جامعة الملك سعود
King Saud University

Bioinformatics

Nahla Bakhamis, MSc

OUTLINE

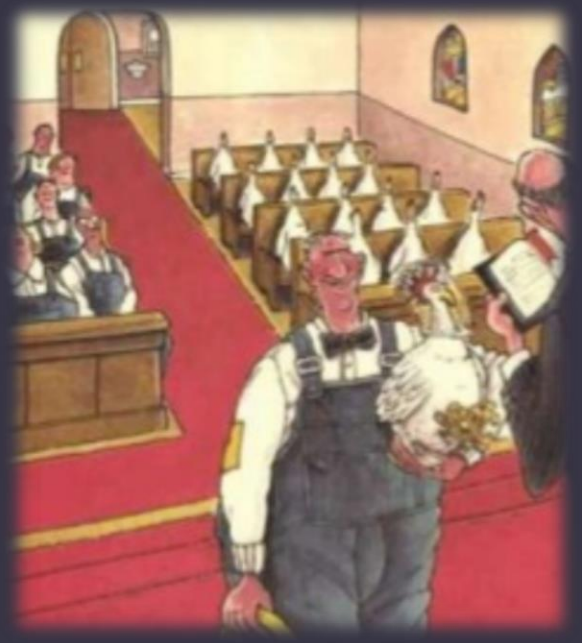
- What is bioinformatics.
- Why bioinformatics
- Types of Data
- Applications
- OMIM workshop
- Primer design workshop

What is bioinformatics ?

- Computational management & analysis of biological data
- Coined by Paulien Hogeweg 1979
- 1980s in genomics and genetics

- Also called; Biocomputing, Systems biology

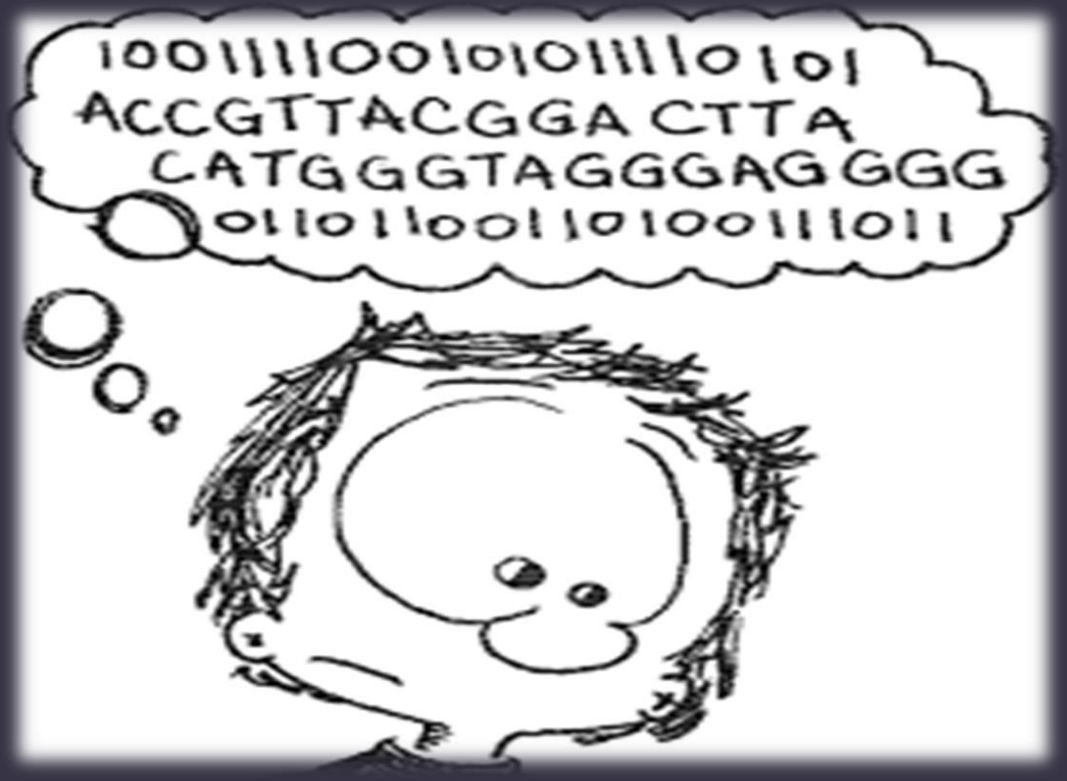
  Computational biology



Aims

- To store maximum amount of data in the internet
- Efficient access/management of data
- Increase understanding of biological process
- Increase research efforts in the field

Why bioinformatics ?



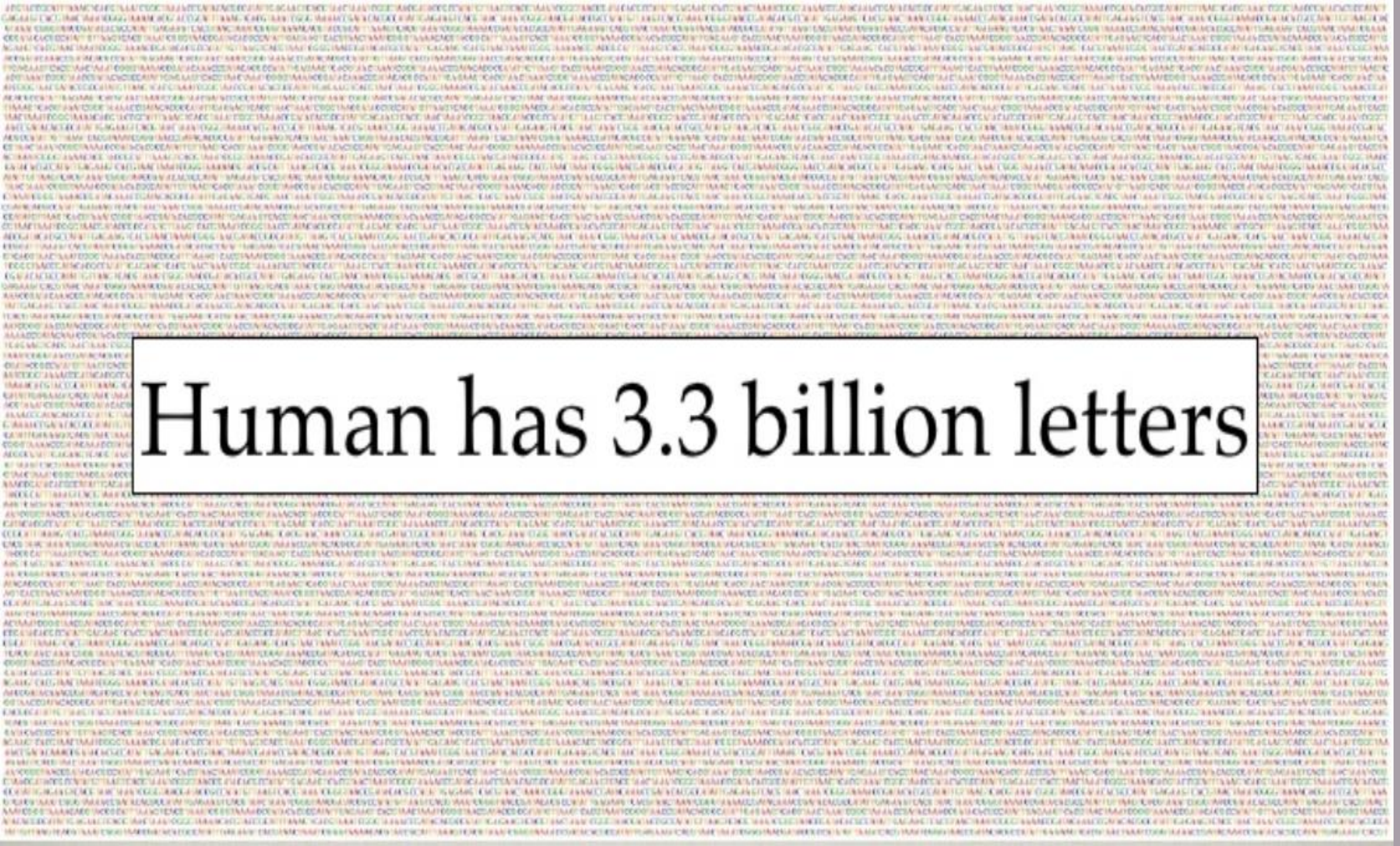
We have the sequence what does it mean ?

ACGTACCGCATTTAAAGTCACGTAAATCGGGTAA
AACCGATACACGCCATATTGAGAAGTCACGTAAC
TAAATCGGGTAACGATACCGCCATATGTTAAGTC
ACGTAAATCGGG[illegible]ACGCCATATTGA
GAAGTCACGTAA[illegible]TAAAACCGATAC
AAACCGATACACGCCATATTGAGAAGTCACGTAA
CTAAATCGGGTAAAACCGATACACGCCATATTGT
TAAGTCACGTAAATCGGGTAACCGATACACGCCA
TATTGAGAAGTCACGTAACTAAATCGGGTAAAAC

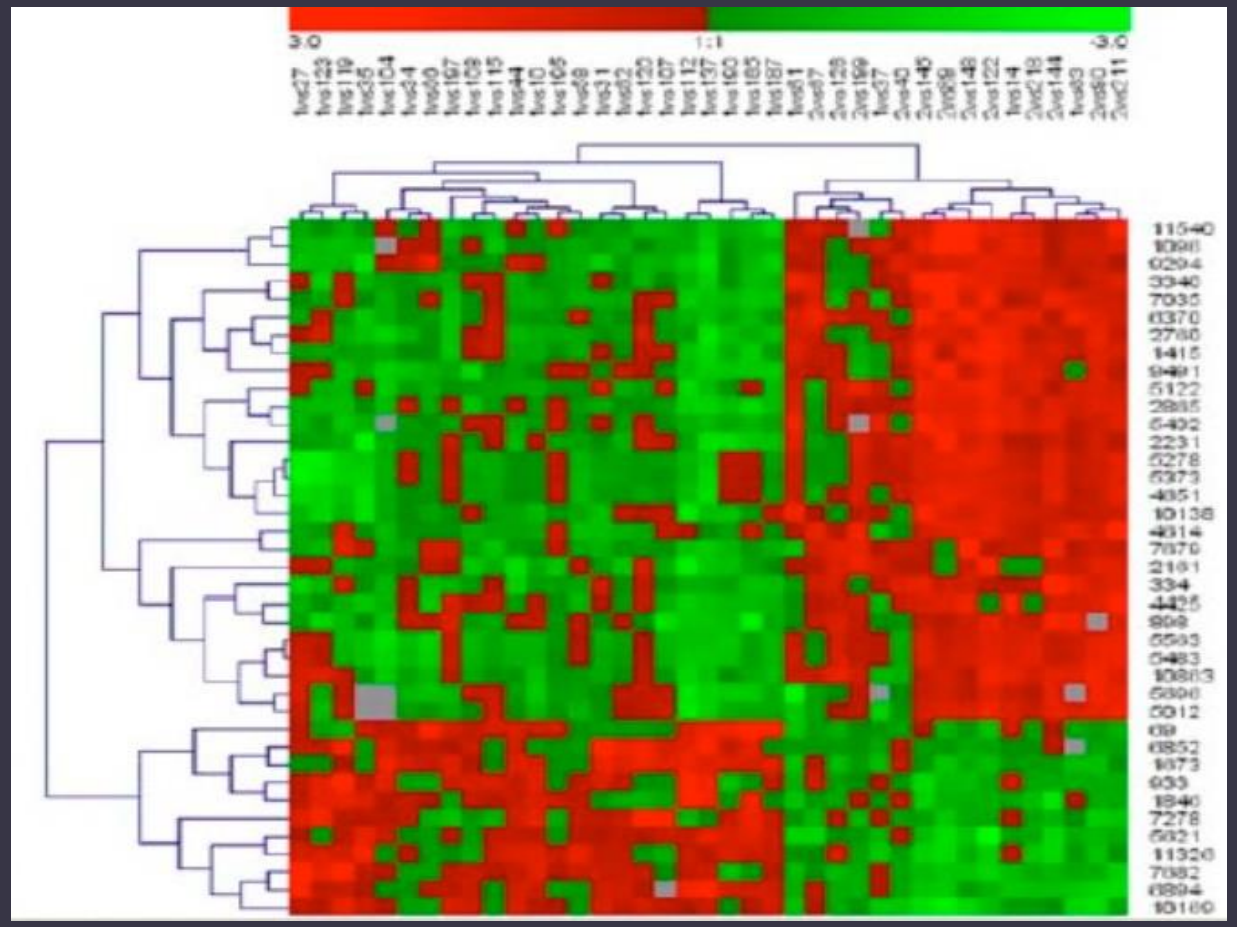
300 letters


5000 letters

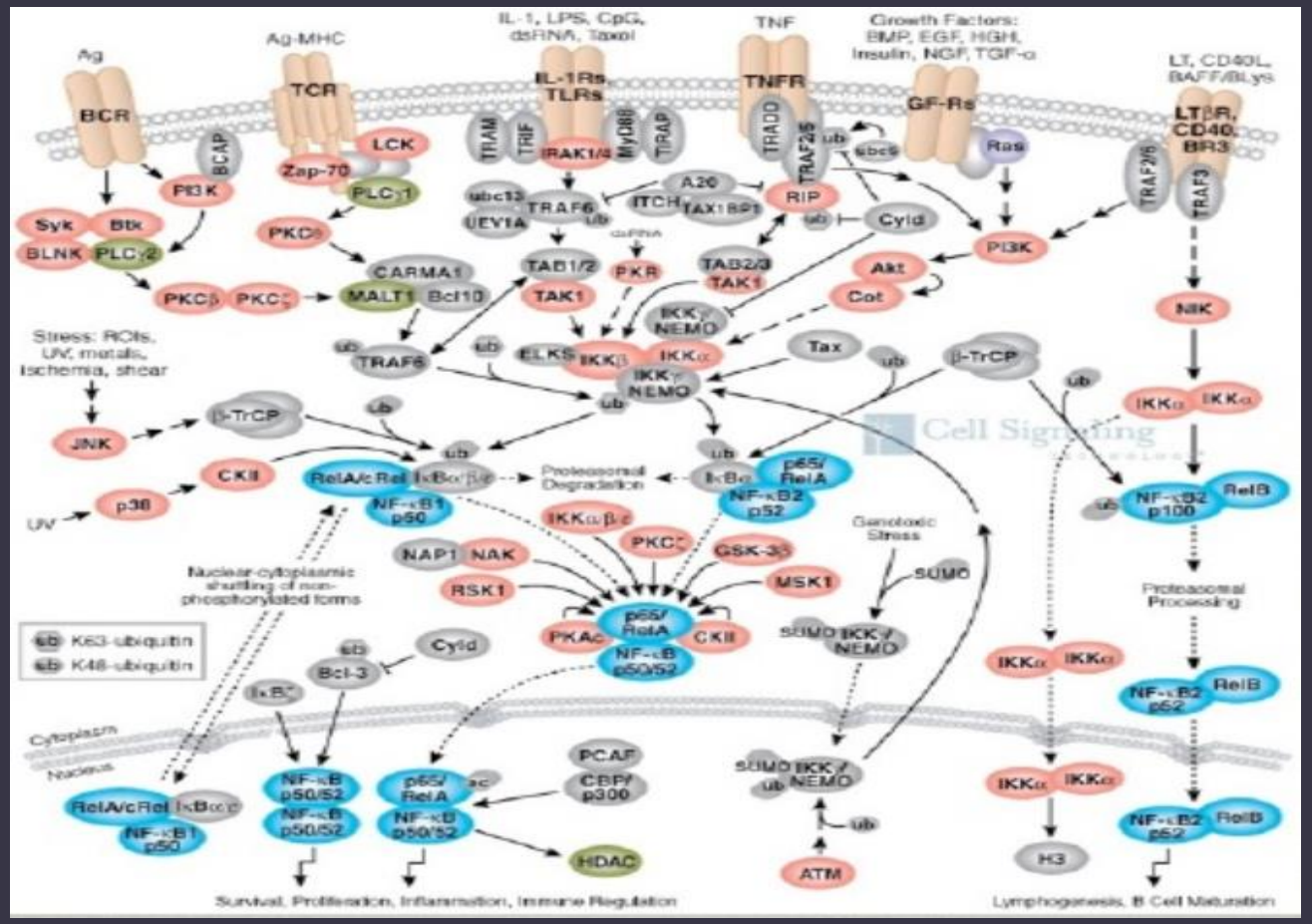

23000 letters


Human has 3.3 billion letters

Not all genes are active



Genes interact with each other



Common activities in bioinformatics

1. Mapping & analysing DNA & protein sequences
2. Aligning and compare different DNA & protein sequence
3. Creating & viewing 3D modules of protein structure

Data classification

- **Primary data:**

  Row/basic data eg. DNA or aa seq (building blocks)

- **Secondary data:**

  arrangement of aa in a protein

- **Tertiary data:**

  more complicated, related to 3D structure of proteins

Unit of information

- DNA (Genome)
- RNA (transcriptome)
- Proteins (Proteome)

- Or genetic, genomic and metabolic

DNA

- Simple seq analysis (database searching)
- Regulatory regions
- Gene finding
- Whole genome annotations
- Comparative genomics between species and strains

DNA

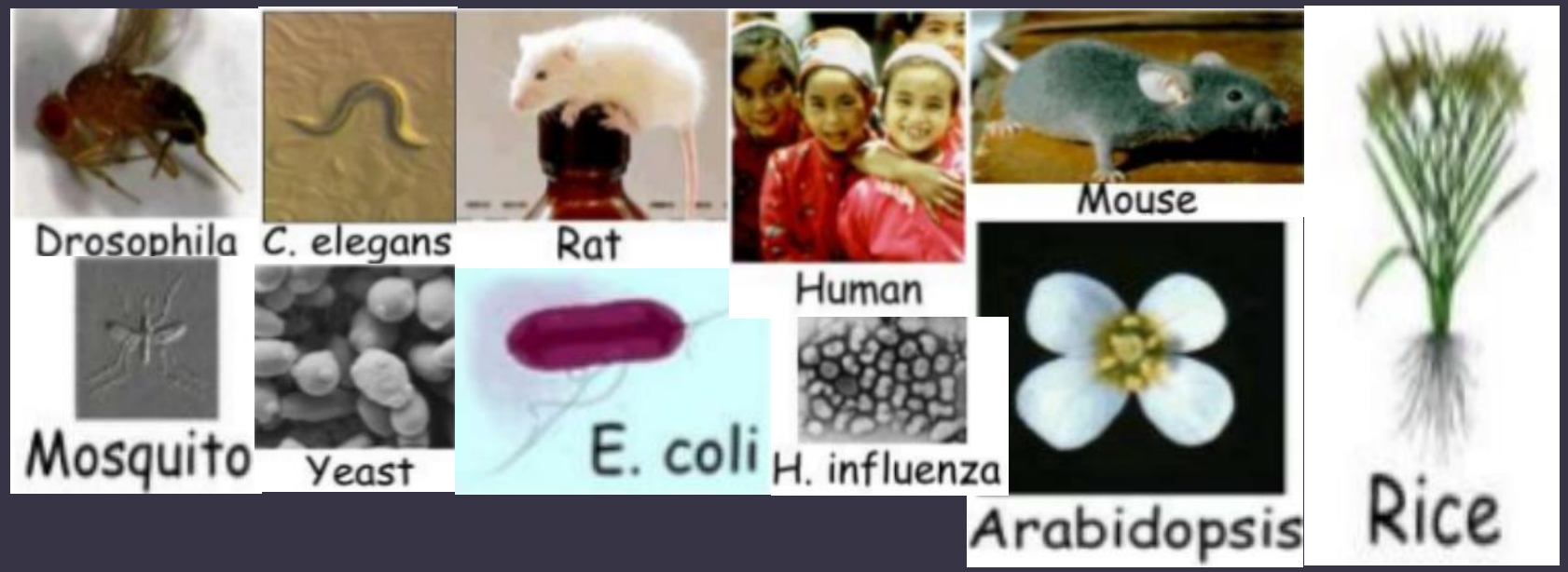
- Row DNA sequence;

  coding or non-coding?

  pares into genes?

```
atggcaattaaaattggtatcaatggt
tttggtcgtatcggccgtatcgtattc
cgtgcagcacaacaccgtgatgacatt
gaagttgtaggtattaacgacttaatc
gacgttgaatacatggcttatatgttg
aaatatgattcaactcacggtcgtttc
gacggcactgttgaagtgaaagatggt
aacttagtggttaatggtaaaactatc
cgtgtaactgcagaacgtgatcca
```

Whole genomes

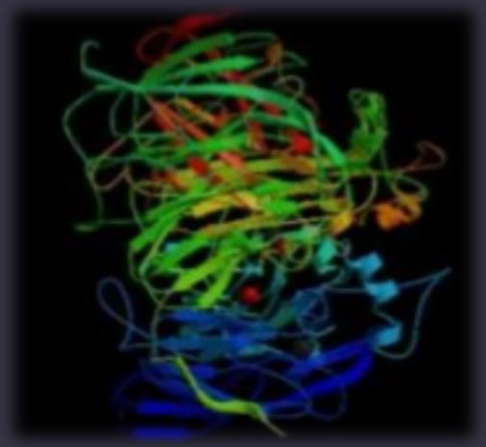


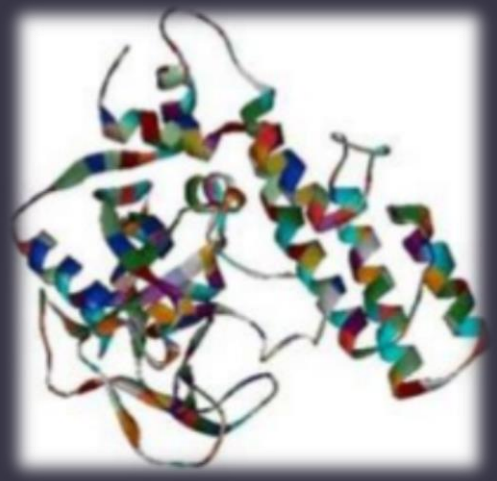
RNA

- Tissue specific expression
- Structure
- Single gene analysis
- Experimental data
- Micro-array and expression array analyses

Protein

- Proteome of an organism
- Mass specific
- 2D,3D 4D structures (interactions)





Other integrative data

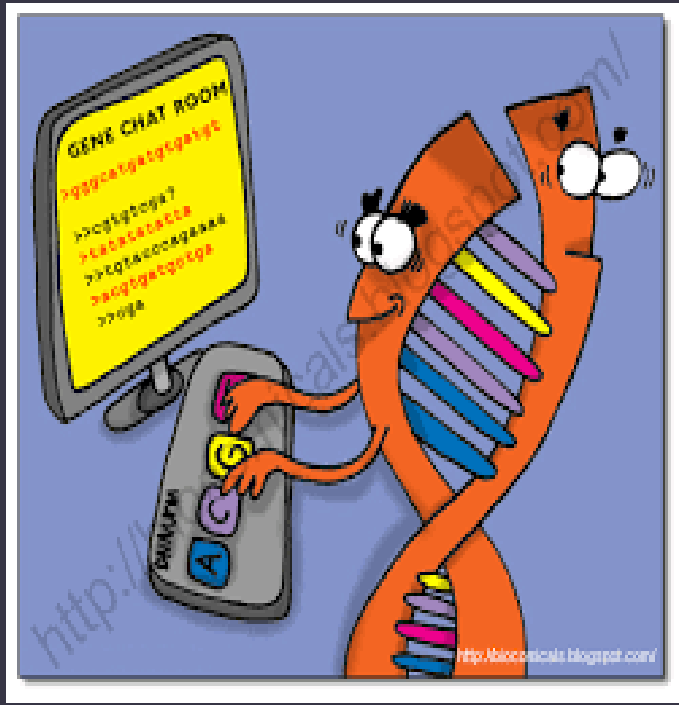
- Metabolic pathways
- Regulatory networks
- Whole organisms phylogeny
- Environments, habitats, ecology

Applications

- **Medical**
  - ✓ understanding life process in healthy & disease states
  - ✓ SNPs
- **Pharmaceutical and biotech industry**
  - ✓ develop new drug or gene/structural base drug design
- **Agricultural applications**
  - ✓ higher yields crops

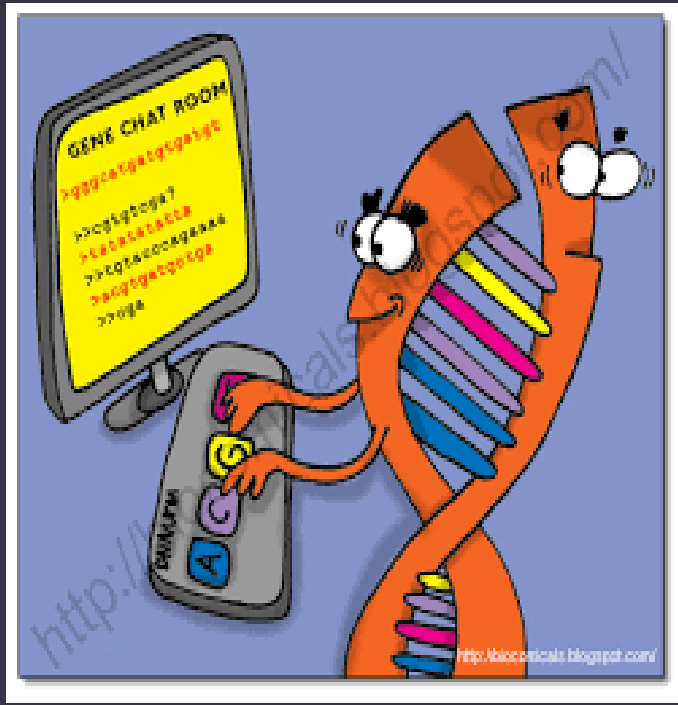
Biological problems computers can help with

- I cloned a gene - is it a known gene?
- Does the sequence match? Is the sequence any good?
- Does it look like anything else in the database?
- Which family does it belong to?
- How can I find more family members?
- I have an orphan receptor, how can I find its ligand?
- The gene I'm interested in was found in another organism, but not mine. How can I look for it?
- I have linkage to a specific region on chromosome x, how do I find genes in that region?



Biological problems computers can help with

- My advisor wants me to construct a chimeric gene - how do I plan primers? How do I check to know that I have the right sequence?
- I have an RNA sequence with poor expression and I'd like to know its structure.
- I have a protein sequence, how can I find out what it's structure and/or function is?
- How can I cluster protein sequences into families of related sequences and develop protein models?
- I'd like to align similar proteins (or DNA) and generate phylogenetic trees.
- How can I find out which other proteins my sequence interacts with?



Software and tools

- Range from simple command-line to more complex programs
- Web-services available

Data bases

- <u>4 majors</u>

1. Nucleotide data bases
2. Protein data bases
3. Whole genome data bases ENSEMBL
4. Specialized data bases

Nucleotide data bases INSDC

International Nucleotide Data Bank Collaborative

- EMBL

  European molecular biology library (Germany)

- Gene bank. US
- DDBJ

  DNA Data Bank in Japan

Collaborate by international Advisory Meeting

Protein data bases

- 3 majors:

1. Sequence (primary)

   UniProt, SwisProt and PIR

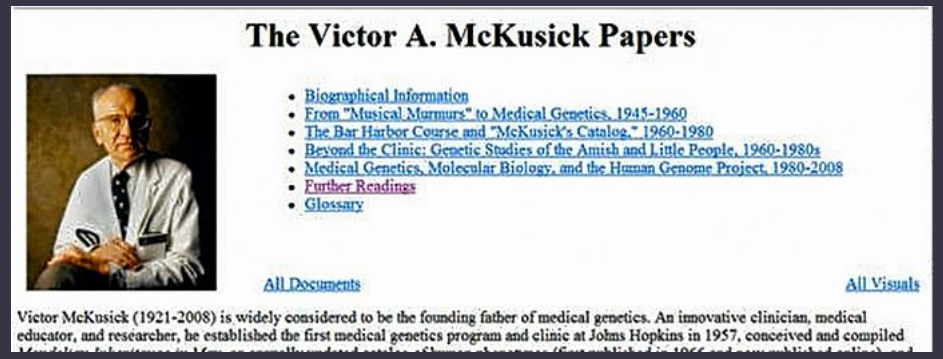
2. Structure

   PDB, SCOP

3. Interactions

Specialized data bases

1. Inherited diseases data bases
   - OMIM (Online Mendelian Inheritance in Man)
   - funded by NHGRI, supported by JHM (copy right)
   - Originally developed by Dr. Victor A. McKusick 1960s
2. Microarray data bases



OMIM

- Focuses on single-gene mendelian disease/disorders/phenotypes

  eg. CF, Sickle cell anemia

- Complex diseases with significant single gene contribution

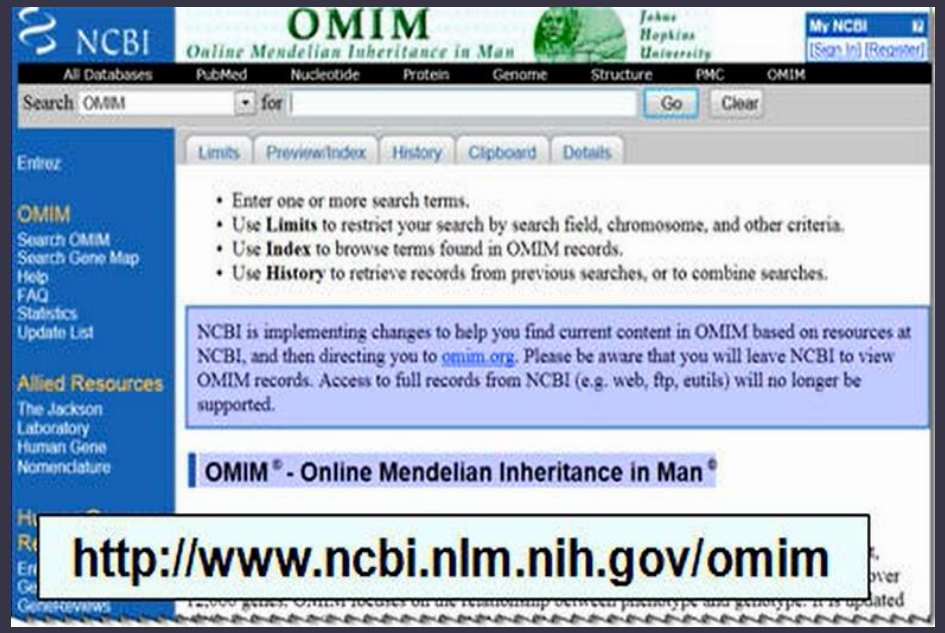
  eg. Complement factor H and age related molecular degeneration

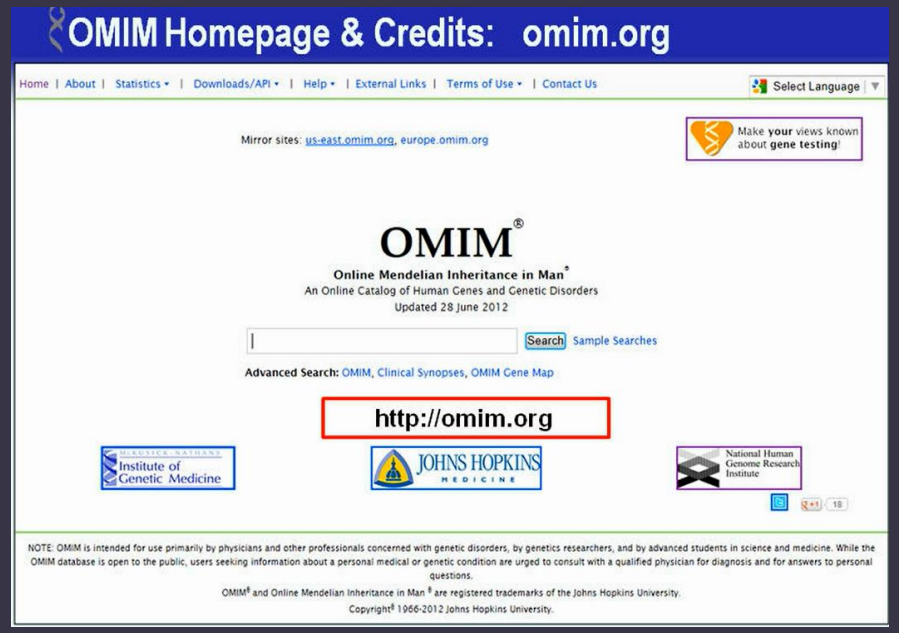
- Descriptions of recurrent deletion and duplication syndromes

  eg. Potocki-Shaffer syndrome, chromosome 10q26deletion syndrome

OMIM (workshop)

- http://omim.org
- http://www.openhelix.com/OMIM





OMIM (workshop)

- <u>You will learn about:</u>

✓ Basic search

✓ Phenotype result

✓ Genotype result

✓ Gene map information

✓ Advanced search

✓ Additional features

✓ Exercises

Primer design

Primer

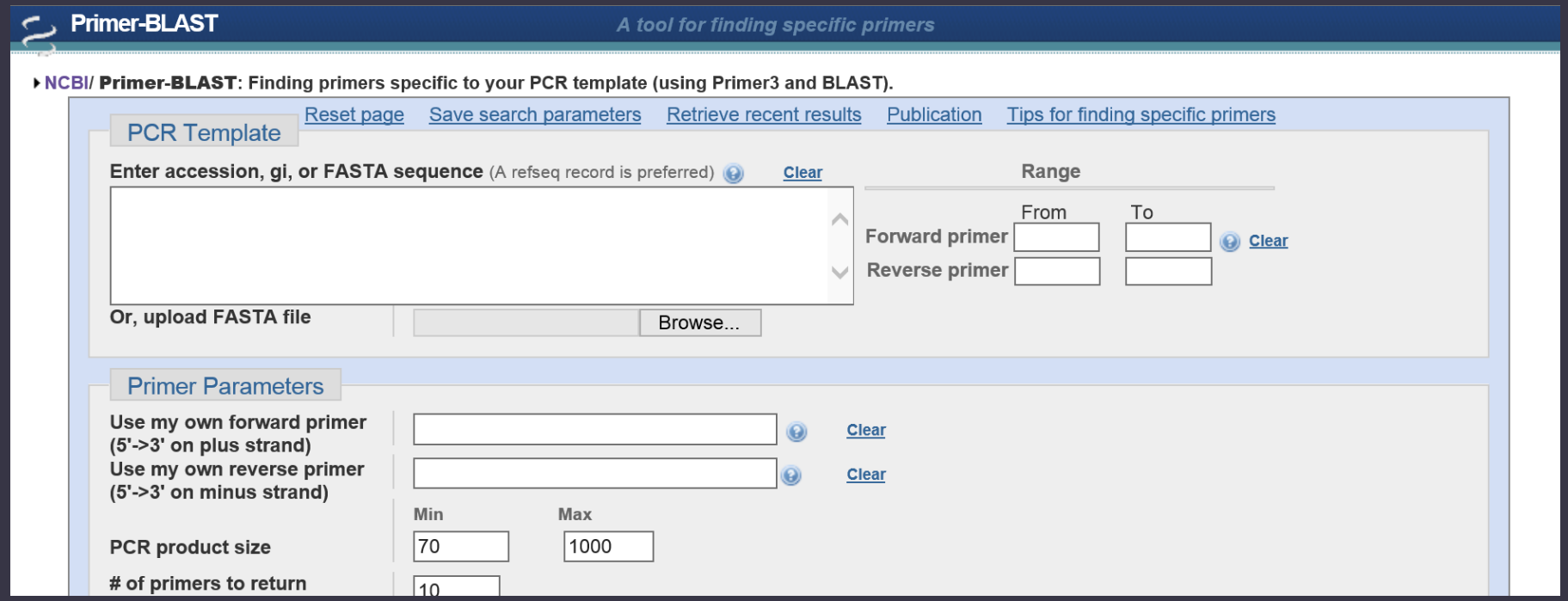
- A strand of nucleic acid serves as starting point for DNA/RNA synthesis
- Why it is required ??
- Polymerase start replication at 3' end of the primer
- PCR and DNA sequencing

Primer design

- NCBI National Centre for Biotechnology Information
- Primer3
- Database of single nucleotide polymorphism dbSNP
- UCSC Genome Browser
- Ensemble Genome Browser

Primer design NCBI

- Tutorial



Primer design

- Points should be taken in consideration:

1. Mononucleotide repeats should be avoided (loop formation)
2. Avoid Primer dimer
3. reverse primer should be the reverse complement of the given seq
4. In TA cloning efficiency can be increased by adding AG tail to 3' & 5' ends

References

- Hunt M. (2006) Real time PCR tutorial - Copyright 2006, The Board of Trustees of the University of South Carolina
- Achuthsankar S Nair Computational Biology & Bioinformatics - A gentle Overview, Communications of Computer Society of India, January 2007
- Aluru, Srinivas, ed. Handbook of Computational Molecular Biology. Chapman & Hall Crc, 2006. ISBN 1584884061.
- Baldi, P and Brunak, S, Bioinformatics: The Machine Learning Approach, 2nd edition. MIT Press, 2001. ISBN 0-262-02506-X
- Barnes, M.R. and Gray, I.C., eds., Bioinformatics for Geneticists, first edition. Wiley, 2003. ISBN 0-470-84394-2
- Baxevanis, A.D. and Ouellette, B.F.F., eds., Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, third edition. Wiley, 2005. ISBN 0-471-47878-4

Patricia, Stock; John, Vanderberg; Itamar, Glazer; Noel, Boemare (2009). "1.6.2. Primers development and virus identification strategies". p. 22. ISBN 978 1 84593 478 1.

Thank you