

# Bioinformatics: why?

- To understand biology
  - e.g. Understand rules underlying biological mechanisms
- To replace other methods/experiments (save costs)
  - e.g. Predict protein structure from sequence
- To understand the past
  - e.g. Evolution of species – phylogenetic analysis
- To predict the future
  - e.g. Medical diagnosis
- To improve organism
  - e.g. Make better beer



# Bioinformatics: how?

- **World Wide Web**

- **Online Database and Tools:** most of them are free of charge, many of them are open source
- **Bioinformatics Journals:** Bioinformatics, PLoS Computational Biology, BMC Bioinformatics, Genome Research, Nucleic Acid Research etc

- **Development**

- Mostly UNIX based
- Computer Programming: BioPerl (also BioPhyton, Bioconductor, BioJava)

# Bioinformatics: how?

- National Center for Biotechnology Information (NCBI)

The image shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI Resources" and "How To" menus, and a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. On the left side, there is a vertical navigation menu with categories like "NCBI Home", "Resource List (A-Z)", "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area is titled "Welcome to NCBI" and includes a brief description of the center's mission. Below this, there are six main sections: "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), and "Research" (Explore NCBI research and collaborative projects). On the right side, there are sections for "Popular Resources" (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and "NCBI News & Blog" (The entire corpus of the Sequence Read Archive (SRA) now live on two cloud platforms!, New ribosomal RNA BLAST databases available on the web BLAST service and for download, NCBI staff to present 3 posters at Advances in Genome Biology and

# Bioinformatics: how?


- European Bioinformatics Institute (EBI) – part of the European Molecular Biology Laboratory (EMBL)


Overview | A to Z | Data submission | Support

## Tools & Data Resources


Search all tools & data resources


### Tools


**Clustal Omega**   
Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.  
**Multiple sequence alignment**

**InterProScan**   
InterProScan searches sequences against InterPro's predictive protein signatures.  
**Protein feature detection**  
**Sequence motif recognition**










### Data resources

**Ensembl**   
Genome browser, API and database, providing access to reference genome annotation

**UniProt**   
A comprehensive resource for protein sequence and functional annotation.

**PDBe**   
The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

### Browse by type

 DNA & RNA	 Gene Expression	 Proteins
 Structures	 Systems	 Chemical biology
 Ontologies	 Literature	 Cross domain

## Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services

# Bioinformatics: definition

**Bioinformatics** is «the **computational handling** and processing of **genetic information**»

Ouzounis CA & Valencia A, 2003

# Handling of genetic information

Apply ( or implement new) efficient techniques for:

- Description
- Storage
- Retrieval
- Interconnectivity

of a huge amount of complex and non-homogenous data

# Handling of genetic information

Attention!

- Origin and Quality of biological data
- Data Annotation: automated or expert-based (curated)
- Interconnectivity
- User friendly interface

# Processing of genetic information

GOAL 1: Answer biological questions, e.g.:

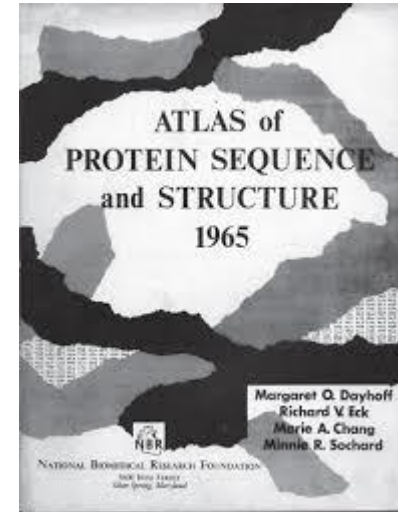
- Does molecule A interact with molecule B?
- What is the 3D structure of molecule X?
- How does the 3D structure of molecule X affect its function?

GOAL 2: Answer other scientific or technical questions, e.g.:

- What is the optimal way to store genetic data in a database?
- How can I test if my results are statistically significant?
- How can I compare DNA, RNA or protein sequence?

# Biological Databases: why?

- Exponential growth in biological data
  - Atlas of protein sequence and structure
    - Vol I, 1965, 65 proteins
    - November 2013
      - Proteins >44 M protein sequences, >95000 structures
      - Nucleotide sequences > 163M
- Computer-readable form: easier analysis, handling and sharing of biological data (well-defined file formats)
- Data are no longer published in a conventional manner, but directly submitted to databases



**ESSENTIAL TOOLS FOR BIOLOGICAL RESEARCH**

# Biological Databases: what?

- **Definition:**

Libraries of life science information, collected from scientific experiments, published literature, high-throughput experiment technology and computational analysis

- **Important features:**

Structured, searchable, updated periodically, cross-referenced

- **Types of databases:**

Nucleotide sequence, protein sequence/structure/motifs, gene expression data, metabolic pathways etc

# FASTA Format

- The most established format representing nucleotide or protein sequences
- First line (Starting with >): Description line
- Remaining lines: sequence (single letter abbreviations)

```
>AAD44166.1 cytochrome b, partial (mitochondrion) [Elephas maximus maximus]  
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSaipYIGTNLV  
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG  
LLILLLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL  
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX  
IENY
```

# Reference Sequence (RefSeq) Database

- Problem

Redundancy is a major problem in primary/archival sequence databases (GenBank/ENA), i.e. There are multiple entries of the same loci

- Solution

Secondary/curated database that culs best available information for each entry = RefSeq

# Reference Sequence (RefSeq) Database

- RefSeq records appear in a similar format as GenBank records from which they are derived but can be distinguished by the accession prefix which contains an underscore symbol (\_) ex. NG\_00007

Accession Prefix	Molecule type
NC_	Genomic; complete genomic molecule
NG_	Genomic; incomplete genomic molecule
NM_	mRNA
NR_	RNA
NP_	Protein; usually associated with NM_ or NC_
XM_	mRNA; predicted molecule
XR_	RNA; predicted molecule
XP_	Protein; predicted model usually associated with XM_

# GenBank vs RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records from same loci	Single records from each molecule of major organism
Records can contradict each other	
No limit to species included	Limited to model organisms
Data Exchange among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein Databases

# Sequence Similarity and Alignment

# Similarity

- Similarity due to what? How do we define similarity?

*Similar due to  
inheritance*



*Similar due to...  
uh...other factors*



# Sequence alignment – evolutionary basis

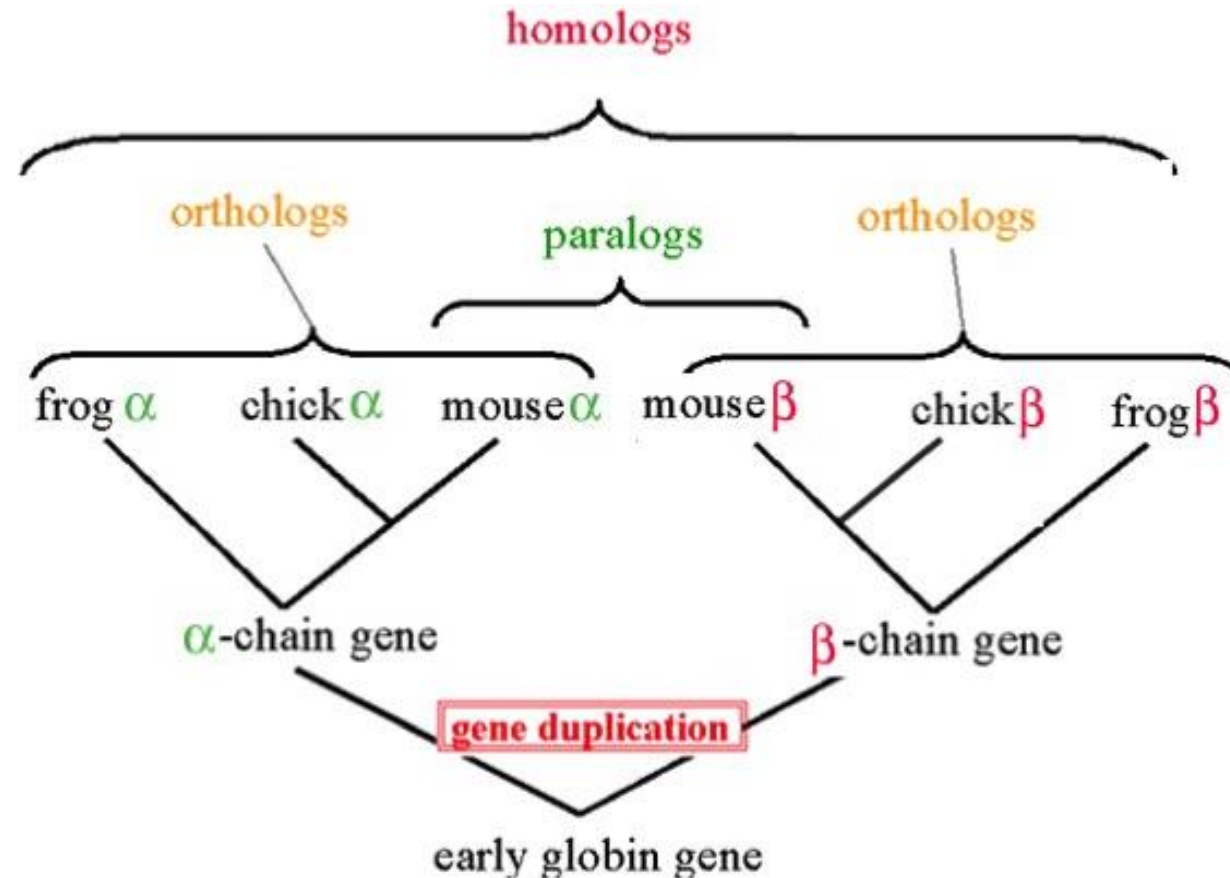
- Find regions of conservation or variability that may be important for their structure and/or function
- Evolutionary history of macromolecules and possible ancestors
- Evolutionary mechanisms that lead to the sequences we observe today

# Sequence alignment – evolutionary basis

- High sequence similarity provides strong **identification** of common evolutionary ancestors (homology) and similar structure/function
- But can also occur due to convergent evolution or because of chance (for short sequences)

# Homology

- Homology: similarity due to descent from common ancestor



# Sequence variations

- Sequences may have diverged from a common ancestor through various types of mutations:
  - Substitutions: **ACTA** → **AGTA**
  - Insertion: **ACTA** → **AC**G**TAA**
  - Deletion: **ACTA** → **ATA**

Indels

# Sequence alignment

## Key Questions:

1. What do we want to align?
2. How do we score an alignment?
3. How do we find the optimal alignment?

# 1. What do we want to align?

- Global alignment: find the best match of both sequences in their entirety (whole sequence)
- Local alignment: find the best sequence match (in some sequences)

# How do we score an alignment?

- Match = reward (+ score), Mismatch= penalty (- score)
- Allow gaps in the alignment = define a gap penalty

- Match score: +1
- Mismatch score: +0
- Gap penalty: -1

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

- Matches:  $18 \times (+1)$
- Mismatches:  $2 \times 0$
- Gaps:  $7 \times (-1)$

**Score = +11**

## 2. How do we score an alignment?

- More questions: Do all amino acids mutate with the same rate and do they have the same effect? NO!

For instance, hydrophilic residues (ex. Arginine) is more likely to be replaced another hydrophilic residue (ex. Glutamine), than it is to be mutated into a hydrophobic residue (ex. Leucine)

### **Substitution Matrix**

# Substitution Matrix

- Probabilistic model for scoring alignments
- Usually takes in account;
  - Evolutionary relationship
  - Structural similarities
  - Physicochemical properties
- Most popular/established substitution matrices:
  - PAM matrices (Dayhoff et al., 1978)
  - BLOSUM matrices (Henikoff & Henikoff, 1992)

## BLOSUM 62 scoring matrix

(positive values are shaded)

<b>A</b>	<b>4</b>																			
<b>R</b>	-1	<b>5</b>																		
<b>N</b>	-2	0	<b>6</b>																	
<b>D</b>	-2	-2	<b>1</b>	<b>6</b>																
<b>C</b>	0	-3	-3	-3	<b>9</b>															
<b>Q</b>	-1	<b>1</b>	0	0	-3	<b>5</b>														
<b>E</b>	-1	0	0	<b>2</b>	-4	<b>2</b>	<b>5</b>													
<b>G</b>	0	-2	0	-1	-3	-2	-2	<b>6</b>												
<b>H</b>	-2	0	<b>1</b>	-1	-3	0	0	-2	<b>8</b>											
<b>I</b>	-1	-3	-3	-3	-1	-3	-3	-4	-3	<b>4</b>										
<b>L</b>	-1	-2	-3	-4	-1	-2	-3	-4	-3	<b>2</b>	<b>4</b>									
<b>K</b>	-1	<b>2</b>	0	-1	-3	<b>1</b>	<b>1</b>	-2	-1	-3	-2	<b>5</b>								
<b>M</b>	-1	-1	-2	-3	-1	0	-2	-3	-2	<b>1</b>	<b>2</b>	-1	<b>5</b>							
<b>F</b>	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	<b>6</b>						
<b>P</b>	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	<b>7</b>					
<b>S</b>	<b>1</b>	-1	<b>1</b>	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	<b>4</b>				
<b>T</b>	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	<b>1</b>	<b>5</b>			
<b>W</b>	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	<b>1</b>	-4	-3	-2	<b>11</b>		
<b>Y</b>	-2	-2	-2	-3	-2	-1	-2	-3	<b>2</b>	-1	-1	-2	-1	<b>3</b>	-3	-2	-2	<b>2</b>	<b>7</b>	
<b>V</b>	0	-3	-3	-3	-1	-2	-2	-3	-3	<b>3</b>	<b>1</b>	-2	<b>1</b>	-1	-2	-2	0	-3	-1	<b>4</b>
	<b>A</b>	<b>R</b>	<b>N</b>	<b>D</b>	<b>C</b>	<b>Q</b>	<b>E</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>L</b>	<b>K</b>	<b>M</b>	<b>F</b>	<b>P</b>	<b>S</b>	<b>T</b>	<b>W</b>	<b>Y</b>	<b>V</b>

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

### 3. How do we find the optimal alignment?

- Piece of cake: find and score all possible alignments
- Alignment with dynamic programming
  - Global alignment: Needleman and Wunsch, 1970
  - Local alignment: Smith and Waterman, 1981
- Always find the optimal solution
- Problem: Too expensive (in terms of required calculations) and too slow

# Heuristic methods

- Fast approximation to dynamic programming (i.e. Not optimal but acceptable solutions=
- Example: BLAST, BLAT, Bowtie etc.

Trade off between speed and sensitivity

# Basic Local Alignment Search Tool (BLAST)

- Finds regions of local similarities between sequences
- The programme compare nucleotide or protein sequences to sequence databases and calculates the statistical significant of matches.
- BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families

# BLAST

## Steps:

1. Break the query (sequence of interest) into words and score each word (for protein; using a substitution matrix)
2. Select high scoring words and search for similar words in a database
3. Extend words (hits) and find High Scoring Pairs (HSPs)
4. Are HSPs statistically significant? ( E-value)

# E-Value

- E-value: Expectation value that indicates the number of alignments with a score that one can expect to find by chance in a database.
- E-value depends on the database size & query length.
- The closer the E-value to 0, the better the alignment is.
- E.g.:  $E=1e-2$  ( $= 1 \times 10^{-2} = 0.01$ )
- The lower the E-value (close to zero), more significant match.
- Statistics of sequence similarity score