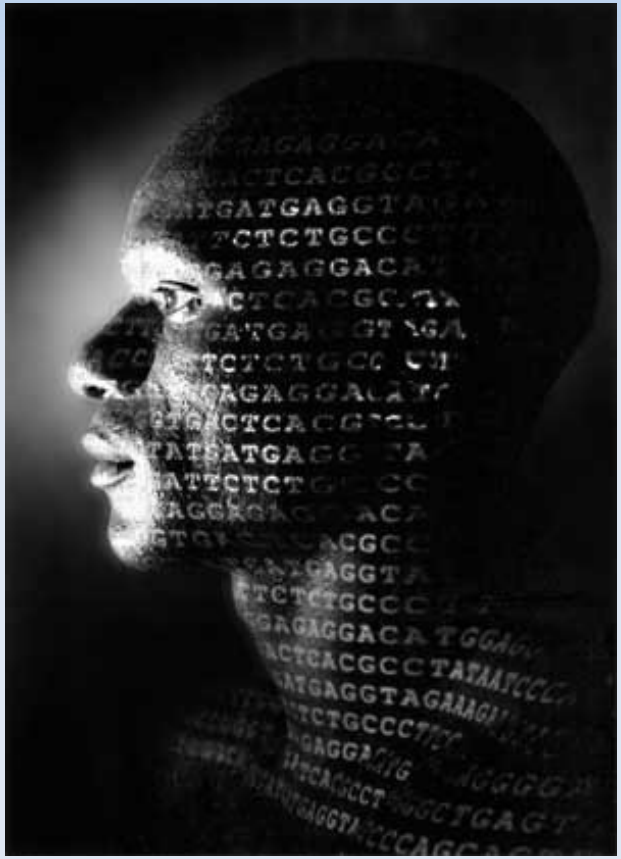
Bioinformatics

Eszter Ari: ELTE, Dept. of Genetics
arieszter@gmail.com

Tamás Korcsmáros



http://falco.elte.hu/bioinfo/english_2010/

username: ***Bioinfo***
password: ***binf***

Marking

? What kind of marks do you need ?

A) A script exam at the end of the semester

AND

B) Solve a bioinformatic exercise at home:

- You can choose a theme / problem
- Or we give you some alternative themes
- And solve it with bioinformatics operation
- Write a doc (as a small scientific paper) about it


...therefore, to optimize the learning experience inherent in the Friday exam, give great heed to Baker v. Carr, study ad infinit... go online and pr... compare & ... expansive detail...
LEXIS®
LEXIS
What law professors say.
Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah Blah-Blah-Blah-Friday exam Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah-Blah Blah-Bla... Blah-Blah-Blah Blah-B... Blah-Blah-Blah-... Blah Blah
LEXIS®
LEXIS
What law students hear.

Syllabus

1. What is bioinformatics? Using computers to solve biological problems. *Eszter Ari*
2. Molecular biology databases. *EA*
3. Sequence comparison, manipulation and alignment. *EA*
4. Sequence similarity searching. *EA*
5. Molecular phylogeny. *EA*
6. The first step of research: Searching and handeling scientific papers. Pulication databases and software. *Tamás Korcsmáros*
7. Bioinformatics of networks, systems biology. Network databases, network analyses. *TK*
8. Structural bioinformatics

What is bioinformatics?

- (Molecular) biology + computers

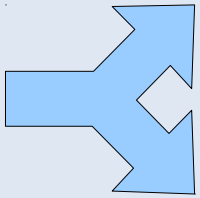
Bioinformatics:

- **In a wider sense:** Computational methods of biology: every kind of biological data processing and evaluation made by computers.
  for examlpe: in the area of supraindividual biology and brain research too
- **In a closer sense:** Computational methods for molecular biology.
- **The primary goal of bioinformatics:** is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (*e.g., pattern recognition, data mining, machine learning algorithms, and visualization*) to achieve this goal.

What is bioinformatics?

- **Major research efforts** in the field include *sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modeling of evolution*.
- It can help to design experimental works too → reduces the costs of laboratory experiments.
- http://www.bioinformatics.org/wiki/Bioinformatics

Bioinformatics



Science: basic and applied science

Business: biotechnological and pharmaceutic industry $$$

Bioinformatics is a tool and scientific research area too

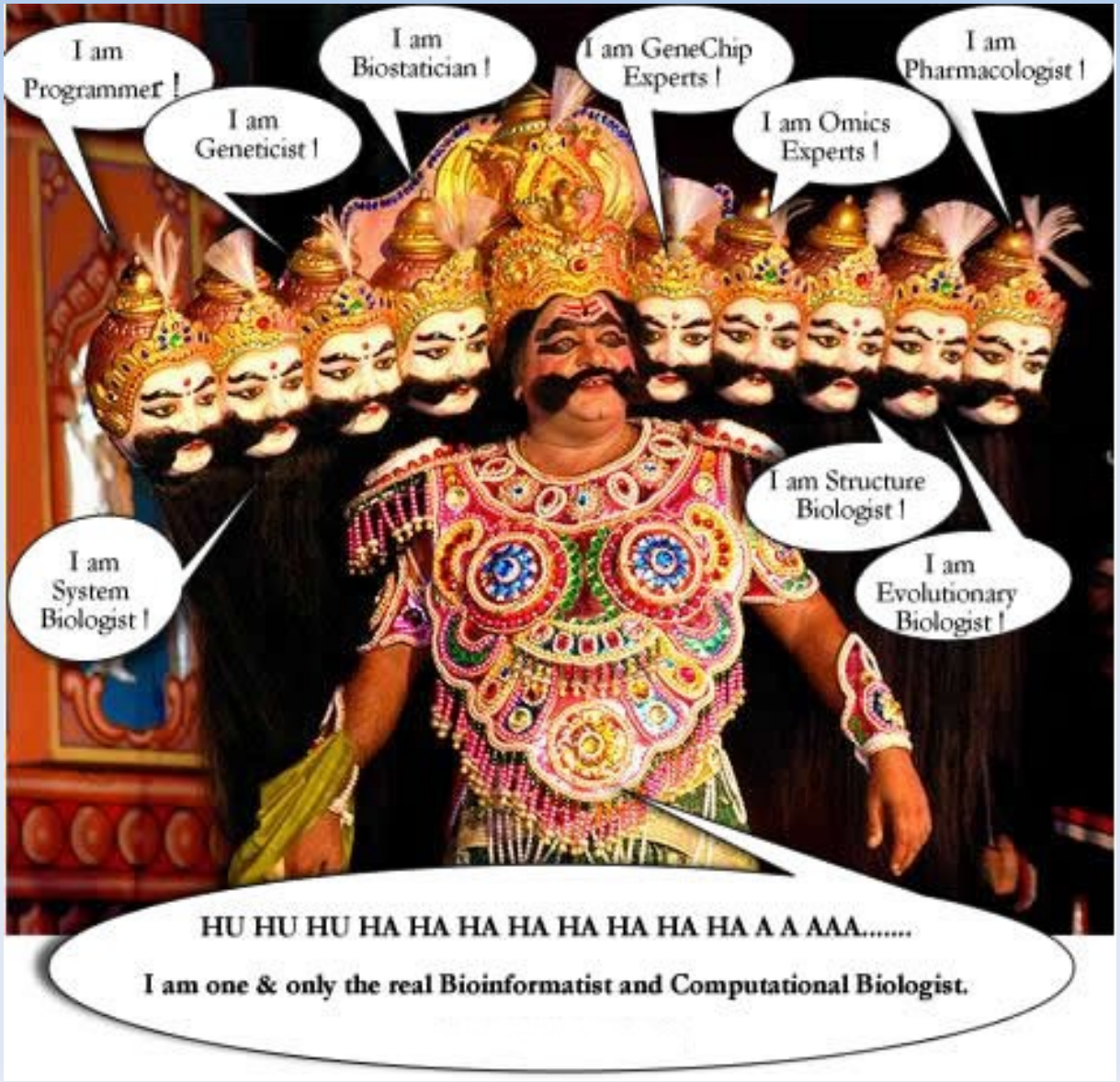
- Bioinformatics as a *tool*:
  - Processing large data
  - Handling great comutational problems
  - Helps to ascertain the structure or function of a molecule
  - The results of bioinformatics should be treated as presumptions till we don't have the experimental evidences
  - So it can't replace the biological experiments. But helps and gives some idea how to design them.
  - ...
- as a *discipline*:
  - Developmenting algorithms
  - Creating databases
  - ...

Bioinformatics go ahead fast

- A lot of biological data → developing databases, algorithms continuously
- A lot of new softwares: mainly on the internet
  - *open source softwares* are specific to bioinformatics
  - Linux, Mac, Windows, INTERNET
  - perl, phyton, java, C++...
- “Low cost” scientific research area

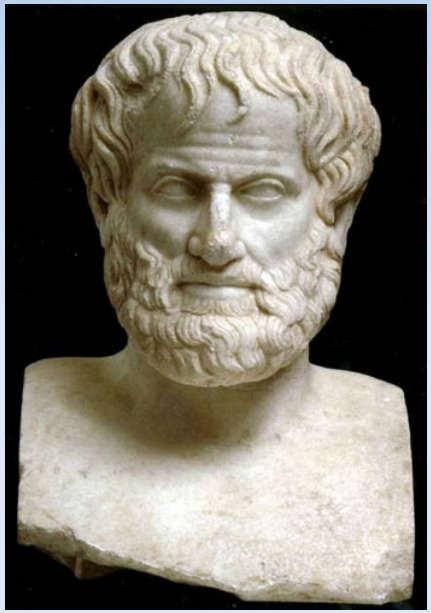
The main area of utilization of bioinformatics

- Genetics
- Omics: genomics, transcriptomics, proteomics
- Biostatistics
- Evolutionary biology
- Structural biology
- Pharmaceutics
- Systems biology
- Ecology
- Anatomy
- ...



The history of bioinformatcs

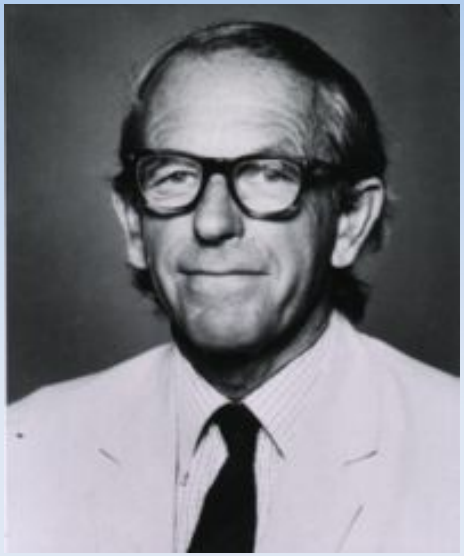
- Already the ancient Greeks ...



  - But bioinformatics is not as old... it's a modern science
  - The date of born of molecular biology is the discoverer of the structure of the DNA molecule in 1953.
    - Bioinformatics mostly based on molecular genetics

Even earlier

- Frederick Sanger, Hans Tuppy 1951. The amino acid sequence of the phenylalanyl chain of insulin. Biochem J. 49:481-490.
  - Frederick Sanger (1918. aug. 13. -) English biochemist, won Nobel prizes twice, discovered the insulin.
  - The main information source for bioinformatician was the protein sequencing by the end of the 70s.

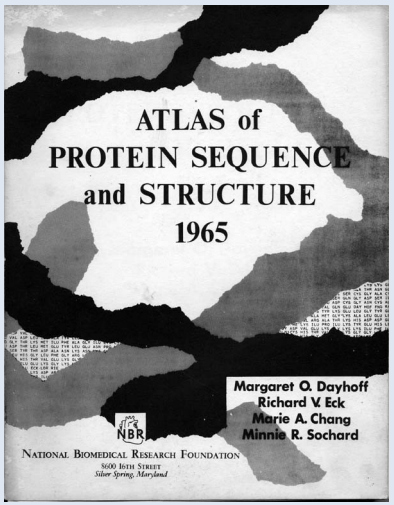


Sanger

Tuppy

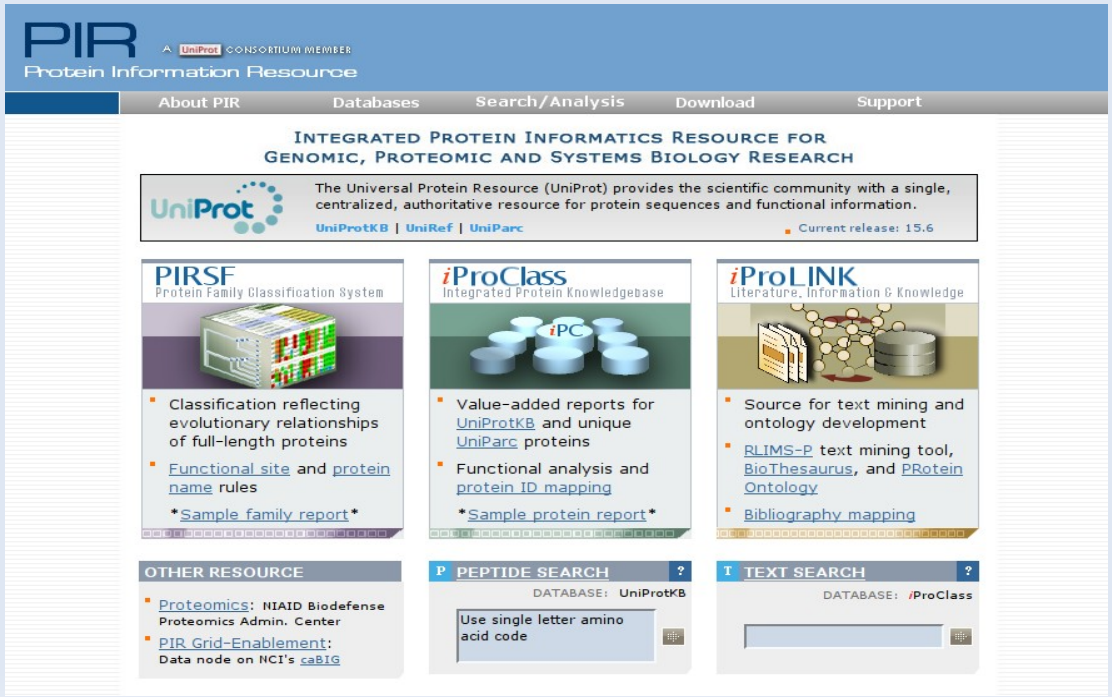
The first bioinformatician

- Margaret Oakley Dayhoff, 1925-1983.
  - National Biomedical Research Foundation (1960. NY State)
  - Atlas of Protein Sequence and Structure
  - Protein superfamilies
  - **PAM matrices:** is a set of matrices used to score sequence alignments. Caltulated from observed mutations in 71 families of closely related proteins. (70s)



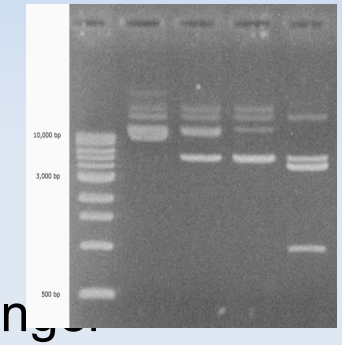
At the 60s, 70s, the beginning of the 80s:
Golden age of protein sequencing

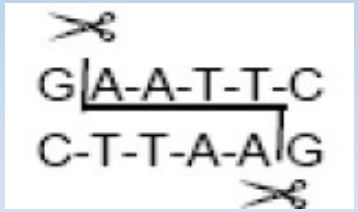
- More and more sequences gathered
- 1984. Atlas → PIR (Protein Information Resource) databank: http://pir.georgetown.edu/
- 1986. SwissPROT → UniProt: http://www.uniprot.org/



The conditions for development of bioinformatics

- New technologies for molecular biology:
  - restriction endonucleases → genetic engineering
  - gel electrophoresis
  - DNA hybridization
  - cloning: DNA, cDNA
  - PCR
  - chain-termination DNA sequencing method of Sanger
  - CHIP
  - Next generation sequencing thechologies
  - → great amount of data
- The development of computers:
  - greater computational capacity
  - internet





Some classical tasks of bioinformatics

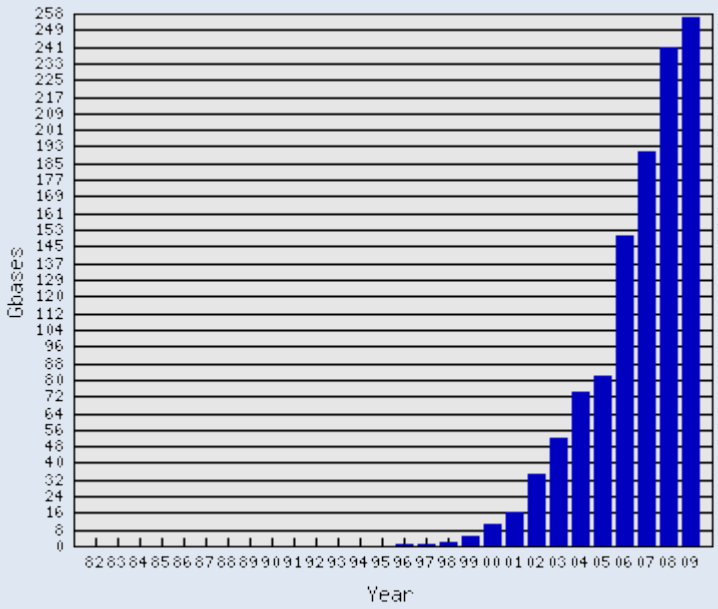
- Sequence examination:
  - Sequence alignment (even genome assembly)
  - Statistical analyses: (e.g. CG ratio, the No. of genes)
  - Genom annotation: ORF and gene finding, searching for exon-intron borders, mapping promoter regions, looking for mobil genetic elements
  - Creating DNA and protein databases: sequences, structures, function, connections, publications
  - Comparing sequences – genomes → e.g. function prediction

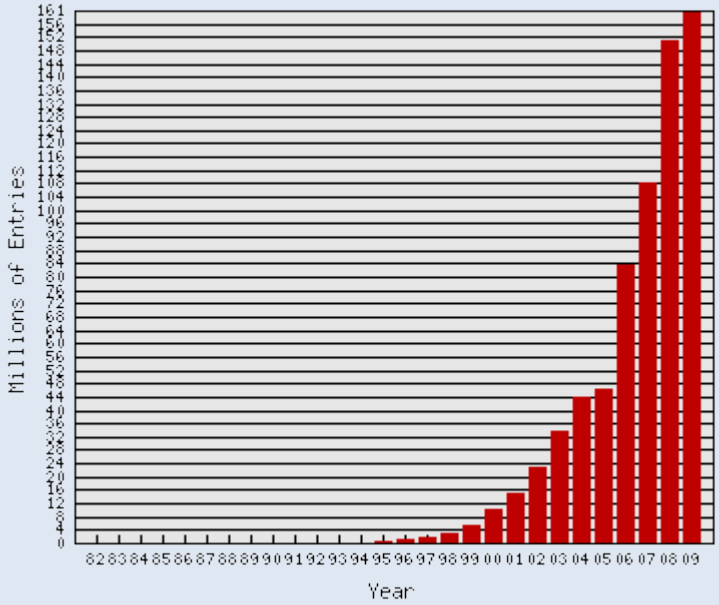
New tasks of bioinformatics

- Connectiong all these informations → networks
  - e.g. protein – protein interactions, signaling pathways

The growth of databases

- The growth of EMBL and NCBI GenBank DNA databases are exponential
- the size of databases doubles in every 9 month
- Novadays: 3 million new sequences / month
- The average length of a sequence from a database is 1000 bps

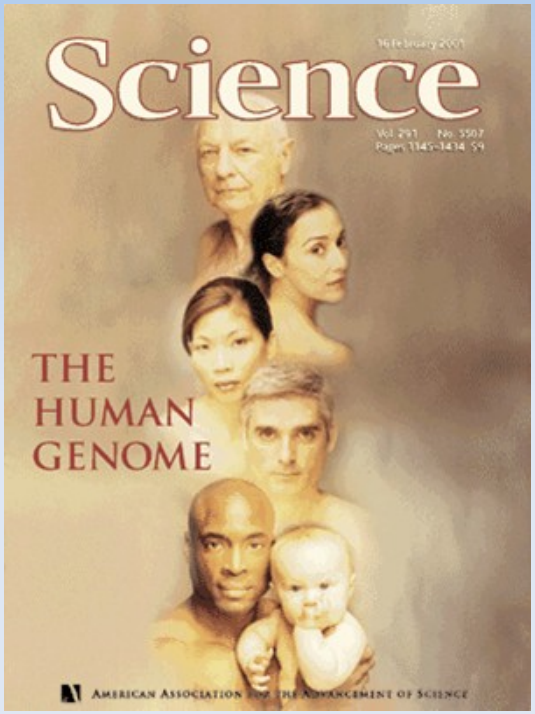




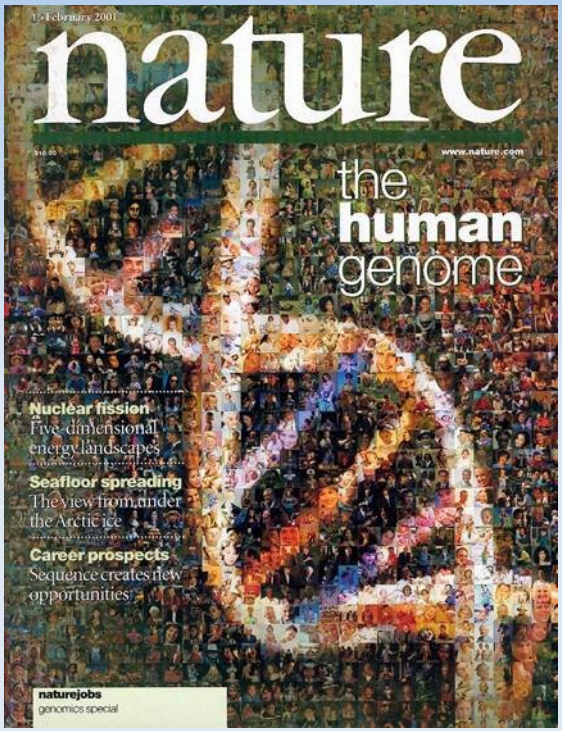
The start of the Human Genome Project

- 1977-1982. Sequencing virus genomes (fX174, lambda, SV40)
- ~1985. Starting of the yeast and C. elegans genome project
- 1988. HGP raises up
- 1990. Human Genome Organisation (HUGO)
- ~1990. Shotgun sequencing, autamatized sequencing of EST-s
- 1995. the yeast and C. elegans maps are ready, HGP starts

A Human Genome: 2000. June 20.



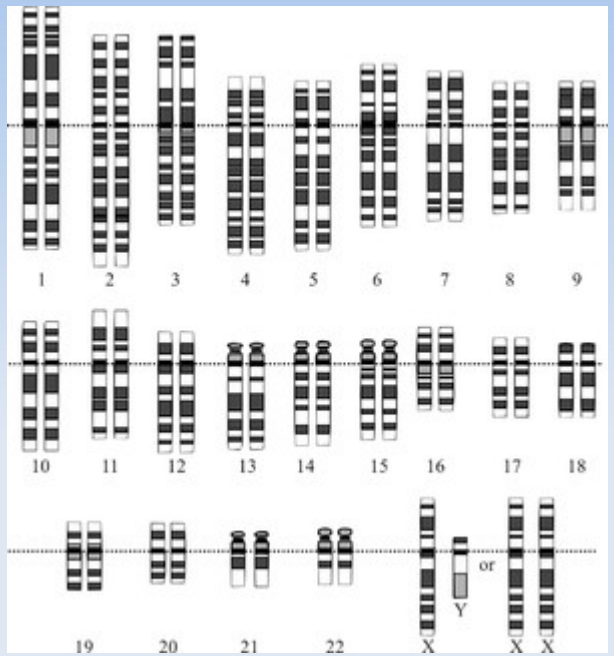
- Dr. Craig Venter
- Celera Genomics



- Dr. Francis Collins & Sir John Sulston
- International Human Genome Sequencing Consortium (IHGSC) (NCBI)

The human genome project

- The two copy is nearly the same (except some repeat sequecnes).
- approximately 3 giga bp = 3.000.000.000 bp = 3×10^9
- The number of our genes is around 23 – 25000 which is much smaller than it has been thought before (around 100.000).

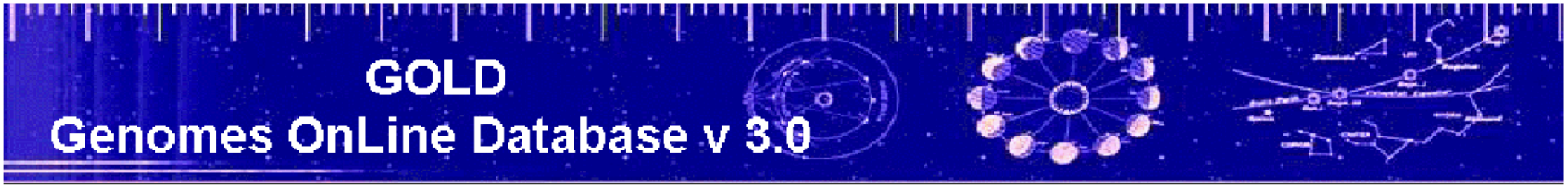


The results of HGP

- The genome of each men is the same in 99.9%
- The protein coding region of our genome is just 1.5%
- We don't know the function of 50% of the genes
- Methodological break-throughs:
  - automatized DNA sequencing
  - PCR
  - bioinformatic improvements

Genome programs

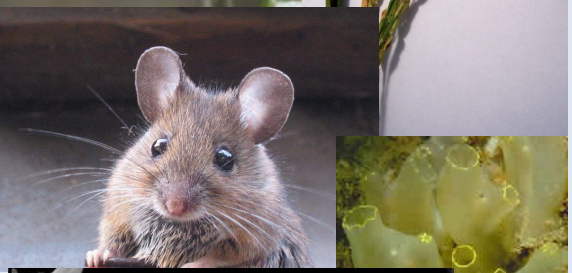
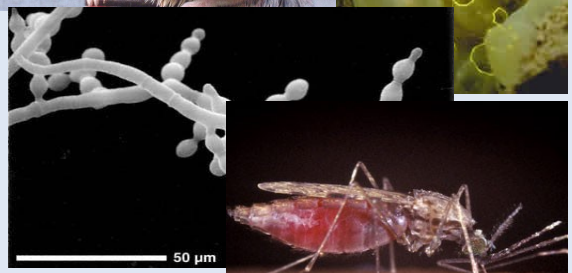
- http://www.genomesonline.org/



Contact: Genomesonline	Last Update: 2010-09-03	Location www.genomesonline.org
1364 Complete Published	Search GOLD: 8225 genome projects	240 Microbial
190 Archaeal Ongoing	4882 Bacterial Ongoing	1549 Eukaryal Ongoing
GOLD RSS Feeds	Click to save all data: DOWNLOAD	METAGENOME CLASSIFICATION
PROJECT TYPE DISTRIBUTION	SEQUENCING STATUS DISTRIBUTION	PHYLOGENETIC DISTRIBUTION

Some important complete genomes



- *Haemophilus influenzae* 1,830 kbp. TIGR 1995 Science 269, 496-512
- *Escherichia coli* K12 4,638 kbp. U. of Wisconsin 1997 Science 277, 1453-1474
- *Saccharomyces cerevisiae* S288C 12,057 kbp. International Collaboration 1997 Nature 387, 5-105
- *Caenorhabditis elegans* 12,069 kbp. Washington University & Sanger Center 1998 Science 282, 1126-1132
- *Arabidopsis thaliana* 115,428 kbp. International Coll. 2000 Nature 408,796-815
- *Homo sapiens* 3000 mbp. International Collaboration 2001 Nature 409,860-921
- *Oryza sativa japonica* 420,000 kbp. Syngenta 2002 Science 296, 92-100
- *Oryza sativa indica* E 420,000 Beijing Genomics Inst. 2002 Science 296, 79-92
- *Mus musculus* 3000 mbp. International Collaboration 2002 Nature 420, 520-62
- *Ciona intestinalis* 116,700 kbp. Joint Genome Institute 2002 Science 298,2157-67
- *Neurospora crassa* 43,000 kbp. Whitehead Institute 2003 Nature 422, 859-68
- *Anopheles gambiae* 228,223 kbp. Celera 2003
- Chicken, cow, dog, etc.

Bioinformatics software packages

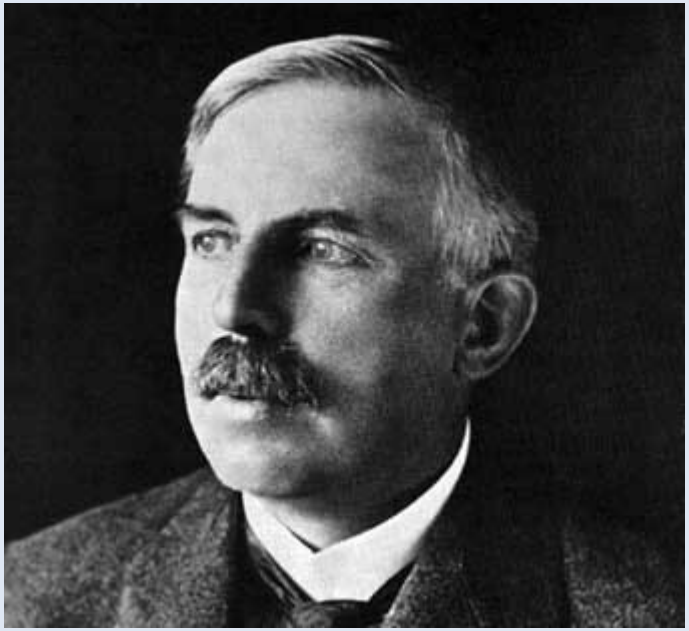
- GCG (pay)
- EMBOSS
- Unipro UGENE
- etc...

- online software collections:
  - EMBL-EBI: http://www.ebi.ac.uk/Tools/
  - NCBI: http://www.ncbi.nlm.nih.gov/
  - Mobyle@pasteur: http://mobyle.pasteur.fr/cgi-bin/portal.py

The science in the past and today...

„All science is either physics, or stamp collecting.”

Rutherford, chemist and physicist, 1871-1937



Biological science in the past and today...

- Past (not too far): The main goal for a scientist was to produce high quality data.

- Today: The main goal for a scientist to interpret the great mass of HQ biological data.

Differences between the two approaches

- Financial: → miniaturization, multiplexization, paralell processing, automatization
- Scale: gene sequencing → genome sequencing, investigate the expression of a gene → microarray, etc.
- Logic: researches based on hypotheses → data mining = searching for questions for the observed data.

Suggested literature

- David W. Mount: Bioinformatics – Sequence and Genome analysis *(online: Google Books)*
- Des Higgins and Willie Taylor: Bioinformatics – Sequence, Structure and Databanks *(online: Google Books)*
- T K Attwood & D J Parry-Smith: Introduction to bioinformatics
- etc...

Thank you for your attention