Introduction to Bioinformatics

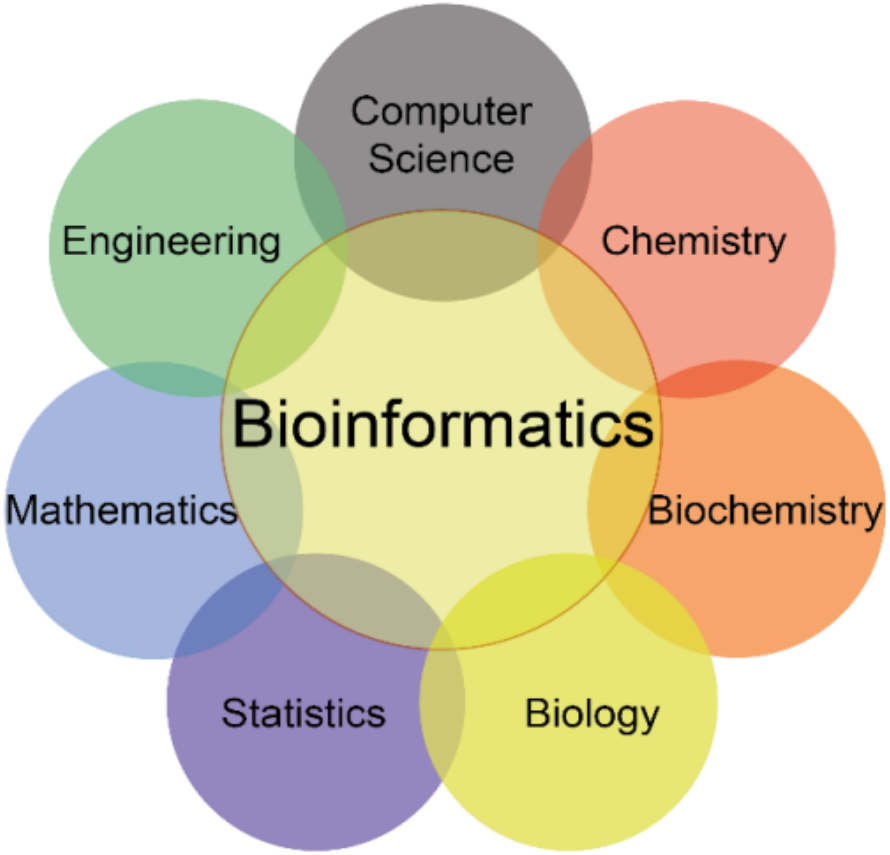
For Biochemistry Sem IV

Dr. Kusum Yadav
Assistant Professor
Department of Biochemistry
University of Lucknow, Lucknow
Mob-9452490044 Email-anukusum@gmail.com

Bioinformatics

➢ ***Bioinformatics is a branch of science that integrates computer science, mathematics and statistics, chemistry and engineering for analysis, exploration, integration and exploitation of biological sciences data, in Research and Development.***

➢ ***Bioinformatics deals with storage, retrieval, analysis and interpretation of biological data using computer based software and tools.***



History of Bioinformatics

- Bioinformatics emerged in mid 1990s.
- From 1965-78 Margaret O. Dayhoff established first database of protein sequences, published annually as series of volume entitled “Atlas of protein sequence and structure”.
- During 1977 DNA sequences began to accumulate slowly in literature and it became more common to predict protein sequences by translating sequenced genes than by direct sequencing of proteins.
- Thus number of uncharacterised proteins began to increase.
- In 1980, there were enough DNA sequences to justify the establishment of the first nucleotide sequence database, GenBank at National Centre for Biotechnology Information (NCBI), USA. NCBI served as primary databank provider for information.

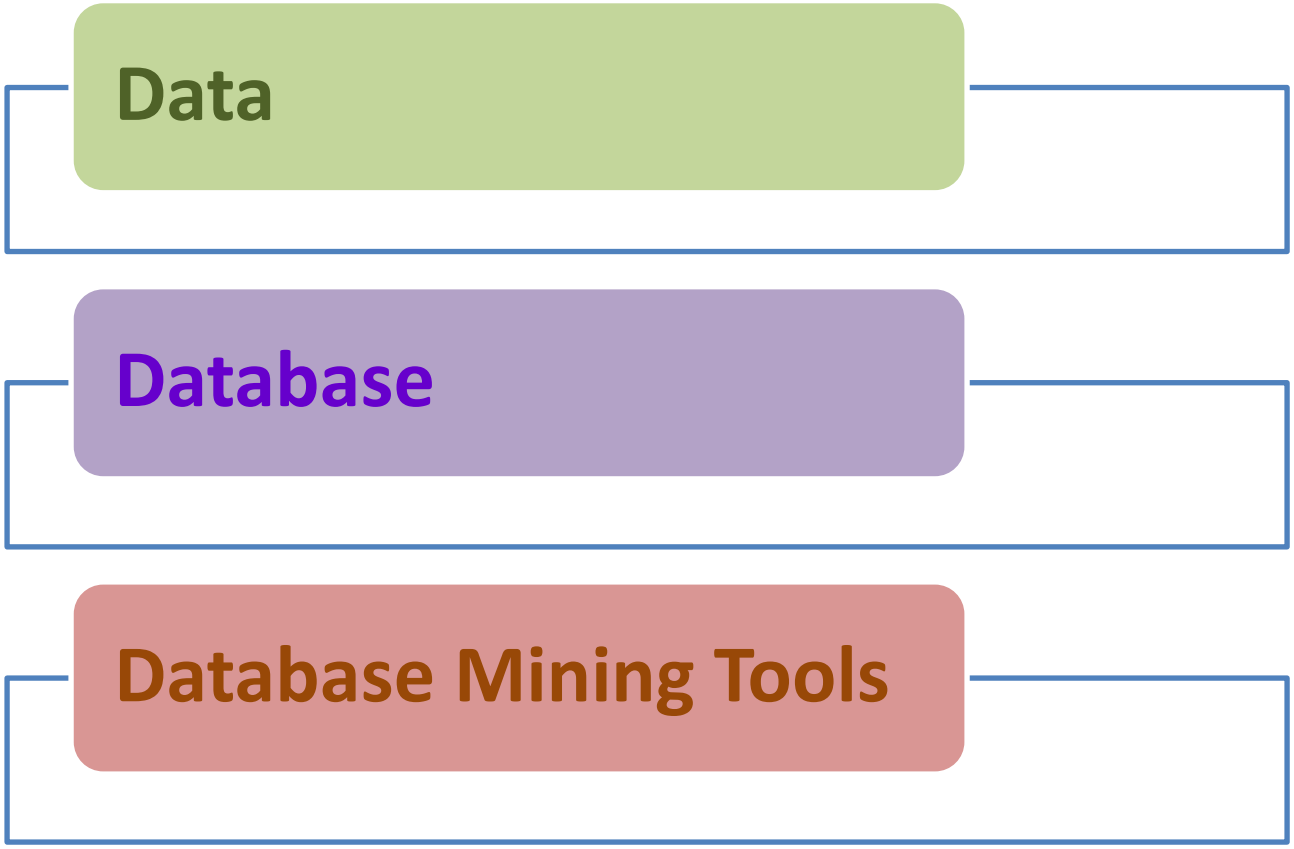
History of Bioinformatics (contd..)

- The European Molecular Biology Laboratory (EMBL) established at European Bioinformatics Institute (EBI) in 1980. The aim of this data library was to collect, organize and distribute nucleotide sequence data and related information.
- In 1986 DNA Data Bank was established by GemonNet, Japan.
- In 1984, the National Biomedical Research Foundation (NBRF) established the protein information Resource (PIR).
- All these data banks operate in close collaboration and regularly exchange data.
- Management and analysis of the rapidly accumulating sequence data required new computer software and statistical tools.
- This attracted scientists from computer science and mathematics to the fast emerging field of bioinformatics.

Objectives of Bioinformatics

1. **Development of new algorithms and statistics** for assessing the relationships among large sets of biological data.
2. **Application of these tools** for the analysis and interpretation of the various biological data.
3. **Development of database** for an efficient storage, access and management of the large body of various biological information.

Components of Bioinformatics



Data

- **Nucleic Acid Sequences**
  - Raw DNA Sequences
  - Genomic sequence tags (GSTs)
  - cDNA sequences
  - Expressed sequence tags (ESTs)
  - Organellar DNA sequences
  - RNA Sequences
- **Protein sequences**
- **Protein structures**
- **Metabolic pathways**
- **Gel pictures**
- **Literature**

Databases

A database is a vast collection of data pertaining to a specific topic e.g. nucleotide sequence, protein sequence etc., in an electronic environment.

- **They are heart of bioinformatics.**
- **Computerized storehouse of data (records).**
- **Allows extraction of specified records.**
- **Allows adding, changing, removing, and merging of records.**
- **Uses standardized formats.**

Databases: Types

- Sequence Databases
- Structural Databases
- Enzyme Databases
- Micro-array Databases
- Clinical Database
- Pathway Databases
- Chemical Databases
- Integrated Databases
- Bibliographic Databases

Nucleotide Sequence Databases

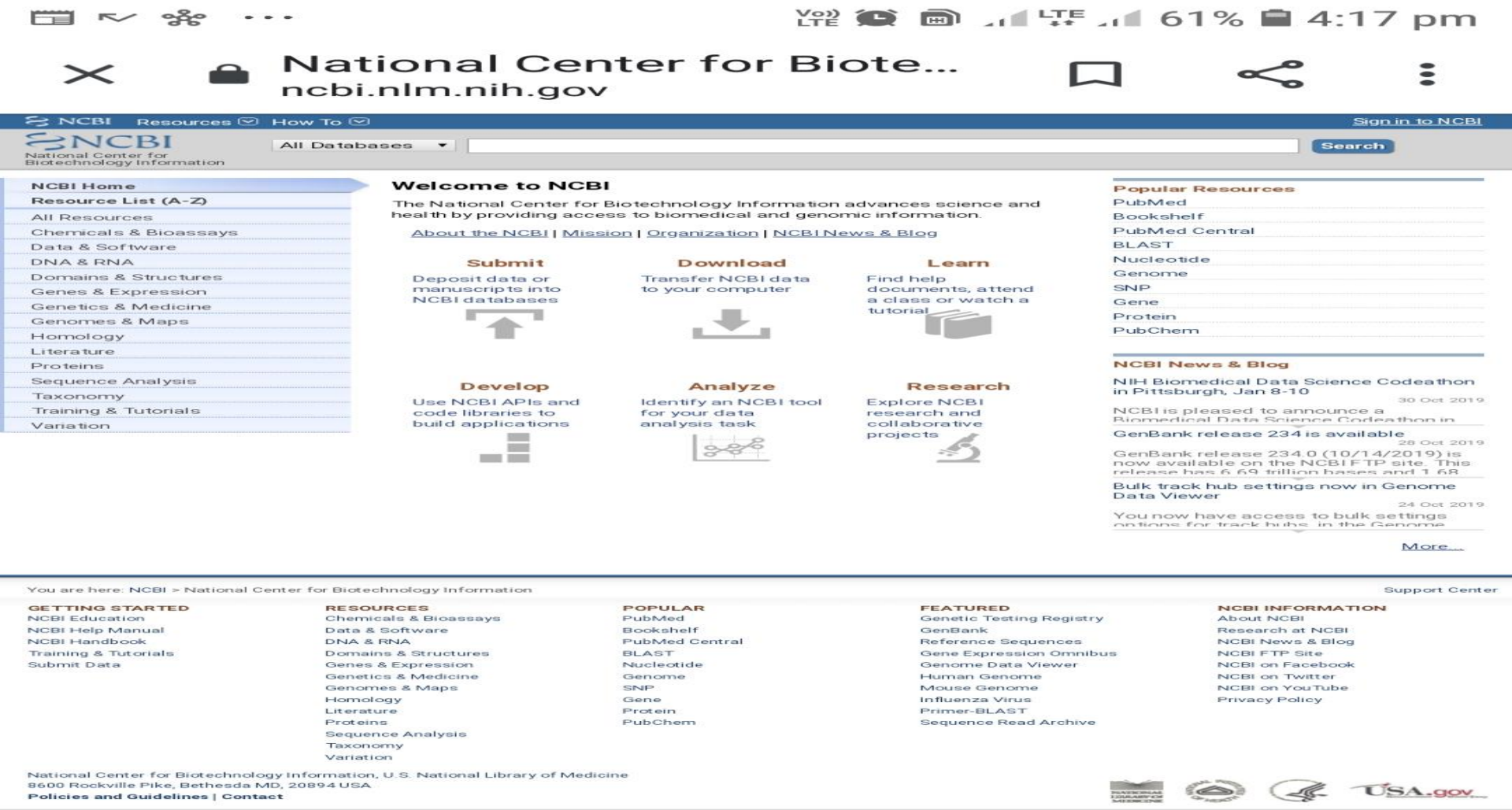
- **NCBI - GenBank:** (www.ncbi.nlm.nih.gov/GenBank)
- **EMBL:** (www.ebi.ac.uk/embl)
- **DDBJ:** (www.ddbj.nig.ac.jp)

The 3 databases are updated and exchanged on a daily basis and the accession numbers are consistent.

There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database.

The International Nucleotide Sequence Database Collaboration (INSD)

National Center for Biotechnology Information (NCBI)



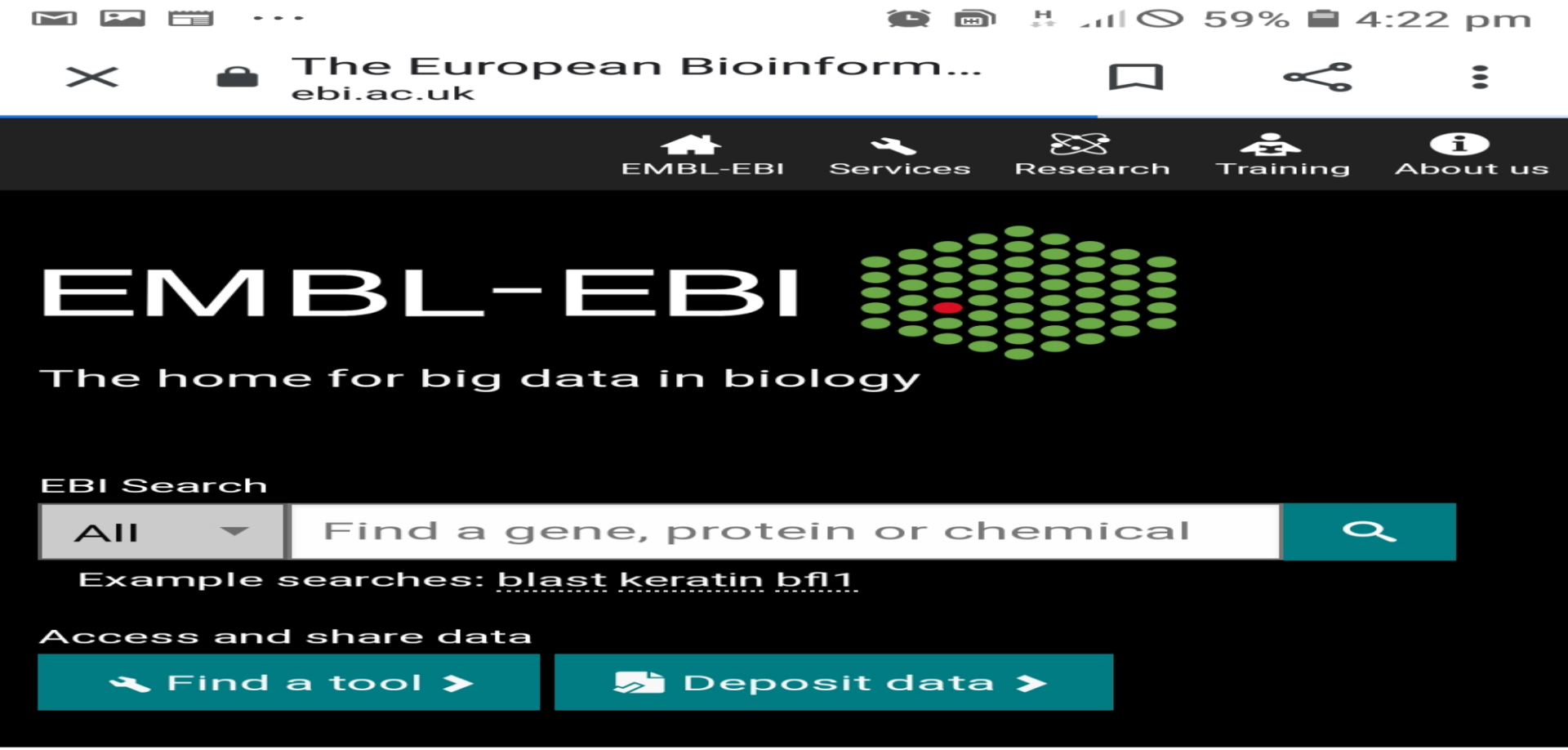
EMBL Database

European Molecular Biology Laboratory (EMBL) :

- **Maintained by *European Bioinformatics Institute* (EBI)**

  - ✓ **GSS (genome survey sequences)**
  - ✓ **HTC (high-throughput c-DNA sequences)**
  - ✓ **HTG (high-throughput genomic sequences)**
  - ✓ **EST (expressed sequence tag)**
  - ✓ **Patents**

European Bioinformatics Institute (EBI)



We are EMBL-EBI

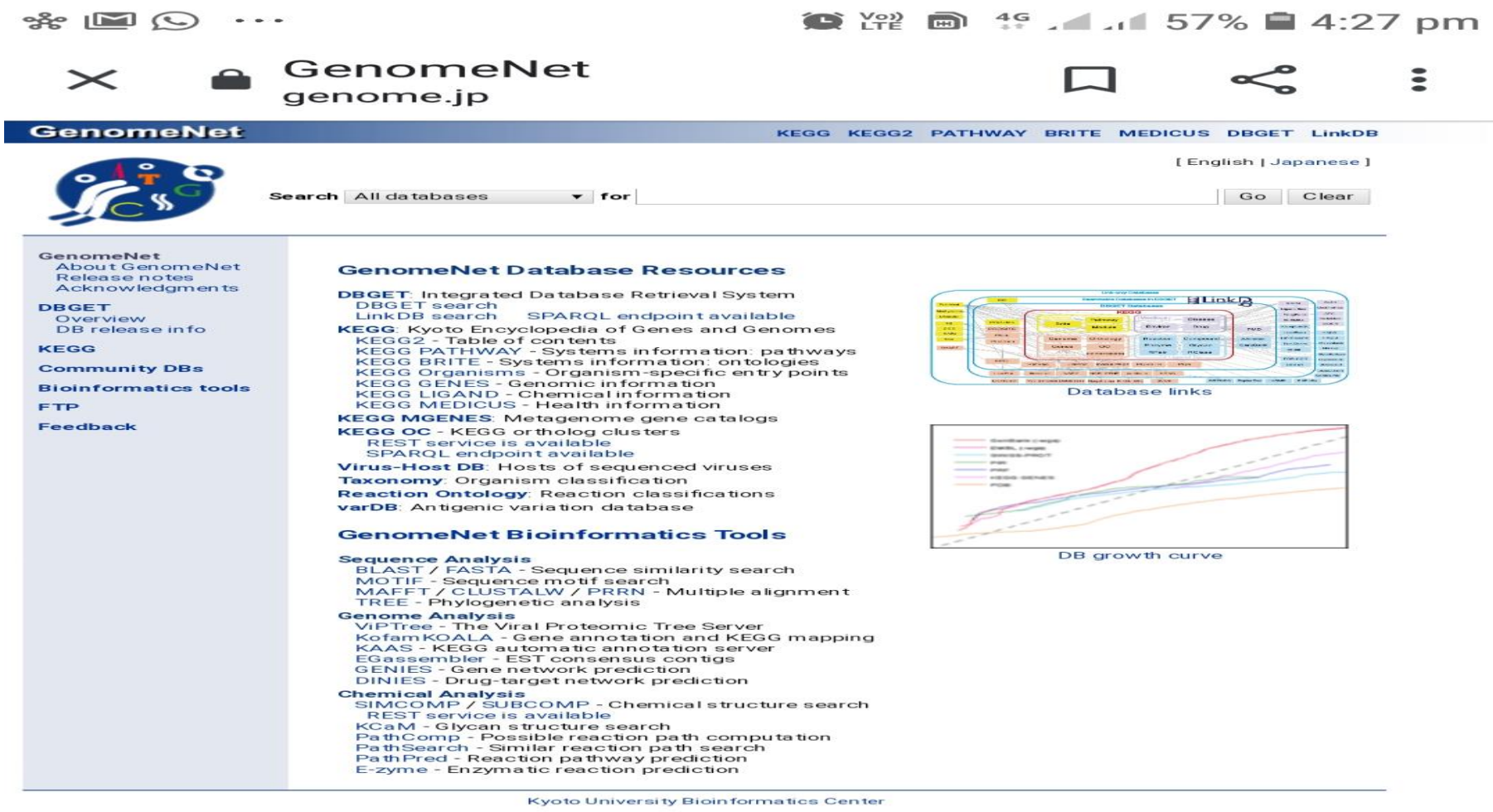
The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact.

Data resources

DDBJ (DNA Database of GenomNet, Japan)

- Developed in 1986 as a collaboration with EMBL and GenBank.
- Produced, maintained and distributed by the National Institute of Genetics, Japan.
- Sequences is submitted via Web based data submission tool.

GenomeNet, Japan



Other Databases

- **ESTs - Expressed Sequence Tags**
  - **dbEST (http://www.ncbi.nlm.nih.gov/dbEST)**
    - **GenBank subset with additional EST-specific data**
    - **Implemented in a Sybase relational database**
- **SNPs - Single Nucleotide Polymorphisms**
  - **dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)**
    - **Very similar to dbEST in philosophy and implementation**
- **Many commercial databases**
  - **Celera, Incyte, etc.**

Protein Databases

Protein sequence database

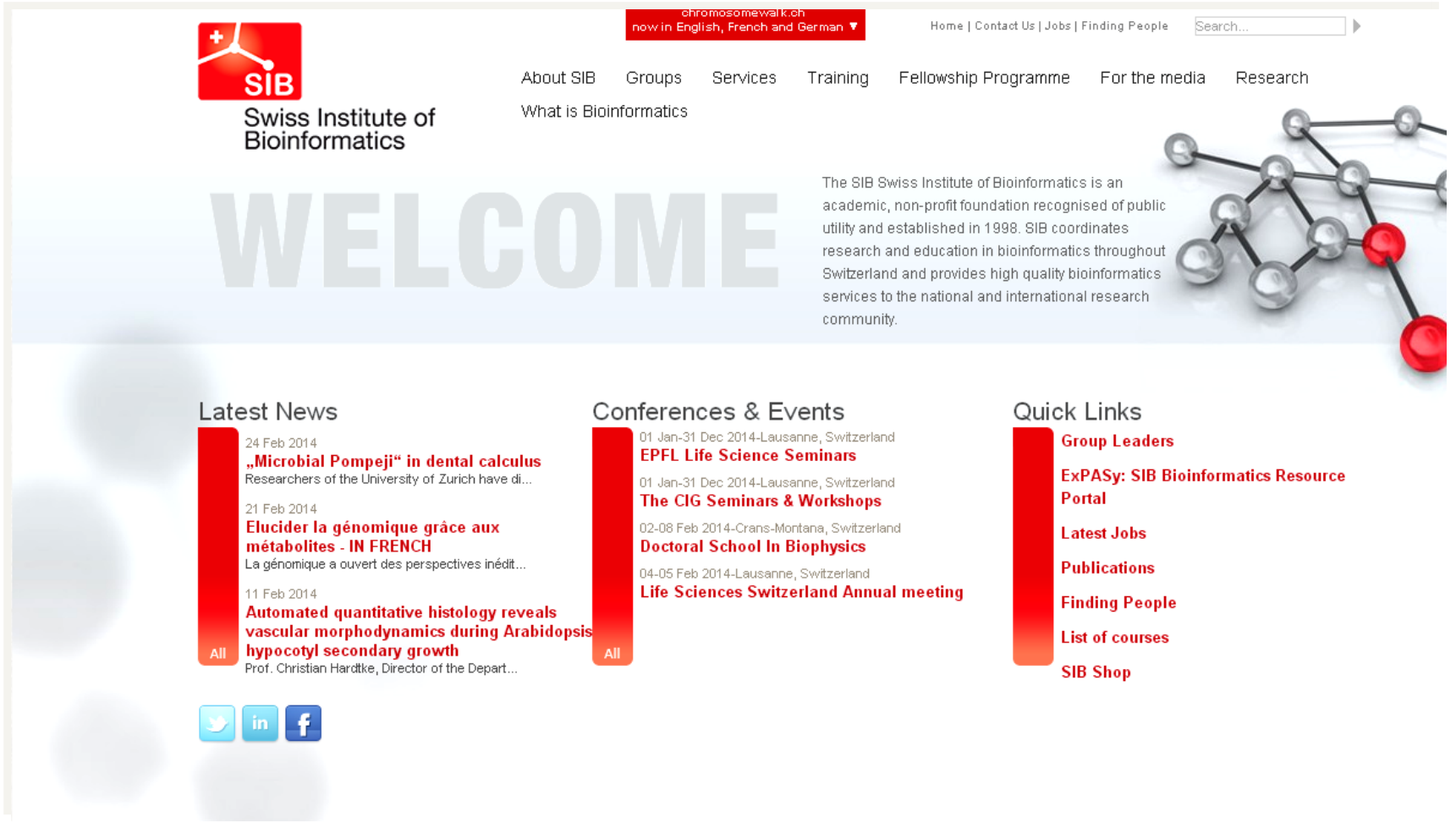
- Functions as repository of raw data: two types
- **Primary**
- **Secondary**

Protein structure database

Primary databases

1. **SWISS-PROT: Groups at Swiss Institute of Bioinformatics (SIB).**
   - It annotate the sequences
   - Describe protein functions
   - Its domain structures
   - Its post translations modifications
   - Provides high level of annotation
   - Minimum level of redundancy
   - High level of integration with other databases
2. **TrEMBL:**
   - Computer annotated supplements of SWISS-PROT that contains all the translations of EMBL nucleotide entries not yet integrated in SWISS-PROT.
2. **PIR: Protein Information Resource, a division of NBRF in US.**
   - Collaborated with Munich Information Centre for Protein Sequences (MIPS) and Japanese International Protein Sequence Database (JIPID).
   - One an search for entries
   - Do sequence similarity
   - PIR also produces MRL-3D (db of sequences extracted from 3D structures in PDB)

Swiss-Prot


chromosomewalk.ch
now in English, French and German
Home | Contact Us | Jobs | Finding People
Search...
SIB
Swiss Institute of Bioinformatics
About SIB
Groups
Services
Training
Fellowship Programme
For the media
Research
What is Bioinformatics
WELCOME
The SIB Swiss Institute of Bioinformatics is an academic, non-profit foundation recognised of public utility and established in 1998. SIB coordinates research and education in bioinformatics throughout Switzerland and provides high quality bioinformatics services to the national and international research community.
Latest News
24 Feb 2014
„Microbial Pompeji“ in dental calculus
Researchers of the University of Zurich have di...
21 Feb 2014
Elucider la génomique grâce aux métabolites - IN FRENCH
La génomique a ouvert des perspectives inédit...
11 Feb 2014
Automated quantitative histology reveals vascular morphodynamics during Arabidopsis hypocotyl secondary growth
Prof. Christian Hardtke, Director of the Depart...
All
Conferences & Events
01 Jan-31 Dec 2014-Lausanne, Switzerland
EPFL Life Science Seminars
01 Jan-31 Dec 2014-Lausanne, Switzerland
The CIG Seminars & Workshops
02-08 Feb 2014-Crans-Montana, Switzerland
Doctoral School In Biophysics
04-05 Feb 2014-Lausanne, Switzerland
Life Sciences Switzerland Annual meeting
All
Quick Links
Group Leaders
ExPASy: SIB Bioinformatics Resource Portal
Latest Jobs
Publications
Finding People
List of courses
SIB Shop

Secondary databases

- Secondary db compile and filter sequence data from different primary db.
- These db contain information derived from protein sequences and help the user determine whether a new sequence belong to a known protein family.

1. PROSITE:

- db of short protein sequence patterns and profiles that characterise biologically significant sites in proteins
- It is based on regular expressions describing characteristic sequences of specific protein families and domains.
- It is part of SWISS-PROT, and maintained in the same way

2. PRINTS

- PRINTS provides a compendium of protein fingerprints (groups of conserved motifs that characterise a protein family)
- Now has a relational version, "PRINTS-S“

3. BLOCKS

- BLOCK patterns without gaps in aligned protein families defined by PROSITE, found by pattern searching and statistical sampling algorithms.
- Automatically determined un-gapped conserved segments

4. Pfam

- Db of protein families defined as domains
- For each domain, it contains a multiple alignment of a set of defining sequences and the other sequences in SWISS-PROT and TrEMBL that can be matched to the alignment.

Protein Structural Database

1. PDB (Protein Data Bank):

- Main db of 3D structures of biological macromolecules (determined by X-ray crystallography and NMR).
- PDB entrys contain the atomic coordinates, and some structural parameters connected with the atoms or computed from the structures (secondary structure).
- PDB provide primary archive of all 3D structures for macromolecules such as proteins, DNA, RNA and various complexes.

2. SCOP (Structural Classification of Proteins):

- Db was started to with objective to classify protein 3D structures in a hierarchical scheme of structural classes.
- It is based on data in a primary db, but adds information through analysis and organization (such as classification of 3D structures into hierarchical scheme of folds, super-families and families)

3. CATH (Class, architecture, topology, homologous super-family):

- CATH perform hierarchical classification of protein domain structures.
- Clusters proteins at four major structural levels

Enzyme Database

- **BRENDA** [BRaunshchweig ENzyme DAtabase] (www.brenda.uni-koeln.de)

- **Enzyme, a part of ExPaSy (Expert Protein Analysis System, the proteomic server of Swiss Institute of Bioinformatics)**

Clinical Databases

Generally contain information from the Human

Human Gene Mutation Database, Cardiff, UK:
http://www.hgmd.org

Registers known mutations in the human genome and the diseases they cause.

OMIM database
Online Mendelian Inheritance in Man
http://www.ncbi.nlm.nih.gov/Omim

The OMIM database contains abstracts and texts describing genetic disorders to support genomics efforts and clinical genetics. It provides gene maps, and known disorder maps in tabular listing formats. Contains keyword search.

Kyoto Encyclopedia of Genes and Genomes (KEGG) www.genome.jp/kegg/

Database and associated software which integrates several databases such as,

- **Pathway database**
- **Genes database**
- **Genome database**
- **Drug database**
- **Reaction database**
- **Compound database**
- **KO database etc.**

Bibliographic Databases

Used for searching for reference articles

PubMed

1. It enables user to do keyword searches, provides links to a selection of full articles, and has text mining capabilities, e.g. provides links to related articles, and GenBank entries, among others.

2. It contains entries for more than 30 million abstracts of scientific publications.

Database Mining Tools (Analysis Tools)

Utilization of various databases requires the use of suitable search engines and analysis tools. These tools are called Database mining tools and the process of data utilization is known as database mining. **Some Analysis Tools are as follows:**

Analysis Tool	Function
BLAST (NCBI, USA)	Used to analyse sequence information and detect homologous sequences
ENTREZ (NCBI, USA)	Used to access literature (abstracts), sequence and structure db
DNAPLOT (EBI, UK)	Sequence alignment tool
LOCUS LINK (NCBI, USA)	Assessing information on homologous genes
LIGAND (GenomNet, Japan)	A chemical db, allows search for a combination of enzymes and links to all publically accessible db.
BRITE (GenomNet, Japan)	Biomolecular relations information transmission and expression db; links to all publically accessible db.
TAXONOMY BROWSER (NCBI, USA)	Taxonomic classification of various species as well as genetic information
STRUCTURE	It support Molecular Modelling Database (MMDB) and software tools for structure analysis

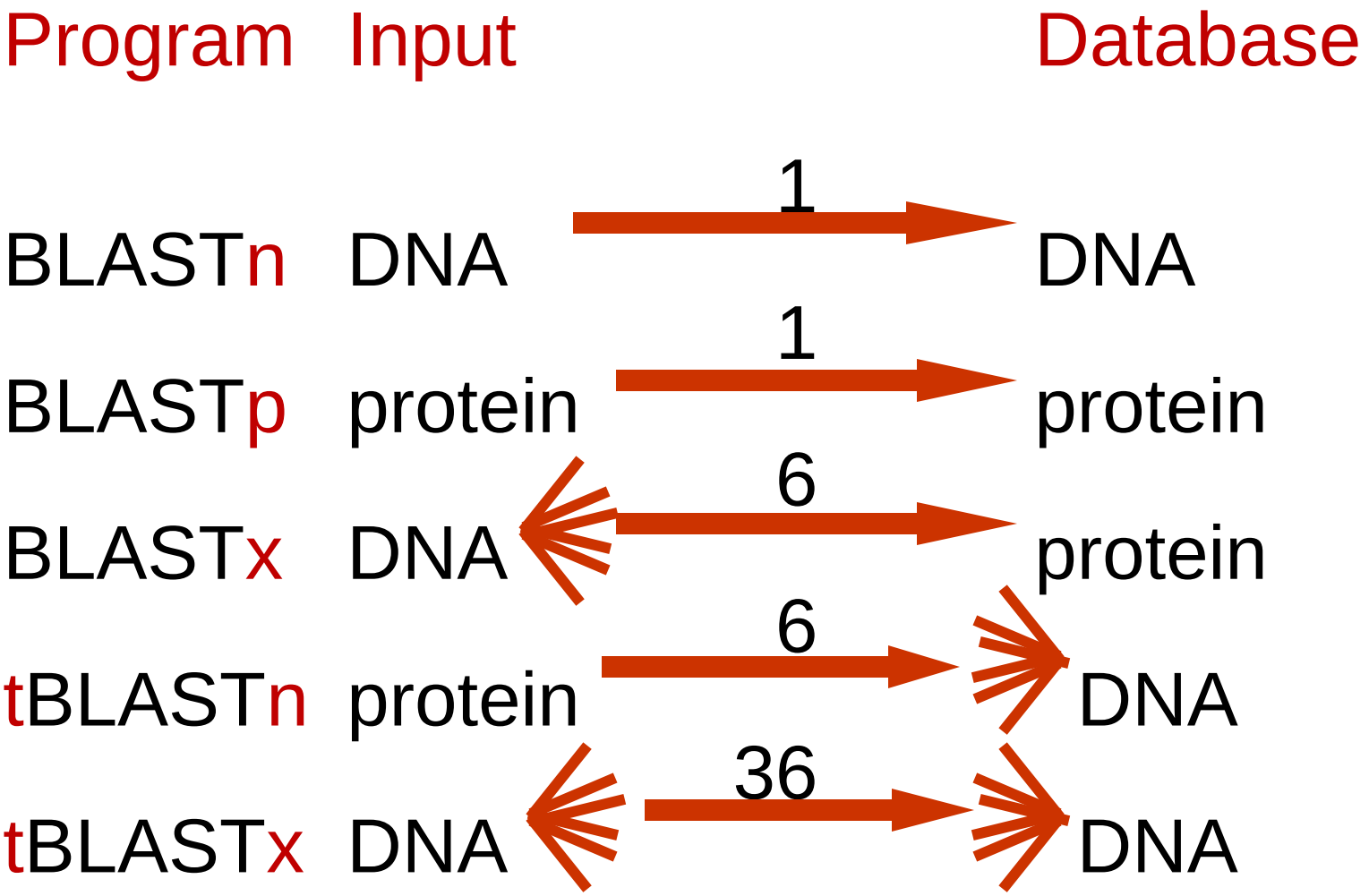
BLAST

(Basic Local Alignment Search Tool) for Homology Analyses

- **BLASTn**
  - **Nucleotide query vs nucleotide database**
- **BLASTp**
  - **protein query vs protein database**
- **BLASTx**
  - **automatic 6-frame translation of nucleotide query vs protein database**
  - **If you have a DNA sequence and you want to now what protein (if any) it encodes, you can perform BLASTx search.**
- **tBLASTn**
  - **protein query vs automatic 6-frame translation of nucleotide database**
  - **You can use this program to ask whether a DNA or ESTs db contains a nucleotide sequence encoding a protein that matches your protein of interest.**
- **tBLASTx**
  - **automatic 6-frame translation of nucleotide query vs automatic 6-frame translation of nucleotide database**

BLAST

(Basic Local Alignment Search Tool) for Homology Analyses



SEQUENCE ALIGNMENT

What is Sequence Alignment ?

A sequence alignment is a way of arranging the sequences of DNA or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Definitions

Similarity
The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

Identity
The extent to which two sequences are invariant.

Conservation
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Types of alignment

- **Pairwise alignment**
- **Multiple Alignment**

Pairwise alignment

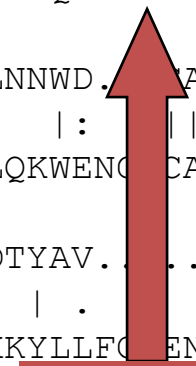
- **The process of lining up two sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.**
- **Pairwise sequence alignment is the most fundamental operation of bioinformatics.**

Pairwise alignment of retinol-binding protein 4 and b-lactoglobulin

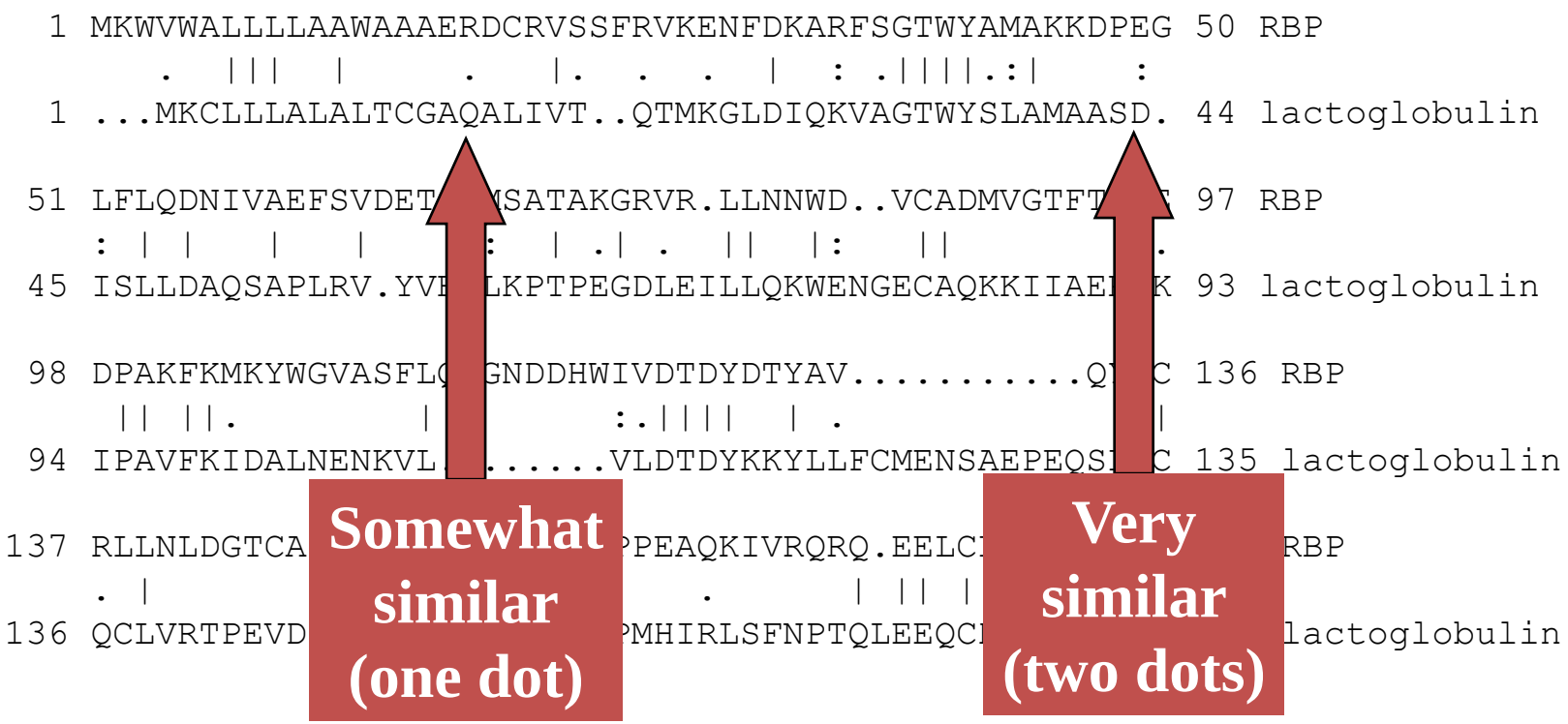
```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Pairwise alignment of retinol-binding protein and b-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
       .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD.VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |: ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV.........QYSC 136 RBP
     || ||.        |        :.||||  | .           .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIV          RQYRLIV 185 RBP
    . |       |     | :    ||   .
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSF.....  178 lactoglobulin
```



Pairwise alignment of retinol-binding protein and β-lactoglobulin

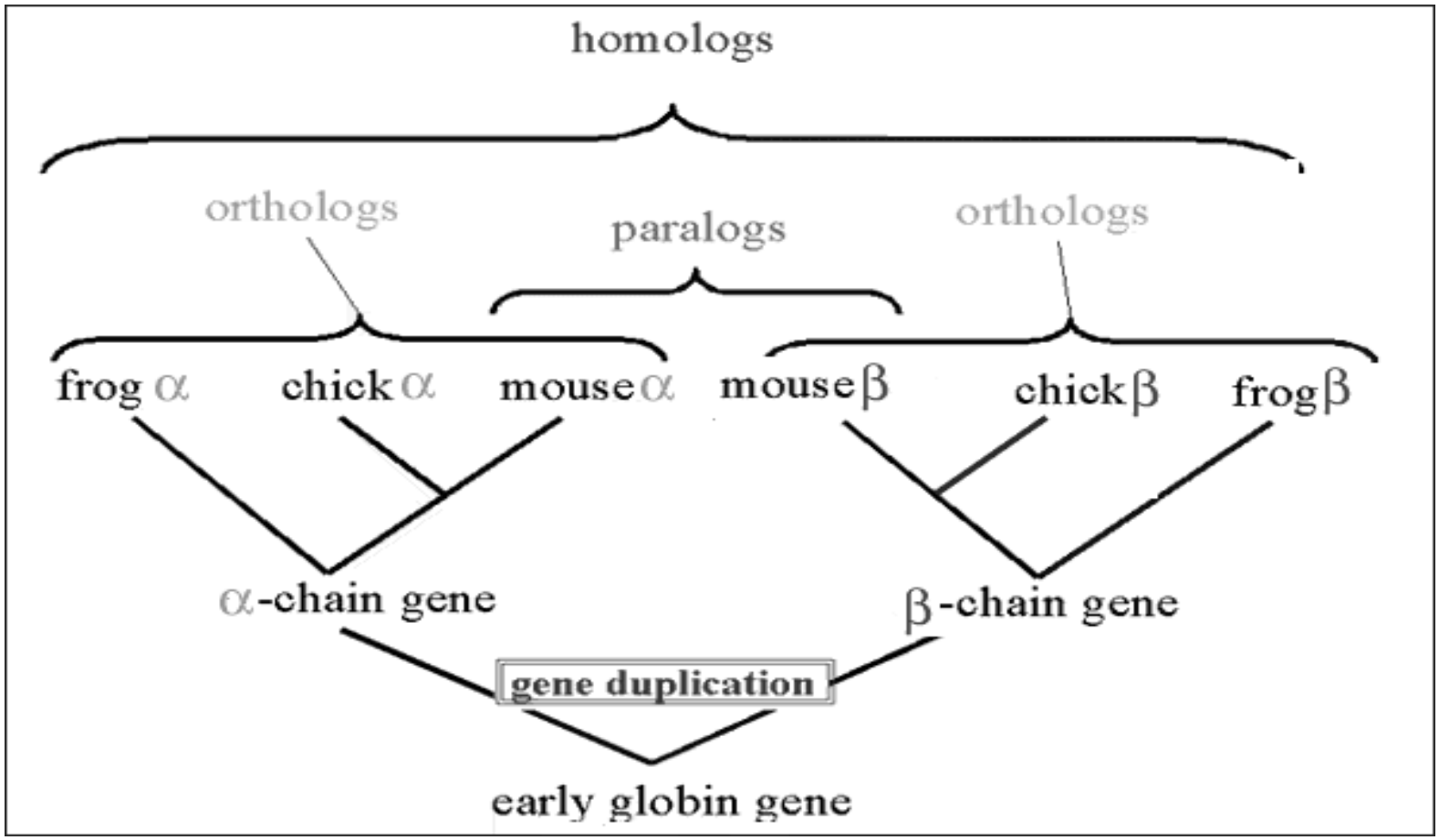


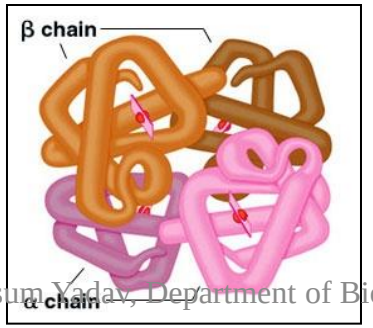
Pairwise alignment of retinol-binding protein and b-lactoglobulin

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136 RBP
     || ||.        |        :.||||  | .            .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSR NGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDK LKALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```

Internal gap

Terminal gap


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene


β chain
α chain

Sequence Analyses for relatedness

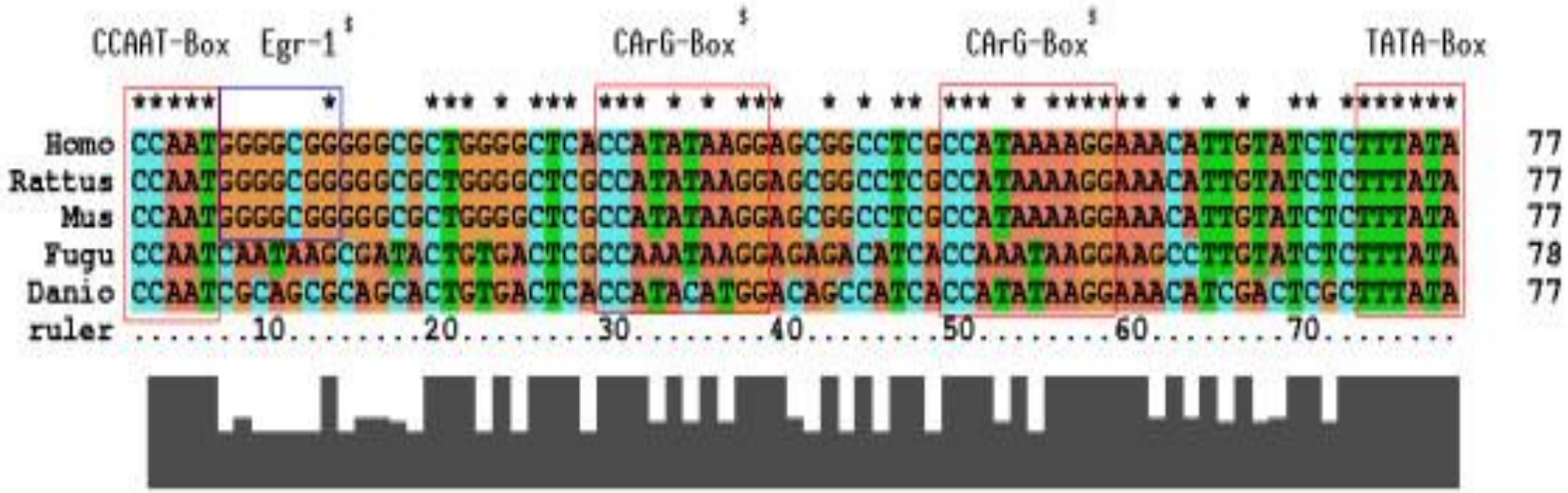
- **Homologs:** similar sequences in different organisms derived from a common ancestor sequence.
- **Orthologs :** homologous sequences in different related species that arose from a common ancestral gene during speciation.

  Orthologs are presumed to have similar biological function. e.g. Human and rats myoglobins both transport oxygen in muscle
- **Paralogs:** homologous genes within the same organism

  e.g. human α and β globins are paralogs. Paralogs are the result of gene duplication events
- **Xenologs:** similar sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

Multiple sequence Alignment

- Partial or complete alignment of three or more related proteins/ nucleotide sequences
- Conserved domain analysis
- Primer Designing



Tools of Multiple Alignment

- **CLUSTALW**
- **T-Coffee**
- **MUSCLE**
- **KALIGN**
- **CLC & GCG WorkBench**

Various categories of Analyses

1. **Analysis of a single gene (protein) sequence**
   - Similarity with other known genes
   - Phylogenetic trees; evolutionary relationships
   - Identification of well-defined domains in the sequence
   - Sequence features (physical properties, binding sites, modification sites)
   - Prediction of sub-cellular localization
   - Prediction of protein secondary and tertiary structures

2. Analysis of whole genomes

- Location of variuos genes on the chromosomes, correlation with function or evolution
- Expansion/duplication of gene families
- Which gene families are present, which missing?
- Presence or absence of biochemical pathways
- Identification of "missing" enzymes
- Large-scale events in the evolution of organisms

3. Analysis of genes and genomes with respect to function (Functional Annotation)

- **Transcriptomics : Expression analysis; micro array data (mRNA/transcript analyses)**
- **Proteomics; protein qualitative and quantitative analyses, covalent modifications**
- **Comparison and analysis of biochemical pathways**
- **Deletion or mutant genotypes vs phenotypes**
- **Identification of essential genes, or genes involved in specific processes**

4. Comparative genomics

- **Identifying pathogen specific unique targets for designing novel drugs.**

Phylogenetic Analysis

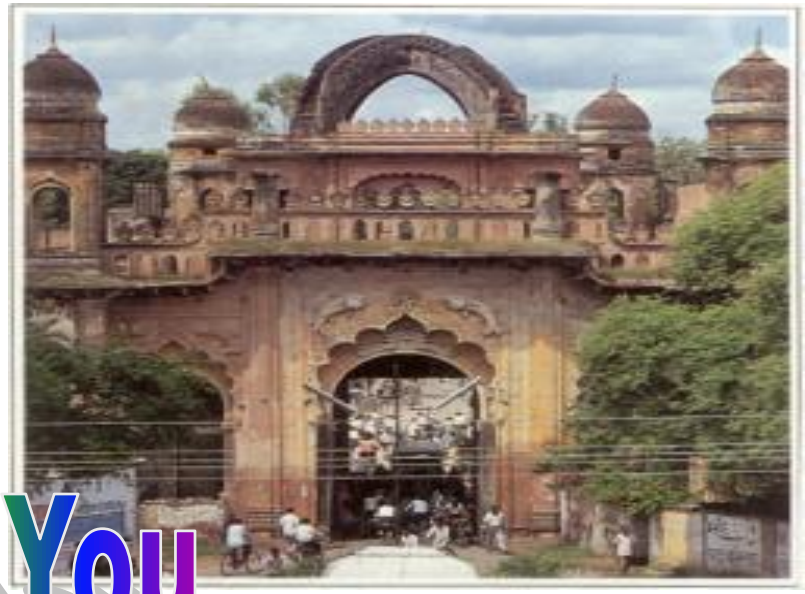
- **The phylogenetic trees aim at reconstructing the history of successive divergence which took place during the evolution, between the considered sequences and their common ancestor.**

- **Nucleic acid and protein sequences are used to infer Phylogenetic relationships**

- **Molecular phylogeny methods allow the suggestion of phylogenetic trees, from a given set of aligned sequences.**

Phylogenetic Analysis Tools

- ✓MEGA
- ✓PHYLIP
- ✓PAUP
- ✓Treeview
- ✓ODEN
- ✓PHYLOWIN
- ✓TREECON
- ✓DENDRON



Thank You