGATSQSYKQFIEALRERLRGGLIHDIPVLPDPTTLQERNRYIT
VELSNSDTESIEVGIDVTNAYVVAYRAGTQSYFLRDAPSSASDYLFTGTDQHS
LPFYGTYGDLERWAHQSRQQIPLGLQALTHGISFFRSGGNDNEEKARTLIVII
QMVAEAARFRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGVQESVQDT
FPNQVTLTNIRNEPVIVDSLSHPTVAVLALMLFVCNPPNIVEKSKICSSRYEP
TVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGK
```



Rozdělení molekulárně biologických databází

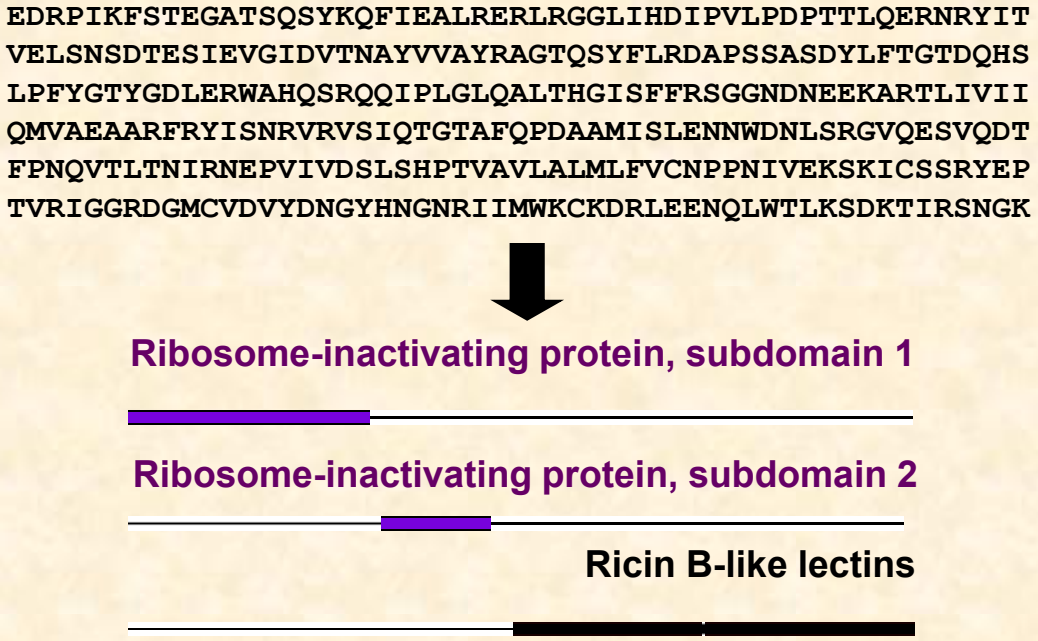
- **Databáze:**
  Primární
  Sekundární
  Strukturní



Primární databáze obsahují anotované sekvence NA nebo proteinů.

Rozdělení molekulárně biologických databází

- **Databáze:**
  - Primární
  - **Sekundární**
  - Strukturní



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

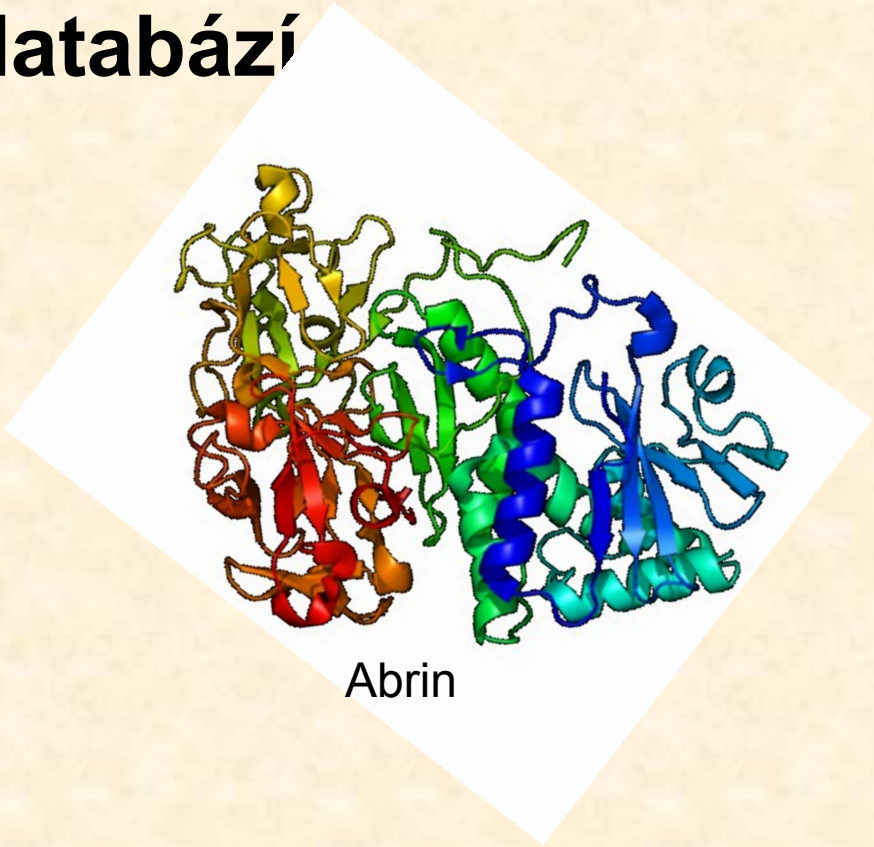
Rozdělení molekulárně biologických databází

- **Databáze:**

  Primární

  Sekundární

  Strukturní



Abrin

Obsahují struktury proteinů (nukleových kyselin) a jejich anotace.

Rozdělení molekulárně biologických databází

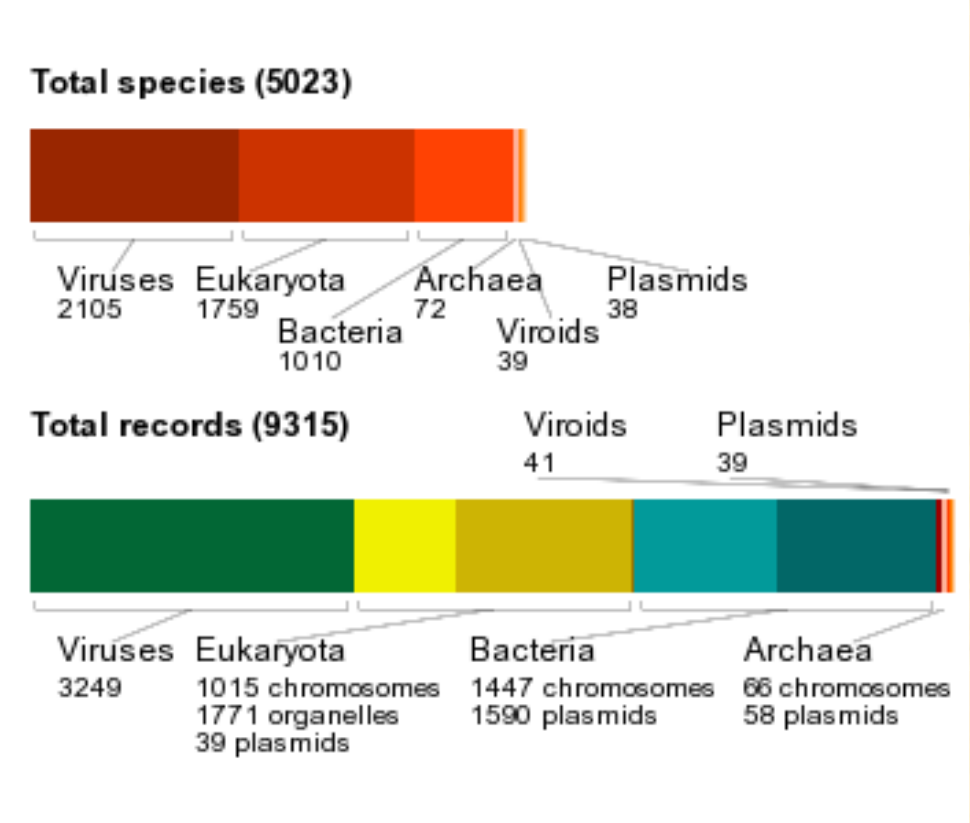
- **Databáze:**

  Primární

  Sekundární

  Strukturní

Genomové zdroje



Rozdělení molekulárně biologických databází

- **Databáze:**

  Specializované

  Univerzální

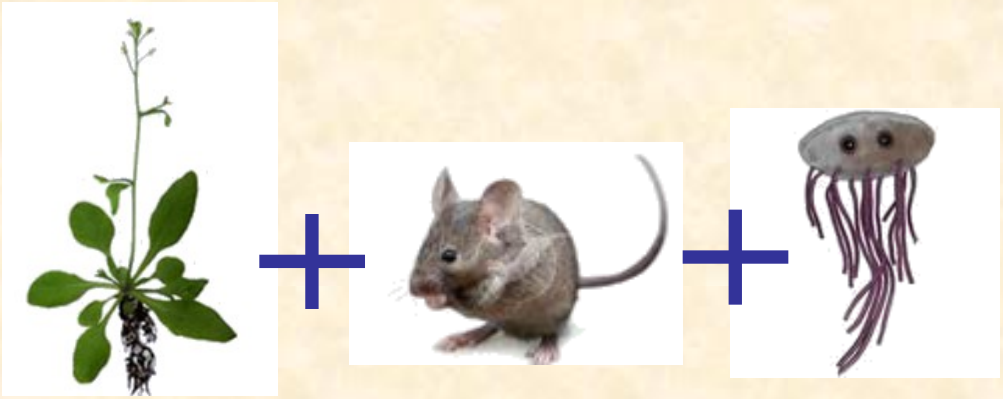
Specializované databáze obsahují informace o určité proteinové rodině nebo skupině proteinů, případně o určitém organismu.

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.

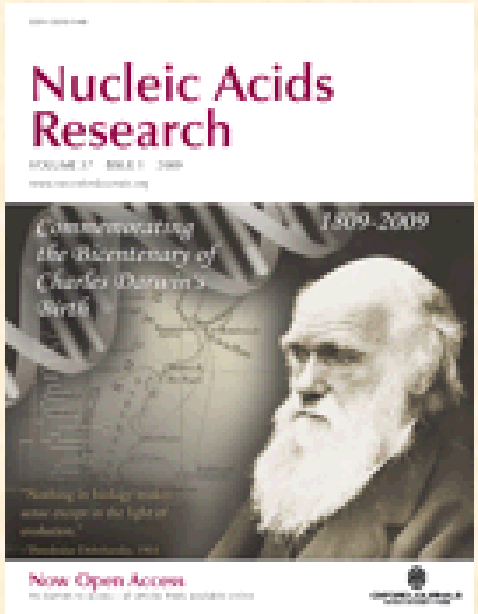
  Eliminace redundantních dat.

  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

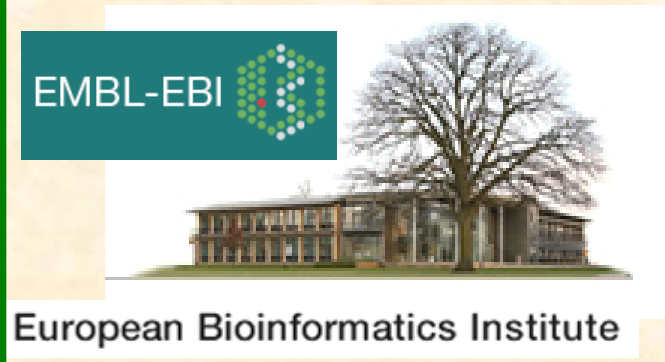
1170 databází

EBI/NCBI/CIB

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

EBI

Evropský institut pro bioinformatiku



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

NCBI

Národní centrum pro biotechnologické informace

National Center for Biotechnology Information

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

CIB

Centrum pro informační biologii

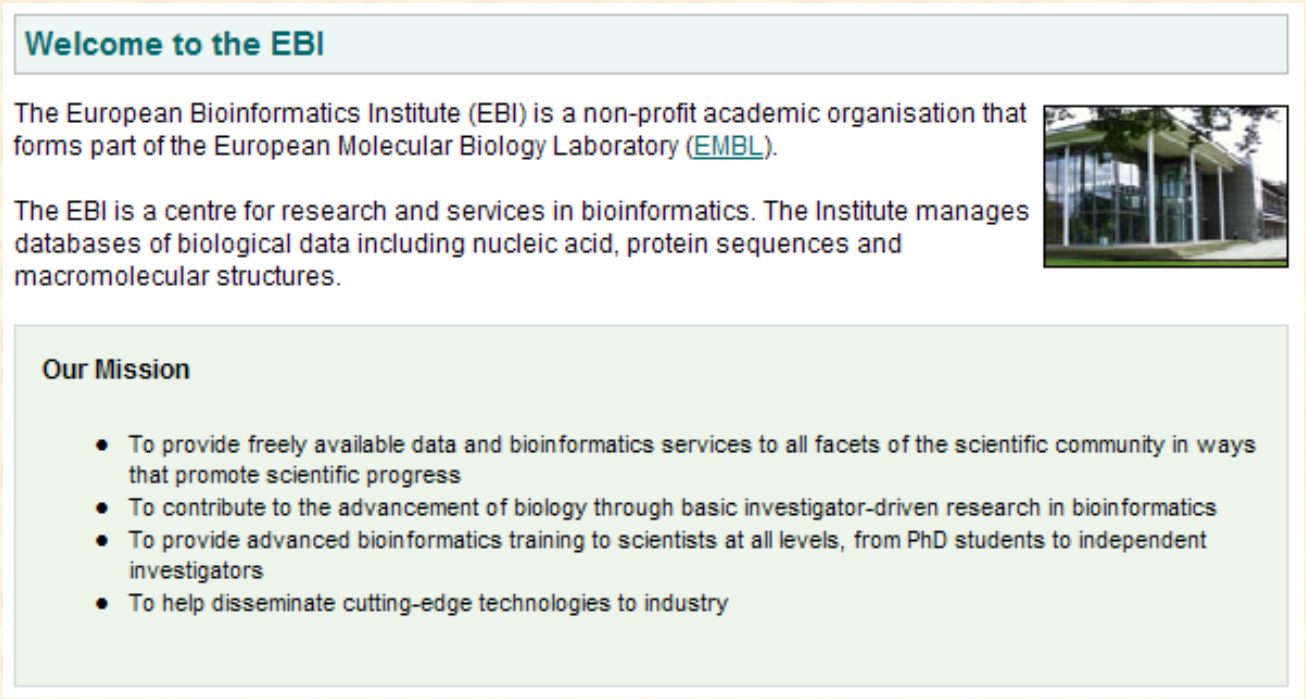
Center for Information Biology

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.**



Welcome to the EBI

The European Bioinformatics Institute (EBI) is a non-profit academic organisation that forms part of the European Molecular Biology Laboratory (EMBL).

The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures.

Our Mission

- To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress
- To contribute to the advancement of biology through basic investigator-driven research in bioinformatics
- To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators
- To help disseminate cutting-edge technologies to industry

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 Hea**

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/

Information/Database

- DNA Data Bank of Japan
- National BioResource Project - Information Site
- WFCC-MIRCEN World Data Centre for Microorganisms
- Genetic Resources Database (SHIGEN)
- Nematode Gene Expression Database
- Mouse Microsatellite Database
- Rice Genome Database (Oryzabase)
- E.coli Genome Database (PEC)

- Virtual Museum of Genetics(Japanese)
- DNA Data Bank of Japan
- National BioResource Project
- Genome Netwaork Project

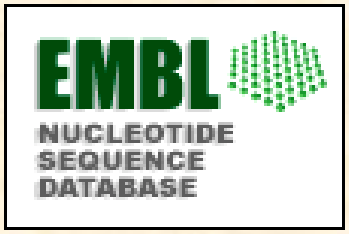
Primární databáze NA

- **EMBL - Evropský institut pro bioinformatiku**

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

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

EMBL



- **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.

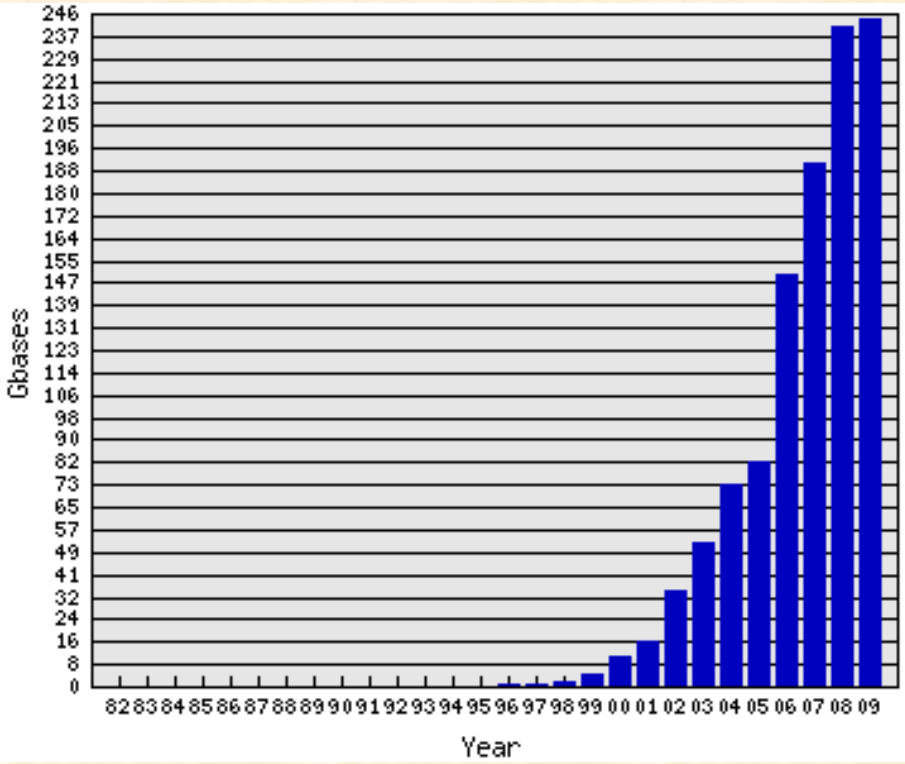
This morning = 21.1.2009

EMBL-EBI

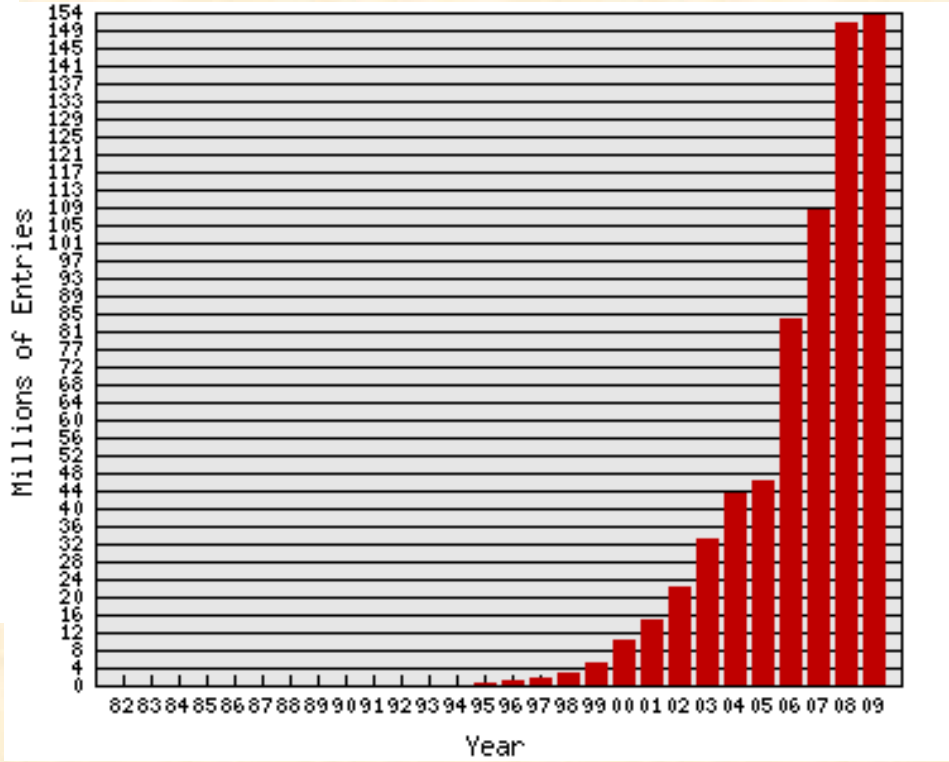
EMBL

EMBL NUCLEOTIDE SEQUENCE DATABASE

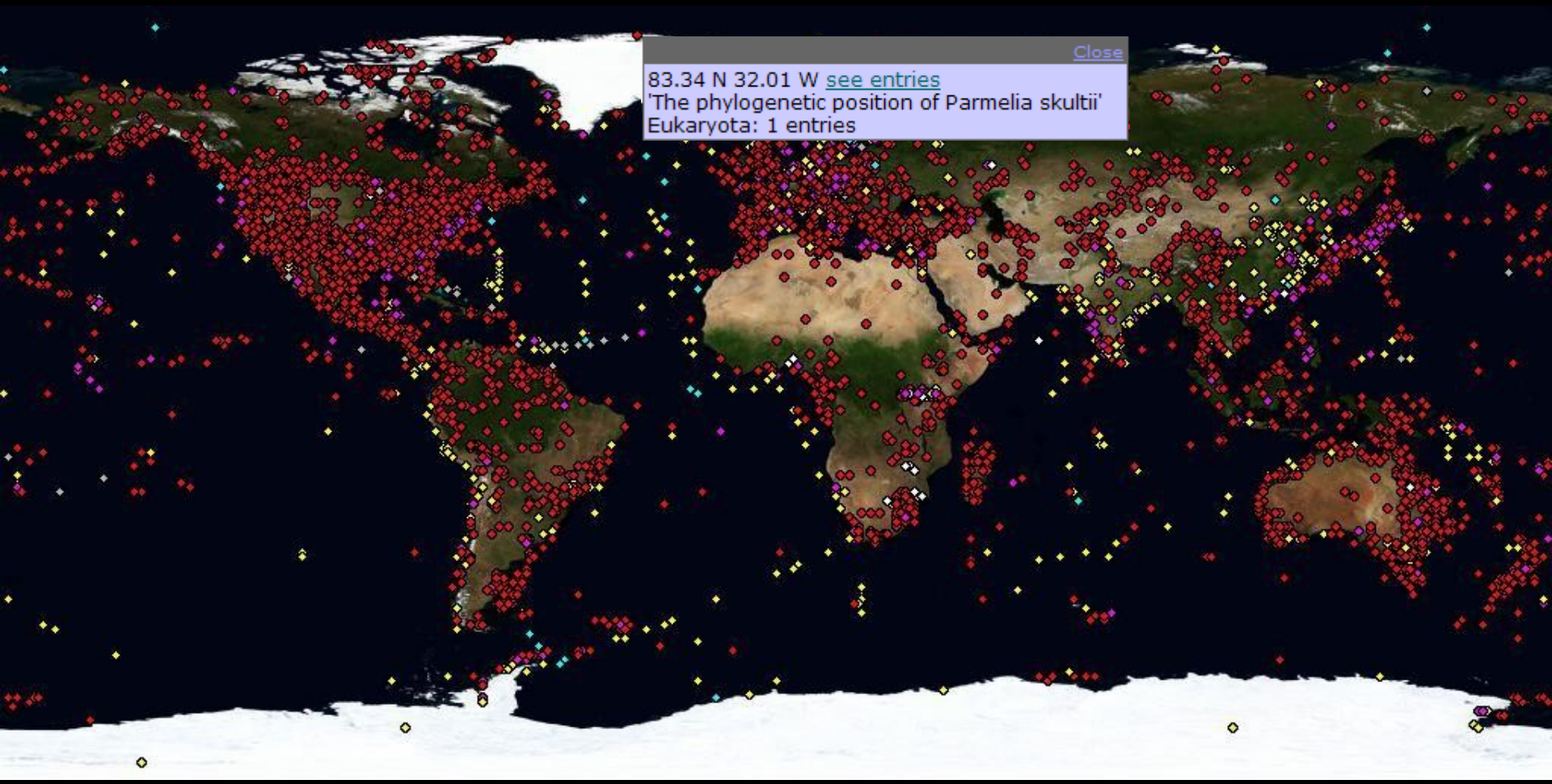
Total nucleotides
(current 244,322,213,780)



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; Trifolieae; 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="TRE361"
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="HSSP:P26205"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
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 ctgtttgtta ttagctcatt        60
cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
ttggccaaga atactcccaa aggggaaagtt gagcggaggc ataaatcacg aaggaatcaa      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 Thoughput CDNA sequencing
HTG	High Thoughput Genome sequencing
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TPA	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.

NCBI 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

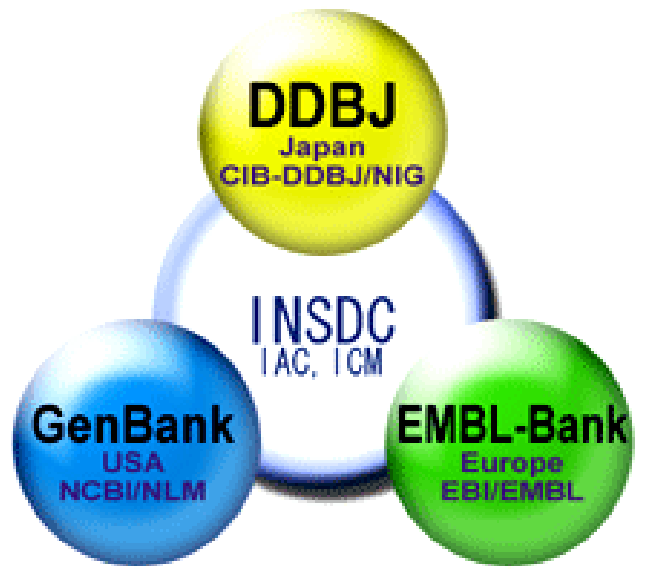
http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

DDBJ
DNA Data Bank of Japan

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



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

http://www.insdc.org/

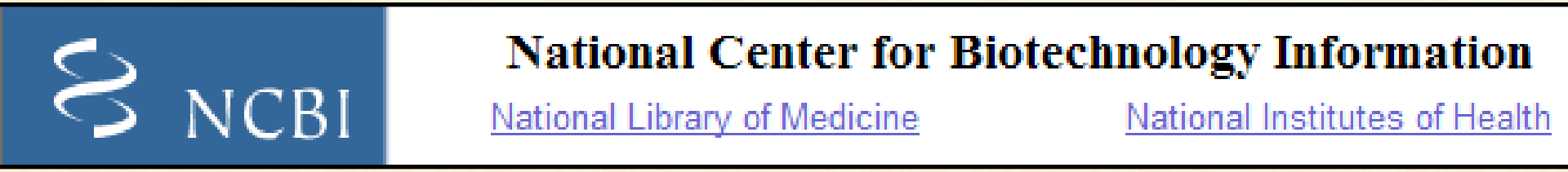
Primární databáze proteinů

- **Univerzální databáze:**

  „Skladiště“ sekvencí – sequence repository

  Manuálně spravovaná – curated database

Příklad: **GenBank *versus* RefSeq**



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 úrověň 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.

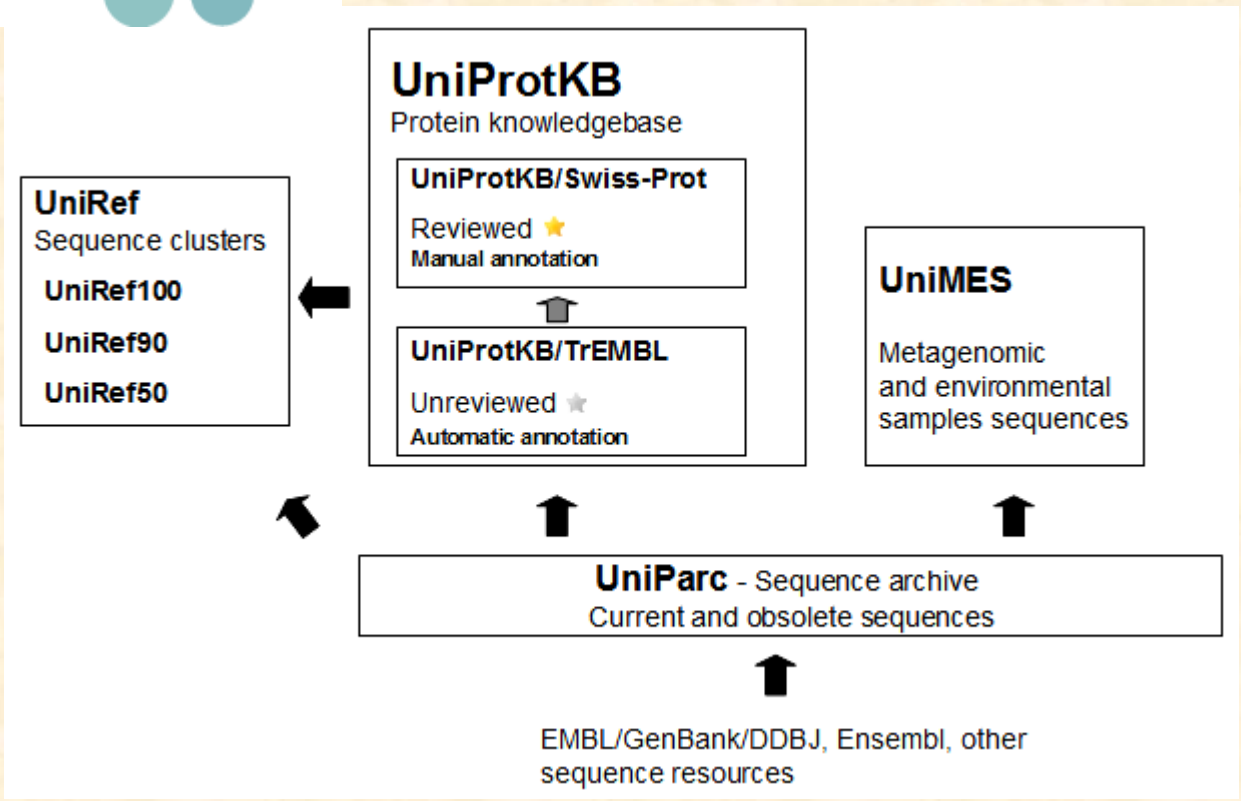
  Eliminace redundantních dat.

  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.**

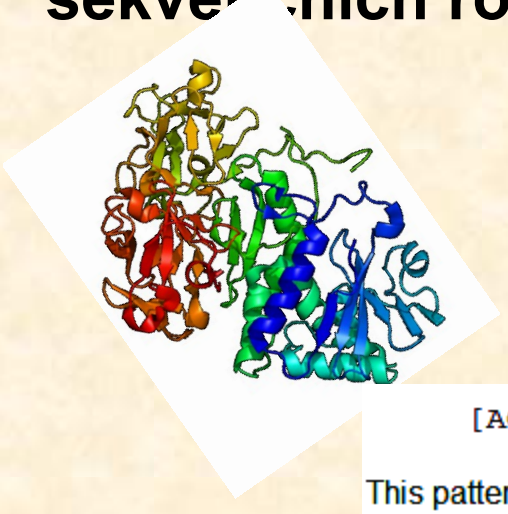
Sekundární databáze NA a proteinů

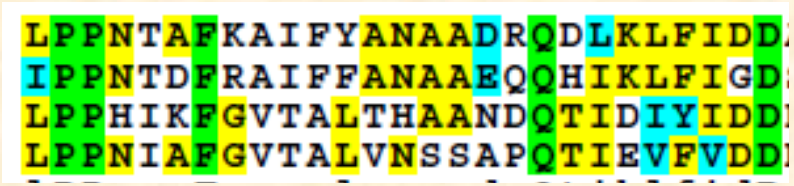
- **Databáze mohou obsahovat:**

  Proteinové DOMÉNY odvozené ze známých struktur

  Proteinové sekvence seřazené do SEKVENČNÍCH RODIN

  CHARAKTERISTICKÉ MOTIVY odvozené z těchto sekvenčních rodin.






```
[AC]-x-V-x(4)-{ED}.
```

This pattern is translated as: [Ala 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/

Table View | Raw Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

SEQUENCE: Sequence_1 CRC64: B08AB341813AD2EE LENGTH: 382 aa

InterPro IPR000772 Domain

Ricin B lectin

PF00652		Ricin_B_lectin
SM00458		RICIN
PS50231		RICIN_B_LECTIN

InterPro IPR001574 Family

Ribosome-inactivating protein

PF00161		RIP
SSF56371		Ribosome_inactivat_prot

InterPro IPR008997 Domain

Ricin B-related lectin

SSF50370		RicinB_like

InterPro IPR016139 Domain

Ribosome-inactivating protein, subdomain 2

G3DSA:4.10.470.10		Ribosome_inactivat_prot_sub2

InterPro IPR017989 Family

Ribosome-inactivating protein subgroup

PR00396		SHIGARICIN

Sekundární databáze NA a proteinů

- **Sekundární proteinové databáze:**

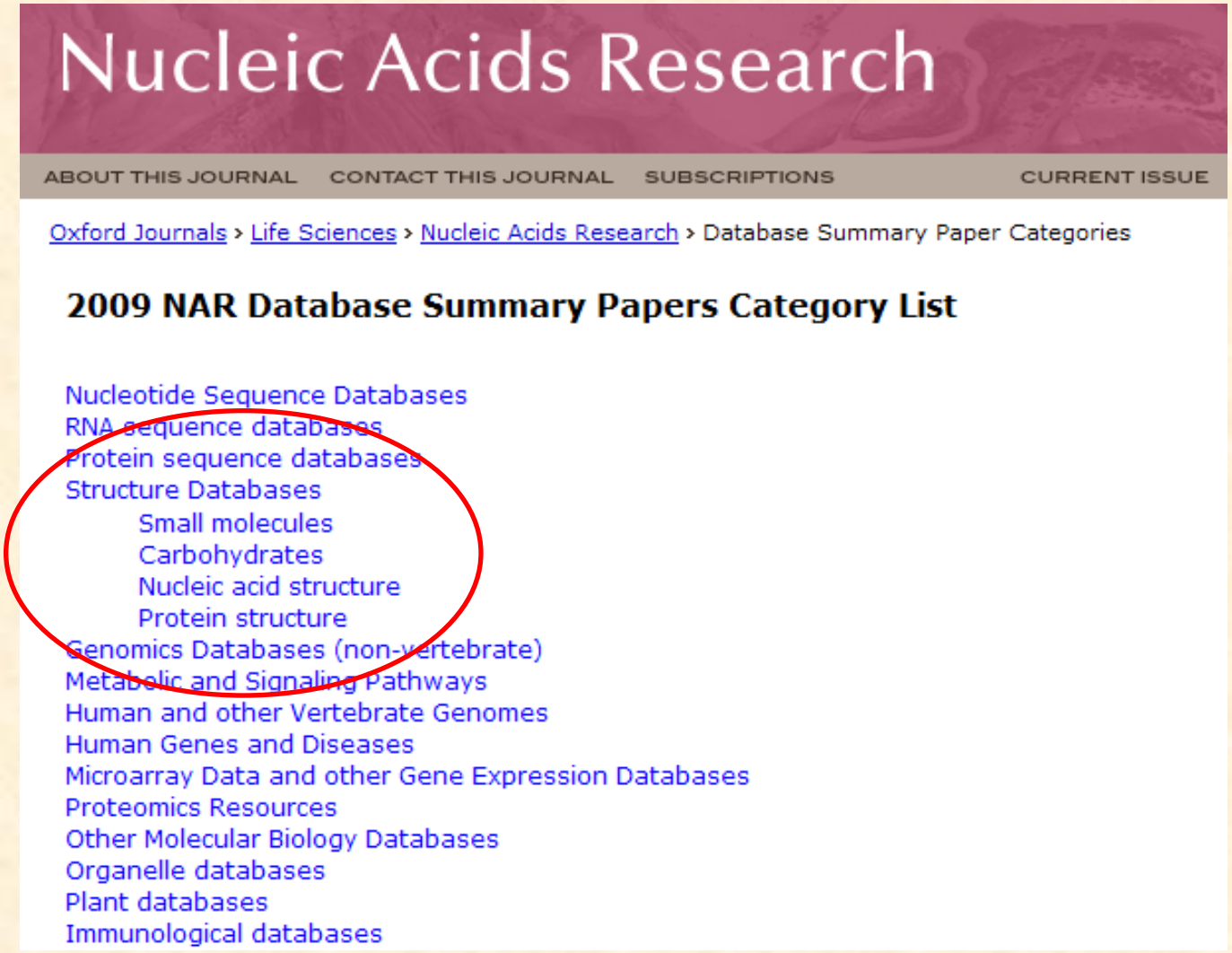
  PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMS

  V současné době sdruženy do integrované klasifikační databáze proteinů InterPro

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

- **Sekundární databáze NA**
  TRANSFAC

Strukturní databáze



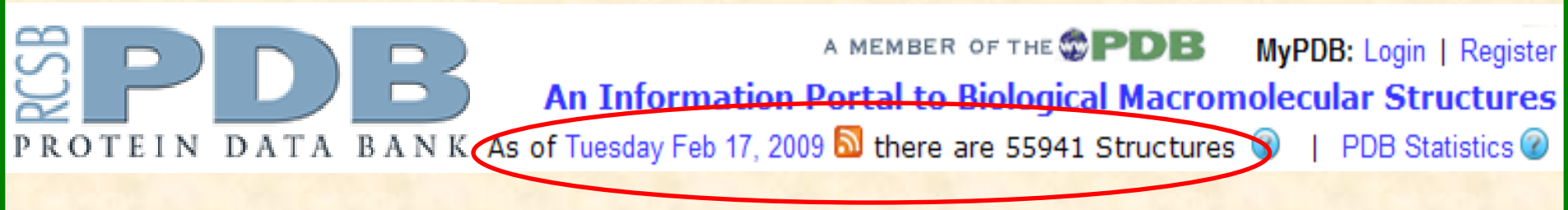
http://www3.oup.co.uk/nar/database/a/

Strukturní databáze proteinů

2009 NAR Database Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
- Small molecules
- Carbohydrates
- Nucleic acid structure
- Protein structure
  - 3D-Genomics
  - 3DID - 3D interacting domains
  - ArchDB
  - AS-ALPS
  - ASTRAL
  - AutoPSI
  - BANMOKI
  - BioMagResBank
  - CADB - Conformational Angles DataBase of Proteins
  - CATH
  - CC+
  - CE
  - CoC Central
  - ColiSNP
  - Columba
  - ConSurf-DB
  - CPDB
  - CSA - Catalytic Site Atlas
  - DisProt - Database of Protein Disorder
  - DMAPS
  - Dockground
  - DomIns - Database of Domain Insertions
  - DSDBASE - Disulfide Database
  - DSMM - a Database of Simulated Molecular Motions
  - E-MSD - EBI-Macromolecular Structure Database
  - eF-site - Electrostatic surface of Functional site
  - EzCatDB
  - FireDB
  - FSN
  - Gene3D
  - Genomic Threading Database
  - GTOP - Genomes To Protein structures
  - HOMSTRAD - Homologous Structure Alignment Database
  - HotSprint
  - IMGT/3Dstructure-DB
  - IMOTdb
  - JAIL
  - Jenalib: Jena Library of Biological Macromolecules
  - KineticDB
  - LPFC
  - MALISAM
  - MegaMotifbase
  - MMDB
  - ModBase
  - MolMovDB - Database of Macromolecular Movements
  - PASS2
  - PDB
  - PDB-REPRDB
  - PDBselect
  - PDBsum

PDB - Protein Data Bank



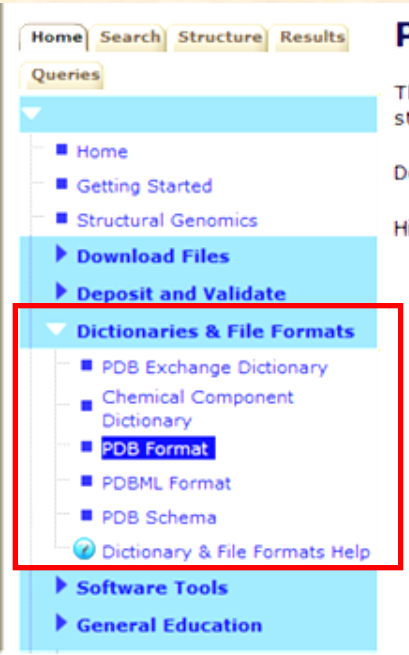
- **Databáze obsahuje experimentálně získané struktury proteinů, nukleových kyselin a komplexů informačních biomakromolekul.**

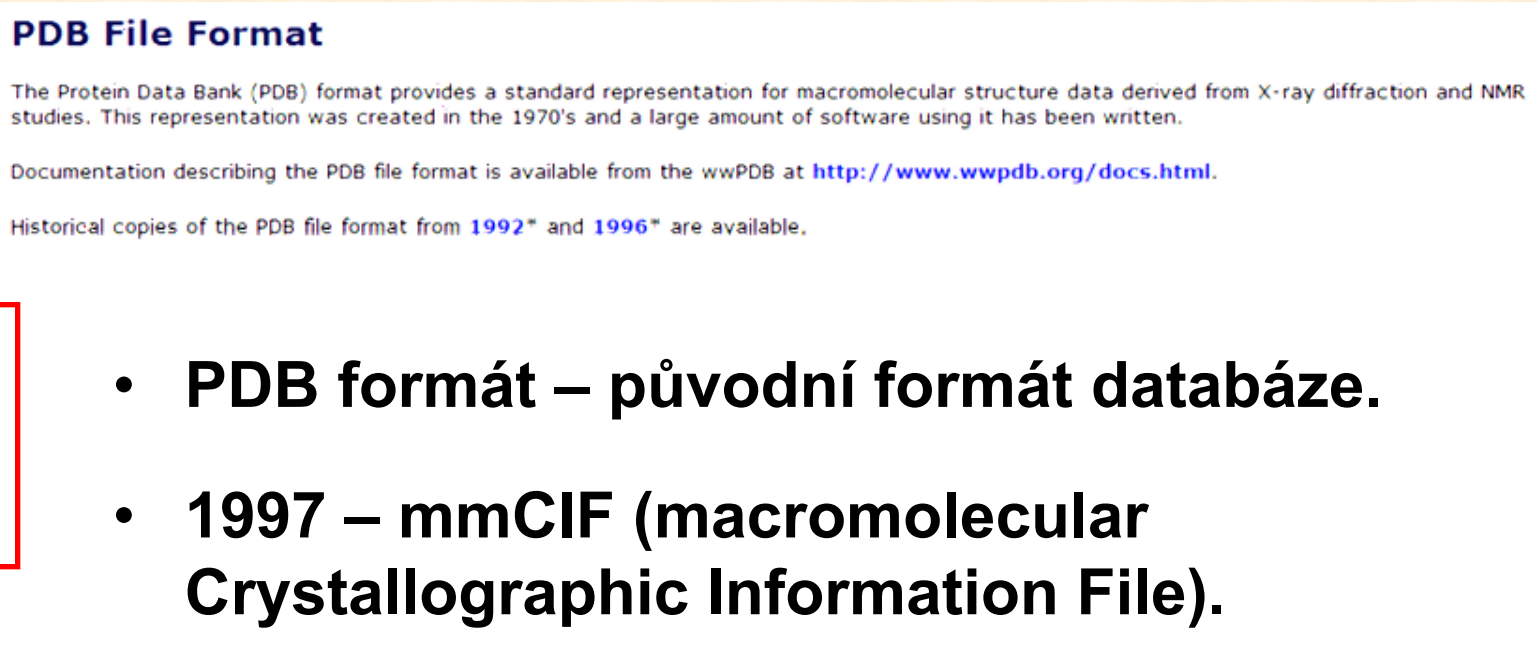
PDB Current Holdings Breakdown

		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.**

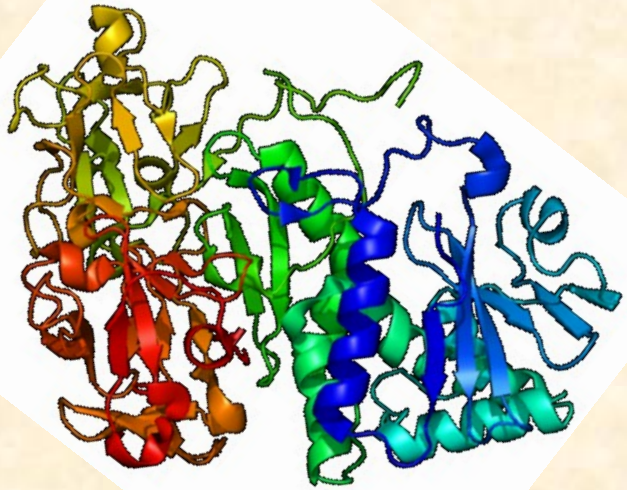
PDB formát

The ATOM records present the atomic coordinates for standard amino acids and nucleotides. They also present the occupancy and temperature factor for each atom. Non-polymer chemical coordinates use the HETATM record type. The element symbol is always present on each ATOM record; charge is optional.

Changes in ATOM/HETATM records result from the standardization atom and residue nomenclature. This nomenclature is described in the Chemical Component Dictionary (ftp://ftp.wwpdb.org/pub/pdb/data/monomers).

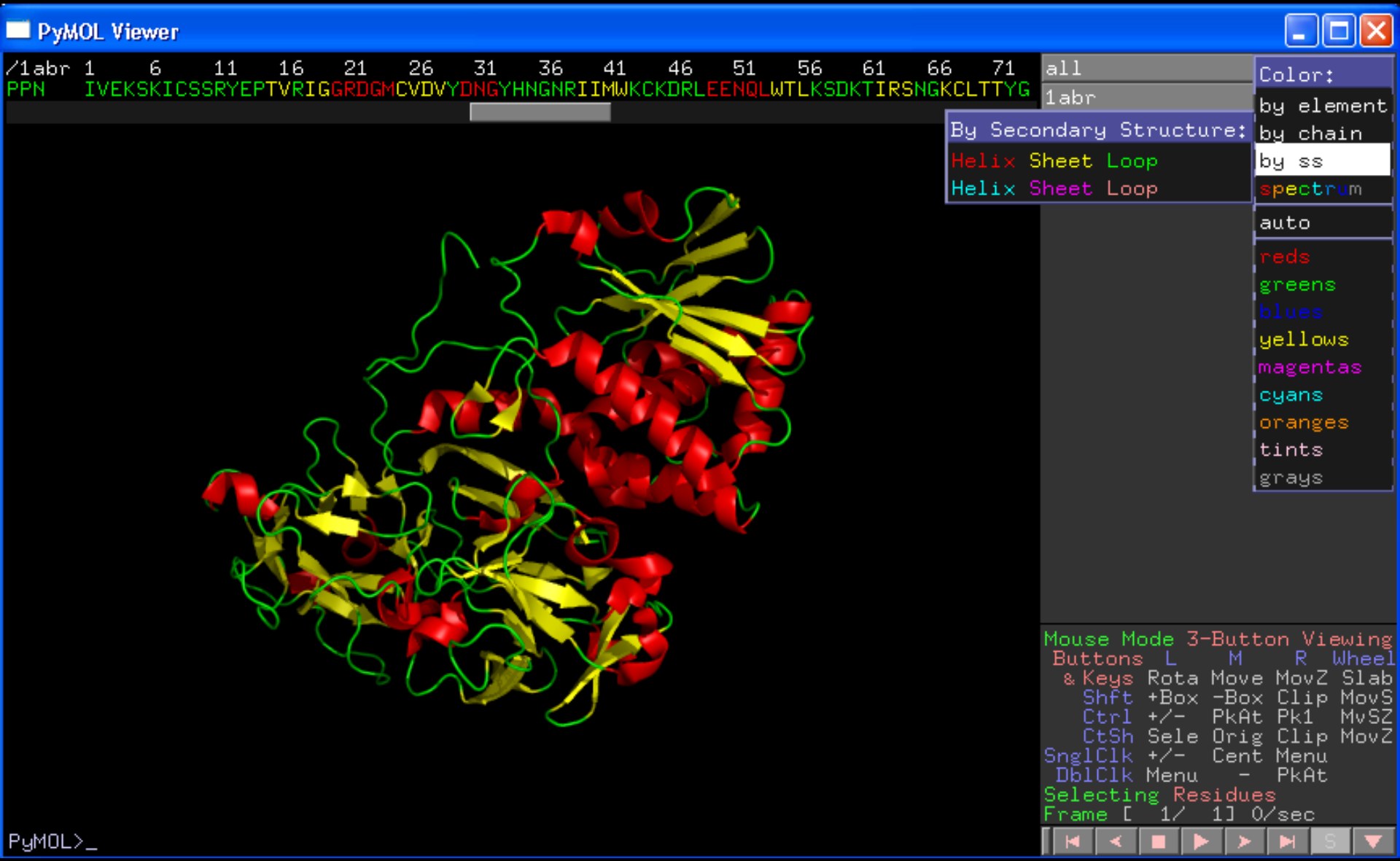
Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Ortho
47 - 54	Real(8.3)	z	Ortho
55 - 60	Real(6.2)	occupancy	Occup
61 - 66	Real(6.2)	tempFactor	Tempe
77 - 78	LString(2)	element	Eleme
79 - 80	LString(2)	charge	Charg



Abrin

```
ATOM      2  CA  GLU A   1      64.373  11.709  60.583  1.00 79.99           C
ATOM      3  C   GLU A   1      63.512  10.438  60.597  1.00 79.31           C
ATOM      4  O   GLU A   1      63.540   9.685  61.574  1.00 79.23           O
ATOM      5  CB  GLU A   1      63.805  12.754  59.603  1.00 79.36           C
ATOM      6  CG  GLU A   1      62.880  13.819  60.228  1.00 78.52           C
ATOM      7  CD  GLU A   1      61.525  13.275  60.676  1.00 78.50           C
ATOM      8  OE1 GLU A   1      60.915  12.482  59.923  1.00 77.14           O
ATOM      9  OE2 GLU A   1      61.064  13.659  61.776  1.00 77.48           O
ATOM     10  H1  GLU A   1      66.078  10.648  60.914  1.00 20.00           H
ATOM     11  H2  GLU A   1      65.776  10.893  59.265  1.00 20.00           H
ATOM     12  H3  GLU A   1      66.387  12.177  60.222  1.00 20.00           H
```


PyMOL Viewer
/1abr 1 6 11 16 21 26 31 36 41 46 51 56 61 66 71
PPN IVEKSKICSSRYEPTVRIGGRDGMCVDVYDNGYHNGNRIIMWKCKDRLEENQLWTLKSDKTIRSNGKCLTTYG
all
1abr
By Secondary Structure:
Helix Sheet Loop
Helix Sheet Loop
Color:
by element
by chain
by ss
spectrum
auto
reds
greens
blues
yellows
magentas
cyans
oranges
tints
grays
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
Frame [1/ 1] 0/sec
PyMOL>_

Strukturní databáze NA

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE

Oxford Journals › Life Sciences › Nucleic Acids Research › Database Summary Paper

2009 NAR Database Summary Papers

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
  - Small molecules
  - Carbohydrates
  - Nucleic acid structure
    - Greglist
    - GRSDB
    - ITS2
    - MeRNA
    - NCIR - Non-Canonical Interactions in RNA
    - NDB
    - NTDB
    - QuadBase
    - Rfam
    - RNA FRABASE
    - RNA SSTRAND
    - RNAJunction
    - SARS-CoV RNA SSS
    - SCOR - Structural Classification Of RNA
    - Vir-Mir db
  - Protein structure

NDB - Nucleic Acid Database

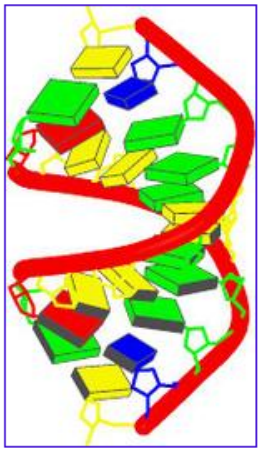
WELCOME TO THE
NUCLEIC ACID DATABASE
a repository of three-dimensional structural information about nucleic acids

Number of Released Structures:
4089 Structures
Last Update: 15-Jan-2009

NDB ID: AD0001

NMR Atlas X-Ray Atlas

Title:	STRUCTURE OF A DNA IN LOW SALT CONDITIONS D (GACCGCGGTC)
Molecular Description:	5'-D(GpApCpCpGpCpGpGpTpC)-3'
Structural Features:	A DOUBLE HELIX
Nucleic Acid Sequence:	Chain A: (DG) (DA) (DC) (DC) (DG) (DC) (DG) (DG) (DT) (DC)
Primary Citation:	Finley, J.B., Luo, M. X-ray crystal structures of half the human papilloma virus E2 binding site: d(GACCGCGGTC). *Nucleic Acids Res.* , **26**, pp. 5719 - 5727, 1998.
Experimental Information:	X-RAY DIFFRACTION
Space Group:	P 6_1 2 2
Cell Constants:	a = 38.444 b = 38.444 c = 80.175 (Ångstroms) α = 90.00 β = 90.00 γ = 120.00 (degrees)
Crystallization Conditions:	Method: VAPOR DIFFUSION Drop: WATER, MPD, Spermine HCl, Na Cacodylate Reservoir: WATER, MPD
Refinement:	The structure was refined using the X-PLOR program. The R value is 21.9 for 1779 reflections in the resolution range 5.000 to 2.200 Ångstroms with Fobs > 3.000 sigma(Fobs).



Biological Unit 1

Other Views
Asymmetric Unit
Crystal Packing
Enlarge Biological Unit 1

http://ndbserver.rutgers.edu/

# of Nucleic Acid Containing Structures

X-Ray Atlas

- Gallery Index
- Index Listing [text only]

NMR Atlas

- Gallery Index
- Index Listing [text only]
- Sorted Galleries
- Musical Atlas
- About this Atlas

The NDB Atlas provides summary information and images for each structure in the database. These images provide many looks at the varied structures of nucleic acids.

The Atlas is first divided by experimental type, and then by structure type. Features include:

- images of the asymmetric and biological units, and crystal packing pictures for nucleic acid structures from X-ray crystallographic experiments
- images of the average and ensemble structure from NMR experiments
- links to coordinate files, experimental data files
- tables of derived data, including torsion angles and hydrogen bonding classifications
- special features for RNA structures, including images of secondary and tertiary structure

A more detailed description of the NDB Atlas features is available at "About this Atlas"

Genomové zdroje

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE

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

2009 NAR Database Summary Papers Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
MGD - Mouse Genome Database
TIGR Gene Indices
Genome annotation terms, ontologies and nomenclature
Taxonomy and identification
General genomics databases
Viral genome databases
Prokaryotic genome databases
Unicellular eukaryotes genome databases
Fungal genome databases
Invertebrate genome databases

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

A service of the U.S. National Library of Medicine and the National Institutes of Health

PubMed is a service of the U.S. National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles back to 1948. PubMed includes links to full text articles and other related resources.

NCBI

Entrez, The Life Sciences Search Engine

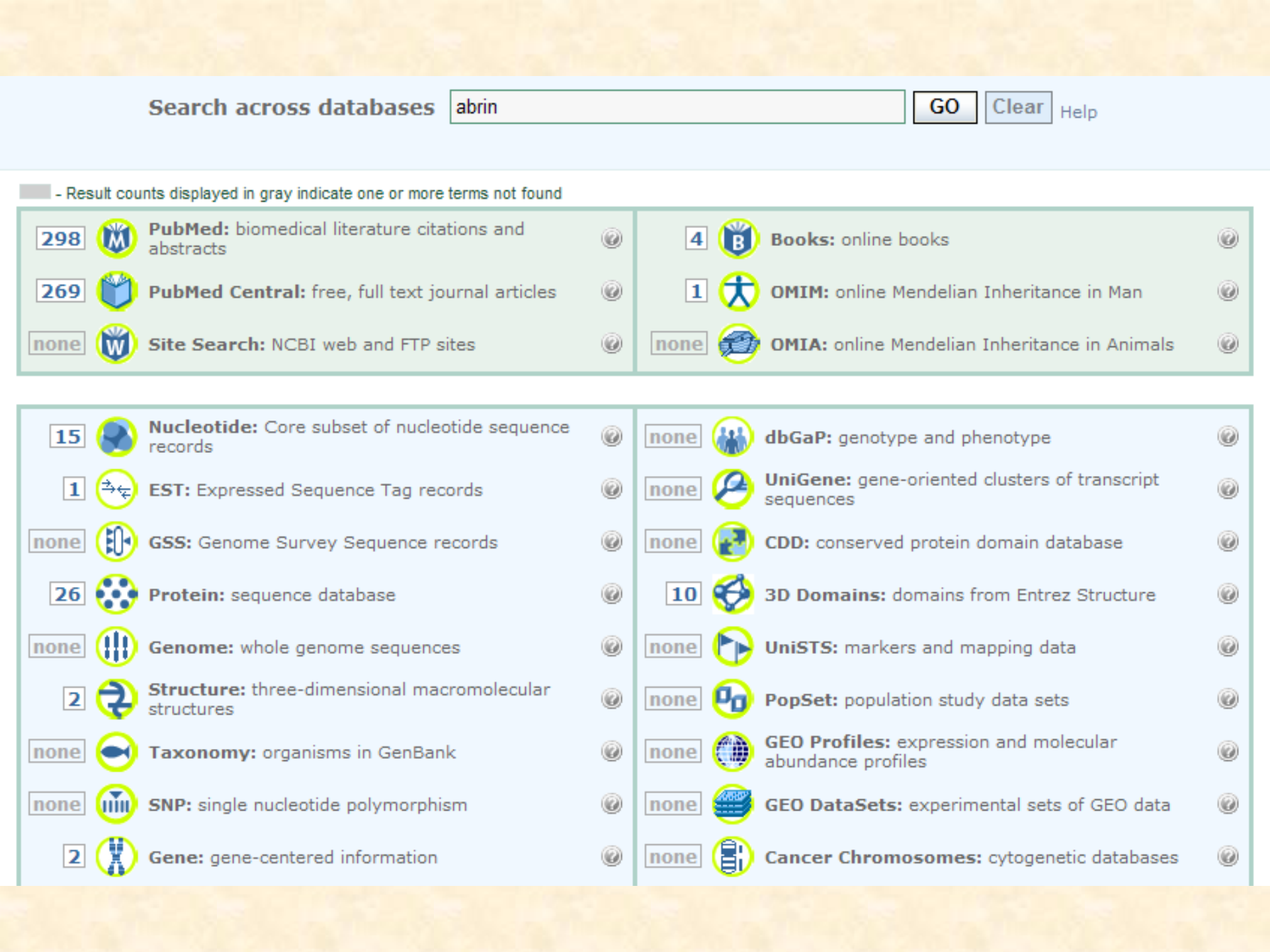
HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

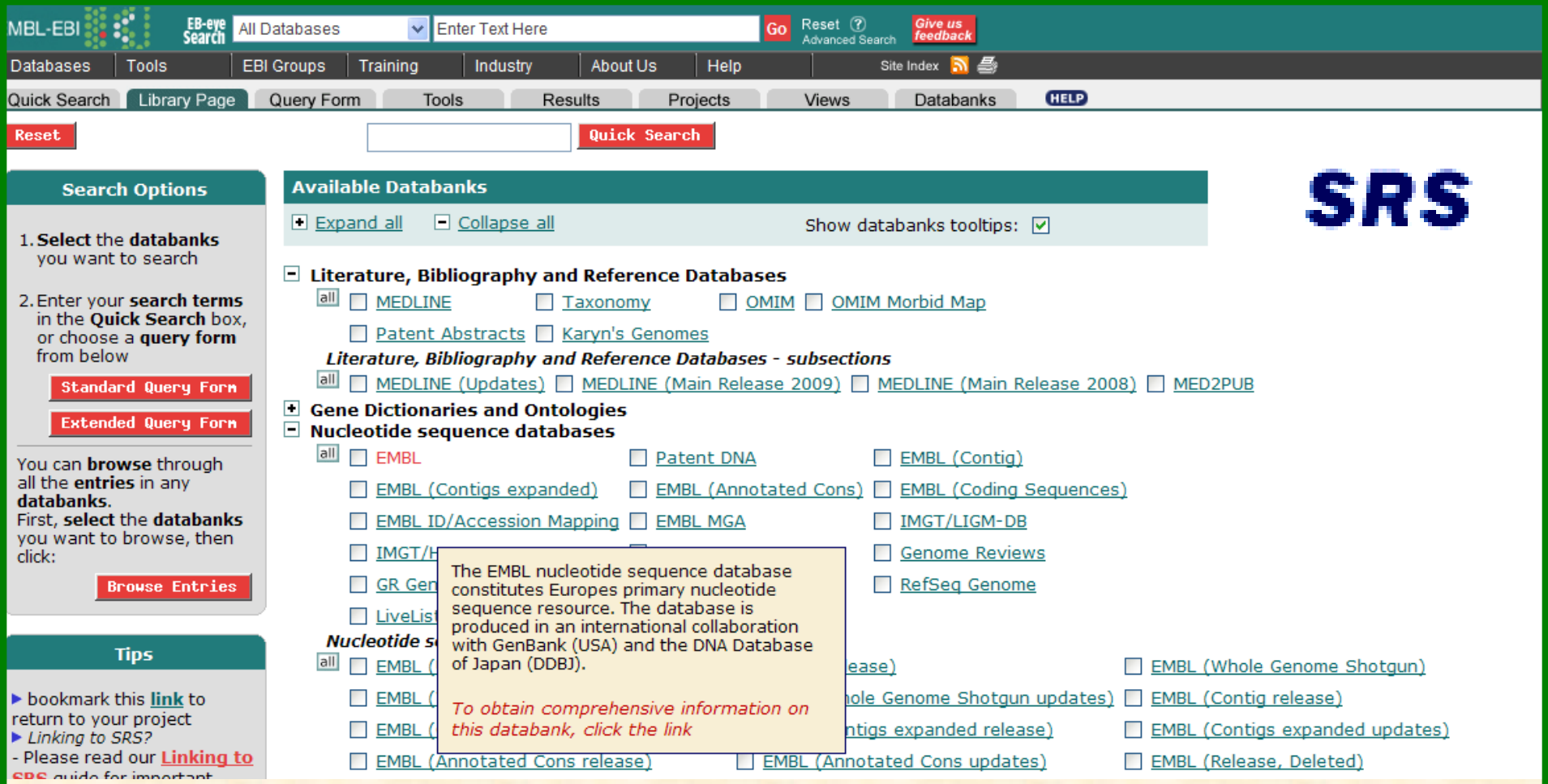
Search across databases [ ] GO Clear Help

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

- **Nucleotide:** 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


Search across databases
abrin
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
298 PubMed: biomedical literature citations and abstracts
269 PubMed Central: free, full text journal articles
none Site Search: NCBI web and FTP sites
4 Books: online books
1 OMIM: online Mendelian Inheritance in Man
none OMIA: online Mendelian Inheritance in Animals
15 Nucleotide: Core subset of nucleotide sequence records
1 EST: Expressed Sequence Tag records
none GSS: Genome Survey Sequence records
26 Protein: sequence database
none Genome: whole genome sequences
2 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
none SNP: single nucleotide polymorphism
2 Gene: gene-centered information
none dbGaP: genotype and phenotype
none UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
10 3D Domains: domains from Entrez Structure
none UniSTS: markers and mapping data
none PopSet: population study data sets
none GEO Profiles: expression and molecular abundance profiles
none GEO DataSets: experimental sets of GEO data
none Cancer Chromosomes: cytogenetic databases



- **Textové vyhledávání v databázích**

EBI– SRS

Sequence Retrieval System

http://srs.ebi.ac.uk/

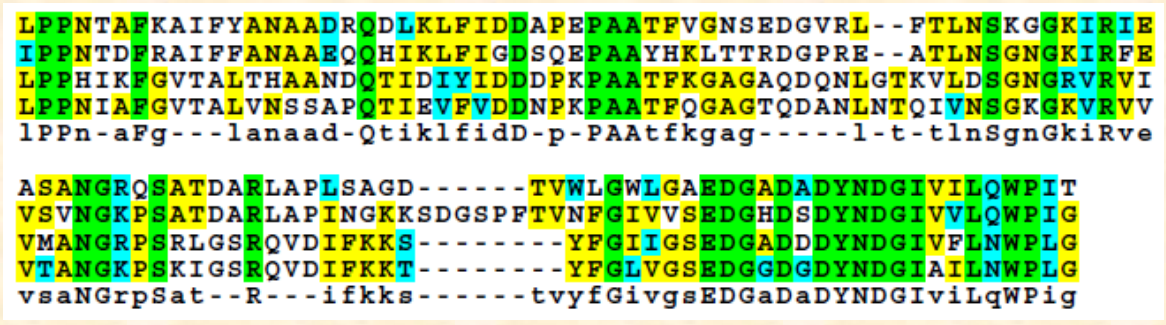
Vyhledávací systémy

- **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