Bioinformatika és genomanalízis az orvostudományban

Biológiai adatbázisok

Cserző Miklós

2020

https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdlUwc3B4eDFCZkNKQT09

A mai előadás

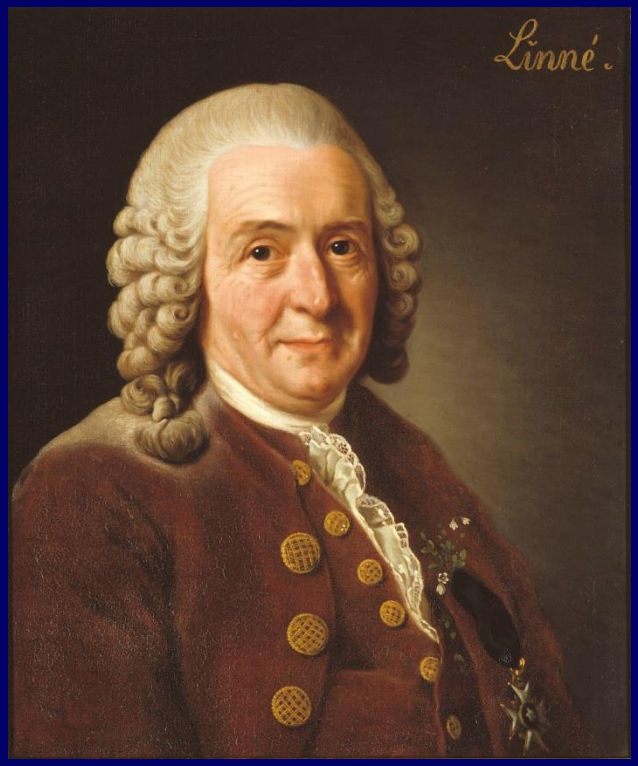
- Mi az adatbázis
- A biológia kapcsolata az adatbázisokkal
- Az adatbázisok típusai
- Adatbázis formátumok, adatbázisok szerkezete
- Néhány fontos biológiai adatbázis
- Hibák az adatbázisokban

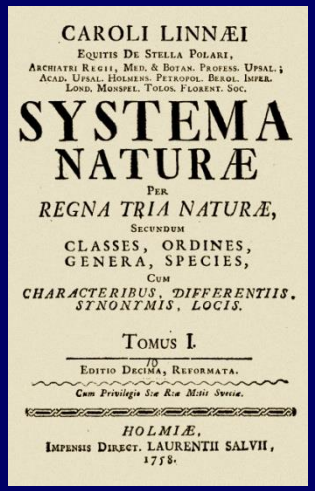
Mi az adatbázis

- Nyers adat – információ
- Adatbázis – struktúrált információ
- A struktúrának köszönhetően
  - Kereshető
  - Rendezhető
  - Szűrhető

A biológia és az adatbázisok

- A XVIII. század – világutazók kora
- Rengeteg új faj felfedezése, leírása
- Rendszer nékül áttekinthetetlen az adattömeg
- Carl von Linne rendszertana




CAROLI LINNÆI
EQUITIS DE STELLA POLARI,
ARCHIATRI REGII, MED. & BOTAN. PROFESS. UPSAL.;
ACAD. UPSAL. HOLMENS. PETROPOL. BEROL. IMPER.
LOND. MONSPEL. TOLOS. FLORENT. SOC.

SYSTEMA NATURÆ

PER
REGNA TRIA NATURÆ,
SECUNDUM
CLASSES, ORDINES,
GENERA, SPECIES,
CUM
CHARACTERIBUS, DIFFERENTIIS,
SYNONYMIS, LOCIS.

TOMUS I.

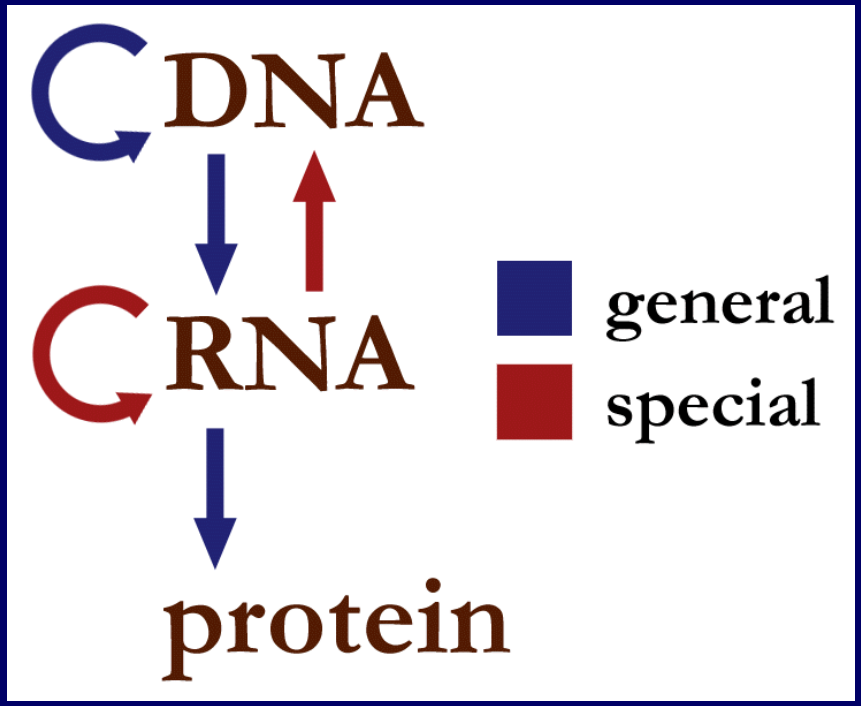
EDITIO DECIMA, REFORMATA.

Cum Privilegio Sæ Reæ Mtis Sveciæ.

HOLMIÆ,
IMPENSIS DIRECT. LAURENTII SALVII,
1758.

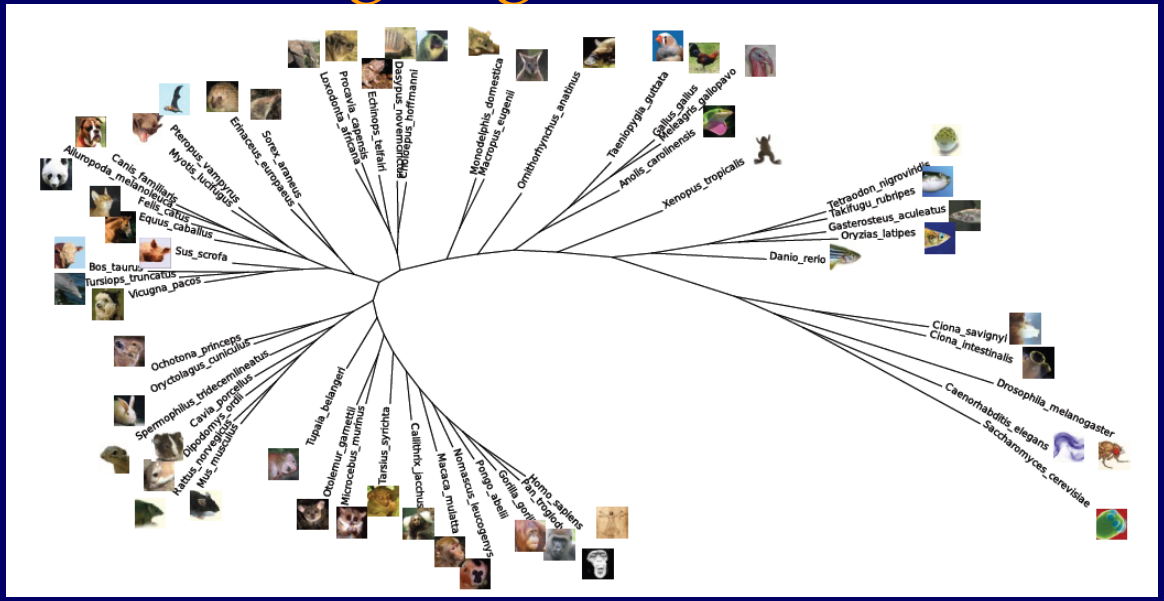
A molekuláris biológia központi dogmája

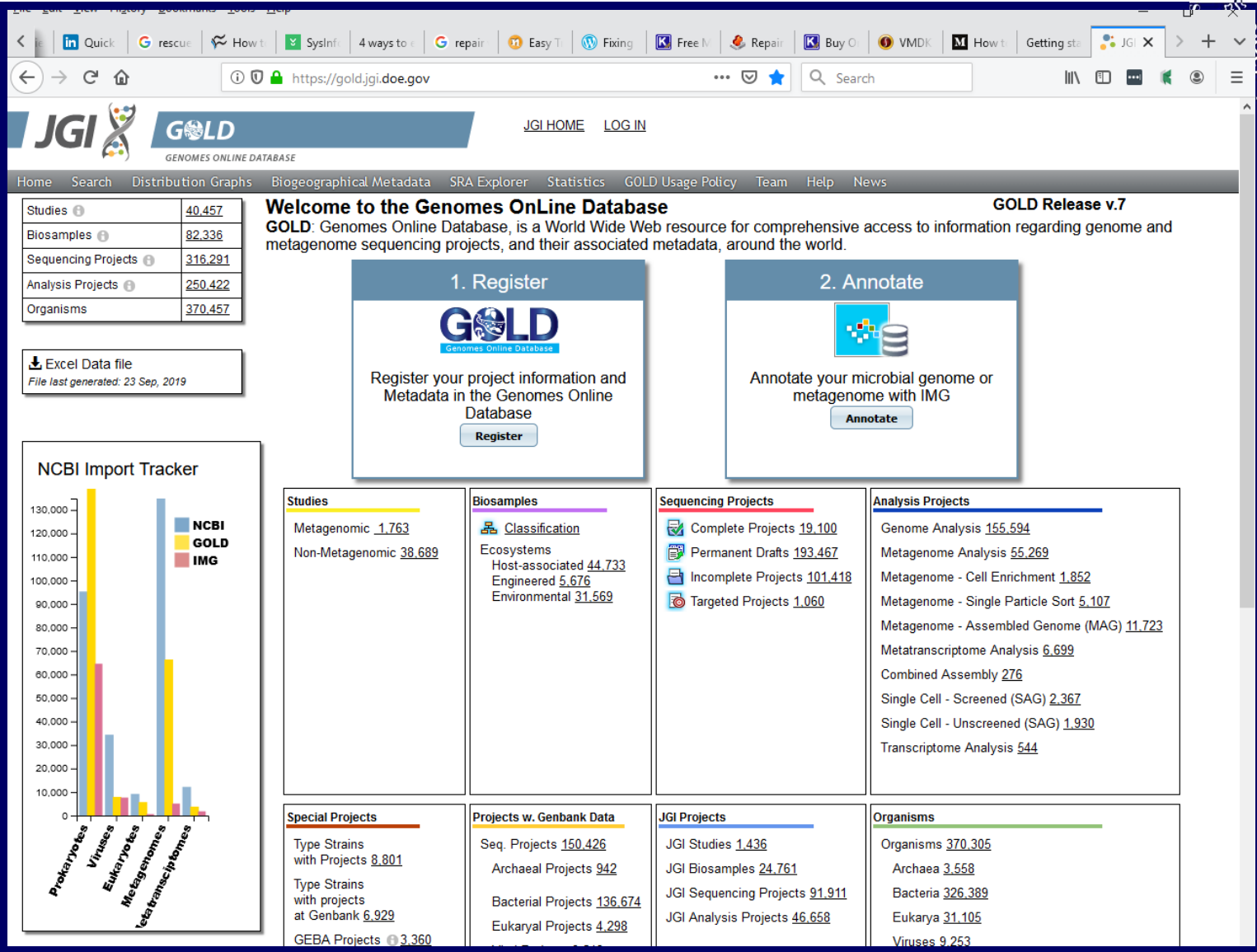
- DNS – az információt a szekvenciájában tárolja
- RNS – az információt közvetíti
- Fehérje – a szekvenciális információ végső megjelenése, az információt használja

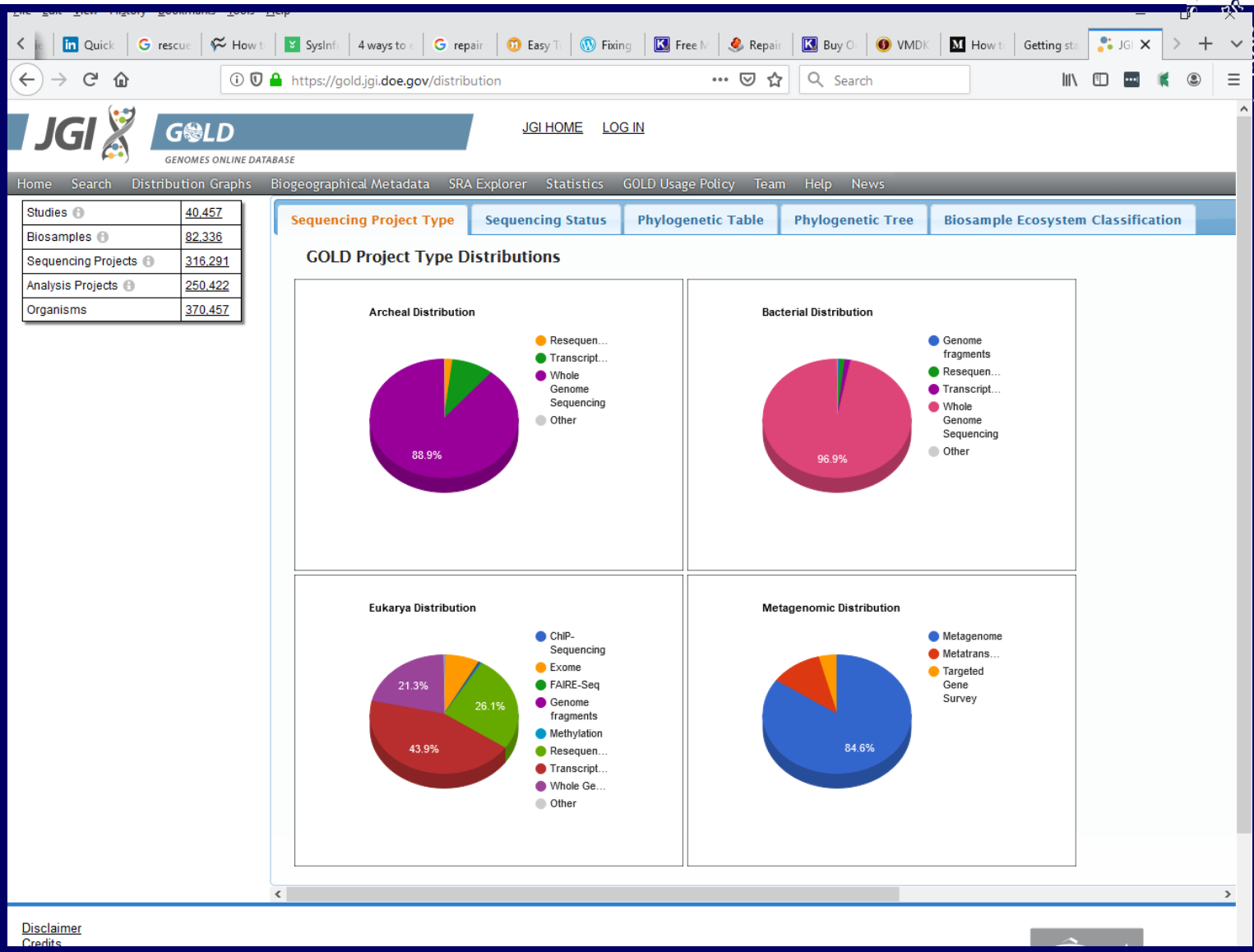


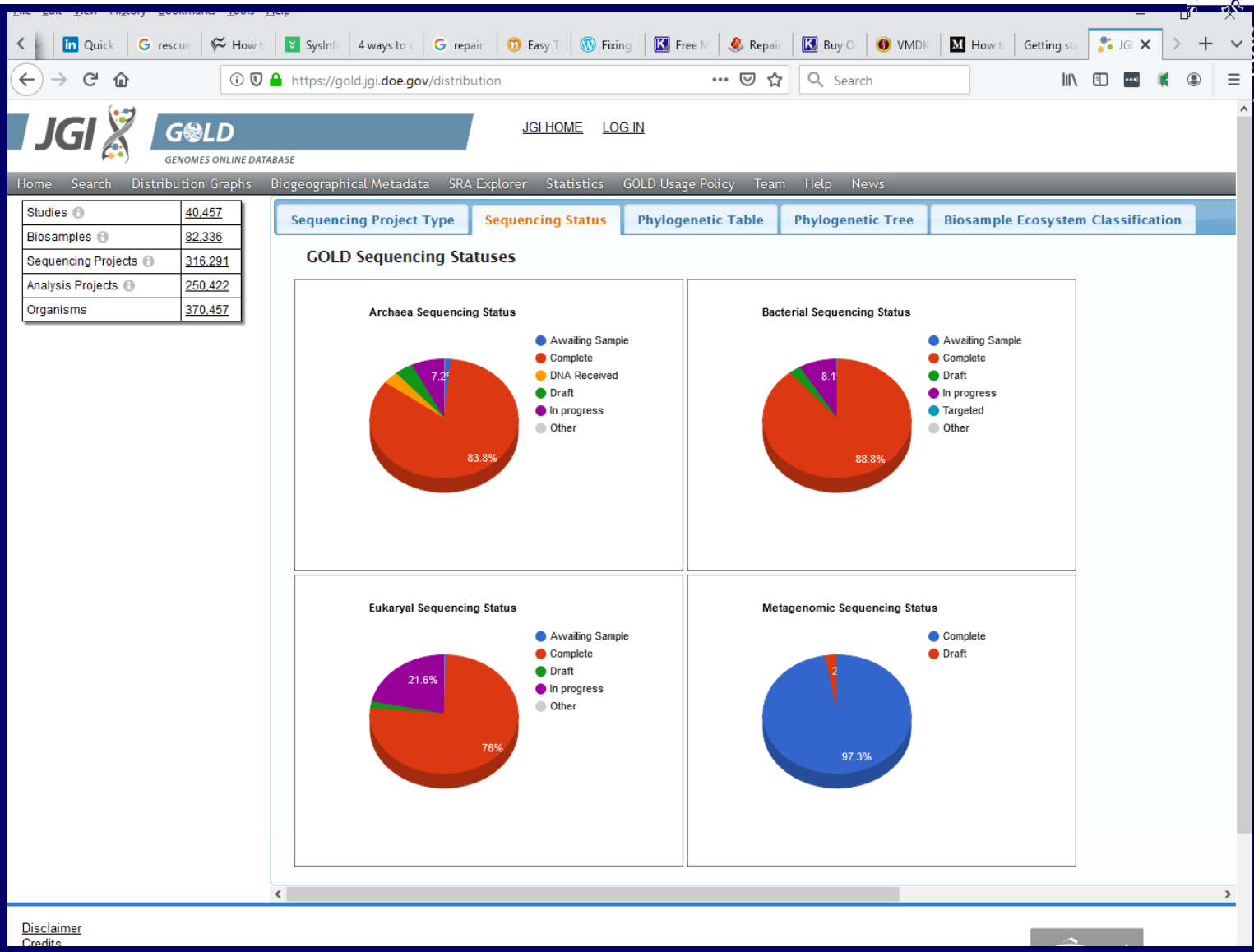
Szekvencia adatbázisok

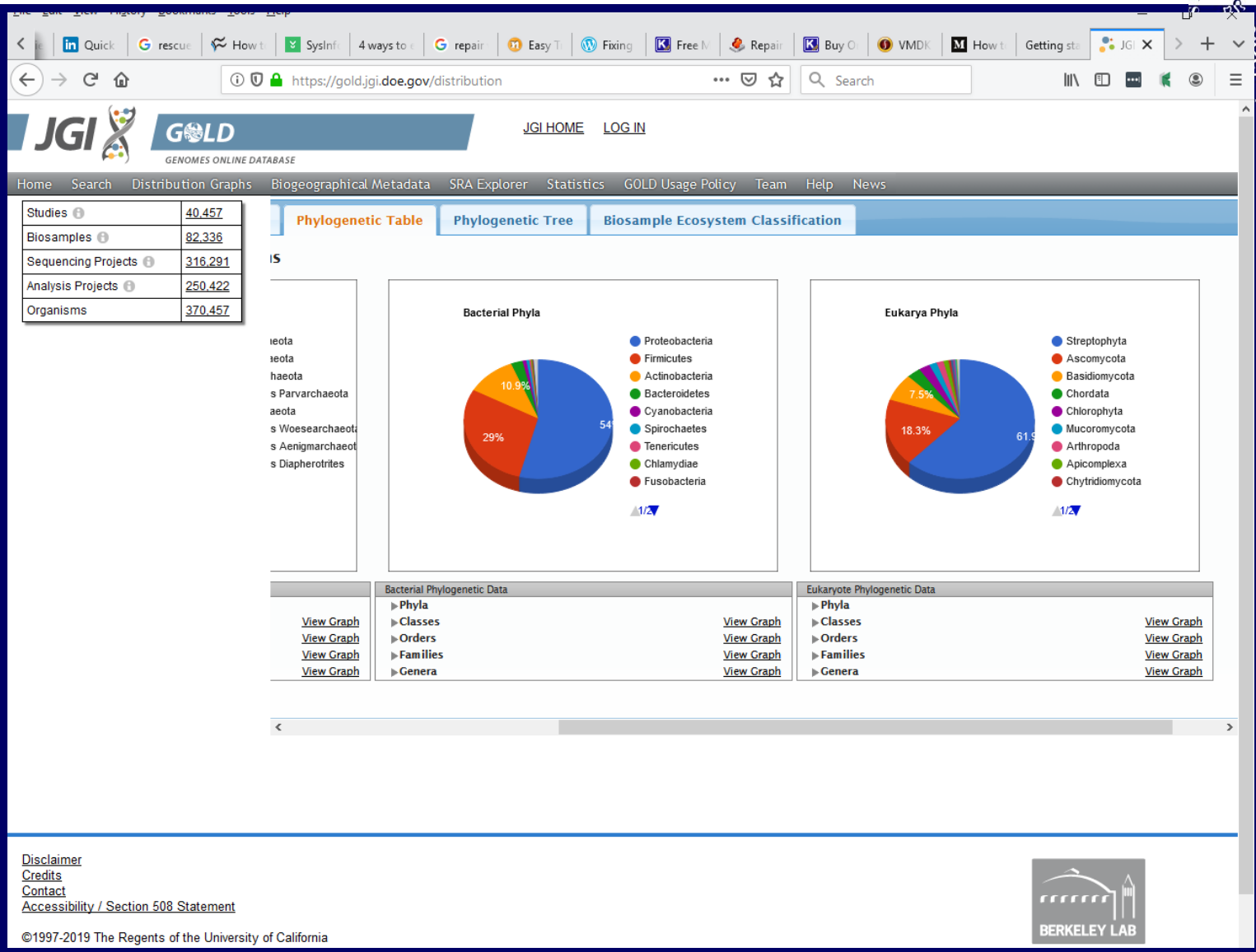
- Rengeteg megszekvenált genom
- Többsége virális és bakteriális
- Eukarióták közül is rengeteg
  - Gerincesek
  - Ízeltlábúak
  - Növények




https://gold.jgi.doe.gov
JGI
GOLD
GENOMES ONLINE DATABASE
JGI HOME
LOG IN
Home
Search
Distribution Graphs
Biogeographical Metadata
SRA Explorer
Statistics
GOLD Usage Policy
Team
Help
News
Studies 40,457
Biosamples 82,336
Sequencing Projects 316,291
Analysis Projects 250,422
Organisms 370,457
Excel Data file
File last generated: 23 Sep, 2019
Welcome to the Genomes OnLine Database
GOLD Release v.7
GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.
1. Register
Register your project information and Metadata in the Genomes Online Database
Register
2. Annotate
Annotate your microbial genome or metagenome with IMG
Annotate
NCBI Import Tracker
NCBI
GOLD
IMG
Prokaryotes
Viruses
Eukaryotes
Metagenomes
Metatranscriptomes
Studies
Metagenomic 1,763
Non-Metagenomic 38,689
Biosamples
Classification
Ecosystems
Host-associated 44,733
Engineered 5,676
Environmental 31,569
Sequencing Projects
Complete Projects 19,100
Permanent Drafts 193,467
Incomplete Projects 101,418
Targeted Projects 1,060
Analysis Projects
Genome Analysis 155,594
Metagenome Analysis 55,269
Metagenome - Cell Enrichment 1,852
Metagenome - Single Particle Sort 5,107
Metagenome - Assembled Genome (MAG) 11,723
Metatranscriptome Analysis 6,699
Combined Assembly 276
Single Cell - Screened (SAG) 2,367
Single Cell - Unscreened (SAG) 1,930
Transcriptome Analysis 544
Special Projects
Type Strains with Projects 8,801
Type Strains with projects at Genbank 6,929
GEBA Projects 3,360
Projects w. Genbank Data
Seq. Projects 150,426
Archaeal Projects 942
Bacterial Projects 136,674
Eukaryal Projects 4,298
JGI Projects
JGI Studies 1,436
JGI Biosamples 24,761
JGI Sequencing Projects 91,911
JGI Analysis Projects 46,658
Organisms
Organisms 370,305
Archaea 3,558
Bacteria 326,389
Eukarya 31,105
Viruses 9,253


https://gold.jgi.doe.gov/distribution
JGI
GOLD
GENOMES ONLINE DATABASE
JGI HOME
LOG IN
Home
Search
Distribution Graphs
Biogeographical Metadata
SRA Explorer
Statistics
GOLD Usage Policy
Team
Help
News
Studies 40,457
Biosamples 82,336
Sequencing Projects 316,291
Analysis Projects 250,422
Organisms 370,457
Sequencing Project Type
Sequencing Status
Phylogenetic Table
Phylogenetic Tree
Biosample Ecosystem Classification
GOLD Project Type Distributions
Archeal Distribution
88.9%
Resequen...
Transcript...
Whole Genome Sequencing
Other
Bacterial Distribution
96.9%
Genome fragments
Resequen...
Transcript...
Whole Genome Sequencing
Other
Eukarya Distribution
21.3%
26.1%
43.9%
ChIP-Sequencing
Exome
FAIRE-Seq
Genome fragments
Methylation
Resequen...
Transcript...
Whole Ge...
Other
Metagenomic Distribution
84.6%
Metagenome
Metatrans...
Targeted Gene Survey
Disclaimer
Credits


https://gold.jgi.doe.gov/distribution
JGI
GOLD
GENOMES ONLINE DATABASE
JGI HOME
LOG IN
Home
Search
Distribution Graphs
Biogeographical Metadata
SRA Explorer
Statistics
GOLD Usage Policy
Team
Help
News
Studies 40,457
Biosamples 82,336
Sequencing Projects 316,291
Analysis Projects 250,422
Organisms 370,457
Sequencing Project Type
Sequencing Status
Phylogenetic Table
Phylogenetic Tree
Biosample Ecosystem Classification
GOLD Sequencing Statuses
Archaea Sequencing Status
Awaiting Sample
Complete
DNA Received
Draft
In progress
Other
83.8%
Bacterial Sequencing Status
Awaiting Sample
Complete
Draft
In progress
Targeted
Other
88.8%
Eukaryal Sequencing Status
Awaiting Sample
Complete
Draft
In progress
Other
21.6%
76%
Metagenomic Sequencing Status
Complete
Draft
97.3%
Disclaimer
Credits


https://gold.jgi.doe.gov/distribution
JGI
GOLD
GENOMES ONLINE DATABASE
JGI HOME
LOG IN
Home
Search
Distribution Graphs
Biogeographical Metadata
SRA Explorer
Statistics
GOLD Usage Policy
Team
Help
News
Studies 40,457
Biosamples 82,336
Sequencing Projects 316,291
Analysis Projects 250,422
Organisms 370,457
Phylogenetic Table
Phylogenetic Tree
Biosample Ecosystem Classification
Bacterial Phyla
Proteobacteria
Firmicutes
Actinobacteria
Bacteroidetes
Cyanobacteria
Spirochaetes
Tenericutes
Chlamydiae
Fusobacteria
10.9%
29%
Eukarya Phyla
Streptophyta
Ascomycota
Basidiomycota
Chordata
Chlorophyta
Mucoromycota
Arthropoda
Apicomplexa
Chytridiomycota
7.5%
18.3%
Bacterial Phylogenetic Data
Eukaryote Phylogenetic Data
Phyla
Classes
Orders
Families
Genera
View Graph
Disclaimer
Credits
Contact
Accessibility / Section 508 Statement
©1997-2019 The Regents of the University of California
BERKELEY LAB

Biológiai adatbázisok fajtái

- Tartalom eredete szerint:
  - Elsődleges – közvetlen, kisérletes adatok
  - Származtatott – az előbbiek feldolgozása
- Elérés szerint:
  - Szabad elérésű – nemzetközi konzorciumok
  - Előfizetéses – üzleti alapú
  - Magán – gyógyszergyári, belső használatra
- Integrált adatbázisok
- Irodalmi adatbázis

Adatbázisok elérése

- Internet – első számú forrás
  - Régebben nyomtatásban, mágnes-szalagon és CD-n
- Sok szaklapnak van adatbázis szekciója
- „Nucleic Acid Research" januári száma
- „Database" Oxford University Press
- Az adatbázisokat rendszeresen frissítik
- Új adatbázisokat hoznak létre

Adatbázisok felépítése

- Az adatbázisok bejegyzésekből állnak (record)
- A bejegyzések mezőkre oszlanak (field)
- A mezők sorrendje meghatározott
- A mezők tartalma szintén
- Excell táblázat szerű, vagy sablon, vagy űrlap

Adatbázis formátumok

- „MS Word is **NOT** a database format!!!” – (EBI: FAQ)
- Az adatbázisok tipikusan egyszerű szöveg formátumúak („.txt” MS Word vagy ‘notepad’)
- Közvetlenül is olvashatók, de nem feltétlenül ebben a formában kerülnek felhasználásra
- Sok alkalmazás néhány szabványos formátumban várja a bemenő adatokat

A FASTA formátum

- Bejegyzés=szekvencia
- Első karakter:
  - ‘;’: komment, kerüld
  - ‘>’: címsor, utána azonosító
- A többi a szekvencia maga
- 120 karakternél ne legyen hosszabb sor
- ‘*’-gal végződik, vagy a következő címsorral

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

A nukleinsav adatbázis

- International Nucleotide Sequence Database Collaboration: ENA, NCBI, DDBJ
- Elsődleges adatbázis, ingyenes
- Közös formátum
- Közös azonosító kódrendszer
- Csak az egyik helyre kell beküldeni – a többi automatikusan szinkronizál
- Honlap: http://www.ebi.ac.uk/ena/home

EMBL formátum

Mező azonosító

A mező tartalma:

```
ID   X64011; SV 1; linear; genomic DNA; STD; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   Listeria ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
```

Tulajdonságok:

A szekvencia:

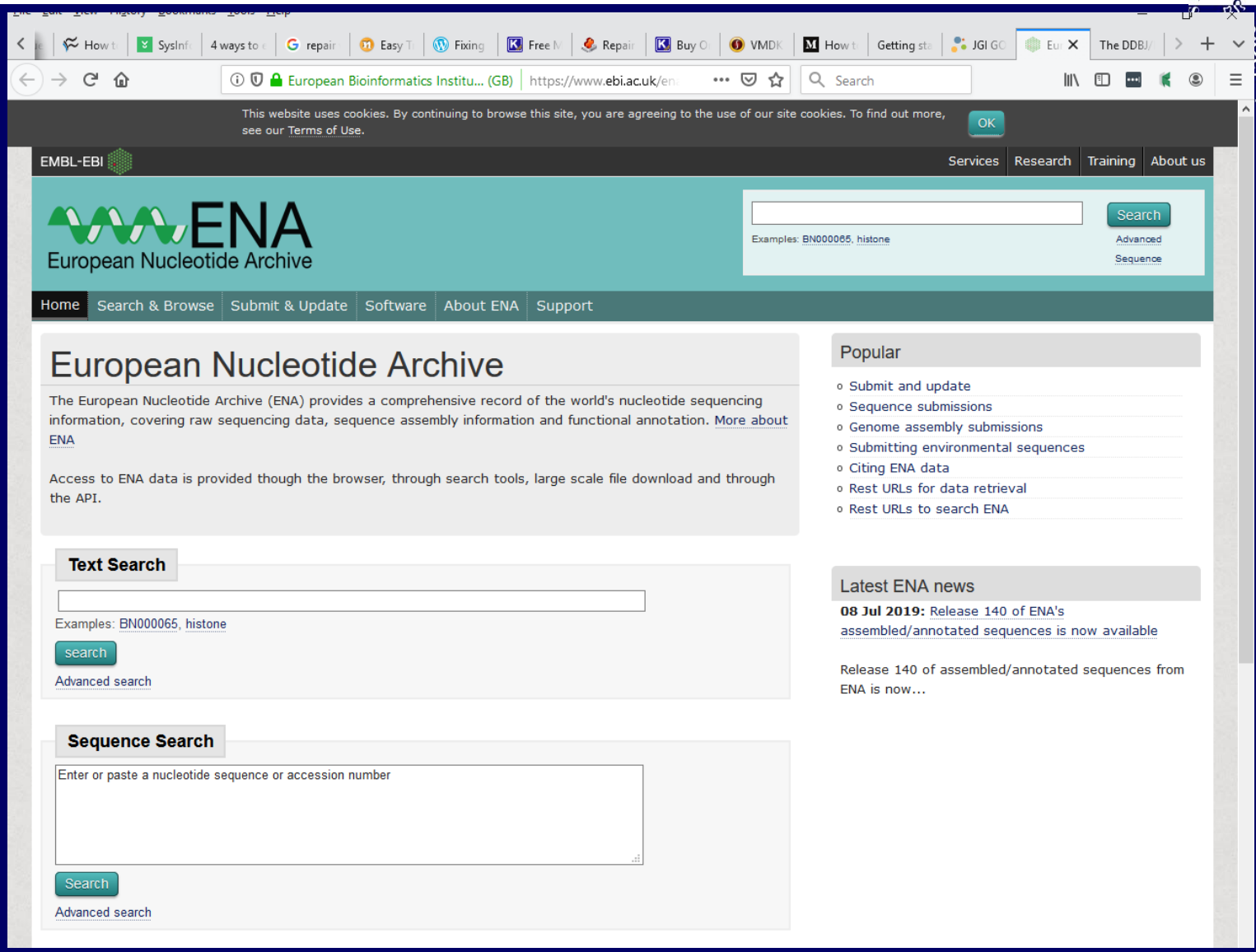
Itt a vége:

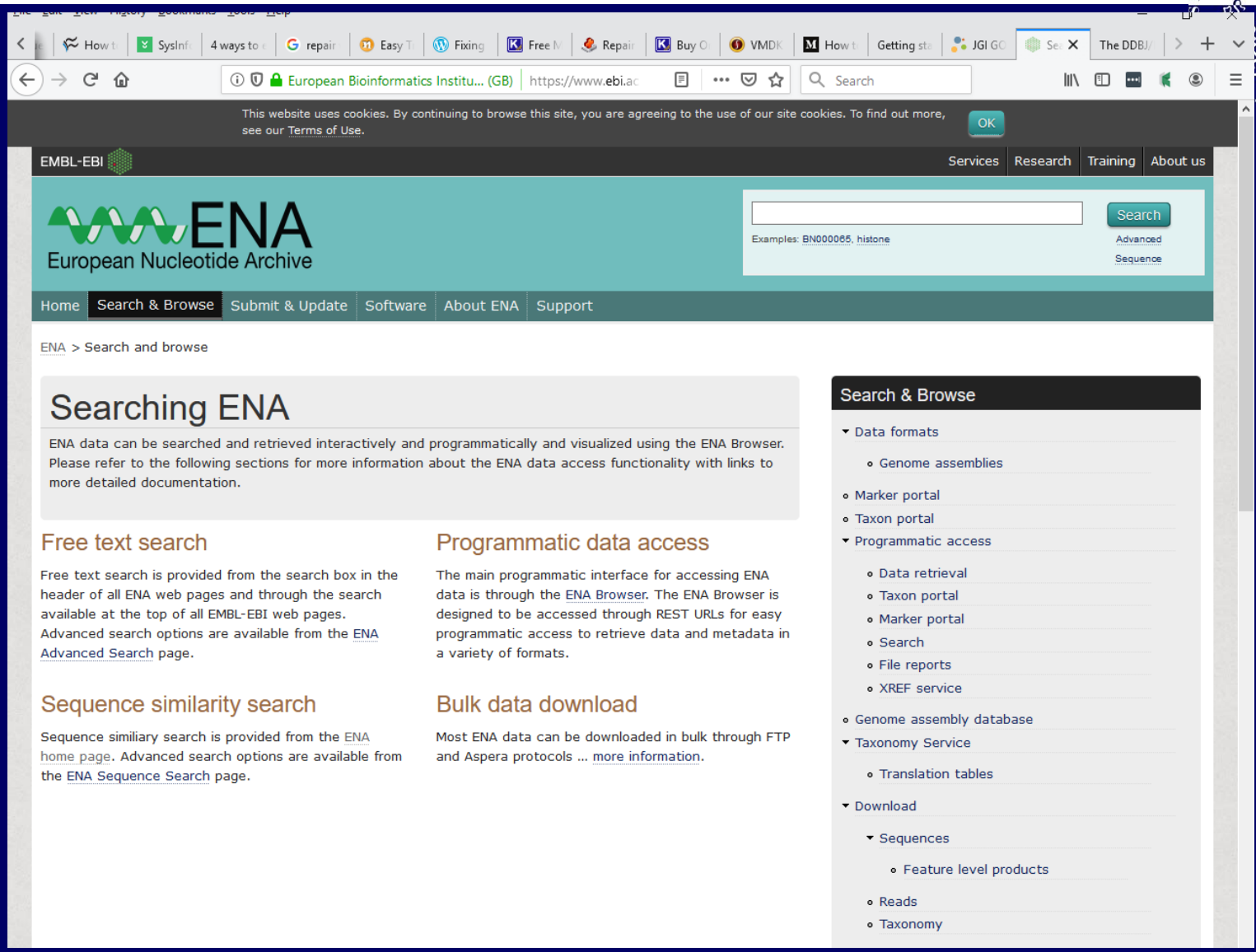
```
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT                   /mol_type="genomic DNA"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /db_xref="GOA:P28763"
FT                   /db_xref="HSSP:P00448"
FT                   /db_xref="InterPro:IPR001189"
FT                   /db_xref="UniProtKB/Swiss-Prot:P28763"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat   60
     gtaatttctt .....                                                   120
//
```

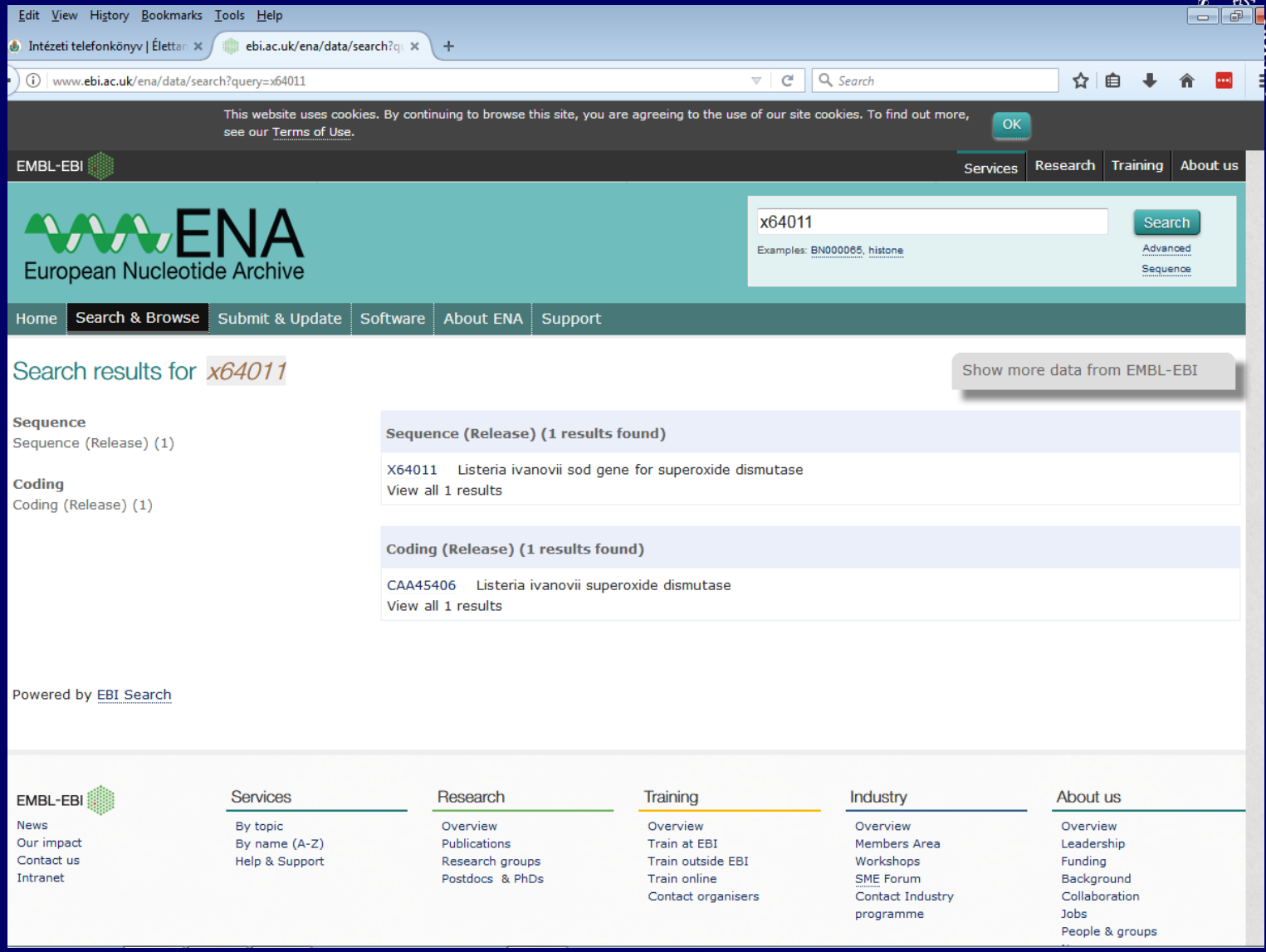
GeneBank formátum

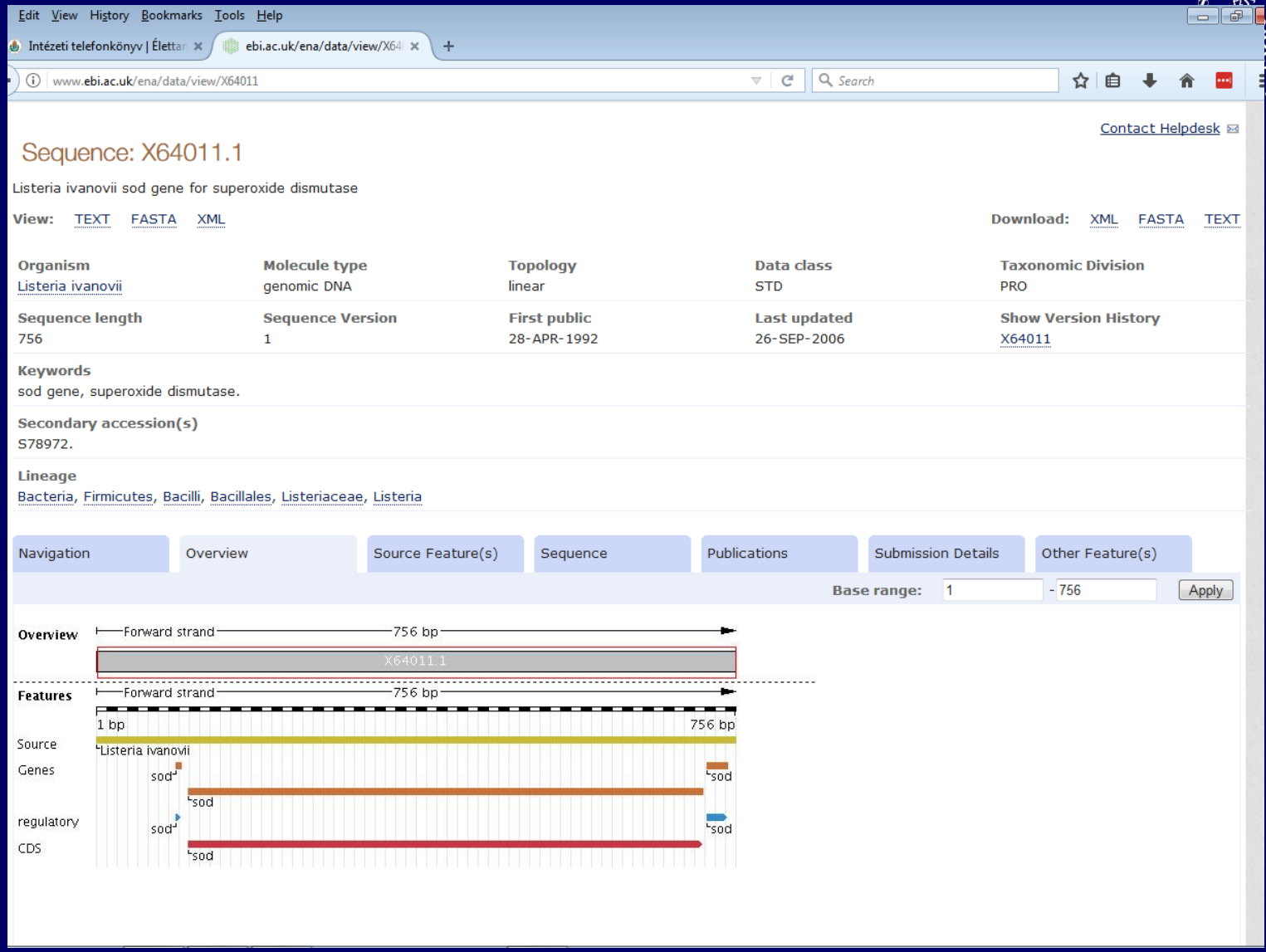
```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase gene from Listeria ivanovii by
            functional complementation in Escherichia coli and characterization
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322 (1992)
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft, Institut f. Mikrobiologie,
            Universitaet Wuerzburg, Biozentrum Am Hubland, 8700 Wuerzburg, FRG
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
```

Kicsit más formátum
Könnyű átalakítani
A teljes leírás:
http://www.insdc.org/documents/feature_table.html


This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.
OK
EMBL-EBI
Services
Research
Training
About us
ENA
European Nucleotide Archive
Search
Examples: BN000065, histone
Advanced
Sequence
Home
Search & Browse
Submit & Update
Software
About ENA
Support
European Nucleotide Archive
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA
Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.
Text Search
Examples: BN000065, histone
search
Advanced search
Sequence Search
Enter or paste a nucleotide sequence or accession number
Search
Advanced search
Popular
Submit and update
Sequence submissions
Genome assembly submissions
Submitting environmental sequences
Citing ENA data
Rest URLs for data retrieval
Rest URLs to search ENA
Latest ENA news
08 Jul 2019: Release 140 of ENA's assembled/annotated sequences is now available
Release 140 of assembled/annotated sequences from ENA is now...

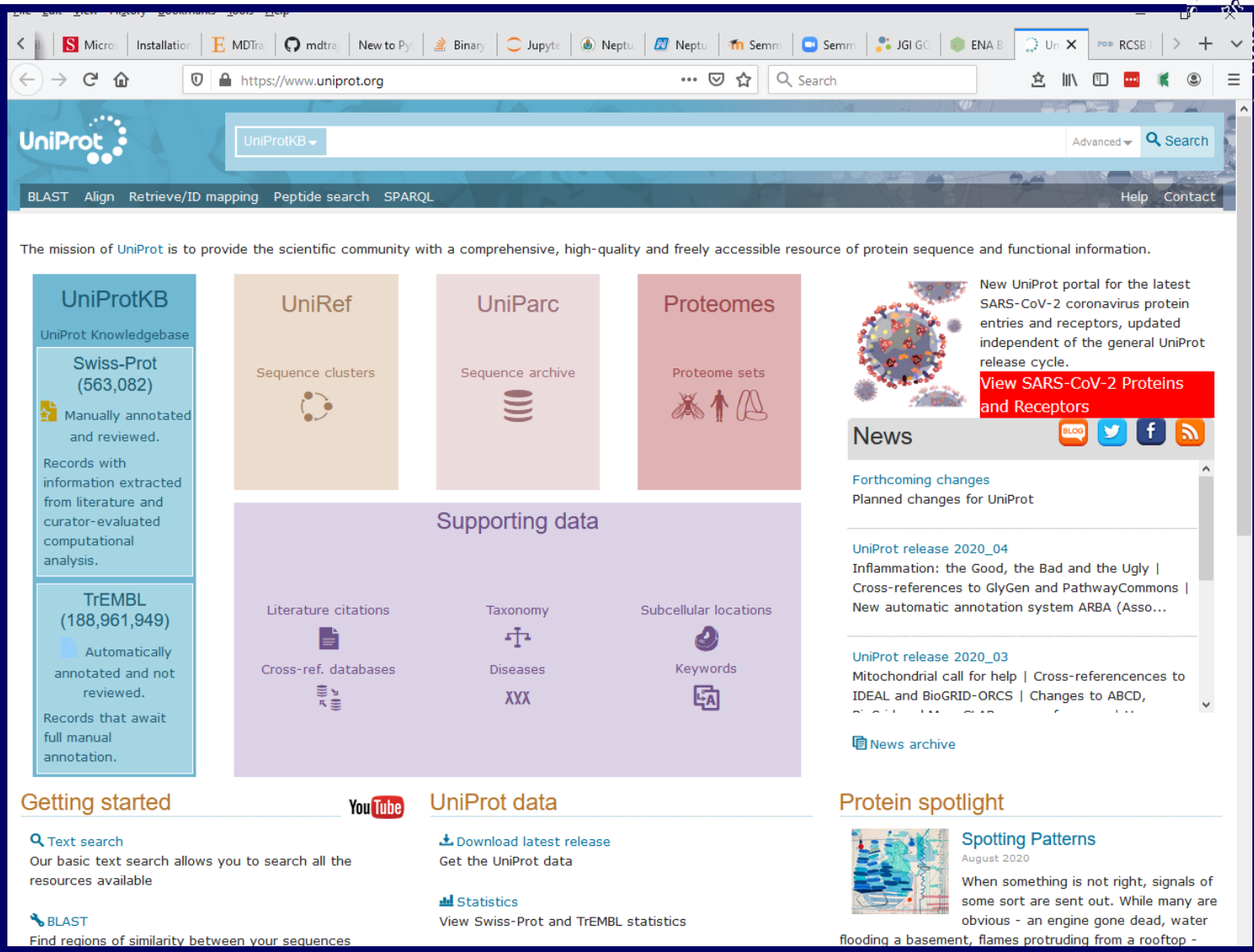

This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.
OK
EMBL-EBI
Services
Research
Training
About us
ENA
European Nucleotide Archive
Search
Examples: BN000065, histone
Advanced
Sequence
Home
Search & Browse
Submit & Update
Software
About ENA
Support
ENA > Search and browse
Searching ENA
ENA data can be searched and retrieved interactively and programmatically and visualized using the ENA Browser. Please refer to the following sections for more information about the ENA data access functionality with links to more detailed documentation.
Free text search
Free text search is provided from the search box in the header of all ENA web pages and through the search available at the top of all EMBL-EBI web pages. Advanced search options are available from the ENA Advanced Search page.
Programmatic data access
The main programmatic interface for accessing ENA data is through the ENA Browser. The ENA Browser is designed to be accessed through REST URLs for easy programmatic access to retrieve data and metadata in a variety of formats.
Sequence similarity search
Sequence similiary search is provided from the ENA home page. Advanced search options are available from the ENA Sequence Search page.
Bulk data download
Most ENA data can be downloaded in bulk through FTP and Aspera protocols ... more information.
Search & Browse
Data formats
Genome assemblies
Marker portal
Taxon portal
Programmatic access
Data retrieval
Taxon portal
Marker portal
Search
File reports
XREF service
Genome assembly database
Taxonomy Service
Translation tables
Download
Sequences
Feature level products
Reads
Taxonomy


Edit View History Bookmarks Tools Help
Intézeti telefonkönyv | Élettan
ebi.ac.uk/ena/data/search?q
www.ebi.ac.uk/ena/data/search?query=x64011
Search
This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.
OK
EMBL-EBI
Services
Research
Training
About us
ENA
European Nucleotide Archive
x64011
Search
Examples: BN000065, histone
Advanced
Sequence
Home
Search & Browse
Submit & Update
Software
About ENA
Support
Search results for x64011
Show more data from EMBL-EBI
Sequence
Sequence (Release) (1)
Coding
Coding (Release) (1)
Sequence (Release) (1 results found)
X64011 Listeria ivanovii sod gene for superoxide dismutase
View all 1 results
Coding (Release) (1 results found)
CAA45406 Listeria ivanovii superoxide dismutase
View all 1 results
Powered by EBI Search
EMBL-EBI
News
Our impact
Contact us
Intranet
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Postdocs & PhDs
Training
Overview
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Overview
Members Area
Workshops
SME Forum
Contact Industry programme
About us
Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups


Edit View History Bookmarks Tools Help
Intézeti telefonkönyv | Élettan
ebi.ac.uk/ena/data/view/X64
www.ebi.ac.uk/ena/data/view/X64011
Search
Contact Helpdesk
Sequence: X64011.1
Listeria ivanovii sod gene for superoxide dismutase
View: TEXT FASTA XML
Download: XML FASTA TEXT
Organism
Listeria ivanovii
Molecule type
genomic DNA
Topology
linear
Data class
STD
Taxonomic Division
PRO
Sequence length
756
Sequence Version
1
First public
28-APR-1992
Last updated
26-SEP-2006
Show Version History
X64011
Keywords
sod gene, superoxide dismutase.
Secondary accession(s)
S78972.
Lineage
Bacteria, Firmicutes, Bacilli, Bacillales, Listeriaceae, Listeria
Navigation
Overview
Source Feature(s)
Sequence
Publications
Submission Details
Other Feature(s)
Base range: 1 - 756 Apply
Overview
Forward strand
756 bp
X64011.1
Features
Forward strand
756 bp
1 bp
756 bp
Source
Listeria ivanovii
Genes
sod
sod
sod
sod
regulatory
sod
sod
CDS
sod

UniProt adatbázis

- Konzorcium üzemelteti:
- A fehérjékkel kapcsolatos információk fő forrása
- Az adatbázis nem homogén:
  - SwissProt: egyedileg annotált és ellenőrzött
  - TrEMBL: automatikusan annotált adatok
- Ingyenesen elérhető: http://www.uniprot.org/

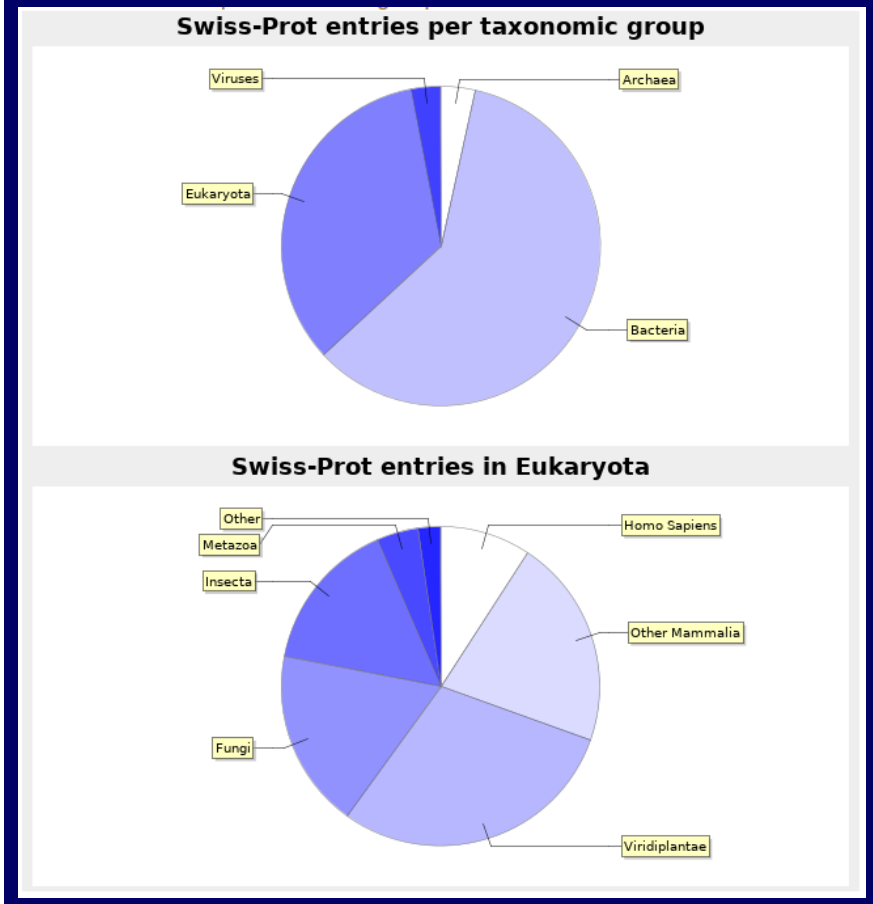

https://www.uniprot.org
UniProt
UniProtKB
Search
BLAST Align Retrieve/ID mapping Peptide search SPARQL
Help Contact
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
UniProtKB
UniProt Knowledgebase
Swiss-Prot (563,082)
Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.
TrEMBL (188,961,949)
Automatically annotated and not reviewed.
Records that await full manual annotation.
UniRef
Sequence clusters
UniParc
Sequence archive
Proteomes
Proteome sets
Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords
New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle.
View SARS-CoV-2 Proteins and Receptors
News
Forthcoming changes
Planned changes for UniProt
UniProt release 2020_04
Inflammation: the Good, the Bad and the Ugly | Cross-references to GlyGen and PathwayCommons | New automatic annotation system ARBA (Asso...
UniProt release 2020_03
Mitochondrial call for help | Cross-referencences to IDEAL and BioGRID-ORCS | Changes to ABCD,
News archive
Getting started
Text search
Our basic text search allows you to search all the resources available
BLAST
Find regions of similarity between your sequences
UniProt data
Download latest release
Get the UniProt data
Statistics
View Swiss-Prot and TrEMBL statistics
Protein spotlight
Spotting Patterns
August 2020
When something is not right, signals of some sort are sent out. While many are obvious - an engine gone dead, water flooding a basement, flames protruding from a rooftop -

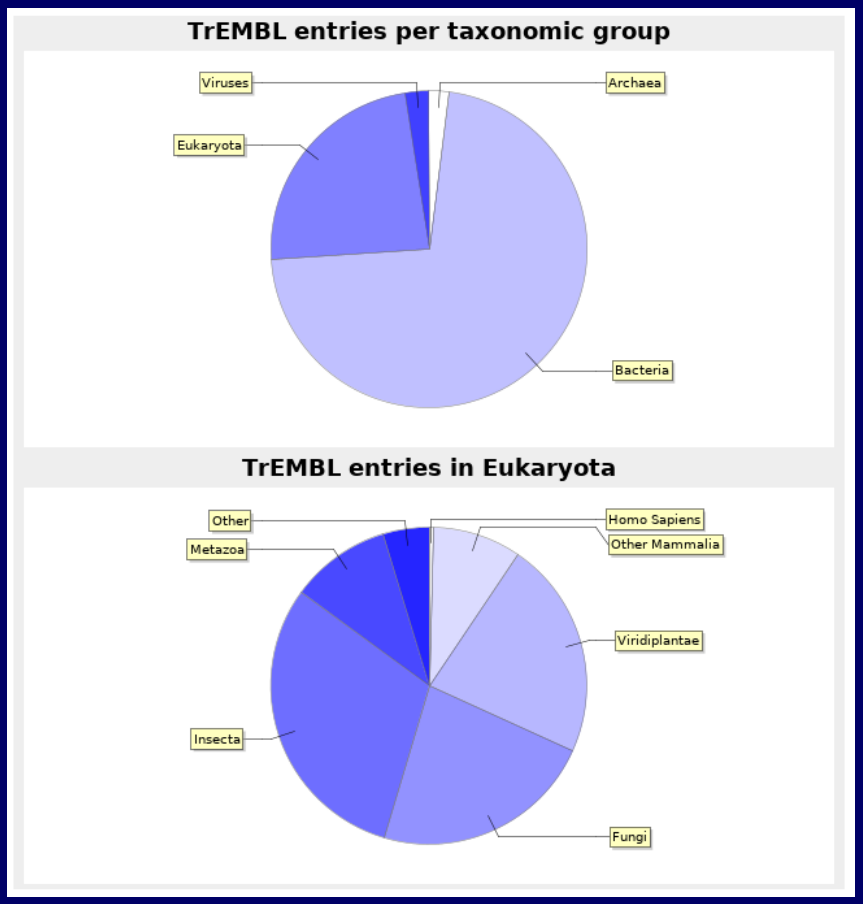
A legfrissebb változat adatai

	Number of entries
New entries	289
Updated entries	187,296
Unchanged entries	373,238
Total	560,823
Entries with updated sequences	29
With a fragmented AA sequence	9,175
With known alternative products	25,328

	Number of entries
New entries	4,227,499
Updated entries	20,902,200
Unchanged entries	146,371,789
Total	171,501,488
Entries with updated sequences	18,613
With a fragmented AA sequence	16,359,479
With known alternative products	0

A fehérjék megoszlása:

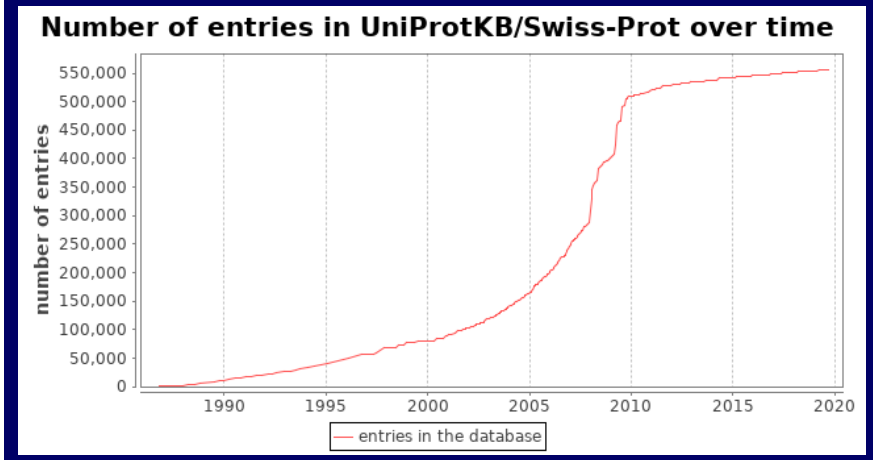


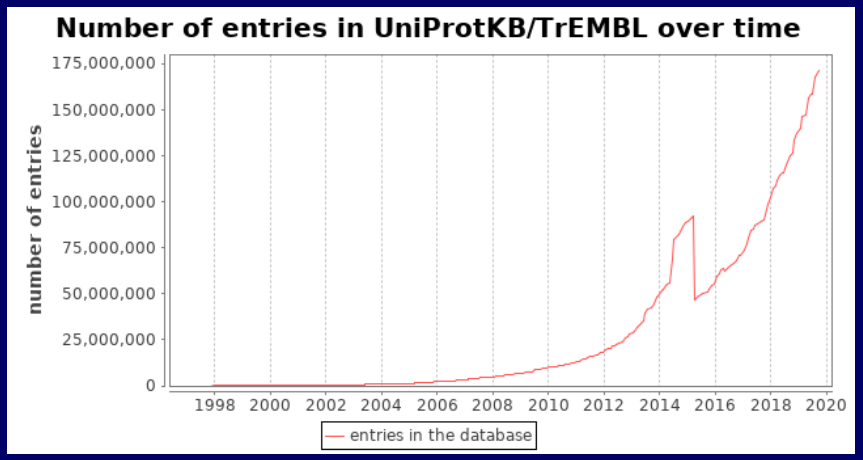


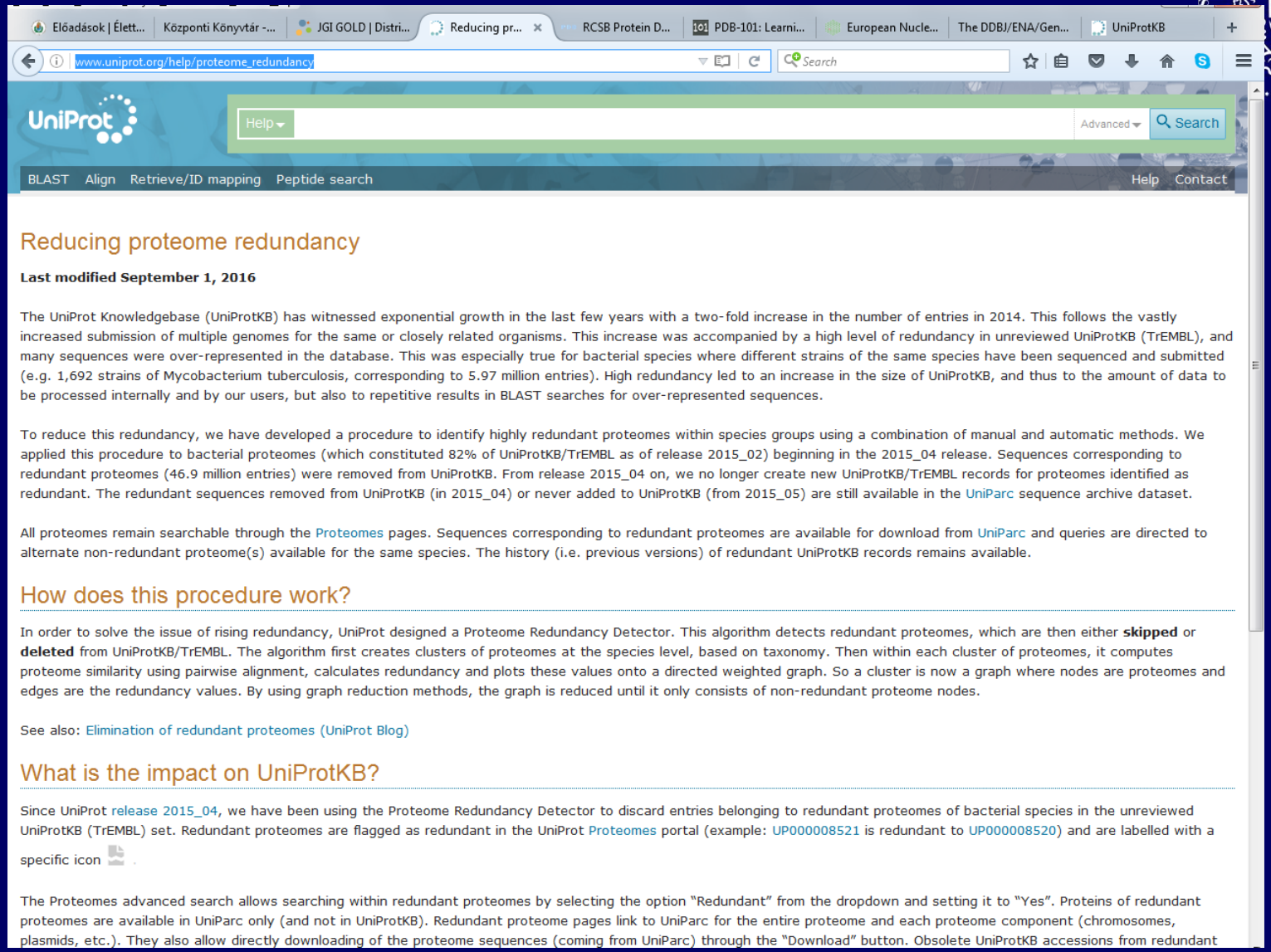
Mennyire biztos?

Evidence:	**Swiss-Prot**	**TrEMBL**
protein level	17.7%	0.12%
transcript level	10.3%	0.93%
from homology	69.2%	24.60%
predicted	2.4%	74.35%
uncertain	0.3%	0%

Az adatbázis gyarapodása







Előadások | Élett... | Központi Könyvtár -... | JGI GOLD | Distri... | Reducing pr... | RCSB Protein D... | PDB-101: Learni... | European Nucle... | The DDBJ/ENA/Gen... | UniProtKB

www.uniprot.org/help/proteome_redundancy

UniProt

Help | Advanced | Search

BLAST | Align | Retrieve/ID mapping | Peptide search | Help | Contact

Reducing proteome redundancy

Last modified September 1, 2016

The UniProt Knowledgebase (UniProtKB) has witnessed exponential growth in the last few years with a two-fold increase in the number of entries in 2014. This follows the vastly increased submission of multiple genomes for the same or closely related organisms. This increase was accompanied by a high level of redundancy in unreviewed UniProtKB (TrEMBL), and many sequences were over-represented in the database. This was especially true for bacterial species where different strains of the same species have been sequenced and submitted (e.g. 1,692 strains of Mycobacterium tuberculosis, corresponding to 5.97 million entries). High redundancy led to an increase in the size of UniProtKB, and thus to the amount of data to be processed internally and by our users, but also to repetitive results in BLAST searches for over-represented sequences.

To reduce this redundancy, we have developed a procedure to identify highly redundant proteomes within species groups using a combination of manual and automatic methods. We applied this procedure to bacterial proteomes (which constituted 82% of UniProtKB/TrEMBL as of release 2015_02) beginning in the 2015_04 release. Sequences corresponding to redundant proteomes (46.9 million entries) were removed from UniProtKB. From release 2015_04 on, we no longer create new UniProtKB/TrEMBL records for proteomes identified as redundant. The redundant sequences removed from UniProtKB (in 2015_04) or never added to UniProtKB (from 2015_05) are still available in the UniParc sequence archive dataset.

All proteomes remain searchable through the Proteomes pages. Sequences corresponding to redundant proteomes are available for download from UniParc and queries are directed to alternate non-redundant proteome(s) available for the same species. The history (i.e. previous versions) of redundant UniProtKB records remains available.

How does this procedure work?

In order to solve the issue of rising redundancy, UniProt designed a Proteome Redundancy Detector. This algorithm detects redundant proteomes, which are then either **skipped** or **deleted** from UniProtKB/TrEMBL. The algorithm first creates clusters of proteomes at the species level, based on taxonomy. Then within each cluster of proteomes, it computes proteome similarity using pairwise alignment, calculates redundancy and plots these values onto a directed weighted graph. So a cluster is now a graph where nodes are proteomes and edges are the redundancy values. By using graph reduction methods, the graph is reduced until it only consists of non-redundant proteome nodes.

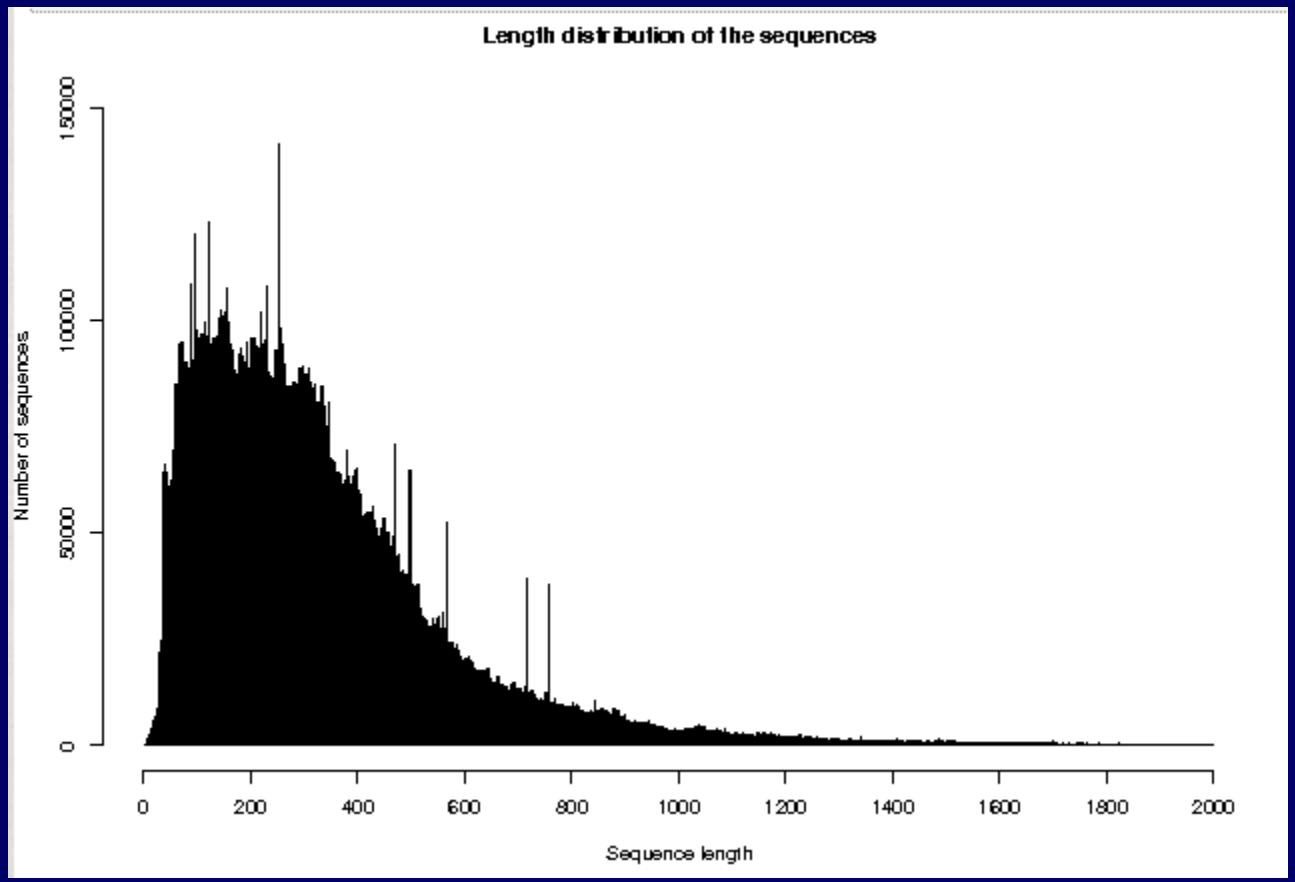
See also: Elimination of redundant proteomes (UniProt Blog)

What is the impact on UniProtKB?

Since UniProt release 2015_04, we have been using the Proteome Redundancy Detector to discard entries belonging to redundant proteomes of bacterial species in the unreviewed UniProtKB (TrEMBL) set. Redundant proteomes are flagged as redundant in the UniProt Proteomes portal (example: UP000008521 is redundant to UP000008520) and are labelled with a specific icon .

The Proteomes advanced search allows searching within redundant proteomes by selecting the option "Redundant" from the dropdown and setting it to "Yes". Proteins of redundant proteomes are available in UniParc only (and not in UniProtKB). Redundant proteome pages link to UniParc for the entire proteome and each proteome component (chromosomes, plasmids, etc.). They also allow directly downloading of the proteome sequences (coming from UniParc) through the "Download" button. Obsolete UniProtKB accessions from redundant

A fehérjék hosszának eloszlása:



Az adsatbázis formátuma:

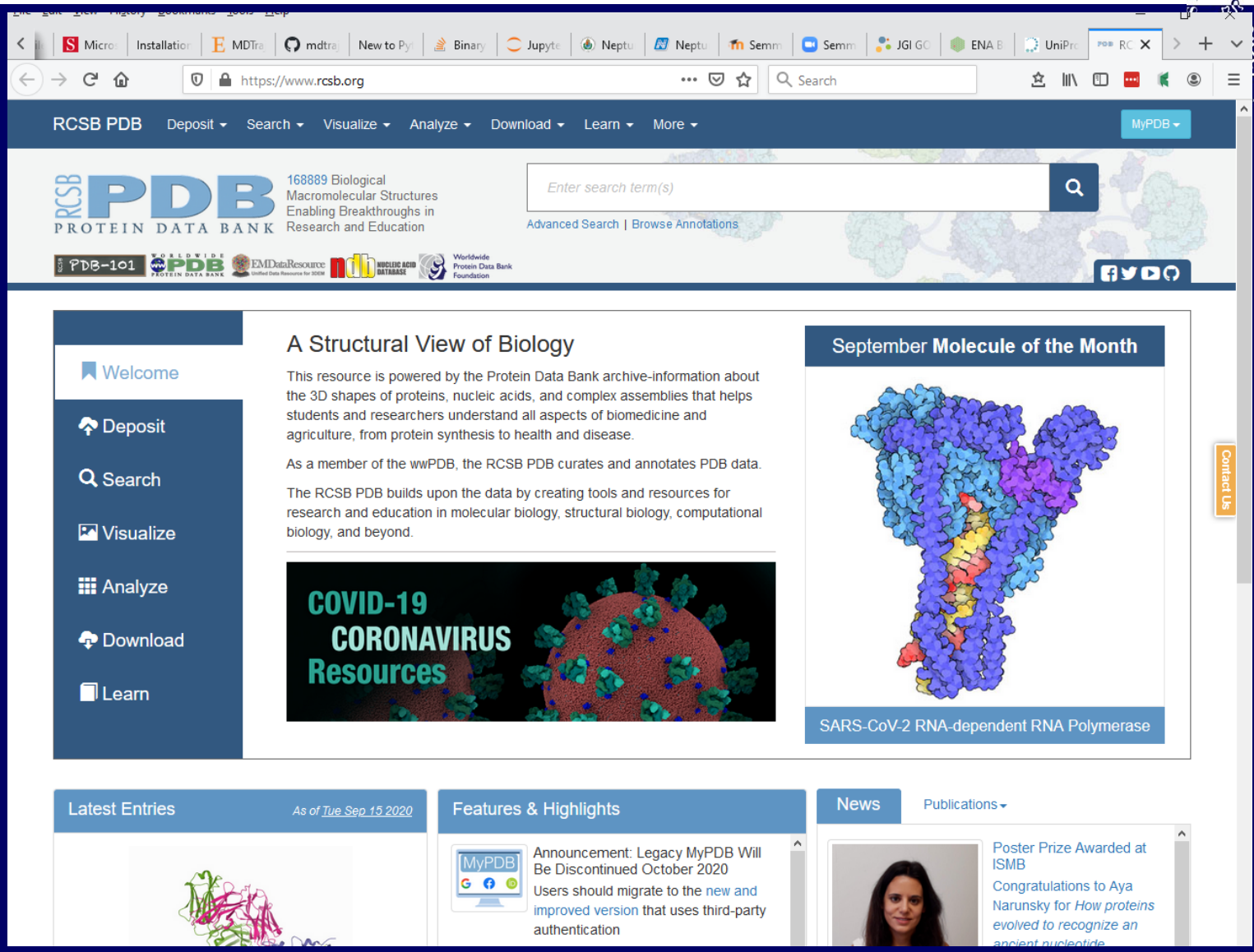
```
ID   SHH_HUMAN               Reviewed;         462 AA.
AC   Q15465; A4D247; Q75MC9;
DT   15-JUL-1999, integrated into UniProtKB/Swiss-Prot.
DT   01-NOV-1996, sequence version 1.
DT   27-JUL-2011, entry version 137.
DE   RecName: Full=Sonic hedgehog protein;
DE            Short=SHH;
DE   AltName: Full=HHG-1;
DE   Contains:
DE     RecName: Full=Sonic hedgehog protein N-product;
DE   Contains:
DE     RecName: Full=Sonic hedgehog protein C-product;
DE   Flags: Precursor;
GN   Name=SHH;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   TISSUE=Fetal lung;
RX   MEDLINE=96070431; PubMed=7590746; DOI=10.1006/geno.1995.1104;
RA   Marigo V., Roberts D.J., Lee S.M.K., Tsukurov O., Levi T.,
RA   Gastier J.M., Epstein D.J., Gilbert D.J., Copeland N.G., Seidman C.E.,
RA   Jenkins N.A., Seidman J.G., McMahon A.P., Tabin C.;
RT   "Cloning, expression, and chromosomal location of SHH and IHH: two
RT   human homologues of the Drosophila segment polarity gene hedgehog.";
RL   Genomics 28:44-51(1995).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RA   Tate G., Kishimoto K., Mitsuya T.;
RT   "Expression of Sonic hedgehog and its receptor Patched/Smoothened in
RT   human cancer cell lines and embryonic organs.";
RL   J. Biochem. Mol. Biol. Biophys. 4:27-34(2000).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RG   NIEHS SNPs program;
RL   Submitted (SEP-2003) to the EMBL/GenBank/DDBJ databases.
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX   MEDLINE=22737999; PubMed=12853948; DOI=10.1038/nature01782;
RA   Hillier L.W., Fulton R.S., Fulton L.A., Graves T.A., Pepin K.H.,
RA   Wagner-McPherson C., Layman D., Maas J., Jaeger S., Walker R.,
RA   Wylie K., Sekhon M., Becker M.C., O'Laughlin M.D., Schaller M.E.,
RA   Fewell G.A., Delehaunty K.D., Miner T.L., Nash W.E., Cordes M., Du H.,
RA   Sun H., Edwards J., Bradshaw-Cordum H., Ali J., Andrews S., Isak A.,
RA   Vanbrunt A., Nguyen C., Du F., Lamar B., Courtney L., Kalicki J.,
```

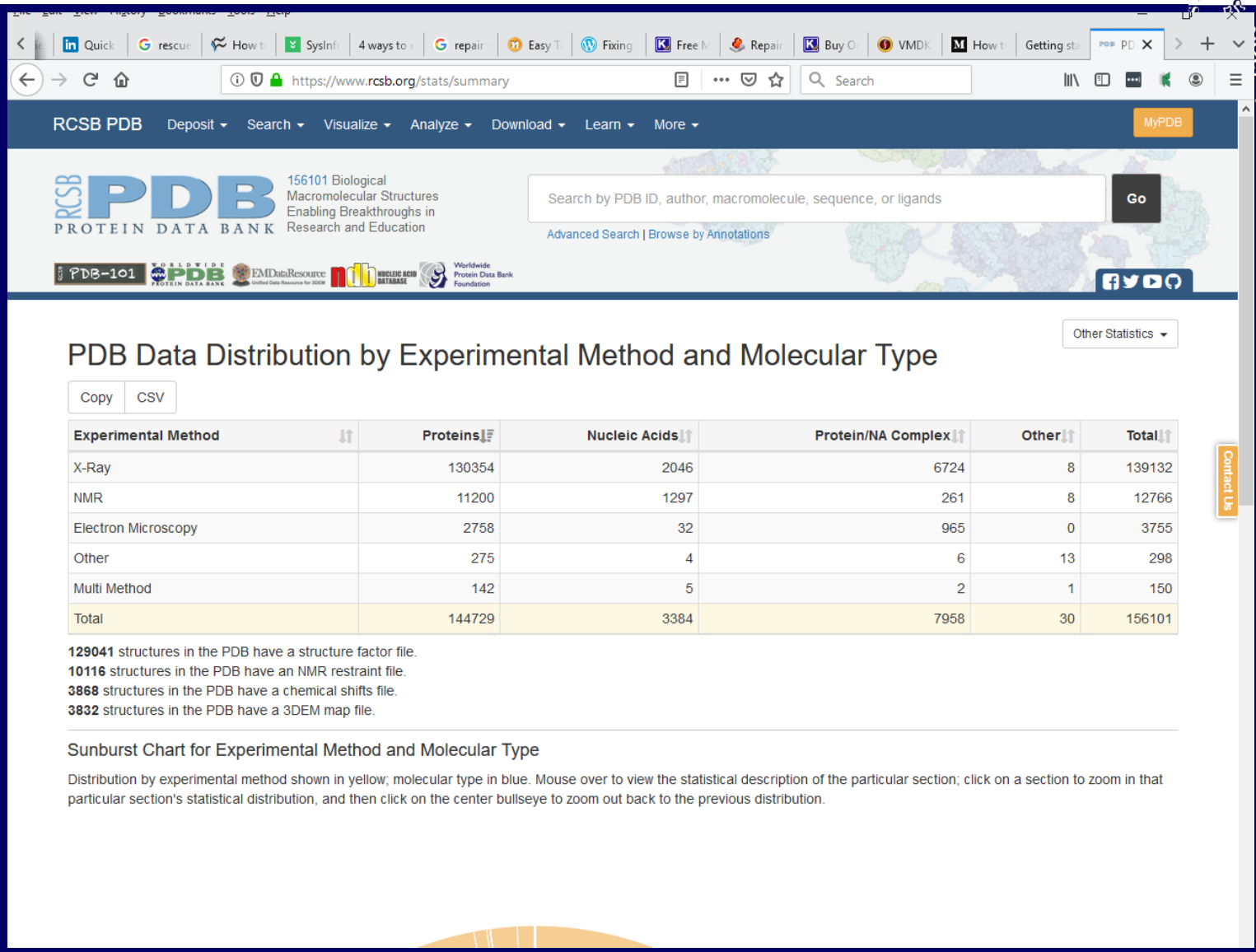
```
FT   VARIANT      401    408       Missing (in ocular coloboma).
FT                                 /FTId=VAR_017884.
FT   VARIANT      402    409       Missing (in HPE3; uncertain
FT                                 pathogenicity).
FT                                 /FTId=VAR_062643.
FT   VARIANT      404    408       Missing (in HPE3; familial).
FT                                 /FTId=VAR_009175.
FT   VARIANT      405    409       Missing (in HPE3; uncertain
FT                                 pathogenicity).
FT                                 /FTId=VAR_062644.
FT   VARIANT      411    411       G -> GG (in HPE3; uncertain
FT                                 pathogenicity).
FT                                 /FTId=VAR_062645.
FT   VARIANT      416    416       T -> A (in HPE3; uncertain
FT                                 pathogenicity).
FT                                 /FTId=VAR_062646.
FT   VARIANT      424    424       P -> A (in HPE3; familial).
FT                                 /FTId=VAR_009176.
FT   VARIANT      435    435       Y -> N (in HPE3).
FT                                 /FTId=VAR_062647.
FT   VARIANT      436    436       S -> L (in HPE3; sporadic).
FT                                 /FTId=VAR_009177.
FT   VARIANT      456    456       G -> R (in HPE3; uncertain
FT                                 pathogenicity).
FT                                 /FTId=VAR_062648.
FT   MUTAGEN       24     24       C->S: Abolishes palmitoylation.
FT   STRAND        47     51
FT   TURN          72     75
FT   STRAND        84     86
FT   HELIX        100    116
FT   STRAND       122    126
FT   STRAND       145    150
FT   HELIX        155    157
FT   HELIX        161    167
FT   STRAND       172    174
FT   STRAND       181    184
SQ   SEQUENCE   462 AA;  49607 MW;  DD687AFA582A4749 CRC64;
     MLLLARCLLL VLVSSLLVCS GLACGPGRGF GKRRHPKKLT PLAYKQFIPN VAEKTLGASG
     RYEGKISRNS ERFKELTPNY NPDIIFKDEE NTGADRLMTQ RCKDKLNALA ISVMNQWPGV
     KLRVTEGWDE DGHHSEESLH YEGRAVDITT SDRDRSKYGM LARLAVEAGF DWVYYESKAH
     IHCSVKAENS VAAKSGGCFP GSATVHLEQG GTKLVKDLSP GDRVLAADDQ GRLLYSDFLT
     FLDRDDGAKK VFYVIETREP RERLLLTAAH LLFVAPHNDS ATGEPEASSG SGPPSGGALG
     PRALFASRVR PGQRVYVVAE RDGDRRLLPA AVHSVTLSEE AAGAYAPLTA QGTILINRVL
     ASCYAVIEEH SWAHRAFAPF RLAHALLAAL APARTDRGGD SGGGDRGGGG GRVALTAPGA
     ADAPGAGATA GIHWYSQLLY QIGTWLLDSE ALHPLGMAVK SS
//
```

SZÜNET

PDB adatbázis

- Nemzetközi konzorcium üzemeltetei
- Ingyenesen elérhető:
  http://www.rcsb.org/pdb/home/home.do
  Elsődleges térszerkezeti adatbázis


https://www.rcsb.org
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
168889 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search term(s)
Advanced Search | Browse Annotations
PDB-101
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
A Structural View of Biology
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.
COVID-19 CORONAVIRUS Resources
September Molecule of the Month
SARS-CoV-2 RNA-dependent RNA Polymerase
Contact Us
Latest Entries
As of Tue Sep 15 2020
Features & Highlights
Announcement: Legacy MyPDB Will Be Discontinued October 2020
Users should migrate to the new and improved version that uses third-party authentication
News
Publications
Poster Prize Awarded at ISMB
Congratulations to Aya Narunsky for How proteins evolved to recognize an ancient nucleotide



https://www.rcsb.org/stats/summary

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB

RCSB PDB PROTEIN DATA BANK — 156101 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands — Go

Advanced Search | Browse by Annotations

PDB-101 | Worldwide PDB | EMDataResource | Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Other Statistics

PDB Data Distribution by Experimental Method and Molecular Type

Copy | CSV

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complex	Other	Total
X-Ray	130354	2046	6724	8	139132
NMR	11200	1297	261	8	12766
Electron Microscopy	2758	32	965	0	3755
Other	275	4	6	13	298
Multi Method	142	5	2	1	150
Total	144729	3384	7958	30	156101

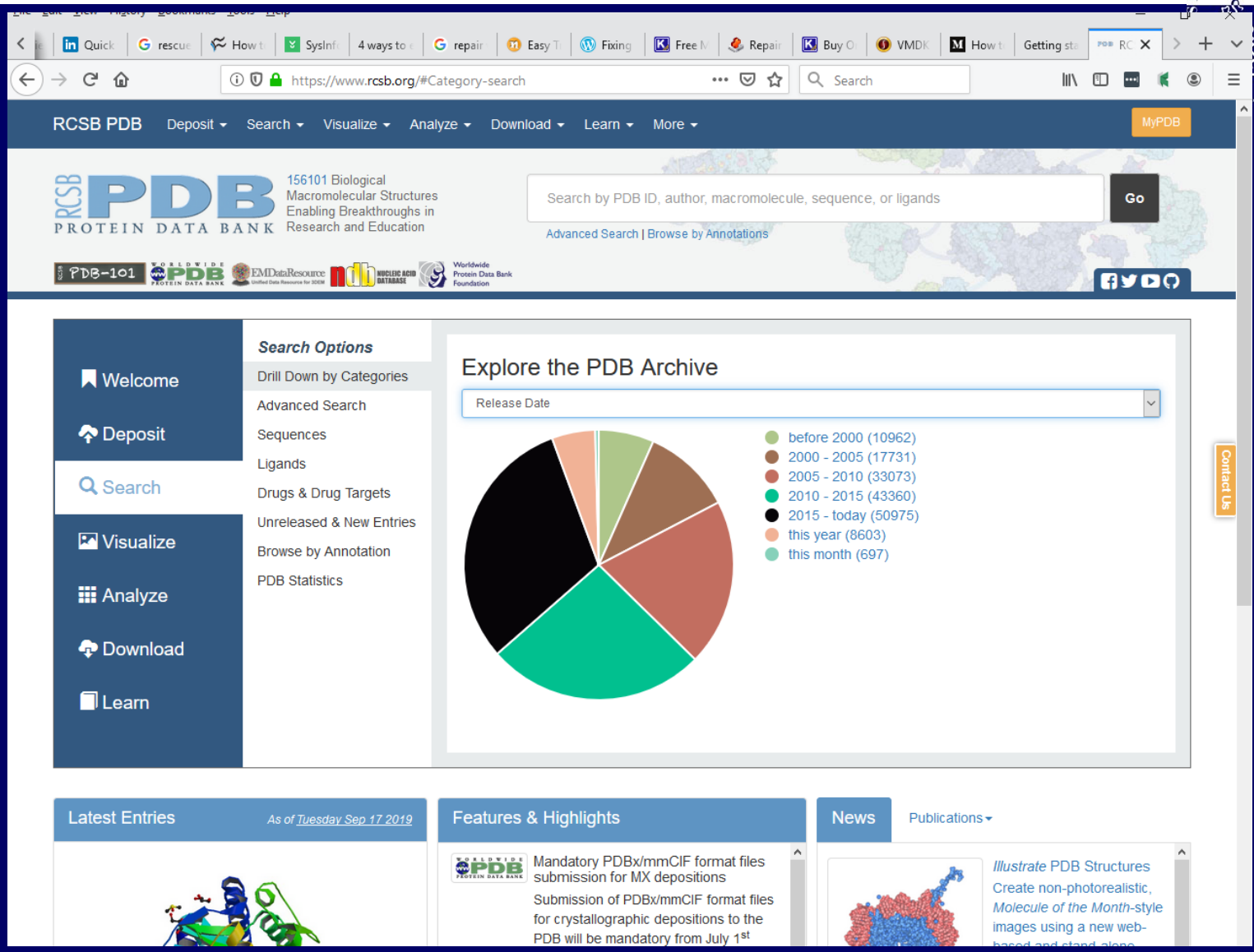
129041 structures in the PDB have a structure factor file.
10116 structures in the PDB have an NMR restraint file.
3868 structures in the PDB have a chemical shifts file.
3832 structures in the PDB have a 3DEM map file.

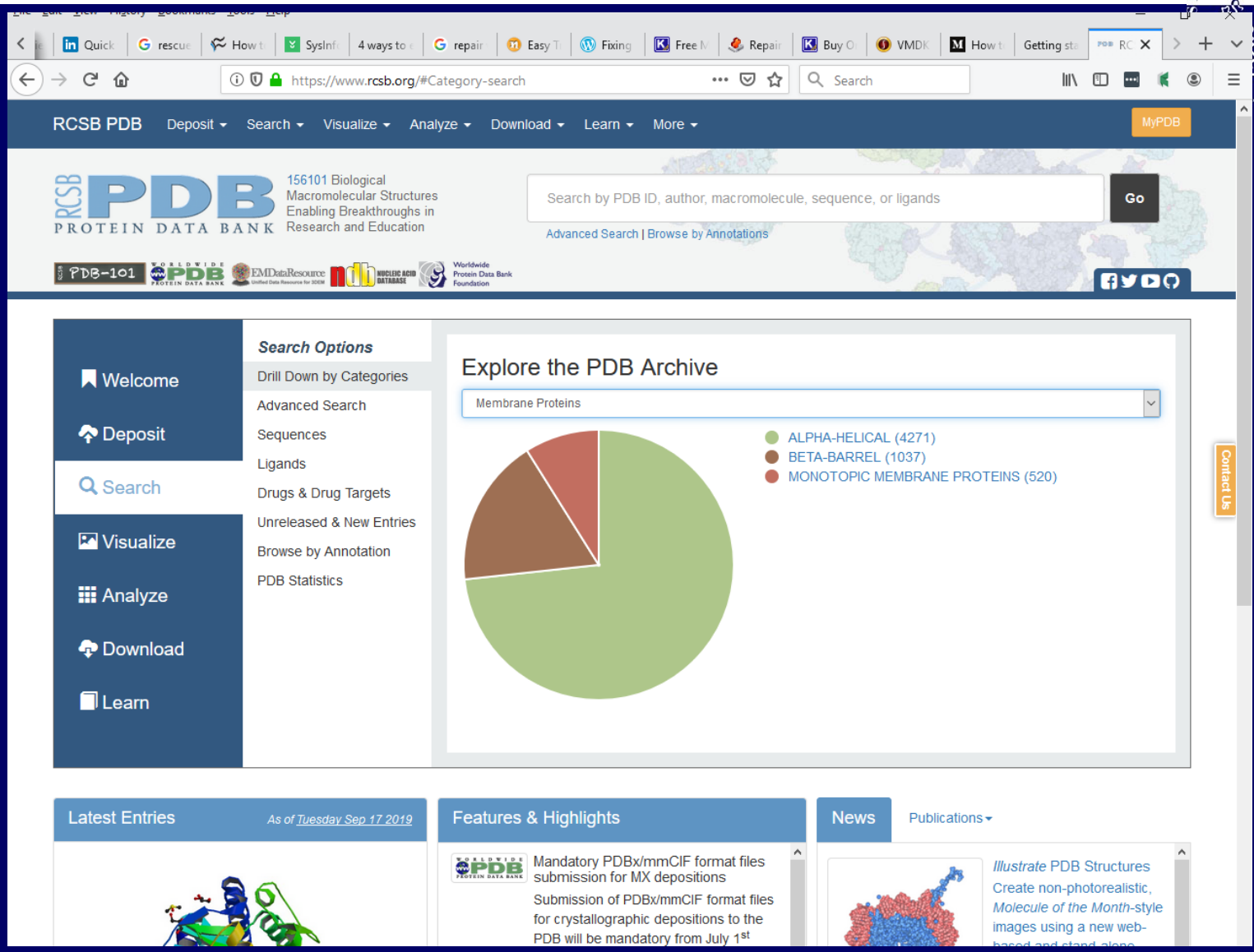
Sunburst Chart for Experimental Method and Molecular Type

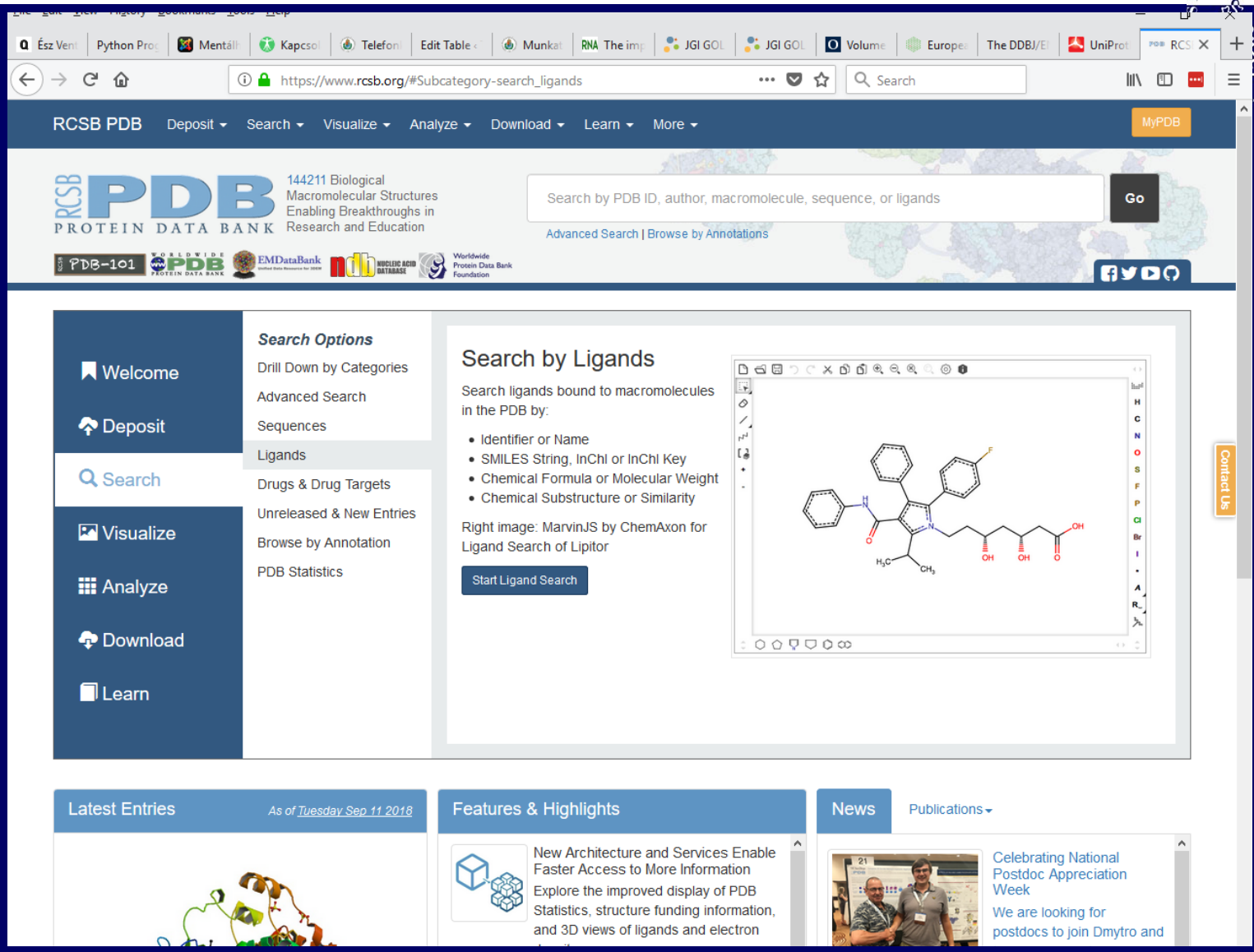
Distribution by experimental method shown in yellow; molecular type in blue. Mouse over to view the statistical description of the particular section; click on a section to zoom in that particular section's statistical distribution, and then click on the center bullseye to zoom out back to the previous distribution.

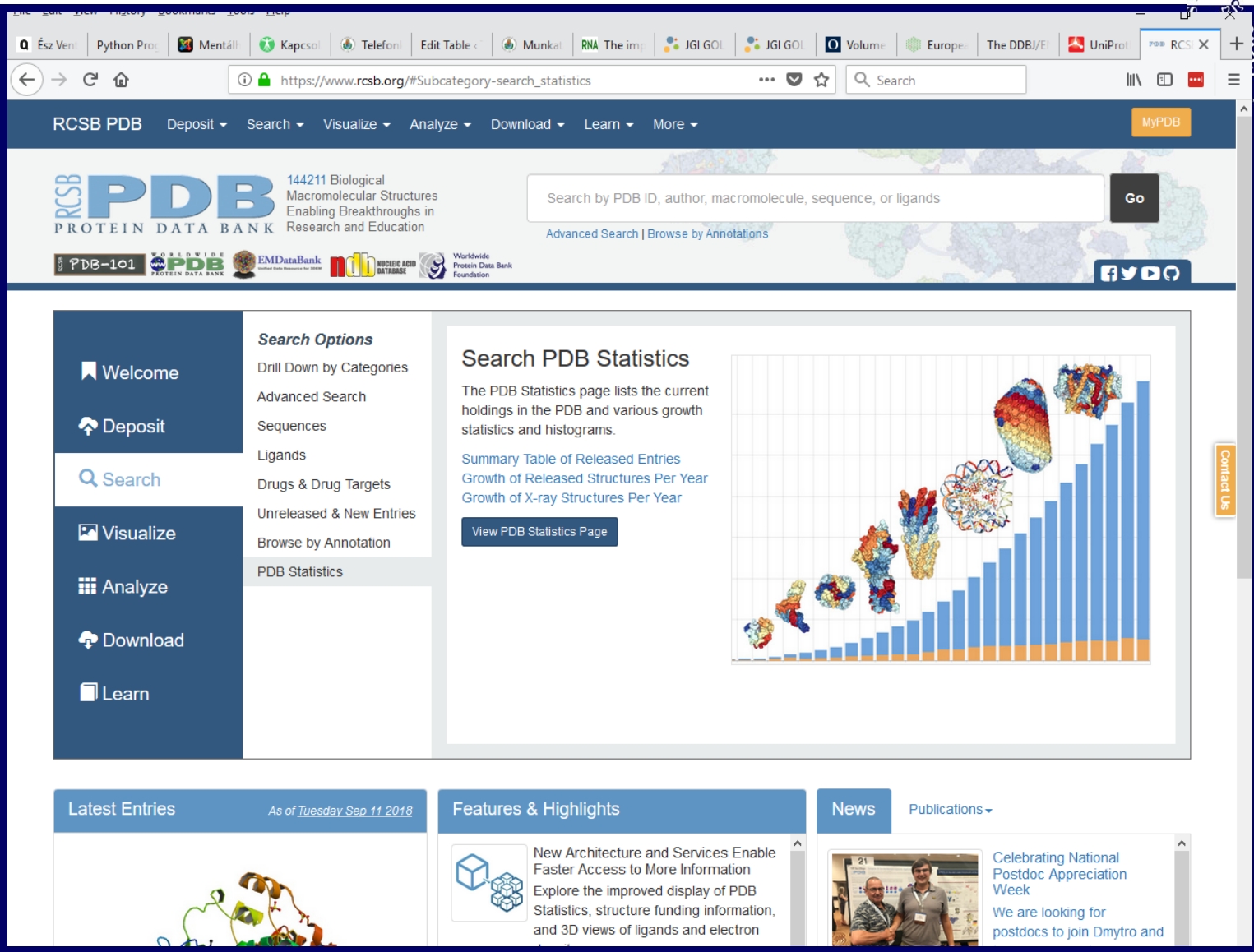
Contact Us


https://www.rcsb.org/#Category-search
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
156101 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Search Options
Drill Down by Categories
Advanced Search
Sequences
Ligands
Drugs & Drug Targets
Unreleased & New Entries
Browse by Annotation
PDB Statistics
Explore the PDB Archive
Organism
Homo sapiens (45332)
Escherichia coli (9752)
Mus musculus (6627)
Saccharomyces cerevisiae (4373)
synthetic construct (4203)
Rattus norvegicus (3120)
Bos taurus (2946)
Other (81034)
Contact Us
Latest Entries
As of Tuesday Sep 17 2019
Features & Highlights
Mandatory PDBx/mmCIF format files submission for MX depositions
Submission of PDBx/mmCIF format files for crystallographic depositions to the PDB will be mandatory from July 1st
News
Publications
Illustrate PDB Structures
Create non-photorealistic, Molecule of the Month-style images using a new web-


https://www.rcsb.org/#Category-search
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
156101 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Search Options
Drill Down by Categories
Advanced Search
Sequences
Ligands
Drugs & Drug Targets
Unreleased & New Entries
Browse by Annotation
PDB Statistics
Explore the PDB Archive
Release Date
before 2000 (10962)
2000 - 2005 (17731)
2005 - 2010 (33073)
2010 - 2015 (43360)
2015 - today (50975)
this year (8603)
this month (697)
Contact Us
Latest Entries
As of Tuesday Sep 17 2019
Features & Highlights
Mandatory PDBx/mmCIF format files submission for MX depositions
Submission of PDBx/mmCIF format files for crystallographic depositions to the PDB will be mandatory from July 1st
News
Publications
Illustrate PDB Structures
Create non-photorealistic, Molecule of the Month-style images using a new web-


https://www.rcsb.org/#Category-search
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
156101 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Search Options
Drill Down by Categories
Advanced Search
Sequences
Ligands
Drugs & Drug Targets
Unreleased & New Entries
Browse by Annotation
PDB Statistics
Explore the PDB Archive
Membrane Proteins
ALPHA-HELICAL (4271)
BETA-BARREL (1037)
MONOTOPIC MEMBRANE PROTEINS (520)
Contact Us
Latest Entries
As of Tuesday Sep 17 2019
Features & Highlights
Mandatory PDBx/mmCIF format files submission for MX depositions
Submission of PDBx/mmCIF format files for crystallographic depositions to the PDB will be mandatory from July 1st
News
Publications
Illustrate PDB Structures
Create non-photorealistic, Molecule of the Month-style images using a new web-


https://www.rcsb.org/#Subcategory-search_ligands
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
144211 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Search Options
Drill Down by Categories
Advanced Search
Sequences
Ligands
Drugs & Drug Targets
Unreleased & New Entries
Browse by Annotation
PDB Statistics
Search by Ligands
Search ligands bound to macromolecules in the PDB by:
Identifier or Name
SMILES String, InChI or InChI Key
Chemical Formula or Molecular Weight
Chemical Substructure or Similarity
Right image: MarvinJS by ChemAxon for Ligand Search of Lipitor
Start Ligand Search
Contact Us
Latest Entries
As of Tuesday Sep 11 2018
Features & Highlights
New Architecture and Services Enable Faster Access to More Information
Explore the improved display of PDB Statistics, structure funding information, and 3D views of ligands and electron
News
Publications
Celebrating National Postdoc Appreciation Week
We are looking for postdocs to join Dmytro and

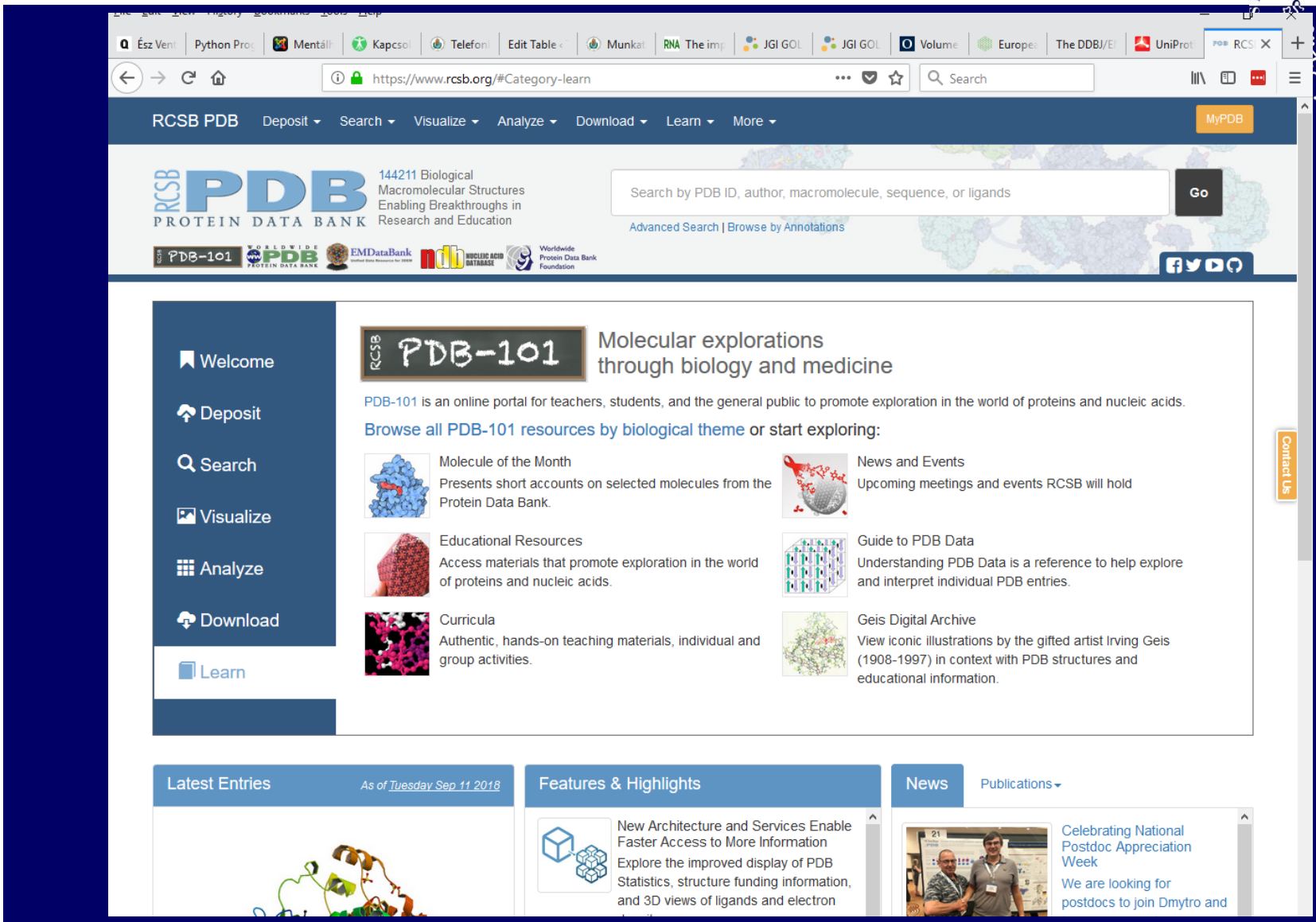

https://www.rcsb.org/#Subcategory-search_statistics
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
144211 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Search Options
Drill Down by Categories
Advanced Search
Sequences
Ligands
Drugs & Drug Targets
Unreleased & New Entries
Browse by Annotation
PDB Statistics
Search PDB Statistics
The PDB Statistics page lists the current holdings in the PDB and various growth statistics and histograms.
Summary Table of Released Entries
Growth of Released Structures Per Year
Growth of X-ray Structures Per Year
View PDB Statistics Page
Contact Us
Latest Entries
As of Tuesday Sep 11 2018
Features & Highlights
New Architecture and Services Enable Faster Access to More Information
Explore the improved display of PDB Statistics, structure funding information, and 3D views of ligands and electron
News
Publications
Celebrating National Postdoc Appreciation Week
We are looking for postdocs to join Dmytro and

```
HEADER    RNA/RNA BINDING PROTEIN                 12-APR-10   3MJ0
TITLE     CRYSTAL STRUCTURE OF DROSOPHIA AGO-PAZ DOMAIN IN COMPLEX WITH 3'-END
TITLE    2 2'-O-METHYLATED RNA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: PROTEIN ARGONAUTE-2;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: UNP RESIDUES 601-723, PAZ DOMAIN;
COMPND   5 ENGINEERED: YES;
COMPND   6 MOL_ID: 2;
COMPND   7 MOLECULE: RNA (5'-R(*CP*GP*UP*UP*AP*CP*GP*CP*(OMU))-3');
COMPND   8 CHAIN: B;
COMPND   9 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: DROSOPHILA MELANOGASTER;
SOURCE   3 ORGANISM_COMMON: FRUIT FLY;
SOURCE   4 ORGANISM_TAXID: 7227;
SOURCE   5 GENE: AGO2, CG7439;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21(DE3)GOLD;
SOURCE   9 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PET28-SMT3;
SOURCE  11 MOL_ID: 2;
SOURCE  12 SYNTHETIC: YES;
SOURCE  13 OTHER_DETAILS: THIS SEQUENCE WAS SYNTHESIZED.
KEYWDS    ARGONAUT, PAZ DOMAIN, 3'-END 2'-O-METHYLATED SSRNA, RNA-RNA BINDING
KEYWDS   2 PROTEIN COMPLEX
EXPDTA    X-RAY DIFFRACTION
AUTHOR    Y.HUANG,L.DING,J.MA
REVDAT   1   25-MAY-11 3MJ0    0
JRNL        AUTH   Y.HUANG,L.DING,J.MA
JRNL        TITL   CRYSTAL STRUCTURE OF DROSOPHILA AGO2 PAZ DOMAIN IN COMPLEX
JRNL        TITL 2 WITH 3'-END 2'-O-METHYLATED RNA
JRNL        REF    TO BE PUBLISHED
JRNL        REFN
REMARK   2
REMARK   2 RESOLUTION.    2.31 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : PHENIX (PHENIX.REFINE: 1.5_2)
REMARK   3   AUTHORS     : PAUL ADAMS,PAVEL AFONINE,VICENT CHEN,IAN
REMARK   3               : DAVIS,KRESHNA GOPAL,RALF GROSSE-
REMARK   3               : KUNSTLEVE,LI-WEI HUNG,ROBERT IMMORMINO,
REMARK   3               : TOM IOERGER,AIRLIE MCCOY,ERIK MCKEE,NIGEL
REMARK   3               : MORIARTY,REETAL PAI,RANDY READ,JANE
REMARK   3               : RICHARDSON,DAVID RICHARDSON,TOD ROMO,JIM
REMARK   3               : SACCHETTINI,NICHOLAS SAUTER,JACOB SMITH,
REMARK   3               : LAURENT STORONI,TOM TERWILLIGER,PETER
REMARK   3               : ZWART
REMARK   3
REMARK   3    REFINEMENT TARGET : MLHL
REMARK   3
```

```
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): 46.54
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): 2.30
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: 50 MM MES, PH 5.8, 100 MM AMMONIUM
REMARK 280  SULFATE, 10 MM MGCL2, 16% PEG 8000 AND 5% GLYCEROL, VAPOR
REMARK 280  DIFFUSION, HANGING DROP, TEMPERATURE 293K
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY
REMARK 290 SYMMETRY OPERATORS FOR SPACE GROUP: C 2 2 21
REMARK 290
REMARK 290      SYMOP   SYMMETRY
REMARK 290     NNNMMM   OPERATOR
REMARK 290       1555   X,Y,Z
REMARK 290       2555   -X,-Y,Z+1/2
REMARK 290       3555   -X,Y,-Z+1/2
REMARK 290       4555   X,-Y,-Z
REMARK 290       5555   X+1/2,Y+1/2,Z
REMARK 290       6555   -X+1/2,-Y+1/2,Z+1/2
REMARK 290       7555   -X+1/2,Y+1/2,-Z+1/2
REMARK 290       8555   X+1/2,-Y+1/2,-Z
REMARK 290
REMARK 290     WHERE NNN -> OPERATOR NUMBER
REMARK 290           MMM -> TRANSLATION VECTOR
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY TRANSFORMATIONS
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE ATOM/HETATM
REMARK 290 RECORDS IN THIS ENTRY TO PRODUCE CRYSTALLOGRAPHICALLY
REMARK 290 RELATED MOLECULES.
REMARK 290   SMTRY1   1  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   1  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   1  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   2 -1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   2  0.000000 -1.000000  0.000000        0.00000
REMARK 290   SMTRY3   2  0.000000  0.000000  1.000000       19.33050
REMARK 290   SMTRY1   3 -1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   3  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   3  0.000000  0.000000 -1.000000       19.33050
REMARK 290   SMTRY1   4  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   4  0.000000 -1.000000  0.000000        0.00000
REMARK 290   SMTRY3   4  0.000000  0.000000 -1.000000        0.00000
REMARK 290   SMTRY1   5  1.000000  0.000000  0.000000       34.60550
REMARK 290   SMTRY2   5  0.000000  1.000000  0.000000       58.39800
REMARK 290   SMTRY3   5  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   6 -1.000000  0.000000  0.000000       34.60550
REMARK 290   SMTRY2   6  0.000000 -1.000000  0.000000       58.39800
REMARK 290   SMTRY3   6  0.000000  0.000000  1.000000       19.33050
```

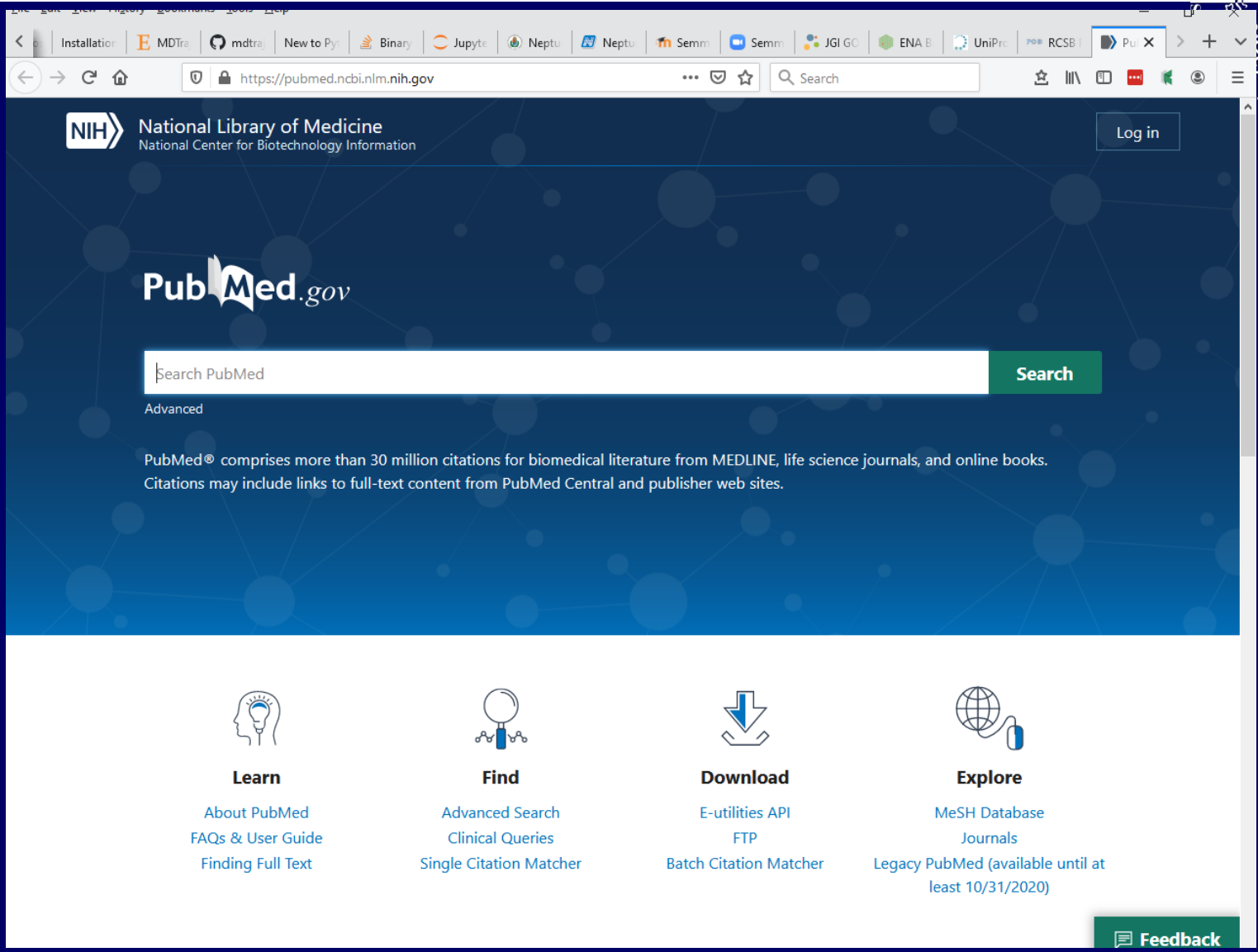
```
REMARK 500 REMARK: NULL
DBREF  3MJ0 A  601   723  UNP    Q9VUQ5   AGO2_DROME     601    723
DBREF  3MJ0 B    1     9  PDB    3MJ0     3MJ0             1      9
SEQADV 3MJ0 SER A  600  UNP  Q9VUQ5              EXPRESSION TAG
SEQRES   1 A  124  SER SER MET PRO MET ILE GLU TYR LEU GLU ARG PHE SER
SEQRES   2 A  124  LEU LYS ALA LYS ILE ASN ASN THR THR ASN LEU ASP TYR
SEQRES   3 A  124  SER ARG ARG PHE LEU GLU PRO PHE LEU ARG GLY ILE ASN
SEQRES   4 A  124  VAL VAL TYR THR PRO PRO GLN SER PHE GLN SER ALA PRO
SEQRES   5 A  124  ARG VAL TYR ARG VAL ASN GLY LEU SER ARG ALA PRO ALA
SEQRES   6 A  124  SER SER GLU THR PHE GLU HIS ASP GLY LYS LYS VAL THR
SEQRES   7 A  124  ILE ALA SER TYR PHE HIS SER ARG ASN TYR PRO LEU LYS
SEQRES   8 A  124  PHE PRO GLN LEU HIS CYS LEU ASN VAL GLY SER SER ILE
SEQRES   9 A  124  LYS SER ILE LEU LEU PRO ILE GLU LEU CYS SER ILE GLU
SEQRES  10 A  124  GLU GLY GLN ALA LEU ASN ARG
SEQRES   1 B    9    C   G   U   U   A   C   G   C OMU
MODRES 3MJ0 OMU B    9    U  O2'-METHYLURIDINE 5'-MONOPHOSPHATE
HET    OMU  B   9      21
HETNAM     OMU O2'-METHYLURIDINE 5'-MONOPHOSPHATE
FORMUL   2  OMU    C10 H15 N2 O9 P
FORMUL   3  HOH   *33(H2 O)
HELIX    1   1 MET A  604  SER A  612  1                                   9
HELIX    2   2 ASN A  622  TYR A  625  5                                   4
HELIX    3   3 SER A  626  ARG A  635  1                                  10
HELIX    4   4 PRO A  644  GLN A  648  5                                   5
HELIX    5   5 ILE A  678  SER A  684  1                                   7
HELIX    6   6 PRO A  709  GLU A  711  5                                   3
SHEET    1   A 4 SER A 601  PRO A 603  0
SHEET    2   A 4 CYS A 713  GLU A 716 -1  O  ILE A 715   N  MET A 602
SHEET    3   A 4 ASN A 638  TYR A 641 -1  N  VAL A 640   O  SER A 714
SHEET    4   A 4 ARG A 652  ARG A 655 -1  O  ARG A 652   N  TYR A 641
SHEET    1   B 3 GLY A 658  PRO A 663  0
SHEET    2   B 3 HIS A 695  ASN A 698 -1  O  CYS A 696   N  SER A 660
SHEET    3   B 3 LEU A 707  LEU A 708 -1  O  LEU A 708   N  LEU A 697
SHEET    1   C 2 THR A 668  GLU A 670  0
SHEET    2   C 2 LYS A 675  THR A 677 -1  O  VAL A 676   N  PHE A 669
LINK         O3'   C B   8                 P   OMU B   9     1555   1555  1.63
CRYST1   69.211  116.796   38.661  90.00  90.00  90.00 C 2 2 21       8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.014449  0.000000  0.000000        0.00000
SCALE2      0.000000  0.008562  0.000000        0.00000
SCALE3      0.000000  0.000000  0.025866        0.00000
```

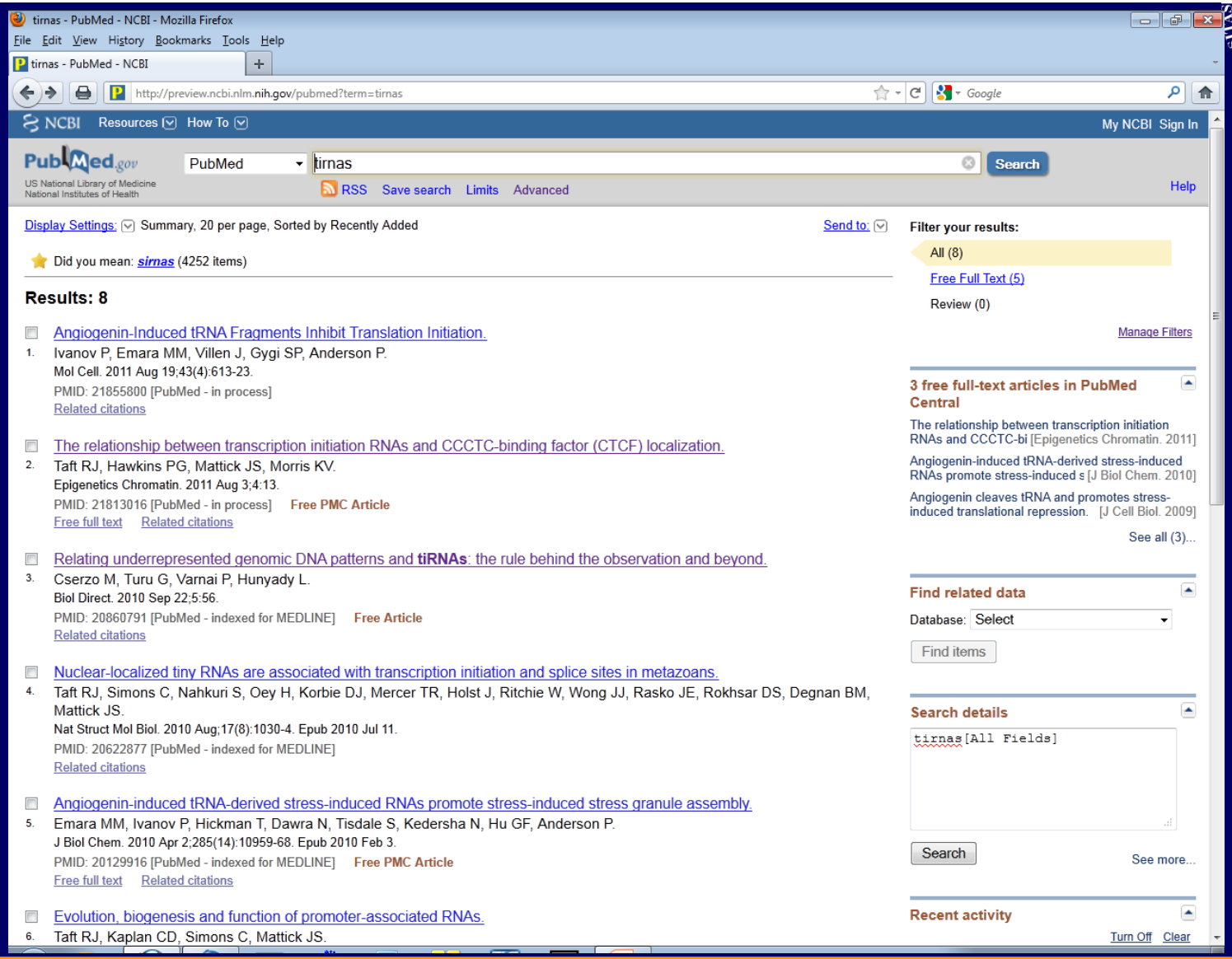
```
ATOM      1  N   SER A 600       6.283   8.861  -8.154  1.00 65.33           N
ATOM      2  CA  SER A 600       6.119   9.334  -6.784  1.00 68.88           C
ATOM      3  C   SER A 600       7.307   8.914  -5.918  1.00 67.53           C
ATOM      4  O   SER A 600       8.330   8.478  -6.445  1.00 70.29           O
ATOM      5  CB  SER A 600       4.801   8.828  -6.187  1.00 74.92           C
ATOM      6  OG  SER A 600       4.734   7.412  -6.204  1.00 74.42           O
ATOM      7  N   SER A 601       7.163   9.037  -4.599  1.00 62.45           N
ATOM      8  CA  SER A 601       8.284   8.835  -3.687  1.00 61.77           C
ATOM      9  C   SER A 601       7.857   8.445  -2.278  1.00 65.47           C
ATOM     10  O   SER A 601       6.713   8.671  -1.876  1.00 66.71           O
ATOM     11  CB  SER A 601       9.130  10.107  -3.598  1.00 63.47           C
ATOM     12  OG  SER A 601       8.456  11.097  -2.843  1.00 60.71           O
ATOM     13  N   MET A 602       8.801   7.879  -1.525  1.00 60.84           N
ATOM     14  CA  MET A 602       8.575   7.513  -0.132  1.00 55.67           C
ATOM     15  C   MET A 602       9.552   8.232   0.804  1.00 52.46           C
ATOM     16  O   MET A 602      10.734   8.346   0.502  1.00 51.93           O
ATOM     17  CB  MET A 602       8.688   5.997   0.040  1.00 53.12           C
ATOM     18  CG  MET A 602       8.692   5.547   1.492  1.00 60.07           C
ATOM     19  SD  MET A 602       8.480   3.773   1.748  1.00 67.00           S
ATOM     20  CE  MET A 602      10.063   3.122   1.252  1.00 64.41           C
ATOM     21  N   PRO A 603       9.050   8.722   1.944  1.00 50.54           N
ATOM     22  CA  PRO A 603       9.879   9.431   2.925  1.00 48.35           C
ATOM     23  C   PRO A 603      10.992   8.549   3.437  1.00 52.20           C
ATOM     24  O   PRO A 603      10.730   7.427   3.839  1.00 58.02           O
ATOM     25  CB  PRO A 603       8.911   9.703   4.072  1.00 52.67           C
ATOM     26  CG  PRO A 603       7.553   9.661   3.455  1.00 55.93           C
ATOM     27  CD  PRO A 603       7.637   8.656   2.349  1.00 53.63           C
ATOM     28  N   MET A 604      12.215   9.057   3.441  1.00 51.47           N
ATOM     29  CA  MET A 604      13.333   8.303   3.971  1.00 52.57           C
ATOM     30  C   MET A 604      13.072   7.851   5.388  1.00 55.38           C
ATOM     31  O   MET A 604      13.595   6.829   5.819  1.00 58.51           O
ATOM     32  CB  MET A 604      14.627   9.115   3.912  1.00 51.24           C
ATOM     33  CG  MET A 604      15.308   9.073   2.547  1.00 50.14           C
ATOM     34  SD  MET A 604      15.618   7.377   1.994  1.00 53.74           S
ATOM     35  CE  MET A 604      16.596   6.758   3.344  1.00 52.26           C
ATOM     36  N   ILE A 605      12.264   8.608   6.120  1.00 54.74           N
ATOM     37  CA  ILE A 605      11.974   8.251   7.499  1.00 52.91           C
ATOM     38  C   ILE A 605      11.214   6.923   7.558  1.00 61.84           C
ATOM     39  O   ILE A 605      11.549   6.020   8.326  1.00 59.30           O
ATOM     40  CB  ILE A 605      11.154   9.335   8.189  1.00 55.89           C
ATOM     41  CG1 ILE A 605      12.020  10.575   8.434  1.00 51.55           C
ATOM     42  CG2 ILE A 605      10.608   8.809   9.504  1.00 60.92           C
ATOM     43  CD1 ILE A 605      13.120  10.358   9.463  1.00 54.08           C
ATOM     44  N   GLU A 606      10.216   6.816   6.702  1.00 59.72           N
ATOM     45  CA  GLU A 606       9.446   5.578   6.594  1.00 61.38           C
ATOM     46  C   GLU A 606      10.371   4.511   6.022  1.00 63.26           C
ATOM     47  O   GLU A 606      10.590   3.464   6.630  1.00 63.43           O
```


https://www.rcsb.org/#Category-learn
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB PROTEIN DATA BANK
144211 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
PDB-101
Molecular explorations through biology and medicine
PDB-101 is an online portal for teachers, students, and the general public to promote exploration in the world of proteins and nucleic acids.
Browse all PDB-101 resources by biological theme or start exploring:
Molecule of the Month
Presents short accounts on selected molecules from the Protein Data Bank.
News and Events
Upcoming meetings and events RCSB will hold
Educational Resources
Access materials that promote exploration in the world of proteins and nucleic acids.
Guide to PDB Data
Understanding PDB Data is a reference to help explore and interpret individual PDB entries.
Curricula
Authentic, hands-on teaching materials, individual and group activities.
Geis Digital Archive
View iconic illustrations by the gifted artist Irving Geis (1908-1997) in context with PDB structures and educational information.
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
Latest Entries
As of Tuesday Sep 11 2018
Features & Highlights
New Architecture and Services Enable Faster Access to More Information
Explore the improved display of PDB Statistics, structure funding information, and 3D views of ligands and electron
News
Publications
Celebrating National Postdoc Appreciation Week
We are looking for postdocs to join Dmytro and
Contact Us

PubMed adatbázis

- Az NCBI szolgáltatja az adatbázist
- Az elérés ingyenes: http://www.ncbi.nlm.nih.gov/pubmed
- A tudományos közlemények fő forrása
- Integrált adatbázis: a felületen keresztül fehérje és nukleinsav adatbázisok is elérhetők


https://pubmed.ncbi.nlm.nih.gov
Search
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
PubMed.gov
Search PubMed
Search
Advanced
PubMed® comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books.
Citations may include links to full-text content from PubMed Central and publisher web sites.
Learn
About PubMed
FAQs & User Guide
Finding Full Text
Find
Advanced Search
Clinical Queries
Single Citation Matcher
Download
E-utilities API
FTP
Batch Citation Matcher
Explore
MeSH Database
Journals
Legacy PubMed (available until at least 10/31/2020)
Feedback


tirnas - PubMed - NCBI - Mozilla Firefox
File Edit View History Bookmarks Tools Help
tirnas - PubMed - NCBI
http://preview.ncbi.nlm.nih.gov/pubmed?term=tirnas
NCBI Resources How To My NCBI Sign In
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
tirnas
Search
RSS Save search Limits Advanced
Help
Display Settings: Summary, 20 per page, Sorted by Recently Added
Send to:
Did you mean: sirnas (4252 items)
Results: 8
Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation.
1. Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P.
Mol Cell. 2011 Aug 19;43(4):613-23.
PMID: 21855800 [PubMed - in process]
Related citations
The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization.
2. Taft RJ, Hawkins PG, Mattick JS, Morris KV.
Epigenetics Chromatin. 2011 Aug 3;4:13.
PMID: 21813016 [PubMed - in process] Free PMC Article
Free full text Related citations
Relating underrepresented genomic DNA patterns and tiRNAs: the rule behind the observation and beyond.
3. Cserzo M, Turu G, Varnai P, Hunyady L.
Biol Direct. 2010 Sep 22;5:56.
PMID: 20860791 [PubMed - indexed for MEDLINE] Free Article
Related citations
Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans.
4. Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE, Rokhsar DS, Degnan BM, Mattick JS.
Nat Struct Mol Biol. 2010 Aug;17(8):1030-4. Epub 2010 Jul 11.
PMID: 20622877 [PubMed - indexed for MEDLINE]
Related citations
Angiogenin-induced tRNA-derived stress-induced RNAs promote stress-induced stress granule assembly.
5. Emara MM, Ivanov P, Hickman T, Dawra N, Tisdale S, Kedersha N, Hu GF, Anderson P.
J Biol Chem. 2010 Apr 2;285(14):10959-68. Epub 2010 Feb 3.
PMID: 20129916 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations
Evolution, biogenesis and function of promoter-associated RNAs.
6. Taft RJ, Kaplan CD, Simons C, Mattick JS.
Filter your results:
All (8)
Free Full Text (5)
Review (0)
Manage Filters
3 free full-text articles in PubMed Central
The relationship between transcription initiation RNAs and CCCTC-bi [Epigenetics Chromatin. 2011]
Angiogenin-induced tRNA-derived stress-induced RNAs promote stress-induced s [J Biol Chem. 2010]
Angiogenin cleaves tRNA and promotes stress-induced translational repression. [J Cell Biol. 2009]
See all (3)...
Find related data
Database: Select
Find items
Search details
tirnas[All Fields]
Search
See more...
Recent activity
Turn Off Clear

The relationship between transcription... [Epigenetics Chromatin. 2011] - PubMed - NCBI - Mozilla Firefox

File Edit View History Bookmarks Tools Help

The relationship between transcription.....

http://preview.ncbi.nlm.nih.gov/pubmed/21813016

NCBI Resources How To My NCBI Sign In

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed Search

Limits Advanced Help

Display Settings: Abstract Send to:

Read free full text at BioMed Central FREE full text article in PubMed Central

Epigenetics Chromatin. 2011 Aug 3;4:13.

The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization.

Taft RJ, Hawkins PG, Mattick JS, Morris KV.

Department of Molecular and Experimental Medicine, The Kellogg School of Science and Technology, The Scripps Research Institute, La Jolla, CA 92037, USA. kmorris@scripps.edu.

Abstract

ABSTRACT:

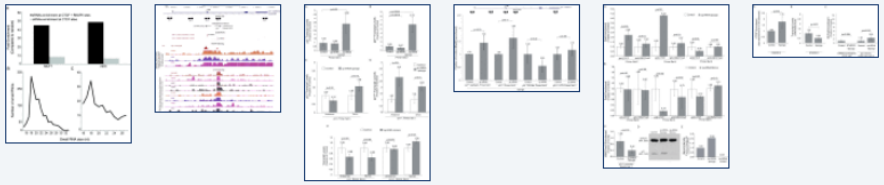
BACKGROUND: Transcription initiation RNAs (tiRNAs) are nuclear localized 18 nucleotide RNAs derived from sequences immediately downstream of RNA polymerase II (RNAPII) transcription start sites. Previous reports have shown that tiRNAs are intimately correlated with gene expression, RNA polymerase II binding and behaviors, and epigenetic marks associated with transcription initiation, but not elongation.

RESULTS: In the present work, we show that tiRNAs are commonly found at genomic CCCTC-binding factor (CTCF) binding sites in human and mouse, and that CTCF sites that colocalize with RNAPII are highly enriched for tiRNAs. To directly investigate the relationship between tiRNAs and CTCF we examined tiRNAs originating near the intronic CTCF binding site in the human tumor suppressor gene, p21 (cyclin-dependent kinase inhibitor 1A gene, also known as CDKN1A). Inhibition of CTCF-proximal tiRNAs resulted in increased CTCF localization and increased p21 expression, while overexpression of CTCF-proximal tiRNA mimics decreased CTCF localization and p21 expression. We also found that tiRNA-regulated CTCF binding influences the levels of trimethylated H3K27 at the alternate upstream p21 promoter, and affects the levels of alternate p21 (p21alt) transcripts. Extending these studies to another randomly selected locus with conserved CTCF binding we found that depletion of tiRNA alters nucleosome density proximal to sites of tiRNA biogenesis.

CONCLUSIONS: Taken together, these data suggest that tiRNAs modulate local epigenetic structure, which in turn regulates CTCF localization.

PMID: 21813016 [PubMed - in process] PMCID: PMC3170176 Free PMC Article

Images from this publication. See all images (7) Free text



LinkOut - more resources

Related citations

Critical role for the transcription regulator CCCTC-binding factor in the control of Th; [J Immunol. 2009]

Regulation of eye development by transcription control of CCCTC binding factor [J Biol Chem. 2004]

CCCTC-binding factor (CTCF) and cohesin influence the geno [Proc Natl Acad Sci U S A. 2011]

Review CCCTC-binding factor: to loop or to bridge. [Cell Mol Life Sci. 2009]

Review Modulation of thyroid hormone receptor silencing function by co- [Biochem Soc Trans. 2000]

See reviews...

See all...

Related information

Related Citations

References for this PMC Article

Free in PMC

Recent activity

Turn Off Clear

The relationship between transcription initiation RNAs and CCCTC-binding factor ... PubMed

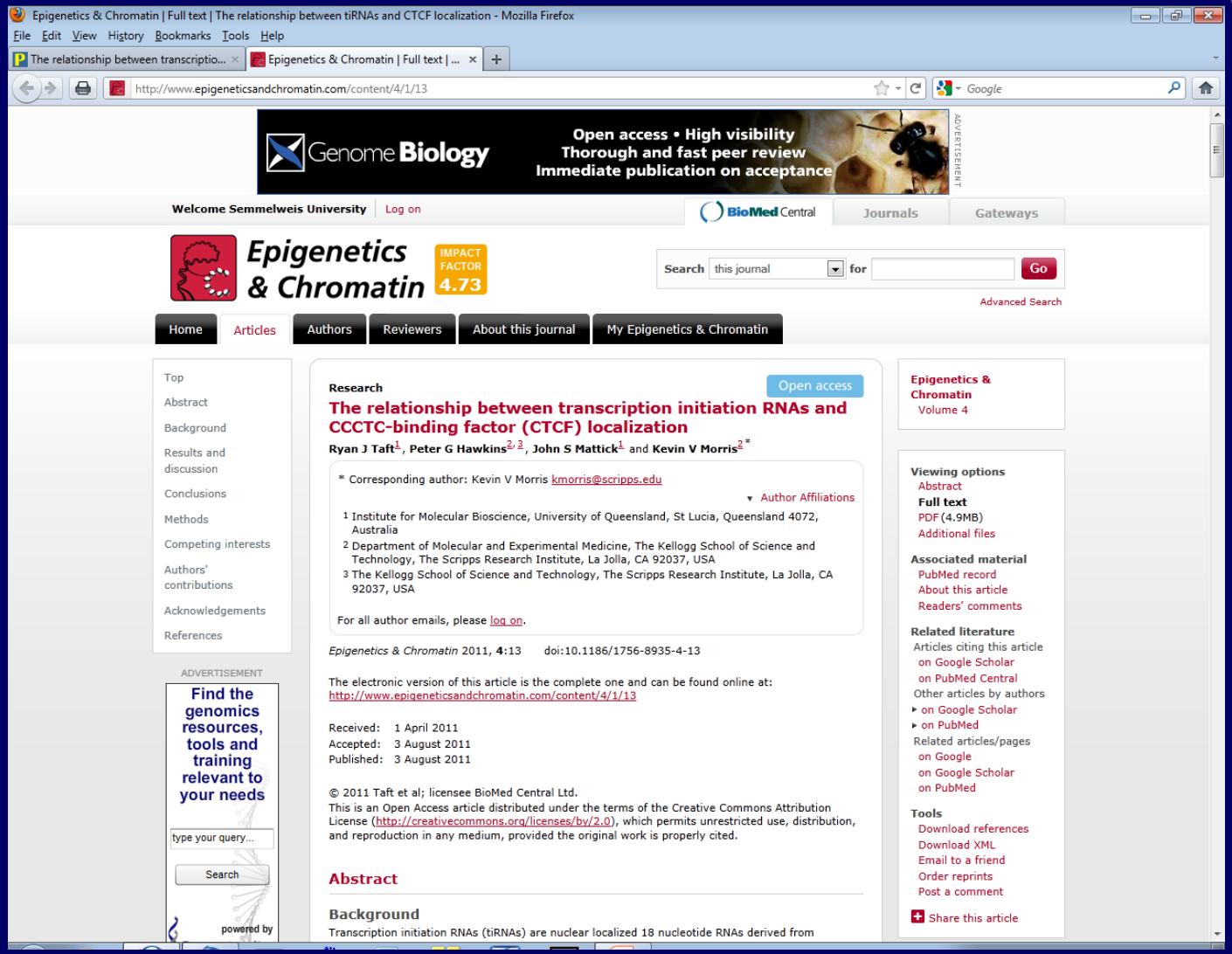
tirnas (8) PubMed

"Sci Am"[jour] (4626) PubMed

Dayhoff M[Author] (42) PubMed

Computer analysis of protein evolution. PubMed

See more...



Epigenetics & Chromatin | Full text | The relationship between tiRNAs and CTCF localization - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.epigeneticsandchromatin.com/content/4/1/13

Genome Biology — Open access • High visibility Thorough and fast peer review Immediate publication on acceptance

Welcome Semmelweis University | Log on

BioMed Central | Journals | Gateways

Epigenetics & Chromatin — IMPACT FACTOR 4.73

Search this journal for Go — Advanced Search

Home | Articles | Authors | Reviewers | About this journal | My Epigenetics & Chromatin

Top
Abstract
Background
Results and discussion
Conclusions
Methods
Competing interests
Authors' contributions
Acknowledgements
References

ADVERTISEMENT

Find the genomics resources, tools and training relevant to your needs

type your query...

Search

powered by

Research

Open access

The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization

Ryan J Taft[1], Peter G Hawkins[2,3], John S Mattick[1] and Kevin V Morris[2]*

* Corresponding author: Kevin V Morris kmorris@scripps.edu

Author Affiliations

1 Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

2 Department of Molecular and Experimental Medicine, The Kellogg School of Science and Technology, The Scripps Research Institute, La Jolla, CA 92037, USA

3 The Kellogg School of Science and Technology, The Scripps Research Institute, La Jolla, CA 92037, USA

For all author emails, please log on.

Epigenetics & Chromatin 2011, **4**:13 doi:10.1186/1756-8935-4-13

The electronic version of this article is the complete one and can be found online at: http://www.epigeneticsandchromatin.com/content/4/1/13

Received: 1 April 2011
Accepted: 3 August 2011
Published: 3 August 2011

© 2011 Taft et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background

Transcription initiation RNAs (tiRNAs) are nuclear localized 18 nucleotide RNAs derived from

Epigenetics & Chromatin
Volume 4

Viewing options
Abstract
Full text
PDF (4.9MB)
Additional files

Associated material
PubMed record
About this article
Readers' comments

Related literature
Articles citing this article
on Google Scholar
on PubMed Central
Other articles by authors
on Google Scholar
on PubMed
Related articles/pages
on Google
on Google Scholar
on PubMed

Tools
Download references
Download XML
Email to a friend
Order reprints
Post a comment

Share this article


```
PMID- 21813016
OWN - NLM
STAT- In-Data-Review
DA  - 20110912
IS  - 1756-8935 (Electronic)
IS  - 1756-8935 (Linking)
VI  - 4
DP  - 2011
TI  - The relationship between transcription initiation RNAs and CCCTC-binding factor
      (CTCF) localization.
PG  - 13
AB  - ABSTRACT: BACKGROUND: Transcription initiation RNAs (tiRNAs) are nuclear
      localized 18 nucleotide RNAs derived from sequences immediately downstream of RNA
      polymerase II (RNAPII) transcription start sites. Previous reports have shown
      that tiRNAs are intimately correlated with gene expression, RNA polymerase II
      binding and behaviors, and epigenetic marks associated with transcription
      initiation, but not elongation. RESULTS: In the present work, we show that tiRNAs
      are commonly found at genomic CCCTC-binding factor (CTCF) binding sites in human
      and mouse, and that CTCF sites that colocalize with RNAPII are highly enriched
      for tiRNAs. To directly investigate the relationship between tiRNAs and CTCF we
      examined tiRNAs originating near the intronic CTCF binding site in the human
      tumor suppressor gene, p21 (cyclin-dependent kinase inhibitor 1A gene, also known
      as CDKN1A). Inhibition of CTCF-proximal tiRNAs resulted in increased CTCF
      localization and increased p21 expression, while overexpression of CTCF-proximal
      tiRNA mimics decreased CTCF localization and p21 expression. We also found that
      tiRNA-regulated CTCF binding influences the levels of trimethylated H3K27 at the
      alternate upstream p21 promoter, and affects the levels of alternate p21 (p21alt)
      transcripts. Extending these studies to another randomly selected locus with
      conserved CTCF binding we found that depletion of tiRNA alters nucleosome density
      proximal to sites of tiRNA biogenesis. CONCLUSIONS: Taken together, these data
      suggest that tiRNAs modulate local epigenetic structure, which in turn regulates
      CTCF localization.
AD  - Department of Molecular and Experimental Medicine, The Kellogg School of Science
      and Technology, The Scripps Research Institute, La Jolla, CA 92037, USA.
      kmorris@scripps.edu.
FAU - Taft, Ryan J
AU  - Taft RJ
FAU - Hawkins, Peter G
AU  - Hawkins PG
FAU - Mattick, John S
AU  - Mattick JS
FAU - Morris, Kevin V
AU  - Morris KV
LA  - eng
```

Hibák az adatbázisban

- Az adatbázisok felhasználják egymás adatait
- Ha egy hiba bekerül – nem tudni, milyen további hibákat eredményez
- Ördögi körök veszélye
- A hibákat kezelni kell:
  - Rendszeres frissítések
  - A frissítés előtti állapot is látszik az új mellett
- Az adatokat nem lehet kritika nélkül elfogadni

A hibák eredete:

- Kisérleti technikából eredő hiba
  - Az adatok hosszú idő óta gyűlnek
  - Közben változik a technika
  - Változik a megbízhatóság
- Hibás predikció
  - A másodlagos adatbázisok predikción alapulnak
  - 99%-os hatékonyság – az emberi genom esetén 250 gént hibásan annotálunk!
- Hiányos tudás
  - 'Aki keres az talál.' – pl. cisPro

Lehetséges megoldás

- Emberi beavatkozás – teljesen reménytelen
  - Túl sok az adat és vannak emberi hibák is
- Kollektív bölcsesség – wikipédia
  - Kiegyenlíti a hibákat, de új problémákat generál
- A meglévő predikciós módszerek javítása
  - Igen lassú és körülményes
- Külön hibajavító – ellenőrző módszerek kidolgozása
  - Hibajavítás szakértői rendszerrel

A MisPred adatbázis

- Patthy László nevéhez fűződik
- Elérhető: http://mispred.com/
- Elve: a fehérje annotáció különböző elemei közt van-e konfliktus?
  1. Extracelluláris domén van, de nincs szignál peptid
  2. Extracelluláris és citoplaznatikus domén van, de nincs transzmembrán szakasz
  3. Nukleáris és extracelluláris domén is van
  4. Domén méret jelentős eltérése a család többi tagjához képest
  5. Egy gén csak egy kromoszómán lehet jelen


European Nucleotide Ar...
The DDBJ/EMBL/GenBank F...
Current Release Statistic...
RCSB Protein Data Bank...
tirnas - PubMed - NCBI
Home
www.mispred.com
gold
Log In
MisPred
HOME ABOUT MISPRED SEARCH MISPRED ANALYZE YOUR SEQUENCE RECENT RESULTS CONTACT AND FEEDBACK
MisPred 8.0 (May 2014)
The MisPred database is a collection of protein sequences identified by the MisPred pipeline as erroneous (incomplete, abnormal or mispredicted). The MisPred database may be searched in various ways (SEARCH MISPRED) and you may also analyze your sequence using the MisPred pipeline to decide whether your sequence violates some basic rules of protein sequences (ANALYZE YOUR SEQUENCE).
Citing MisPred
If you find MisPred useful, please consider citing the references that describe this work:
Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008; 9:353. http://www.ncbi.nlm.nih.gov/pubmed/18752676
Nagy A. and Patthy L. MisPred: a resource for identification of erroneous protein sequences in public databases. Database (2013) Vol. 2013: article ID bat053; doi:10.1093/database/bat053 http://database.oxfordjournals.org/content/2013/bat053.full
NIH National Innovation Office
The FixPred9 project is funded by the National Innovation Office of Hungary through grant FixPred9 "Exploiting genome information", contract number TECH_09_A1-FixPred9.
Új Magyarország FEJLESZTÉSI TERV
This page Copyright © 2011 by Combit Számítástechnikai Zrt.

European Nucleotide Ar... | The DDBJ/EMBL/GenBank F... | Current Release Statistic... | RCSB Protein Data Bank... | tirnas - PubMed - NCBI | MisPred: a resource for iden...

database.oxfordjournals.org/content/2013/bat053.full

OXFORD JOURNALS — CONTACT US | MY BASKET | MY ACCOUNT

DATABASE The Journal of Biological Databases and Curation

ABOUT THIS JOURNAL | CONTACT THIS JOURNAL | SUBSCRIPTIONS | CURRENT ISSUE | ARCHIVE | SEARCH

Institution: Semmelweis Ote | Sign In as Personal Subscriber

Oxford Journals › Science & Mathematics › Database › Volume 2013 › 10.1093/database/bat053

MisPred: a resource for identification of erroneous protein sequences in public databases

Alinda Nagy and László Patthy

Author Affiliations

*Corresponding author: Tel: +361 279 3100; Fax: +361 466 5465; Email: patthy.laszlo@ttk.mta.hu

Received March 21, 2013.
Revision received June 12, 2013.
Accepted June 13, 2013.

Abstract

Correct prediction of the structure of protein-coding genes of higher eukaryotes is still a difficult task; therefore, public databases are heavily contaminated with mispredicted sequences. The high rate of misprediction has serious consequences because it significantly affects the conclusions that may be drawn from genome-scale sequence analyses of eukaryotic genomes. Here we present the MisPred database and computational pipeline that provide efficient means for the identification of erroneous sequences in public databases. The MisPred database contains a collection of abnormal, incomplete and mispredicted protein sequences from 19 metazoan species identified as erroneous by MisPred quality control tools in the UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, NCBI/RefSeq and EnsEMBL databases. Major releases of the database are automatically generated and updated regularly. The database (http://www.mispred.com) is easily accessible through a simple web interface coupled to a powerful query engine and a standard web service. The content is completely or partially downloadable in a variety of formats.

Database URL: http://www.mispred.com

« Previous | Next Article » Table of Contents

This Article

Database (2013) 2013 : bat053 doi: 10.1093/database/bat053

This article is Open Access
Abstract *Free*
» Full Text (HTML) *Free*
Full Text (PDF) *Free*
Supplementary Data

Classifications
Database Tool

Services
Alert me when cited
Alert me if corrected
Alert me if commented
Find similar articles
Similar articles in PubMed
Add to my archive
Download citation
Request Permissions

Responses
Citing Articles
Google Scholar
PubMed
Share

Search this journal: GO
Advanced »

Current Content
Volume 2014

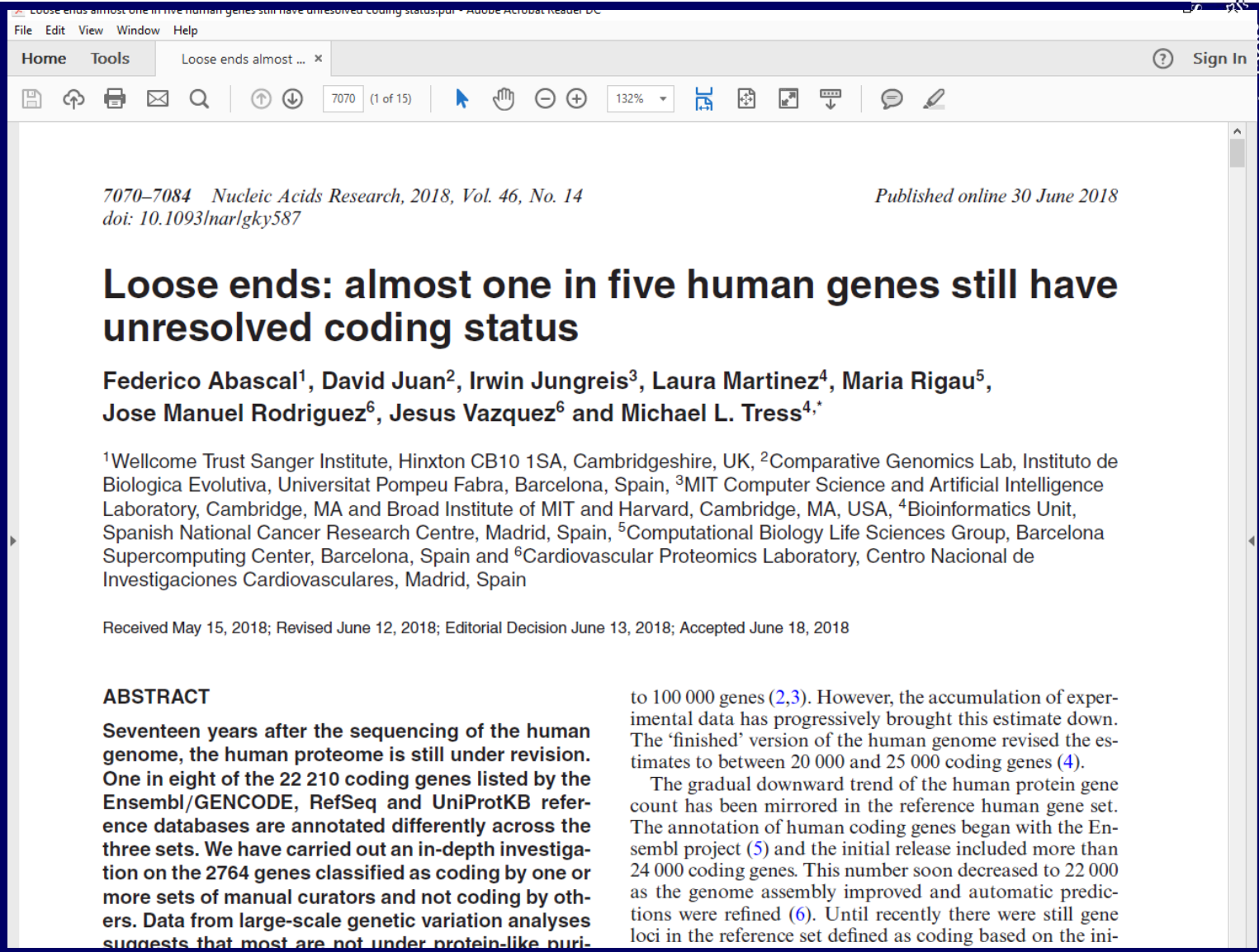
Alert me to new content

The Journal
About the journal
Rights & permissions
Recent Comments
We are mobile – find out more
This journal is a member of the Committee on Publication Ethics (COPE)

BioCreative Virtual Issue
BIOCURATION VIRTUAL ISSUE
BioMart virtual issue

S.E. received 1 new message
Re: [ensembl-dev] mysql --ssl-ca option Andy Yates
Hi, I think this will require a hack of the code in a few locations. As far as I am a...

Oxford Index | About the Index



7070–7084 Nucleic Acids Research, 2018, Vol. 46, No. 14
doi: 10.1093/nar/gky587

Published online 30 June 2018

Loose ends: almost one in five human genes still have unresolved coding status

Federico Abascal[1], David Juan[2], Irwin Jungreis[3], Laura Martinez[4], Maria Rigau[5], Jose Manuel Rodriguez[6], Jesus Vazquez[6] and Michael L. Tress[4,*]

[1]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK, [2]Comparative Genomics Lab, Instituto de Biologica Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain, [3]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA and Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Bioinformatics Unit, Spanish National Cancer Research Centre, Madrid, Spain, [5]Computational Biology Life Sciences Group, Barcelona Supercomputing Center, Barcelona, Spain and [6]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain

Received May 15, 2018; Revised June 12, 2018; Editorial Decision June 13, 2018; Accepted June 18, 2018

ABSTRACT

Seventeen years after the sequencing of the human genome, the human proteome is still under revision. One in eight of the 22 210 coding genes listed by the Ensembl/GENCODE, RefSeq and UniProtKB reference databases are annotated differently across the three sets. We have carried out an in-depth investigation on the 2764 genes classified as coding by one or more sets of manual curators and not coding by others. Data from large-scale genetic variation analyses suggests that most are not under protein-like puri-

to 100 000 genes (2,3). However, the accumulation of experimental data has progressively brought this estimate down. The 'finished' version of the human genome revised the estimates to between 20 000 and 25 000 coding genes (4).

The gradual downward trend of the human protein gene count has been mirrored in the reference human gene set. The annotation of human coding genes began with the Ensembl project (5) and the initial release included more than 24 000 coding genes. This number soon decreased to 22 000 as the genome assembly improved and automatic predictions were refined (6). Until recently there were still gene loci in the reference set defined as coding based on the ini-

Mit tanultunk ma?

- Az adatbázis struktúrált információ
- Bejegyzésekből és mezőkből áll
- Formátuma kötött
- De azért rugalmas is
- Az adatok hibásak lehetnek – kritikával használjuk

Feladat 2.

Tervezzünk adatbázisformátumot, amely alkalmas a budapesti kávézók (fagyizók, italmérések,...) nyilvántartására.

Tárgymező: ‘Kurzus – Feladat 2.’