BIOINFORMATICS -1

Concepts of Genomics & Bioinformatics (6 hours)
Genomics & personalized medicine,
Introduction to Bioinformatics,
DNA sequence as data
Sequence alignment using comparative tools
homology, phylogeny, Mining for data in relevant databases

Introducing and defining bioinformatics and its main applications
Building the necessary foundations to understand the problems that bioinformaticians tackle
Providing an introduction to some of the main bioinformatics tools in use

An introduction to biological databases

What is a database ?

- A collection of
  - structured
  - searchable (index)
  - updated periodically (release)
  - cross-referenced data

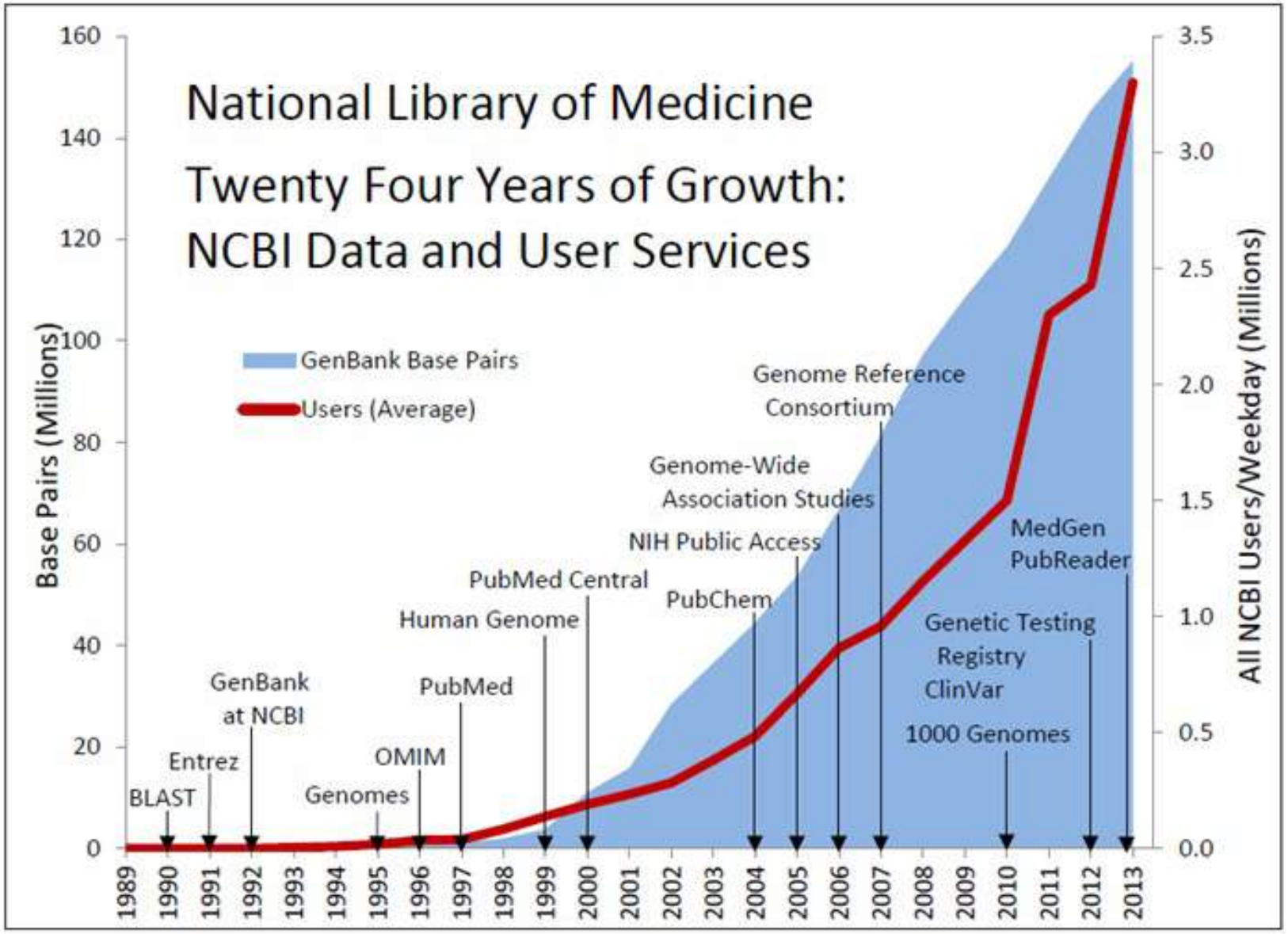
- Includes also associated tools (software) necessary for db access/query, db updating, db information insertion, db information deletion….

Why biological databases ?

- Data (genomic sequences, 3D structures….) are no longer published in a conventional manner, but directly submitted to databases.

- Update frequency: daily to annually

Bioinformatics: A rapidly growing discipline



What is Bioinformatics?
National Center for Biotechnology (NCBI) definition

“Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned”

Bioinformatics vs. Computational Biology

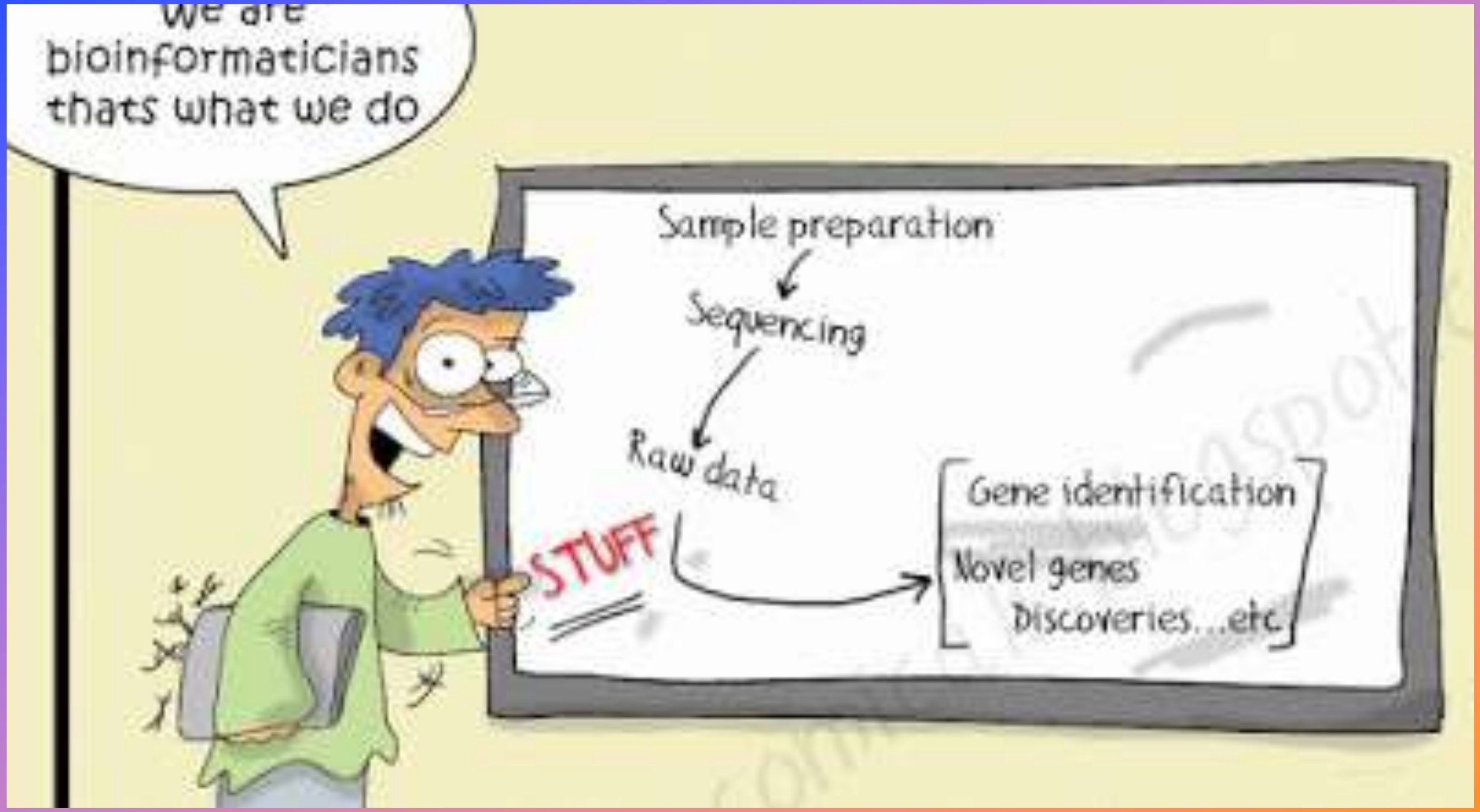
Computational Biology = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living sytems. It is about science

Bioinformatics = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering

In practice: Computational Biology ⇐⇒ Bioinformatics

Russ B. Altman

http://rbaltman.wordpress.com/2009/02/18/bioinformatics-computational-biology-same-no


we are
bioinformaticians
thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc

THE BIOINFORMATICIAN

Who isn't a true bioinformatician?

- A computer scientist who dabbles in biology
- A biologist who dabbles in computer science

So, who is a true bioinformatician?

A scientist with enough knowledge of both fields to be able to:

1. understand the biological problem at hand
2. come up with a computational solution to it

Bioinformatics is interdisciplinary

Biology

Molecular Biology

Computer Science

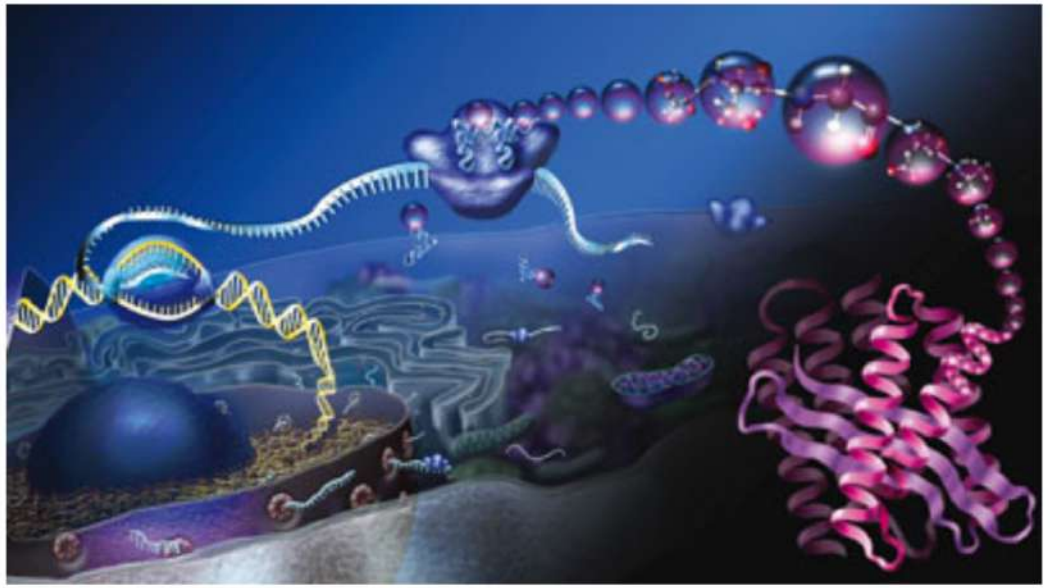
Probability Theory

Chemistry

Physics

.....

A quick primer on the molecules of life



DNA $\rightarrow$ RNA $\rightarrow$ proteins

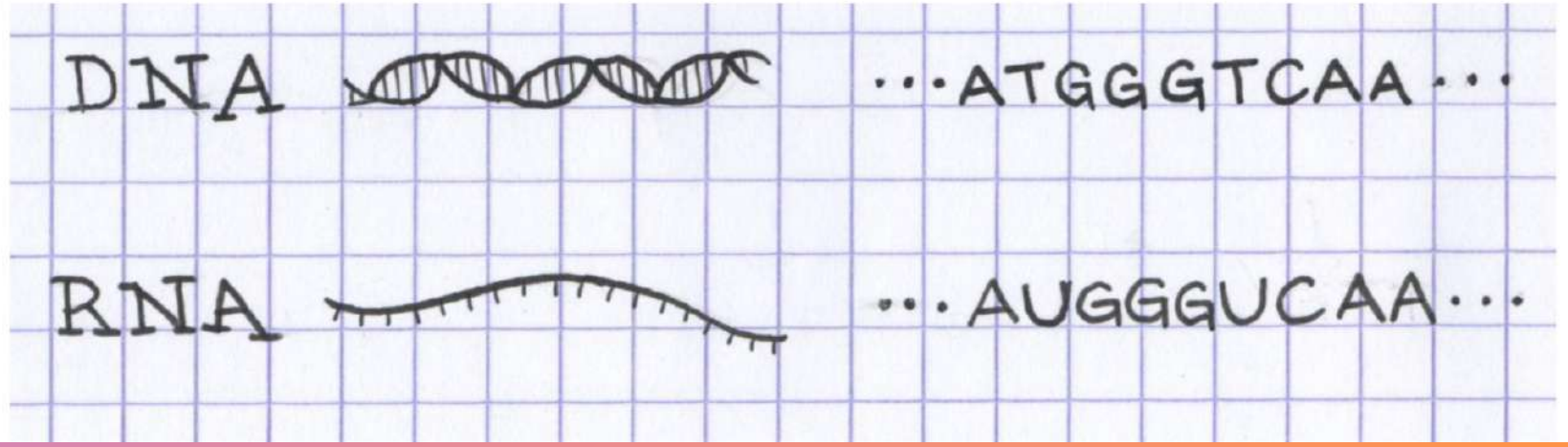
DNA: The blueprint of the cell

RNA: Intermediary molecule needed to make proteins, but endowed with other functions as well

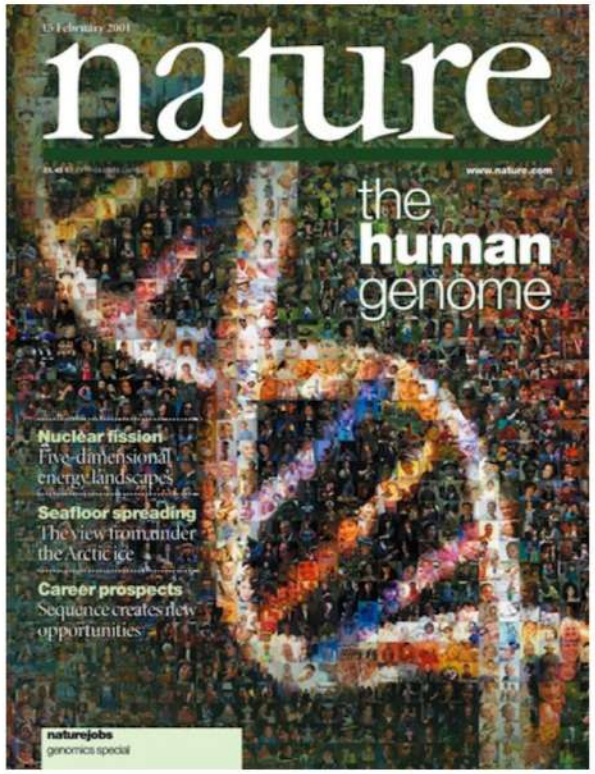
Proteins: Highly sophisticated machines, the workhorse of Cell

The strings of life

- DNA – 3 billion letters (2m. long) in human, size of the alphabet: 4
- RNA – size of the alphabet: 4
- Proteins – size of the alphabet: 20



DNA Sequencing



- DNA sequencing technology enables us to identify the sequence of letters (called nucleotides) that make up the DNA string

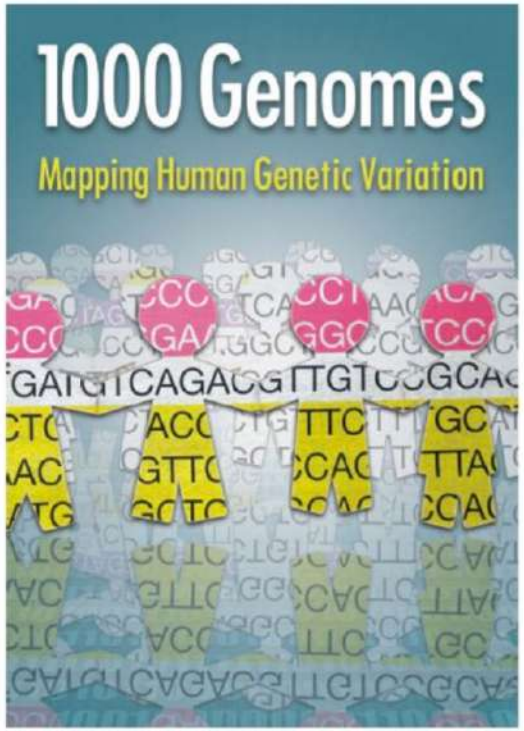
In 2001 the first draft of the human genome was released

One genome is not enough

Differences in DNA make us different

```
Reference genome:   ...ACCGTTACGCGAAAG...

Individual A:       ...AGCGTTACGCGAAAG...

Individual B:       ...ATCGTTACGCGAAAG...

Individual C:       ...ATCGTTA---GAAAG...

Individual D:
...ATCGTTACGCGAAAG...ACCGTTACGCGAAAG...
```

The 1000Genomes Project



- A collaborative effort among research groups in the US, UK, Germany, and China
- Over 2,500 genomes of individuals from Europe, Africa, Asia, and the Americas have been sequenced

What can we do with this information?

Some applications:

- 1 understand the molecular bases of human variation
- 2 identify the regions of DNA that are highly conserved across healthy individuals (and therefore potentially disease-causing or lethal if mutated)
- 3 filter out common variants that are unlikely to be associated with disease (to increase the power of genome-wide association studies)
- 4 identify population-specific rare variants

But....
Data is only useful if we have the conceptual framework and practical tools to interpret it!

Role of Bioinformatics in the 1000 Genomes Project

Bioinformatics is involved at every step:

1. Calling the variants (that is, determining whether at a given position in the DNA string two individuals differ from each other)

2. Computing the degree of conservation among individuals for given positions

3. Predicting the effect of a variant on protein structure and function

4. ...

Cancer Genome Atlas

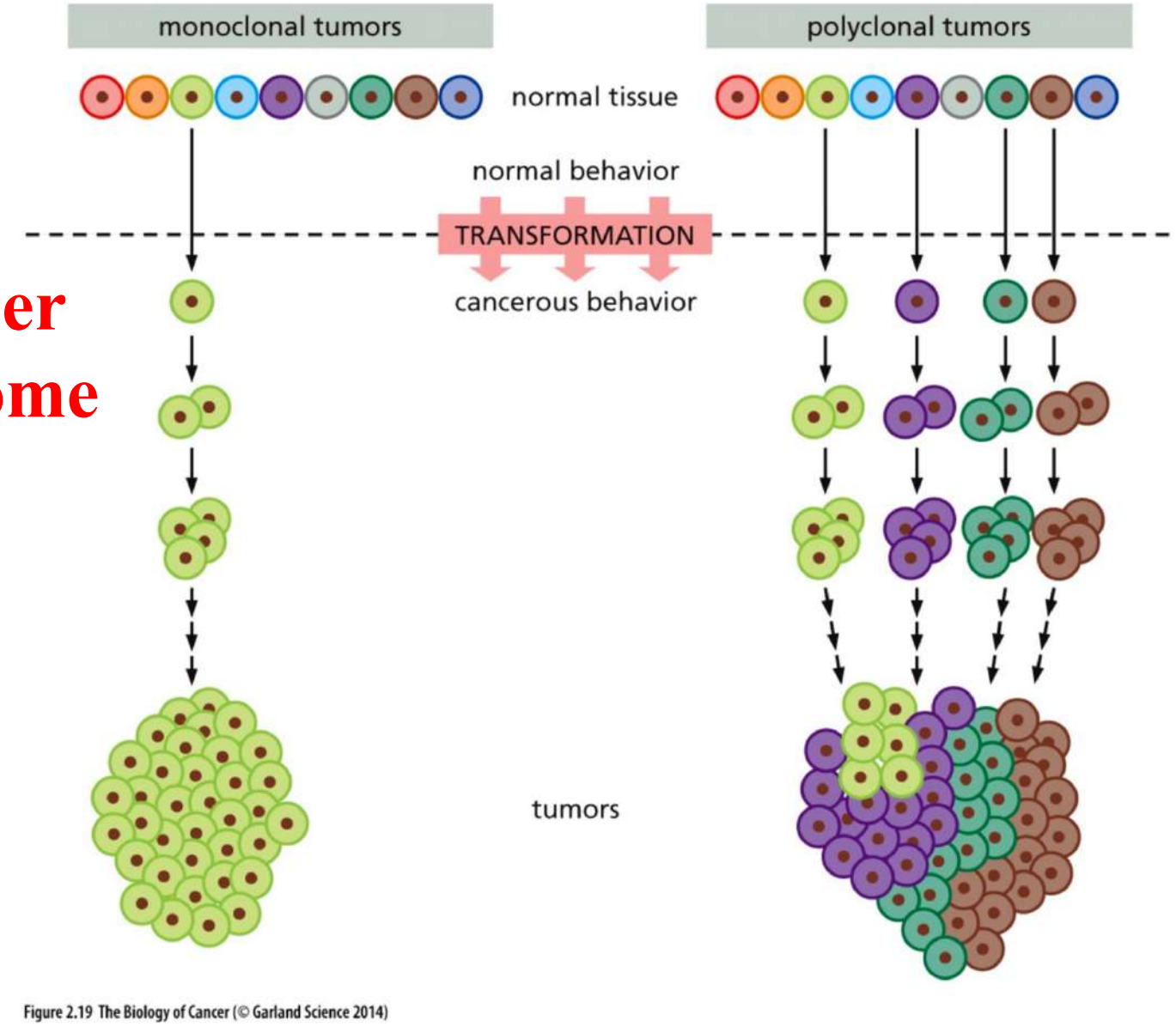


Figure 2.19 The Biology of Cancer (© Garland Science 2014)

Over 30 types of human cancers have been sequenced, in thousands of patients

Other types of data have been collected (for example, how much of a gene is expressed by a given patient's cancer cell)

All the data is publicly available

Questions where bioinformatic approaches are vital:

1. Which mutations are driving the formation of cancers? (most mutations are thought to be of the “passenger” type, i.e., not contributing to cancer directly)

2. What are the functional consequences of a given mutation on protein structure and function?

3. How can we identify the most “actionable”3 mutations in a patient?

...

Bioinformatics Databases

The ten important bioinformatics databases

GenBank/DDJB/EMBL	www.ncbi.nlm.nih.gov	Nucleotide sequences
Ensembl	www.ensembl.org **	Human/mouse genome
PubMed	www.ncbi.nlm.nih.gov	Literature references
NR	www.ncbi.nlm.nih.gov	Protein sequences
Swiss-Prot	www.expasy.org	Protein sequences
InterPro	www.ebi.ac.uk	Protein domains
OMIM	www.ncbi.nlm.nih.gov	Genetic diseases
Enzymes	www.expasy.org	Enzymes
PDB	www.rcsb.org/pdb/	Protein structures
KEGG	www.genome.ad.jp	Metabolic pathways

Ideal minimal content of a sequence database entry

- Sequences !!
- Accession number (AC) (unique identifier, specific to a database)
- Taxonomic data
- References
- ANNOTATION/CURATION
- Keywords
- Cross-references
- Documentation

Database 1: nucleotide sequences

- Main nucleic acid sequence databases are –

NCBI **database** (www.ncbi.nlm.nih.gov/)

European Molecular Biology Laboratory (EMBL) **database** (www.ebi.ac.uk/embl/

DNA Database of Japan (DDBJ) **database** (www.ddbj.nig.ac.jp/)

« different views of the same data set »

EMBL/GenBank/DDBJ

- Serve as **archives**

  containing all public sequences derived from:

  - Genome projects (> 80 % of entries)
  - Sequencing centers (cDNAs, ESTs…)
  - Individual scientists (15 % of entries)
  - Patent offices (i.e. European Patent Office, EPO)

NCBI

Sequences in the NCBI Sequence Database (or EMBL/DDBJ) are identified by an accession number. Unique number that is only associated with one sequence.

Example:

- ❑ Accession number NC_001477 is for the DEN-1 Dengue virus genome sequence.
- ❑ The accession number is what identifies the sequence.
- ❑ It is reported in scientific papers describing that sequence.

Sequence itself, for each sequence the NCBI database (or EMBL/DDBJ databases) also stores some additional *annotation* data, such as
Name of the species it comes from
References to publications describing that sequence, etc.

Some of this annotation data was added by the person who sequenced a sequence and submitted it to the NCBI database, while some may have been added later by a human curator working for NCBI.

NCBI

The NCBI database contains several sub-databases, the most important of which are:

- the NCBI Nucleotide database: contains DNA and RNA sequences
- the NCBI Protein database: contains protein sequences
- EST: contains ESTs (expressed sequence tags), which are short sequences derived from mRNAs
- the NCBI Genome database: contains DNA sequences for whole genomes
- PubMed: contains data on scientific publications

Searching for an accession number in the NCBI database

Obtain a FASTA file containing the DNA sequence corresponding to a particular accession number, eg. accession number NC_001477 (the DEN-1 Dengue virus genome sequence)

FASTA format is a file format commonly used to store sequence information.
The first line starts with the character '>' followed by a name and/or description for the sequence.
Subsequent lines contain the sequence itself.

```
>mysequence1
ACATGAGACAGACAGACCCCCAGAGACAGACCCCTAGACACAGAGAGAG
TATGCAGGACAGGGTTTTTGCCCAGGGTGGCAGTATG
```

A FASTA file can contain more than one sequence. If a FASTA file contains many sequences, then for each sequence it will have a header line starting with '>' followed by the sequence itself.

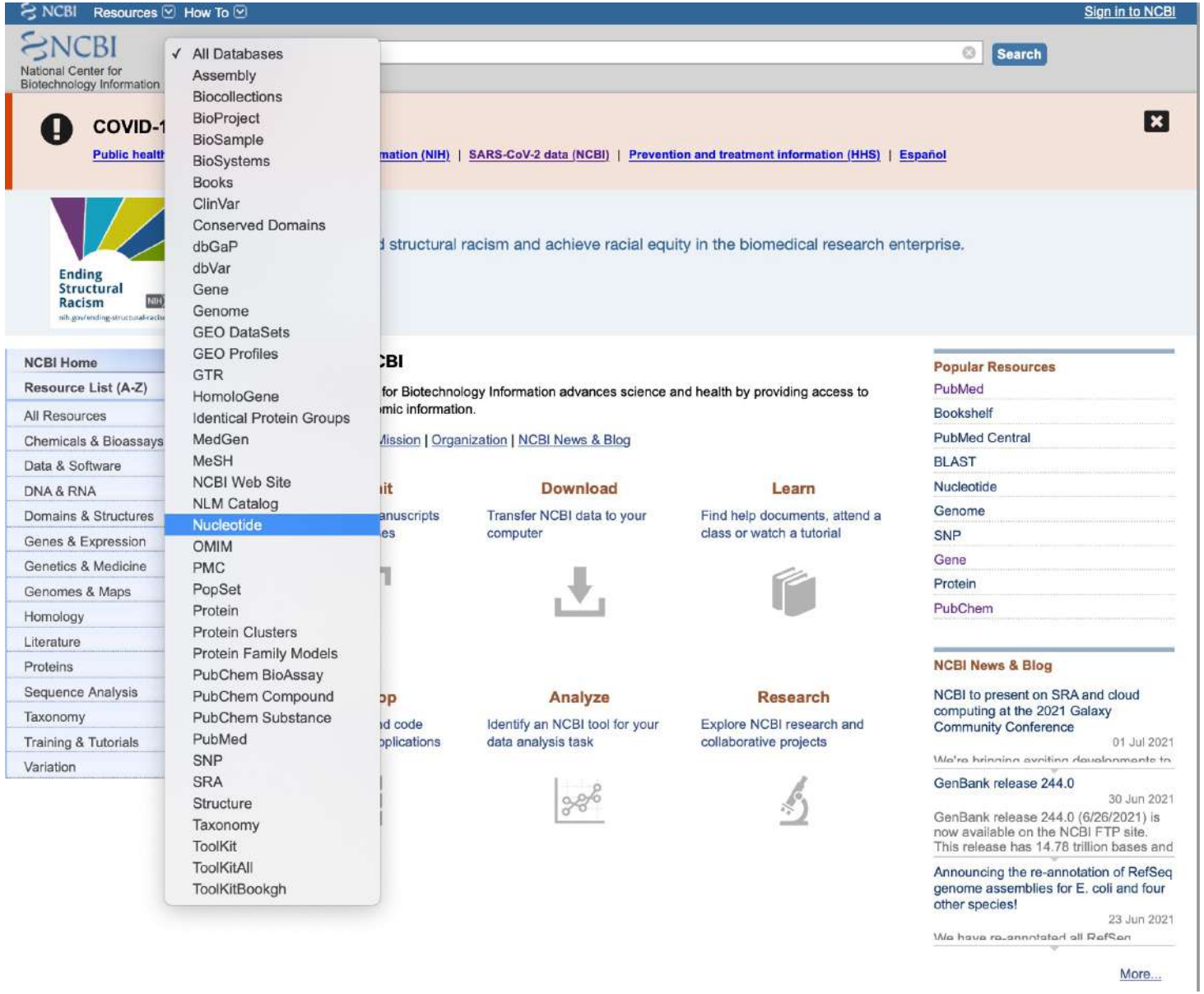
```
>mysequence1
ACATGAGACAGACAGACCCCCAGAGACAGACCCCTAGACACAGAGAGAG
TATGCAGGACAGGGTTTTTGCCCAGGGTGGCAGTATG
>mysequence2
AGGATTGAGGTATGGGTATGTTCCCGATTGAGTAGCCAGTATGAGCCAG
AGTTTTTTACAAGTATTTTTCCCAGTAGCCAGAGAGAGAGTCACCCAGT
ACAGAGAGC
```

NCBI Sequence Format (NCBI Format)

To view the NCBI entry for the DEN-1 Dengue virus (which has accession NC_001477), follow these steps:

1.Go to the NCBI website (www.ncbi.nlm.nih.gov).
2.Search for the accession number.
3.On the results page, if your sequence corresponds to a nucleotide (DNA or RNA) sequence, you should see a hit in the Nucleotide database, and you should click on the word ‘Nucleotide’ to view the NCBI entry for the hit. Likewise, if your sequence corresponds to a protein sequence, you should see a hit in the Protein database, and you should click on the word ‘Protein’ to view the NCBI entry for the hit.
4.After you click on ‘Nucleotide’ or ‘Protein’ in the previous step, the NCBI entry for the accession will appear.

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



Dengue virus 1, complete genome

NCBI Reference Sequence: NC_001477.1

FASTA Graphics

Go to:

```
LOCUS       NC_001477              10735 bp ss-RNA     linear   VRL 03-MAY-2019
DEFINITION  Dengue virus 1, complete genome.
ACCESSION   NC_001477
VERSION     NC_001477.1
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Dengue virus 1
  ORGANISM  Dengue virus 1
            Viruses; Riboviria; Orthornavirae; Kitrinoviricota; Flasuviricetes;
            Amarillovirales; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10735)
  AUTHORS   Puri,B., Nelson,W.M., Henchal,E.A., Hoke,C.H., Eckels,K.H.,
            Dubois,D.R., Porter,K.R. and Hayes,C.G.
  TITLE     Molecular analysis of dengue virus attenuation after serial passage
            in primary dog kidney cells
  JOURNAL   J. Gen. Virol. 78 (PT 9), 2287-2291 (1997)
   PUBMED   9292016
REFERENCE   2  (bases 1 to 10735)
  AUTHORS   McKee,K.T. Jr., Bancroft,W.H., Eckels,K.H., Redfield,R.R.,
            Summers,P.L. and Russell,P.K.
  TITLE     Lack of attenuation of a candidate dengue 1 vaccine (45AZ5) in
            human volunteers
  JOURNAL   Am. J. Trop. Med. Hyg. 36 (2), 435-442 (1987)
   PUBMED   3826504
REFERENCE   3  (bases 1 to 10735)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (01-AUG-2000) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   4  (bases 1 to 10735)
  AUTHORS   Puri,B. and Nelson,W.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-FEB-1997) Inf. Dis. Dept, Naval Medical Research
            Institute, 8901 Wisconsin Ave., Bethesda, MD 20889-5607, USA
COMMENT     VALIDATED REFSEQ: This record has undergone validation or
            preliminary review. The reference sequence was derived from U88536.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..10735
                     /organism="Dengue virus 1"
                     /mol_type="genomic RNA"
                     /db_xref="taxon:11053"
                     /clone="45AZ5"
                     /type="1"
```

Change region shown

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

NCBI Virus
Retrieve, view, and download dengue virus genomic and protein sequences.

Articles about the POLY gene
Nucleotide-dependent dynamics of the Dengue NS3 helicase [Biochim Biophys Acta Proteins ...]
Tracking dengue virus type 1 genetic diversity during lineage replacement in [PLoS One. 2019]
Infectivity of Dengue Virus Serotypes 1 and 2 Is Correlated with E-Protein Intrin [Structure. 2019]
See all...

Reference sequence information
RefSeq protein product
See the reference protein sequence for polyprotein (NP_059433.1).

More about the gene POLY
POLY gene
Also Known As: DV1_gp1, polyprotein gene

Related information
Assembly

NCBI

The NCBI entry for an accession contains a lot of information about the sequence, such as papers describing it, features in the sequence, etc.

The ‘DEFINITION’ field gives a short description for the sequence.

The ‘ORGANISM’ field in the NCBI entry identifies the species that the sequence came from.

The ‘REFERENCE’ field contains scientific publications describing the sequence.

The ‘FEATURES’ field contains information about the location of features of interest inside the sequence, such as regulatory sequences or genes that lie inside the sequence.

The ‘ORIGIN’ field gives the sequence itself.

NCBI
RefSeq

When carrying out searches of the NCBI database, it is important to bear in mind that the database may contain redundant sequences for the same gene that were sequenced by different laboratories (because many different labs have sequenced the gene, and submitted their sequences to the NCBI database).

There are also many different types of nucleotide sequences and protein sequences in the NCBI database.
With respect to nucleotide sequences, some may be entire genomic DNA sequences, some may be mRNAs, and some may be lower quality sequences such as expressed sequence tags (ESTs, which are derived from parts of mRNAs), or DNA sequences of contigs from genome projects.

NCBI database often contains redundant information for a gene, contains sequences of varying quality, and contains both uncurated and curated data.
As a result, NCBI has made a special database called RefSeq (reference sequence database), which is a subset of the NCBI database.
The data in RefSeq is manually curated, is high quality sequence data, and is non-redundant; this means that each gene (or splice-form of a gene, in the case of eukaryotes), protein, or genome sequence is only represented once.

NCBI
RefSeq

The data in RefSeq is curated and is of much higher quality than the rest of the NCBI Sequence Database.

RefSeq does not cover all species, and is not comprehensive for the species that are covered so far.

One can easily tell that a sequence comes from RefSeq because its accession number starts with particular sequence of letters.
That is, accessions of RefSeq sequences corresponding to protein records usually start with ‘NP_’, and accessions of RefSeq curated complete genome sequences usually start with ‘NC_’ or ‘NS_’.

Querying the NCBI Database

You may need to interrogate the NCBI Database to find particular sequences or a set of sequences matching given criteria, such as:

- The sequence with accession NC_001477
- The sequences published in *Nature* **460**:352-358
- All sequences from *Chlamydia trachomatis*
- Sequences submitted by Matthew Berriman
- Flagellin or fibrinogen sequences
- The glutamine synthetase gene from *Mycobacteriuma leprae*
- The upstream control region of the *Mycobacterium leprae dnaA* gene
- The sequence of the *Mycobacterium leprae* DnaA protein
- The genome sequence of *Trypanosoma cruzi*
- All human nucleotide sequences associated with malaria

There are two main ways that you can query the NCBI database to find these sets of sequences.
The first possibility is to carry out searches on the NCBI website.
The second possiblity is to carry out searches from R.

Querying the NCBI Database via the NCBI Website

If you are carrying out searches on the NCBI website, to narrow down your searches to specific types of sequences or to specific organisms, you will need to use "search tags".

For example, the search tags "[PROP]" and "[ORGN]" let you restrict your search to a specific subset of the NCBI Sequence Database, or to sequences from a particular taxon, respectively. Here is a list of useful search tags, which we will explain how to use below:

Search tag	Example	Restricts your search to sequences:
[AC]	NC_001477[AC]	With a particular accession number
[ORGN]	Fungi[ORGN]	From a particular organism or taxon
[PROP]	biomol_mRNA[PROP]	Of a specific type (eg. mRNA) or from a specific database (eg. RefSeq)
[JOUR]	Nature[JOUR]	Described in a paper published in a particular journal
[VOL]	531[VOL]	Described in a paper published in a particular journal volume
[PAGE]	27[PAGE]	Described in a paper with a particular start-page in a journal
[AU]	"Smith J"[AU]	Described in a paper, or submitted to NCBI, by a particular author

To carry out searches of the NCBI database, you first need to go to the NCBI website, and type your search query into the search box at the top.
For example, to search for all sequences from Fungi, you would type “Fungi[ORGN]” into the search box on the NCBI website.

You can combine the search tags above by using “AND”, to make more complex searches.
For example, to find all mRNA sequences from Fungi, you could type “Fungi[ORGN] AND biomol_mRNA[PROP]” in the search box on the NCBI website.

Likewise, you can also combine search tags by using “OR”, for example, to search for all mRNA sequences from Fungi or Bacteria, you would type “(Fungi[ORGN] OR Bacteria[ORGN]) AND biomol_mRNA[PROP]” in the search box. Note that you need to put brackets around “Fungi[ORGN] OR Bacteria[ORGN]” to specify that the word “OR” refers to these two search tags.

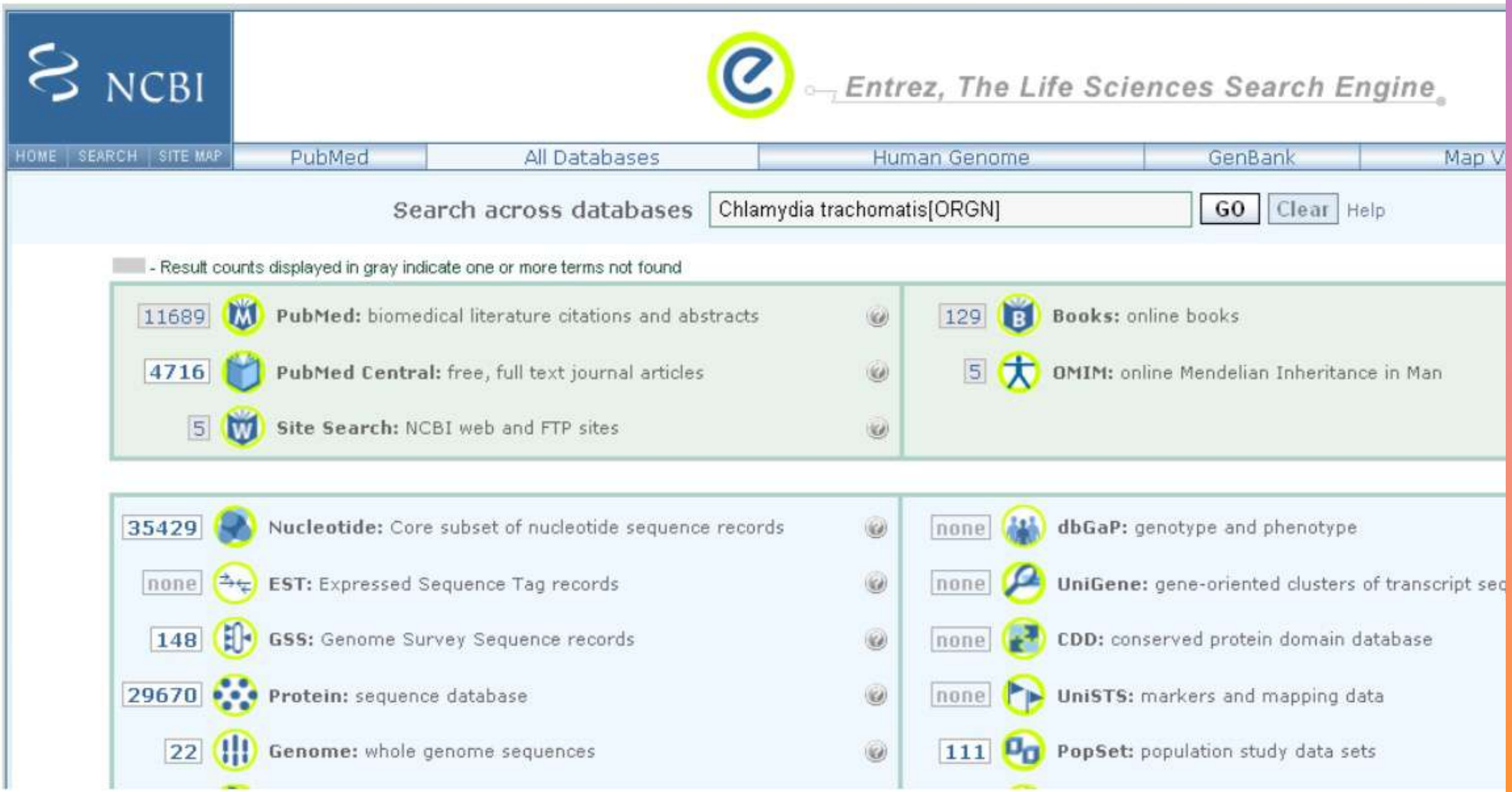
Examples of searches, some of them made by combining search terms using “AND”:

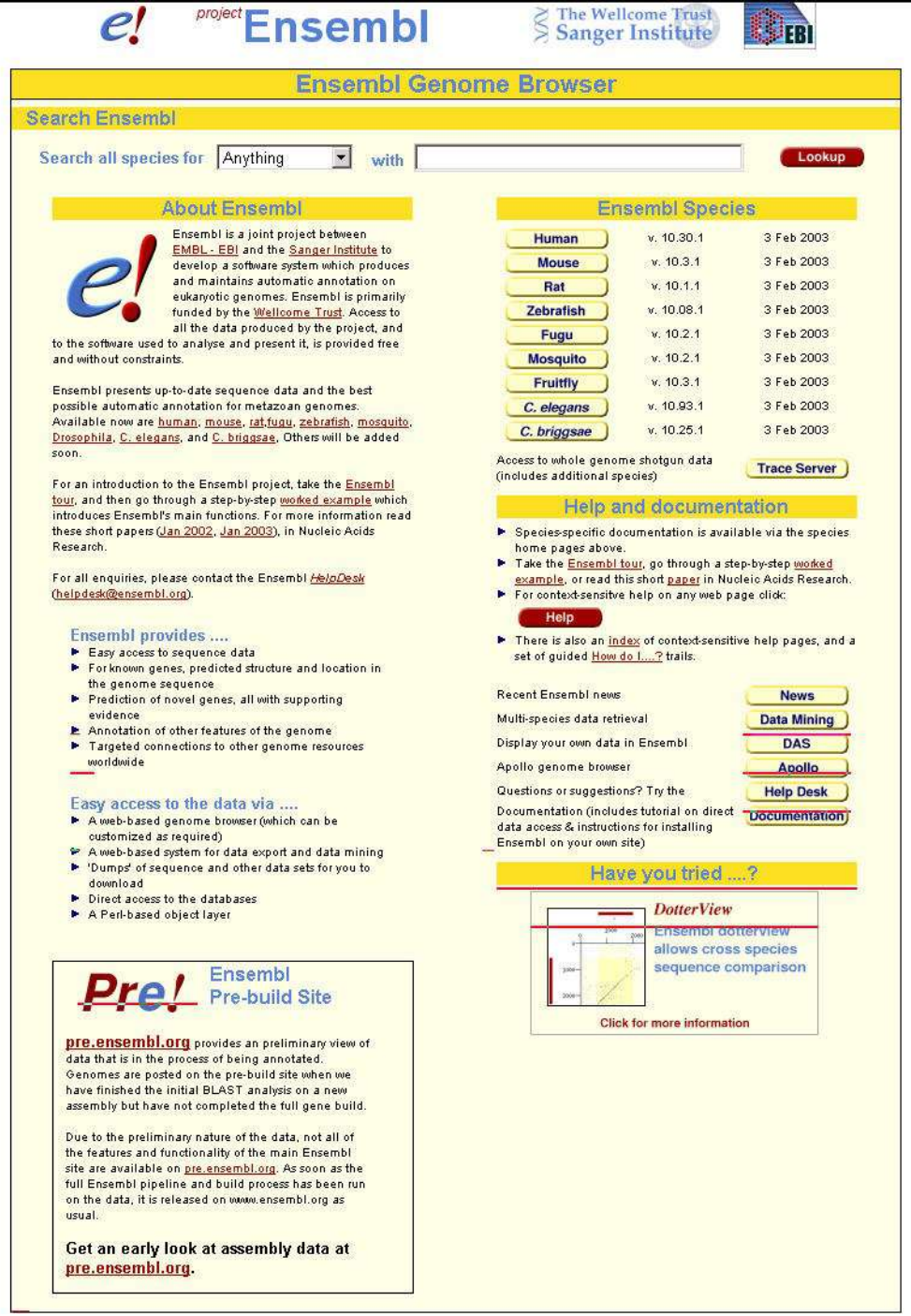
Typed in the search box	Searches for sequences:
NC_001477[AC]	With accession number NC_001477
Nature[JOUR] AND 460[VOL] AND 352[PAGE]	Published in *Nature* **460**:352–358
“Chlamydia trachomatis”[ORGN]	From the bacterium *Chlamydia trachomatis*
“Berriman M”[AU]	Published in a paper, or submitted to NCBI, by M. Berriman
flagellin OR fibrinogen	Which contain the word ‘flagellin’ or ‘fibrinogen’ in their NCBI record
“Mycobacterium leprae”[ORGN] AND dnaA	Which are from *M. leprae*, and contain “dnaA” in their NCBI record
“Homo sapiens”[ORGN] AND “colon cancer”	Which are from human, and contain “colon cancer” in their NCBI record
“Homo sapiens”[ORGN] AND malaria	Which are from human, and contain “malaria” in their NCBI record
“Homo sapiens”[ORGN] AND biomol_mrna[PROP]	Which are mRNA sequences from human
“Bacteria”[ORGN] AND srcdb_refseq[PROP]	Which are RefSeq sequences from Bacteria
“colon cancer” AND srcdb_refseq[PROP]	From RefSeq, which contain “colon cancer” in their NCBI record

Note that if you are searching for a phrase such as “colon cancer” or “Chlamydia trachomatis, you need to put the phrase in inverted commas when typing it into the search box. This is because if you type the phrase in the search box without using inverted commas, the search will be for NCBI records that contain either of the two words ‘colon’ or ‘cancer’ (or either of the two words ‘Chlamydia’ or ‘trachomatis’), not necessarily both words.

As mentioned above, the NCBI database contains several sub-databases, including the NCBI Nucleotide database and the NCBI Protein database. If you go to the NCBI website, and type one of the search queries above in the search box at the top of the page, the results page will tell you how many matching NCBI records were found in each of the NCBI sub-databases.

For example, if you search for “Chlamydia trachomatis[ORGN]”, you will get matches to proteins from C. trachomatis in the NCBI Protein database, matches to DNA and RNA sequences from C. trachomatis in the NCBI Nucleotide database, matches to whole genome sequences for C. trachomatis strains in the NCBI Genome database


NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map V
Search across databases
Chlamydia trachomatis[ORGN]
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
11689 PubMed: biomedical literature citations and abstracts
4716 PubMed Central: free, full text journal articles
5 Site Search: NCBI web and FTP sites
129 Books: online books
5 OMIM: online Mendelian Inheritance in Man
35429 Nucleotide: Core subset of nucleotide sequence records
none EST: Expressed Sequence Tag records
148 GSS: Genome Survey Sequence records
29670 Protein: sequence database
22 Genome: whole genome sequences
none dbGaP: genotype and phenotype
none UniGene: gene-oriented clusters of transcript seq
none CDD: conserved protein domain database
none UniSTS: markers and mapping data
111 PopSet: population study data sets



Ensembl provides a bioinformatics framework to organise biology around the sequences of large genomes.

Available now are:
human, mouse, rat, fugu, zebrafish, mosquito, Drosophila, *C. elegans*, and *C. briggsae*,

http://www.ensembl.org/

Swiss-Prot

Created in 1986 by Amos Bairoch

Collaboration between the Swiss Institute of Bioinformatic (Geneva, SIB/ISB, www.expasy.org) and the European Institute of Bioinformatic (EBI, UK, www.ebi.ac.uk)

- Manually curated sequences
- Manual annotation (based on scientific publications, personal communications, software tools, …)
- Traceable information (experimental qualifiers)
- Link between sequences and biological information

TrEMBL (TRanslation of EMBL)

- We cannot cope with the speed with which new data is coming out AND we do not want to dilute the quality of Swiss-Prot

  -> TrEMBL was created in 1996;

  Collaboration between the European Institute of Bioinformatic (EBI, UK, www.ebi.ac.uk) and Swiss Institute of Bioinformatic (Geneva).

- Contains all what is **not yet** in Swiss-Prot.

- Well-structured Swiss-Prot-like resource.

- The Universal Protein Resource Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins.

It consists of two sections:

Swiss -Prot

- Reviewed
- Manually annotated
- Records with information extracted from literature and curator-evaluated computational analysis.

Tr-EMBL (Transalted European Molecular Biological Labratory)

- Unreviewed
- Computationally annotated
- Records that await full manual annotation.

TrEMBL

Sp- TrEMBL
(SWISS PROT-TrEMBL)

contains sequences, which will eventually be incorporated into SWISS-PROT

REM-TrEMBL
(Remaining TrEMBL)

contains those sequences which will not be incorporated into SWISS-PROT.

For eg synthetic sequences, patent application sequences, fragments of less than 8 amino acids and coding sequences where there is strong experimental evidence that the sequence does not code for a real protein.

TrEMBL: A computer-annotated supplement to Swiss-PROT

- TrEMBL (translation of EMBL nucleotide sequence database) in 1996..

Why TrEMBL ?

- Increased data flow from genome projects to the sequence databases.
- To maintain the high annotation quality.
- To make sequences available as quickly as possible..
- TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the nucleotide sequence databases, except for CDS already included in Swiss-PROT.
- It also contains protein sequences extracted from the literature and protein sequences submitted directly by the user community.

Current Release Statistics

UniProtKB/TrEMBL PROTEIN DATABASE RELEASE 2021_03 STATISTICS

1. INTRODUCTION

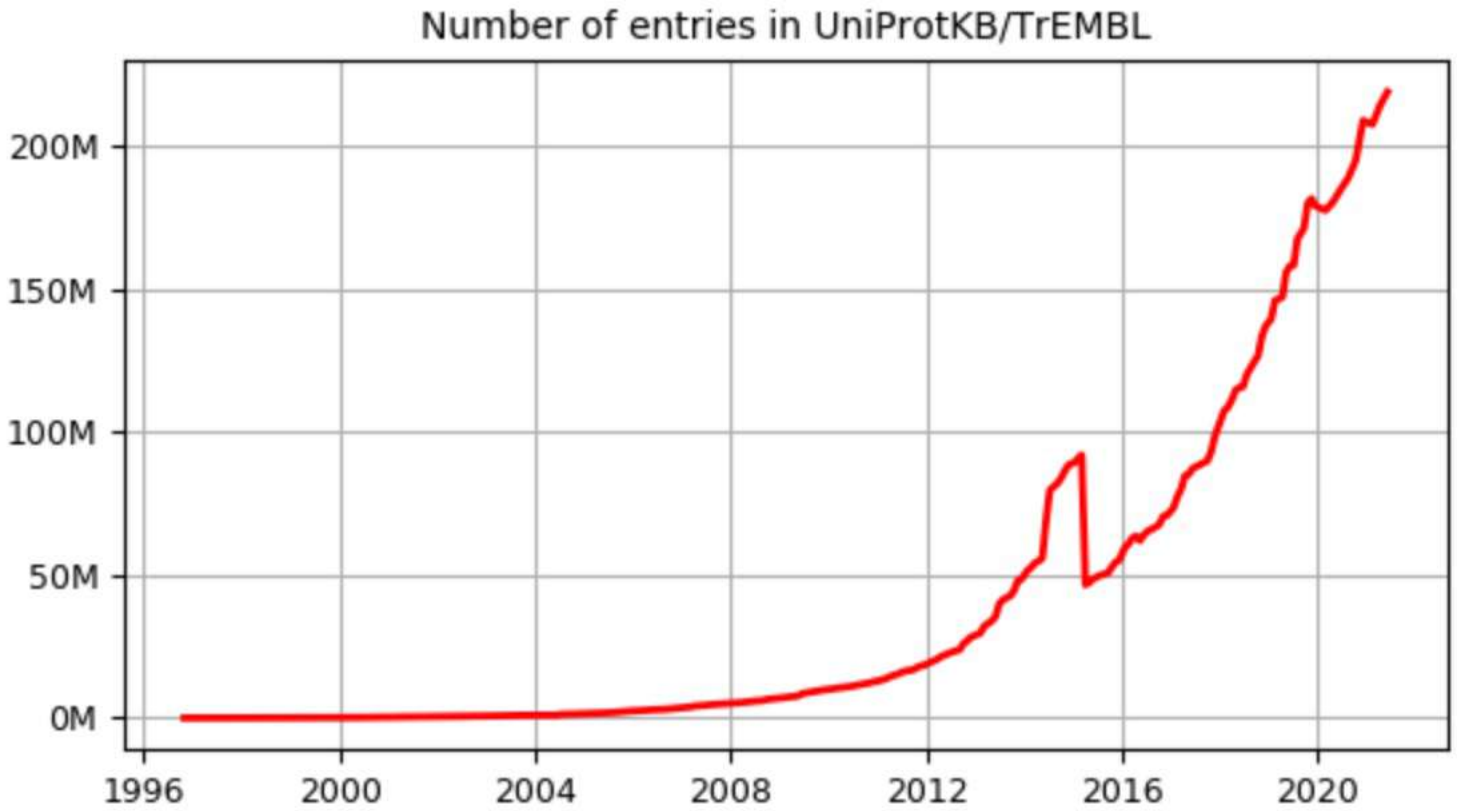
Release 2021_03 of 02-Jun-2021 of UniProtKB/TrEMBL contains 219174961 sequence entries, comprising 75271144009 amino acids.

14881314 sequences have been added since release 2021_02, the sequence data of 85669 existing entries has been updated and the annotations of 83429017 entries have been revised. This represents an increase of 7%.

Number of fragments: 23817622

Protein existence (PE):	entries	%
1: Evidence at protein level	174677	0.08%
2: Evidence at transcript level	1363132	0.62%
3: Inferred from homology	68206006	31.12%
4: Predicted	149431146	68.18%
5: Uncertain	0	0.00%

Growth of the database



Line type / subtype	Total number	Number of entries	Average per entry	Rank	Category
Cross-references (DR)	2562321326		11.69		
ABCD	397	397	<0.01	111	Protocols and materials databases
Allergome	3861	3150	<0.01	90	Protein family/group databases
Antibodypedia	74701	74609	<0.01	61	Protocols and materials databases
ArachnoServer	197	197	<0.01	123	Organism-specific databases
Araport	26808	26670	<0.01	74	Organism-specific databases
BMRB	322	322	<0.01	113	3D structure databases
BRENDA	17400	16635	<0.01	78	Enzyme and pathway databases
Bgee	483818	483760	<0.01	43	Gene expression databases
BindingDB	2248	2248	<0.01	100	Chemistry databases
BioCyc	775827	772073	<0.01	40	Enzyme and pathway databases
BioGRID	1	1	<0.01	143	Protein-protein interaction databases
BioGRID-ORCS	59690	59348	<0.01	64	Miscellaneous databases
BioMuta	964	964	<0.01	106	Genetic variation databases
CAZy	128703	120463	<0.01	53	Protein family/group databases
CDD	35077356	31773869	0.16	14	Family and domain databases
CGD	20791	20725	<0.01	77	Organism-specific databases
CLAE	444	444	<0.01	110	Protein family/group databases
COMPLUYEAST-2DPAGE	4	4	<0.01	138	2D gel databases
CORUM	255	255	<0.01	117	Protein-protein interaction databases
CPTAC	23	16	<0.01	131	Proteomic databases
CTD	1748970	1746725	0.01	35	Organism-specific databases
CarbonylDB	229	229	<0.01	119	PTM databases
ChEMBL	1156	1153	<0.01	105	Chemistry databases
ChiTaRS	173080	173078	<0.01	49	Miscellaneous databases
CollecTF	190	190	<0.01	124	Gene expression databases
ComplexPortal	226	176	<0.01	120	Protein-protein interaction databases
ConoServer	157	157	<0.01	126	Organism-specific databases
DIP	3090	3089	<0.01	93	Protein-protein interaction databases
DNASU	99813	99693	<0.01	54	Protocols and materials databases
DisProt	205	205	<0.01	121	Family and domain databases
DrugBank	764	453	<0.01	109	Chemistry databases
DrugCentral	201	201	<0.01	122	Chemistry databases
ELM	89	89	<0.01	127	Protein-protein interaction databases
EMBL	262238767	209159786	1.20	3	Sequence databases
EPD	12577	12577	<0.01	80	Proteomic databases
ESTHER	82788	82457	<0.01	59	Protein family/group databases
Ensembl	5219800	5069499	0.02	28	Genome annotation databases
EnsemblBacteria	86305412	79278265	0.39	9	Genome annotation databases
EnsemblFungi	5852464	5711940	0.03	27	Genome annotation databases
EnsemblMetazoa	1728503	1638699	0.01	36	Genome annotation databases
EnsemblPlants	3585504	3273742	0.02	31	Genome annotation databases
EnsemblProtists	1672596	1585634	0.01	37	Genome annotation databases
EvolutionaryTrace	5805	5805	<0.01	85	Miscellaneous databases
ExpressionAtlas	764097	764080	<0.01	41	Gene expression databases
FlyBase	28543	28475	<0.01	73	Organism-specific databases
GO	382224173	138766474	1.74	2	Ontologies

Gene3D	106513669	85961827	0.49	8	Family and domain databases
GeneCards	1313	1301	<0.01	102	Organism-specific databases
GeneDB	94292	92798	<0.01	56	Genome annotation databases
GeneID	13530551	13015459	0.06	23	Genome annotation databases
GeneTree	3947940	3947616	0.02	29	Phylogenomic databases
Genevisible	15463	15462	<0.01	79	Gene expression databases
GenomeRNAi	31803	31803	<0.01	72	Miscellaneous databases
GlyConnect	47	47	<0.01	130	PTM databases
GlyGen	15	15	<0.01	132	PTM databases
Gramene	3829294	3176039	0.02	30	Genome annotation databases
GuidetoPHARMACOLOGY	4	4	<0.01	139	Chemistry databases
HAMAP	22641435	22352120	0.10	16	Family and domain databases
HGNC	57003	56901	<0.01	65	Organism-specific databases
HOGENOM	17282013	17281575	0.08	20	Phylogenomic databases
IDEAL	9	9	<0.01	135	Family and domain databases
InParanoid	2143791	2143791	0.01	33	Phylogenomic databases
IntAct	23726	23726	<0.01	75	Protein-protein interaction databases
InterPro	566255232	166656803	2.58	1	Family and domain databases
KEGG	21017643	20570801	0.10	17	Genome annotation databases
LegioList	2496	2483	<0.01	96	Organism-specific databases
Leproma	1271	1269	<0.01	103	Organism-specific databases
MEROPS	308747	308746	<0.01	46	Protein family/group databases
MGI	63652	63263	<0.01	62	Organism-specific databases
MINT	2812	2812	<0.01	94	Protein-protein interaction databases
MalaCards	6	6	<0.01	137	Organism-specific databases
MaxQB	41248	41248	<0.01	71	Proteomic databases
MetOSite	335	335	<0.01	112	PTM databases
MoonDB	1	1	<0.01	144	Protein family/group databases
MoonProt	56	56	<0.01	129	Protein family/group databases
NIAGADS	259	259	<0.01	116	Organism-specific databases
OGP	3	3	<0.01	141	2D gel databases
OMA	9251047	9250807	0.04	26	Phylogenomic databases
OpenTargets	54876	54824	<0.01	67	Organism-specific databases
OrthoDB	18513604	18513482	0.08	18	Phylogenomic databases
PANTHER	48531595	46722291	0.22	12	Family and domain databases
PATRIC	14980819	14965950	0.07	21	Genome annotation databases
PCDDB	14	14	<0.01	133	3D structure databases
PDB	56630	23546	<0.01	66	3D structure databases
PDBsum	48624	21319	<0.01	70	3D structure databases
PHI-base	4684	4255	<0.01	87	Miscellaneous databases

PIR	161792	129576	<0.01	51	Sequence databases
PIRSF	17344544	17169148	0.08	19	Family and domain databases
PRIDE	339259	339259	<0.01	45	Proteomic databases
PRINTS	27472221	24609609	0.13	15	Family and domain databases
PRO	2265	2265	<0.01	99	Miscellaneous databases
PROSITE	108748834	71240783	0.50	7	Family and domain databases
PaxDb	248198	248198	<0.01	47	Proteomic databases
PeptideAtlas	162253	162253	<0.01	50	Proteomic databases
PeroxiBase	2581	2565	<0.01	95	Protein family/group databases
Pfam	215429775	152989899	0.98	4	Family and domain databases
PharmGKB	3110	3110	<0.01	92	Organism-specific databases
PhosphoSitePlus	2139	2139	<0.01	101	PTM databases
PhylomeDB	424017	424017	<0.01	44	Phylogenomic databases
PlantReactome	1213	870	<0.01	104	Enzyme and pathway databases
PomBase	2	2	<0.01	142	Organism-specific databases
ProMEX	2385	2385	<0.01	98	Proteomic databases
Proteomes	214900921	193968656	0.98	5	Miscellaneous databases
ProteomicsDB	93750	93639	<0.01	57	Proteomic databases
PseudoCAP	4370	4366	<0.01	88	Organism-specific databases
REBASE	94323	88538	<0.01	55	Protein family/group databases
REPRODUCTION-2DPAGE	62	61	<0.01	128	2D gel databases
RGD	21962	21055	<0.01	76	Organism-specific databases
RNAct	2494	2494	<0.01	97	Miscellaneous databases
Reactome	137083	47567	<0.01	52	Enzyme and pathway databases
RefSeq	52689555	50944317	0.24	11	Sequence databases
SABIO-RK	839	839	<0.01	108	Enzyme and pathway databases
SASBDB	158	158	<0.01	125	3D structure databases

SFLD	1656325	1330490	0.01	38	Family and domain databases
SGD	7	7	<0.01	136	Organism-specific databases
SIGNOR	4	4	<0.01	140	Enzyme and pathway databases
SMART	53380818	39925689	0.24	10	Family and domain databases
SMR	2206900	2206900	0.01	32	3D structure databases
STRING	12248519	12248219	0.06	25	Protein-protein interaction databases
SUPFAM	142359705	111979057	0.65	6	Family and domain databases
SWISS-2DPAGE	1	1	<0.01	145	2D gel databases
SignaLink	4150	4150	<0.01	89	Enzyme and pathway databases
SwissLipids	9	9	<0.01	134	Chemistry databases
SwissPalm	3392	3392	<0.01	91	PTM databases
TAIR	11574	11513	<0.01	81	Organism-specific databases
TCDB	8551	8533	<0.01	83	Protein family/group databases
TIGRFAMs	42872216	39446949	0.20	13	Family and domain databases
TopDownProteomics	272	272	<0.01	115	Proteomic databases
TreeFam	509782	509736	<0.01	42	Phylogenomic databases
TubercuList	943	942	<0.01	107	Organism-specific databases
UCSC	90930	90666	<0.01	58	Genome annotation databases
UniLectin	236	236	<0.01	118	Protein family/group databases
UniPathway	14185068	13131824	0.06	22	Enzyme and pathway databases
VEuPathDB	1861894	1692427	0.01	34	Organism-specific databases
VGNC	231072	231008	<0.01	48	Organism-specific databases
WBParaSite	939692	928304	<0.01	39	Genome annotation databases
World-2DPAGE	311	306	<0.01	114	2D gel databases
WormBase	62080	61706	<0.01	63	Organism-specific databases
Xenbase	52361	50722	<0.01	69	Organism-specific databases
ZFIN	54127	53953	<0.01	68	Organism-specific databases
dictyBase	7985	7763	<0.01	84	Organism-specific databases
eggNOG	13519462	13065588	0.06	24	Phylogenomic databases
euHCVdb	75267	75264	<0.01	60	Organism-specific databases
iPTMnet	5189	5189	<0.01	86	PTM databases
jPOST	11348	11348	<0.01	82	Proteomic databases

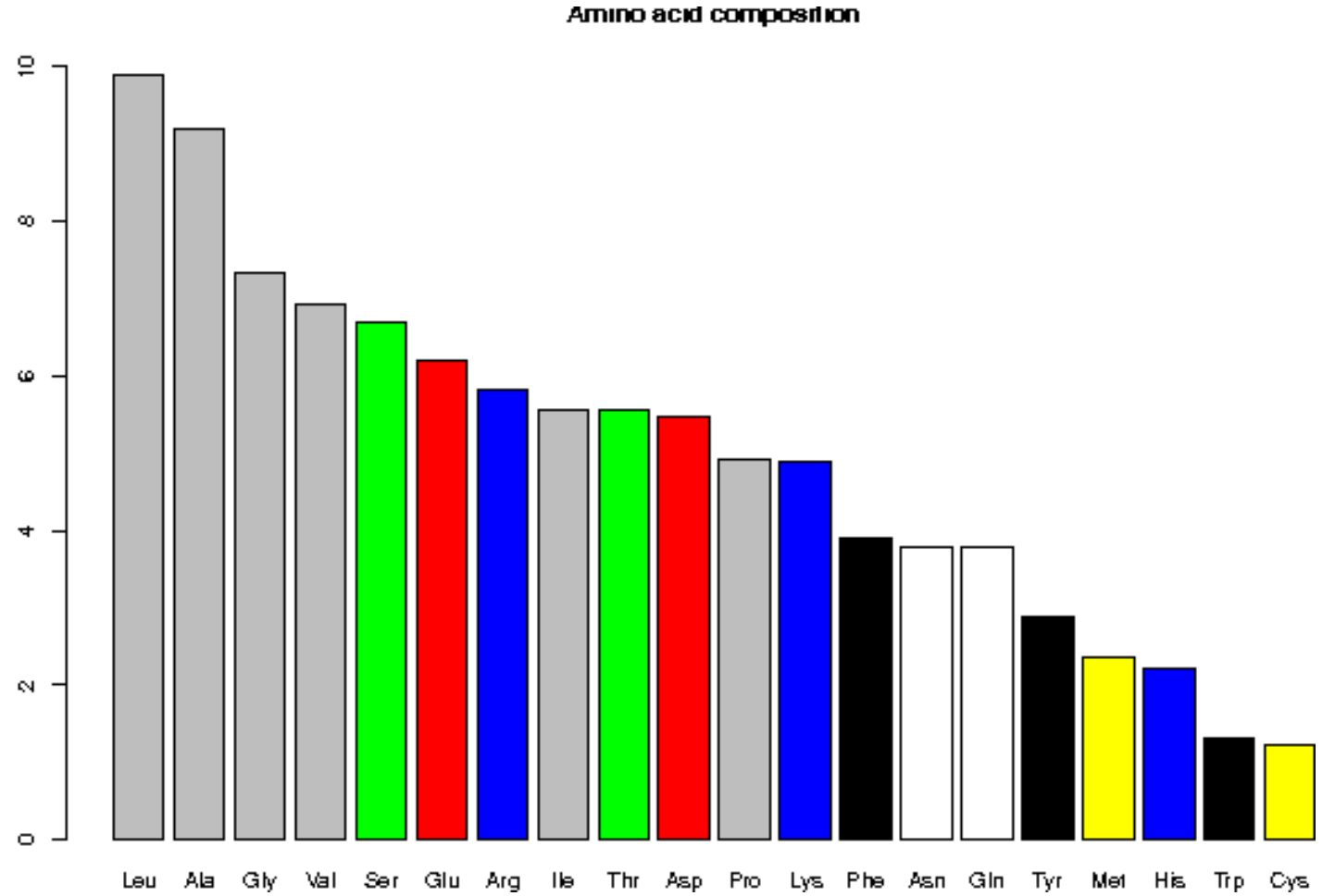
Line type / subtype	Total number	Number of entries	Average per entry	Rank
Features (FT)	685140936		3.13	
ACT_SITE	14276311	8759683	0.07	11
BINDING	30217691	8515268	0.14	6
CARBOHYD	53186	39872	<0.01	25
CHAIN	16030018	15813913	0.07	9
COILED	27462602	18384916	0.13	7
COMPBIAS	58972775	25535713	0.27	4
CROSSLNK	65593	60229	<0.01	24
DISULFID	4062562	1086932	0.02	15
DNA_BIND	1681123	1650919	0.01	18
DOMAIN	157295983	111039362	0.72	2
INIT_MET	82741	82740	<0.01	22
INTRAMEM	2072	1784	<0.01	27
LIPID	441563	255749	<0.01	21
METAL	24524331	6329602	0.11	8
MOD_RES	4188048	3709291	0.02	14
MOTIF	2357359	1615074	0.01	17
NON_STD	17306	16496	<0.01	26
NON_TER	36674267	25065264	0.17	5
NP_BIND	12140520	7539026	0.06	12
PEPTIDE	1183	850	<0.01	28
PROPEP	71112	71112	<0.01	23
REGION	80580849	47306559	0.37	3
REPEAT	9293032	2163870	0.04	13
SIGNAL	15523339	15523326	0.07	10
SITE	3708768	2270669	0.02	16
TOPO_DOM	465062	212350	<0.01	20
TRANSIT	240	240	<0.01	29
TRANSMEM	184188306	40548580	0.84	1
ZN_FING	762994	587921	<0.01	19

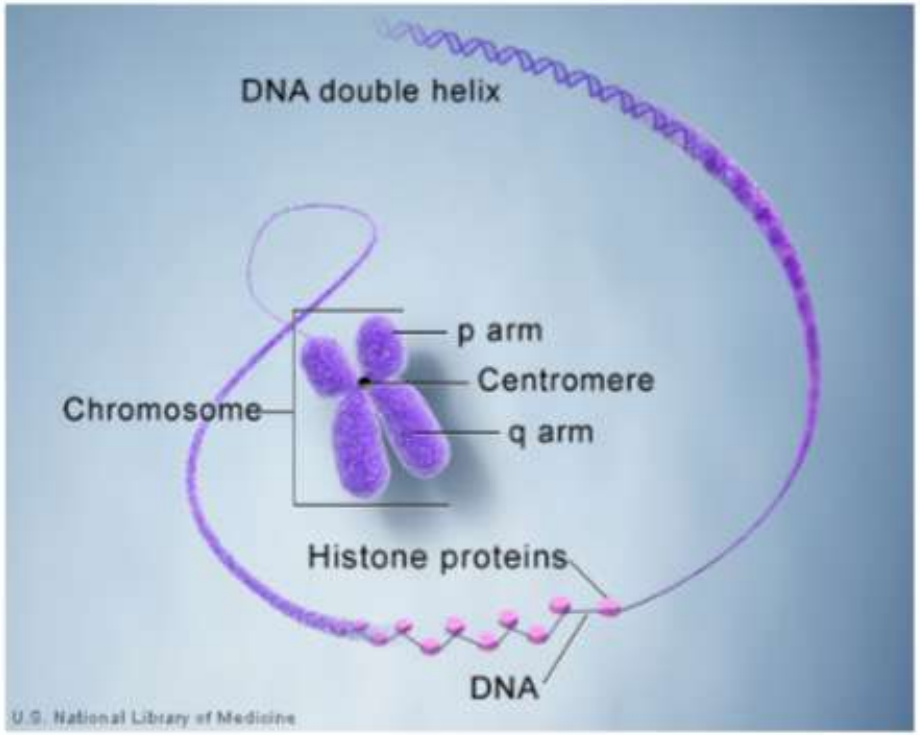
Total number of feature keys: 29

AMINO ACID COMPOSITION

5.1 Composition in percent for the complete database

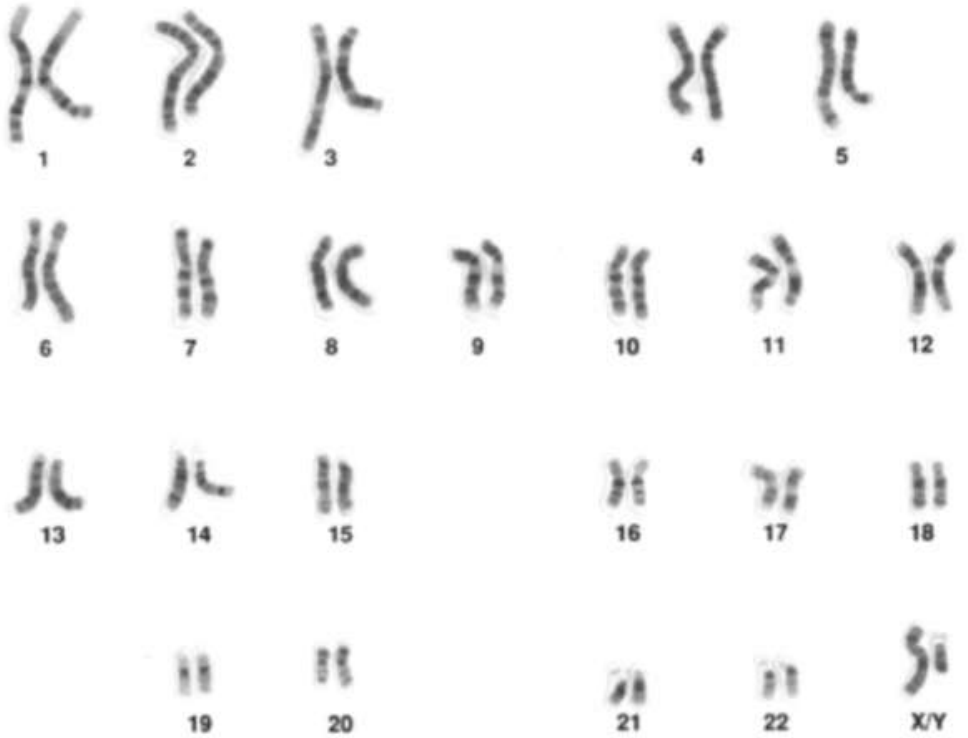
Ala (A) 9.20	Gln (Q) 3.77	Leu (L) 9.90	Ser (S) 6.68
Arg (R) 5.83	Glu (E) 6.20	Lys (K) 4.89	Thr (T) 5.54
Asn (N) 3.77	Gly (G) 7.34	Met (M) 2.35	Trp (W) 1.30
Asp (D) 5.48	His (H) 2.20	Phe (F) 3.89	Tyr (Y) 2.88
Cys (C) 1.23	Ile (I) 5.56	Pro (P) 4.93	Val (V) 6.93





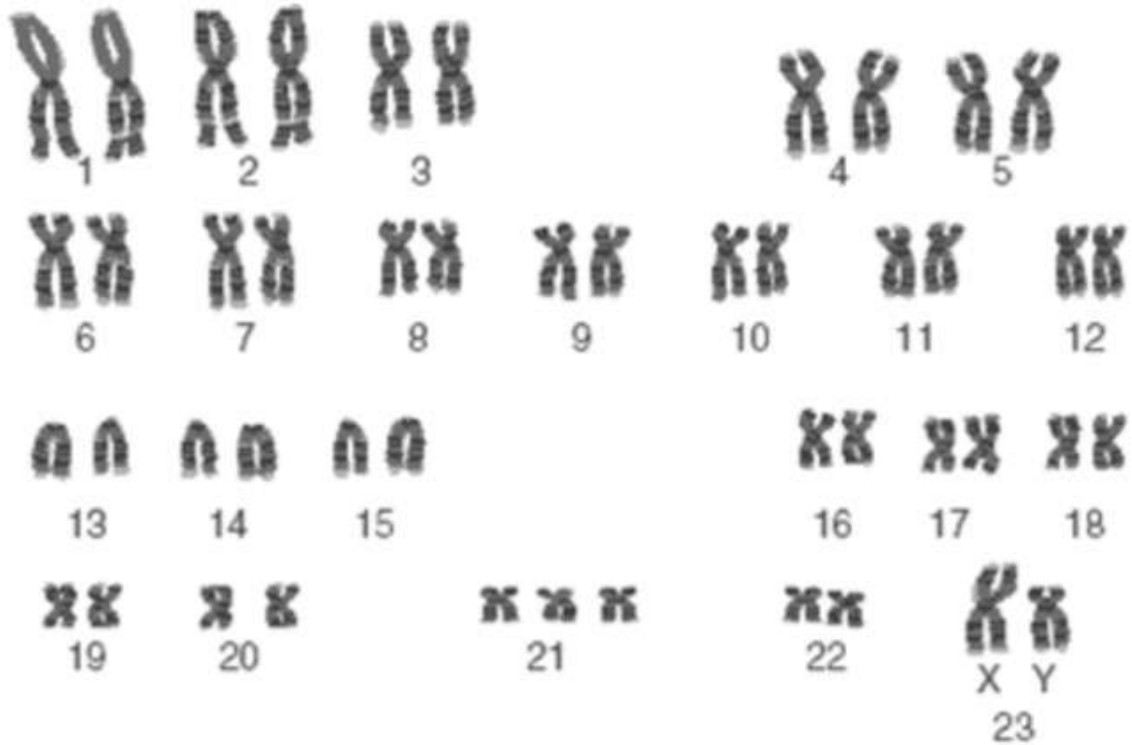
- Chromosomes contain DNA wrapped around proteins known as histones
- Chromosomes have two identical parts known as chromatids
- The point where the two chromatids touch is the centromere
- The p-arm is also known as the short arm
- The q-arm is also known as the long arm

Karyotype



- The complete set of chromosomes in the cells of an organism is its karyotype

What is different in this karyotype?

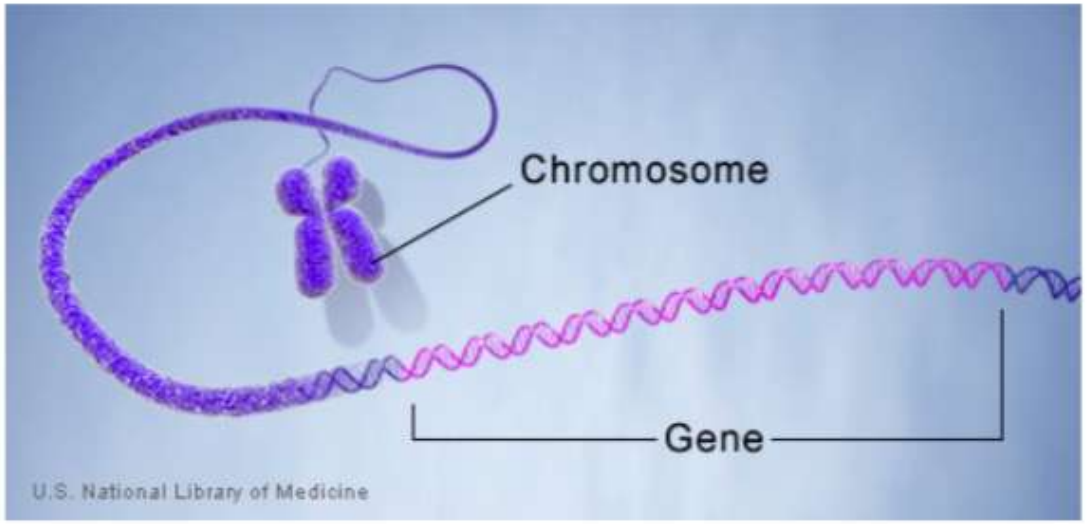


Trisomy 21, aka Down Syndrome

Number of chromosomes in different species

Homo sapiens (human)	46
Mus musculus (house mouse)	40
Drosophila melanogaster (fruit fly)	8
Caenorhabditis elegans (microscopic roundworm)	12
Saccharomyces cerevisiae (budding yeast)	32
Arabidopsis thaliana (plant in the mustard family)	10
Xenopus laevis (South African clawed frog)	36
Canis familiaris (domestic dog)	78
Gallus gallus (chicken)	78
Zea mays (corn or maize)	20
Muntiacus reevesi (the Chinese muntjac, a deer)	23
Muntiacus muntjac (its Indian cousin)	6
Myrmecia pilosula (an ant)	2
Parascaris equorum var. univalens (parasitic roundworm)	2
Cambarus clarkii (a crayfish)	200
Equisetum arvense (field horsetail, a plant)	216

The gene



Definition

The gene is the basic unit of inheritance and specifies a trait. Genes are just a portion of the DNA molecule (on average, 1000bp long), and often – but not always! – code for a protein

GENE

“The gene is the basic physical unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify traits. Genes are arranged, one after another, on structures called chromosomes. A chromosome contains a single, long DNA molecule, only a portion of which corresponds to a single gene. Humans have approximately 20,000 genes arranged on their chromosomes.”

Mutations in genes can lead to genetic diseases.

Some examples are: CFTR (cystic fibrosis), breast cancer (BRCA1), sickle cell anemia (hemoglobin gene), ...

In cancer, most mutations are acquired during the lifetime of the individual (that is, they are not inherited). BRCA1 is a good exception to this.

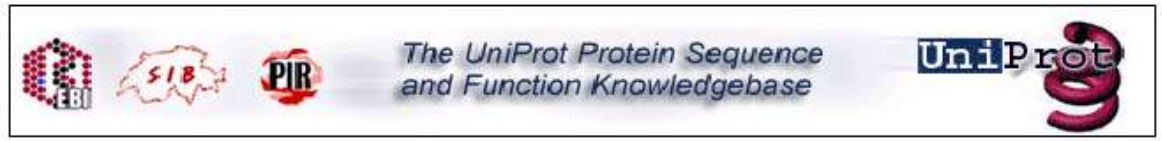
Bioinformatics approaches play an important role in helping understand the effect of mutations on gene function and disease

United Protein Databases

Site Map Site Search

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support



The United Protein Databases (UniProt) project will create a central database of protein sequence and function by joining the forces of the SWISS-PROT, TrEMBL and PIR protein database activities.

The project is funded by the U.S. National Human Genome Research Institute (NHGRI), in cooperation with five other institutes and centers at the National Institutes of Health (NIH) (Grant Number: 1 U01 HG02712-01). Read the Press release.

The broad, long-term objectives of this project are:

- To provide a stable and comprehensive resource for information on proteins, their sequences and their functions.
- To enable scientists to use these data to identify and analyze genes and their products and to make queries across databases containing complementary information.
- To provide efficient and unencumbered access to the Database.

The specific aims are:

- To develop and maintain a central database of curated protein sequences with annotations of sequence and functional information.
- To facilitate use of the database by providing user-friendly interfaces, tools for simple and complex queries and for retrieval of large datasets, down-loadable database records in defined, parsable format, and user support services.
- To provide the flexibility and adaptability needed to be responsive to the changing needs of the scientific community.

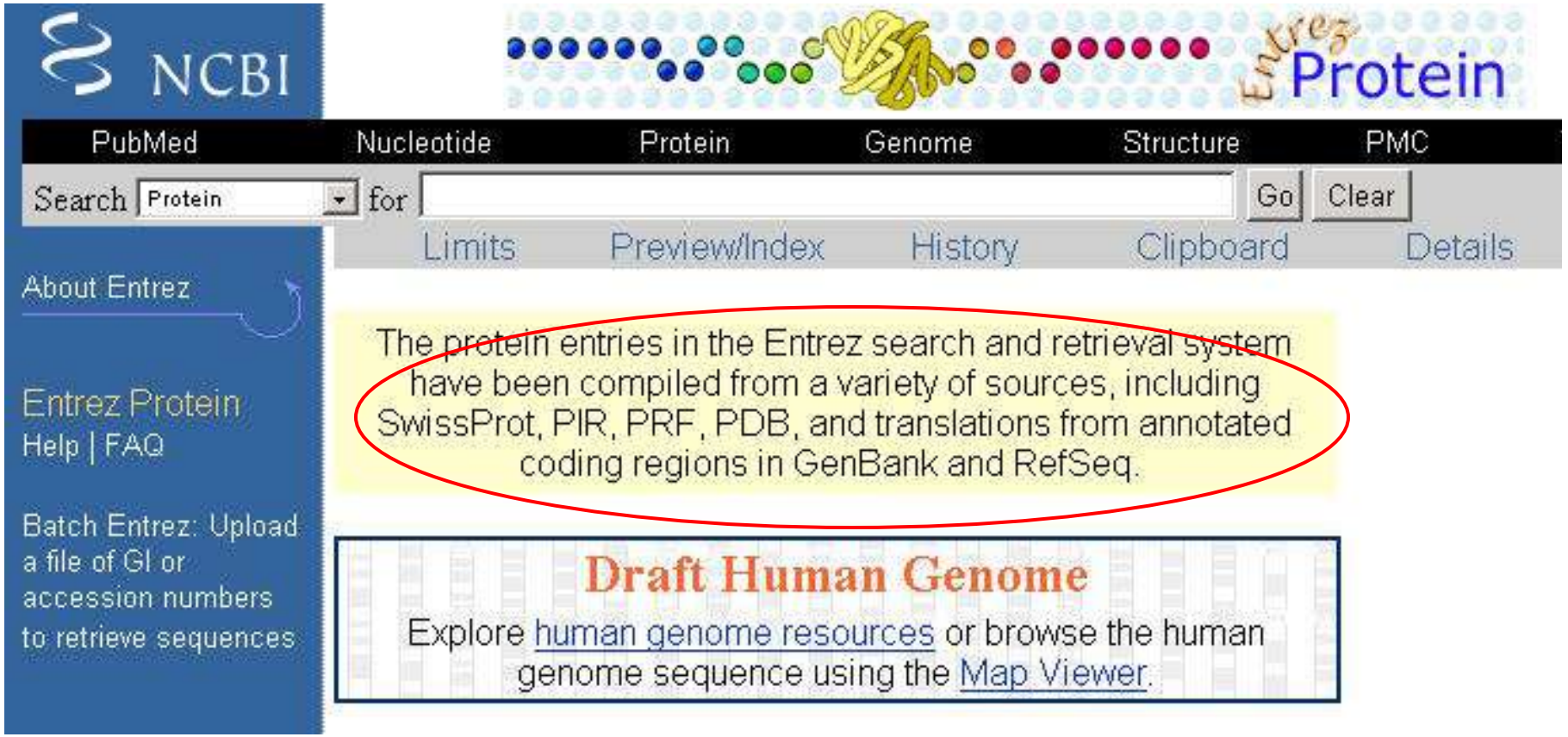
Performance Sites:

- **European Bioinformatics Institute (EBI)** - Hinxton, Cambridge, UK
- **Protein Information Resource (PIR)** - Georgetown University Medical Center (GUMC) & National Biomedical Research Foundation (NBRF), Washington, D.C., USA
- **Swiss Institute of Bioinformatics (SIB)** - Geneva, Switzerland

http://pir.georgetown.edu/uniprot/

Protein sequences: « NR database »

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein



Some important remarks

The AC number jungle

Type of record	Sample Accession Format
GenBank/EMBL/DDBJ	One letter followed by five digits: e.g. U12345 Two letters followed by 6 digits: e.g. AF123456
Swiss-Prot/TrEMBL	One letter (O, P, Q) and five digits/letters: e.g. P12345
RefSeq nucleotide	Two letters, underscore bar and six digit: e.g. mRNA NM_000492 e.g. genomic NT_000907
RefSeq protein	e.g. NP_00483
RefSeq prediction	e.g. XM_000483 e.g. XP_000467
PDB (protein structure)	One digit followed by three letters: e.g. 1TUP

Swiss-Prot / TrEMBL: a minimal of redundancy

Human EPO: Blastp against Swiss-Prot/TrEMBL

sp	P01588	EPO_HUMAN Erythropoietin precursor (Epoetin) [EPO] [Ho...	389	e-107
tn	AAP22357	Hypothetical protein EPO [EPO] [Homo sapiens (Human)]	389	e-107
sp	P07865	EPO_MACFA Erythropoietin precursor [EPO] [Macaca fasci...	353	7e-97
sp	Q28513	EPO_MACMU Erythropoietin precursor [EPO] [Macaca mulat...	351	3e-96
tr	Q867B1	Erythropoietin [EPO] [Equus caballus (Horse)]	318	3e-86
sp	P33708	EPO_FELCA Erythropoietin precursor [EPO] [Felis silves...	312	1e-84
sp	P29676	EPO_RAT Erythropoietin precursor [EPO] [Rattus norvegi...	303	1e-81
tr	Q9MYM8	Erythropoietin precursor [EPO] [Sus scrofa (Pig)]	302	1e-81
sp	P33709	EPO_SHEEP Erythropoietin precursor [EPO] [Ovis aries (...	300	6e-81
sp	P49157	EPO_PIG Erythropoietin precursor (Fragment) [EPO] [Sus...	298	3e-80
sp	P07321	EPO_MOUSE Erythropoietin precursor [EPO] [Mus musculus...	298	3e-80
sp	P48617	EPO_BOVIN Erythropoietin precursor [EPO] [Bos taurus (...	295	2e-79
tr	Q9GKA3	Erythropoietin [Oryctolagus cuniculus (Rabbit)]	288	4e-77
tr	Q9GKA2	Erythropoietin [Oryctolagus cuniculus (Rabbit)]	287	5e-77
sp	P33707	EPO_CANFA Erythropoietin precursor (Fragment) [EPO] [C...	282	2e-75
tr	Q8HZ88	Erythropoietin (Fragment) [Gorilla gorilla (gorilla)]	259	1e-68
tr	Q8HZ89	Erythropoietin (Fragment) [Pan troglodytes (Chimpanzee)]	258	4e-68
tr	Q8HZ87	Erythropoietin (Fragment) [Pongo pygmaeus (Orangutan)]	245	3e-64
tr	Q8HZ86	Erythropoietin (Fragment) [Macaca sp]	238	3e-62
tr	Q8HZ85	Erythropoietin (Fragment) [Saguinus oedipus (Cotton-to...	216	1e-55

Human EPO: BlastP against NR

→	gi|119526|sp|P01588|EPO_HUMAN Erythropoietin precursor (Epo...	300	7e-81	L
→	gi|312304|emb|CAA26095.1| erythropoietin [Homo sapiens]	299	1e-80	L
→	gi|4503589|ref|NP_000790.1| erythropoietin [Homo sapiens] >...	296	1e-79	L
→	gi|5822016|pdb|1CN4|C Chain C, Erythropoietin Complexed Wit...	285	3e-76	S
→	gi|6137383|pdb|1EER|A Chain A, Crystal Structure Of Human E...	283	9e-76	S
	gi|119527|sp|P07865|EPO_MACFA Erythropoietin precursor >gi|...	275	4e-73	
	gi|2494368|sp|Q28513|EPO_MACMU Erythropoietin precursor >gi...	272	2e-72	
	gi|27806897|ref|NP_776334.1| erythropoietin [Bos taurus] >g...	245	4e-64	L
	gi|21389309|ref|NP_031968.1| erythropoietin [Mus musculus] ...	244	5e-64	L
	gi|8393316|ref|NP_058697.1| erythropoietin [Rattus norvegic...	243	1e-63	L
	gi|204061|gb|AAA41126.1| erythropoietin	243	2e-63	L
→	gi|4261664|gb|AAD13964.1|S65458_1 mutant erythropoietin [Ho...	242	2e-63	
	gi|462017|sp|P33709|EPO_SHEEP Erythropoietin precursor >gi|...	241	6e-63	
	gi|90481|pir||A24902 erythropoietin precursor - mouse	240	8e-63	
	gi|165877|gb|AAA31518.1| erythropoietin	238	4e-62	
	gi|1352380|sp|P49157|EPO_PIG Erythropoietin precursor >gi|2...	236	2e-61	
	gi|8920355|emb|CAB96416.1| erythropoietin [Sus scrofa] >gi|...	236	2e-61	
	gi|25294239|pir||JC7699 erythropoietin - rabbit >gi|1152728...	232	3e-60	
	gi|11527287|gb|AAG36962.1| erythropoietin [Oryctolagus cuni...	231	3e-60	
	gi|2144696|pir||I46083 erythropoietin precursor - cat (frag...	231	4e-60	
	gi|1706680|sp|P33708|EPO_FELCA Erythropoietin precursor >gi...	229	2e-59	
	gi|23379782|gb|AAM76633.1| erythropoietin [Gorilla gorilla]	228	4e-59	
	gi|23379780|gb|AAM76632.1| erythropoietin [Pan troglodytes]	228	6e-59	
	gi|27807634|dbj|BAC55239.1| erythropoietin [Equus caballus]	225	3e-58	
	gi|23379784|gb|AAM76634.1| erythropoietin [Pongo pygmaeus]	218	3e-56	
	gi|23379786|gb|AAM76635.1| erythropoietin [Macaca sp.]	212	2e-54	
	gi|28521143|ref|XP_282615.1| erythropoietin [Mus musculus] ...	206	2e-52	L
	gi|462015|sp|P33707|EPO_CANFA Erythropoietin precursor >gi|...	205	4e-52	
	gi|23379788|gb|AAM76636.1| erythropoietin [Saguinus oedipus]	200	9e-51	
→	gi|6573159|gb|AAF17572.1|AF202311_1 erythropoietin [Homo sa...	190	1e-47	
→	gi|6671275|gb|AAF23133.1| erythropoietin [Homo sapiens]	114	8e-25	
	gi|460893|gb|AAB29659.1| erythropoietin, Epo (N-terminal) [...	70	3e-11	

Righting the wrongs

"Sequences are rarely deposited in a "mature" state; as with all scientific research, DNA and protein annotation is a continual process of learning, revision and corrections."

"Sequencing error rates: ~1 base in 10'000"

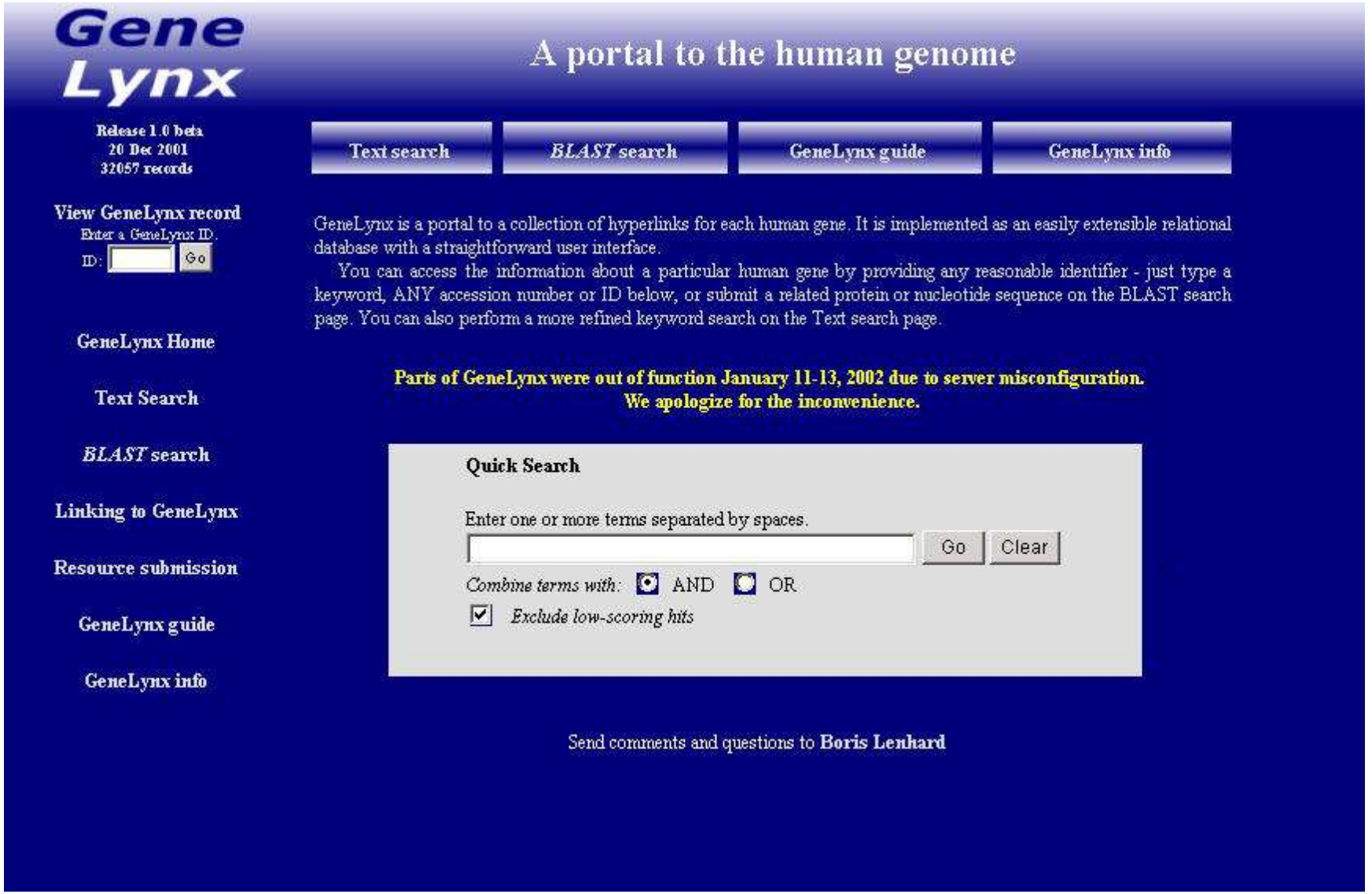
Databases 3: ‘genomics’

- Contain informations on gene chromosomal location (mapping) and nomenclature, and provide links to sequence databases; *has usually no sequence*;
- Exist for most organisms important in life science research; usually species specific.

- Examples: MIM, GDB (human), MGD (mouse), FlyBase (Drosophila), SGD (yeast), MaizeDB (maize), SubtiList (B.subtilis), etc.;

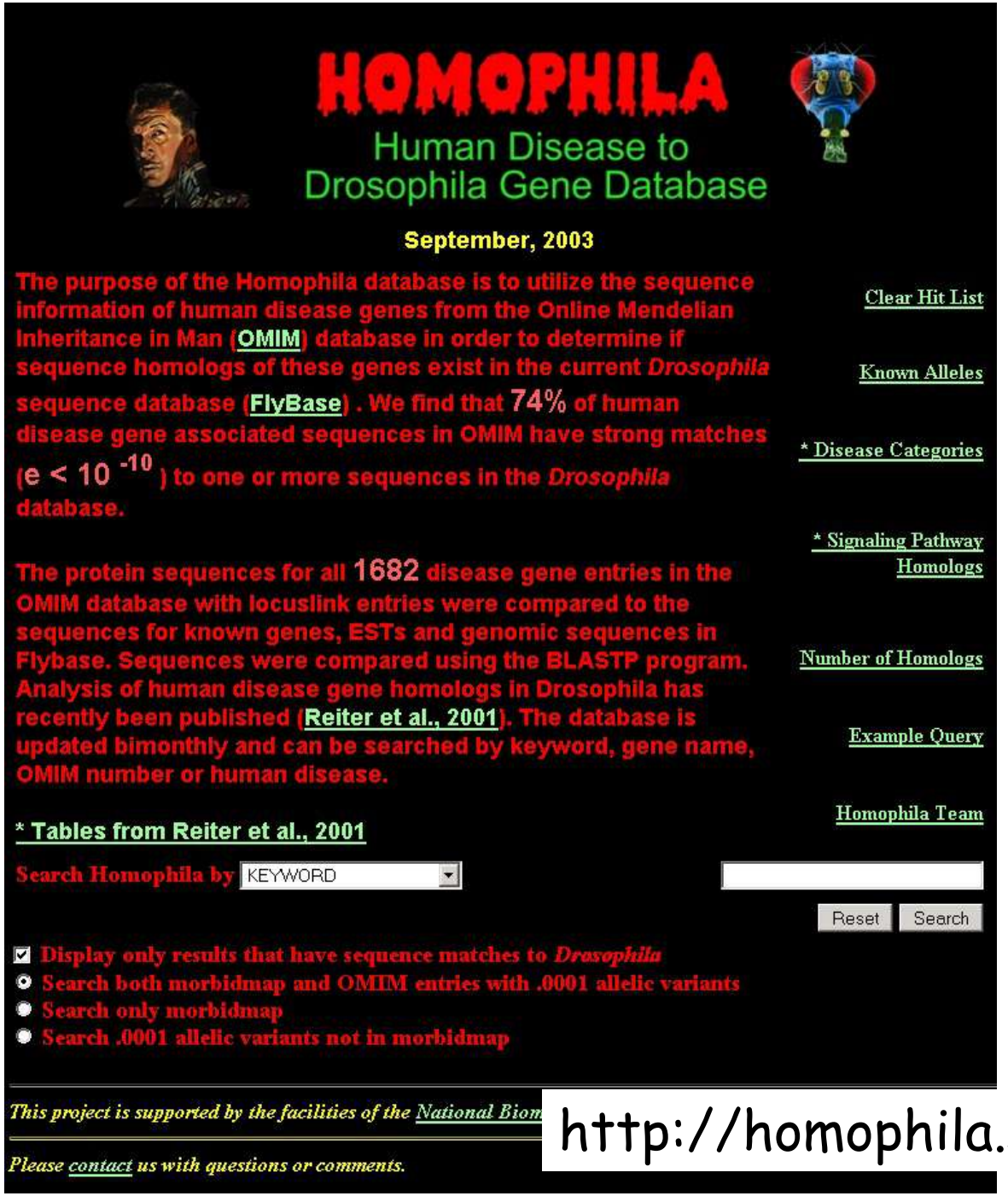
OMIM (http://www.ncbi.nlm.nih.gov/Omim/)

- OMIM™ : Online Mendelian Inheritance in Man
- catalog of human genes and genetic disorders
- contains a summary of literature and reference information. It also contains links to publications and sequence information.



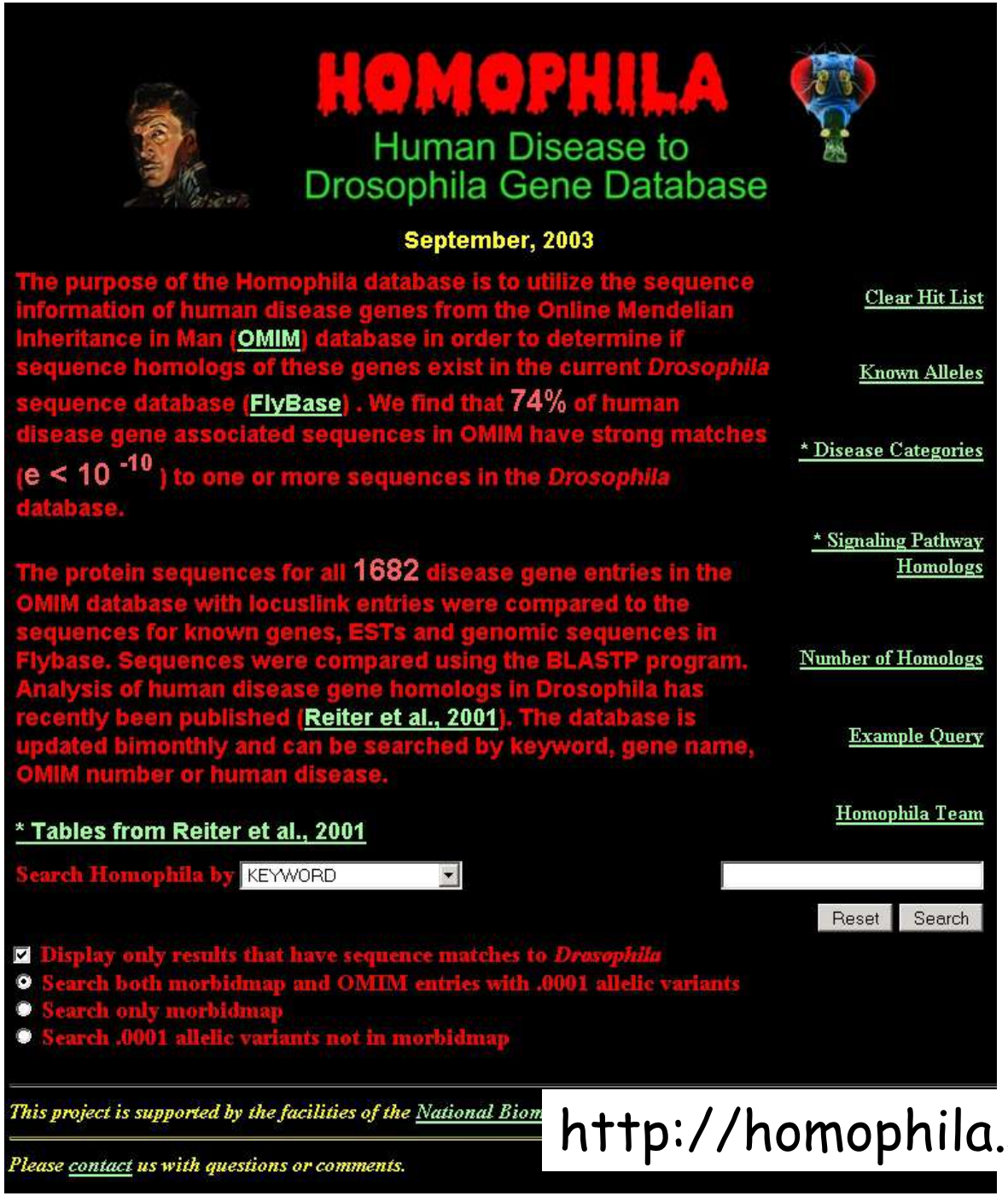
http://www.genelynx.org/

Collections of hyperlinks for each human gene



http://homophila.sdsc.edu/

Collections of hyperlinks for each human gene



http://homophila.sdsc.edu/

Databases 4: mutation/polymorphism

- Contain informations on sequence variations linked or not to genetic diseases;
- Mainly human but: OMIA - Online Mendelian Inheritance in Animals
- **General db**:
  - OMIM
  - HMGD - Human Gene Mutation db
  - SVD - Sequence variation db
  - HGBASE - Human Genic Bi-Allelic Sequences db
  - dbSNP - Human single nucleotide polymorphism (SNP) db
- **Disease-specific db**: most of these databases are either linked to a single gene or to a single disease;
  - p53 mutation db
  - ADB - Albinism db (Mutations in human genes causing albinism)
  - Asthma and Allergy gene db
  - ....

For human (Amos'link)

- HGMD - Human Gene Mutation db
- SVD - EBI Sequence variation db
- HGBASE - Human Genic Bi-Allelic Sequences db
- The SNP consortium
- dbSNP - Human single nucleotide polymorphism (SNP) db

- ALFRED - Allele Frequency Db
- SeattleSNPs - UW-FHCRC Variation Discovery Resource
- PicSNP - Catalog of non-synonymous SNP

- List of mutation databases from OMIM
- List of mutation databases from IMT (Finland)

! Numbering of the mutated amino acid depends on the db (aa no 1 is not necessary the initiator Met !)

Mutation/polymorphism

The SNP consortium (TSC) http://snp.cshl.org/

- Public/private collaboration: Bayer, Roche, IBM, Pfizer, Novartis, Motorola……
- Has to date discovered and characterized nearly 1.5 million SNPs; in addition, the allele frequencies in three major world populations have been determined on a subset of ~57,000 SNPs.

SNPs dbSNP at NCBI http://www.ncbi.nlm.nih.gov/SNP/

- **Collaboration between the National Human Genome Research Institute and the National Center for Biotechnology Information (NCBI)**
- **Mission: central repository for both single base nucleotide subsitutions and short deletion and insertion polymorphisms (several species)**
- **Sept 2003, dbSNP has submissions for 9'557'000 SNPs.**

Chromosome 21 dbSNP http://csnp.isb-sib.ch/

- A joint project between the Division of Medical Genetics of the University of Geneva Medical School and the SIB
- Mission: comprehensive cSNP (Single Nucleotide Polymorphisms within cDNA sequences) database and map of chromosome 21

Database 5: protein domain/family

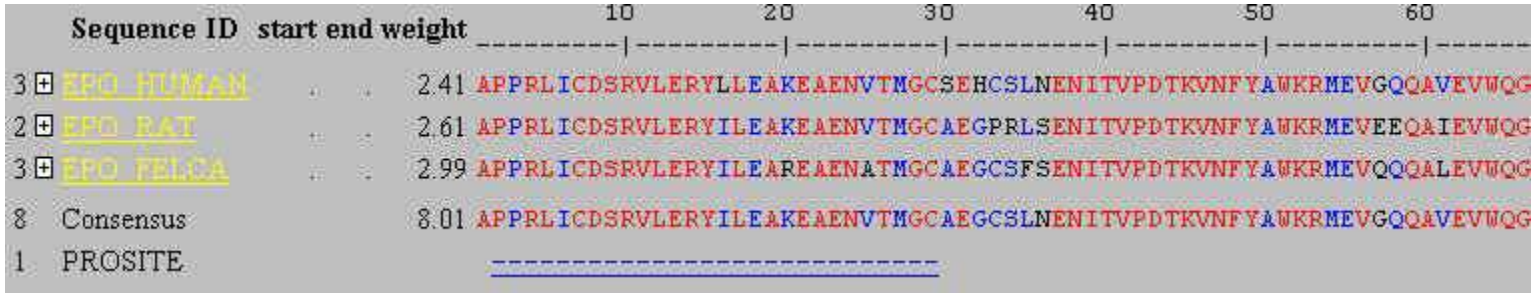
Protein domain/family: some definitions

- Most proteins have « modular » structures
- Estimation: ~ 3 domains / protein
- Protein domains are ideally defined by a specific combination of secondary structures that fold into a characteristic three dimensional (3D) structure.
- Domains not only share a common structure but have also often a similar function that contributes to the global activity of the proteins which contain them.

Protein interaction domains for different type of proteins

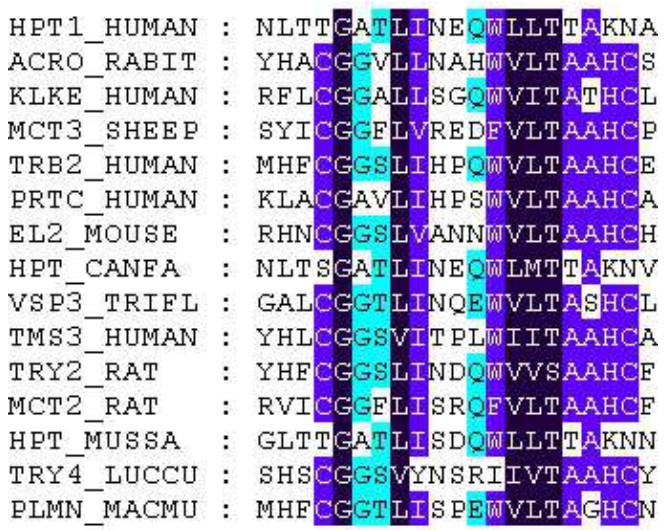
- Domains are more commonly defined as a region of sequence homology

- Domains are identified by multiple sequence alignments



- Domains can be defined by different methods:
  - **Pattern** (regular expression); used for very conserved domains
  - **Profiles** (weighted matrices): two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families; used for less conserved domains
  - **Hidden Markov Model** (HMM); probabilistic models; an other method to generate profiles.

Pattern-Profile



- Pattern[LIVM]-[ST]-A-[STAG]-H-C

  → Yes or no

- Profile:

```
ID   TRYPSIN_DOM; MATRIX.
AC   PS50240;
DT   DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); DEC-2001 (INFO UPDATE).
DE   Serine proteases, trypsin domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=234;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=229;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.0169; R2=0.00836256; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=1134; N_SCORE=9.5; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=775; N_SCORE=6.5; MODE=1; TEXT='?';
MA   /DEFAULT: M0=-9; D=-20; I=-20; B1=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I:       B1=0; BI=-105; BD=-105;
MA             A  B  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
MA   /M: SY='I'; M= -8,-29,-34,-26,  3,-34,-24, 34,-26, 19, 15,-24,-21,-21,-24,-19, -8, 25,-19,  3;
MA   /M: SY='N'; M=  0, 14, 10,  1,-22, -1,  6,-23, -4,-26,-17, 20,-14, -1, -6, 13,  2,-20,-34,-15;
MA   /M: SY='E'; M= -4,  4,  7, 14,-26,-13, -7,-23,  3,-22,-16,  2,  7,  3, -3,  2, -2,-21,-30,-18;
MA   /M: SY='R'; M=-12,  5,  5,  7,-23,-17,  3,-24,  8,-20,-12,  7,-16, 10, 12, -2, -6,-21,-27, -9;
MA   /M: SY='W'; M=-16,-33,-35,-27, 13,-22,-24,-11,-18,-13,-13,-31,-27,-20,-18,-30,-21,-18, 97, 25;
MA   /M: SY='V'; M=  1,-29,-31,-28, -1,-30,-29, 31,-22, 13, 11,-27,-27,-26,-22,-12, -2, 41,-27, -8;
MA   /M: SY='L'; M= -8,-29,-31,-22,  9,-30,-21, 23,-27, 37, 20,-28,-28,-21,-20,-25, -8, 17,-20, -1;
MA   /M: SY='T'; M=  2, -1, -9, -9,-11,-17,-19,-10,-10,-13,-11,  1,-11, -9,-10, 23, 43,  0,-32,-12;
MA   /M: SY='A'; M= 45, -9,-19,-10,-20, -2,-15,-11,-10,-11,-10, -9,-11, -9,-19, 10,  1, -1,-21,-18;
MA   /M: SY='A'; M= 40, -9,-17, -8,-21,  5,-18,-14, -9,-13,-12, -8,-11, -9,-16,  9, -2, -5,-21,-21;
MA   /M: SY='H'; M=-18,  0,  0,  1,-21,-19, 89,-29, -8,-21, -1,  9,-19, 11,  0, -7,-17,-29,-30, 16;
MA   /M: SY='C'; M= -9,-18,-28,-29,-20,-29,-29,-29,-29,-20,-19,-18,-39,-29,-29, -9, -9, -9,-49,-29;
MA   /I:       E1=0; IE=-105; DE=-105;
//
```

→ score/threshold

Database : protein domain/family

- Contains biologically significant « pattern / profiles/ HMM » formulated in such a way that, with appropriate computional tools, it can rapidly and reliably determine to which known family of proteins (if any) a new sequence belongs to

- Used as a tool to identify the function of uncharacterized proteins translated from genomic or cDNA sequences (« functional diagnostic »)

- Either manually curated (i.e. PROSITE, PfamA, PRINTS, SMART, TIGRFAM etc.) or automatically generated (i.e. PfamB, ProDom,…)

Protein domain/family db

PROSITE	Patterns / Profiles
ProDom	Aligned motifs (PSI-BLAST) (Pfam B)
PRINTS	Aligned motifs
Pfam	HMM (Hidden Markov Models)
SMART	HMM
TIGRfam	HMM

Interpro

BLOCKS	Aligned motifs (PSI-BLAST)
SCOP	Structural classification of proteins
CATH	Structural classification of proteins

(see PDB)

Prosite

- Created in 1988 (SIB)
- Contains functional domains fully annotated, based on two methods: patterns and profiles

- Entries are deposited in PROSITE in two distinct files:
  - Pattern/profiles with the list of all matches in Swiss-Prot
  - Documentation

InterPro
www.ebi.ac.uk/interpro

- Search simultaneously many domain databases (PRINTS, PROSITE, Pfam, ProDom, SMART, and TIGRFAMs).

- Contains an unique AC, functional description of the domain and references.

- Links are made back to the relevant member databases.

From a Swiss-Prot entry:

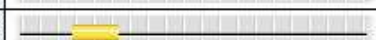
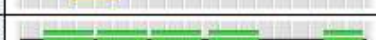
GlycoSuiteDB	P01588; -.
MIM	133170 [NCBI / EBI].
GeneCards	EPO.
GeneLynx	EPO.
InterPro	IPR001323; EPO_TPO. IPR003013; Erythroptn. Graphical view of domain structure.
Pfam	PF00758; EPO_TPO; 1.
PRINTS	PR00272; ERYTHROPTN.
PROSITE	PS00817; EPO_TPO; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
BLOCKS	P01588.
DOMO	P01588.

InterPro Detailed matches for protein

Protein matches for protein EPO_HUMAN(P01588) from the the Swiss-Prot database.

One line is shown per method for the protein. The vertical line are drawn at 10aa intervals.

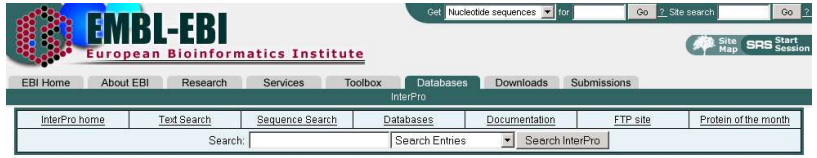
- Go to the Swiss-Prot entry for this protein.
- View the GOA annotation for this protein.
- View the matches in a table

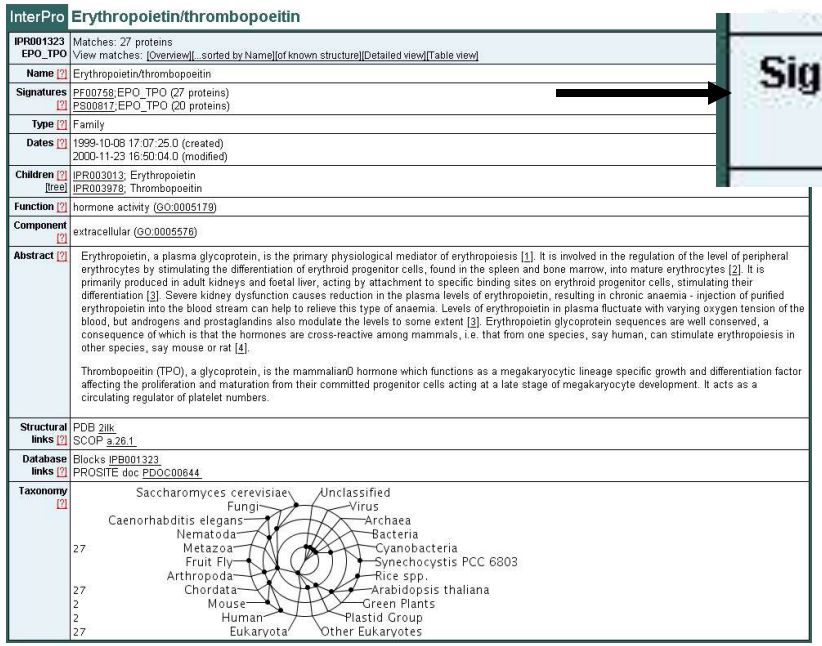
Interpro Entry	Method accession	Graphical match [?]	Method name
IPR001323:	PF00758		EPO_TPO
IPR001323:	PS00817		EPO_TPO
IPR003013:	PR00272		ERYTHROPTN
NONE:	1buyA0		1buyA0
NONE:	d1buya_		d1buya_

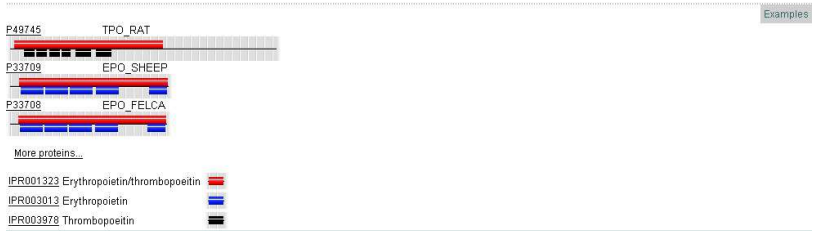
Key:		
Database	True match	False/Uncertain match
SCOP		
Pfam		
CATH		
Prints		
Prosite pattern		

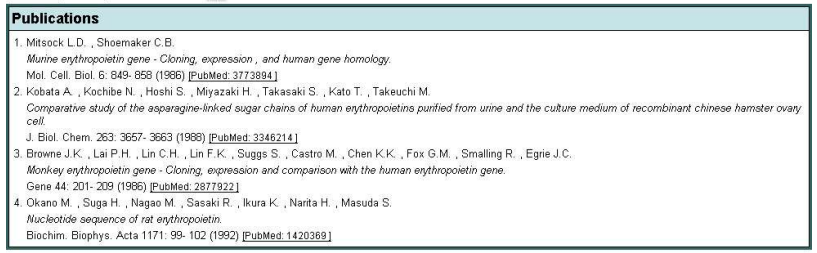
Normal | Printer Friendly | Text | Simple HTML | XML | Curator View

InterPro

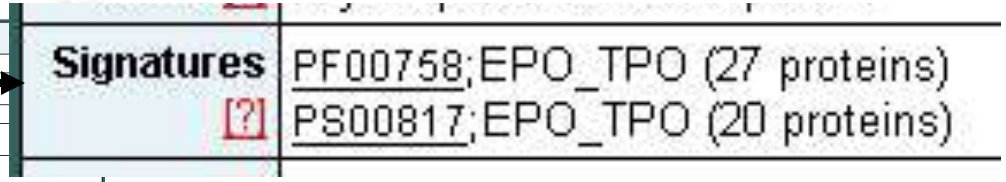

EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go
Site Map SRS Start Session
EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
InterPro
InterPro home Text Search Sequence Search Databases Documentation FTP site Protein of the month
Search: Search Entries Search InterPro


InterPro Erythropoietin/thrombopoeitin
IPR001323 EPO_TPO
Matches: 27 proteins
View matches: [Overview][...sorted by Name][of known structure][Detailed view][Table view]
Name [?] Erythropoietin/thrombopoeitin
Signatures [?] PF00758;EPO_TPO (27 proteins)
PS00817;EPO_TPO (20 proteins)
Type [?] Family
Dates [?] 1999-10-08 17:07:25.0 (created)
2000-11-23 16:50:04.0 (modified)
Children [?] [tree] IPR003013; Erythropoietin
IPR003978; Thrombopoeitin
Function [?] hormone activity (GO:0005179)
Component [?] extracellular (GO:0005576)
Abstract [?]
Erythropoietin, a plasma glycoprotein, is the primary physiological mediator of erythropoiesis [1]. It is involved in the regulation of the level of peripheral erythrocytes by stimulating the differentiation of erythroid progenitor cells, found in the spleen and bone marrow, into mature erythrocytes [2]. It is primarily produced in adult kidneys and foetal liver, acting by attachment to specific binding sites on erythroid progenitor cells, stimulating their differentiation [3]. Severe kidney dysfunction causes reduction in the plasma levels of erythropoietin, resulting in chronic anaemia - injection of purified erythropoietin into the blood stream can help to relieve this type of anaemia. Levels of erythropoietin in plasma fluctuate with varying oxygen tension of the blood, but androgens and prostaglandins also modulate the levels to some extent [3]. Erythropoietin glycoprotein sequences are well conserved, a consequence of which is that the hormones are cross-reactive among mammals, i.e. that from one species, say human, can stimulate erythropoiesis in other species, say mouse or rat [4].
Thrombopoeitin (TPO), a glycoprotein, is the mammalian0 hormone which functions as a megakaryocytic lineage specific growth and differentiation factor affecting the proliferation and maturation from their committed progenitor cells acting at a late stage of megakaryocyte development. It acts as a circulating regulator of platelet numbers.
Structural links [?] PDB 2ilk
SCOP a.26.1
Database links [?] Blocks IPB001323
PROSITE doc PDOC00644
Taxonomy [?]
Saccharomyces cerevisiae
Fungi
Caenorhabditis elegans
Nematoda
Metazoa
Fruit Fly
Arthropoda
Chordata
Mouse
Human
Eukaryota
Unclassified
Virus
Archaea
Bacteria
Cyanobacteria
Synechocystis PCC 6803
Rice spp.
Arabidopsis thaliana
Green Plants
Plastid Group
Other Eukaryotes
27
27
2
2
27


Examples
P49745 TPO_RAT
P33709 EPO_SHEEP
P33708 EPO_FELCA
More proteins...
IPR001323 Erythropoietin/thrombopoeitin
IPR003013 Erythropoietin
IPR003978 Thrombopoeitin


Publications
1. Mitsock L.D. , Shoemaker C.B.
Murine erythropoietin gene - Cloning, expression , and human gene homology.
Mol. Cell. Biol. 6: 849- 858 (1986) [PubMed: 3773894]
2. Kobata A. , Kochibe N. , Hoshi S. , Miyazaki H. , Takasaki S. , Kato T. , Takeuchi M.
Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant chinese hamster ovary cell.
J. Biol. Chem. 263: 3657- 3663 (1988) [PubMed: 3346214]
3. Browne J.K. , Lai P.H. , Lin C.H. , Lin F.K. , Suggs S. , Castro M. , Chen K.K. , Fox G.M. , Smalling R. , Egrie J.C.
Monkey erythropoietin gene - Cloning, expression and comparison with the human erythropoietin gene.
Gene 44: 201- 209 (1986) [PubMed: 2877922]
4. Okano M. , Suga H. , Nagao M. , Sasaki R. , Ikura K. , Narita H. , Masuda S.
Nucleotide sequence of rat erythropoietin.
Biochim. Biophys. Acta 1171: 99- 102 (1992) [PubMed: 1420369]


Normal Printer Friendly Text Simple HTML XML Curator View


Signatures
[?]
PF00758;EPO_TPO (27 proteins)
PS00817;EPO_TPO (20 proteins)

Databases : proteomics

- Contain informations obtained by 2D-PAGE: images of master gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Composed of image and text files
- There is currently no protein Mass Spectrometry (MS) database (not for long…)

Databases : 3D structure

- Contain the spatial coordinates of macromolecules whose 3D structure has been obtained by X-ray or NMR studies

- Proteins represent more than 90% of available structures (others are DNA, RNA, sugars, viruses, protein/DNA complexes…)

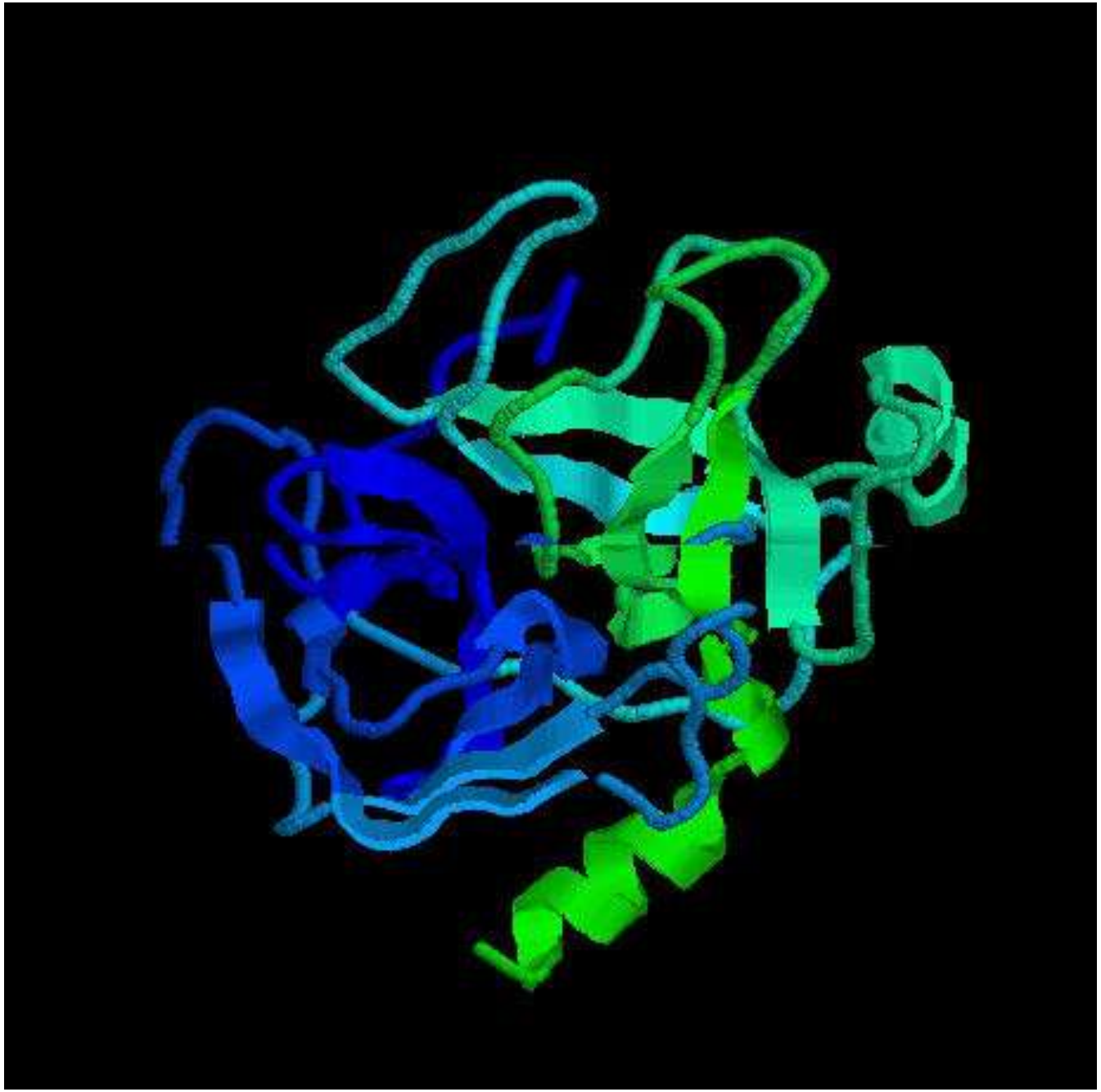
- PDB (Protein Data Bank), SCOP (structural classification of proteins (according to the secondary structures)), BMRB (BioMagResBank; RMN results)

  HSSP: Homology-derived secondary structure of proteins.

  SCOP: Structural classification of proteins

  CATH: hierarchical domain classification of protein structures derived from PDB.

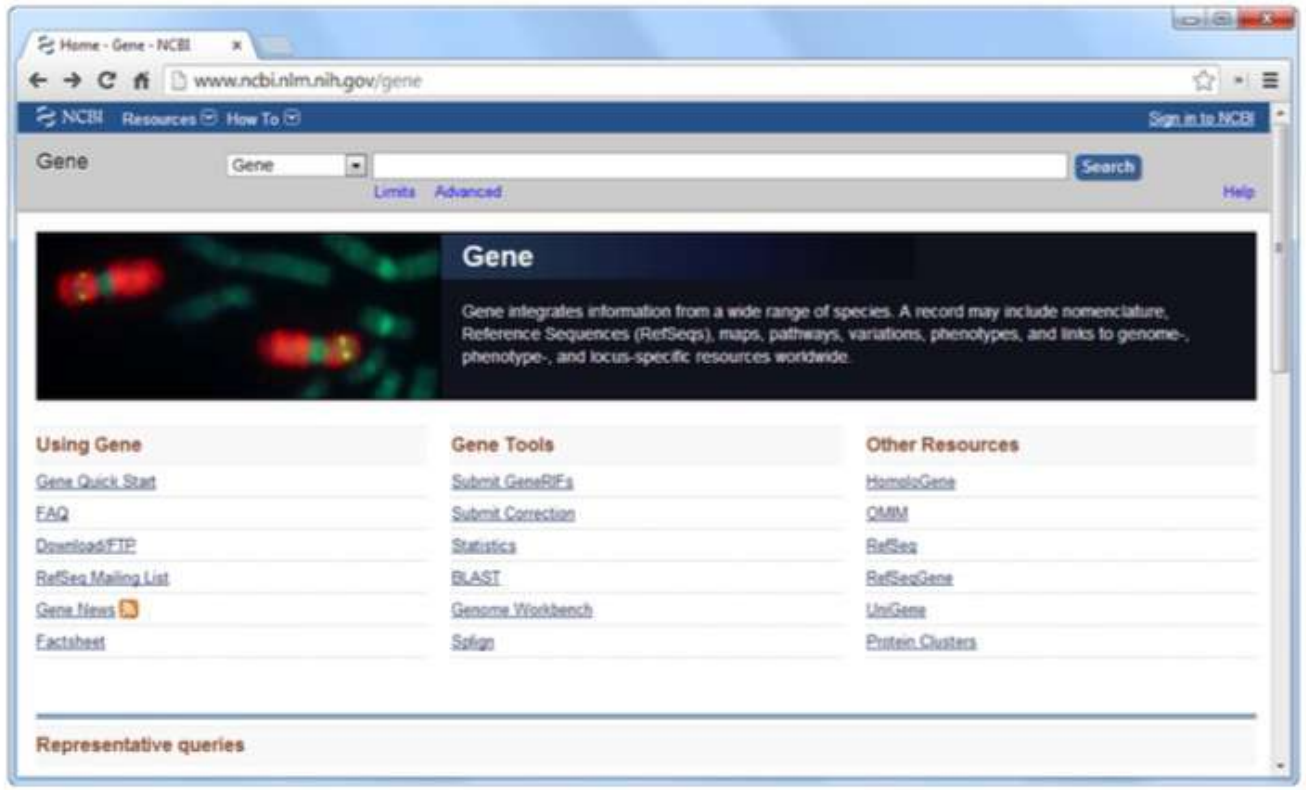
- Future: Homology-derived 3D structure db.



Motivation For Sequence Comparison

Let's use Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) to search for CFTR, the gene implicated in cystic fibrosis:

```
CFTR[sym] AND human[orgn]
```



COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

LEARN MORE

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
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog

NCBI to present on SRA and cloud computing at the 2021 Galaxy Community Conference
01 Jul 2021
We're bringing exciting developments to

GenBank release 244.0
30 Jun 2021
GenBank release 244.0 (6/26/2021) is now available on the NCBI FTP site. This release has 14.78 trillion bases and

Announcing the re-annotation of RefSeq genome assemblies for E. coli and four other species!
23 Jun 2021
We have re-annotated all RefSeq

NCBI Resources How To Sign in to NCBI

Gene Gene CFTR[sym] AND human[orgn] Search

Create RSS Save search Advanced Help

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Full Report Send to: Hide sidebar >>

Showing Current items.

CFTR CF transmembrane conductance regulator [*Homo sapiens* (human)]

Download Datasets

Gene ID: 1080, updated on 28-Jun-2021

Summary

Official Symbol	CFTR provided by HGNC
Official Full Name	CF transmembrane conductance regulator provided by HGNC
Primary source	HGNC:HGNC:1884
See related	Ensembl:ENSG00000001626 MIM:602421
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CF; MRP7; ABC35; ABCC7; CFTR/MRP; TNR-CFTR; dJ760C5.1
Summary	This gene encodes a member of the ATP-binding cassette (ABC) transporter superfamily. The encoded protein functions as a chloride channel, making it unique among members of this protein family, and controls ion and water secretion and absorption in epithelial tissues. Channel activation is mediated by cycles of regulatory domain phosphorylation, ATP-binding by the nucleotide-binding domains, and ATP hydrolysis. Mutations in this gene cause cystic fibrosis, the most common lethal genetic disorder in populations of Northern European descent. The most frequently occurring mutation in cystic fibrosis, DeltaF508, results in impaired folding and trafficking of the encoded protein. Multiple pseudogenes have been identified in the human genome. [provided by RefSeq, Aug 2017]

Table of contents

Now you try!!

Use Entrez Gene (*http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene*) to find the official full name and chromosomal location of the following genes (choose five)—all are associated with monogenic diseases.*

Gene	Disease	Gene	Disease
PAH	Phenylketonuria (PKU)	F8	Hemophilia A
HBB	Sickle-cell anemia	DMD	Muscular dystrophy
OCA2	Albinism	PHEX	Hypophosphatemic rickets
HTT	Huntington's disease	MECP2	Rett's syndrome
DMPK	Myotonic dystrophy	USP9Y	Spermatogenic failure
NF1	Neurofibromatosis	HEXA	Tay-Sachs disease
PKD1	Polycystic kidney disease	FMR1	Fragile X disease

**http://www.nature.com/scitable/topicpage/rare-genetic-disorders-learning-about-genetic-disease-979*

NCBI Resources How To Sign in to NCBI

Gene

Gene CFTR[sym] Search

Create RSS Save search Advanced Help

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Protein-coding

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status clear
✓ Current

Clear all

Show additional filters

Tabular 20 per page Sort by Relevance Send to:

Hide sidebar >>

Search results

Items: 1 to 20 of 399

<< First < Prev Page 1 of 20 Next > Last >>

See also 12 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
CFTR ID: 1080	CF transmembrane conductance regulator [*Homo sapiens* (human)]	Chromosome 7, NC_000007.14 (117480025..117668665)	ABC35, ABCC7, CF/MRP, MRP7, TNR-CFTR, dJ760C5.1, CFTR	602421
Cftr ID: 12638	cystic fibrosis transmembrane conductance regulator [*Mus musculus* (house mouse)]	Chromosome 6, NC_000072.7 (18170686..18322769)	AW495489, Abcc, Abcc7	
Cftr ID: 24255	CF transmembrane conductance regulator [*Rattus norvegicus* (Norway rat)]	Chromosome 4, NC_051339.1 (46561269..46728759)	RGD1561193	
CFTR ID: 403154	CF transmembrane conductance regulator [*Sus scrofa* (pig)]	Chromosome 18, NC_010460.4 (28627717..28818209, complement)	ABCC7	
cftr ID: 559080	CF transmembrane conductance regulator [*Danio rerio* (zebrafish)]	Chromosome 18, NC_007129.7 (20677184..20722548)	abcc, si:dkey-270i2.2	
cftr.L ID: 373725	cystic fibrosis transmembrane conductance regulator L homeolog [*Xenopus laevis* (African clawed frog)]	Chromosome 3L, NC_054375.1 (75128155..75202388)	XELAEV_18017649mg, abc35, abcc7, cftr, cftr-A, cftr-b, cftr/mrp, mrp7, tnr-cftr, xcftr	
CFTR ID: 443347	CF transmembrane conductance regulator [*Ovis aries* (sheep)]	Chromosome 4, NC_040255.1 (57106498..57293281, complement)		
CFTR ID: 281067	CF transmembrane conductance regulator [*Bos taurus* (cattle)]	Chromosome 4, NC_037331.1 (50743789..50957592, complement)		
CFTR ID: 100049619	cystic fibrosis transmembrane conductance regulator [*Gallus gallus* (chicken)]	Chromosome 1, NC_052532.1 (24715298..24799319, complement)		
CFTR ID: 101672484	CF transmembrane conductance regulator [*Mustela putorius furo* (domestic ferret)]		ABCC7	
CFTR ID: 100009471	CF transmembrane conductance regulator [*Oryctolagus cuniculus* (rabbit)]	Chromosome 7, NC_013675.1 (27524303..27737713, complement)		
Cfm1 ID: 110011	cystic fibrosis modifier 1 [*Mus musculus* (house mouse)]		Cftr, Cftrm	
CFTR ID: 574346	CF transmembrane conductance regulator [*Macaca mulatta* (Rhesus monkey)]	Chromosome 3, NC_041756.1 (143689225..143868872)	ABCC7	

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Mus musculus *(2)*
Polypterus senegalus *(2)*
Mus caroli *(1)*
Mus pahari *(1)*
Mastomys coucha *(1)*
All other taxa *(392)*
More...

Find related data

Database: Select

Find items

Search details

```
CFTR[sym] AND alive[prop]
```

Search See more...

Recent activity

Turn Off Clear

CFTR[sym] AND (alive[prop]) (399) Gene

CFTR CF transmembrane conductance regulator [Homo sapiens] Gene

CFTR[sym] AND human[orgn] AND (alive[prop]) (1) Gene

Dengue virus 1, complete genome Nucleotide

NC_001477 AND (alive[prop]) (1) Gene

Now you try!!

Use Entrez Gene (*http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene*) to determine what other organisms (besides humans) have known copies of the following genes (choose another five).

Gene	Disease	Gene	Disease
PAH	Phenylketonuria (PKU)	F8	Hemophilia A
HBB	Sickle-cell anemia	DMD	Muscular dystrophy
OCA2	Albinism	PHEX	Hypophosphatemic rickets
HTT	Huntington's disease	MECP2	Rett's syndrome
DMPK	Myotonic dystrophy	USP9Y	Spermatogenic failure
NF1	Neurofibromatosis	HEXA	Tay-Sachs disease
PKD1	Polycystic kidney disease	FMR1	Fragile X disease

Questions raised by this simple exercise

- why two (or a lot more) different species of organisms have the same gene?
- why does it matter?

- most (if not all) genes have homologs in other organisms
- how do we establish whether two genes are homologous?
- how do we compare their sequences?
- how do we automate the process?

Definition

A database is an organized collection of data

- again, databases have been around for a long time (an address book is a good example)
- computers are very good at storing large amounts of data in an organized and systematic way
- DBMS (database management systems) help us create, run, and maintain large database systems

Considering the HUGE amounts of data we deal with in Bioinformatics, databases are essential

The CFTR gene in two species: human and rat

First 120 nucleotides of the gene in **rat**:

```
AATTGGAAGCAAATGACATCACCTCAGGACTGAGTAAAAGGGAAGAGCC
AAAAGCATTGAACGGGTCCTGGATATCCAGAAGTCGAGTCCAACCTGAA
CCTGTCCGGACACAGACCTTAG...
```

First 120 nucleotides of the gene in **human**:

```
AATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCG
GCAGGCACCCAGAGTAGTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGA
CGGCCCTAGCAGGGACCCCAGC...
```

The CFTR protein in two species: human and rat

First 40 amino acids of the protein product in **rat**:

`MQKSPLEKASFISKLFFSWTTPILRKGYRHHLELSDIYQA...`

First 40 amino acids of the protein product in **human**:

`MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQI...`

The CFTR protein in two species: human and rat (2)

Obviously, the sequences do look similar. However:

1. how can we effectively compare two (or more) sequences, without having to stare at them?
2. how can we obtain a quantitative measure of similarity?

A good way to compare two (or more) sequences is to align them!

A few general notes on sequence alignments

- pairwise alignment: only two sequences are aligned
- multiple alignment: three or more sequences are aligned
- we can align DNA, RNA, or protein sequences
- what do you think has more information per letter: DNA or proteins?

Hands-on alignment using `ClustalW`

- pick two protein sequences from two different organisms (e.g., mouse and human) coming from the same gene
- copy and past each sequence to

  https://www.ebi.ac.uk/Tools/msa/clustalo/

  https://www.ebi.ac.uk/Tools/msa/
- before each sequence, add a line like this: `>name`
  where `name` can be the organism the protein sequence comes from, or something else
- run the alignment

The FASTA format for DNA, RNA or protein sequences

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

- widely used in Bioinformatics applications
- the first line is the HEADER ($>$ sign $+$ an identifier such as the name of the gene or protein)
- the subsequent lines contain the actual sequence
- more than one sequence can be stored in the same file (see Above)

Sequence alignment

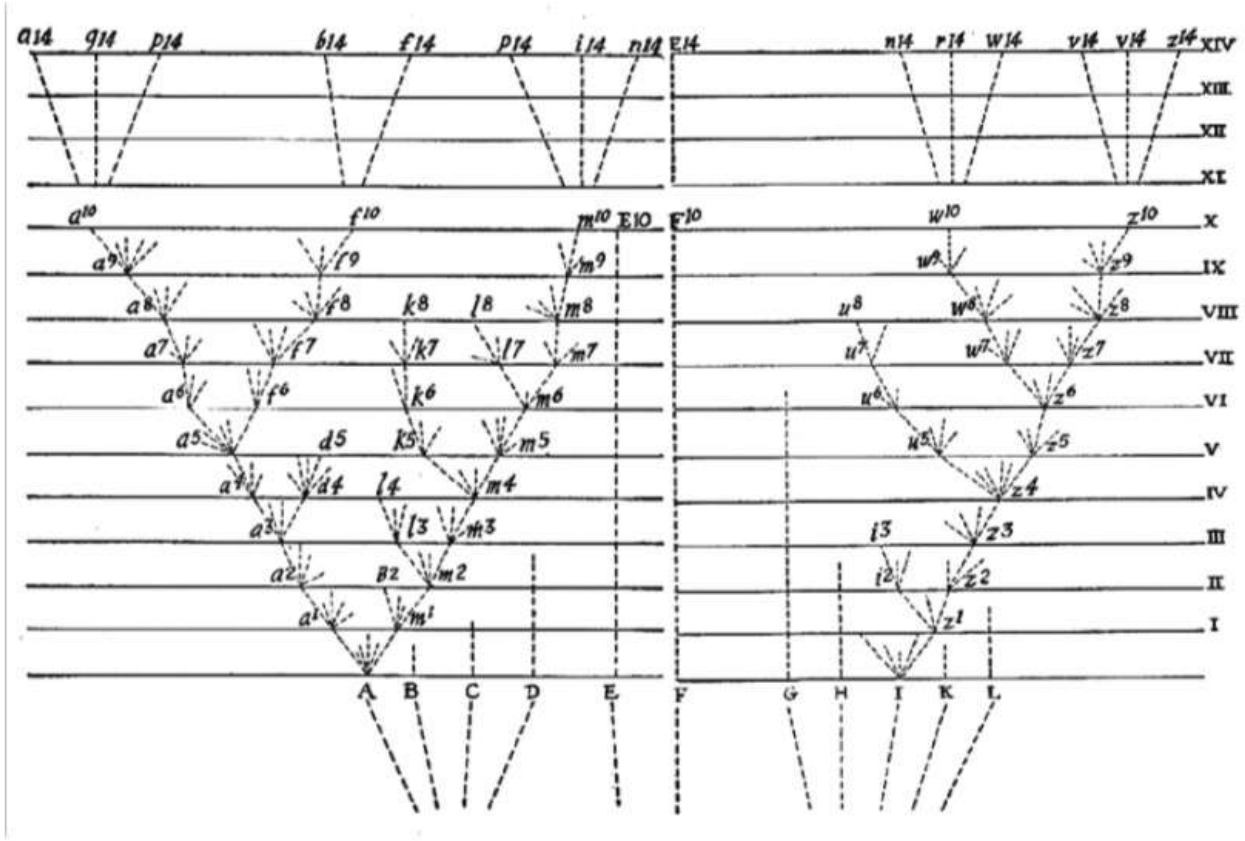
- pick two protein sequences from two di↵erent organisms (e.g., mouse and human) coming from the same gene
- copy and past each sequence to

https://www.ebi.ac.uk/Tools/msa/clustalo/

before each sequence, add a line like this: >name
where name can be the organism the protein sequence comes from, or something else (FASTA format)
run the alignment

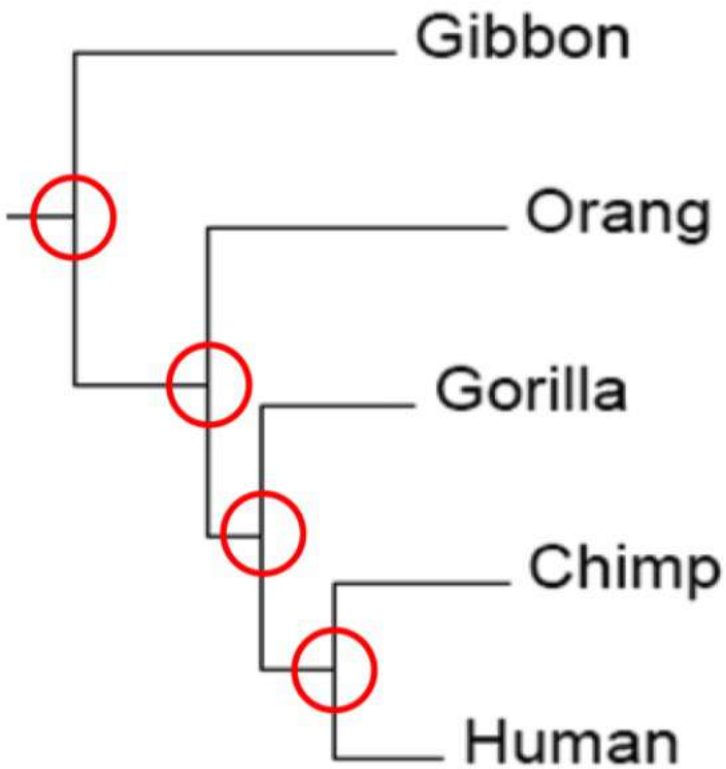
Common ancestry

Darwin used a tree analogy to illustrate how species are related

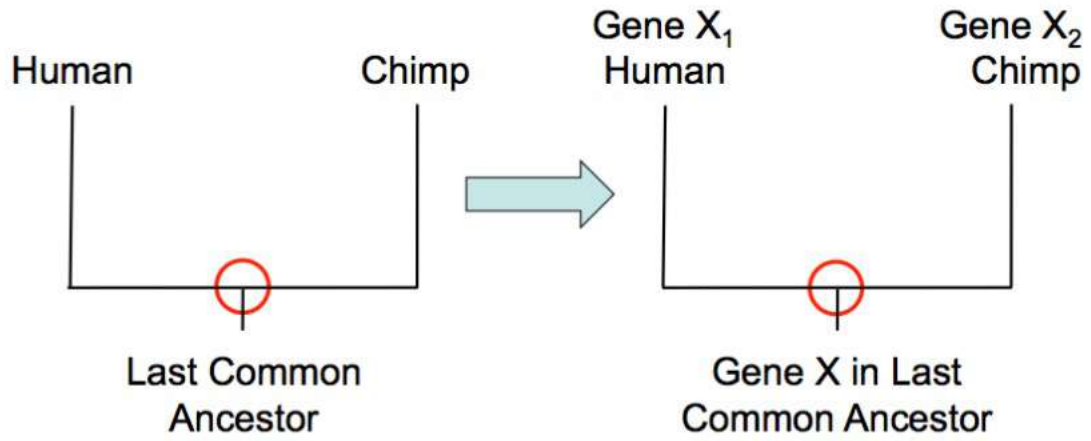


"Descent with modification", from Darwin's *The Origin of the Species*

Common ancestry: partial primate tree



What is homology?

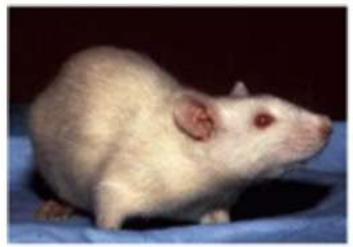


- homology is similarity due to descent from a common ancestor
- genes X_1 and X_2 have diverged from a common ancestor. **They should have similar functions.**

Similarity vs. Homology

Similarity	Homology
Similarity refers to the likeness or % identity between two sequences.	Homology refers to shared ancestry. It is a *hypothesis* that is supported by similarity
Similarity does not imply homology but provides support for it.	Homology usually implies similarity.
Similarity can be quantified! • It is correct to say that two sequences are X% identical • It is correct to say that two sequences have a similarity score of Z • It is generally incorrect to say that two sequences are X% *similar.*	Homology cannot be quantified! (Two sequences are either homologous or they're not.) • If two sequences have a high % identity, it is okay to say they are homologous (the higher the better). • It is incorrect to say two sequences have a homology score of Z. • It is incorrect to say two sequences are X% homologous.

Model organisms

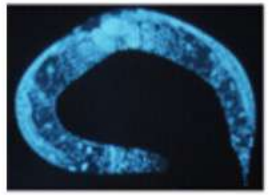


Rat
(*Rattus norvegicus*)

Mouse
(*Mus musculus*)

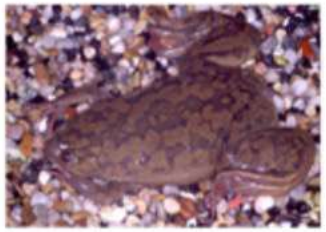
Fruitfly
(*Drosophila melanogaster*)

Plant
(*Arabidopsis thaliana*)



Nematode
(*Caenorhabditis elegans*)

Sea Urchin
(*Strongylocentrotus purpuratus*)



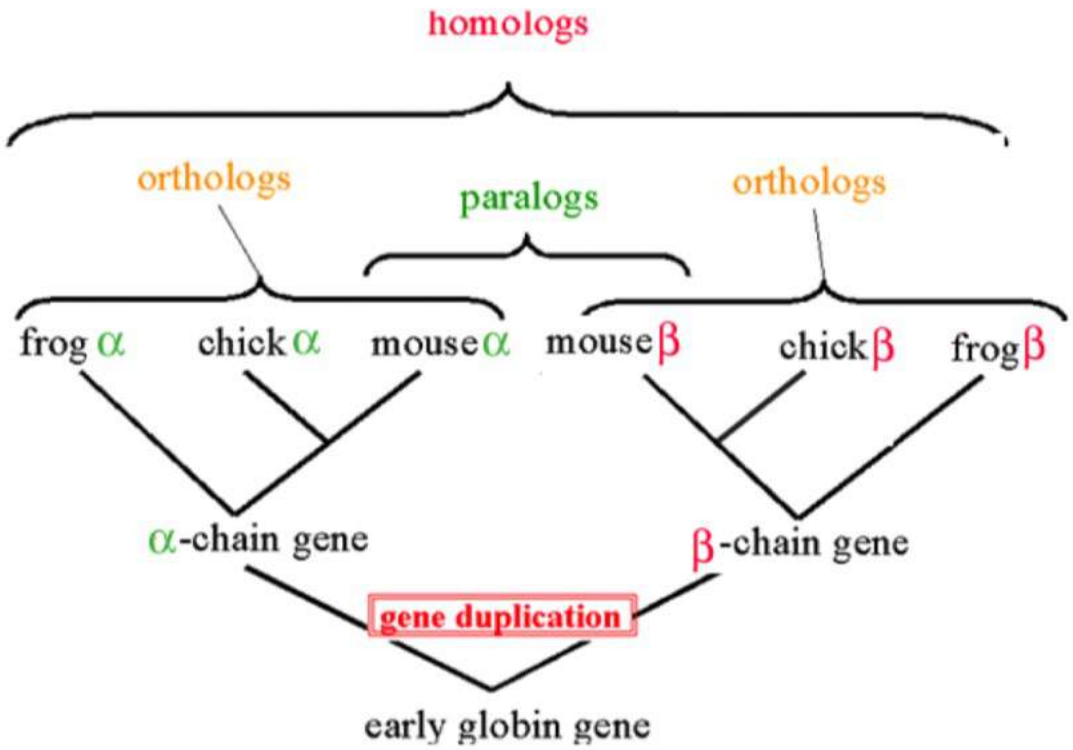
Frog
(*Xenopus laevis*)

All organisms are related: genetic elements in humans are also found in other species.

Homology relationships

- **orthologs:** similar sequences in two different organisms that have arisen due to a speciation event. They likely perform the same (or a highly similar) function
- **paralogs:** similar sequences within a single organism that have arisen due to a gene duplication event. Functionality has diverged
- **xenologs:** similar sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

Example of homology relationships: the globin gene



Figuring out ring out an unknown sequence

```
>Homo sapiens sequence X (function unknown)
CTAAACTGACAAGCCTTCAAGGAGCCCAGGGGTAAGTTAACTTGTCAACGGCATGGTTTAATCCCTTCTT
TACACTTGTGTAAATTTCAGTTACTGGTCATAGAAGGCTTTCAATGTTGAGTGGCCTTTTATTAACATGT
TTATGGTACTGCATAGATACGGGTATTTATTTTACCCTAAGAAGATTTTGAAGTTTAAAAGTACTTAAAC
TATTTGGCAAAGATTTGTTTTTAAAAATCTATTTGGTCAATCTAAATGCATTCATTCTAAAAAATTTTTT
GAACCAGATAAATAAAATTTTTTTTTGACACCAC
```

the sequence above has no known function
what would you do to learn more about it?

We could take a database of sequences with known function and extract the ones that are “suciently similar” to the sequence with unknown function
but, to compute similarity we first need to align the sequences

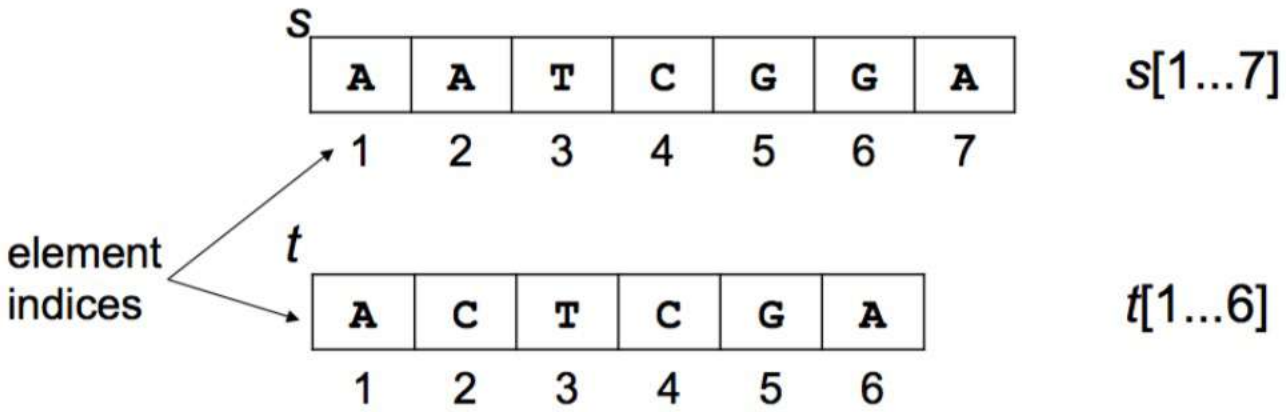
Sequence alignment

homologous sequences are, in general, not identical
because of evolution, three types of events can occur that change the sequences:

1.1 insertions (a nucleotide gets inserted into one sequence)
2.2 deletions (a nucleotide is lost)
3.3 substitutions (a nucleotide is replaced)

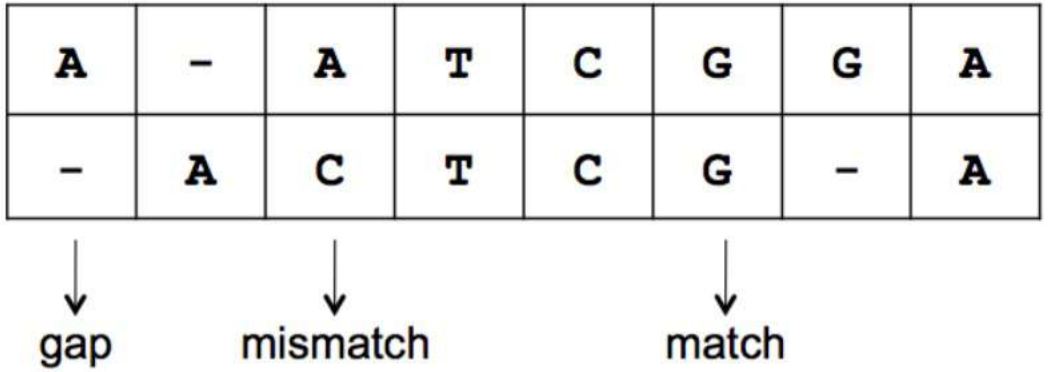
this is of course reflected on protein sequences too (we can have insertions, deletions, and substitutions as for DNA)

An example



Let's assume that the two sequences above are homologous. What kind of changes happened?

Pairwise global alignment



- a global alignment is a $2xn$ table, where n is $\geq$ the length of the longest sequence
- gaps (indicated by a dash) are allowed to make the sequences the same length
- each column in the table receives a score: match, mismatch, or gap

Score of the alignment is the sum of the column-wise scores

Scoring matrices, aka substitution matrices

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

- a scoring matrix gives scores for matches and mismatches
- matches are rewarded, mismatches are penalized
- gaps are a separate parameter, and are usually penalized more than mismatches
- scoring matrices are always symmetric
  (e.g., A $\rightarrow$ T = T$\rightarrow$A)

Scoring matrices for proteins

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Scoring matrices for proteins are more complicated (since there are 20 amino acids vs. 4 nucleotides!)

What is wrong with these scoring matrices

Now you try!

What is wrong with the following nucleotide scoring matrices?

1.

	A	T	C	G
A	-3	2	2	2
T	2	-3	2	2
C	2	2	-3	2
G	2	2	2	-3

Gap = -5

2.

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

Gap = -2

3.

	A	T	C	G
A	2	-1	-1	-1
T	-3	2	-1	-1
C	-3	-3	2	-1
G	-3	-3	-3	2

Gap = -5

Which of the two alignments you think is "better"?

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

Gap = -5

A	–	A	T	C	G	G	A
–	A	C	T	C	G	–	A
-5	-5	-3	2	2	2	-5	2

= -10

A	A	–	T	C	G	G	A
–	A	C	T	C	G	–	A
-5	2	-5	2	2	2	-5	2

= -5

The second alignment is "better" because it has a higher score

Lots of different alignments are usually possible

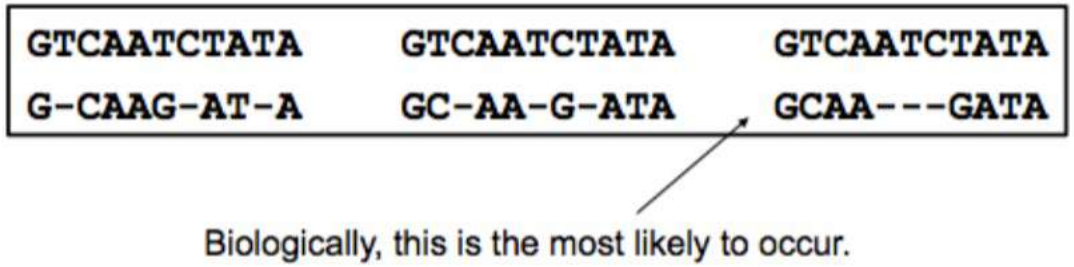
- in general, for a pair of sequences there are usually alternative alignments that are possible
- evolution teaches us that the origin of similarity is homology, so the "correct" alignment is the one that maximizes the similarity between the sequences
- therefore, the "optimal" alignment is the one with the highest (most positive) score

The "search space"

- The "search space" (number of possible gapped alignments) for optimally aligning two sequences is exponential in the length of the sequences, *n*.
- For two sequences 100 bp long (with gaps),

  $\sim 10^{70} =$
  1000000000000000000000000000000000000000000000000000000000000000000000 different alignments.
- Finding the optimal alignment for sequences of this (non-biological meaningful; average gene length is about *n* = 1000!) length would be impossible to do "by hand." → Need automation!!

Gaps tend to come in clusters



- if multiple gaps occur in an alignment, they are more biologically meaningful when they occur in clusters (in blocks)
- in the figure above, the scenario on the right is the most evolutionarily plausible

➢ however, so far we have modeled gaps with a linear function, that is, the gap penalty is simply the sum of the gaps in the alignment

➢ we need a better scheme to score gaps

Affine gaps

With affine gaps we penalize more the first gap in a cluster, and the gaps that **immediately** follows the first receive a smaller penalty

$$wx = g + r(x - 1) \tag{1}$$

where wx is the total gap penalty for a cluster, g is the penalty for opening a gap, r is the penalty for extending a gap, and x is the number of gaps in a cluster

Affine gaps (2)

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

Gap creation = -5
Gap extension = -2

Now together with the substitution matrix, we specify the penalty for opening a gap ("gap creation"), and for extending a gap ("gap extension").

How to score an alignment

```
ACGTCTGATACGCCGTATAGTCTATCT
    ||||| |||   ||  |||||||||
----CTGATTCGC---ATCGTCTATCT
```

Matches: 18 × (2) = 36

Mismatches: 2 × (-3) = -6

Gaps creations: 2 × (-5) = -10

Gap extensions 5 × (-2) = -10

Score: 36 + (-6) + (-10) + (-10) =10

Alignment score using a linear gap penalty was -5 so this alignment is "better."

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

Gap create = -5

Gap extend = -2

Now you try!!

Score the following alignments with the provided scoring matrix using both linear and affine-gap penalties.

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

Gap create = -5

Gap extend = -2

Alignment 1 (Note: Sequence 1 from NM_000207.2, *Homo sapiens* insulin; Sequence 2 from NM_008386.3, *Mus musculus* insulin)

```
Sequence 1  GGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAA
            ||| ||  |||| || |       ||| ||| |||||||  ||||||  | ||| ||||||
Sequence 2  GGGAGGAAGCCCCGGGG------ACCTTCAGACCTTGGCGTTGGAGGTGGCCCGGCAGAA
```

Alignment 2 (Note: Sequence 1 from NM_000492.3, Homo sapiens CFTR; Sequence 2 from NM_001110510.1, *Equus caballus* [horse] CFTR)

```
Sequence 1  AGTAGTAGGTC-TTTGG-CATTAGGAGCTTGAGCCCAGAC-GGCCCTAGCAGGGACCCCA
            |  |||||| | ||||| ||||||||||  |||||||| | | |||  | |||||||||
Sequence 2  A-AAGTAGGGCGTTTGGCCATTAGGAGCGCGAGCCCAGCCTGTCCC--GGAGGGACCCCG
```

- **Alignment 1 (linear gap)**

  Matches = 41 × 2 = 82

  Mismatches = 13 × -3 = -39

  Gaps = 6 × -5 = -30

  Total: 82 + (-39) + (-30) = 13

- **Alignment 2 (linear gap)**

  Matches = 46 × 2 = 92

  Mismatches = 8 × -3 = -24

  Gaps = 6 × -5 = -30

  Total: 82 + (-39) + (-30) = 38

- **Alignment 1 (affine gap)**

  Matches = 41 × 2 = 82

  Mismatches = 13 × -3 = -39

  Gap creations = 1 × -5 = -5

  Gap extensions = 5 × -2 = -10

  Total: 82 + (-39) + (-5) + (-10) = 34

- **Alignment 2 (affine gap)**

  Matches = 46 × 2 = 92

  Mismatches = 8 × -3 = -24

  Gap creations = 5 × -5 = -25

  Gap extensions = 1 × -2 = -2

  Total: 82 + (-24) + (-25) + (-2) = 41

Global vs. local alignment

- under certain circumstances, the most meaningful alignment between two sequences is not a global one, but a local one
- a local alignment only aligns parts of the sequences

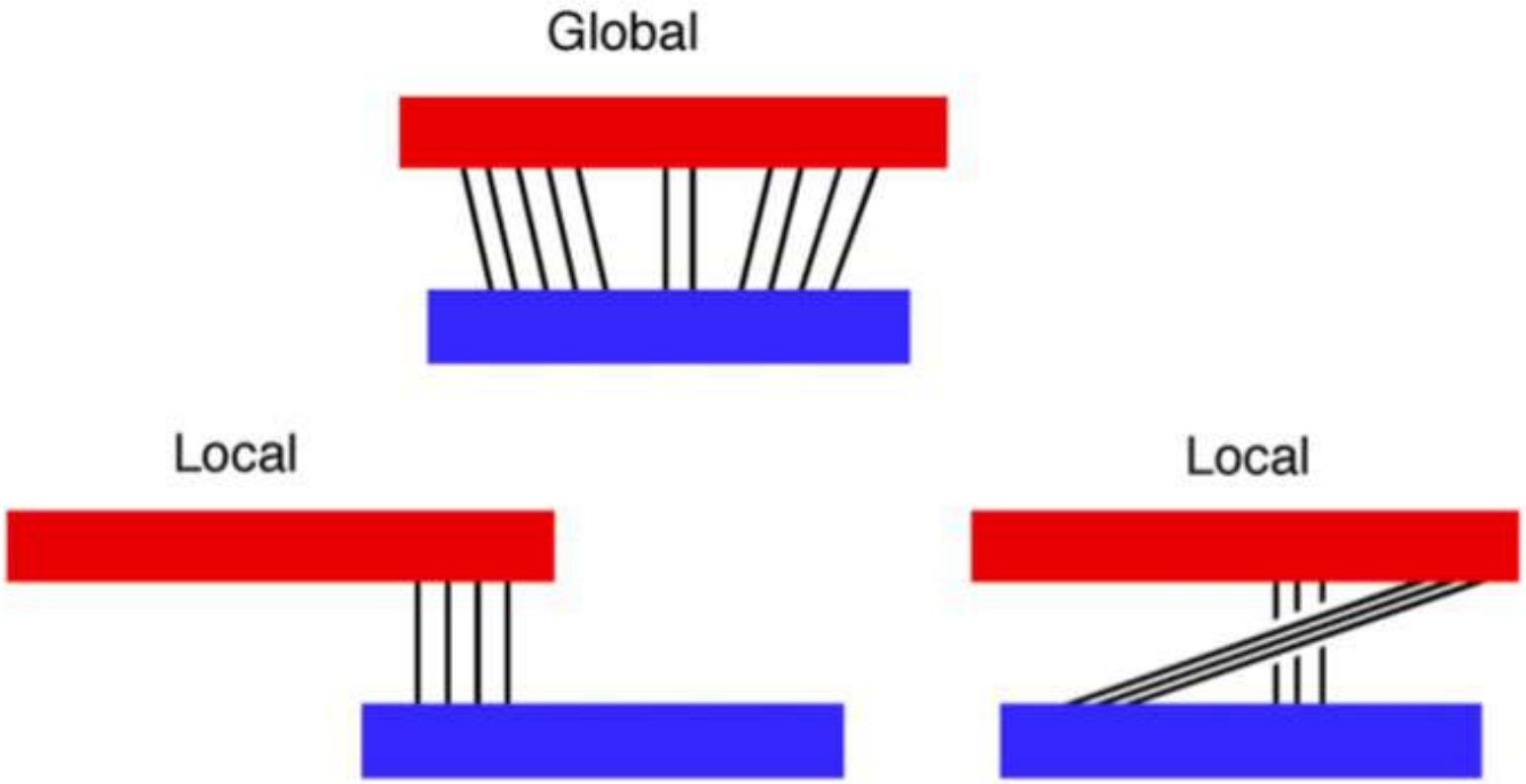
Global Alignment

- a global alignment attempts to align every nucleotide (or amino acid) in both sequences
- it is appropriate if the sequences are similar and roughly of equal sizes
- if the sequences are dissimilar or one is much longer than the other, the alignment can be very “gappy”
- algorithm: Needleman-Wunsch, from the names of the scientists who developed it in 1970

Local Alignment

- look for short regions of similarity within two sequences
- motivated by evolutionary arguments: some regions of a sequence are move conserved than the rest of the sequence, and easier to align
- algorithm: Smith-Waterman (a variant of Needleman-Wunsch), published in 1981

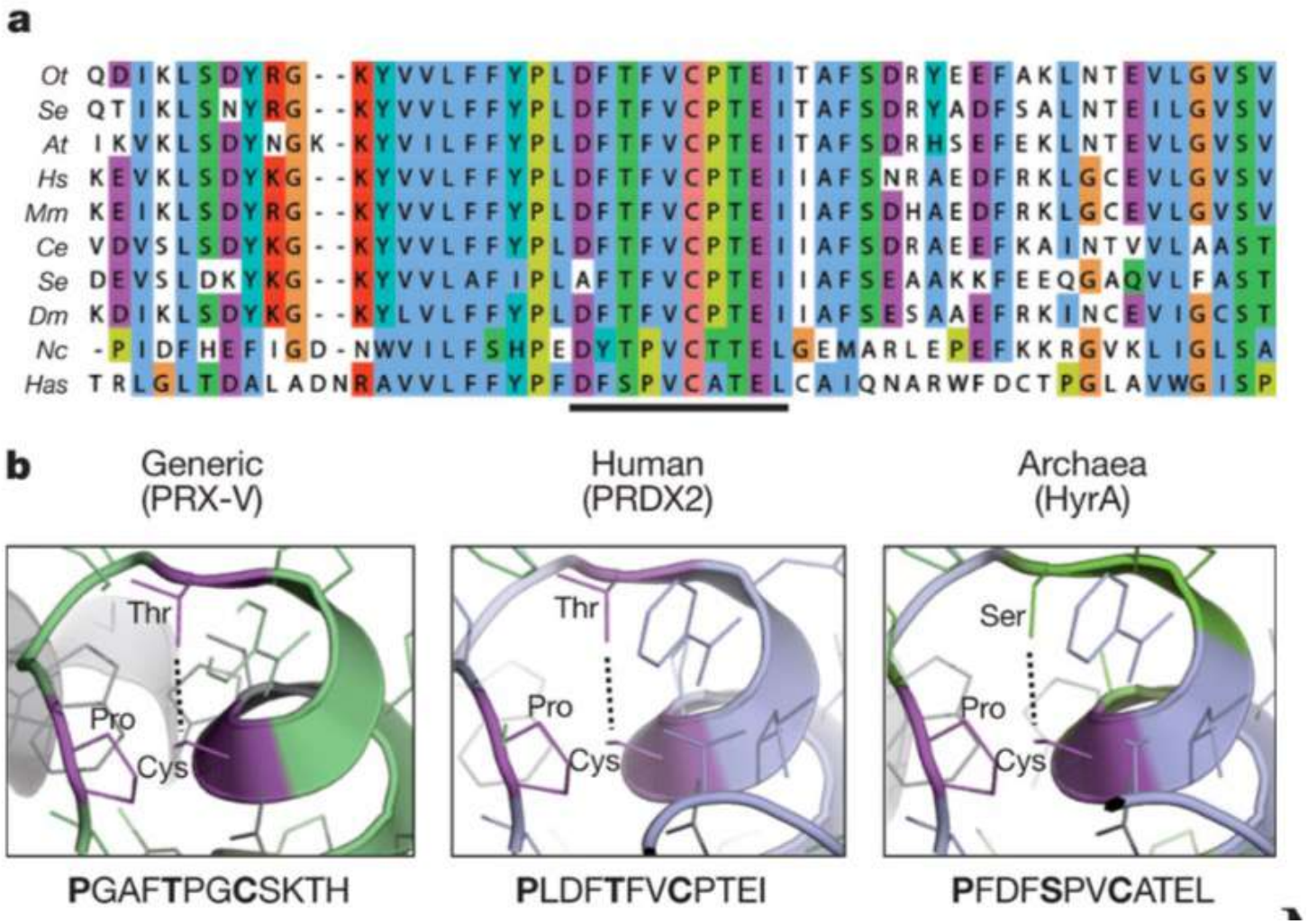
Global vs. Local Alignment



Multiple Sequence Alignment

- A multiple sequence alignment is an alignment that contains more than two sequences.
- Usually, multiple sequence alignments contain tens and often hundreds of sequences.
- For example, we could create a multiple sequence alignment of the CFTR (or another gene) for all species that have a copy of the gene.
- Multiple alignments are computationally very intensive and are usually performed on protein sequences.

How do multiple sequence alignments look like?



What can we do with them

- determine regions in the sequence that are highly conserved (important for structure, function, and disease)
- identify species-specific features of the sequences
- build phylogenetic trees

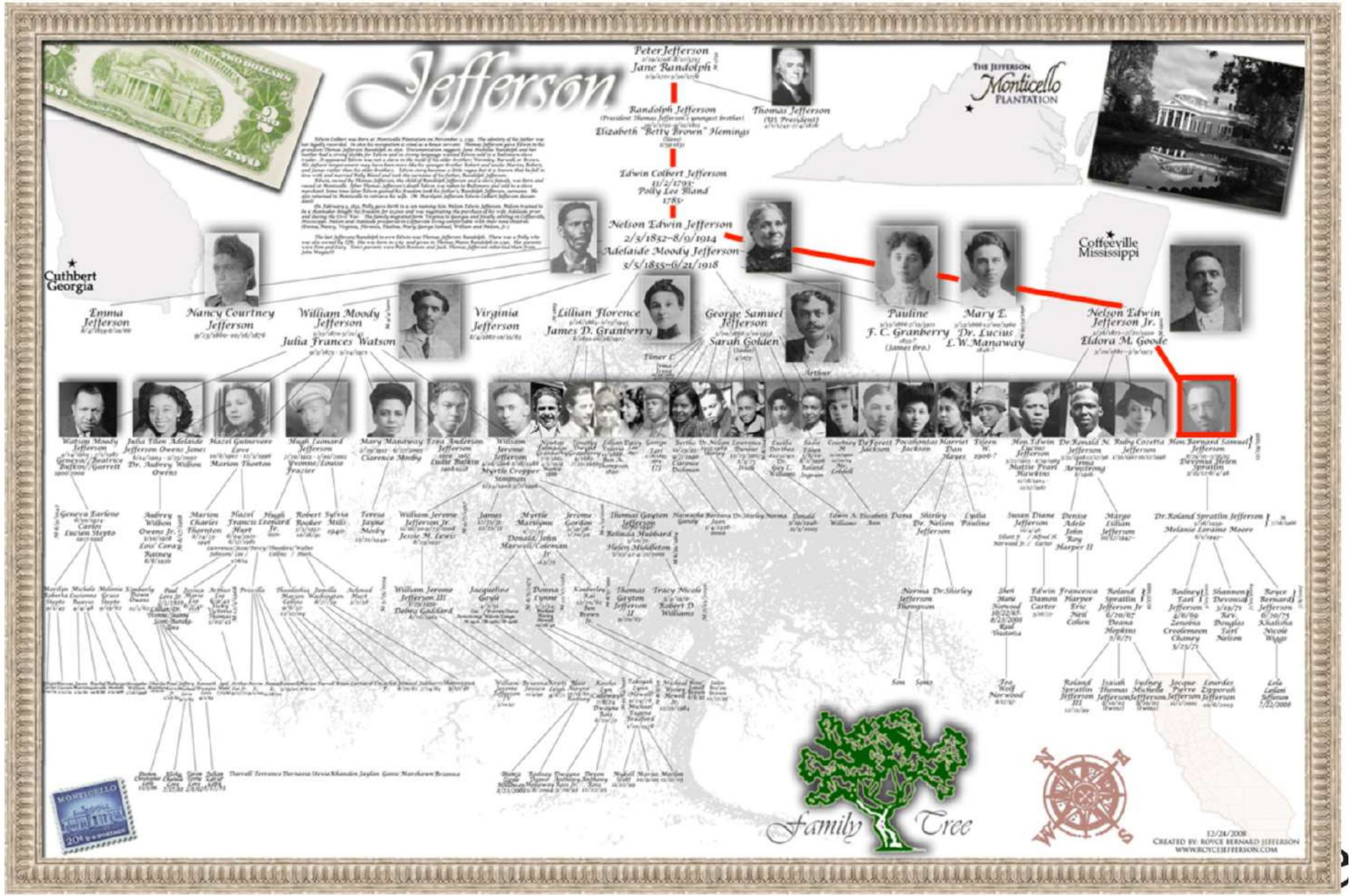
What is Phylogenetics

Definition

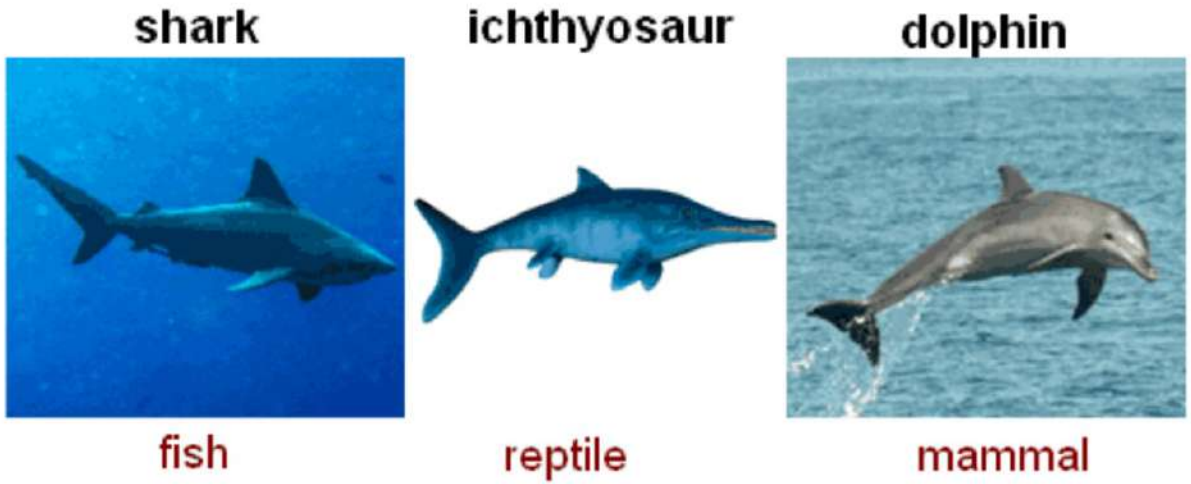
"A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have derived during evolution. The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree. The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related. The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch length."

David Mount, *Bioinformatics*, Cold Spring Harbor Press

An analogy: family trees



Convergent evolution



Definition

Convergent Evolution is defined as the evolutionary process whereby organisms that are not closely related **independently** acquire similar traits as a result of having to adapt to similar environments.

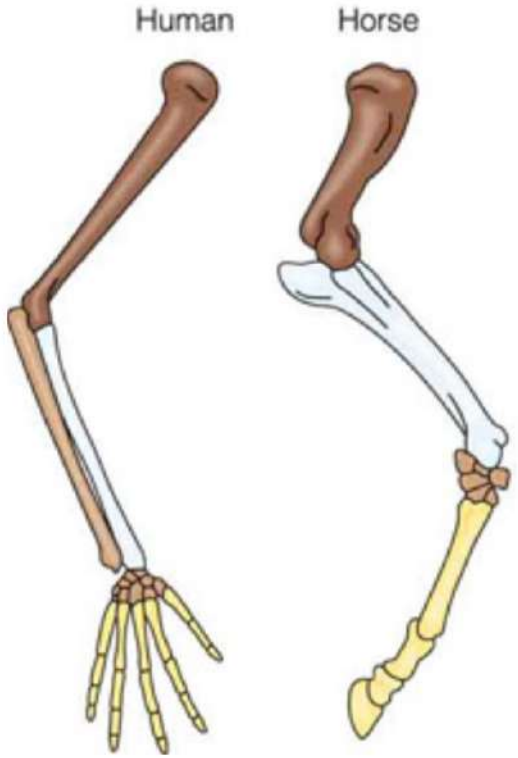
Convergent Evolution

Echidna - An egg-laying mammal (monotreme) that lives in Australia and New Guinea.

Porcupine - A rodent found in the Americas, southern Asia, and Africa..

Echidnas and porcupines are not closely related but both have evolved spines.

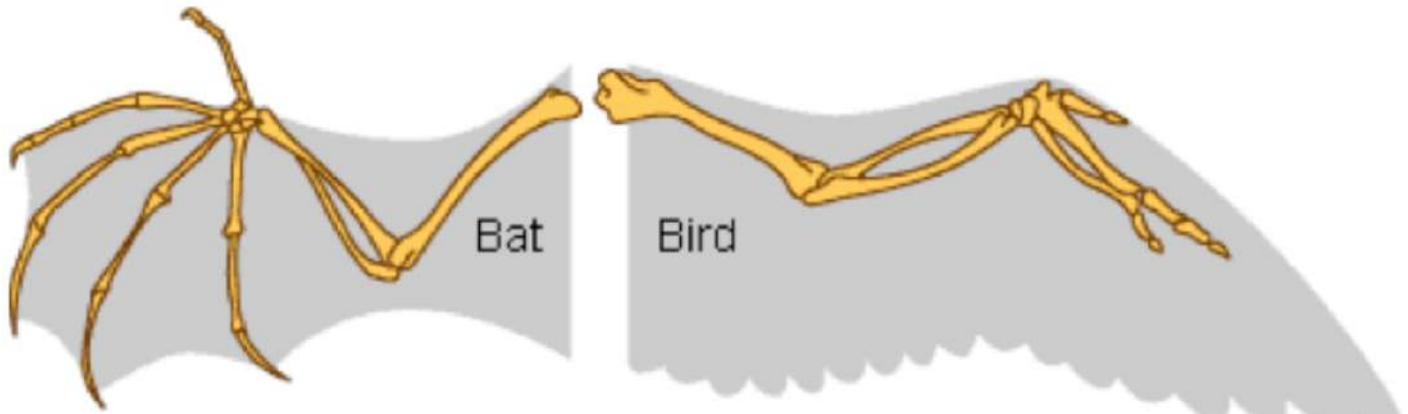
Divergent evolution

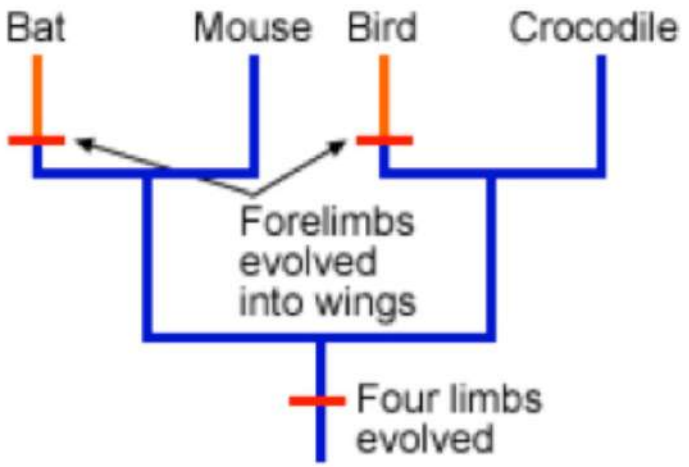


Definition

Divergent Evolution is defined as the process whereby **closely related** species acquire very **different traits**

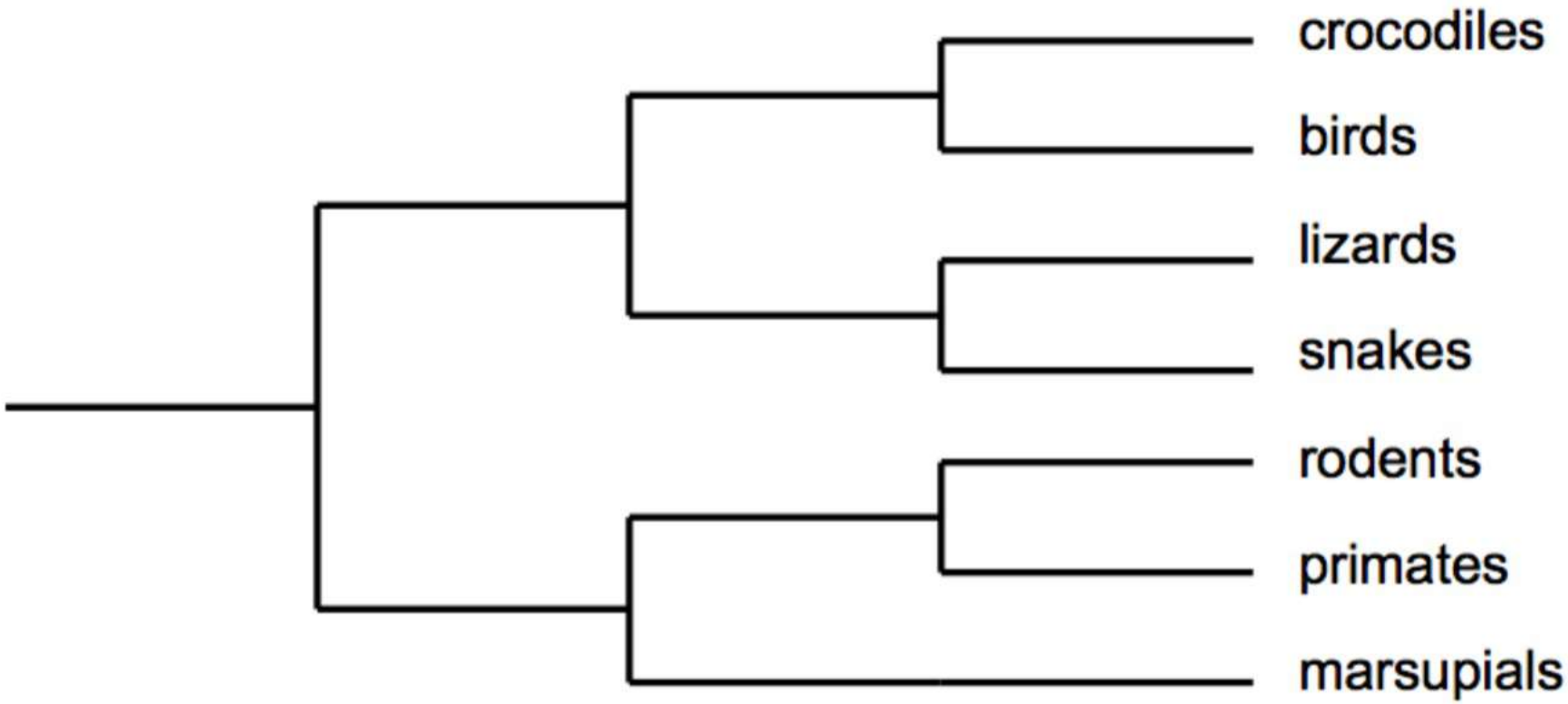
Convergent or divergent evolution?



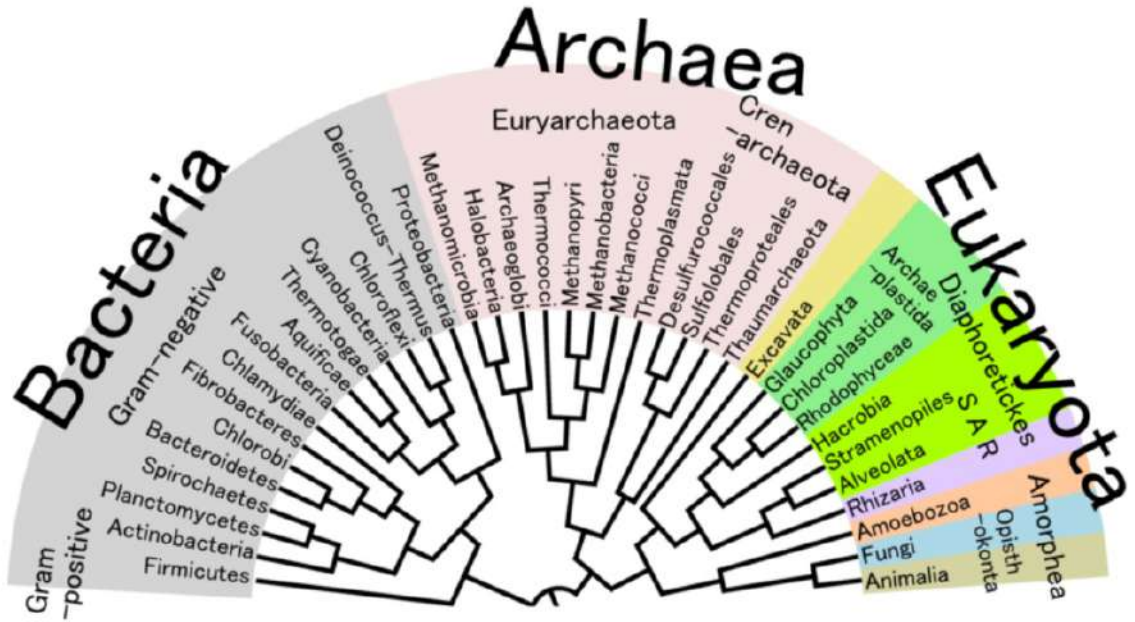


Convergent evolution. Birds and bats are **not** closely related.

Intro to phylogenetic trees

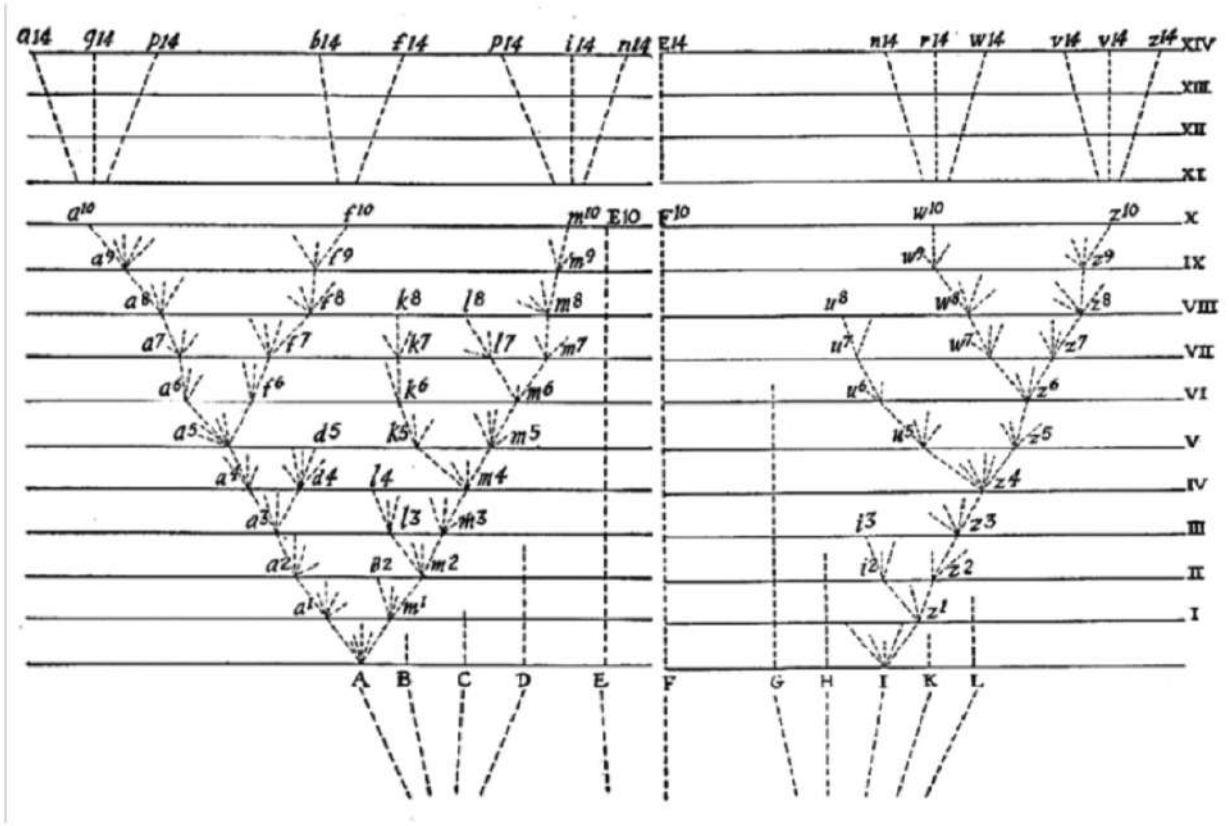


The Tree of Life



- roughly 1.75 million species have been formally described and given official names[1]
- between 5 million and 30 million species are believed to exist today (including microorganisms)

Traditional phylogenetics



Darwin drew this tree by hand, and built phylogenies by comparing the morphological structures of different organisms. He went far beyond traditional taxonomy.

Molecular phylogenetics

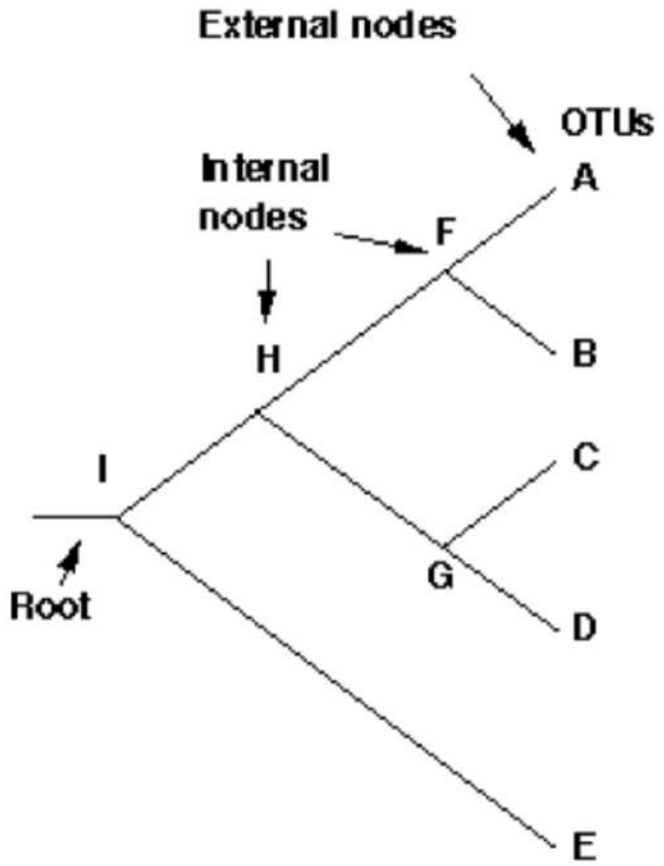
```
Mouse    LQKCSYDEHAKLVQEVTDFAKTCVADESAANCDKSLHTLFGDKLCAIPNLRENYGELADC 114
Rat      LQKCPYEEHIKLVQEVTDFAKTCVADENAENCDKSIHTLFGDKLCAIPKLRDNYGELADC 114
Dog      LQQCPFEDHVKLAKEVTEFAKACAAEESGANCDKSLHTLFGDKLCTVASLRDKYGDMADC 114
Cat      LQQCPFEDHVKLVNEVTEFAKGCVADQSAANCEKSLHELLGDKLCTVASLRDKYGEMADC 114
Human    LQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADC 114
Cow      LQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADC 114
Sheep    LQQCPFDEHVKLVKELTEFAKTCVADESHAGCDKSLHTLFGDELCKVATLRETYGDMADC 114
Pig      LQQCPYEEHVKLVREVTEFAKTCVADESAENCDKSIHTLFGDKLCAIPSLREHYGDLADC 114
Horse    LQQCPFEDHVKLVNEVTEFAKKCAADESAENCDKSLHTLFGDKLCTVATLRATYGELADC 114
Rabbit   LQKCPYEEHAKLVKEVTDLAKACVADESAANCDKSLHDIFGDKICALPSLRDTYGDVADC 114
Chicken  LQRCSYEGLSKLVKDVVDLAQKCVANEDAPECSKPLPSIILDEICQVEKLRDSYGAMADC 117
Frog     LQKCSLEELSKLVNEINDFAKSCTGNDKTPECEKPIGTLFYDKLCADPKVGVNYEWSKEC 117
         **:*. :   **..:: ::*: *.:::.   *.*.:  :: *::*   .:   *    :*
```


Molecular phylogenetics uses molecular data and sequence alignments to **infer** the evolutionary relationships between organisms.

Advantages of using molecular data

1. molecular data is less affected by problems arising from convergent evolution
2. easier to find molecular characters to compare distantly related organisms
3. for some organisms (like bacteria) it is not easy to identify distinct morphological features!
4. the results are **quantitative** and with solid statistical foundations

Phylogenetic trees: terminology



A-E are external nodes (extant)
F-I are internal (ancestral nodes)

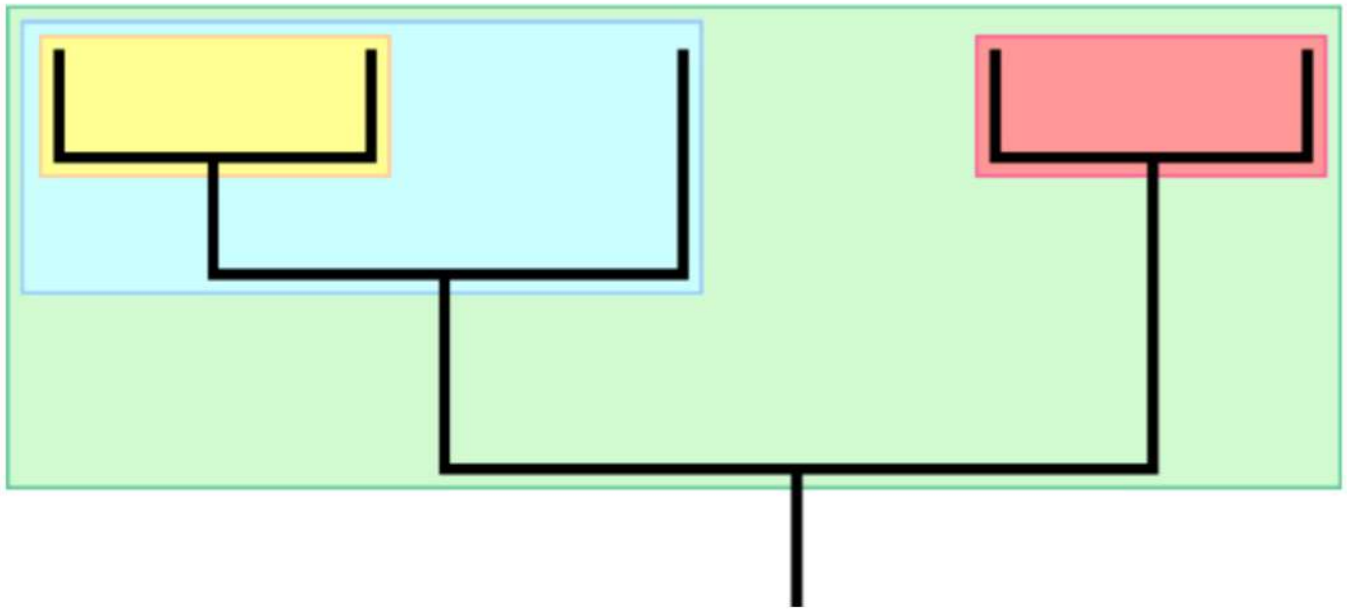
OTUs are operational taxonomic units
They can be: species
populations
individuals
genes
proteins
They are the extant (existing) OTUs
Internal nodes represent ancestral units.

Topology: order of the nodes on the tree

This is an example of a rooted tree. Rooted trees provide information about the direction of evolution.

Clades

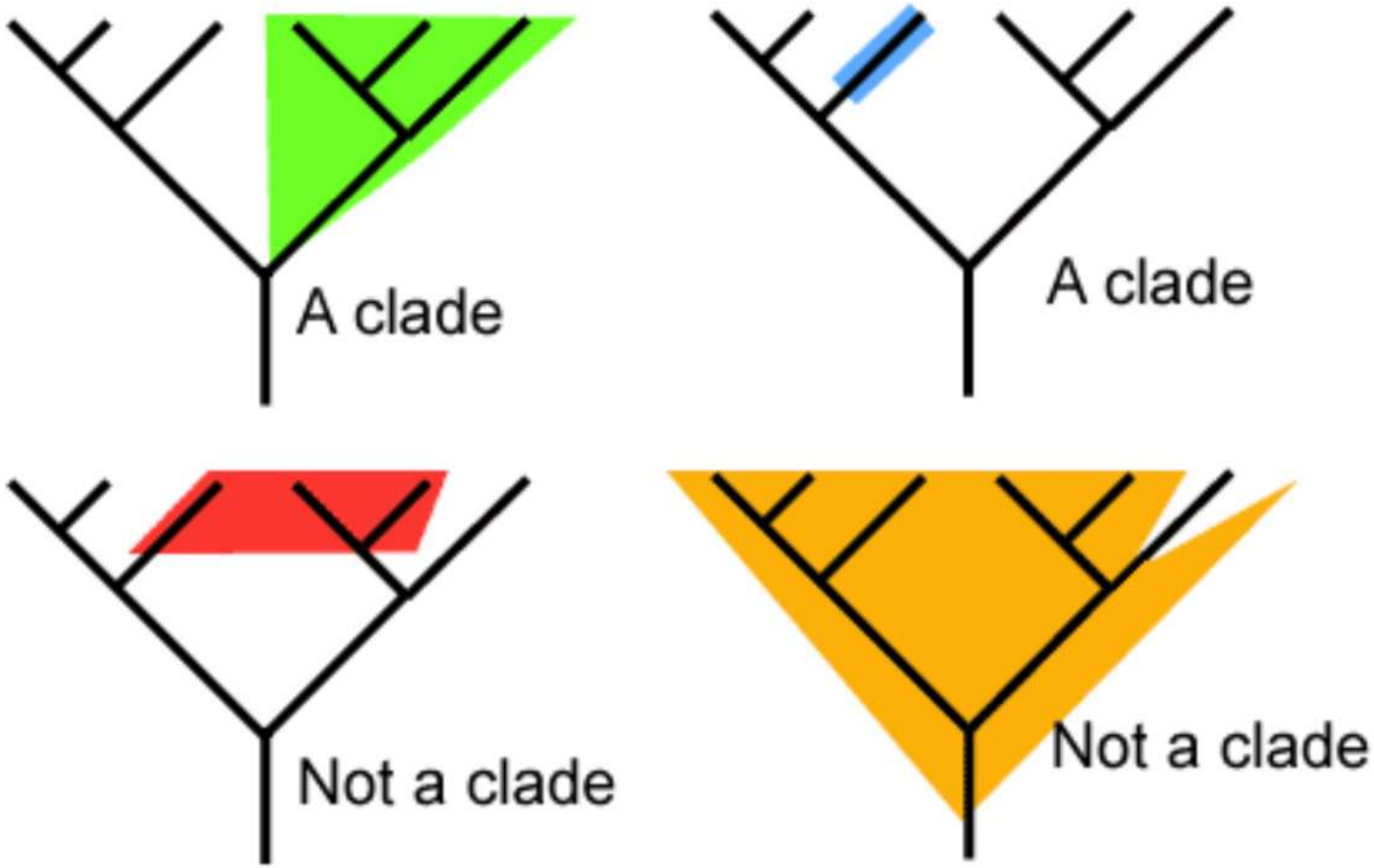
Each colored rectangle below represents a clade:



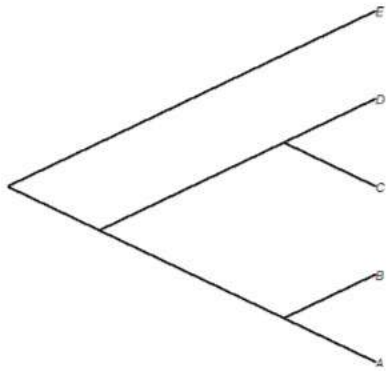
Definition

A clade is a group of Taxonomy Units (organisms, genes, etc.) that includes an ancestor and **all** descendants of that ancestor.

Clades



Tree terminology: bifurcating nodes



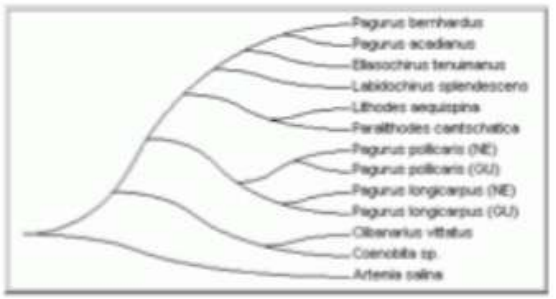
Definition

bifurcating nodes are nodes that have only two lineages that descend from them. It's the most common evolutionary scenario.

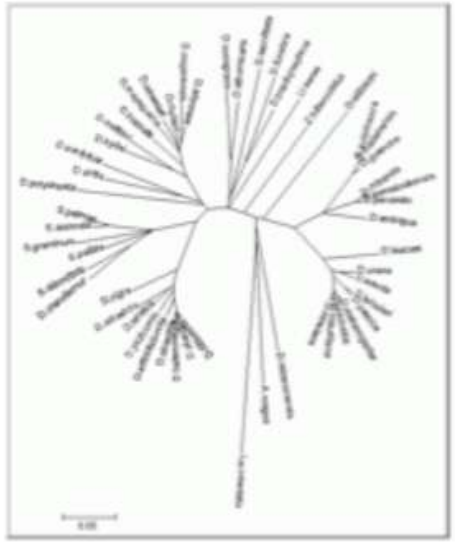
Different representations

Other styles of phylogenetic trees

Style: *curved branch*



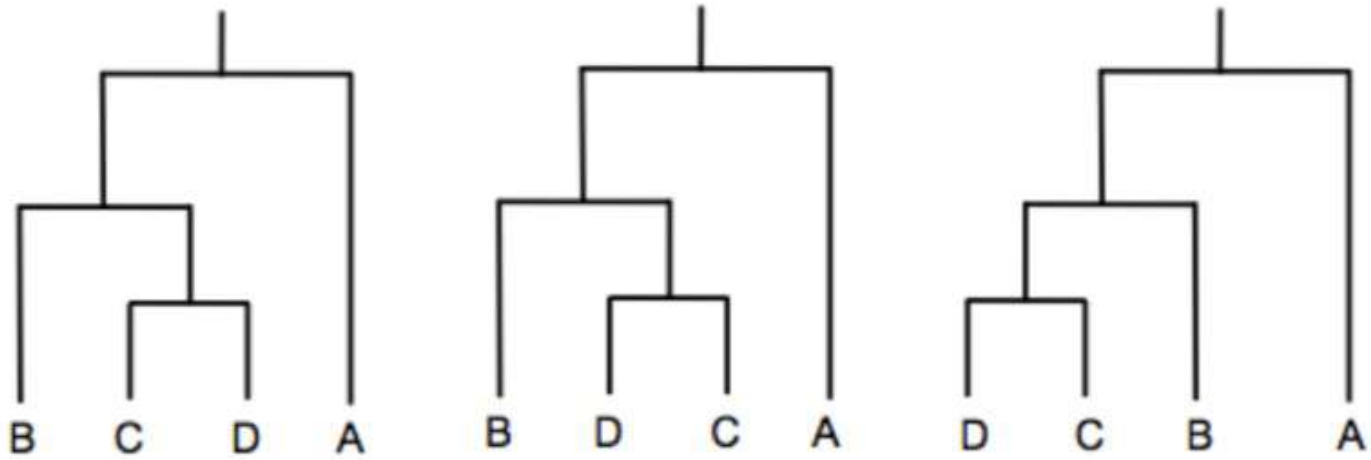
Style: *circular*



Style: *radiation (unrooted with branch lengths)*

Tree topology

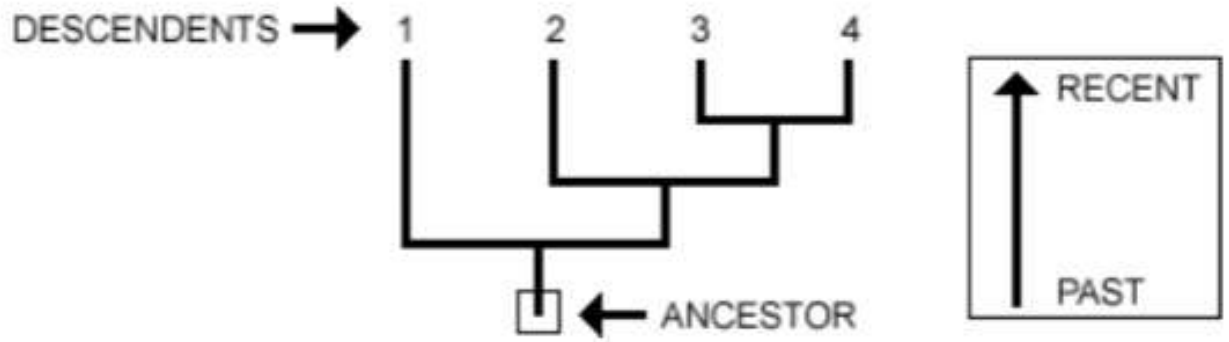
- *Topology*: Branching pattern of a tree.
- Trees can look different but have the same topology.



Style: *rectangular branch*

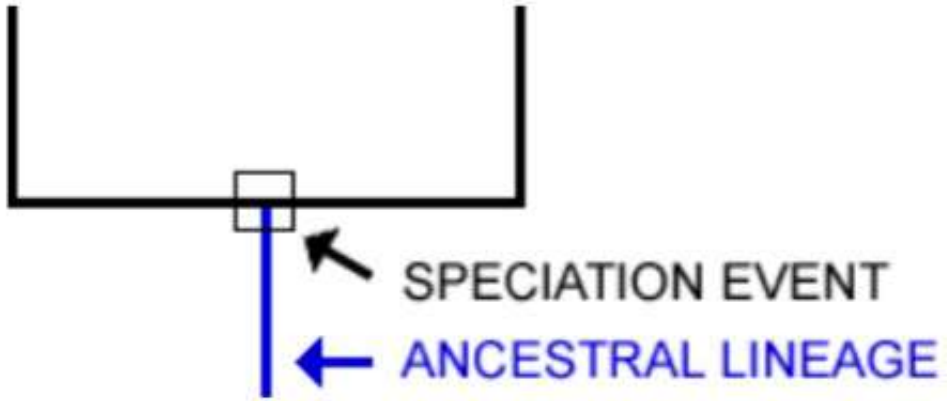
- The three trees above look different but have the same topology (convey exactly the same information).

The time arrow in phylogenetic trees



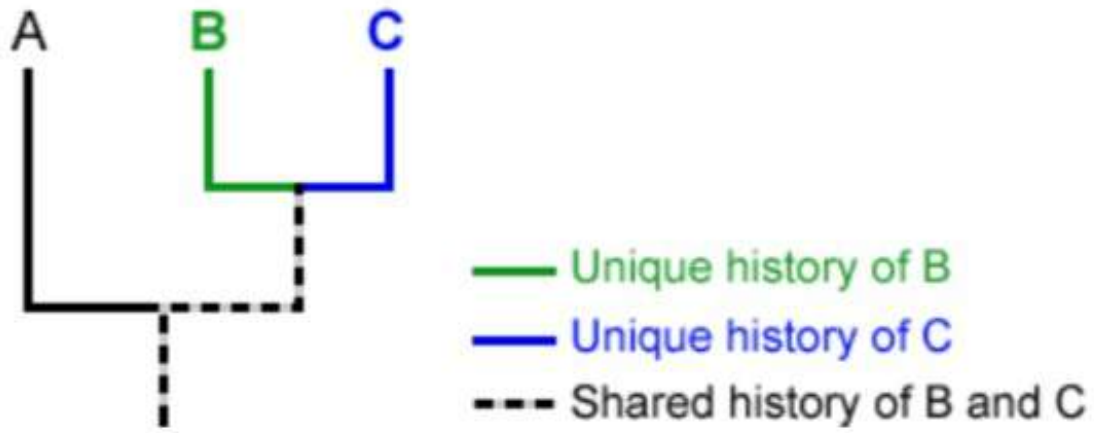
In phylogenetic trees, time flows in the direction of the descendants (the "leaves" in the trees)

Speciation events



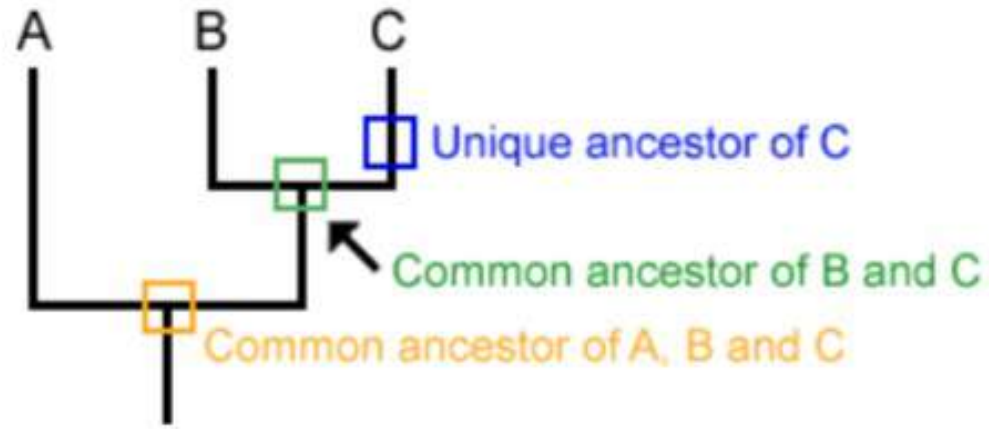
In phylogenetic trees, speciation events (the formation of new species) are represented as branches coming out of a node

Evolutionary histories in phylogenetic trees



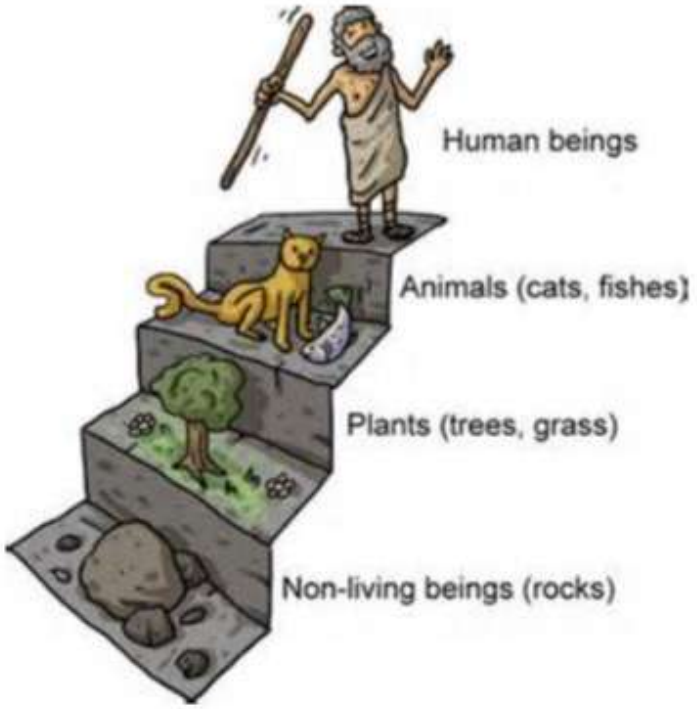
Each lineage has a part of its history that is unique to it alone and parts that are shared with other lineages.

Ancestors in phylogenetic trees



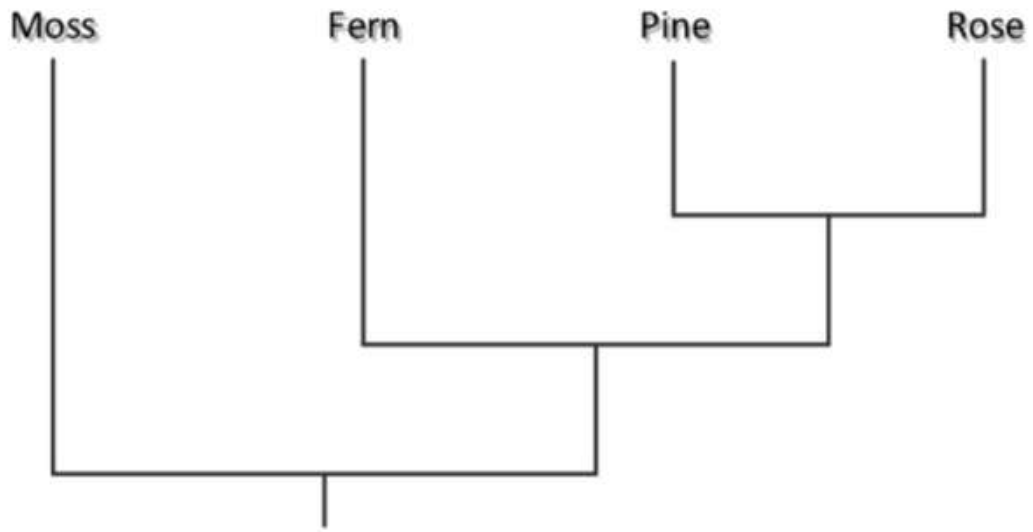
Similarly, each lineage has ancestors that are unique to that lineage and ancestors that are shared with other lineages (common ancestors).

Phylogenies do not represent "advancement" or "progress"



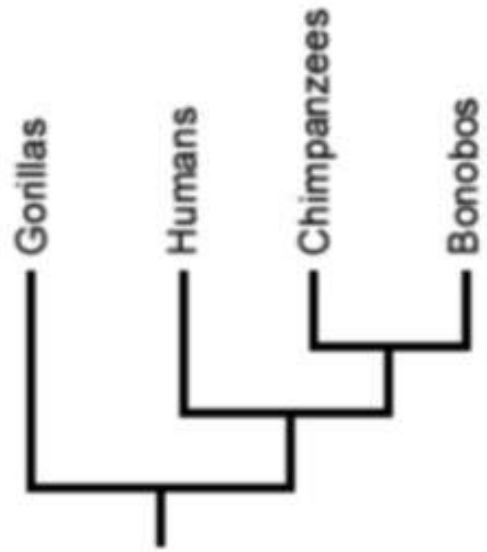
Starting from the ancients philosophers (Aristotle, for example) and up to very recently, we had the erroneous idea that life can be organized on a ladder of lower to higher organisms.

Phylogenies do not represent "advancement" or "progress" (2)



Although mosses branch off early on the tree of life and share many features with the ancestor of all land plants, living moss species are not ancestral to other land plants; nor are they more primitive.
Mosses are the "cousins" of other land plants.

Evolution of humans



1. Humans did not evolve from chimpanzees: humans and chimpanzees are evolutionary "cousins" and share a recent common ancestor that was neither chimpanzee nor human

2. Humans are not "higher" or "more evolved" than other living lineages: since our lineages split, humans and chimpanzees have each evolved traits unique to their own lineages

Generating phylogenetic trees

We can use data that falls into two categories:

1 numerical data (distances between objects)
2 non-numerical data (e.g., anatomical features, DNA sequences, etc.)
non-numerical data like DNA sequences can be converted into numerical data

So, how does a distance matrix look like?

Species	Dog	Bear	Raccoon	Weasel	Seal	Seal Lion	Cat	Monkey
Dog	0	-	-	-	-	-	-	-
Bear	32	0	-	-	-	-	-	-
Raccoon	48	26	0	-	-	-	-	-
Weasel	51	34	42	0	-	-	-	-
Seal	50	29	44	44	0		-	-
Seal Lion	48	33	44	38	24	0	-	-
Cat	98	84	92	86	89	90	0	-
Monkey	148	136	152	142	142	142	148	0

Important!

If alignment data is used, the number of mismatches between the sequences will be a measure of the distance between the sequences (fewer mismatches = closer distance).

UPGMA

- stands for Unweighted Pair Group Method with Arithmetic mean
- simplest method for tree reconstruction
- requires a distance matrix
- creates a rooted tree (many methods don’t)

UPGMA: the basic algorithm

1. build a distance matrix
2. find the shortest distance in the matrix, and cluster the corresponding elements into a composite group (a cluster)
3. recompute a new distance matrix: the distance between the composite group and another element is the average of the distances between each item in the group and the element
4. repeat steps 2 and 3 until only two groups are left

UPGMA: an example

Example 1 - Alignment between five hypothetical homologous DNA sequences (differences between **A** and **B**, and **D** and **E** are indicated)

```
          10         20         30         40         50
A: GTGCTGCACG GCTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
B: ACGCTGCACG GCTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
C: GTGCTGCACG GCTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
D: GTATCACACG ACTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
E: GTATCACATA GCTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

Initial distance matrix

Species	A	B	C	D	E
A	0	-	-	-	-
B	9	0	-	-	-
C	8	11	0	-	-
D	12	15	10	0	-
E	15	18	13	5	0

D and E are "closest." Form a composite group. (See next slide for an important note.)

UPGMA: an example (2)

Distance between remaining species and group (DE) are determined by taking the average distance between its two members (D and E) and all other remaining species. e.g., $d_{(DE)A} = \frac{1}{2}(d_{AD} + d_{AE}) = \frac{1}{2}(12 + 15) = 13.5$

New distance matrix:

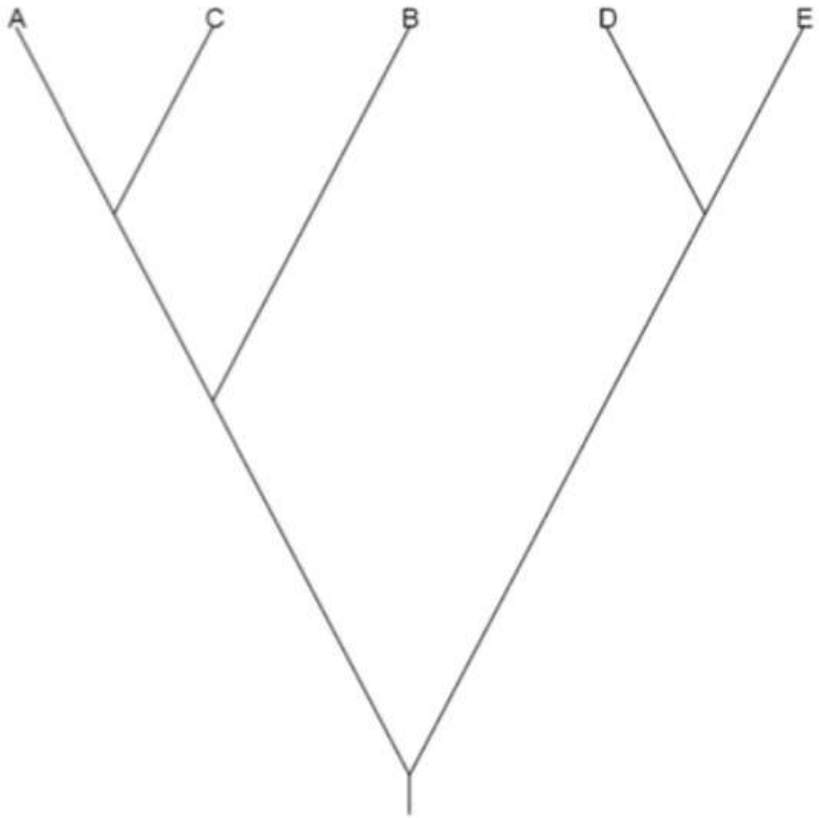
Species	A	B	C	(D,E)
A	0	-	-	-
B	9	0	-	-
C	8	11	0	
(D,E)	13.5	16.5	11.5	0

A and C are "closest." Form a composite group.

When short sequences are compared two or more species can be "closest." In this case, arbitrarily select one and proceed. (Means that at least two trees. both equally correct, could be produced.)

UPGMA: an example (4)

The generated tree




```
(((A,C),B),(D,E));
```

Generated at *http://iubio.bio.indiana.edu/treeapp/treeprint-form.html*