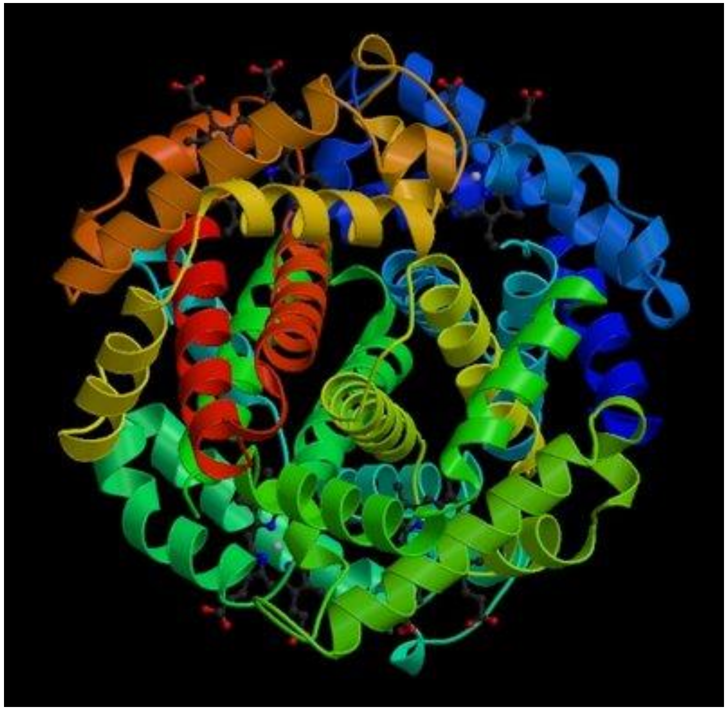
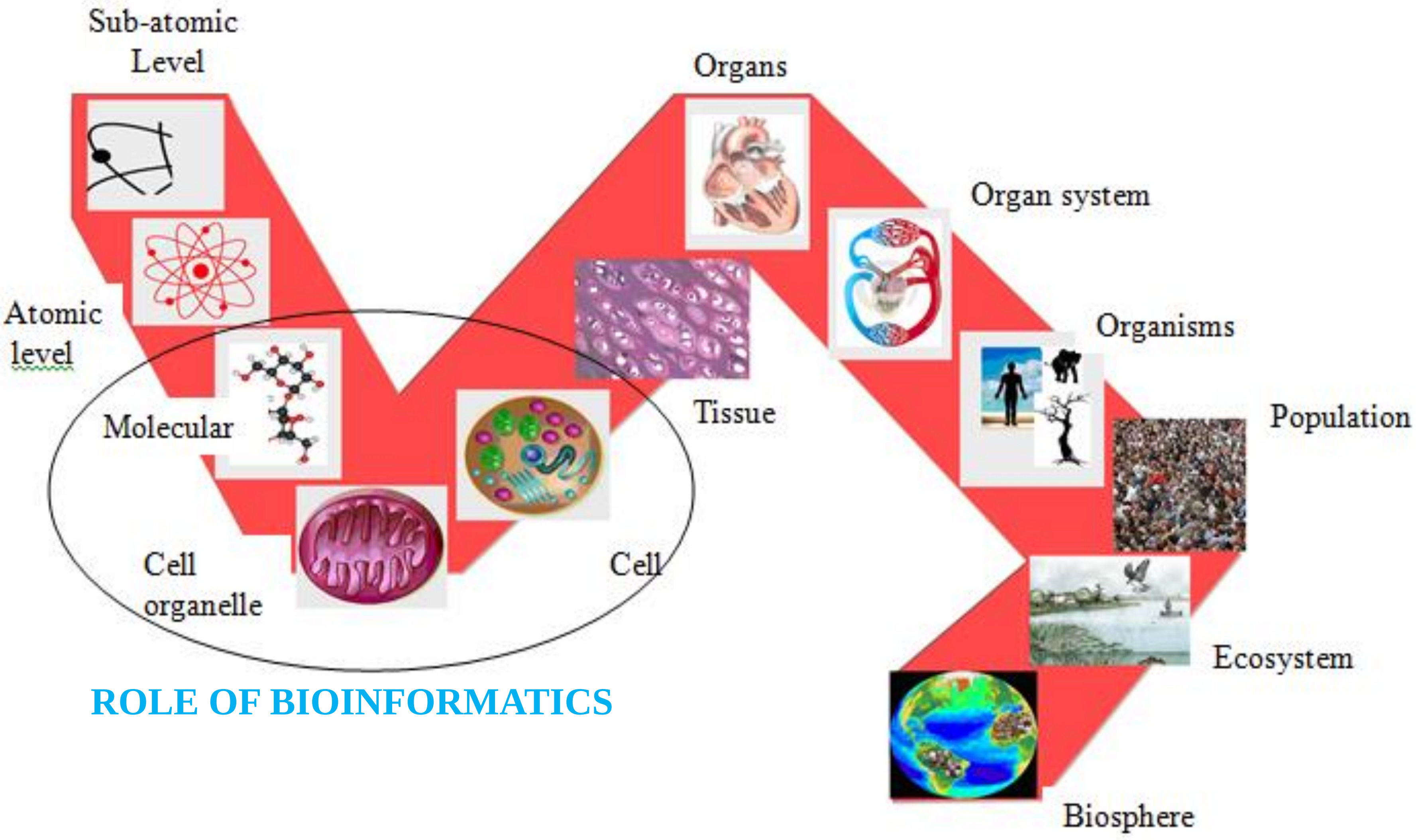
Bioinformatics

Bioinformatics - play with sequences & structures

```
GAATTCTCTTTGGTATCCAATGAAGAAATCGAATCCATACCCATAG
TTCAGGAGAAAATAAGACCGAAGCTGCTCAATTAGGCGCAATTGAT
GTGAAACTTGCCAGCTTACTTCGGCATGTCCTGGTCATTTTGGAAA
CAACCATTATTTAAAGTCGCATTTAAAAAACTTGTTGAAAATATTT
TTCTTTCTGTGGTGCTTTACAAAATCTTGAACTTCTGGAATTGATC
AACGAAATACTGGAATAACAGTTAAAGATCGTGCTGCTTTTAAAAA
ACCAAACAAAGCAAATTCAAGTGTATTGCACCTAATTGCCAAAAAC
ACAATATTCGAAAAATAATAACTTTATATATAATTCGGGTACTACA
TGGATAACAGGCATGTGTTTAATATCTTACAAAATCTTCCACAAAC
TTAACCCCTTCGAATGCTCATCAAATCGTATCTCCCGAAAATGTCT
TATCTTACTTCCACCACATAATCTACGAACTATCAATGTTTATGAT
GTTTGTTAACAAGTGATTTGAATCTGATAATGCGAAGAGTTGCTAA
GCAAAAATACAAAAAAATCTTGGATTCTATCGATAACAGCCGAGGTG
TACAAATAAAAAGCTTACTTTGGATACTTTGACAGGTGGACACTCA
TGCGAAGTTATATTAATGGCAAACGTATTCCTGAGACTGCCAGAGC
TCTATGAATAAAACTGGCTTTATTGAAGTACCATCTTACATTTTAA
TGTTGTCTTTTATAATCACGTTACGAAAGATAACATACTCAAAAGT
AAGCTTTTCTAACATATATCAAAAGTGATCATAATTCTGAAAATCC
GATTTAGCACAGAAGAATGGATATTTAACCTTGGCTCCTAATTTCG
AAAAAGGAAAGAGGAAGGTGGTTTTGTAACTATTTGCAGACATCCA
```



ORGANIZATION OF LIFE



ROLE OF BIOINFORMATICS

WHAT IS BIOINFORMATICS?

Computational Biology/Bioinformatics is the application of computer sciences and allied technologies to answer the questions of Biologists, about the mysteries of life.

It has evolved to serve as the bridge between:

- ✓ Observations (**data**) in diverse biologically-related disciplines and
- ✓ The derivations of understanding (**information**)

APPLICATIONS OF BIOINFORMATICS

- ➢ Computer Aided Drug Design
- ➢ Microarray Bioinformatics
- ➢ Proteomics
- ➢ Genomics
- ➢ Biological Databases
- ➢ Phylogenetics
- ➢ Systems Biology

WHAT IS A BIO-SEQUENCE?

DNA, RNA or protein information represented as a series of bases (or amino acids) that appear in bio-molecules. The method by which a bio-sequence is obtained is called ***Bio-sequencing.***

GTCCTGATAAGTCAGTGTCTC
GAGTCTAGCTTCTGTCCATGC
ATCATGTCCATGTTCTAGTCA
GATAGTTGATTCTAGTGTCCT

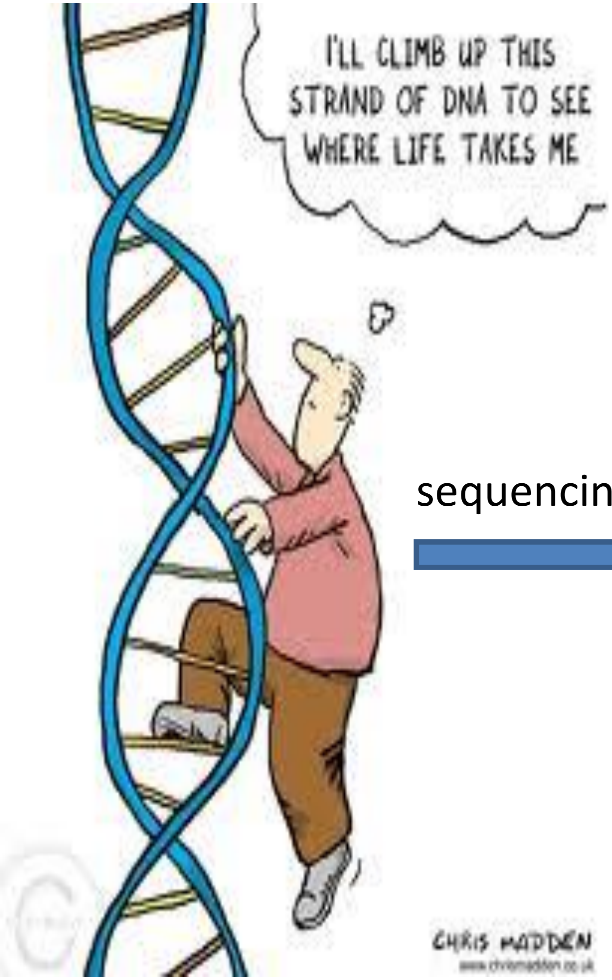
DNA/ RNA SEQUENCE

TPPUQWRDCCLKSWCUWM
ESPWYZWEGHILDDFPTCTW
CCDTWCUWGHISTDTKKSUN
RGHPPHHLDTWQESRNDCQE

PROTEIN SEQUENCE

WHAT IS SEQUENCE ALIGNMENT?

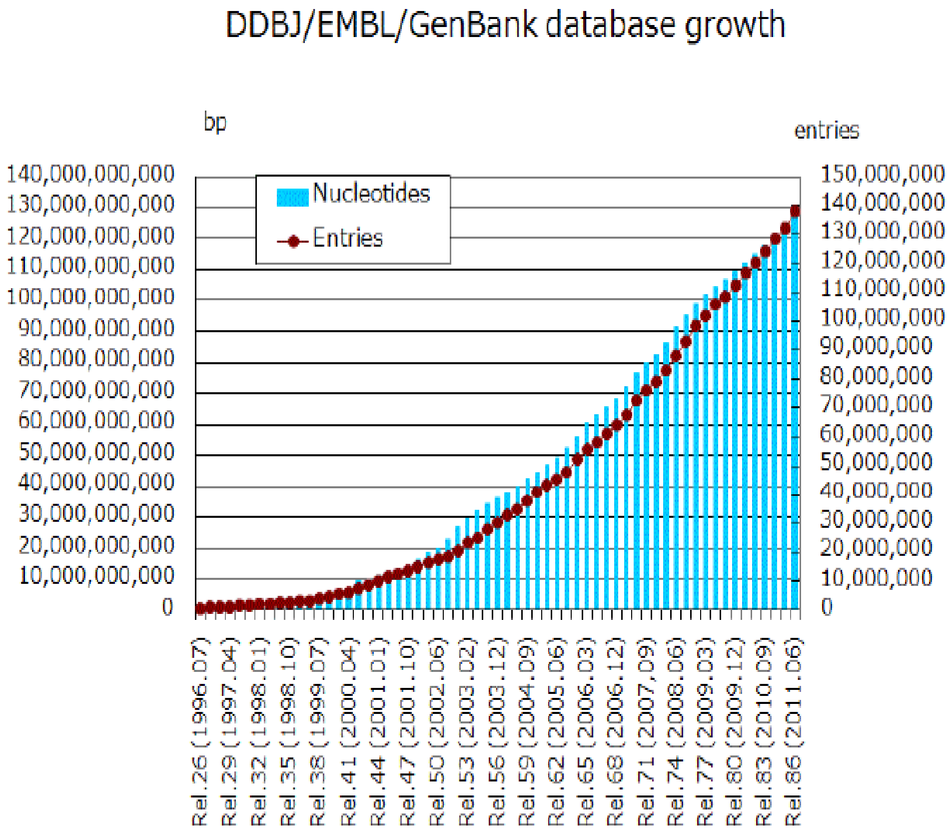
Arranging DNA/protein sequences side by side to study the extent of their similarity



sequencing

CRISIS AFTER DATA EXPLOSION!!

DATA EXPLOSION TREND



SOLUTION??

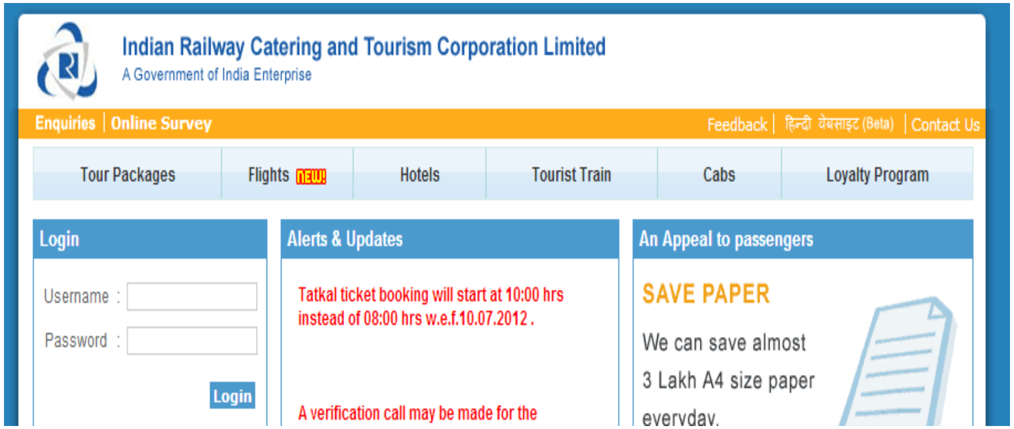
BIOLOGICAL DATABASES

BIOLOGICAL DATABASES

WHAT IS A DATABASE?

A structured set of data held in a computer, esp. one that is accessible in various ways.

POPULAR DATABASE WEBSITES

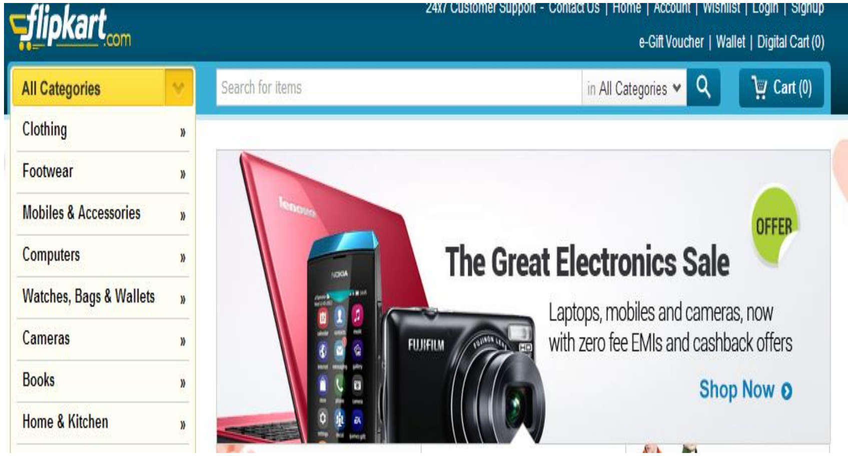


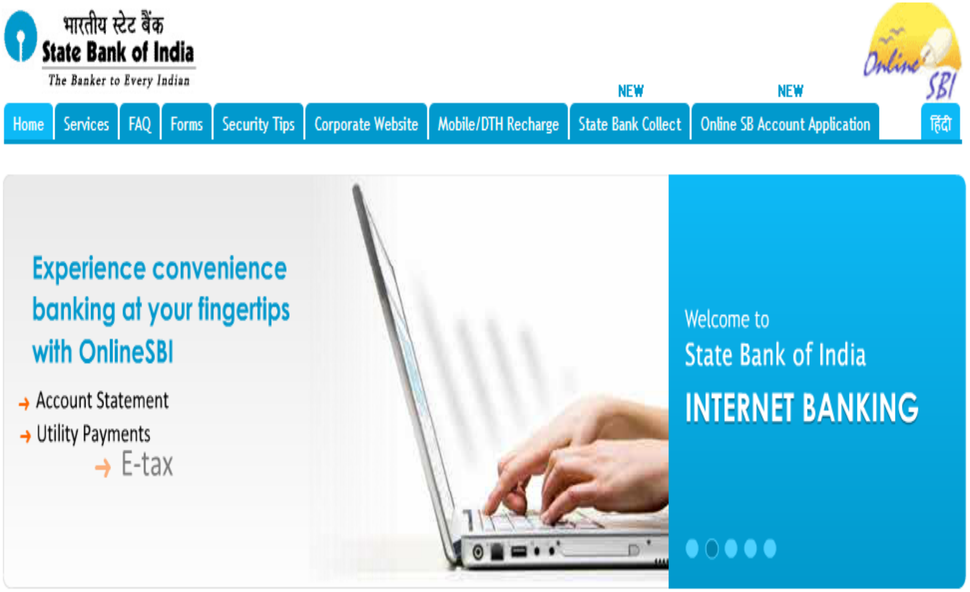
Tatkal ticket booking will start at 10:00 hrs instead of 08:00 hrs w.e.f.10.07.2012 .

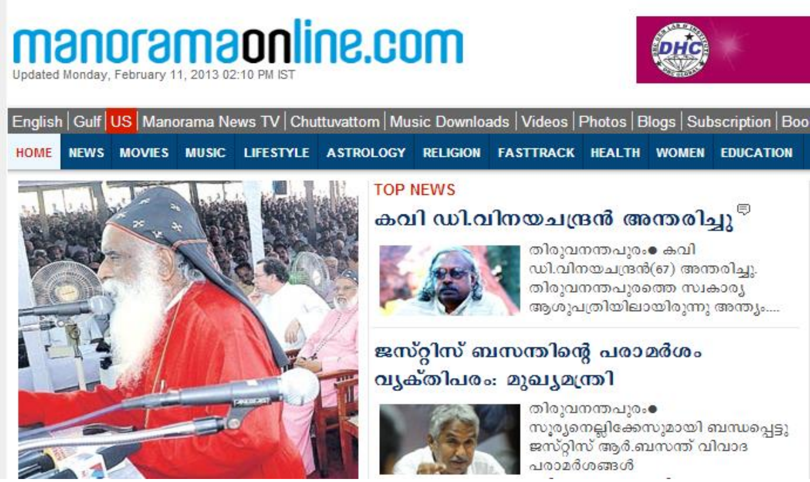
A verification call may be made for the

SAVE PAPER

We can save almost 3 Lakh A4 size paper everyday.







കവി ഡി.വിനയചന്ദ്രൻ അന്തരിച്ചു

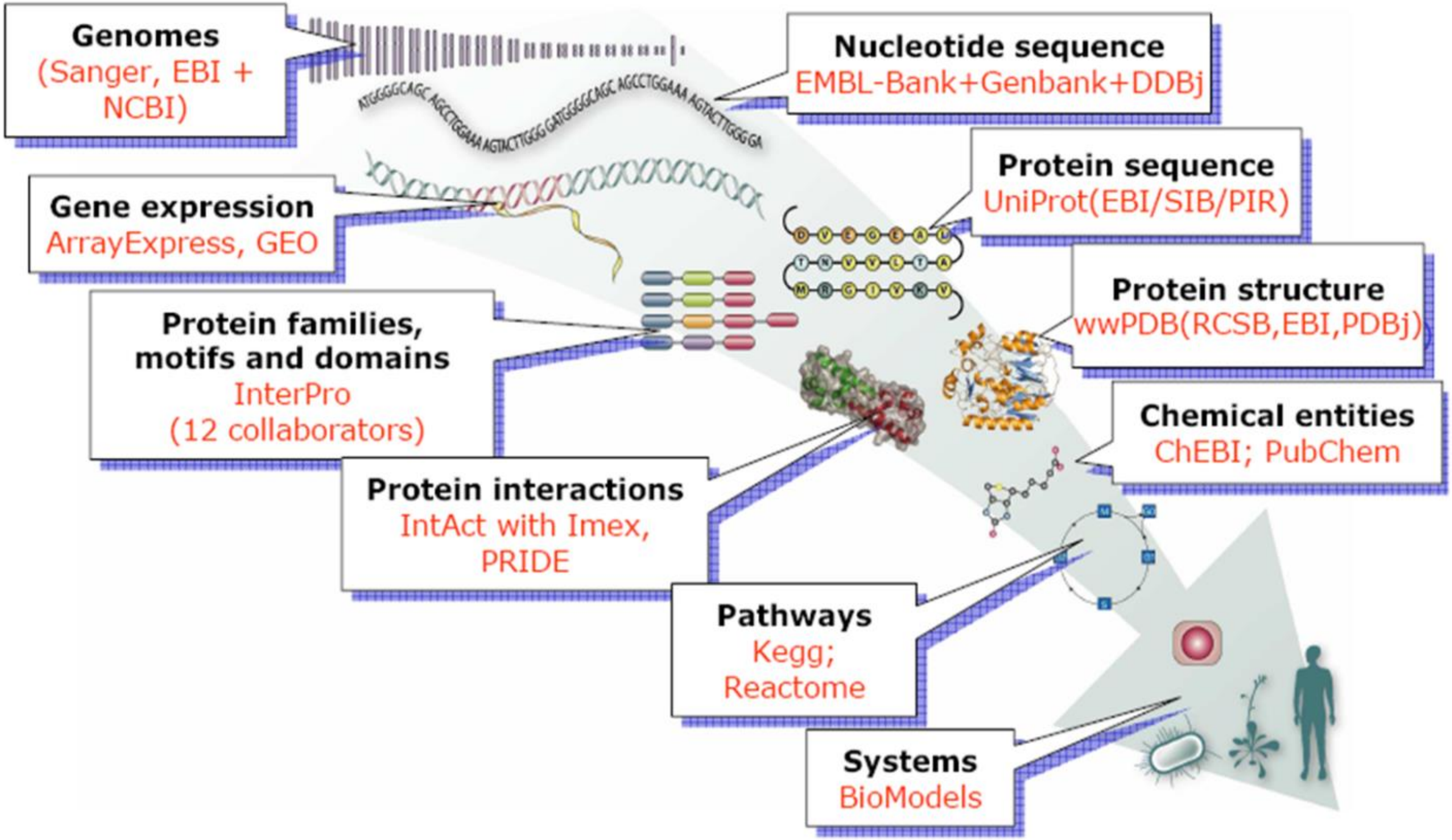
തിരുവനന്തപുരം● കവി ഡി.വിനയചന്ദ്രൻ(67) അന്തരിച്ചു. തിരുവനന്തപുരത്തെ സ്വകാര്യ ആശുപത്രിയിലായിരുന്നു അന്ത്യം....

ജസ്റ്റിസ് ബസന്തിന്റെ പരാമർശം വ്യക്തിപരം: മുഖ്യമന്ത്രി

തിരുവനന്തപുരം● സൂര്യനെല്ലിക്കേസുമായി ബന്ധപ്പെട്ടു ജസ്റ്റിസ് ആർ.ബസന്ത് വിവാദ പരാമർശങ്ങൾ

BIOLOGICAL DATABASES

Core Molecular Data



Genomes
(Sanger, EBI + NCBI)

Nucleotide sequence
EMBL-Bank+Genbank+DDBj

Gene expression
ArrayExpress, GEO

Protein sequence
UniProt(EBI/SIB/PIR)

Protein families, motifs and domains
InterPro
(12 collaborators)

Protein structure
wwPDB(RCSB,EBI,PDBj)

Chemical entities
ChEBI; PubChem

Protein interactions
IntAct with Imex, PRIDE

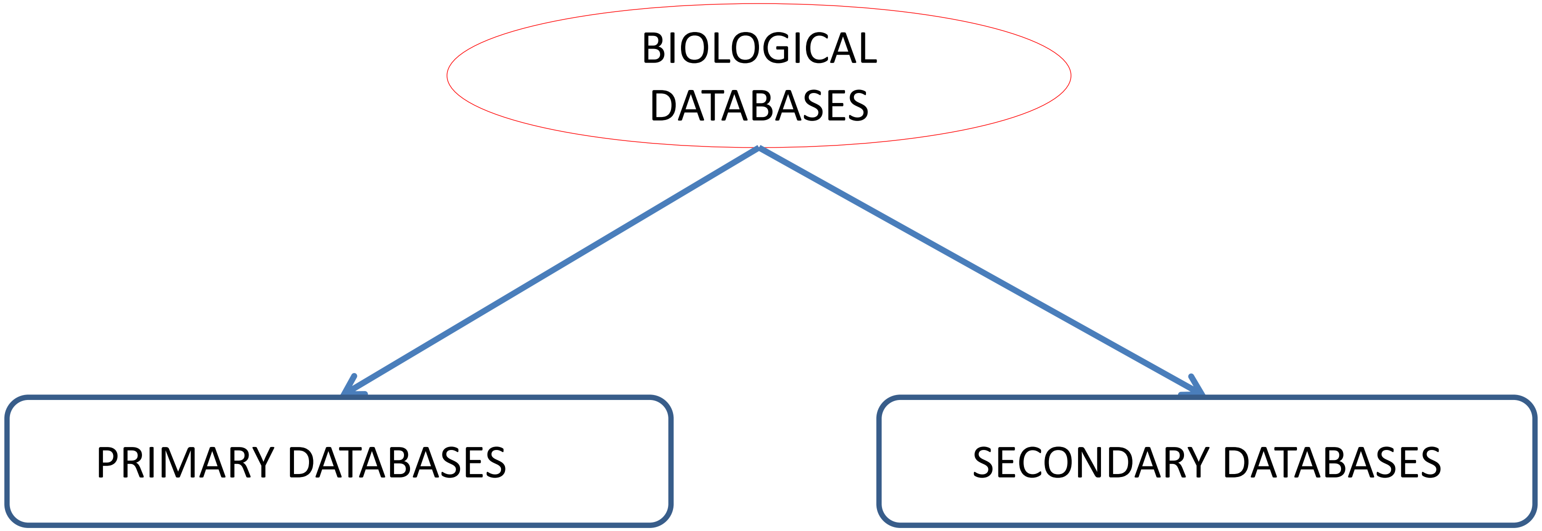
Pathways
Kegg;
Reactome

Systems
BioModels

CLASSIFICATION OF BIOLOGICAL DATABASES

- **Based on data source**
- **Based on data type**

➢ BASED ON DATA SOURCE



BIOLOGICAL DATABASES

PRIMARY DATABASES

SECONDARY DATABASES

- First-hand information of experimental data from scientists and researchers
- Data not edited or validated
- Raw and original submission of data
- Made available to public for annotation

- Derived from information gathered in primary database
- Data is manually curated and annotated
- Data of highest quality as it is double checked

PRIMARY DATABASES

	Database	Website
1.	NCBI (National Centre for Biotechnology Information)	www.ncbi.nlm.nih.gov
2.	DDBJ (DNA Data Bank of Japan)	www.ddbj.nig.ac.jp
3.	EMBL(European Molecular Biology Laboratory)	www.ebi.ac.uk/embl
4.	PIR (Protein Information Resource)	www.pir.georgetown.edu
5.	PDB (Protein Data Bank)	www.rcsb.org/pdb
6.	NDB(Nucleotide Data Bank)	www.ndbserver.rutgers.edu
7.	SwissProt (Protein- only sequence database)	www.expasy.ch

SECONDARY DATABASES

	Database	Website
1.	PROSITE (Protein domains, families, functional sites)	www.expasy.org/prosite
2.	Pfam (Protein families)	www.sanger.ac.uk/pfam
3.	SCOP (Structural Classification Of Proteins)	www.scop.mrc-lmb.cam.ac.uk/scop
4.	CATH (Class, Architecture, Topology, Homologous Super Family of Proteins)	www.cathdb.info
5.	OMIM (Online Mendelian Inheritance in Man)	www.ncbi.nlm.nih/omim
6.	KEGG (Kyoto Encyclopedia of Genes and Genome)	www.genome.jp/kegg/pathway.html
7.	MetaCyc (Enzyme Metabolic Pathways)	www.metacyc.org

- Based on type of data

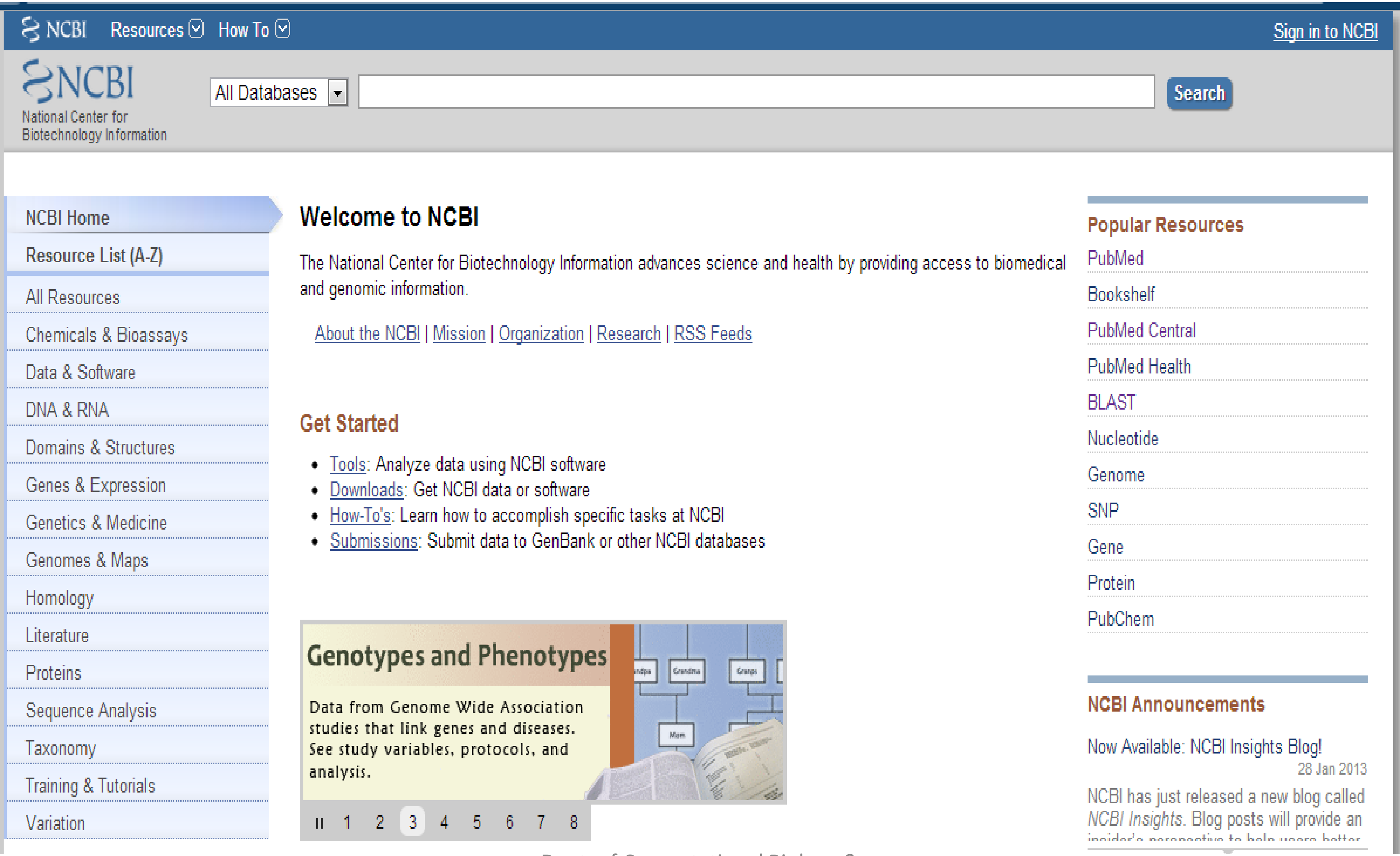
BASED ON THE TYPE OF DATA

BIOLOGICAL DATABASES

- NUCLEOTIDE SEQUENCE DATABASE
- PROTEIN SEQUENCE DATABASE
- GENOME DATABASE
- GENE EXPRESSION DATABASE
- ENZYME DATABASE
- STRUCTURE DATABASE
- PROTEIN INTERACTION DATABASE
- PATHWAY DATABASE
- LITERATURE DATABASE

NUCLEOTIDE SEQUENCE DATABASES

NCBI- National Centre for Biotechnology Information



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 | RSS Feeds

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

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

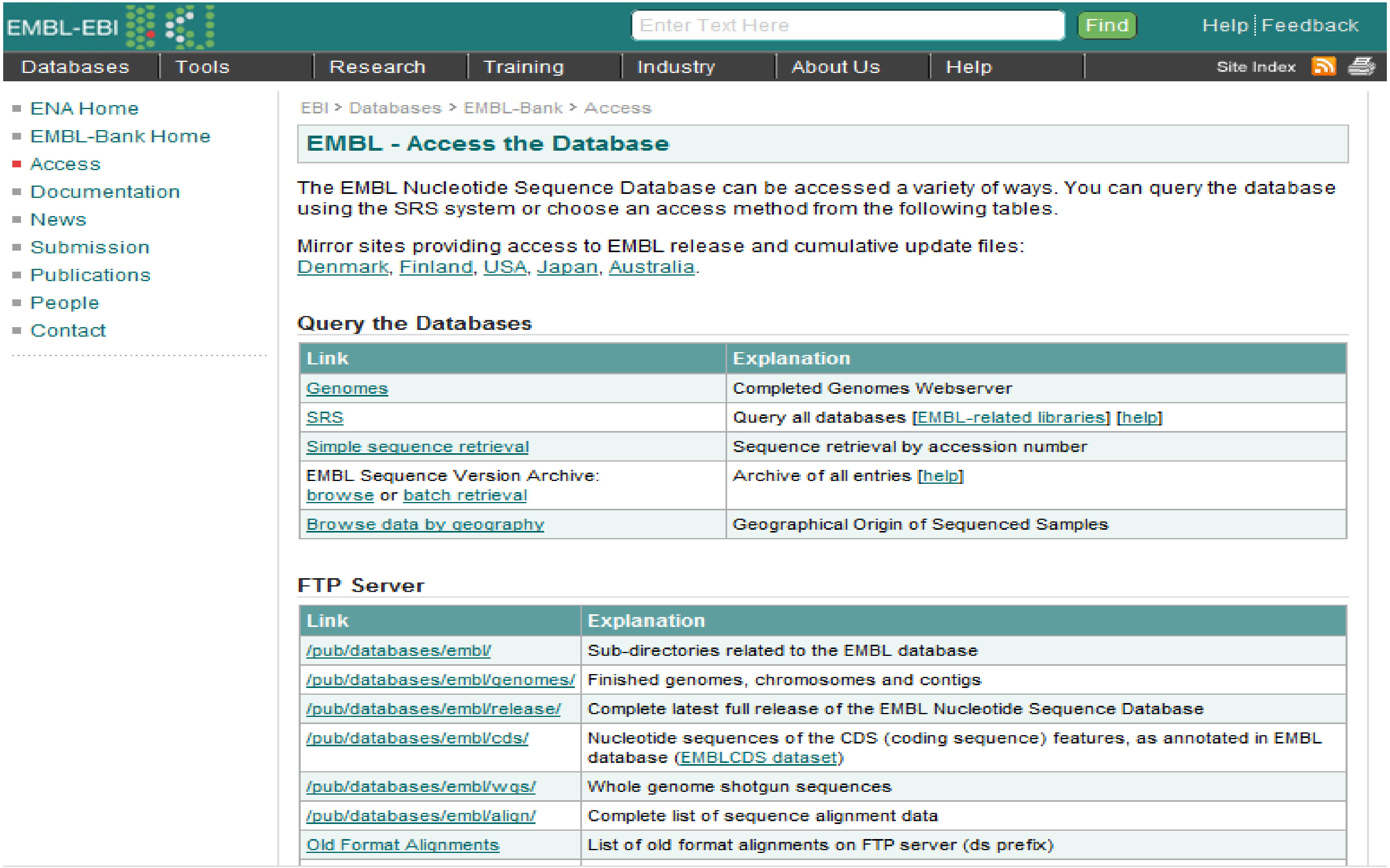
NCBI Announcements

Now Available: NCBI Insights Blog!

28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an

EMBL – European Molecular Biology Lab



EMBL - Access the Database

The EMBL Nucleotide Sequence Database can be accessed a variety of ways. You can query the database using the SRS system or choose an access method from the following tables.

Mirror sites providing access to EMBL release and cumulative update files:
Denmark, Finland, USA, Japan, Australia.

Query the Databases

Link	Explanation
Genomes	Completed Genomes Webserver
SRS	Query all databases [EMBL-related libraries] [help]
Simple sequence retrieval	Sequence retrieval by accession number
EMBL Sequence Version Archive: browse or batch retrieval	Archive of all entries [help]
Browse data by geography	Geographical Origin of Sequenced Samples

FTP Server

Link	Explanation
/pub/databases/embl/	Sub-directories related to the EMBL database
/pub/databases/embl/genomes/	Finished genomes, chromosomes and contigs
/pub/databases/embl/release/	Complete latest full release of the EMBL Nucleotide Sequence Database
/pub/databases/embl/cds/	Nucleotide sequences of the CDS (coding sequence) features, as annotated in EMBL database (EMBLCDS dataset)
/pub/databases/embl/wgs/	Whole genome shotgun sequences
/pub/databases/embl/align/	Complete list of sequence alignment data
Old Format Alignments	List of old format alignments on FTP server (ds prefix)

DDBJ- DNA DATA BANK OF JAPAN

DDBJ : DNA Data Bank of Japan

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

Hot Topics

- 2012.02.14 DDBJ provides a new entry retrieval system from Feb.15, 2012
- 2012.02.13 Change of creating links to DDBJ entries
- 2012.02.02 Release of EST data of eucaly (*Eucalyptus camaldulensis*)

Maintenance

- 2012.02.20 Temporary suspension of databank activities(as of Feb. 15)
- 2012.02.14 (Feb. 25) NIG and DDBJ network temporary down
- 2012.02.13 (resume)(2012/2/23-27) "SAKURA" data submission system temporary unavailable

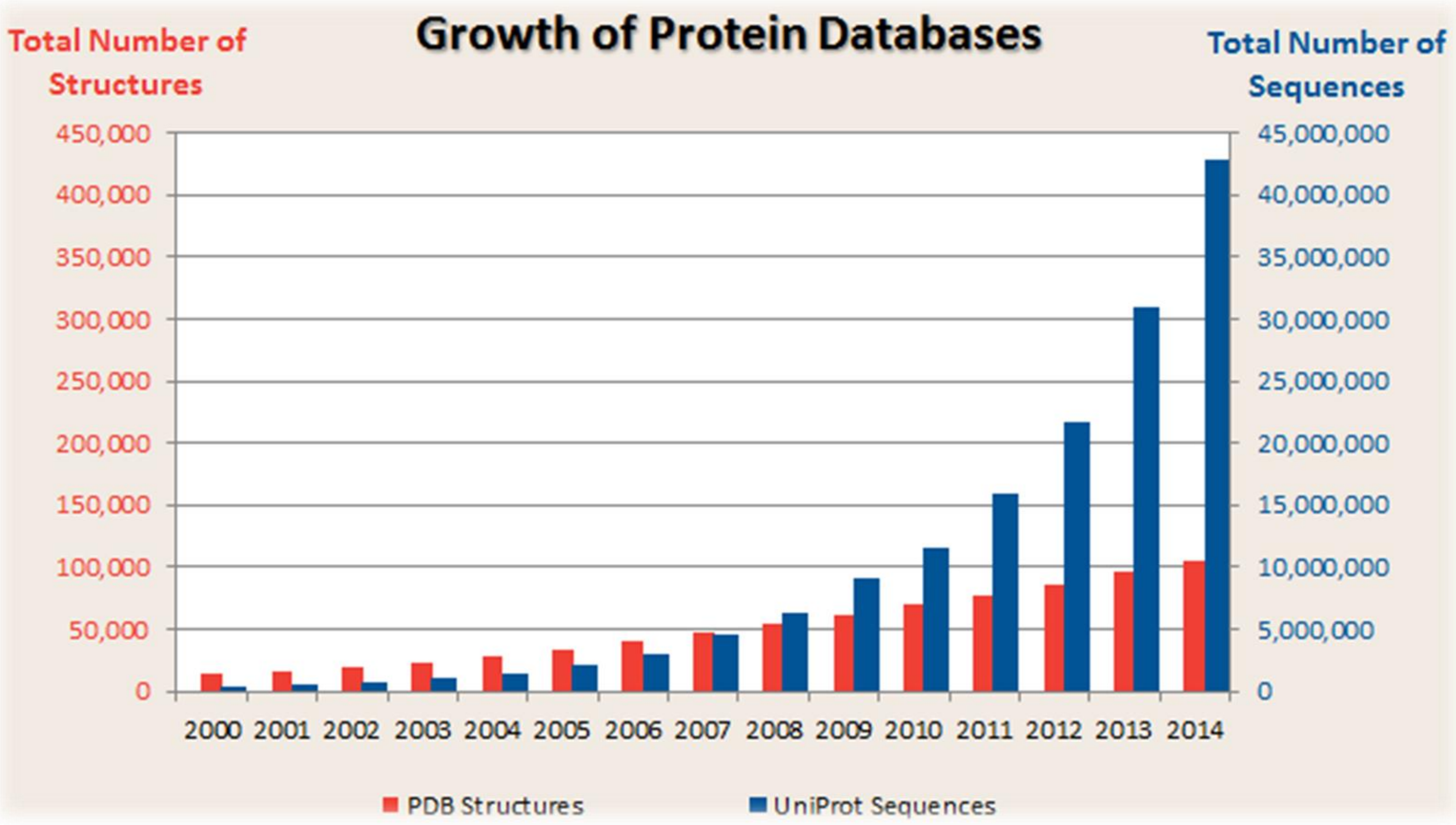
Information

- News for computer system replace in Mar., 2012 (title list)

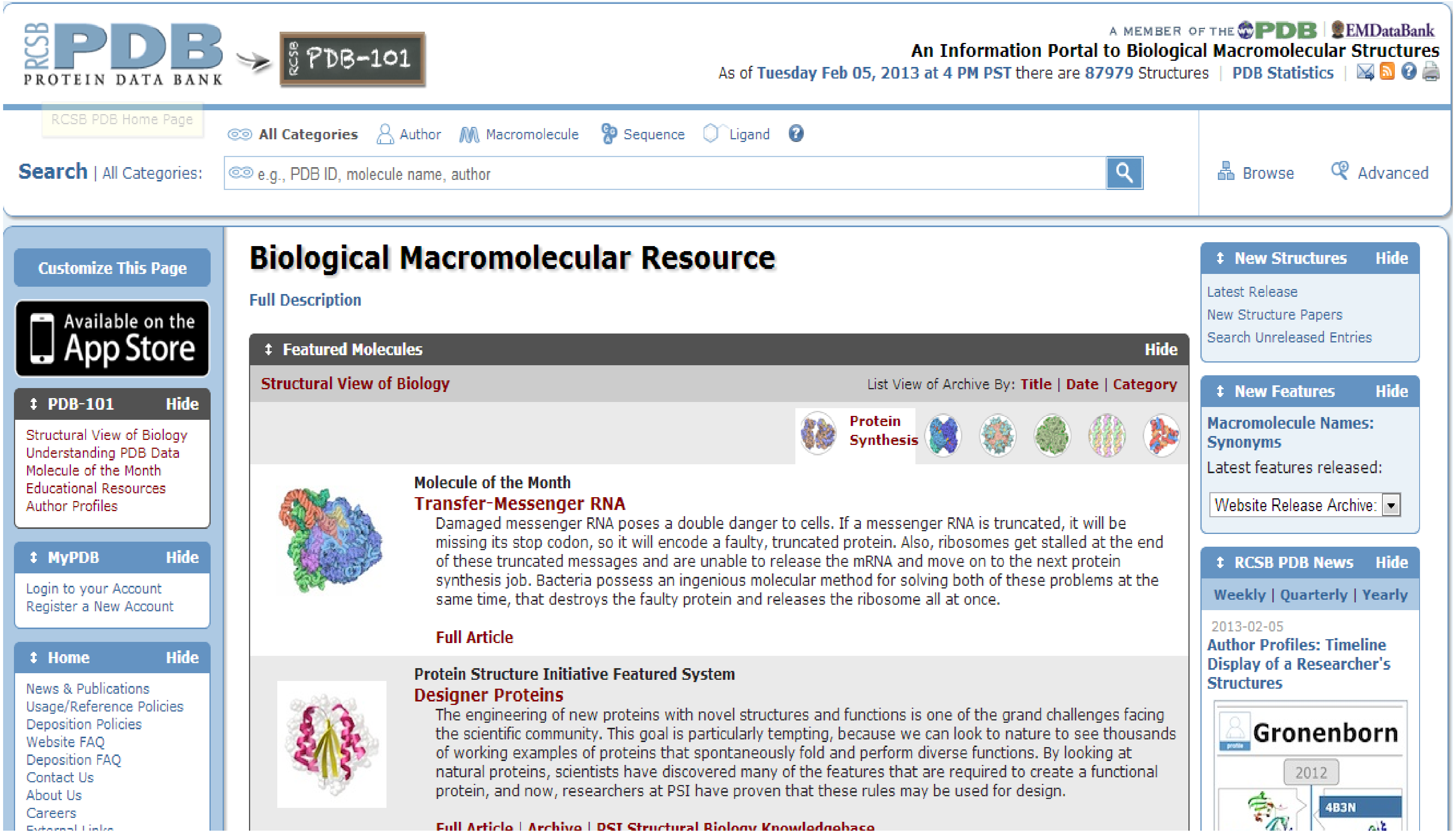
Sequence Data Submission

FTP- Web API

PROTEIN SEQUENCE DATABASE


Growth of Protein Databases
Total Number of Structures
Total Number of Sequences
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
🟥 PDB Structures
🟦 UniProt Sequences

PDB- PROTEIN DATA BANK



RCSB PDB PROTEIN DATA BANK → PDB-101

A MEMBER OF THE wwPDB | EMDataBank

An Information Portal to Biological Macromolecular Structures

As of Tuesday Feb 05, 2013 at 4 PM PST there are 87979 Structures | PDB Statistics

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

- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links

Biological Macromolecular Resource

Full Description

Featured Molecules Hide

Structural View of Biology — List View of Archive By: Title | Date | Category

Protein Synthesis

Molecule of the Month

Transfer-Messenger RNA

Damaged messenger RNA poses a double danger to cells. If a messenger RNA is truncated, it will be missing its stop codon, so it will encode a faulty, truncated protein. Also, ribosomes get stalled at the end of these truncated messages and are unable to release the mRNA and move on to the next protein synthesis job. Bacteria possess an ingenious molecular method for solving both of these problems at the same time, that destroys the faulty protein and releases the ribosome all at once.

Full Article

Protein Structure Initiative Featured System

Designer Proteins

The engineering of new proteins with novel structures and functions is one of the grand challenges facing the scientific community. This goal is particularly tempting, because we can look to nature to see thousands of working examples of proteins that spontaneously fold and perform diverse functions. By looking at natural proteins, scientists have discovered many of the features that are required to create a functional protein, and now, researchers at PSI have proven that these rules may be used for design.

Full Article | Archive | PSI Structural Biology Knowledgebase

New Structures Hide

- Latest Release
- New Structure Papers
- Search Unreleased Entries

New Features Hide

Macromolecule Names: Synonyms

Latest features released:

Website Release Archive: ▾

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2013-02-05

Author Profiles: Timeline Display of a Researcher's Structures

PATHWAY DATABASES

KEGG- KYOTO ENCYCLOPEDIA OF GENES AND GENOMES

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

Select prefix: map

Enter keywords:

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps (see new maps, change history, and last updates) representing our knowledge on the molecular interaction and reaction networks for:

- **0. Global Map**
- **1. Metabolism**
  - Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid Glycan
  - Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Overview
- **2. Genetic Information Processing**
- **3. Environmental Information Processing**
- **4. Cellular Processes**
- **5. Organismal Systems**
- **6. Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

- **7. Drug Development**

Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretaion of higher-level systemic functions.

- Search objects in KEGG pathways
- Search&Color objects in KEGG pathways

GENOME DATABASE

WORMBASE : has the entire genome of *C. elegans* and other nematodes

WormBase Release WS229

Web Site Directory

About this release Release Notes, New/changed genes
General Searches WormBase Class Browser, Wormbase Query Language Search, AQL Search
Sequences *C. elegans* Genome, *C. briggsae* Genome, **Gene**, Blast / Blat, e-PCR, Gene Ontology, Synteny Viewer,
Cells and Gene Expression
Cell and Pedigree, Neurons, Expression Pattern, Expression profile
Genetics, Strains, and Phenotypes Genetic interval, Rearrangements, Balancers, Clone, Allele, SNPs, Markers, and Strains, Strain Report, Phenotypes, RNAi
Batch Queries WormMart [about...], Batch Genes, Batch Sequences
Downloads and Data Mining **Bulk Downloads**, Linking to WormBase and Data Mining...
Community **Worm Meetings**, Worm Community Discussion Forum, WormBase Wiki, Mailing Lists
About WormBase User's Guide, FAQs, C. elegans gene and protein nomenclature, Newsletters, Mission Statement, Citing Wormbase, Positions available, People, Acceptable Use Policy
WormBase News (RSS)

News and Notes

» 14 Feb 2012: ModENCODE Data Access Webinar, March 1st 2012

The next webinar held by the Data Coordination Center for modENCODE is scheduled on Thursday Mar 1, 12:00 EST, Topic: "Understanding worm and fly transcriptome data." This webinar will introduce participants to the transcriptome analysis data generated in Robert Waterston's lab (*C. elegans*) and Susan Celniker's lab (*D. melanogaster*). Space is limited so advance registration is necessary. Please email help@modencode.org to reserve your spot. More details and the schedule are available on the modENCODE wiki.

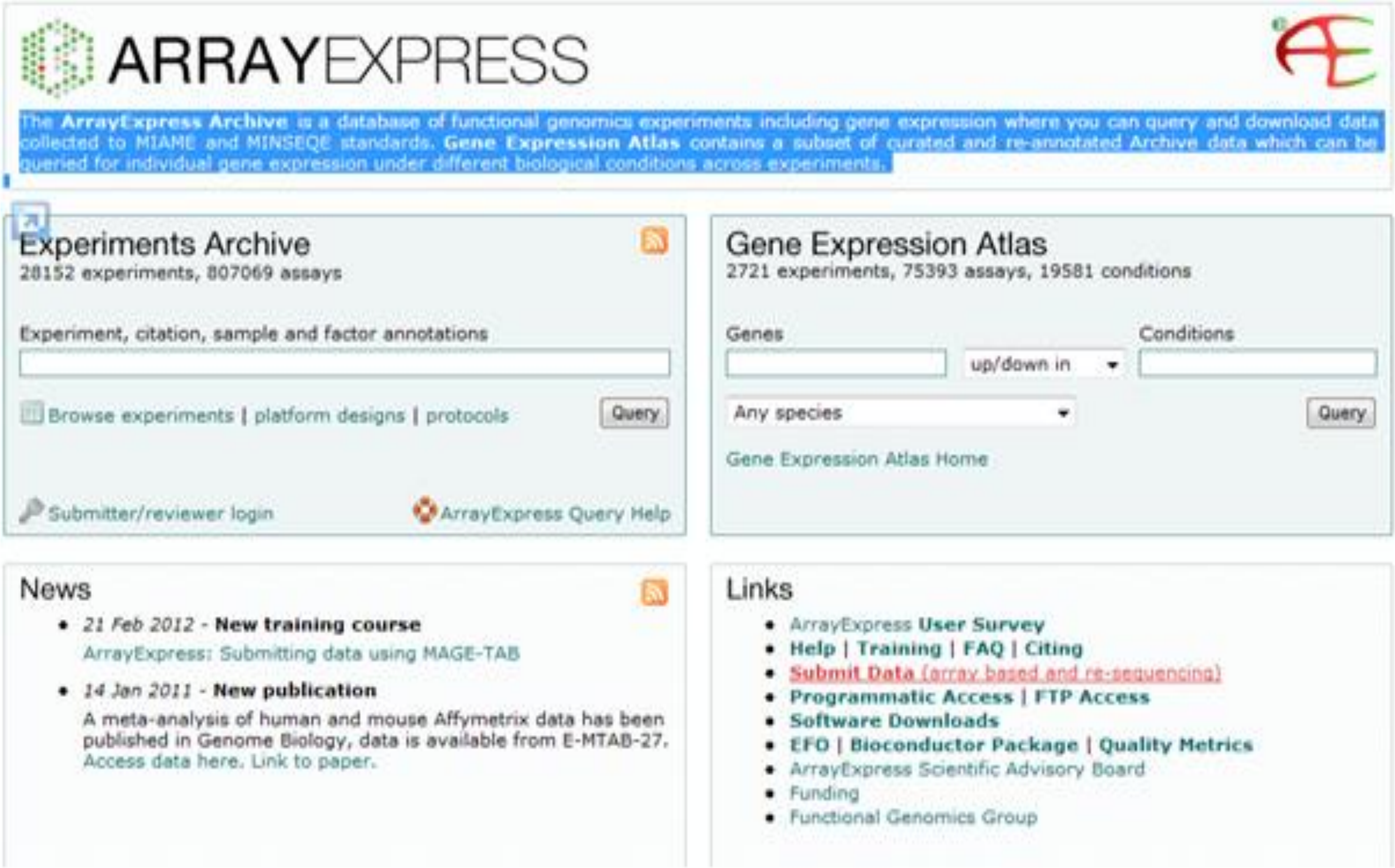
» 04 Feb 2012: New cloud AMI of WormBase released for WS229

We just finished bundling the latest Amazon EC2 AMI of WormBase for WS229. This AMI includes an updated genome browser, several new species, and of course all the databases and software necessary to run your own instance of WormBase. Search for WormBase on EC2 or for the AMI ID: ami-e9994980.

» 04 Feb 2012: New release of WormBase: WS229

WormBase has been updated to the WS229 release of the database.

GENE EXPRESSION DATABASE



The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive

28152 experiments, 807069 assays

Experiment, citation, sample and factor annotations

Gene Expression Atlas

2721 experiments, 75393 assays, 19581 conditions

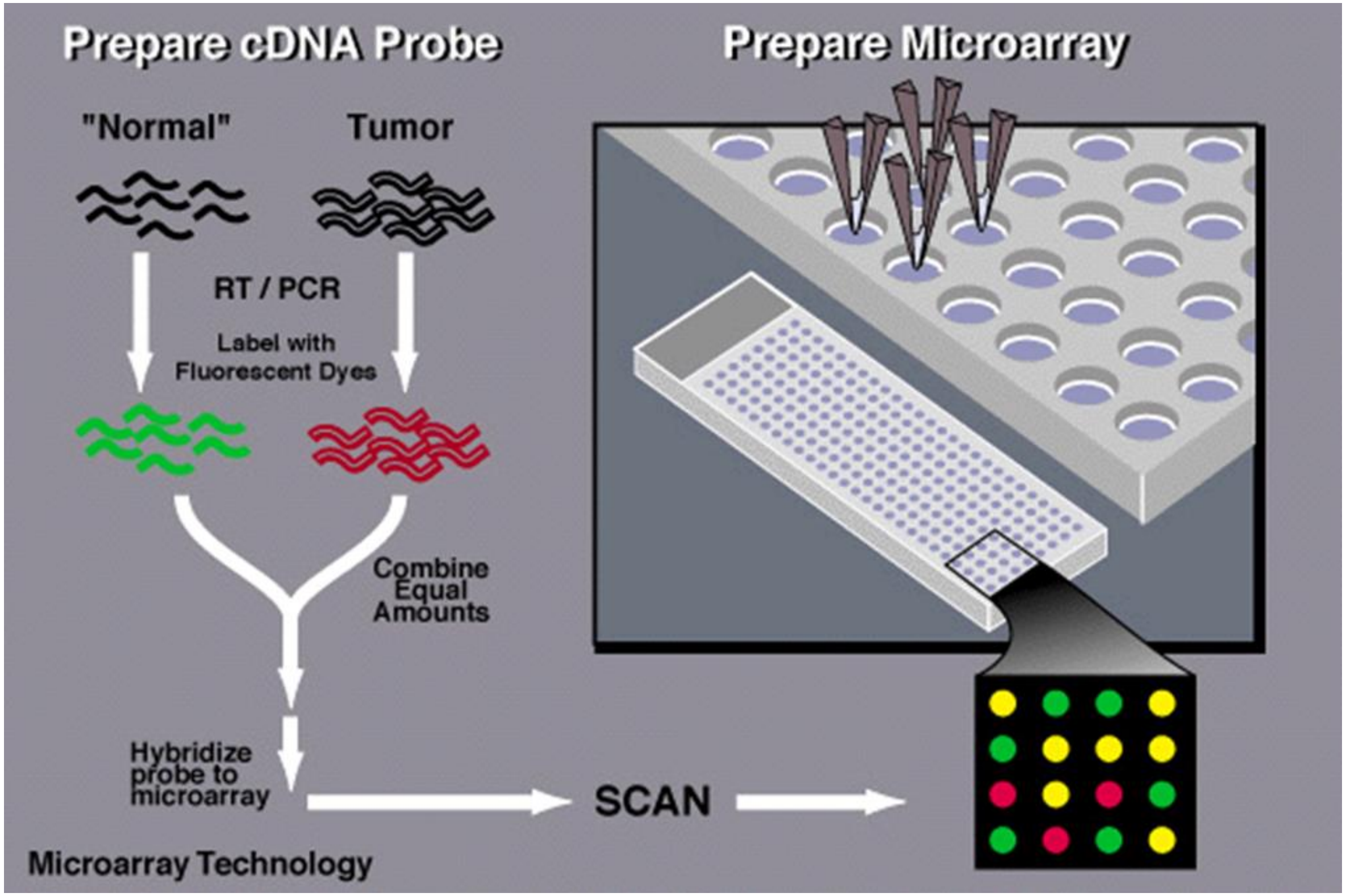
News

- *21 Feb 2012* - **New training course**
  ArrayExpress: Submitting data using MAGE-TAB
- *14 Jan 2011* - **New publication**
  A meta-analysis of human and mouse Affymetrix data has been published in Genome Biology, data is available from E-MTAB-27. Access data here. Link to paper.

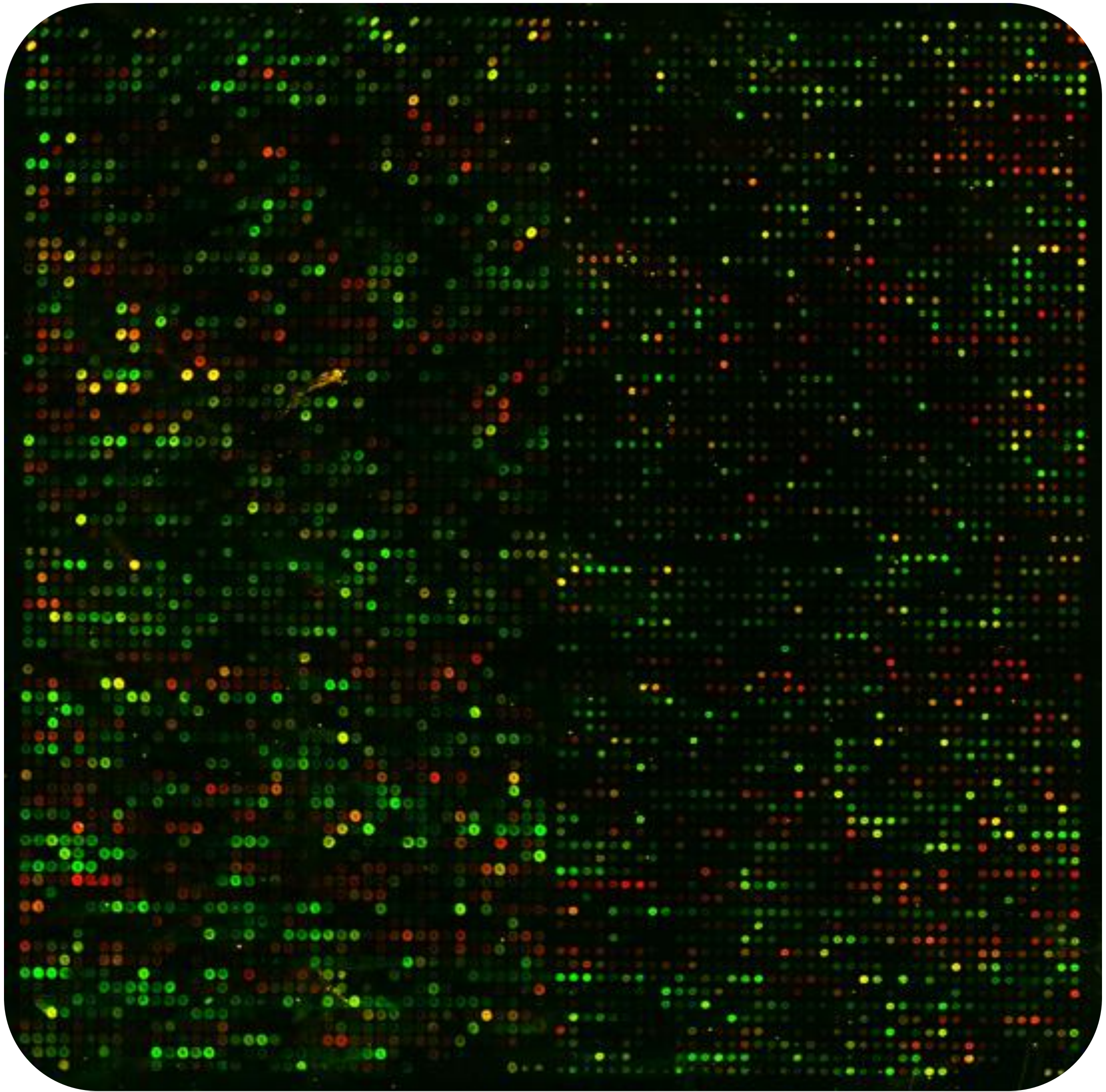
Links

- ArrayExpress **User Survey**
- **Help | Training | FAQ | Citing**
- **Submit Data** (array based and re-sequencing)
- **Programmatic Access | FTP Access**
- **Software Downloads**
- **EFO | Bioconductor Package | Quality Metrics**
- ArrayExpress Scientific Advisory Board
- Funding
- Functional Genomics Group

Microarrays provide a means to measure gene expression

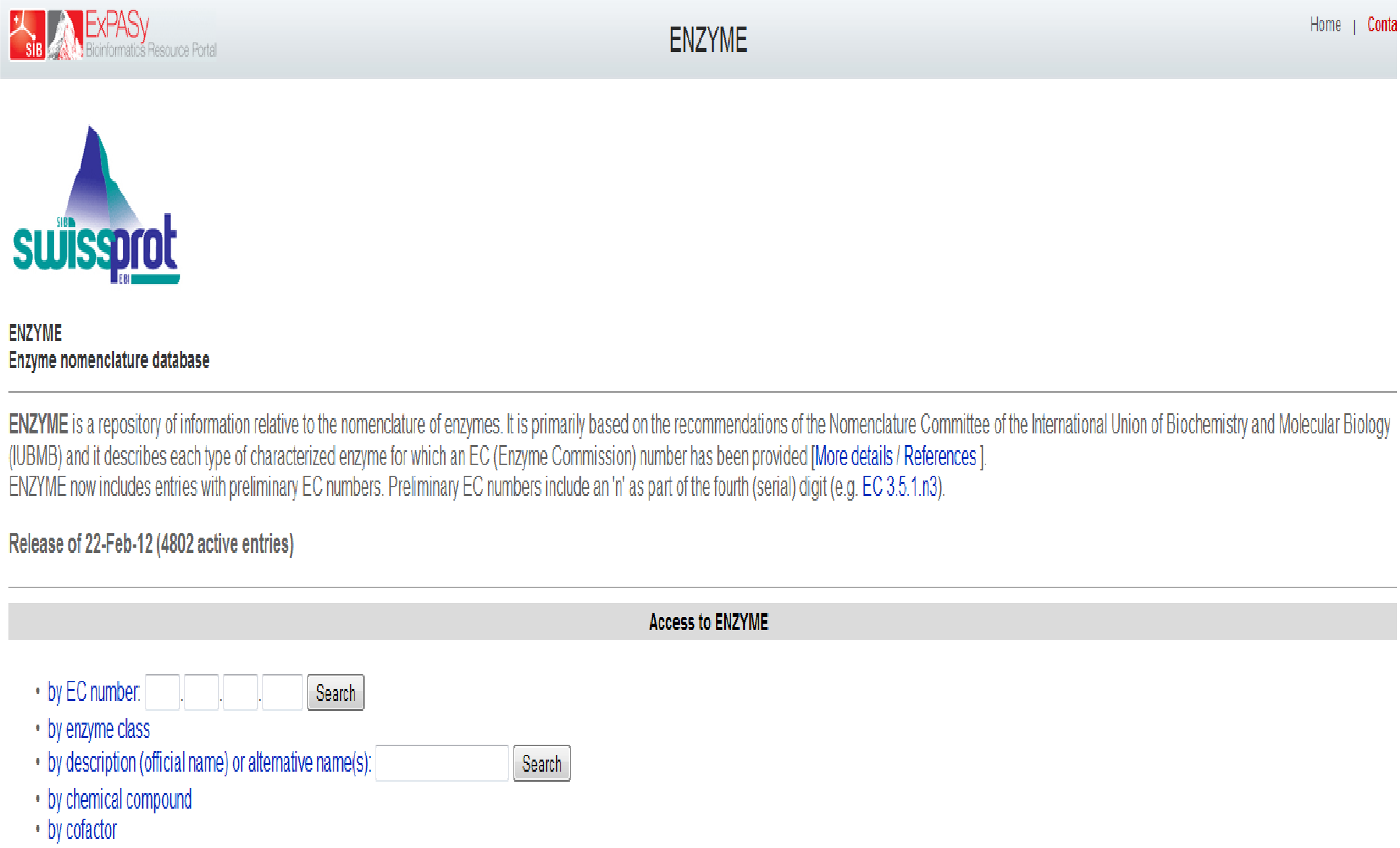


Yeast genome on a chip



ENZYME DATABASE

ENZYME DATABASE OF ExPaSy server



ENZYME

ENZYME
Enzyme nomenclature database

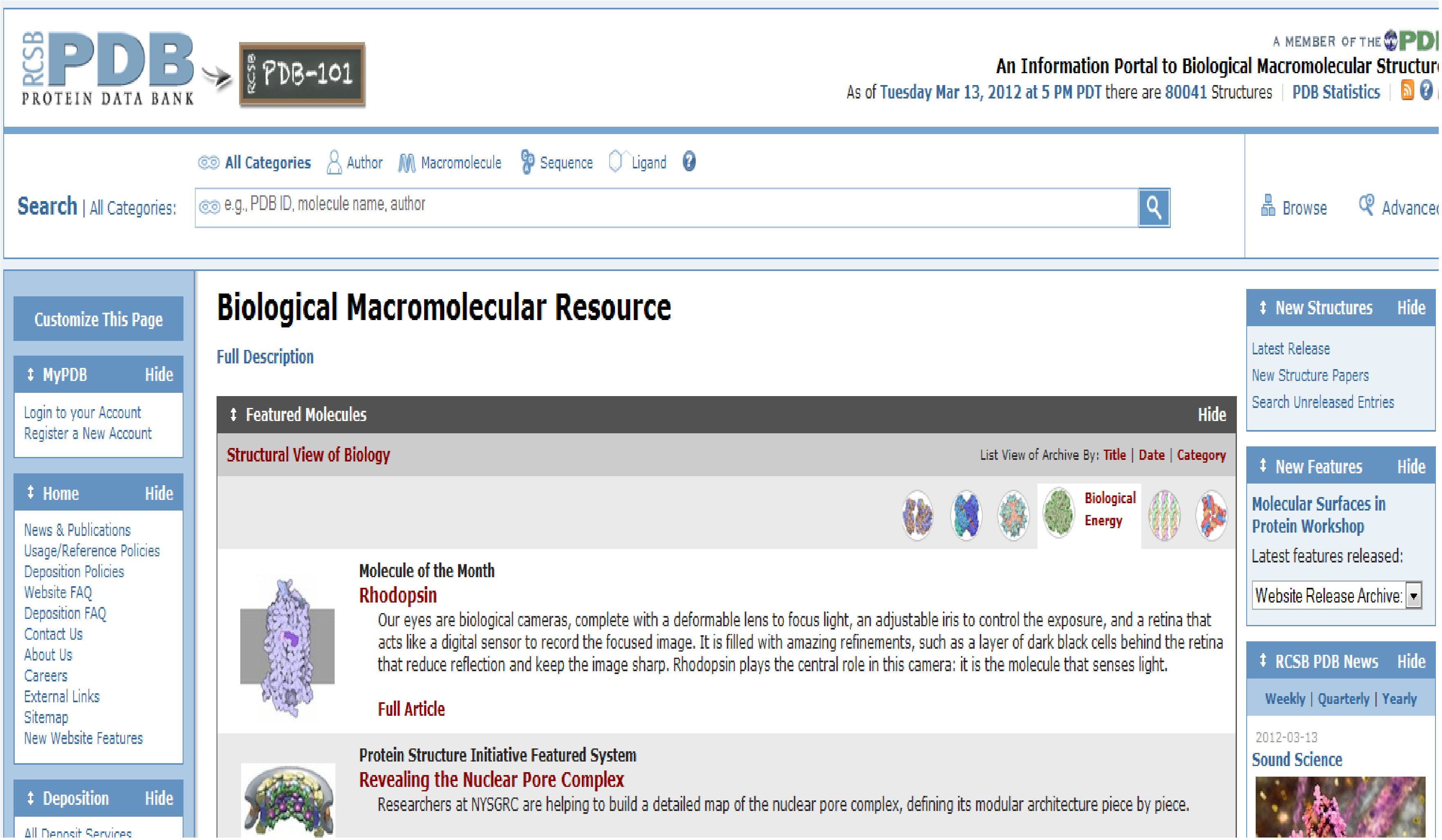
ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [More details / References].
ENZYME now includes entries with preliminary EC numbers. Preliminary EC numbers include an 'n' as part of the fourth (serial) digit (e.g. EC 3.5.1.n3).

Release of 22-Feb-12 (4802 active entries)

Access to ENZYME

- by EC number: ☐ . ☐ . ☐ . ☐ Search
- by enzyme class
- by description (official name) or alternative name(s): ☐ Search
- by chemical compound
- by cofactor

STRUCTURE DATABASE



RCSB PDB PROTEIN DATA BANK

RCSB PDB-101

A MEMBER OF THE PDI

An Information Portal to Biological Macromolecular Structur

As of Tuesday Mar 13, 2012 at 5 PM PDT there are 80041 Structures | PDB Statistics |

All Categories | Author | Macromolecule | Sequence | Ligand

Search | All Categories: e.g., PDB ID, molecule name, author

Browse | Advance

Customize This Page

MyPDB Hide

- Login to your Account
- Register a New Account

Home Hide

- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Deposition Hide

- All Deposit Services

Biological Macromolecular Resource

Full Description

Featured Molecules Hide

Structural View of Biology

List View of Archive By: Title | Date | Category

Biological Energy

Molecule of the Month

Rhodopsin

Our eyes are biological cameras, complete with a deformable lens to focus light, an adjustable iris to control the exposure, and a retina that acts like a digital sensor to record the focused image. It is filled with amazing refinements, such as a layer of dark black cells behind the retina that reduce reflection and keep the image sharp. Rhodopsin plays the central role in this camera: it is the molecule that senses light.

Full Article

Protein Structure Initiative Featured System

Revealing the Nuclear Pore Complex

Researchers at NYSGRC are helping to build a detailed map of the nuclear pore complex, defining its modular architecture piece by piece.

New Structures Hide

- Latest Release
- New Structure Papers
- Search Unreleased Entries

New Features Hide

Molecular Surfaces in Protein Workshop

Latest features released:

Website Release Archive:

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2012-03-13

Sound Science

LITERATURE DATABASE

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Target your results using the improved Limits page!

Click the Limits tab to easily:

- Add an author or journal to your search.
- Limit to citations with links to free full text.
- Select multiple languages, publication types, and do lots more.

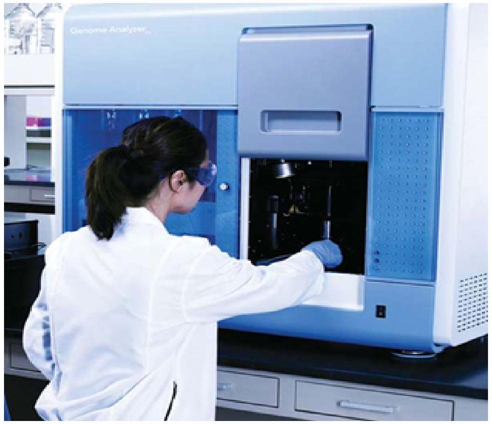
Read the PubMed Help to explore other PubMed search options.

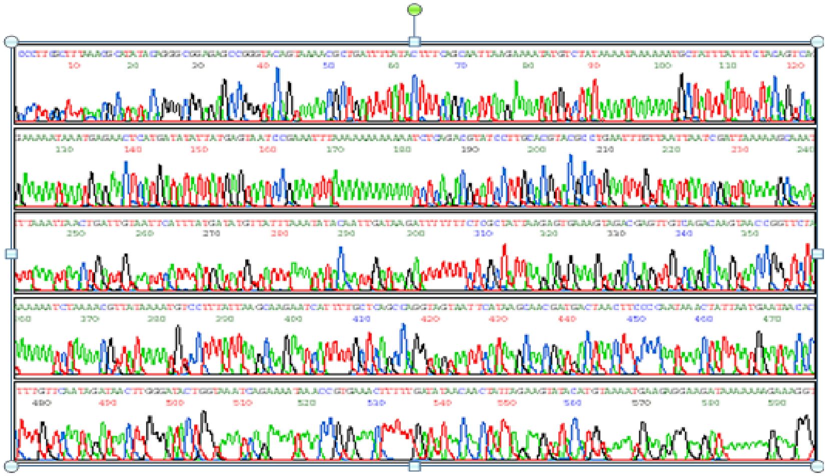
PubMed is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

Use of Databases in Biology- Sequence Analysis

Where do we get these sequences from?

Through genome sequencing projects





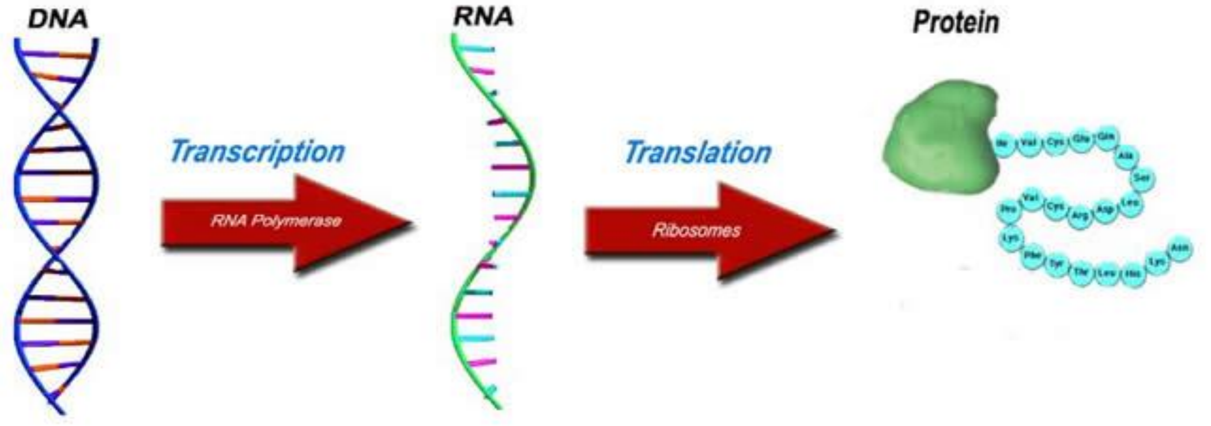
```
AGCGCATCGTCTCAGGTCAGCGCGGCGCTGCTACTCGAGCTCCGAGAGGCCCACGGCC
GGCGTCCACCATGTTGTCGCGAGCGGCGTGCAGGTGAGGCTCCGGATAGCGGGCGCGG
CGGCCGGTCGCGGCCCCGGCGGGGCCTCGGTGGCTTGCCGCTGCTGCCCTCCACTGGC
CCGCTCGCGCTGGCCCCGCGGGCTCCAGGACACCGGCCTGCGGGCAGGTGGTGGGCCG
AGCGCGCCCAGGTGCCGCCGGATCCGCTGGACCGCAGCTTGAGTTTCAGGCTGGAGAC
TGGACGGCCGGCTGGCCGACGGGCATGGAGAGACGTGAATAAGAAATGTGTAGTATG
TCAGGGGCTGAAGTCCTATCGAGGAAATTGTGCATAGTAGGAATGACTTGATTTTCT
TAAGTTACCTTCATTCTCCCTGAGGTATCATGATGGCATCTTTCTCATAGCAGAGGGT
```

- Submit sequences to biological databases
- Biological databases helps in efficient manipulation of large data sets
- Provides improved search sensitivity, search efficiency
- Joining of multiple data sets
- **Databases allows the users to analyse the biological data sets**

✓ **DNA**

✓ **RNA**

✓ **Proteins**



Analysis of Nucleic acids & Protein Sequences

- **Sequence Analysis**

Process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods

- ✓ To understand its features, function, structure, or evolution
- ✓ To assign function to genes & proteins by the studying the similarities between the compared sequences

Methodologies include:

- ❖ *Sequence alignment*
- ❖ *Searches against biological databases*

- **Sequence analysis in molecular biology includes a very wide range of relevant topics:**

- The comparison of sequences in order to find similarity, infer if they are related (**homologous**)
- Identification of active sites, gene structures, reading frames etc.
- Identification of sequence differences and variations – SNP, Point mutations, identify genetic markers
- Revealing the evolution and genetic diversity of sequences and organisms
- Identification of molecular structure from sequence alone

Sequence Alignment

Relationships between these sequences are usually discovered by

- aligning them together
- assigning a score to the alignments

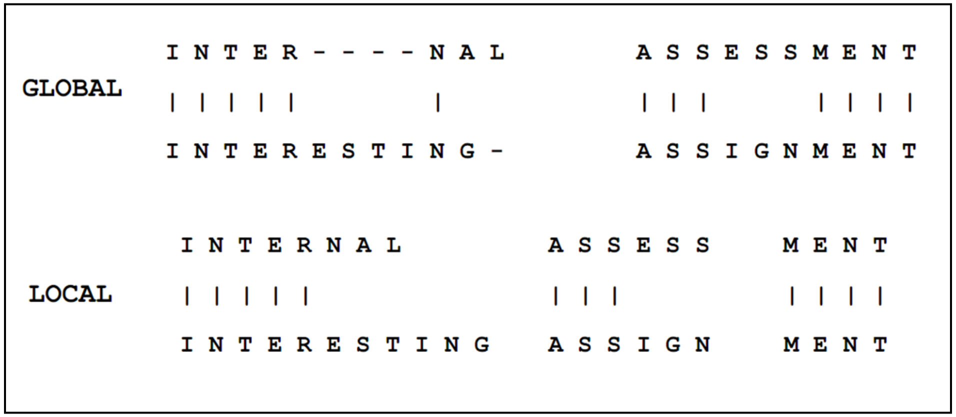
Two main types of sequence alignment:

- Pair-wise sequence alignment - compares only two sequences at a time
- Multiple sequence alignment- compares many sequences

Two important algorithms for aligning pairs of sequences :

- Needleman-Wunsch algorithm
- Smith-Waterman algorithm

- **Popular tools for sequence alignment include:**
  - ✓ Pair-wise alignment - BLAST
  - ✓ Multiple alignment - ClustalW, MUSCLE, MAFFT, T-Coffee etc.
- **Alignment methods:**
  - ✓ Local alignments - Needleman–Wunsch algorithm
  - ✓ Global alignments - Smith-Waterman algorithm



```
GLOBAL   INTER----NAL   ASSESSMENT
         |||||    |     |||   ||||
         INTERESTING-   ASSIGNMENT

LOCAL    INTERNAL   ASSESS   MENT
         |||||      |||      ||||
         INTERESTING  ASSIGN   MENT
```

Local Alignment

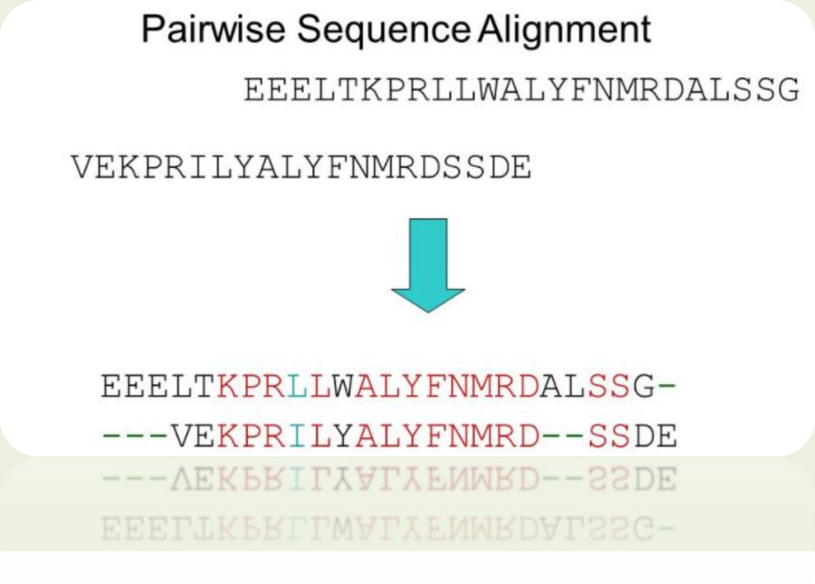
```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
```

Pair-wise alignment

- Used to find the best-matching piecewise (local or global) alignments of two query sequences
- Can only be used between two sequences at a time



Pairwise Sequence Alignment

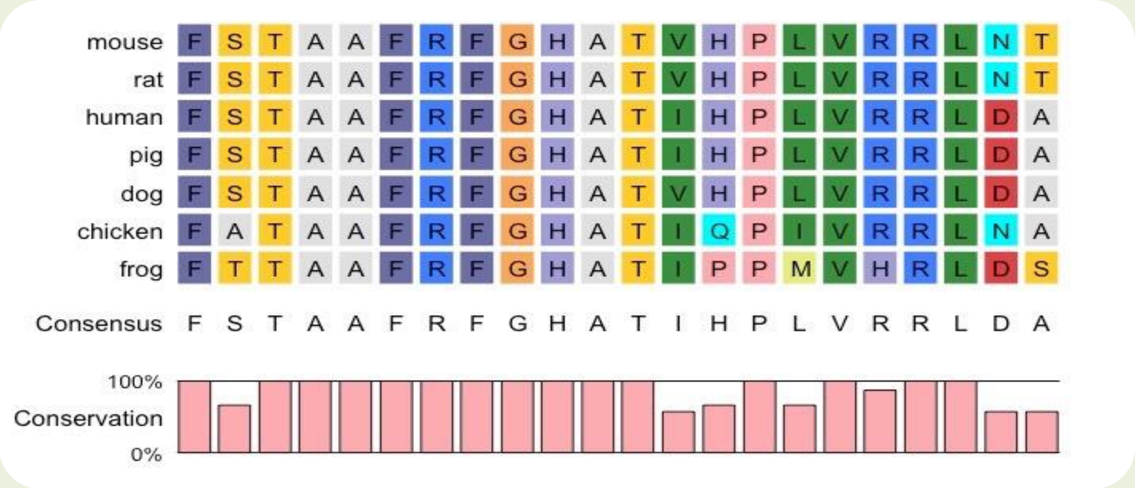
```
        EEELTKPRLLWALYFNMRDALSSG

VEKPRILYALYFNMRDSSDE

EEELTKPRLLWALYFNMRDALSSG-
---VEKPRILYALYFNMRD--SSDE
```

Multiple Sequence Alignment

- Is an extension of pairwise alignment to incorporate more than two sequences at a time
- Align all of the sequences in a given query set
- Often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related
- Alignments helps to establish evolutionary relationships by constructing phylogenetic trees



mouse	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	N	T
rat	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	N	T
human	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A
pig	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A
dog	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	D	A
chicken	F	A	T	A	A	F	R	F	G	H	A	T	I	Q	P	I	V	R	R	L	N	A
frog	F	T	T	A	A	F	R	F	G	H	A	T	I	P	P	M	V	H	R	L	D	S
Consensus	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A

Sequence Analaysis Tools

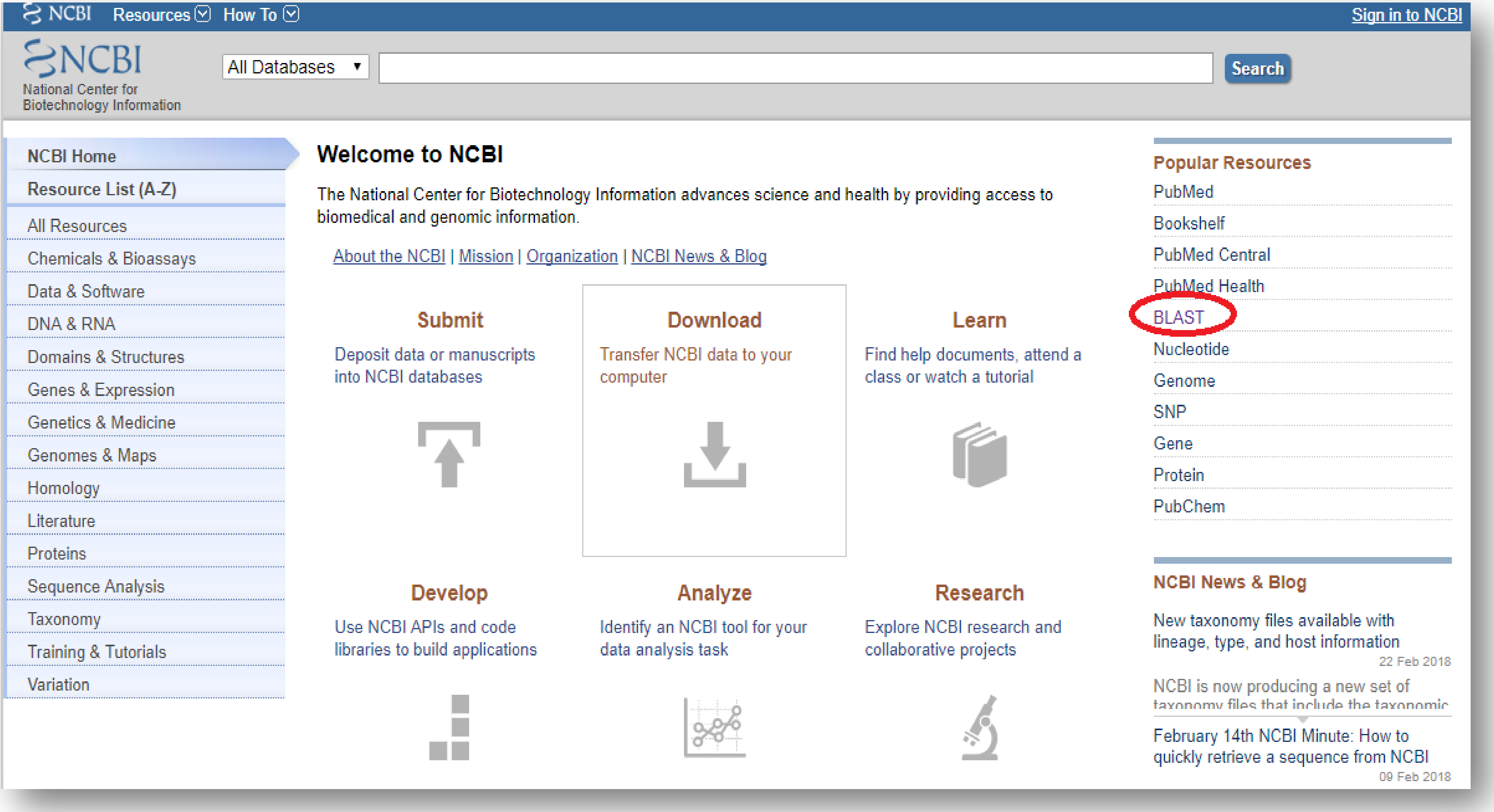
Pair-wise alignment - BLAST

- **Basic Local Alignment Search Tool (BLAST)**
- Developed by Research staff at NCBI/GenBank as a new way to perform seq. similarity search
- Available as free service over internet
- Very fast ,Accurate and sensitive database searching
- Server-NCBI

Types of BLAST Programs:

Types of BLAST	Query	Database
BLAST P	Protein	Protein
BLAST N	Nucleotide	Nucleotide
BLAST X	Nucleotide (translated)	Protein
TBLAST N	Protein	Nucleotide (translated)
TBLAST X	Nucleotide (translated)	Nucleotide (translated)

NCBI -BLAST



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 | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

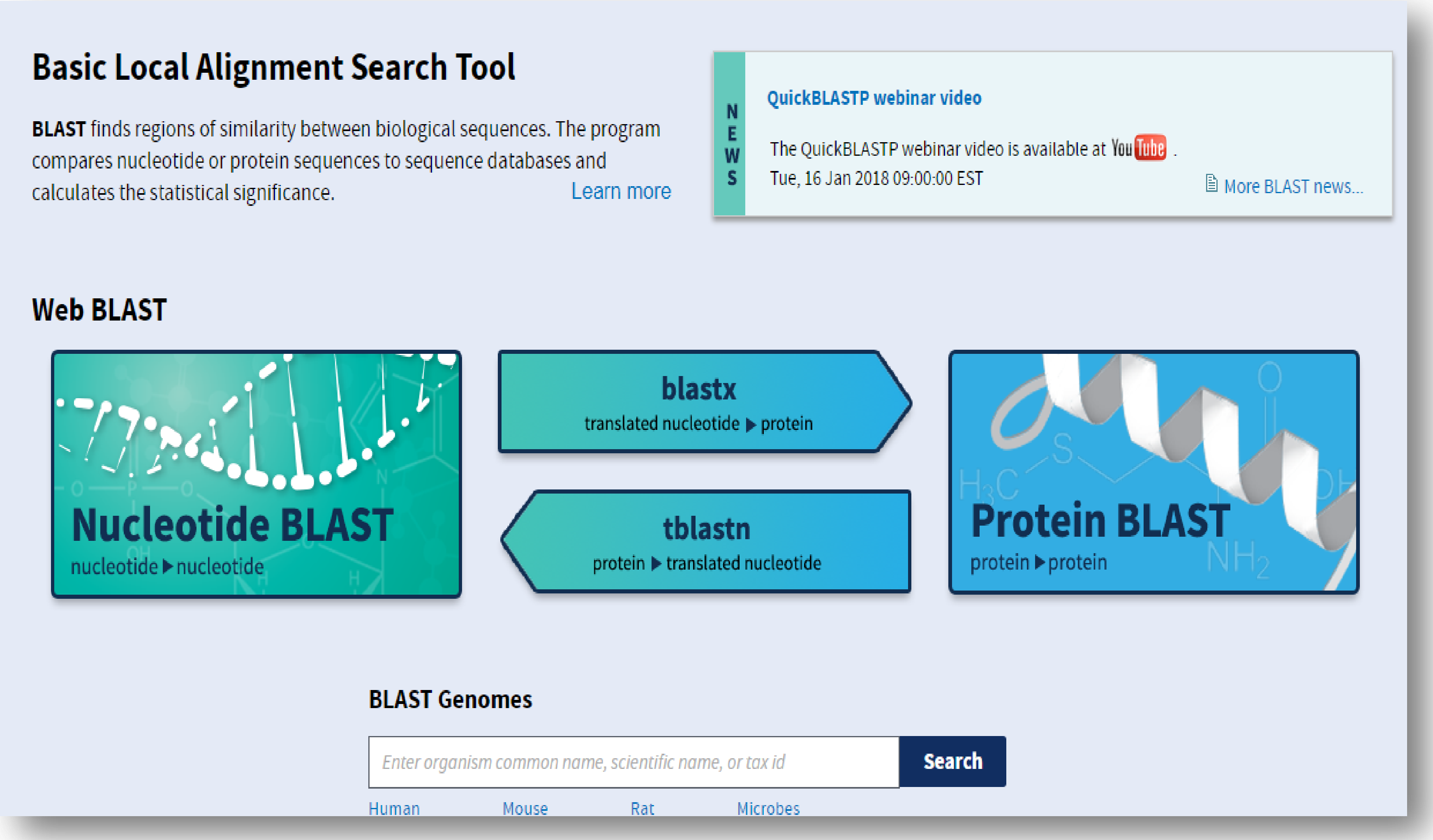
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

New taxonomy files available with lineage, type, and host information
22 Feb 2018

NCBI is now producing a new set of taxonomy files that include the taxonomic

February 14th NCBI Minute: How to quickly retrieve a sequence from NCBI
09 Feb 2018



Basic Local Alignment Search Tool

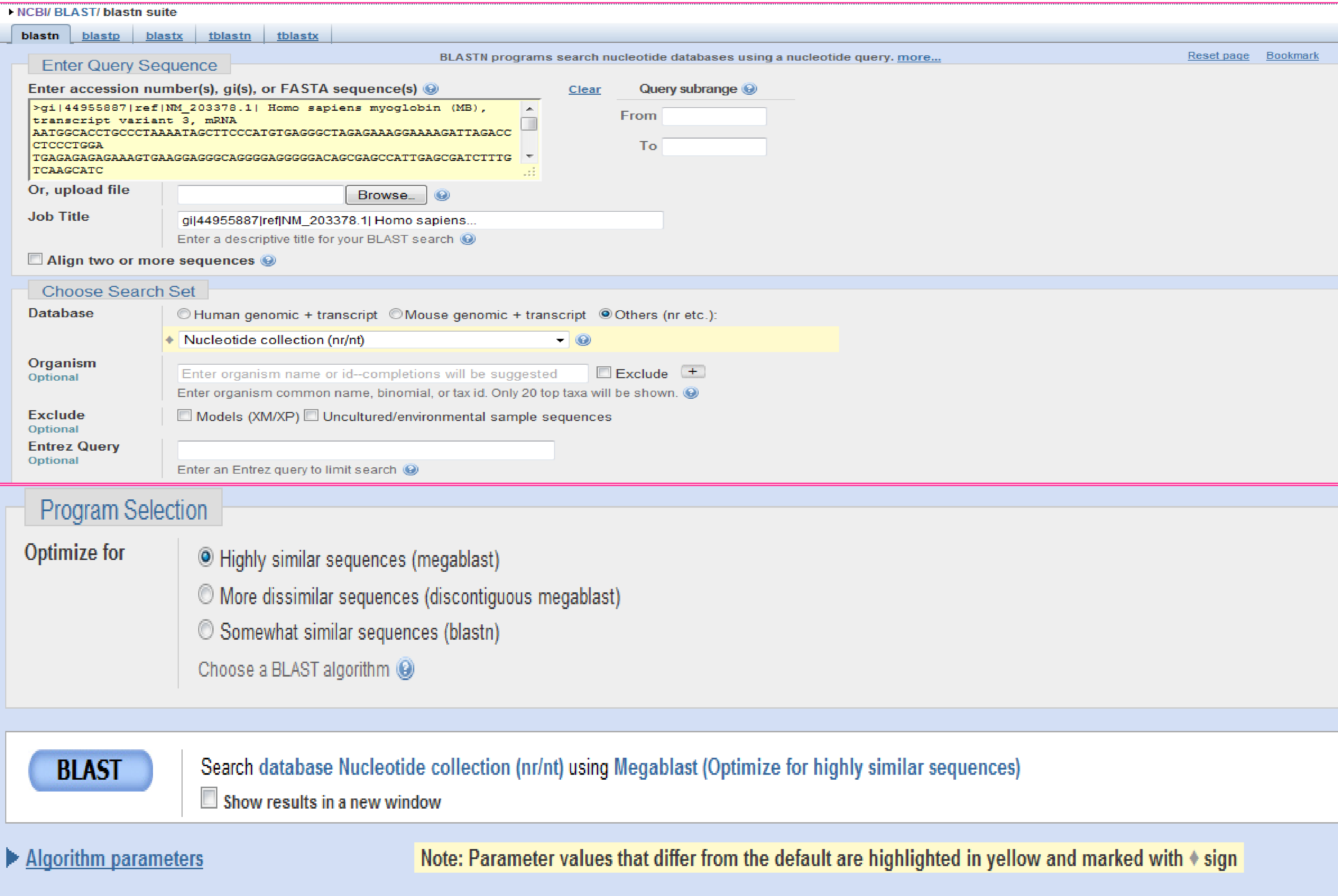
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

QuickBLASTP webinar video

The QuickBLASTP webinar video is available at YouTube .
Tue, 16 Jan 2018 09:00:00 EST

Web BLAST

BLAST Genomes



▸ NCBI/ BLAST/ blastn suite

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page   Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)   Clear

```
>gi|44955887|ref|NM_203378.1| Homo sapiens myoglobin (MB),
transcript variant 3, mRNA
AATGGCACCTGCCCTAAAATAGCTTCCCATGTGAGGGCTAGAGAAAGGAAAAGATTAGACC
CTCCCTGGA
TGAGAGAGAGAAAGTGAAGGAGGGCAGGGGAGGGGGACAGCGAGCCATTGAGCGATCTTTG
TCAAGCATC
```

Query subrange
From
To

Or, upload file   Browse...

Job Title: gi|44955887|ref|NM_203378.1| Homo sapiens...
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):
♦ Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query (Optional):
Enter an Entrez query to limit search

Program Selection

Optimize for
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

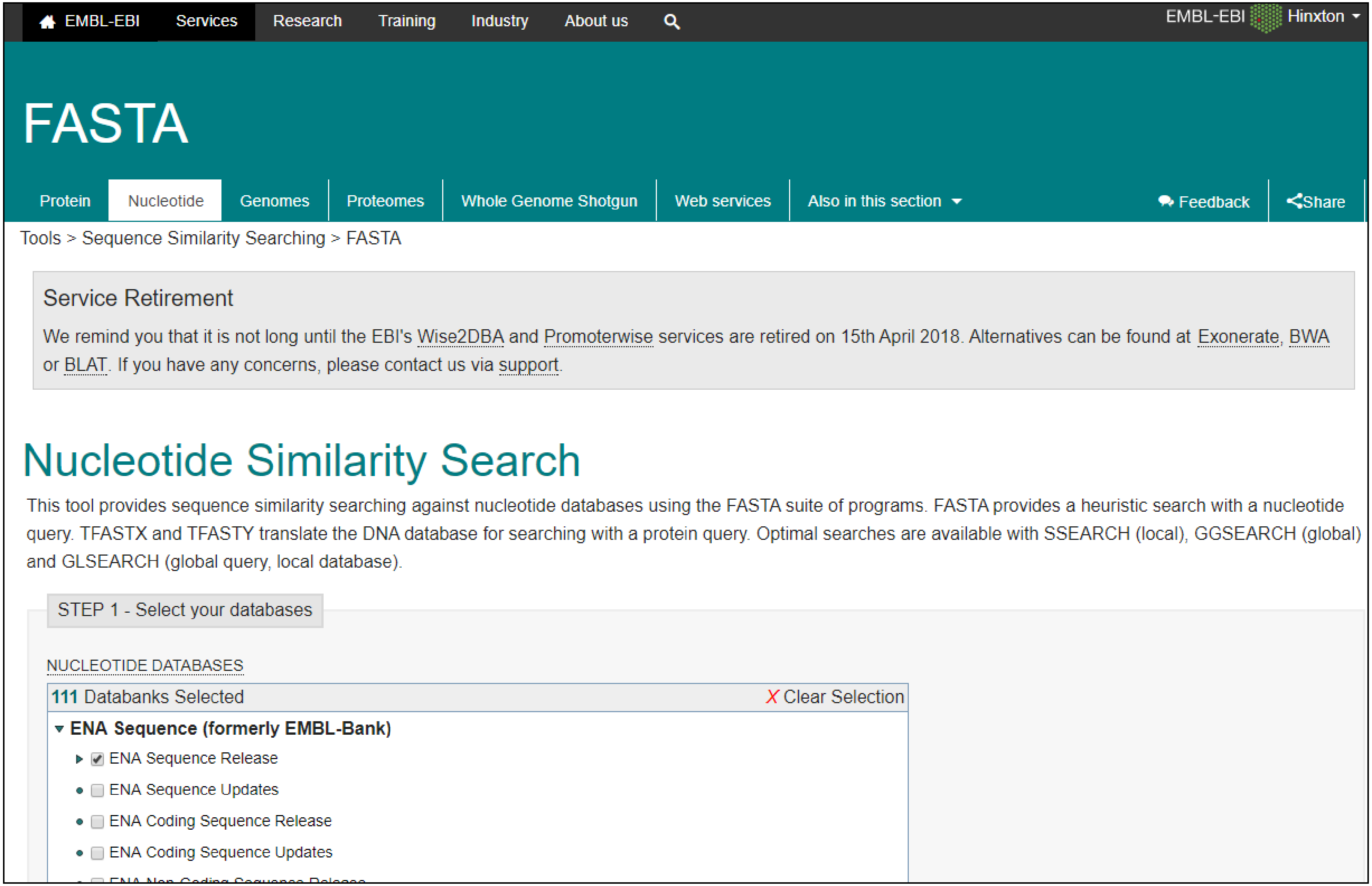
☐ Show results in a new window

▸ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

FASTA

- **DNA** & **Protein** sequence alignment software package
- Fast A → “Fast –ALL”
- Works on any Alphabets
- FAST P → Protein
- FAST N → Nucleotide



FASTA

Tools > Sequence Similarity Searching > FASTA

Service Retirement

We remind you that it is not long until the EBI's Wise2DBA and Promoterwise services are retired on 15th April 2018. Alternatives can be found at Exonerate, BWA or BLAT. If you have any concerns, please contact us via support.

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected — X Clear Selection

ENA Sequence (formerly EMBL-Bank)

- ☑ ENA Sequence Release
- ☐ ENA Sequence Updates
- ☐ ENA Coding Sequence Release
- ☐ ENA Coding Sequence Updates

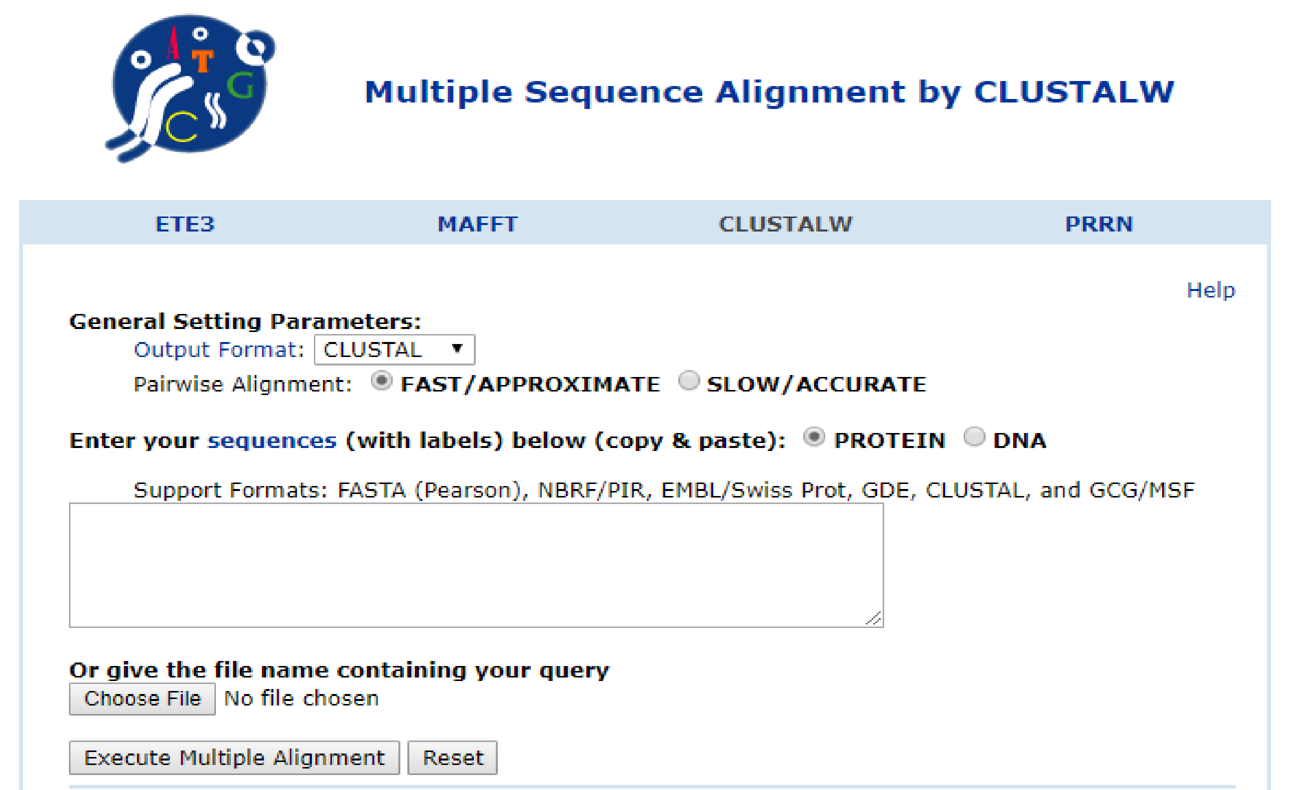
Sequence Analaysis Tools

Multiple alignment - **ClustalW**

- Study the identities, Similarities & Differences
- Study evolutionary relationship
- Identification of conserved sequence regions
- **Useful in predicting –**

❖ Function & structure of proteins

❖ Identifying new members of protein families



Multiple Sequence Alignment by CLUSTALW

General Setting Parameters:

Output Format: CLUSTAL ▾

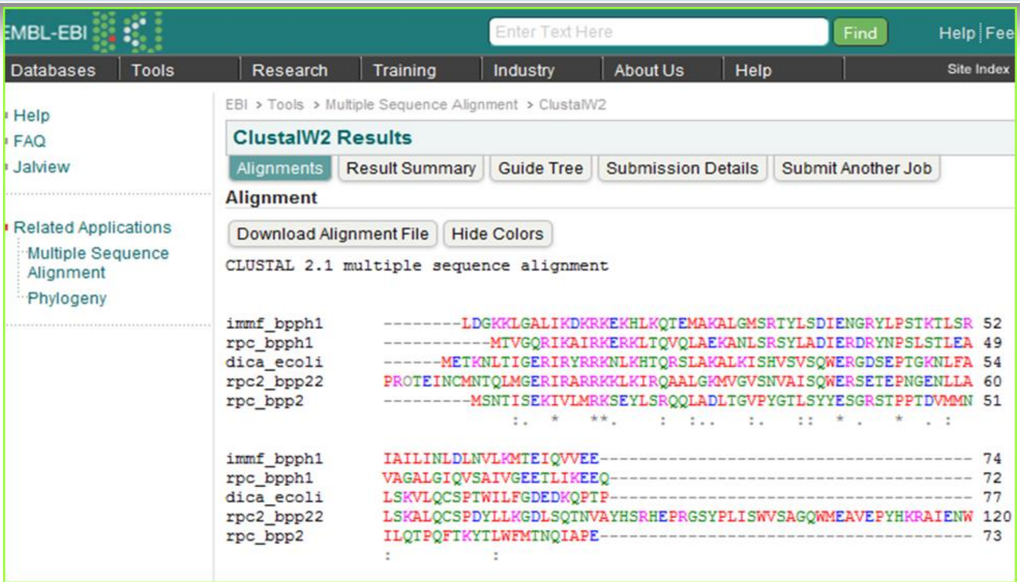
Pairwise Alignment: ● **FAST/APPROXIMATE** ○ **SLOW/ACCURATE**

Enter your sequences (with labels) below (copy & paste): ● **PROTEIN** ○ **DNA**

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose File No file chosen



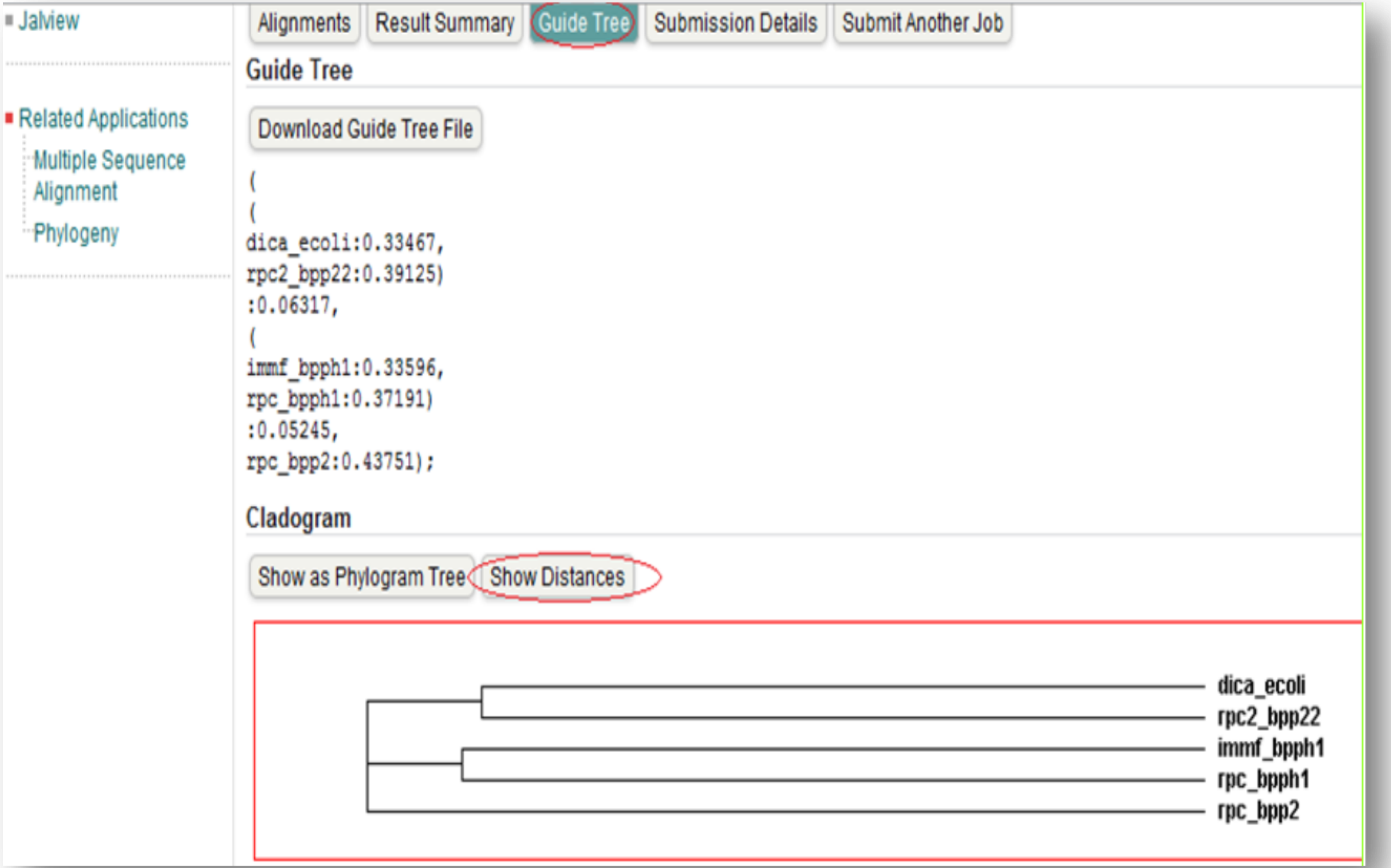
ClustalW2 Results

Alignment

```
CLUSTAL 2.1 multiple sequence alignment


immf_bpph1      --------LDGKKLGALIKDKRKEKHLKQTEMAKALGMSRTYLSDIENGRYLPSTKTLSR 52
rpc_bpph1       -----------MTVGQRIKAIRKERKLTQVQLAEKANLSRSYLADIERDRYNPSLSTLEA 49
dica_ecoli      ------METKNLTIGERIRYRRKNLKHTQRSLAKALKISHVSVSQWERGDSEPTGKNLFA 54
rpc2_bpp22      PROTEINCMNTQLMGERIRARRKKLKIRQAALGKMVGVSNVAISQWERSETEPNGENLLA 60
rpc_bpp2        ---------MSNTISEKIVLMRKSEYLSRQQLADLTGVPYGTLSYYESGRSTPPTDVMMN 51
                             :.  *   **.    :  :..   :.   ::  * .   *  . :

immf_bpph1      IAILINLDLNVLKMTEIQVVEE-------------------------------------- 74
rpc_bpph1       VAGALGIQVSAIVGEETLIKEEQ------------------------------------- 72
dica_ecoli      LSKVLQCSPTWILFGDEDKQPTP------------------------------------- 77
rpc2_bpp22      LSKALQCSPDYLLKGDLSQTNVAYHSRHEPRGSYPLISWVSAGQWMEAVEPYHKRAIENW 120
rpc_bpp2        ILQTPQFTKYTLWFMTNQIAPE-------------------------------------- 73
                :          :
```



Guide Tree

```
(
(
dica_ecoli:0.33467,
rpc2_bpp22:0.39125)
:0.06317,
(
immf_bpph1:0.33596,
rpc_bpph1:0.37191)
:0.05245,
rpc_bpp2:0.43751);
```

Cladogram

Molecular Modelling

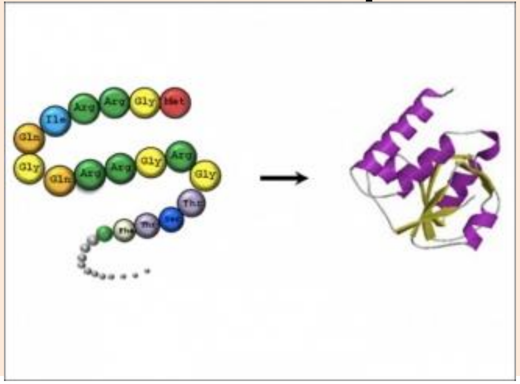
- Includes all methods, theoretical & computational, used to model or mimic the behaviour of molecules
- Helps to study molecular systems ranging from small chemical systems to large biological molecules

The methods are used in the fields of :

- *Computational chemistry*
- *Drug design*
- *Computational biology*
- *Materials science*

Structure Analysis of Proteins

- Researchers predict the 3D structure using protein or molecular modeling
- Experimentally determined protein structures (templates) are used
- To predict the structure of another protein that has a similar amino acid sequence (target)



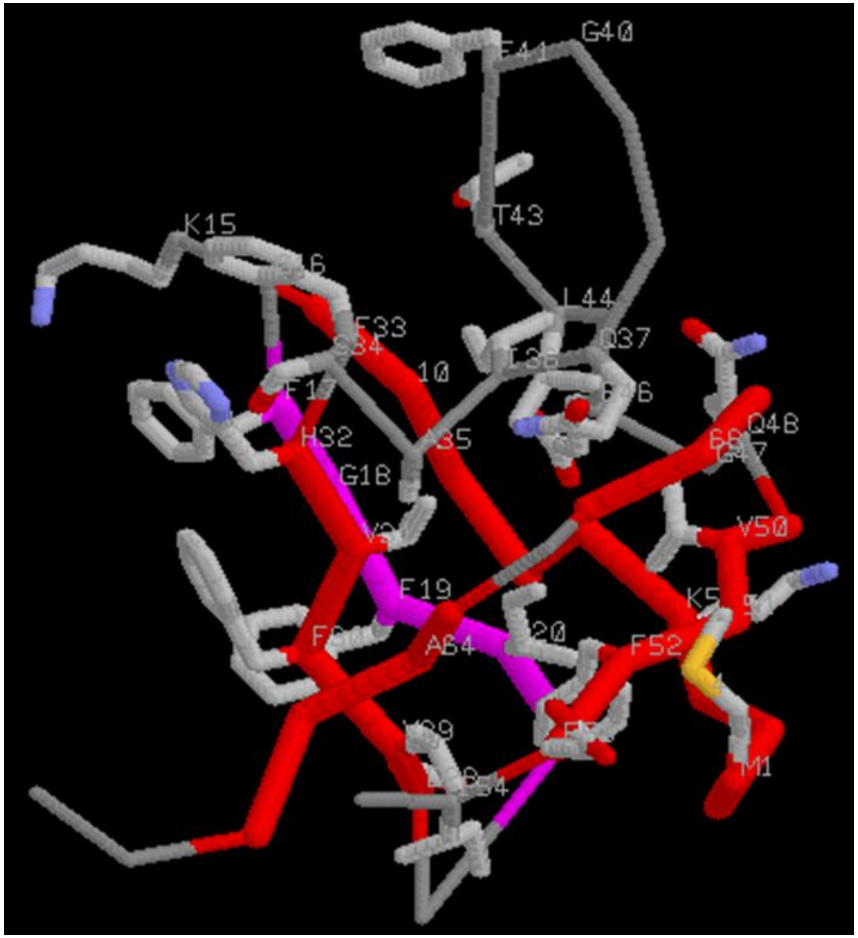
Advantages in Protein Modeling

- **Examining a protein in 3D allows for :**

✓ greater understanding of protein functions

✓ providing a **visual understanding** that cannot always be conveyed through still photographs or descriptions

Example of 3D-Protein Model



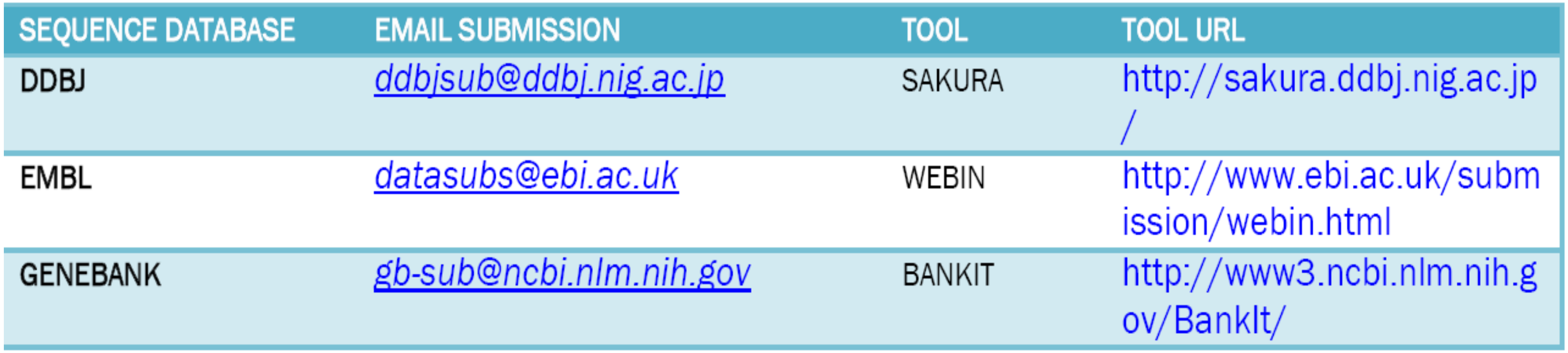
Impact of Bioinformatics in Biology/Biotechnology

- Biological research is the most fundamental research to understand complete mechanism of living system
- The advancements in technologies helps in providing regular updates and contribution to make human life better and better.

✓ Reduced the time consuming experimental procedure

✓ Software development – Bioinformatians & Computational Biologists

✓ Submitting biological sequences to databases



SEQUENCE DATABASE	EMAIL SUBMISSION	TOOL	TOOL URL
DDBJ	*ddbjsub@ddbj.nig.ac.jp*	SAKURA	http://sakura.ddbj.nig.ac.jp/
EMBL	*datasubs@ebi.ac.uk*	WEBIN	http://www.ebi.ac.uk/submission/webin.html
GENEBANK	*gb-sub@ncbi.nlm.nih.gov*	BANKIT	http://www3.ncbi.nlm.nih.gov/BankIt/

Role of Bioinformatics in Biotechnology

- **Genomics**

✓ The study of genes and their expression

✓ Generates vast amount of data from gene sequences, their interrelations & functions

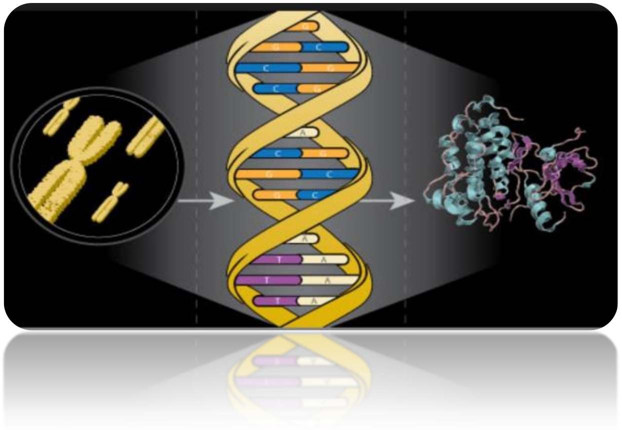
✓ Understand structural genomics, functional genomics and nutritional genomics

- **Proteomics**

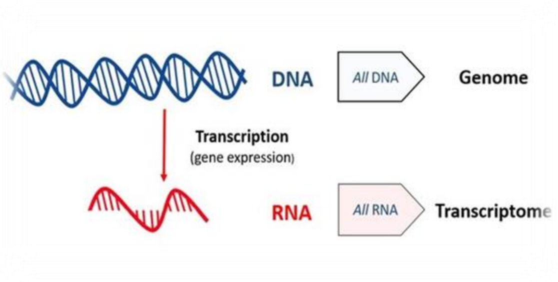
✓ Study of protein structure, function &interactions produced by a particular cell, tissue, or organism

✓ Deals with techniques of genetics, biochemistry and molecular biology

✓ Study protein-protein interactions, protein profiles, protein activity pattern and organelles compositions



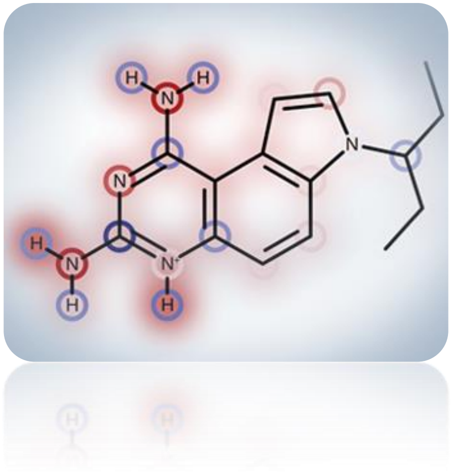
- **Transcriptomics**



✓ Study of sets of all messenger RNA molecules in the cell
✓ Also be called as Expression Profiling- DNA Micro array
✓ RNA sequencing –NGS
✓ Used to analyse the continuously changing cellular transcriptome

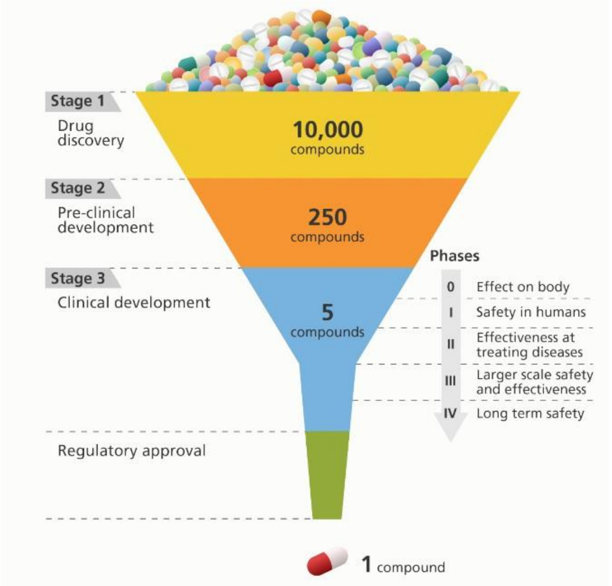
- **Cheminformatics**

✓ Deals with focuses on storing, indexing, searching, retrieving, and applying information about chemical compounds
✓ involves organization of chemical data in a logical form - to facilitate the retrieval of chemical properties, structures & their relationships
✓ Helps to identify and structurally modify a natural product



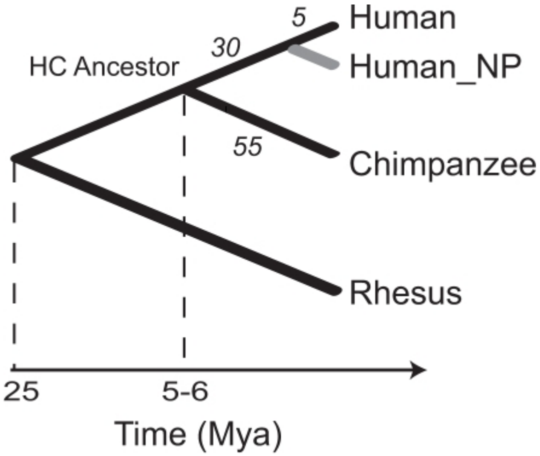
• Drug Discovery

- ✓ Increasingly important role in drug discovery, drug assessment & drug development
- ✓ Computer-aided drug design (CADD)- generate more & more drugs in a short period of time with low risk
- ✓ wide range of drug-related databases & softwares - for various purposes related to drug designing & development process



• Evolutionary Studies

- ✓ Phylogenetics - evolutionary relationship among individuals or group of organisms
- ✓ phylogenetic trees are constructed based on the sequence alignment using various methods



- **Crop Improvement**
  - ✓ Innovations in omics based research improve the plant based research
  - ✓ Understand molecular system of the plant which are used to improve the plant productivity
  - ✓ comparative genomics helps in understanding the genes & their functions, biological properties of each species

- **Biodefense**
  - ✓ Biosecurity of organisms - subjected to biological threats or infectious diseases (Biowar)
  - ✓ Bioinformatics- limited impact on forensic & intelligence operations
  - ✓ Need of more algorithms in bioinformatics for biodefense

- **Bioenergy/Biofuels**
  - ✓ contributing to the growing global demand for alternative sources of renewable energy
  - ✓ progress in algal genomics + ‘omics’ approach - Metabolic pathway & genes – genetically engineered micro algal strains