Bioinformatics Databases

MV 2017

Data in Bioinformatics

- **DNA**- Sequences of nucleotides (ATGC) that contain information in the form of triplet codons having specific reading frames and built-in control segments

- **RNA** sequences (AUGC) mRNA, tRNA, hnRNA

- **Protein sequences**- Strings of Amino-acid sequences (*e. g.,* Aspartate, Glycine, Histidine, Isoleucine, Leucine, Methionine, Serine, Threonine, Valine, Phenyl alanine, Tyrosine)

- **Structure data** (Protein structure-primary secondary, tertiary, Quaternary, 3D views)

- Images of 2D Gel electrophoresis

Bioinformatics Databases

- Databases are convenient system to properly store, search and retrieve any type of data

- Databases are different types based on nature of information and manner (complexity) of data storage

Types of databases

Based on nature of information db are divided into

- 1. Generalized db: DNA, Protein (*e. g.,* NCBI)
  - a. Sequence db: nucleotides or amino acids
  - b. Structure db: structure of macromolecules
- 2. Specialized db: Expressed Sequence Tags (EST), Single Nucleotide Polymorphisms (SNP)

Based on the manner of data storage, db are divided into

1. Primary or abbreviated db: in original form, taken as such from the source. Eg: GenBank, Swiss-Prot

2. Secondary db: value added db with derived information from primary db

3. Composite db: combined primary db

Redundant and Non-redundant db: more than one copy of each sequence

Boutique db: species specific sequence data

- Db entries composed of
  - Core data: original sequence
  - Supplementary data or annotation (source, author, date, method used etc)

- Sequence formats
  - PIR (Protein Information Resource)/NBRF(National Biomedical Res. Foundation) - >P, >N
  - FASTA (Fast Alignment) - >
  - GDE (Genetic Data Environment) - %

Primary Databases

- In original form, taken from the source
- Original submission by researcher
- Contents controlled by the submitter
- Data explosion in 1980s - so started many repositories
1. Nucleic acid sequence db
2. Protein sequence db
3. Metabolite db

Secondary databases

- Derivative db

- Result of analyses of sequences in the primary db

- Secondary db built up from primary db

- Secondary db analyzed in a variety of ways and contain different information in different formats

- Contents of secondary db controlled by a third party

- Eg: Prosite, Prints, Blocks

Nucleic acid sequence databases

- Collection of nucleotide sequences

- Organize and distribute nucleotide sequences from all available source

- In the form of a text file

- Can read by humans and computer

- Many dbs are assembled from several publications, so overlapping fragments of complete sequence

- First sequence - Yeast t-RNA with 77 bases in 1964

NCBI

- National Centre for Biotechnology Information
- Established on November 4, 1988 as part of the National Library of Medicine (NLM) at the National Institute of Health (NIH), USA
- Headquarters in Bethesda, Maryland
- Legislation sponsored by Senator Claude Pepper

Services

- Pubmed
- Genbank
- BLAST
- Entrez

GenBank

- GenBank [®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

- GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC) , which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

International Nucleotide Sequence Database Collaboration (INSDC)

- INSDC consist of

- 1. EMBL

- 2. DDBJ

- 3. GenBank

- Daily exchange of data

- The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted.

What is in it?

- Annotated nucleotide sequences, including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters

- More than 100,000 organisms

- Aminoacid translations (CDS)

EMBL

- The **European Molecular Biology Laboratory** (**EMBL**) is a molecular biology research institution supported by 25 member states, four prospect and two associate member states. EMBL was constituted in 1974 and is an intergovernmental organisation funded by public research money from its member states. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology. EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors.

Stations

- The Laboratory operates from six sites: the main laboratory in Heidelberg, and outstations in Hinxton (the European Bioinformatics Institute (EBI), in England), Grenoble (France), Hamburg (Germany), Monterotondo (near Rome) and Barcelona (Spain).

www.ebi.ac.uk

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
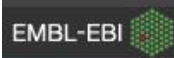
EMBL-EBI

Services | Research | Training | About us

The European Bioinformatics Institute

EMBL-EBI

Other EMBL locations >

The home for big data in biology

At EMBL-EBI, we use bioinformatics — the science of storing, sharing and analysing biological data — to help people everywhere understand how living systems work, and what makes them change.

Find a gene, protein or chemical:

1

Examples: blast, keratin, bfl1, Janet Thornton ...

Explore EMBL-EBI

Services > | Research > | Training > | Industry > | ELIXIR >

European Molecular Biology Laboratory (EMBL)

- From European Bioinformatics Institute (EBI), UK
- Collect and assemble data from
  - -Direct author submission
  - -Genome sequencing groups
  - -Patent application
  - -Literature
- Goal - integrate nucleotide sequence data and annotation into the wealth of bioinformatics resources
- By cross reference and Sequence Retrieval System (SRS) data can be viewed in 200 local stations
- 2494 completed genomes

EMBL

- The roots of the EMBL-EBI lie in the EMBL Nucleotide Sequence Data Library (now known as EMBL-Bank), which was established in 1980 at the EMBL laboratories in Heidelberg, Germany and was the world's first nucleotide sequence database.

- The original goal was to establish a central computer database of DNA sequences, to supplement sequences submitted to journals.

- The EMBL-EBI hosts a number of publicly open, free to use life science resources, including biomedical databases, analysis tools and bio-ontologies. These include:
- ArrayExpress - archive of gene expression experiments
- BioModels Database - a database of computational models relevant to the life sciences
- BioStudies - a database that serves as a generic data archive at EMBL-EBI for biomolecular datasets
- Chemical Entities of Biological Interest (ChEBI) - database and ontology of molecular entities
- European Nucleotide Archive (ENA) - resource of nucleotide sequencing information
- Ensembl project - genome databases for vertebrates and other eukaryotic species (joint with Wellcome Trust Sanger Institute)
- Europe PubMed Central - database offering free access to collection of biomedical research literature

DNA Data Bank of Japan

- Currently, DDBJ Center is in operation at the National Institute of Genetics (NIG) in Mishima, Japan with endorsement of MEXT; Japanese Ministry of Education, Culture, Sports, Science and Technology.
- DDBJ Center is reviewed and advised by its own advisory board, DNA Database Advisory Committee (an outside committee of NIG), and also by the advisory board to INSDC, International Advisory Committee.
- Started in 1986

- It is located at the **National Institute of Genetics** (NIG) in the Shizuoka prefecture of Japan. It is also a member of the INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis.

- These three databanks contain the same data at any given time.

Protein sequence databases

- SWISSPROT, PIR

- UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB).

- It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.

- Since 2002, it is maintained by the UniProt consortium and is accessible via the UniProt website.

UniProtKB/Swiss-Prot

- UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB).
  It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.

- Since 2002, it is maintained by the UniProt consortium and is accessible via the UniProt website.

Swiss-Prot

- Established in 1986 by Dept. of Biochemistry, University of Geneva
- Maintenance by Swiss Institute of Bioinformatics (SIB) and EMBL

- Database composed of 2 parts
    1. Core data - sequence reference and taxonomic details
    2. Annotation - sequence variants, functions, 2º & 3º structures

- Provide high level annotation including functions of the protein
- Maintain high quality and structure - first choice for most research purpose

- Swiss-Prot is supplemented by TrEMBL in 1996 - translated EMBL

- TrEMBL has 2 sections
1. SP-TrEMBL - data included in the Swiss-Prot from EMBL
2. REM-TrEMBL - data which are not included in the Swiss-Prot

- A well-defined manual curation process is essential to ensure that all manually annotated entries are handled in a consistent manner. This process consists of 6 major mandatory steps: (1) sequence curation, (2) sequence analysis, (3) literature curation, (4) family-based curation, (5) evidence attribution, (6) quality assurance and integration of completed entries. Curation is performed by expert biologists using a range of tools that have been iteratively developed in close collaboration with curators.

Protein Sequence Databases

1. **PIR** - **Protein Information Resource**

- Established in 1984 by National Biomedical Research Foundation (NBRF), Washington DC

- Aim - identification and interpretation of protein sequence information

- Investigating evolutionary relationship among proteins

- Help to do search and similarity analysis

- Provide integrated environment for sequence analysis between 3 units

PIR - Protein Information Resource

- Established in 1984 by National Biomedical Research Foundation (NBRF), Washington DC

- Aim - identification and interpretation of protein sequence information

- Investigating evolutionary relationship among proteins

- Help to do search and similarity analysis

- Provide integrated environment for sequence analysis between 3 units

PIR is composed of 3 databases:

1. PSD - protein sequence database

2. NREF - Non-redundant reference database

3. iProClass - provides structural and functional features of proteins

- PIR database split into 4 sections - differ in terms of quality of data and levels of annotation provided
    1. fully classified and annotated entries
    2. preliminary entries, not thoroughly reviewed
    3. unverified entries, not reviewed
    4. genetically engineered sequences

Structure databases

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display

Troponin complex

Tropomyosin

G-actin
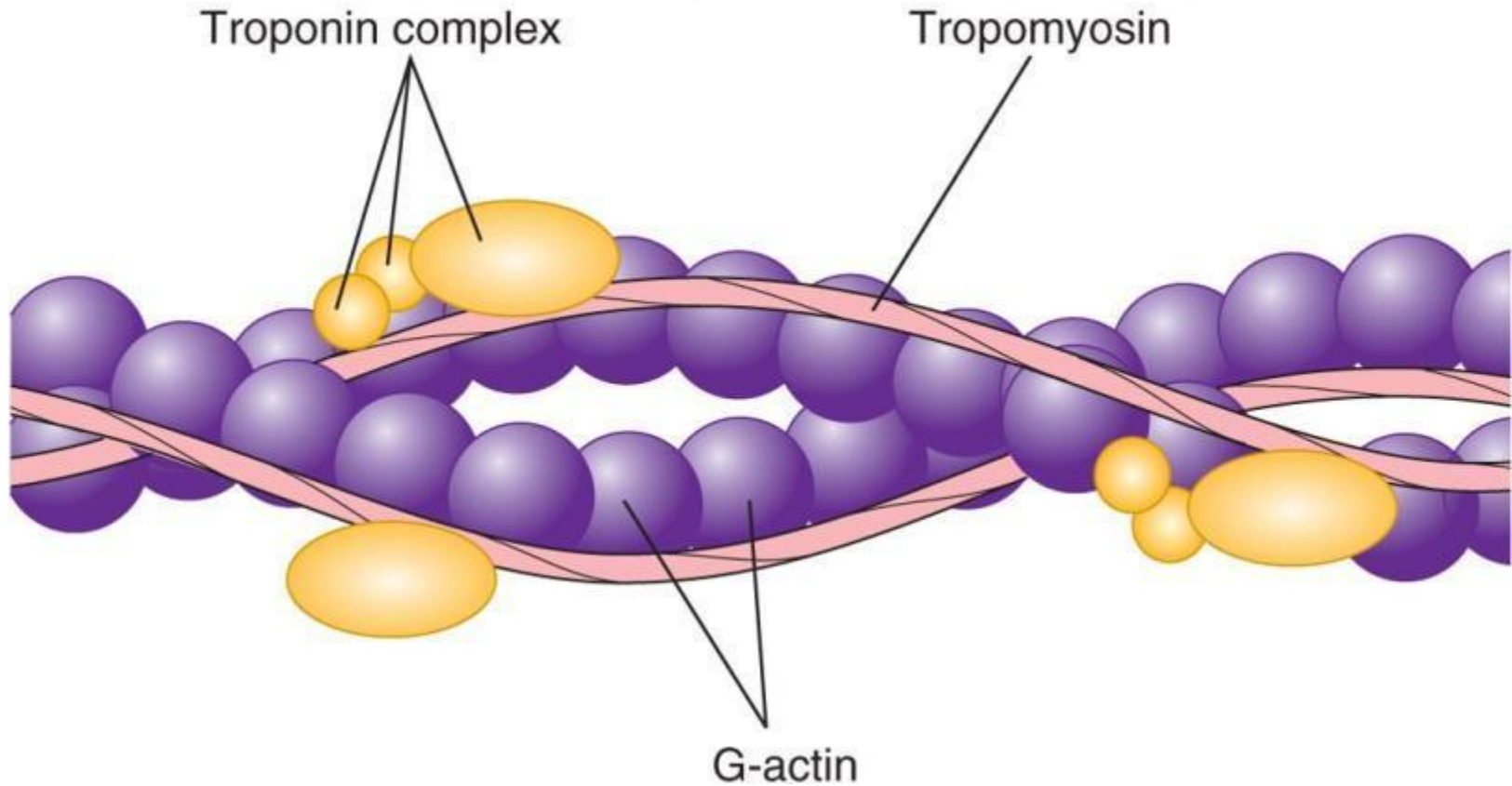
PDB

- # **Protein Data Bank**

- The **Protein Data Bank** (**PDB**) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy

www.rcsb.org/

- The data is freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB).

- The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

- The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB.

- Molecular graphics display, the Brookhaven RAster Display (BRAD), is used to visualize protein structures in 3-D.
- The file format initially used by the PDB was called the PDB file format.
- PDB was initiated in 1968 with the help of BRAD visualization of protein structure and X ray crystallographic studies of proteins
- In October 1998, the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB).

- In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe), RCSB (USA), and PDBj (Japan).

- Each of the three members of wwPDB can act as deposition, data processing and distribution centers for PDB data.

- The data processing refers to the fact that wwPDB staff review and annotate each submitted entry

INSULIN
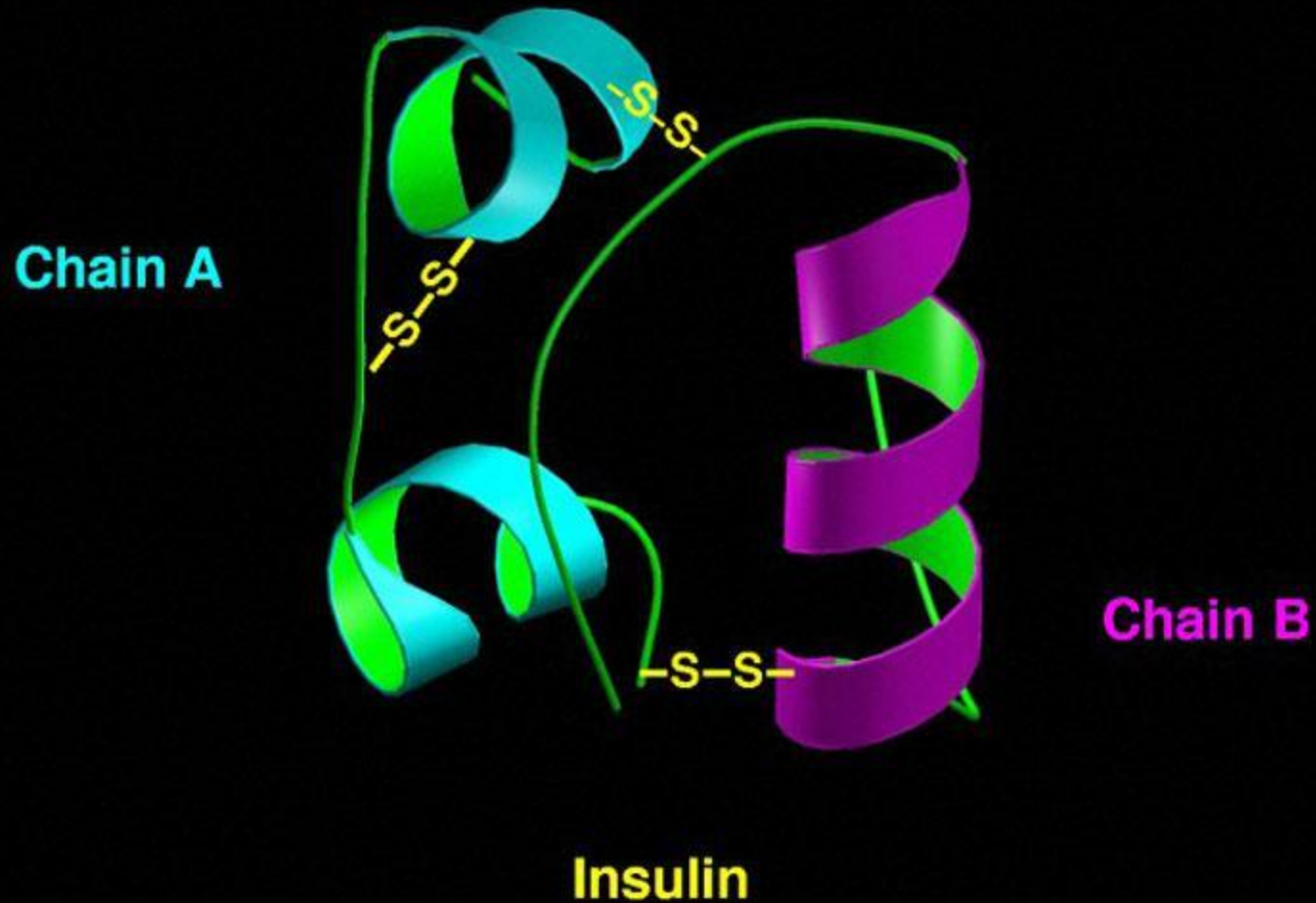
NDB
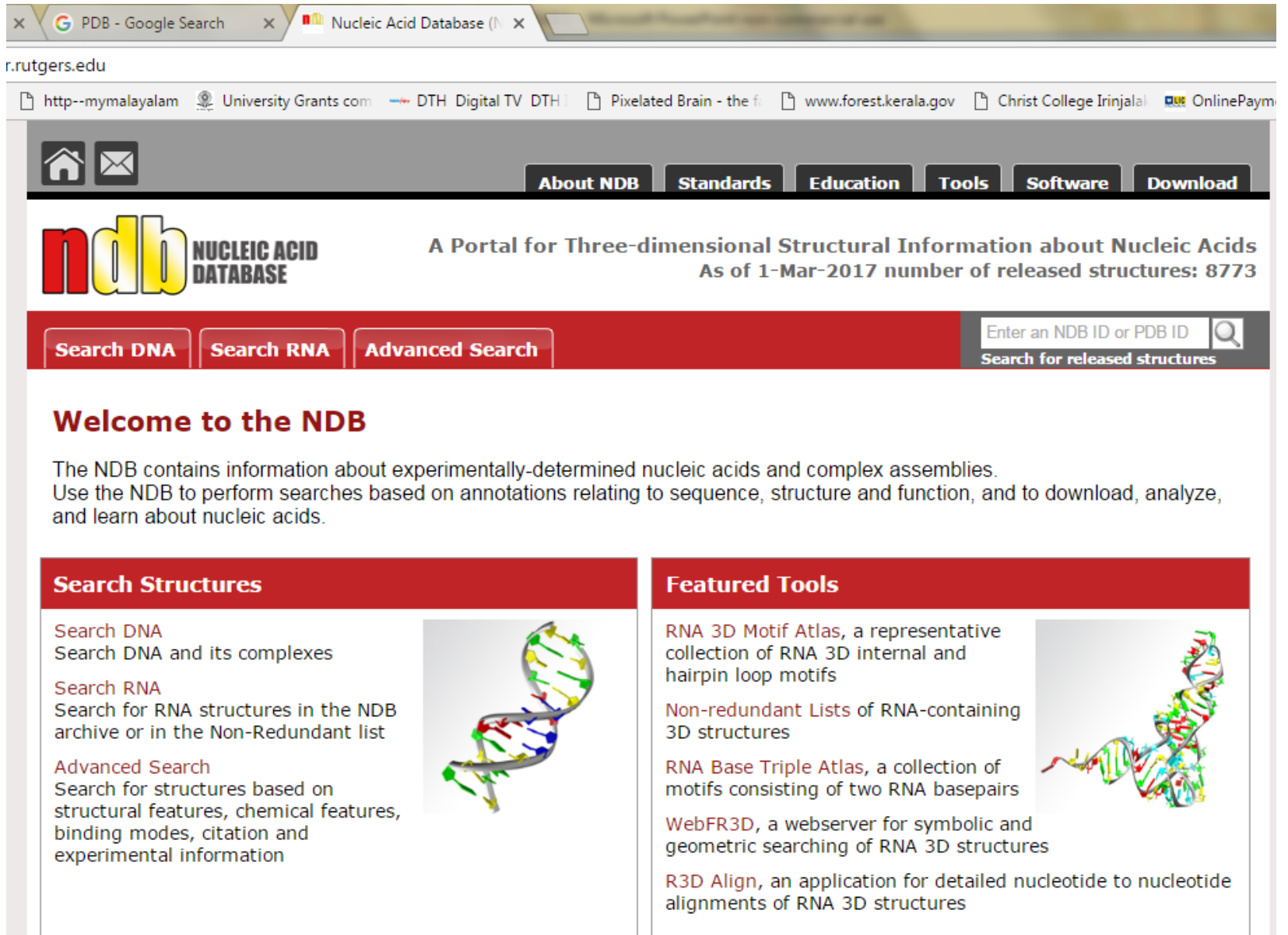
- The Nucleic Acid Database (NDB; Berman et al., 1992) was established in 1991 as a resource for specialists in the field of nucleic acid structure. Over the years, the NDB has developed generalized software for processing, archiving, querying and distributing structural data for nucleic acid-containing structures. The core of the NDB has been its relational database of nucleic acid-containing crystal structures.

- It allows researchers to perform comparative analyses of nucleic acid-containing structures selected from the NDB

- Structures available in the NDB include RNA and DNA oligonucleotides with two or more bases either alone or complexed with ligands, natural nucleic acids such as tRNA and protein±nucleic acid complexes. The archive stores both primary and derived information about the structures

- The primary data include the crystallographic coordinate data, structure factors and information about the experiments used to determine the structures, such as crystallization information, data-collection and refinement statistics.

OMIM

- Online Mendelian Inheritance in Man
- A comprehensive, authoritative and timely compendium of human genes and genetic phenotypes
- The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes.

- Initiated in the early 1960s by Dr. Victor A. Mc Kusick as a catalogue of Mendelian traits and disorders, entitled Mendelian Inheritance in Man
- 1995 internet version