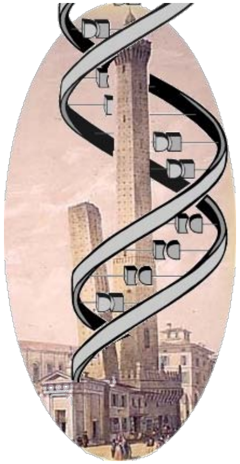
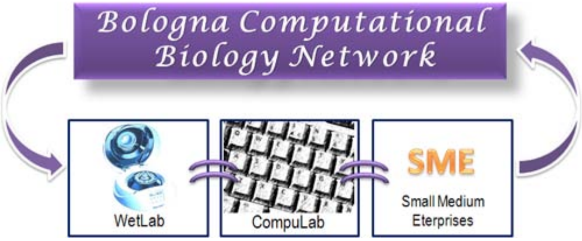
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Biological Databases

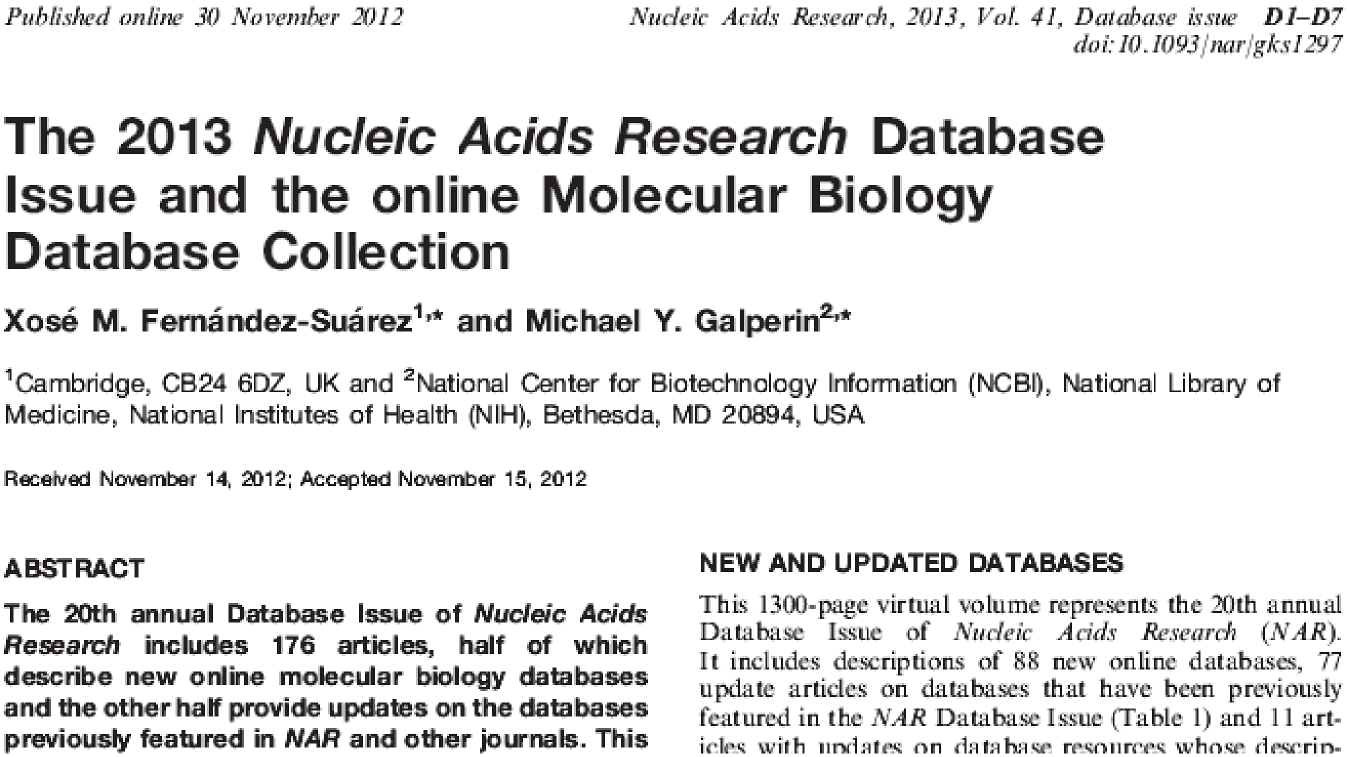


Rita Casadio



BIOCOMPUTING GROUP
University of Bologna, Italy

AIRBBC



Published online 30 November 2012

Nucleic Acids Research, 2013, Vol. 41, Database issue D1–D7
doi:10.1093/nar/gks1297

The 2013 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection

Xosé M. Fernández-Suárez[1],* and Michael Y. Galperin[2],*

[1]Cambridge, CB24 6DZ, UK and [2]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received November 14, 2012; Accepted November 15, 2012

ABSTRACT

The 20th annual Database Issue of *Nucleic Acids Research* includes 176 articles, half of which describe new online molecular biology databases and the other half provide updates on the databases previously featured in *NAR* and other journals. This

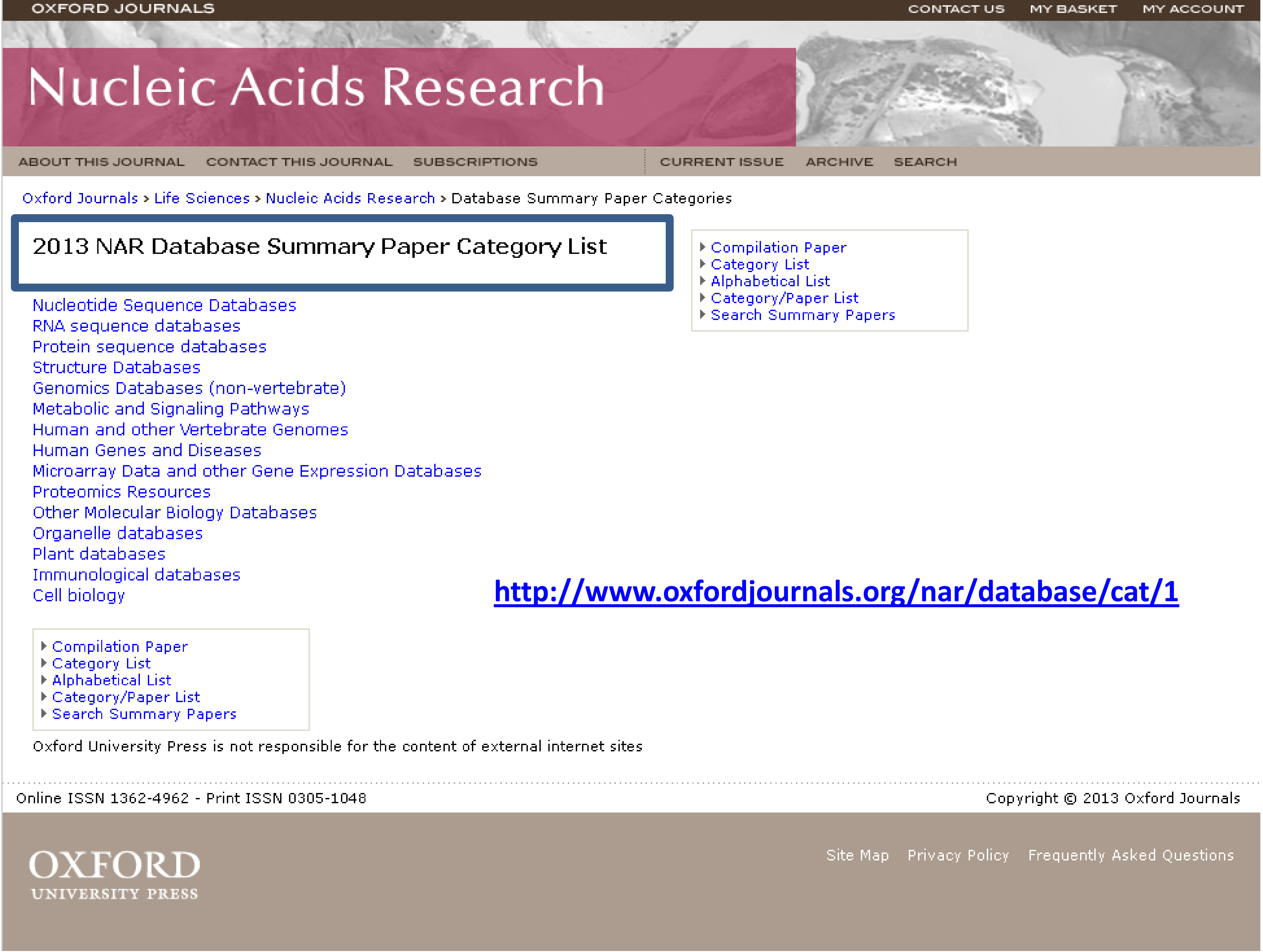
NEW AND UPDATED DATABASES

This 1300-page virtual volume represents the 20th annual Database Issue of *Nucleic Acids Research* (*NAR*). It includes descriptions of 88 new online databases, 77 update articles on databases that have been previously featured in the *NAR* Database Issue (Table 1) and 11 articles with updates on database resources whose descrip-

The NAR online Molecular Biology Database Collection,

available at **http://www.oxfordjournals.org/nar/database/a/**

has been updated and currently lists 1512 online databases.



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

2013 NAR Database Summary Paper Category List

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

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

http://www.oxfordjournals.org/nar/database/cat/1

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

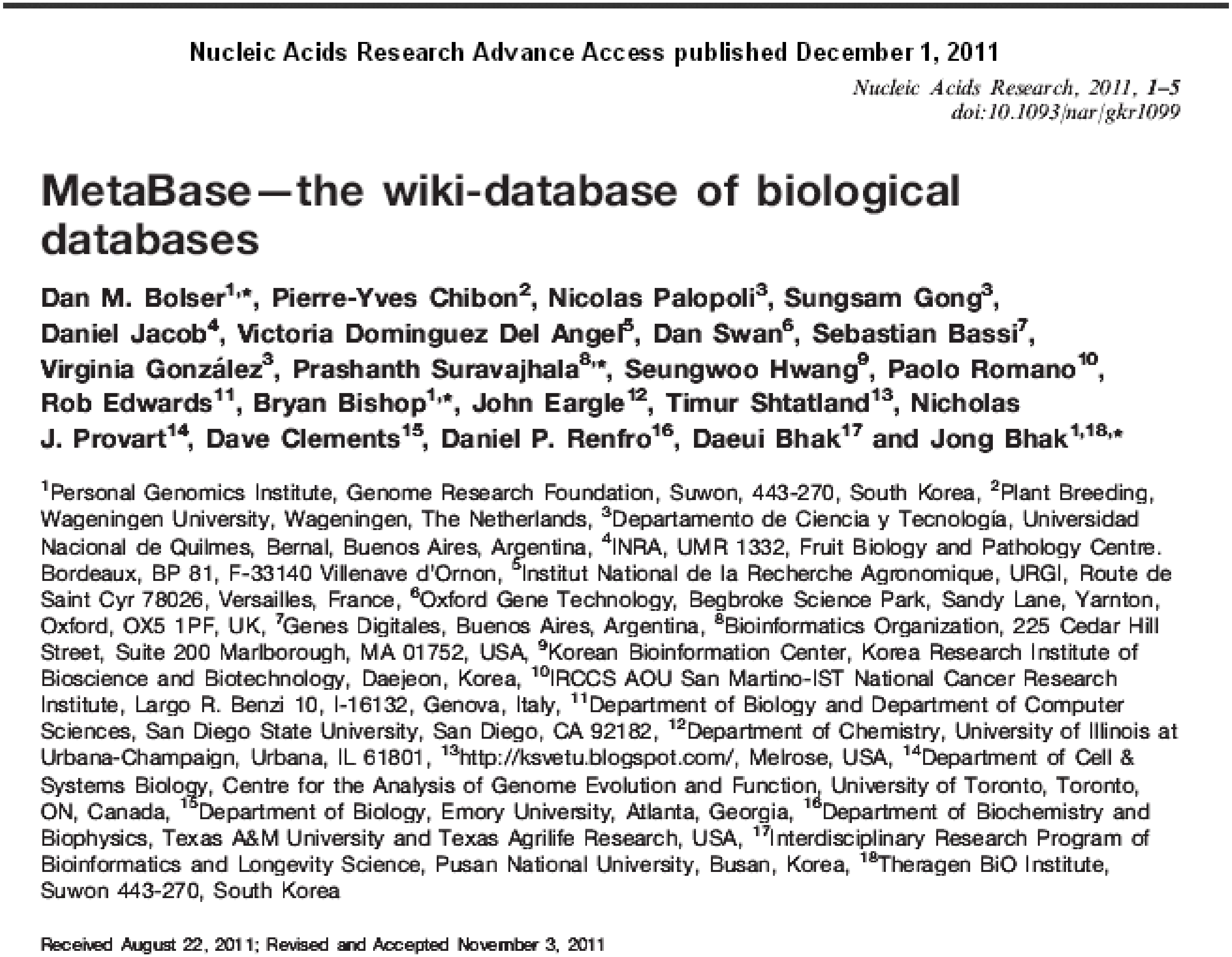
Oxford University Press is not responsible for the content of external internet sites

Online ISSN 1362-4962 - Print ISSN 0305-1048

In total, the NAR online Molecular Biology Database Collection now includes 1512 databases sorted into 15 categories

The NAR issue published on July 1, 1993, was the first one formally labelled as the “Data base Issue”.
It consisted of 24 articles (24 Data bases)

In 20 years the number of data bases increased 63 fold



Nucleic Acids Research Advance Access published December 1, 2011

Nucleic Acids Research, 2011, 1–5
doi:10.1093/nar/gkr1099

MetaBase—the wiki-database of biological databases

Dan M. Bolser[1,*], Pierre-Yves Chibon[2], Nicolas Palopoli[3], Sungsam Gong[3], Daniel Jacob[4], Victoria Dominguez Del Angel[5], Dan Swan[6], Sebastian Bassi[7], Virginia González[3], Prashanth Suravajhala[8,*], Seungwoo Hwang[9], Paolo Romano[10], Rob Edwards[11], Bryan Bishop[1,*], John Eargle[12], Timur Shtatland[13], Nicholas J. Provart[14], Dave Clements[15], Daniel P. Renfro[16], Daeui Bhak[17] and Jong Bhak[1,18,*]

[1]Personal Genomics Institute, Genome Research Foundation, Suwon, 443-270, South Korea, [2]Plant Breeding, Wageningen University, Wageningen, The Netherlands, [3]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Buenos Aires, Argentina, [4]INRA, UMR 1332, Fruit Biology and Pathology Centre. Bordeaux, BP 81, F-33140 Villenave d'Ornon, [5]Institut National de la Recherche Agronomique, URGI, Route de Saint Cyr 78026, Versailles, France, [6]Oxford Gene Technology, Begbroke Science Park, Sandy Lane, Yarnton, Oxford, OX5 1PF, UK, [7]Genes Digitales, Buenos Aires, Argentina, [8]Bioinformatics Organization, 225 Cedar Hill Street, Suite 200 Marlborough, MA 01752, USA, [9]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea, [10]IRCCS AOU San Martino-IST National Cancer Research Institute, Largo R. Benzi 10, I-16132, Genova, Italy, [11]Department of Biology and Department of Computer Sciences, San Diego State University, San Diego, CA 92182, [12]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, [13]http://ksvetu.blogspot.com/, Melrose, USA, [14]Department of Cell & Systems Biology, Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON, Canada, [15]Department of Biology, Emory University, Atlanta, Georgia, [16]Department of Biochemistry and Biophysics, Texas A&M University and Texas Agrilife Research, USA, [17]Interdisciplinary Research Program of Bioinformatics and Longevity Science, Pusan National University, Busan, Korea, [18]Theragen BiO Institute, Suwon 443-270, South Korea

Received August 22, 2011; Revised and Accepted November 3, 2011

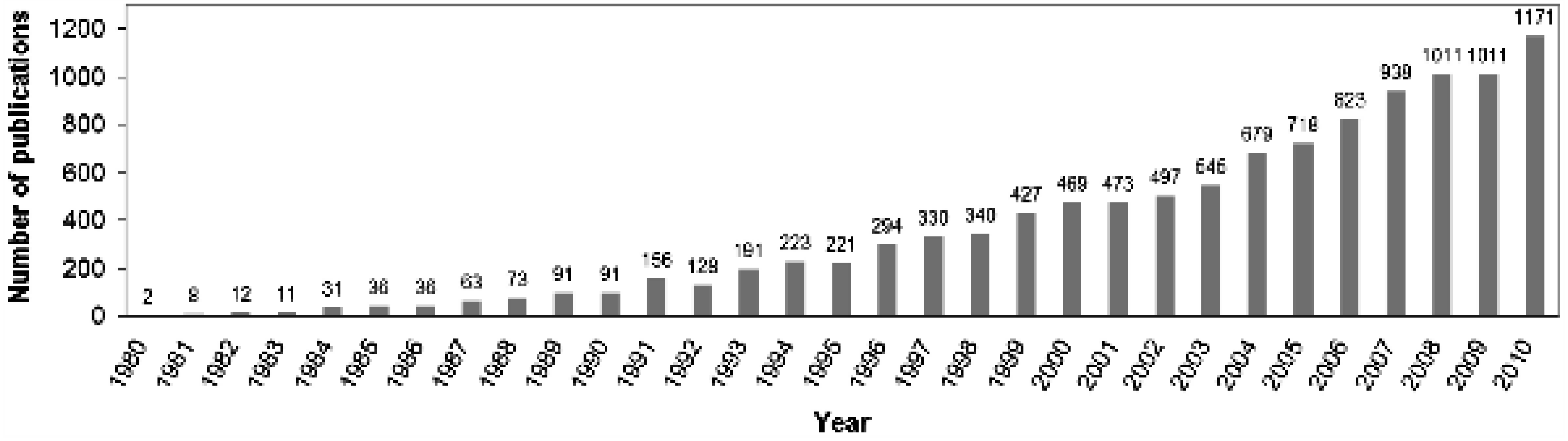
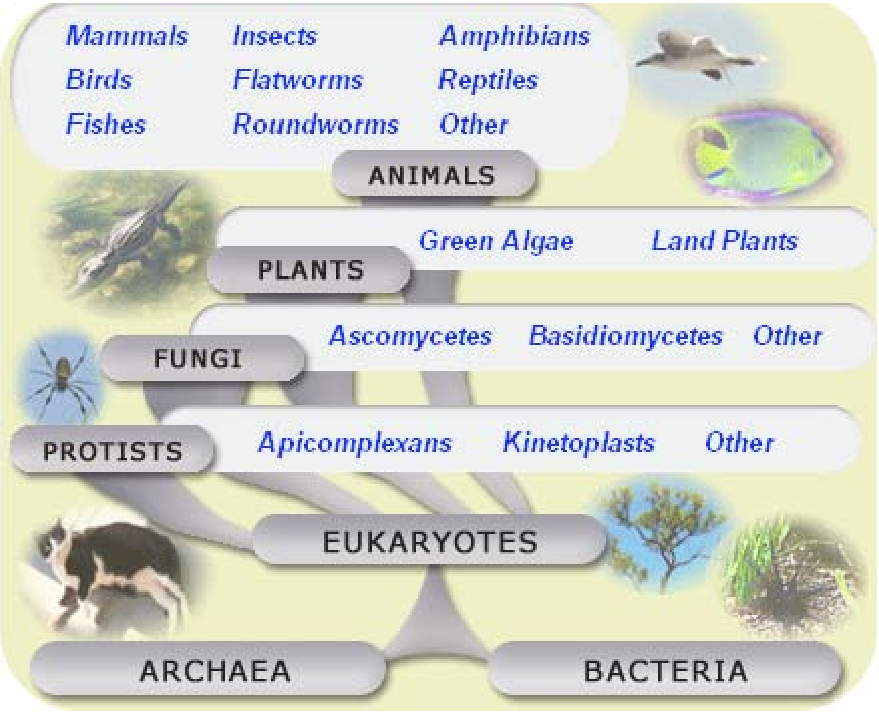


Figure 1. The growth in the number of database publications per year. Each bar shows the number of research articles with the keyword 'database' appearing in the article title in the given year. The count only covers articles indexed in PubMed. The increase shows an exponential trend that will produce nearly 2000 database publications per year by 2015.

Bolser et al., Nucleic Acids Research, 2011, 1–5

http://en.wikipedia.org/wiki/List_of_biological_databases#Genome_databases

The "omic" era-RESULTS



Complete Genomes

Prokaryotes:2474

Eukaryotes: 194

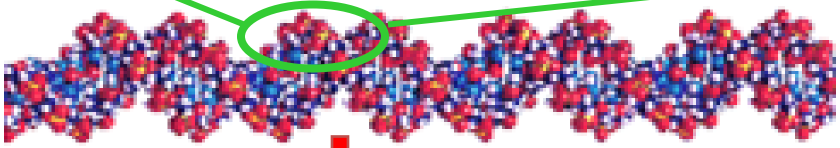
Viruses:3518

http://www.ncbi.nlm.nih.gov/ Update: May 2013

The basic information flow: from DNA to proteins

A,T,C,G

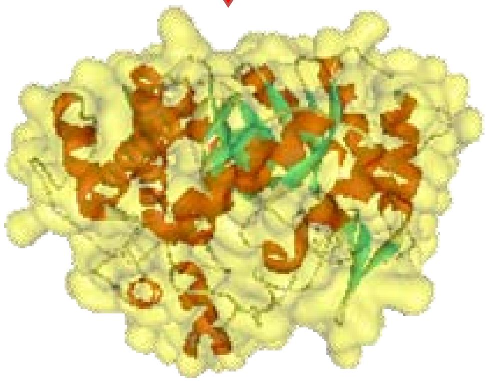
cctgttgatggcgacagggactgtatgctgatctatgctgatgcatgcatgctgactactgatgtgggggctat



From genes…

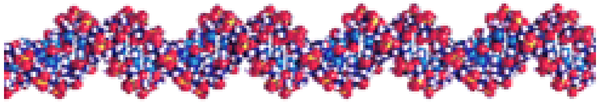
```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
```

A,C,D,E,F,G,H,I,K,LM,N,
P,Q,R,S,T,V,Y,W



…to Proteins

The Data Bases of Biological Sequences and Structures

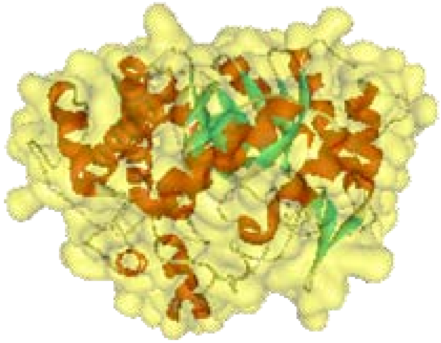


GenBank: 164,136,731 sequences
151,178,979,155 nucleotides

```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

UniProt/Tremble:
33,995,348 sequences
10,924,561,758 residues

UniProt/SwissProt:
540,052 sequences
191,770,152 residues



PDB: 87,651 structures
membrane proteins <2%

≅50 HGE!

Update:
May 2013

The Data Bases of Biological Sequences and Structures

```
>ENA|M34696|M34696.1 S.solfataricus beta-D-galactosidase (lacS)
gene, complete cds. : Location:1..1000
AAGGAGAAACTTGGCAGTTTATAACTTGACAGTAGGTTGTGGAGTGATGACTGGATCAAT
ACTAGGAGGAGTAGCATATAATTACGTTACACAATTTTATAACCCAATATATTCAATAGA
CCTTATGCTTATCCTATCCTCTATTCTAAGATTCTCGGTATCTCCCCTATTCTTGACCAT
AAAAGATACTCGCTCAAAGCTTAAATAATATTAATCATAAATAAAGTCATGTACTCATTT
CCAAATAGCTTTAGGTTTGGTTGGTCCCAGGCCGGATTTCAATCAGAAATGGGAACACCA
GGGTCAGAAGATCCAAATACTGACTGGTATAAATGGGTTCATGATCCAGAAAACATGGCA
GCGGGATTAGTAAGTGGAGATCTACCAGAAAATGGGCCAGGCTACTGGGGAAACTATAAG
ACATTTCACGATAATGCACAAAAAATGGGATTAAAAATAGCTAGACTAAATGTGGAATGG
TCTAGGATATTTCCTAATCCATTACCAAGGCCACAAAACTTTGATGAATCAAAACAAGAT
GTGACAGAGGTTGAGATAAACGAAAACGAGTTAAAGAGACTTGACGAGTACGCTAATAAA
GACGCATTAAACCATTACAGGGAAATATTCAAGGATCTTAAAAGTAGAGGACTTTACTTT
ATACTAAACATGTATCATTGGCCATTACCTCTATGGTTACACGACCCAATAAGAGTAAGA
AGAGGAGATTTTACTGGACCAAGTGGTTGGCTAAGTACTAGAACAGTTTACGAATTCGCT
AGATTCTCAGCTTATATAGCTTGGAAATTCGATGATCTAGTGGATGAGTACTCAACAATG
AATGAACCTAACGTTGTTGGAGGTTTAGGATACGTTGGTGTTAAGTCCGGTTTTCCCCCA
GGATACCTAAGCTTTGAACTTTCCCGTAGGCATATGTATAACATCATTCAAGCTCACGCA
AGAGCGTATGATGGGATAAAGAGTGTTTCTAAAAAAACCAG
```

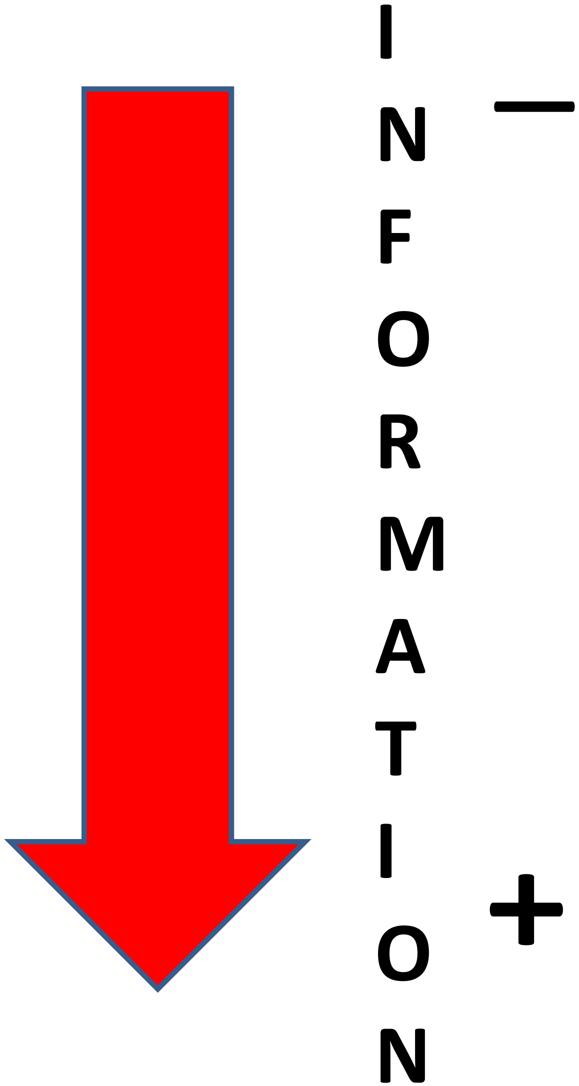
GenBank

```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

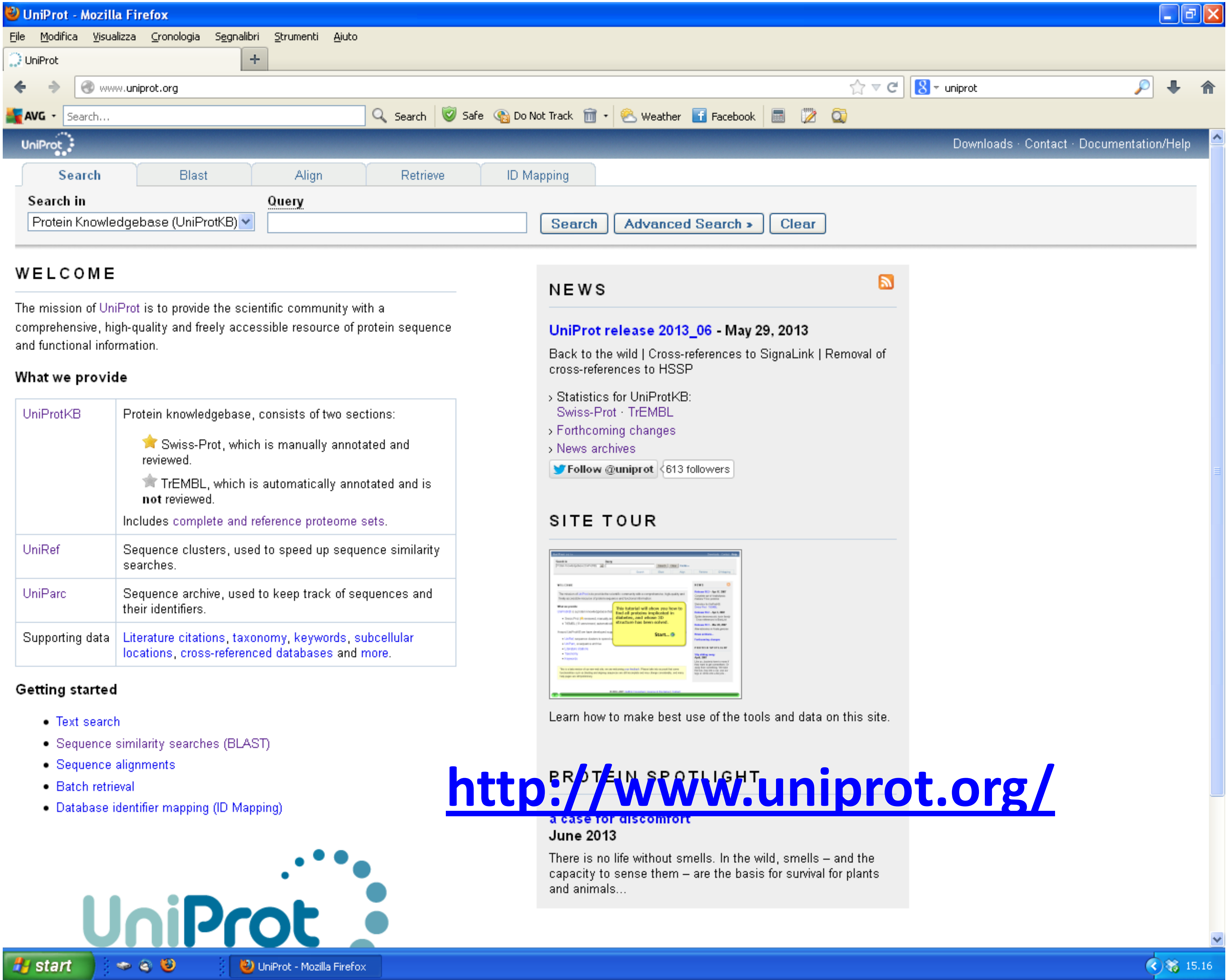
UniProt/SwissProt

PDB

1GOV



Update:
May 2013


UniProt - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
UniProt
www.uniprot.org
uniprot
AVG Search... Search Safe Do Not Track Weather Facebook
UniProt
Downloads · Contact · Documentation/Help
Search Blast Align Retrieve ID Mapping
Search in
Protein Knowledgebase (UniProtKB)
Query
Search
Advanced Search »
Clear
start
UniProt - Mozilla Firefox
15.16

WELCOME

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

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

NEWS

UniProt release 2013_06 - May 29, 2013

Back to the wild | Cross-references to SignaLink | Removal of cross-references to HSSP

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL

› Forthcoming changes

› News archives

Follow @uniprot 613 followers

SITE TOUR

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

PROTEIN SPOTLIGHT

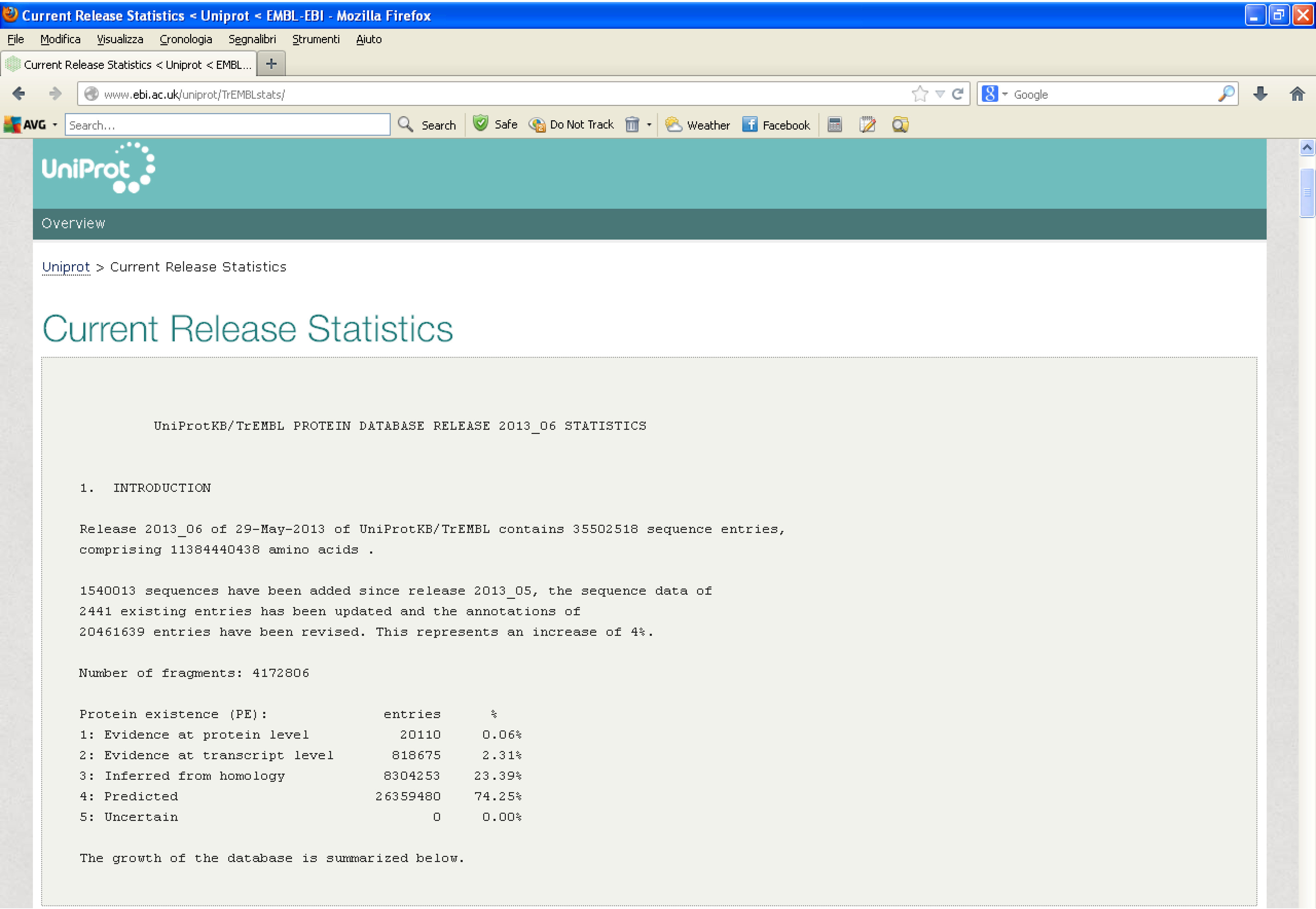
a case for discomfort

June 2013

There is no life without smells. In the wild, smells – and the capacity to sense them – are the basis for survival for plants and animals...

http://www.uniprot.org/

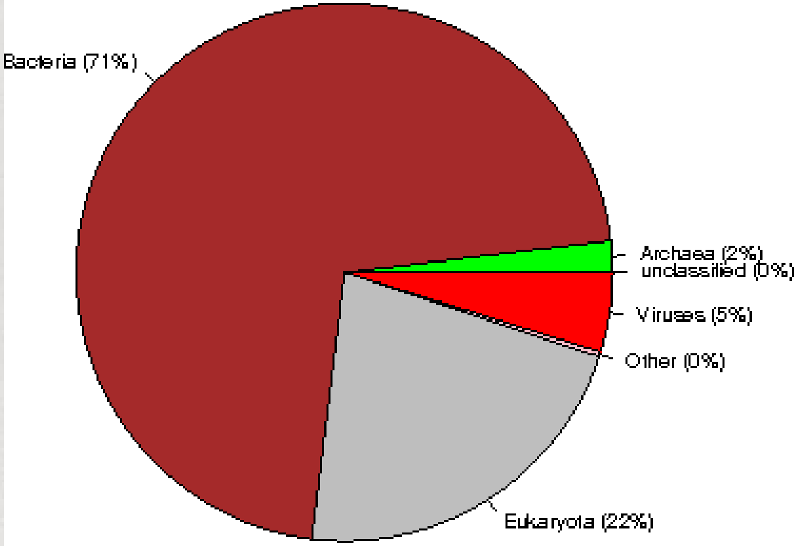
UniProt


Current Release Statistics < Uniprot < EMBL-EBI - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
Current Release Statistics < Uniprot < EMBL...
www.ebi.ac.uk/uniprot/TrEMBLstats/
Google
AVG Search... Search Safe Do Not Track Weather Facebook
UniProt
Overview

Uniprot > Current Release Statistics

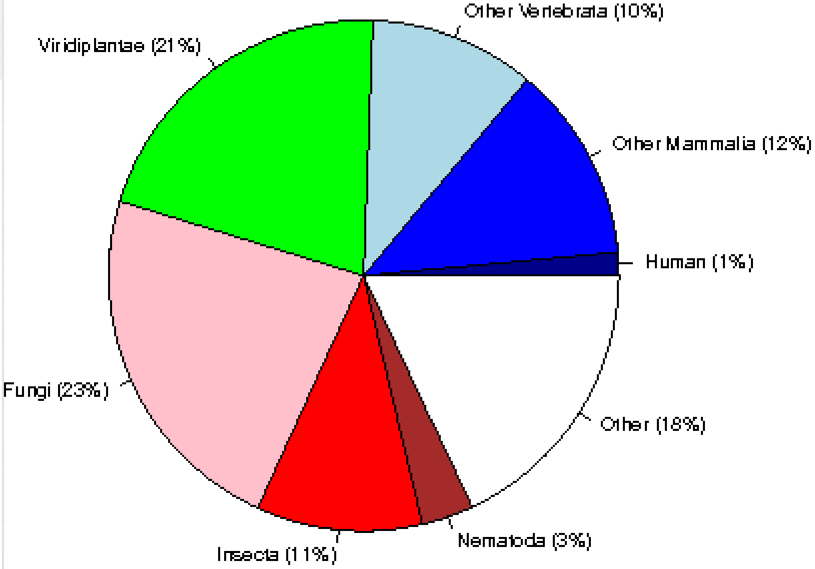
Current Release Statistics

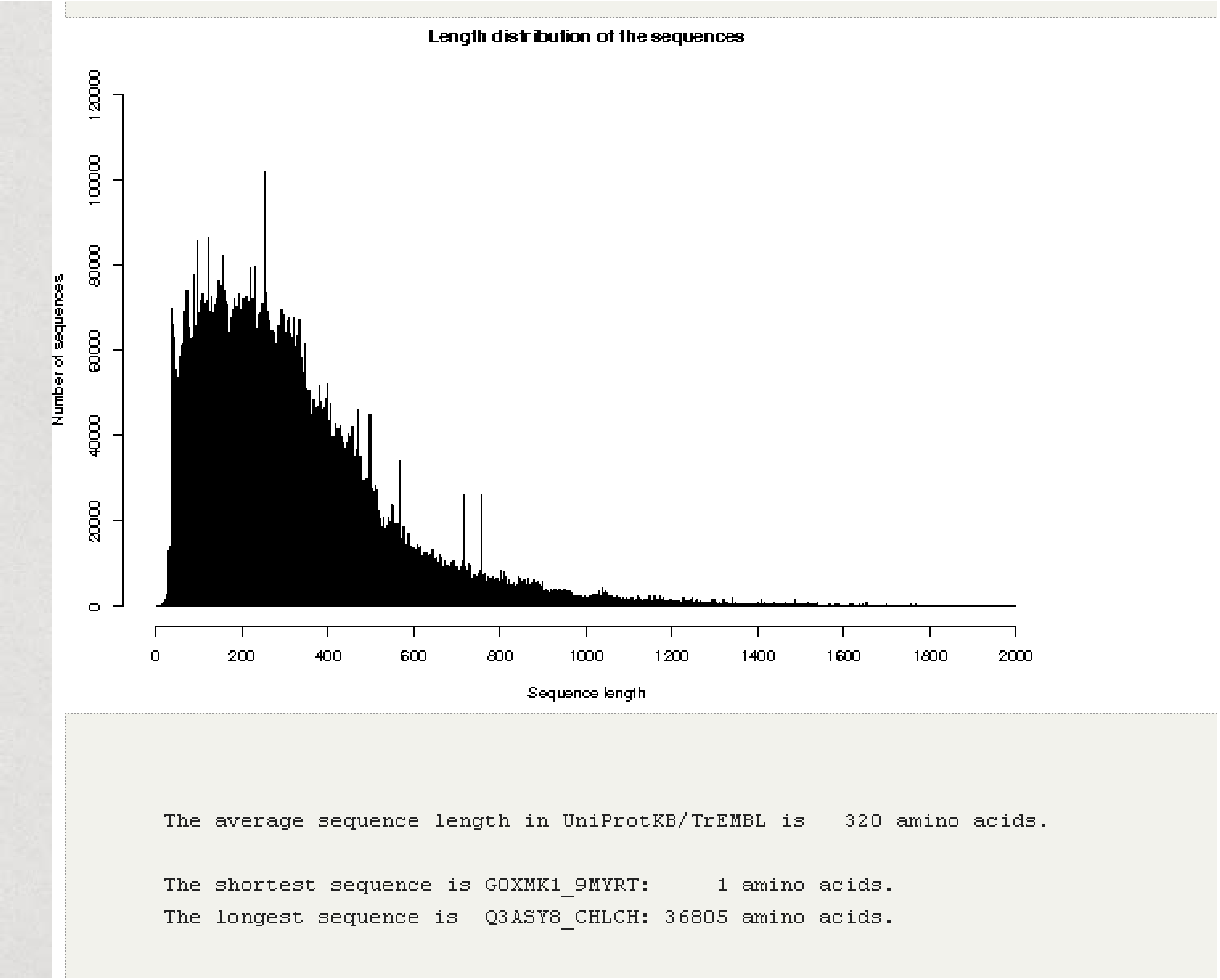
```
          UniProtKB/TrEMBL PROTEIN DATABASE RELEASE 2013_06 STATISTICS


 1.   INTRODUCTION

 Release 2013_06 of 29-May-2013 of UniProtKB/TrEMBL contains 35502518 sequence entries,
 comprising 11384440438 amino acids .

 1540013 sequences have been added since release 2013_05, the sequence data of
 2441 existing entries has been updated and the annotations of
 20461639 entries have been revised. This represents an increase of 4%.

 Number of fragments: 4172806

 Protein existence (PE):               entries      %
 1: Evidence at protein level            20110     0.06%
 2: Evidence at transcript level        818675     2.31%
 3: Inferred from homology             8304253    23.39%
 4: Predicted                         26359480    74.25%
 5: Uncertain                                0     0.00%

 The growth of the database is summarized below.
```



```
Kingdom          sequences (% of the database)
 Archaea            672982 (  2%)
 Bacteria         25261484 ( 71%)
 Eukaryota         7764851 ( 22%)
 Viruses           1699620 (  5%)
 Other              103580 ( <1%)
```

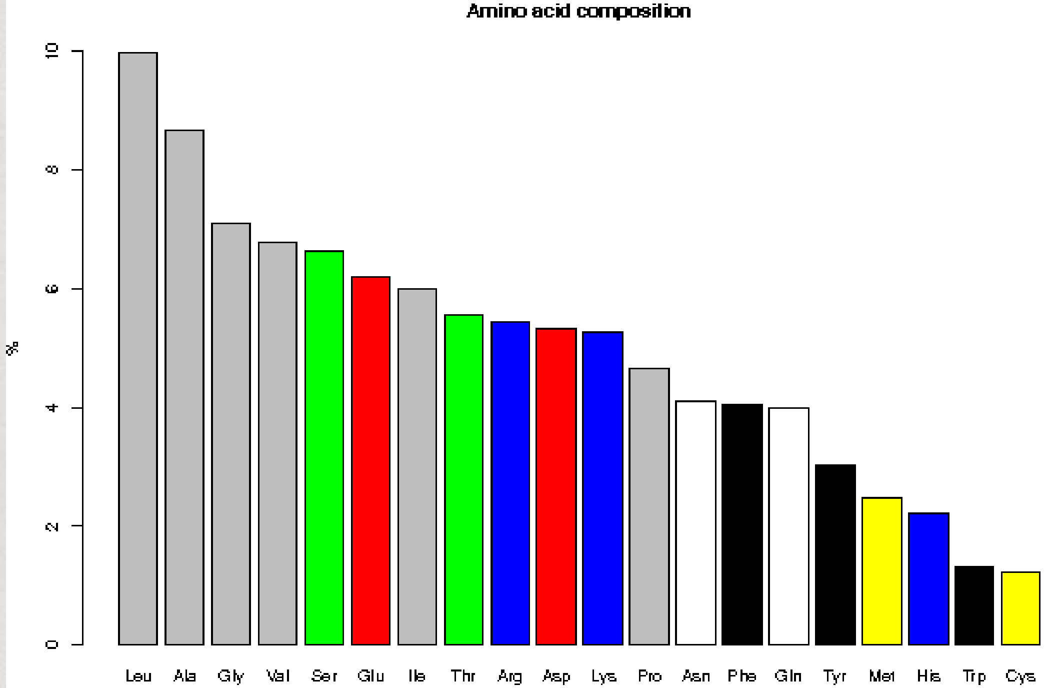
Within Eukaryota:



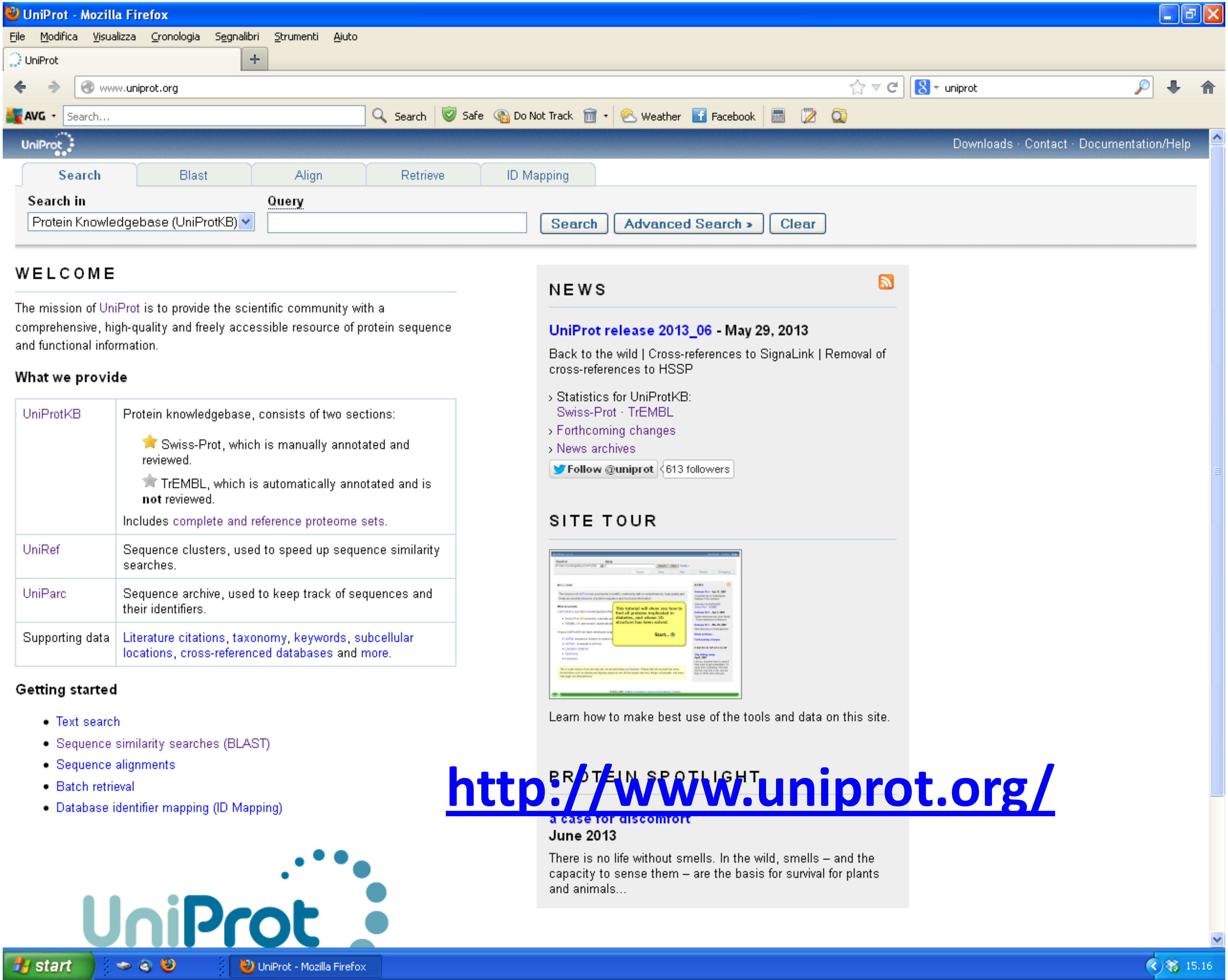


```
The average sequence length in UniProtKB/TrEMBL is   320 amino acids.

The shortest sequence is G0XMK1_9MYRT:     1 amino acids.
The longest sequence is  Q3ASY8_CHLCH: 36805 amino acids.
```

```
Ala (A) 8.66   Gln (Q) 3.98   Leu (L) 9.96   Ser (S) 6.63
Arg (R) 5.43   Glu (E) 6.19   Lys (K) 5.26   Thr (T) 5.55
Asn (N) 4.09   Gly (G) 7.09   Met (M) 2.47   Trp (W) 1.30
Asp (D) 5.33   His (H) 2.20   Phe (F) 4.03   Tyr (Y) 3.03
Cys (C) 1.23   Ile (I) 6.00   Pro (P) 4.65   Val (V) 6.79

Asx (B) 0.000  Glx (Z) 0      Xaa (X) 0.03
```



```
Legend: gray = aliphatic, red = acidic, green = small hydroxy,
        blue = basic, black = aromatic, white = amide, yellow = sulfur


5.2  Classification of the amino acids by their frequency

Leu, Ala, Gly, Val, Ser, Glu, Ile, Thr, Arg, Asp, Lys, Pro, Asn, Phe,
Gln, Tyr, Met, His, Trp, Cys
```


UniProt - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
UniProt
www.uniprot.org
uniprot
AVG Search... Search Safe Do Not Track Weather Facebook
UniProt
Downloads · Contact · Documentation/Help
Search Blast Align Retrieve ID Mapping
Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Advanced Search »
Clear
start
UniProt - Mozilla Firefox
15.16

WELCOME

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

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

NEWS

UniProt release 2013_06 - May 29, 2013

Back to the wild | Cross-references to SignaLink | Removal of cross-references to HSSP

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

Follow @uniprot 613 followers

SITE TOUR

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

PROTEIN SPOTLIGHT

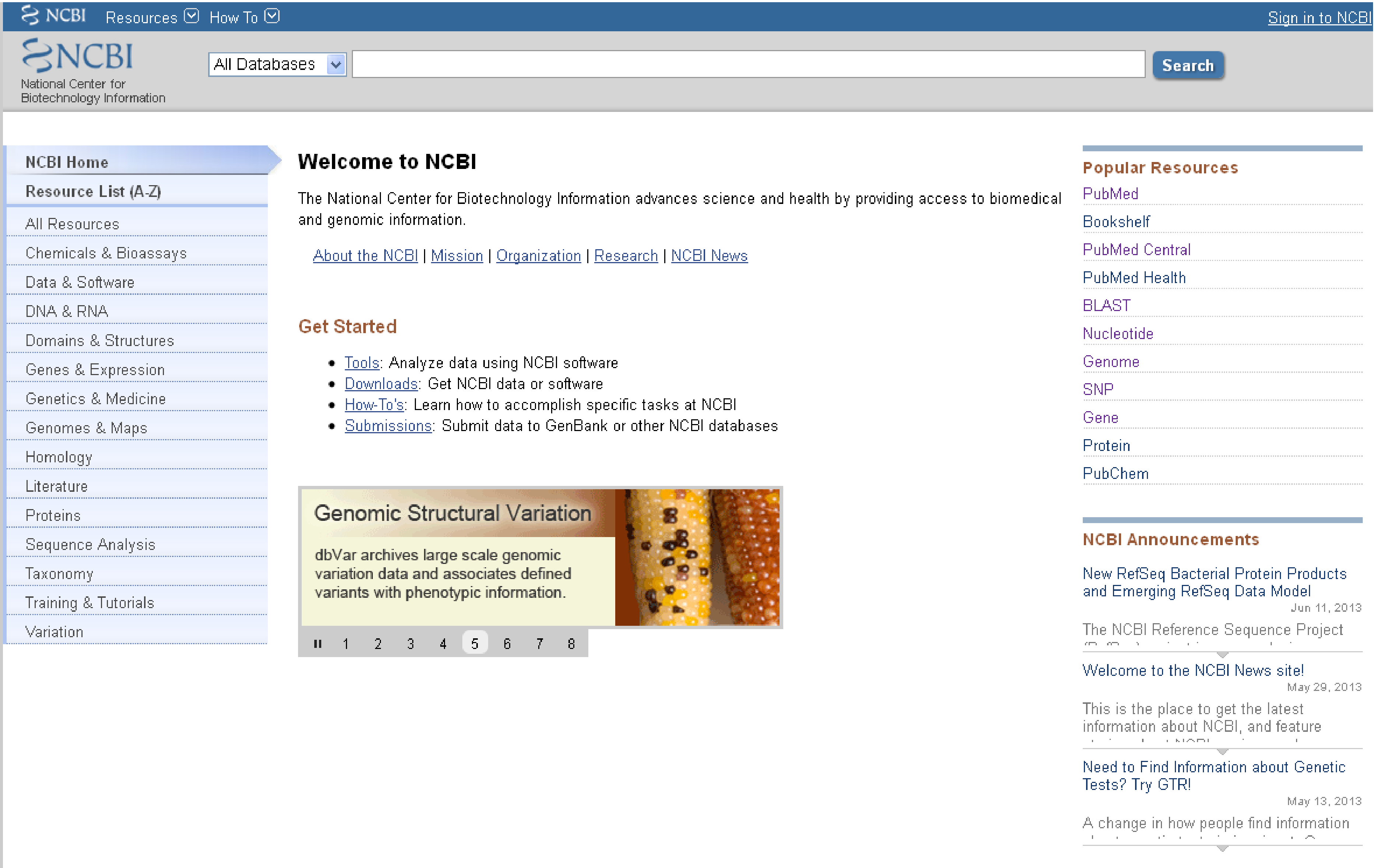
a case for discomfort

June 2013

There is no life without smells. In the wild, smells – and the capacity to sense them – are the basis for survival for plants and animals...

UniProt

http://www.uniprot.org/



All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.

Popular Resources

PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements

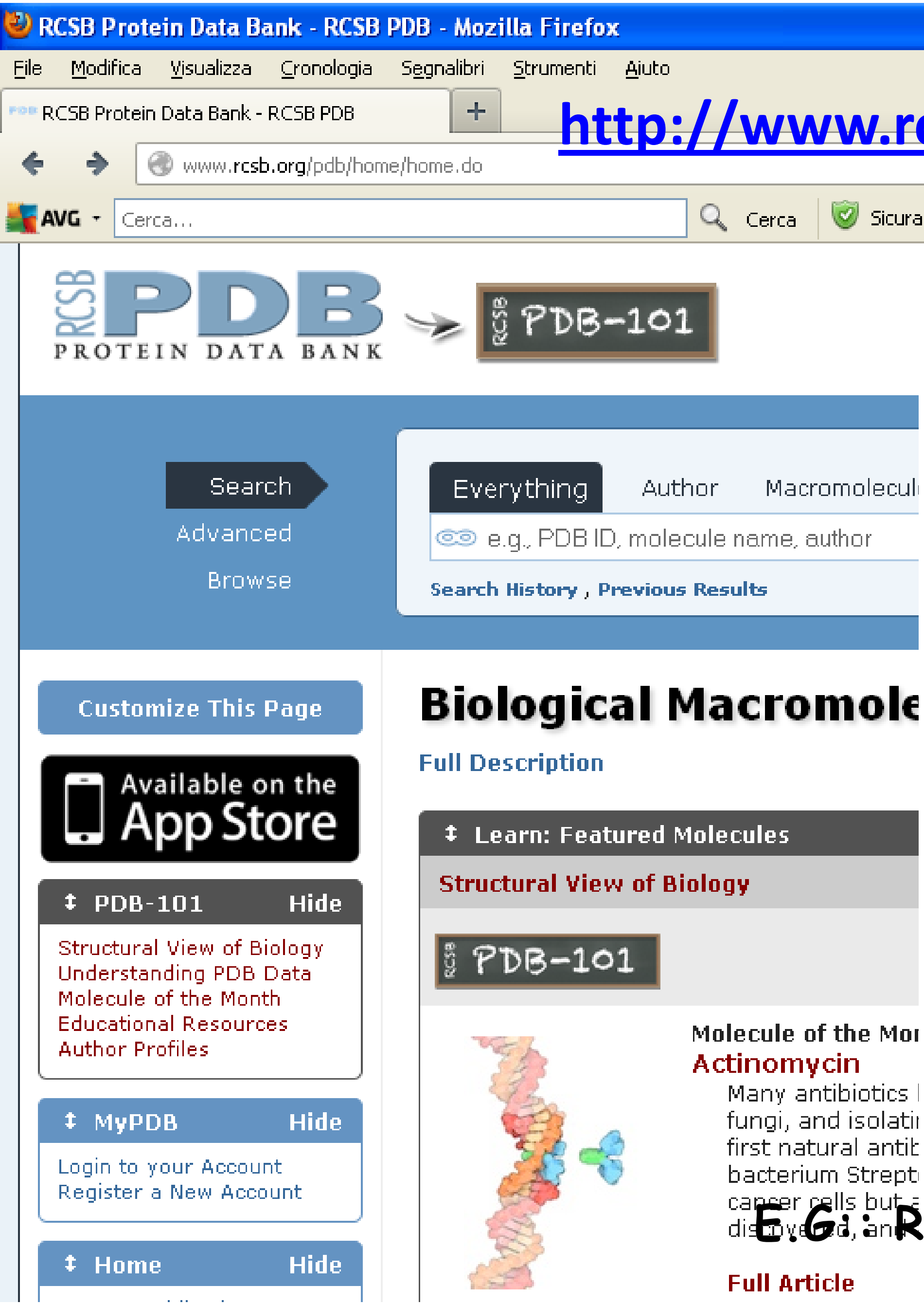
New RefSeq Bacterial Protein Products and Emerging RefSeq Data Model
Jun 11, 2013
The NCBI Reference Sequence Project

Welcome to the NCBI News site!
May 29, 2013
This is the place to get the latest information about NCBI, and feature

Need to Find Information about Genetic Tests? Try GTR!
May 13, 2013
A change in how people find information

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

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



RCSB Protein Data Bank - RCSB PDB - Mozilla Firefox

RCSB PDB PROTEIN DATA BANK → PDB-101

A MEMBER OF

An Information Portal to Biological

As of Tuesday Apr 30, 2013 at 5 PM PDT there are 90206 Structures

Search
Advanced
Browse
Everything
Author
Macromolecul
e.g., PDB ID, molecule name, author
Search History , Previous Results
Customize This Page
Available on the App Store

PDB-101 Hide

Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles

MyPDB Hide

Login to your Account
Register a New Account

Home Hide

Biological Macromole

Full Description

Learn: Featured Molecules

Structural View of Biology

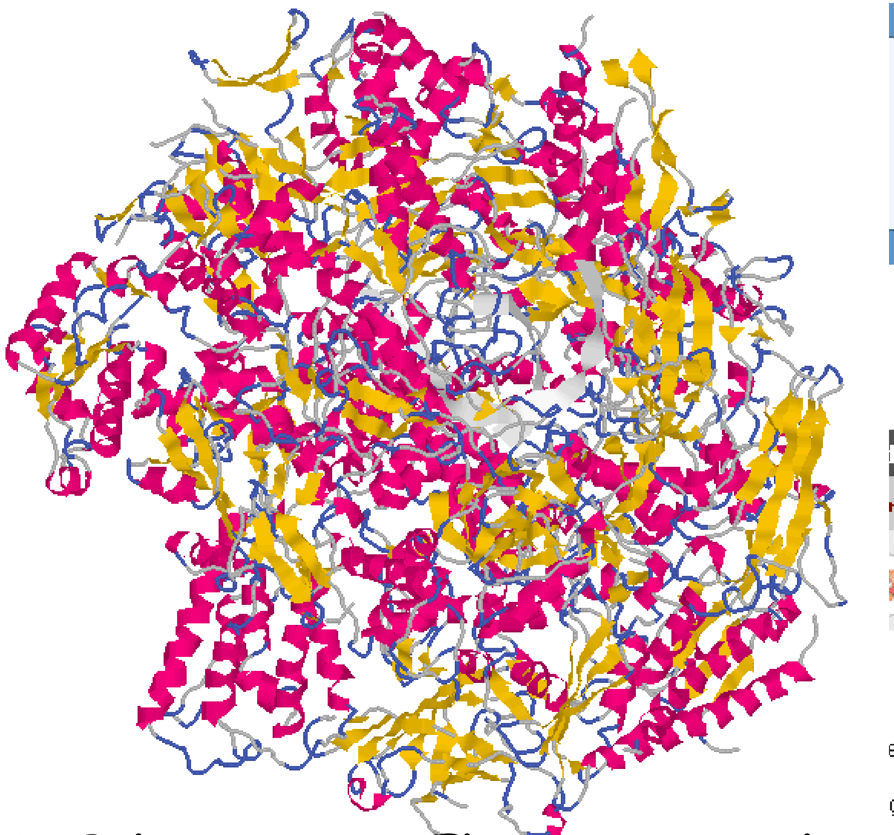
PDB-101

Molecule of the Mor

Actinomycin

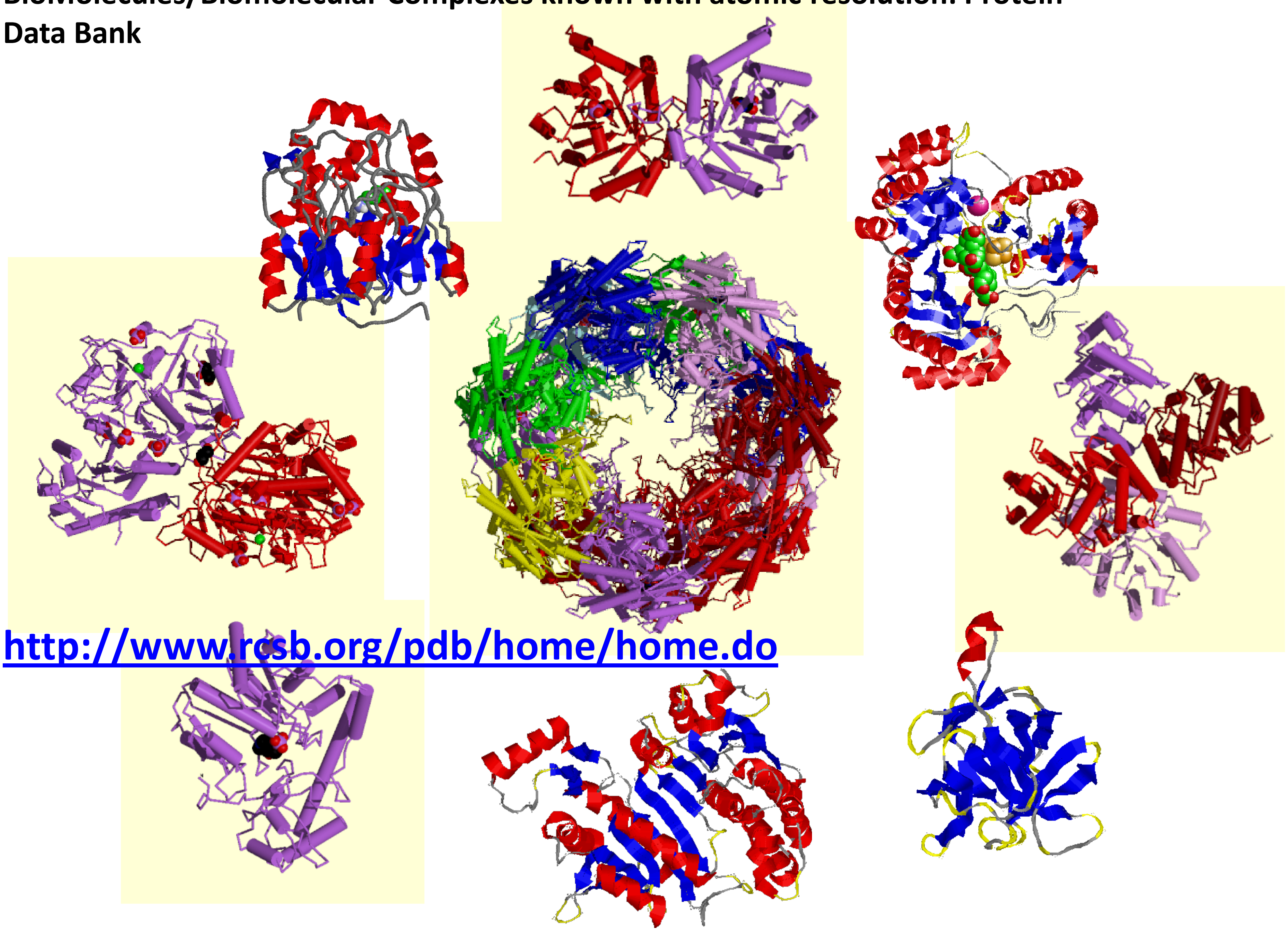
Many antibiotics
fungi, and isolati
first natural antib
bacterium Strept
cancer cells but
discovered, an

Full Article

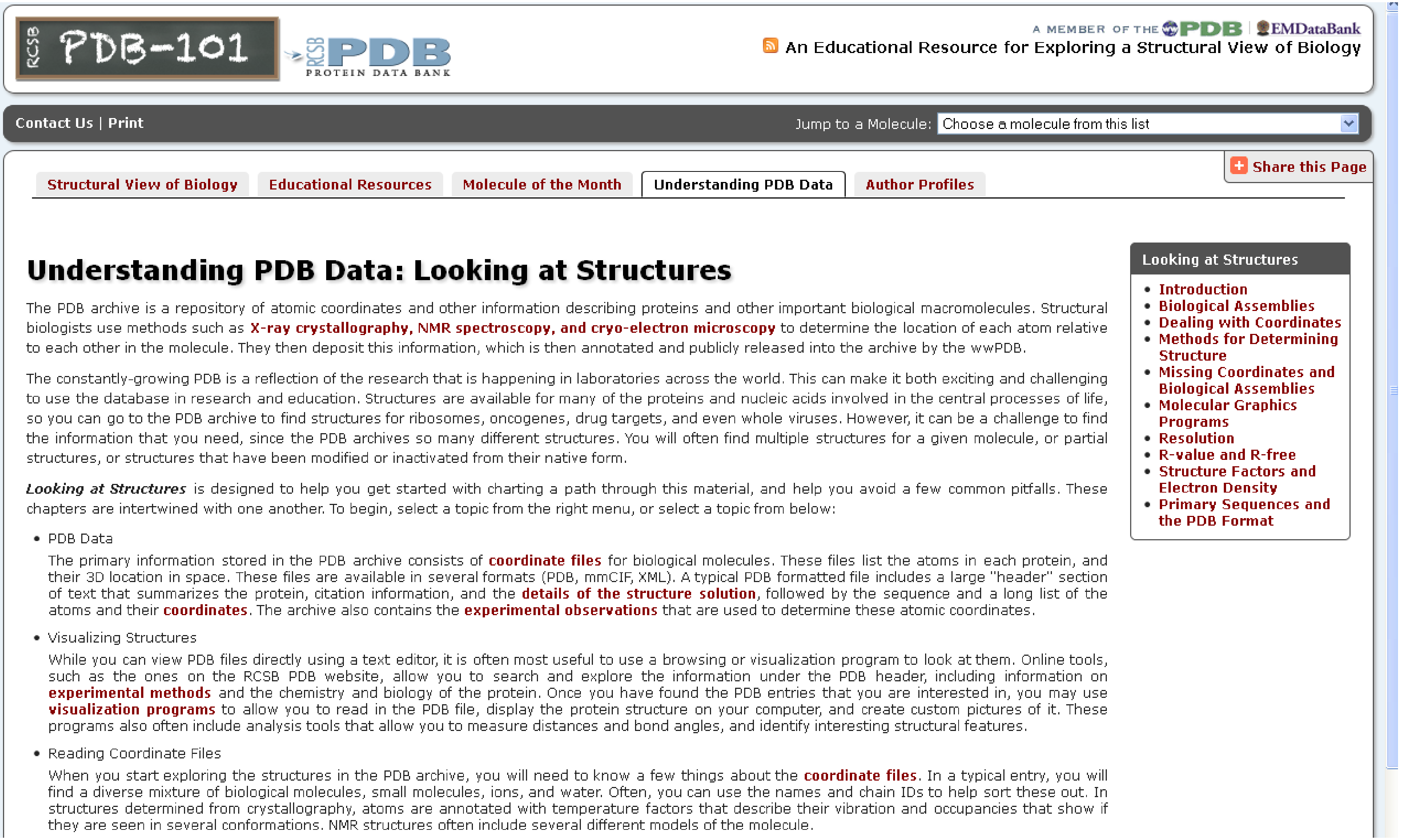


E.G.: RNA Polymerase II Elongation Complex

BioMolecules/Biomolecular Complexes known with atomic resolution: Protein Data Bank



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


PDB-101
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE wwPDB | EMDataBank
An Educational Resource for Exploring a Structural View of Biology
Contact Us | Print
Jump to a Molecule: Choose a molecule from this list
Share this Page
Structural View of Biology
Educational Resources
Molecule of the Month
Understanding PDB Data
Author Profiles

Understanding PDB Data: Looking at Structures

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as **X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy** to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

The constantly-growing PDB is a reflection of the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education. Structures are available for many of the proteins and nucleic acids involved in the central processes of life, so you can go to the PDB archive to find structures for ribosomes, oncogenes, drug targets, and even whole viruses. However, it can be a challenge to find the information that you need, since the PDB archives so many different structures. You will often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form.

Looking at Structures is designed to help you get started with charting a path through this material, and help you avoid a few common pitfalls. These chapters are intertwined with one another. To begin, select a topic from the right menu, or select a topic from below:

- PDB Data

  The primary information stored in the PDB archive consists of **coordinate files** for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). A typical PDB formatted file includes a large "header" section of text that summarizes the protein, citation information, and the **details of the structure solution**, followed by the sequence and a long list of the atoms and their **coordinates**. The archive also contains the **experimental observations** that are used to determine these atomic coordinates.

- Visualizing Structures

  While you can view PDB files directly using a text editor, it is often most useful to use a browsing or visualization program to look at them. Online tools, such as the ones on the RCSB PDB website, allow you to search and explore the information under the PDB header, including information on **experimental methods** and the chemistry and biology of the protein. Once you have found the PDB entries that you are interested in, you may use **visualization programs** to allow you to read in the PDB file, display the protein structure on your computer, and create custom pictures of it. These programs also often include analysis tools that allow you to measure distances and bond angles, and identify interesting structural features.

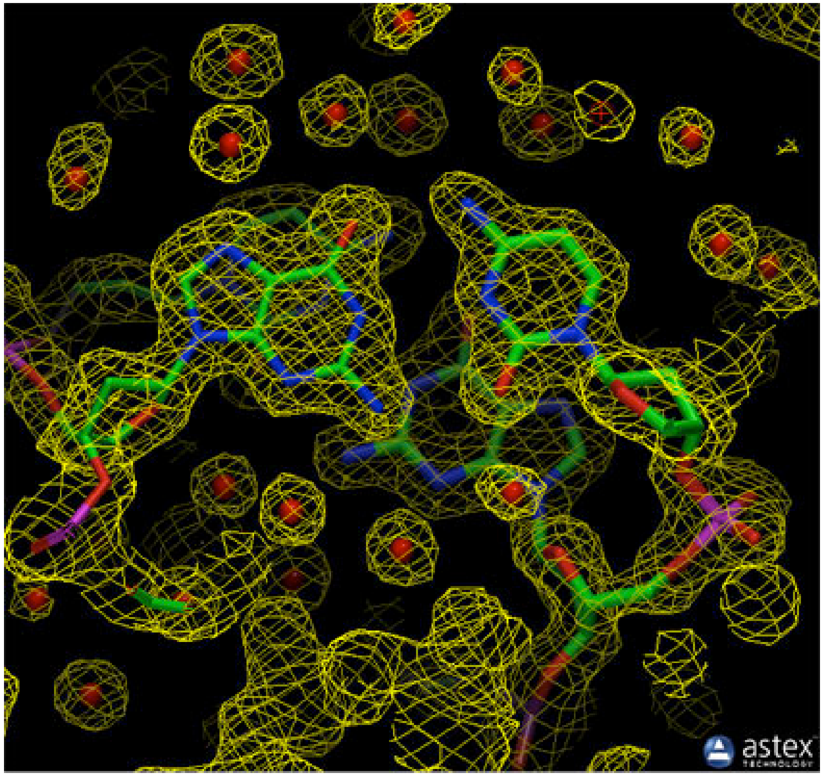
- Reading Coordinate Files

  When you start exploring the structures in the PDB archive, you will need to know a few things about the **coordinate files**. In a typical entry, you will find a diverse mixture of biological molecules, small molecules, ions, and water. Often, you can use the names and chain IDs to help sort these out. In structures determined from crystallography, atoms are annotated with temperature factors that describe their vibration and occupancies that show if they are seen in several conformations. NMR structures often include several different models of the molecule.

Looking at Structures

http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/intro.html

Biological molecule crystals are finicky: some form perfect, well-ordered crystals and others form only poor crystals. The accuracy of the atomic structure that is determined depends on the quality of these crystals. In perfect crystals, we have far more confidence that the atomic structure correctly reflects the structure of the protein. Two important measures of the accuracy of a crystallographic structure are its **resolution,** which measures the amount of detail that may be seen in the experimental data, and the **R-value,** which measures how well the atomic model is supported by the experimental data found in the structure factor file.



The experimental electron density from a structure of DNA is shown here (PDB entry ***196d),*** *along with the atomic model that was generated based on the data. The contours surround regions with high densities of electrons, which correspond to the atoms in the molecule. This picture was created with the Astex viewer, which can be accessed by clicking the "EDS" link on the Structure Summary page for this entry.*

The Data Bases of Biological Sequences and Structures

```
>ENA|M34696|M34696.1 S.solfataricus beta-D-galactosidase (lacS)
gene, complete cds. : Location:1..1000
AAGGAGAAACTTGGCAGTTTATAACTTGACAGTAGGTTGTGGAGTGATGACTGGATCAAT
ACTAGGAGGAGTAGCATATAATTACGTTACACAATTTTATAACCCAATATATTCAATAGA
CCTTATGCTTATCCTATCCTCTATTCTAAGATTCTCGGTATCTCCCCTATTCTTGACCAT
AAAAGATACTCGCTCAAAGCTTAAATAATATTAATCATAAATAAAGTCATGTACTCATTT
CCAAATAGCTTTAGGTTTGGTTGGTCCCAGGCCGGATTTCAATCAGAAATGGGAACACCA
GGGTCAGAAGATCCAAATACTGACTGGTATAAATGGGTTCATGATCCAGAAAACATGGCA
GCGGGATTAGTAAGTGGAGATCTACCAGAAAATGGGCCAGGCTACTGGGGAAACTATAAG
ACATTTCACGATAATGCACAAAAAATGGGATTAAAAATAGCTAGACTAAATGTGGAATGG
TCTAGGATATTTCCTAATCCATTACCAAGGCCACAAAACTTTGATGAATCAAAACAAGAT
GTGACAGAGGTTGAGATAAACGAAAACGAGTTAAAGAGACTTGACGAGTACGCTAATAAA
GACGCATTAAACCATTACAGGGAAATATTCAAGGATCTTAAAAGTAGAGGACTTTACTTT
ATACTAAACATGTATCATTGGCCATTACCTCTATGGTTACACGACCCAATAAGAGTAAGA
AGAGGAGATTTTACTGGACCAAGTGGTTGGCTAAGTACTAGAACAGTTTACGAATTCGCT
AGATTCTCAGCTTATATAGCTTGGAAATTCGATGATCTAGTGGATGAGTACTCAACAATG
AATGAACCTAACGTTGTTGGAGGTTTAGGATACGTTGGTGTTAAGTCCGGTTTTCCCCCA
GGATACCTAAGCTTTGAACTTTCCCGTAGGCATATGTATAACATCATTCAAGCTCACGCA
AGAGCGTATGATGGGATAAAGAGTGTTTCTAAAAAACCAG
```

GenBank

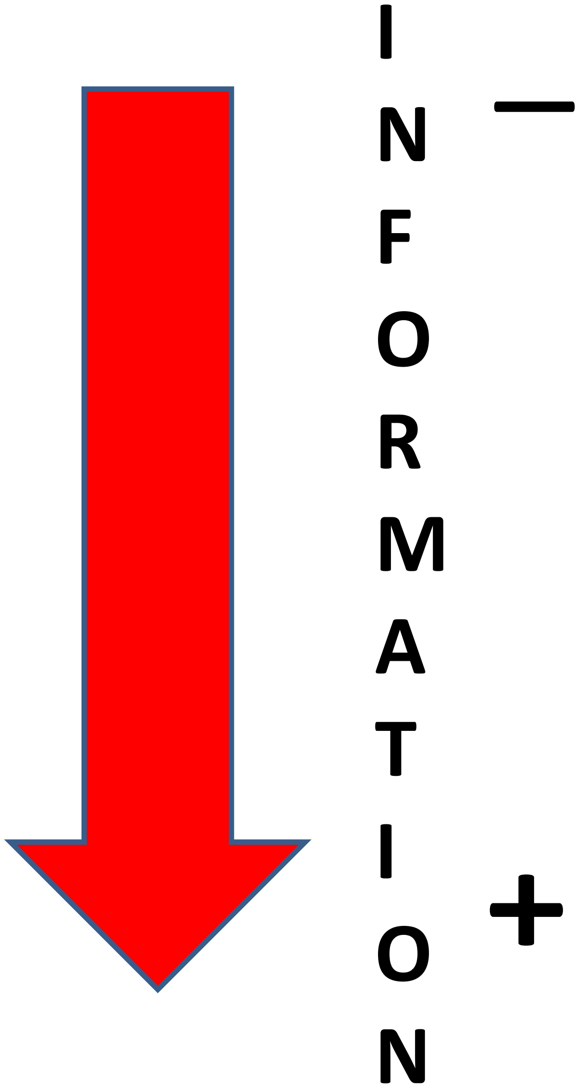
```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

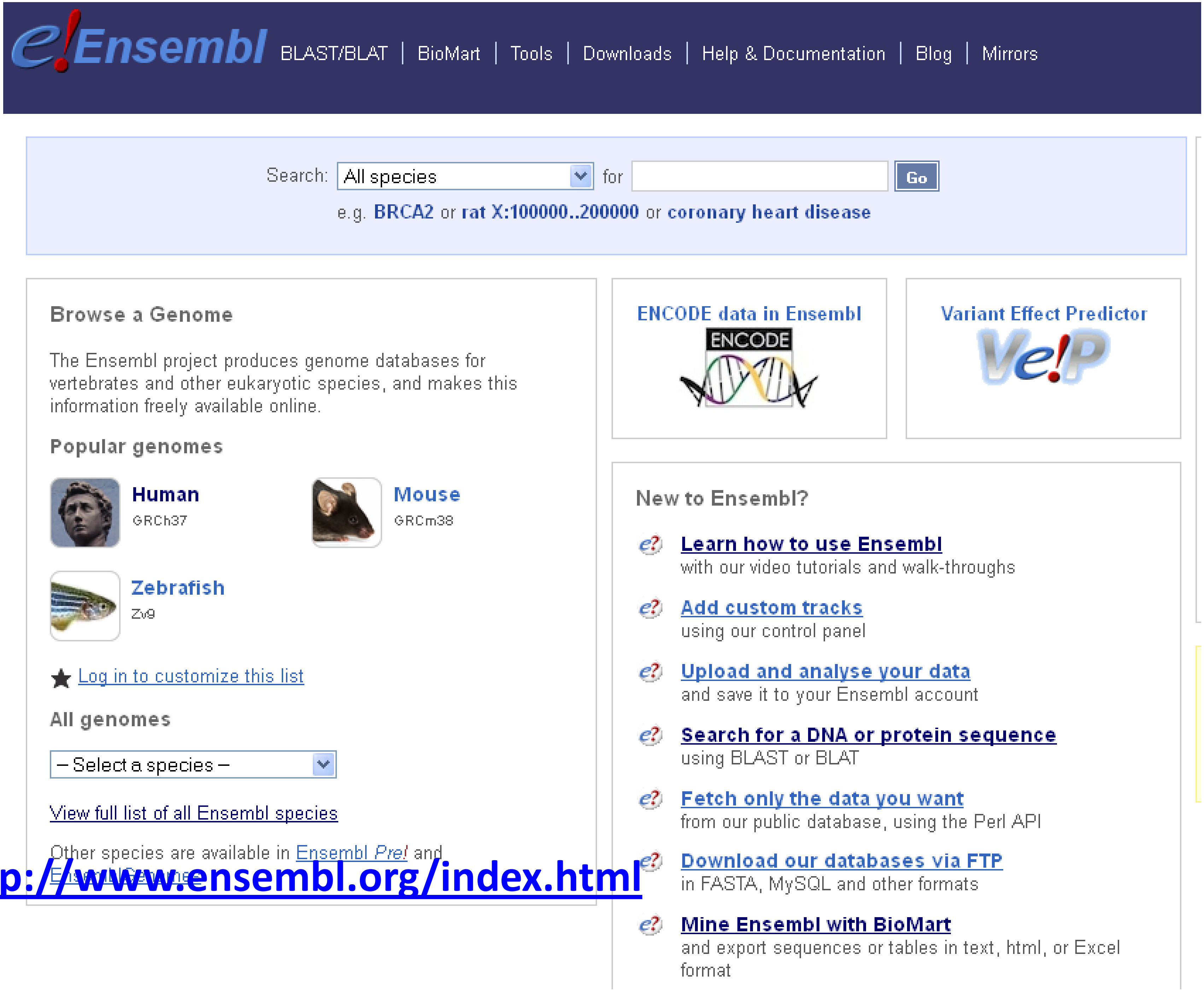
UniProt/SwissProt

PDB

1GOV

Update:
May 2013





e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search: All species for Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

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

Popular genomes

Human GRCh37	Mouse GRCm38
Zebrafish Zv9	

★ Log in to customize this list

All genomes

View full list of all Ensembl species

Other species are available in Ensembl Pre! and Ensembl Genomes

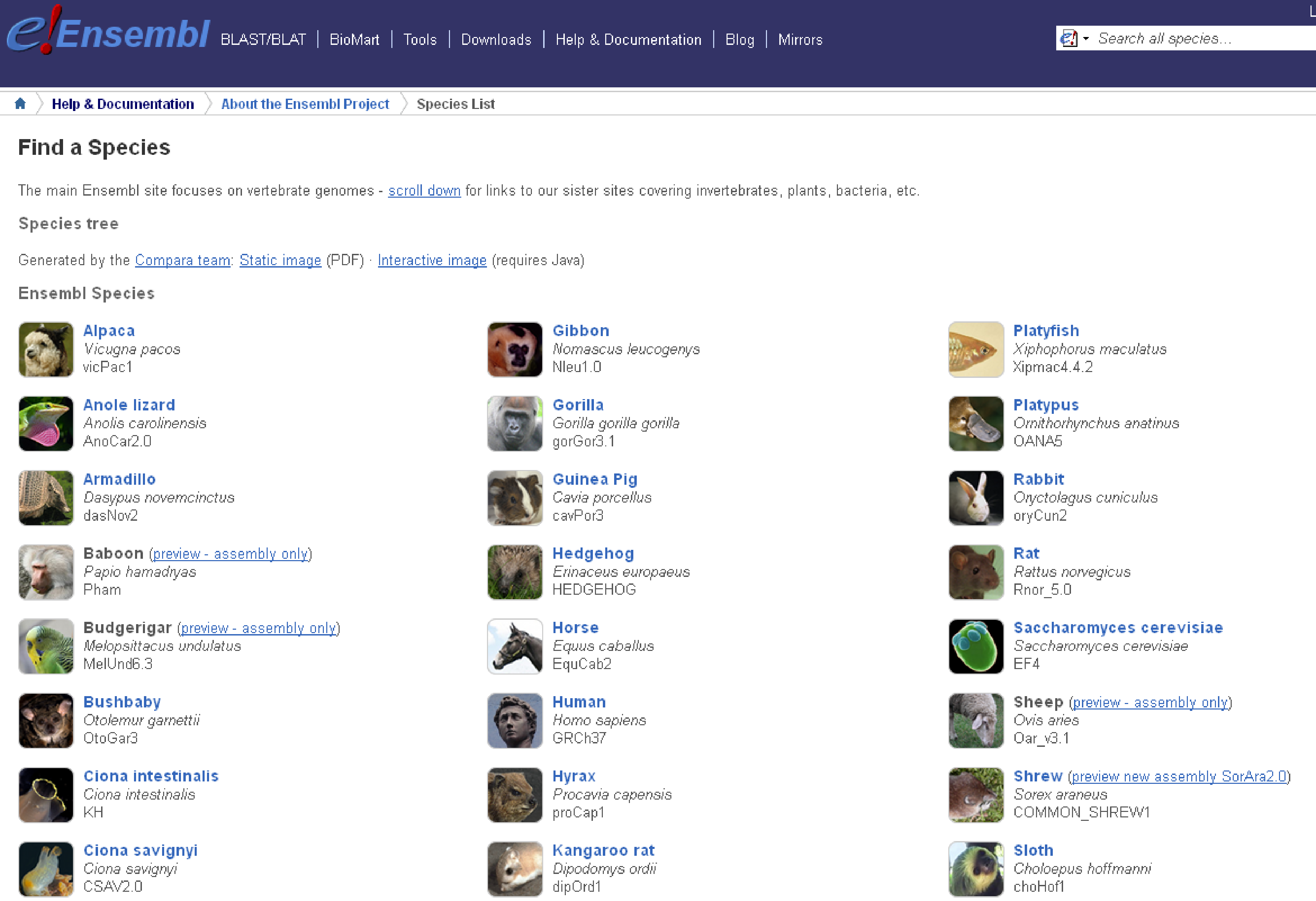
ENCODE data in Ensembl

Variant Effect Predictor

New to Ensembl?

- **Learn how to use Ensembl** with our video tutorials and walk-throughs
- **Add custom tracks** using our control panel
- **Upload and analyse your data** and save it to your Ensembl account
- **Search for a DNA or protein sequence** using BLAST or BLAT
- **Fetch only the data you want** from our public database, using the Perl API
- **Download our databases via FTP** in FASTA, MySQL and other formats
- **Mine Ensembl with BioMart** and export sequences or tables in text, html, or Excel format

http://www.ensembl.org/index.html



Find a Species

The main Ensembl site focuses on vertebrate genomes - scroll down for links to our sister sites covering invertebrates, plants, bacteria, etc.

Species tree

Generated by the Compara team: Static image (PDF) · Interactive image (requires Java)

Ensembl Species

Alpaca	*Vicugna pacos*	vicPac1
Anole lizard	*Anolis carolinensis*	AnoCar2.0
Armadillo	*Dasypus novemcinctus*	dasNov2
Baboon (preview - assembly only)	*Papio hamadryas*	Pham
Budgerigar (preview - assembly only)	*Melopsittacus undulatus*	MelUnd6.3
Bushbaby	*Otolemur garnettii*	OtoGar3
Ciona intestinalis	*Ciona intestinalis*	KH
Ciona savignyi	*Ciona savignyi*	CSAV2.0
Gibbon	*Nomascus leucogenys*	Nleu1.0
Gorilla	*Gorilla gorilla gorilla*	gorGor3.1
Guinea Pig	*Cavia porcellus*	cavPor3
Hedgehog	*Erinaceus europaeus*	HEDGEHOG
Horse	*Equus caballus*	EquCab2
Human	*Homo sapiens*	GRCh37
Hyrax	*Procavia capensis*	proCap1
Kangaroo rat	*Dipodomys ordii*	dipOrd1
Platyfish	*Xiphophorus maculatus*	Xipmac4.4.2
Platypus	*Ornithorhynchus anatinus*	OANA5
Rabbit	*Oryctolagus cuniculus*	oryCun2
Rat	*Rattus norvegicus*	Rnor_5.0
Saccharomyces cerevisiae	*Saccharomyces cerevisiae*	EF4
Sheep (preview - assembly only)	*Ovis aries*	Oar_v3.1
Shrew (preview new assembly SorAra2.0)	*Sorex araneus*	COMMON_SHREW1
Sloth	*Choloepus hoffmanni*	choHof1

http://www.ensembl.org/info/about/species.html

Summing up…..

- Biological data bases are the archives where presently all the available knowledge is stored

- Their usage is necessary for retrieving all the mocelur details of our knowledge

- Data mining requires some bioinformatic skillness