Bioinformatics: course introduction

Filip Železný and Jiří Kléma

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Intelligent Data Analysis lab
http://ida.felk.cvut.cz

A6M33BIN – Biomedical Engineering and Informatics
B4M36BIN – Open Informatics, Bioinformatics

- Purpose of this course:

  Understand the computational problems in bioinformatics, the available types of data and databases, and the algorithms that solve the problems.
- Methods/Prerequisities
  - mainly: probability and statistics, algorithms (complexity classes), programming skills
  - also: discrete math topics (graphs, automata), relational databases
- Lectures may be held in English
  - OI study program open to foreign students
- Purpose of this lecture

  Sneak informal preview of the major bioinformatics topics

Teachers

Doc. Jiří Kléma
CTU Prague, Dept. of Computer Science
klema@fel.cvut.cz

Prof. Filip Železný
CTU Prague, Dept. of Computer Science
zelezny@fel.cvut.cz

Ing. František Malinka
CTU Prague, Dept. of Computer Science
malinfr1@fel.cvut.cz

Other courses

- B4M36MBG – Molekulární biologie a genetika
  - understanding the interactions between the various systems of a cell, including the interactions between the different types of DNA, RNA and protein biosynthesis as well as learning how these interactions are regulated.

Doc. Martin Pospíšek
Charles University, Dept. of Genetics and Microbiology
Laboratory of RNA Biochemistry

Course materials

- Main page

  find a6m33bin on department's courseware page
  http://cw.felk.cvut.cz
- Course largely based on Mark Craven's bioinformatics class page at UW Wisconsin
- Contains a lot of links to useful materials in English
- Links will be also continually added to our CW
- The only Czech bioinformatics book

  Fatima Cvrčková: Úvod do praktické bioinformatiky (Academia, 2006)
  - user-oriented, for biologists/medics, not informaticians

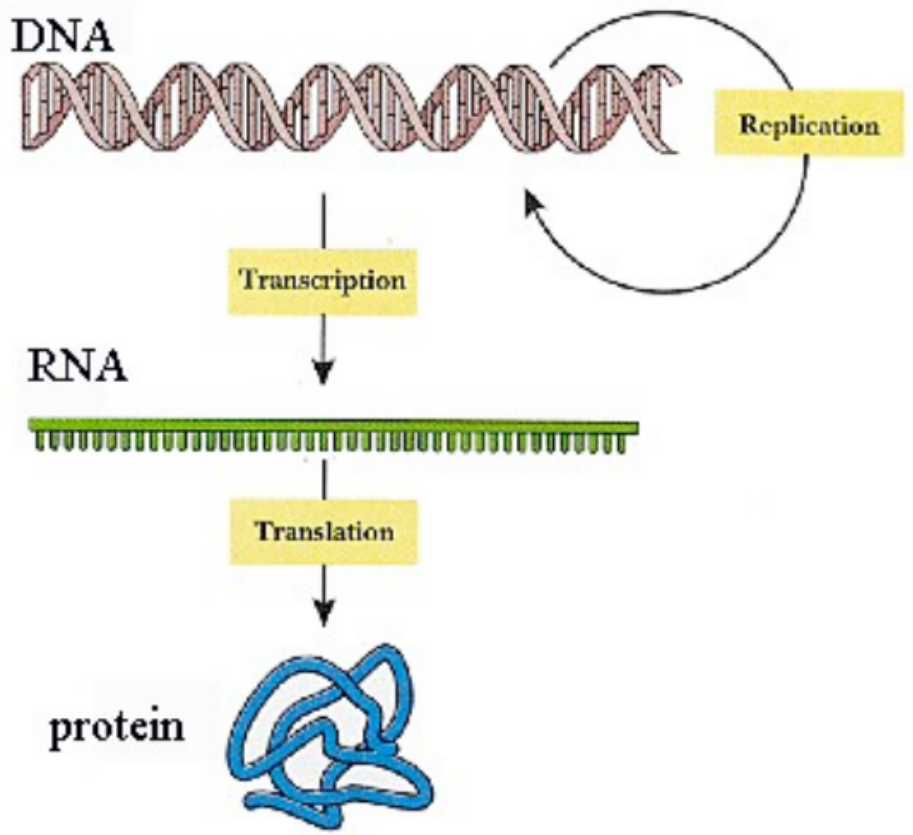
Bioinformatics

- Bioinformatics
  - representation
  - storage
  - retrieval
  - visualization
  - **analysis**

  of gene- and protein-centric biological data
- Not just bio databases!
- Also: computational biology
- Related: systems biology, structural biology

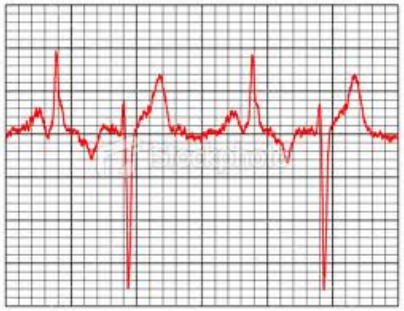
Bioinformatics: Main sources of data

- Information processes inside each cell which govern the entire organism.



Bioinformatics vs. Biomedical Informatics

- Biomedical informatics includes Bioinformatics but also other fields such as



signal analysis

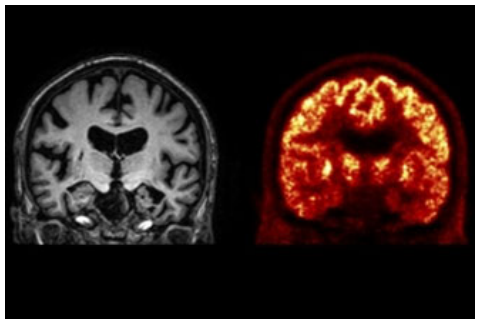
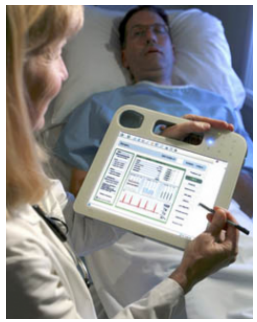


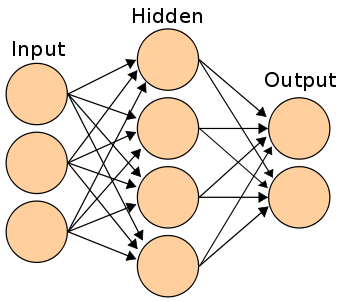
image analysis



healthcare informatics

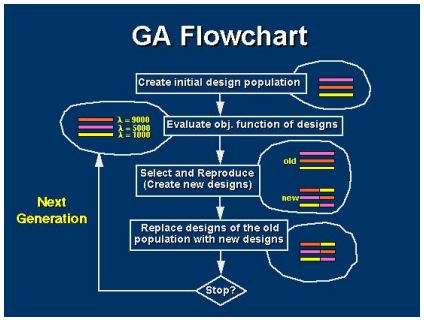
not usually associated with bioinformatics.

Bioinformatics vs. Bio-Inspired Computing

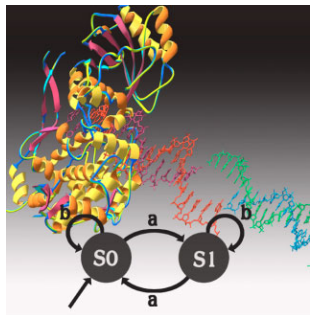


Artificial neural networks

Swarm intelligence



Genetic algorithms



DNA computing

- Also "computers + biology" but **not** bioinformatics

Bioinformatics vs. Bioinformatics

http://www.esoterika.cz/clanek/2992-mimosmyslova_spionaz_dalkove_pozorovani_i_.htm

"Podle definičního třídění ruských vědců rozlišujeme dva obory paranormálních jevů: bioinformatika a bioenergetika. Bioinformatika (tzn. mimosmyslové vnímání, ESP) zahrnuje získávání a výměnu informací mimosmyslovou cestou (nikoli normálními smyslovými orgány). V podstatě rozlišujeme následující formy bioinformace: hypnózu (kontrolu vědomí), telepatii, dálkové vnímání, prekognici, retrokognici, mimotělní zkušenost, "vidění" rukama nebo jinými částmi těla, inspiraci a zjevení."

- **not** bioinformatics

Bioinformatics: Impact

Worldwide

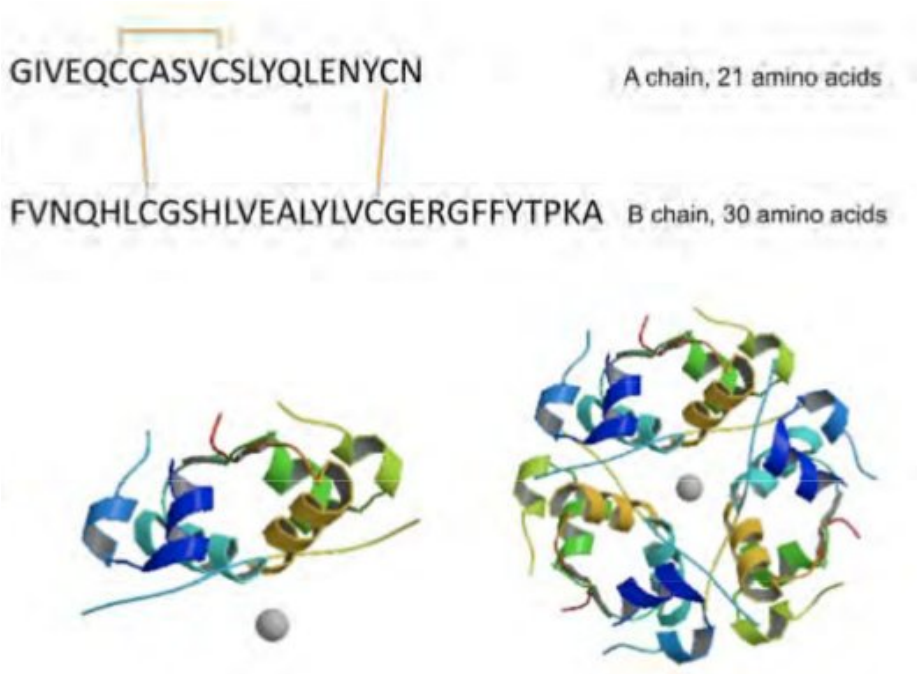
- Basic biological research
- Personalized health care
- Gene-therapy
- Drug discovery
- etc.

Czech landscape

- Small community (FEL, VSCHT, MFF, FI MU, ...)
- High demand (IKEM, IEM, IMB, UHKT, ...)
- come to see our projects

Bioinformatics: origins

- 1950's: Fred Sanger deciphers the sequence of "letters" (amino acids) in the insulin protein
- 51 letters



Bioinformatics: origins

- 2004: Human Genome (DNA) deciphered
- billions of letters (nucleic acids)

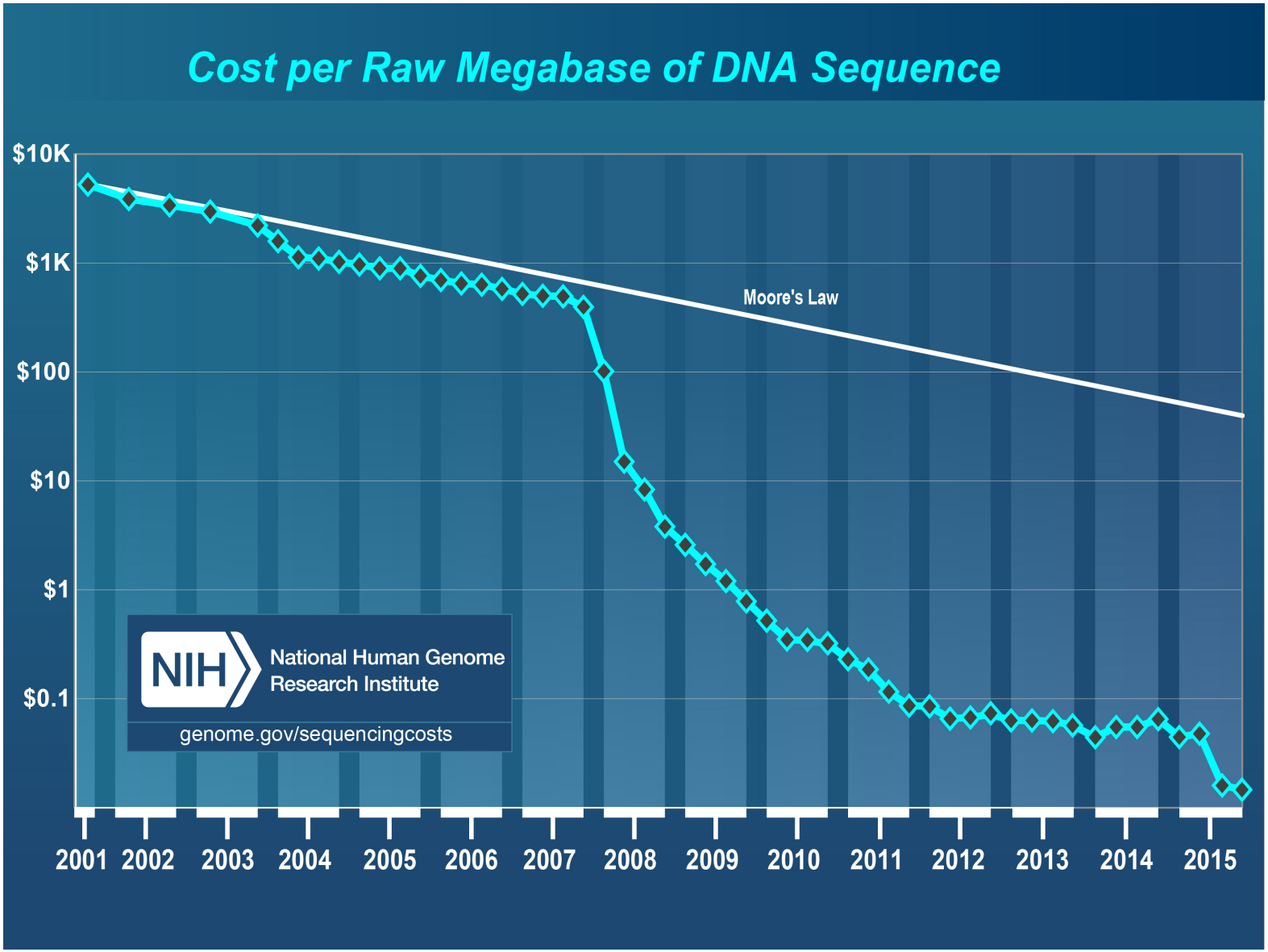
Progress in Sequencing

- Sequencing: reading the letters in the macromolecules of interest

Year	Protein	RNA	DNA	No. of residues
1935	Insulin			1
1945	Insulin			2
1947	Gramicidin S			5
1949	Insulin			9
1955	Insulin			51
1960	Ribonuclease			120
1965		$tRNA_{Ala}$		75
1967		5S RNA		120
1968			Bacteriophage λ	12
1977			Bacteriophage ϕX 174	5,375
1978			Bacteriophage ϕX 174	5,386
1981			Mitochondria	16,569
1982			Bacteriophage λ	48,502
1984			Epstein-Barr virus	172,282
2004			*Homo sapiens*	2.85 billion

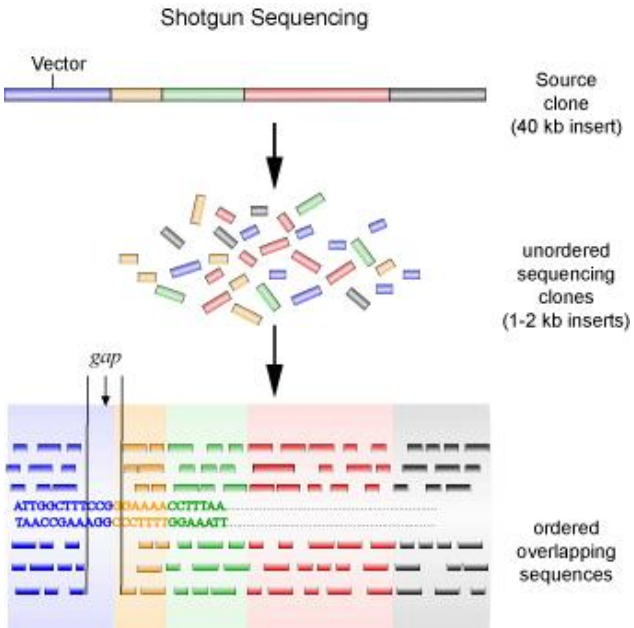
- Work continues: population sequencing (not just 1 individual), variation analysis
- Extinct species (Neandertal genome sequenced in 2010)

DNA sequencing cost



Shotgun sequencing

- DNA letters can be read only small sequences
- Shotgun approach: first shatter DNA into fragments



- Classical bioinformatics problem: assemble a genome from the read sequence fragments
- Shortest superstring problem
- Graph-theoretical formulations (Hamiltonian / Eulerian path finding)

Databases

- Read bio sequences are stored in public databases
- Main umbrella institutes

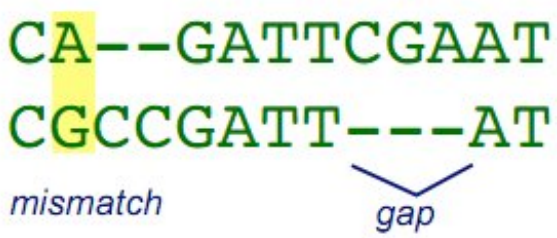
European Bioinformatics Institute (EBI)

US National Center for Biotechnology Information (NCBI)

- Protein databases: Protein Data Bank (PDB), SWISS-PROT, ...
- Gene databases: EMBL, GenBank, Entrez, ...
- Many more
- Mutually interlinked

Database Retrieval by Similarity

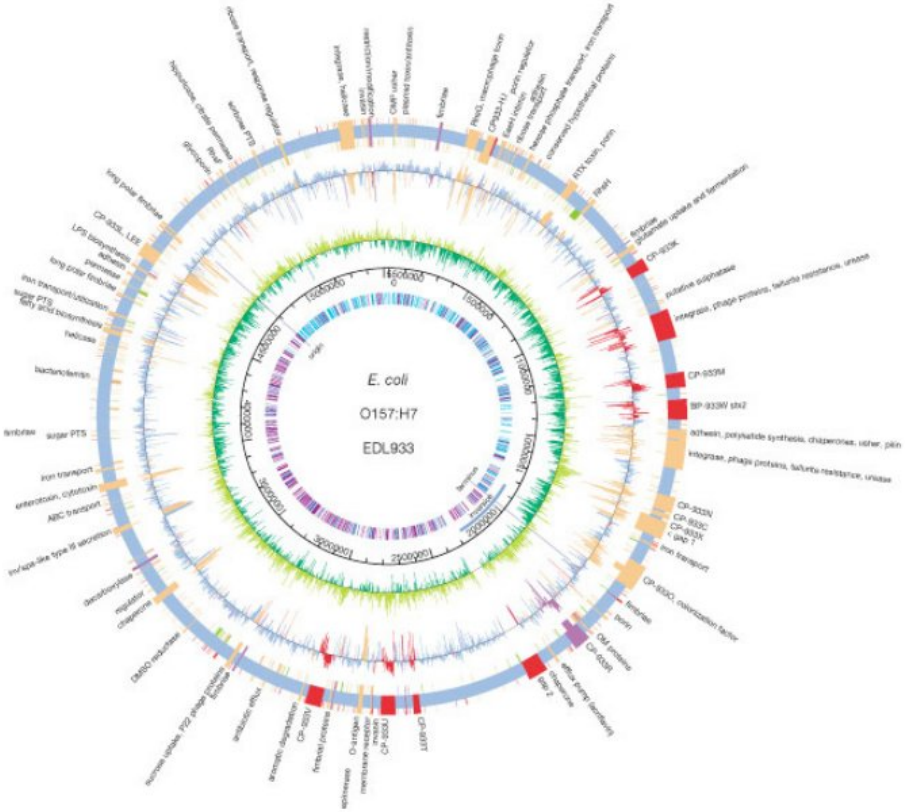
- Typical biologist's problem: retrieve sequences similar to one I have (protein, DNA fragment, ..)
- Sequence similarity may imply homology (descent from a common ancestor) and similar functions
- "Similarity" is tricky: insertions and deletions must be considered



- Bioinformatics problem: find and score the best possible *alignment*
- Dynamic programming, heuristic methods, ...

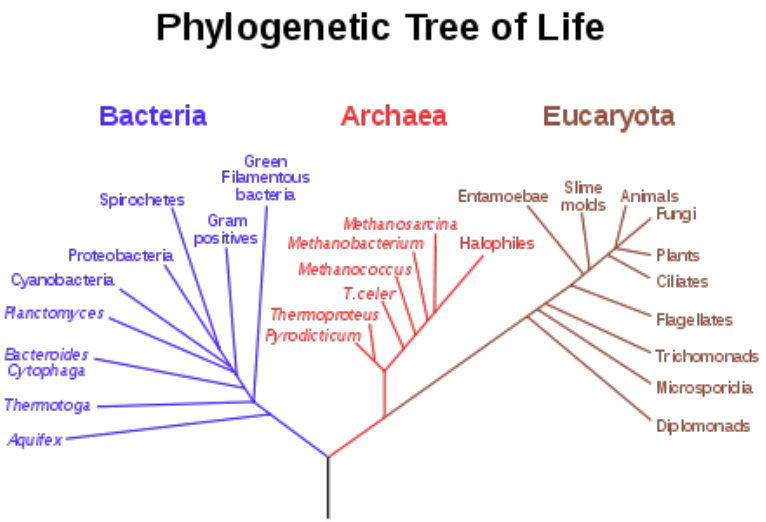
Whole Genome Similarity

- Entire genomes (not just fragments) may be aligned
- Reveal relatedness between organisms
- Further complications come into play
  - variations in repeat numbers
  - inversions
  - etc.



Inference of Phylogenetic Trees

- Given a pairwise similarity function, and a set of genomes, infer the optimal phylogenetic tree of the corresponding organisms
- Application of hierarchical clustering
- A modern approach to replace phenotype-based taxonomy



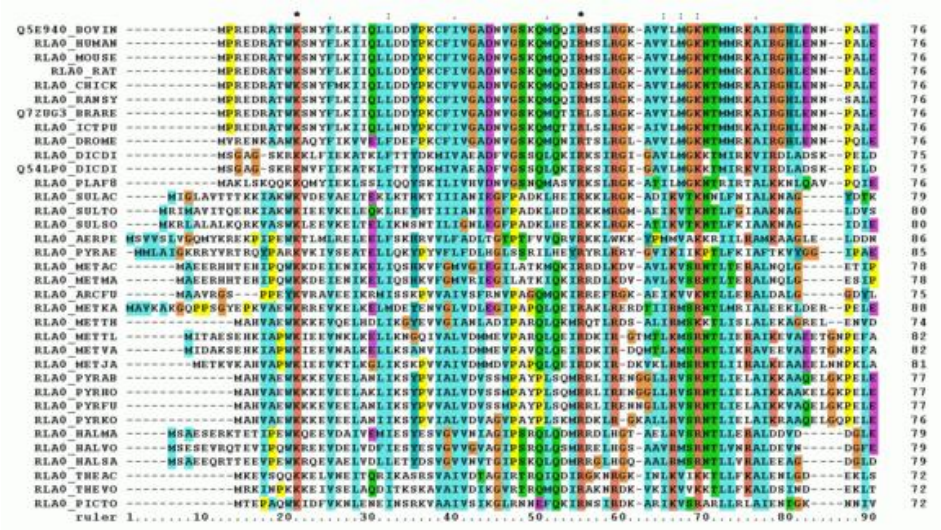

10 **Pocházíme z myši**

Společným předkem všech savců včetně lidí byl tvor podobný větší myši o váze několika set gramů, který se živil hmyzem a žil zhruba 200 tisíc let po vyhynutí dinosaurů před 65 miliony let. K tomuto závěru došla mezinárodní skupina vědců, jež využila nejnovější technologické možnosti výzkumu fosilií a DNA pomocí speciálního softwaru. Trvalo jim to šest let. ■

FOTO BRITANNICA ENCYCLOPEDIA

Multiple Sequence Alignment

- Aligning more than two sequences
- Reveal shared evolutionary origins (conserved domains)



- NP-complete problem (exp time in the number of aligned sequences)

Probabilistic Sequence Models

- specific sites (substrings) on a sequence have specific roles
- e.g. genes or promoters on DNA, active sites on proteins
- How to tell them apart?

these sequences are E. coli promoters

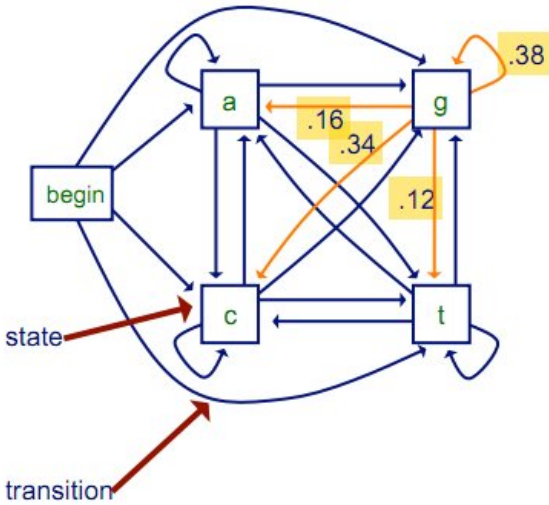
```
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca
accggaagaaaaccgtgacattttaacacgtttgttacaaggtaaaggcgacgccgc
aaattaaaatttttattgacttaggtcactaaatactttaaccaatataggcatagcg
ttgtcataatcgacttgtaaaccaaattgaaaagatttaggtttacaagtctacacc
catcctcgcaccagtcgacgacggtttacgctttacgtatagtggcgacaatttttt
tccagtataatttgttggcataattaagtacgacgagtaaaattacatacctgcccg
acagttatccactattcctgtggataaccatgtgtattagagttagaaaacacgagg
```

these sequences are not promoters

```
atagtctcagagtcttgacctactacgccagcattttggcggtgtaagctaaccatt
aactcaaggctgatacggcgagacttgcgagccttgtccttgcggtacacagcagcg
ttactgtgaacattattcgtctccgcgactacgatgagatgcctgagtgcttccgtt
tattctcaacaagattaaccgacagattcaatctcgtggatggacgttcaacattga
aacgagtcaatcagaccgctttgactctggtattactgtgaacattattcgtctccg
aagtgcttagcttcaaggtcacggatacgaccgaagcgagcctcgtcctcaatggcc
gaagaccacgcctcgccaccgagtagacccttagagagcatgtcagcctcgacaact
```

How can we tell the difference? Is this sequence a promoter?

```
ccatcaaaaaaatattctcaacataaaaaactttgtgtaatacttgtaacgctacat
```

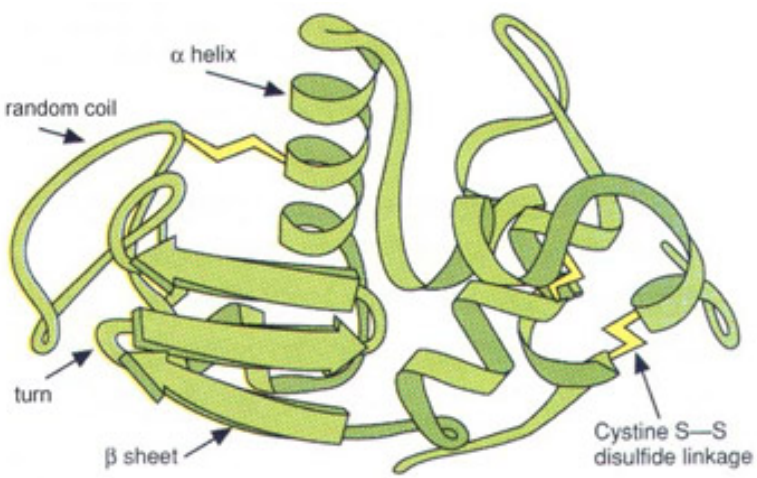


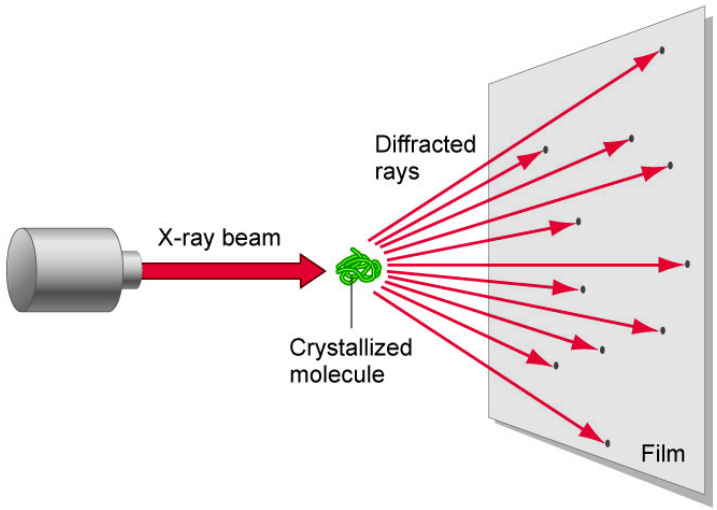
Markov Chain Model

- Each type of site has a different probabilistic model

Protein Spatial Structure

- From the DNA nucleic-acid sequence, the protein amino-acid sequence is constructed by cell machinery
- The protein folds into a complex spatial conformation

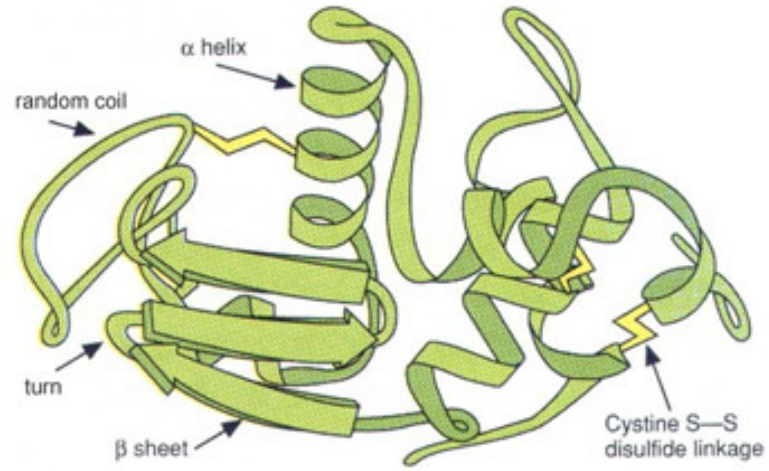




- Spatial conformation can be determined at high cost
- e.g. X-ray crystallography
- Determined structures are deposited in public protein data bases

Protein Structure Prediction

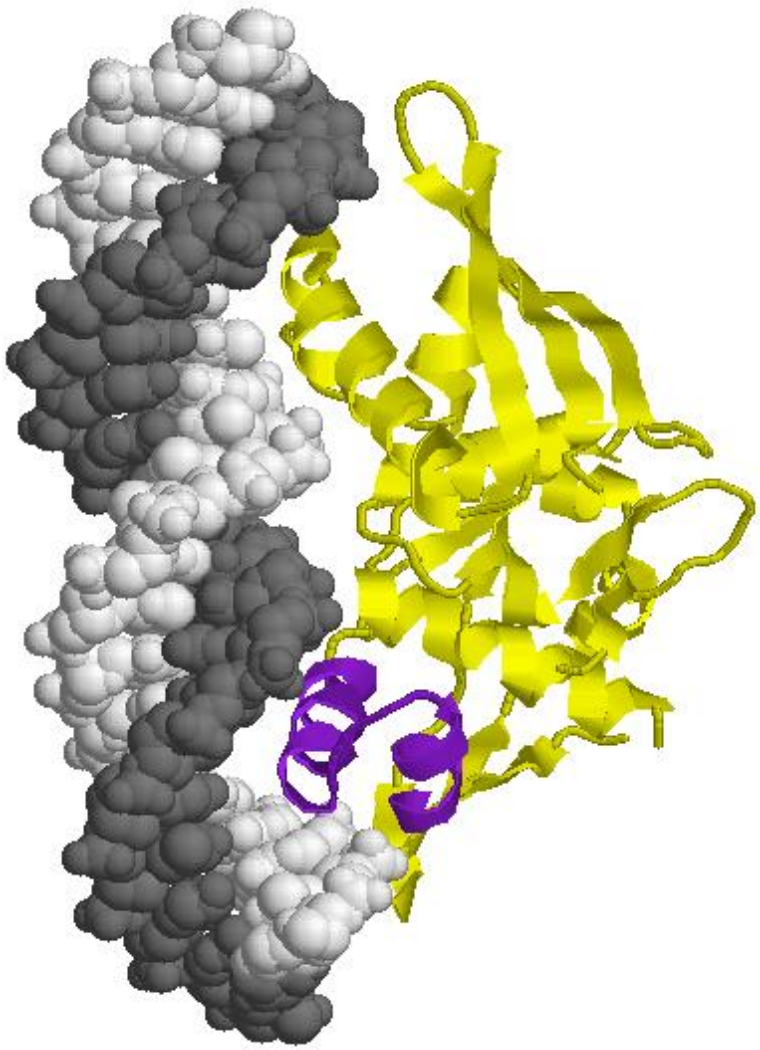
- Can we compute protein structure from sequence?
- At least distinguish α-helices from β-sheets



- Very difficult, not yet solved problem
- Approches include machine learning

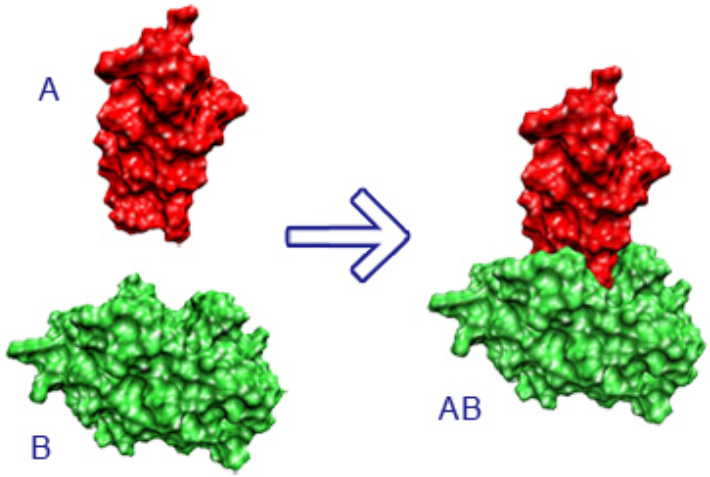
Protein Function Prediction

- Protein function is given by its geometrical conformation
- E.g., ability to bind to DNA or to other proteins
- The *active site* (shown in purple) is most important
- Important machine-learning tasks:
  - prediction of function from structure
  - detection of active sites within structure



Protein Docking Problem

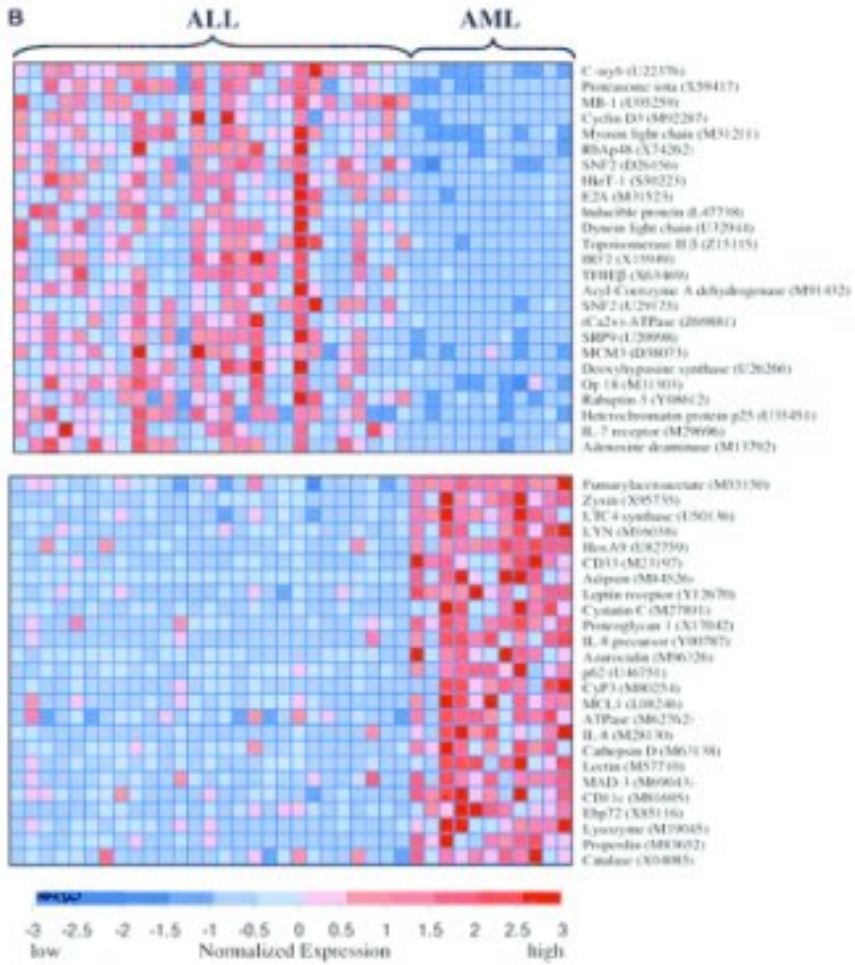
- Proteins interact by *docking*



- Will a protein dock into another protein?
- Optimization problem in a geometrical setting
- Important for novel drug discovery
  - e.g: green - receptor, red - drug
  - the trouble is, the protein may dock also in many unwanted receptors
  - immensely hard computational problems under uncertainty

Gene Expression Analysis

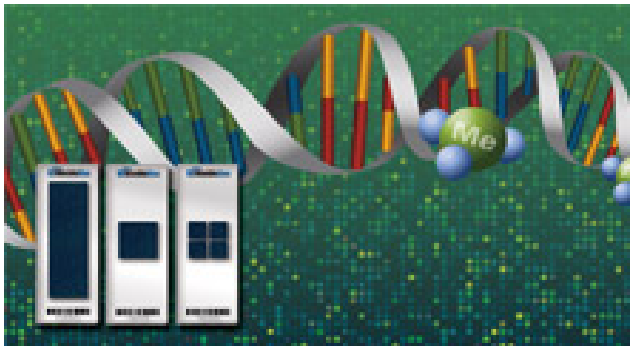
- A gene is *expressed* is the cell produces proteins according to it
- Rate of expression can be measured for thousands of genes simultaneously by *microarrays*
- Can we predict phenotype (e.g. diseases) by gene expression profiling?



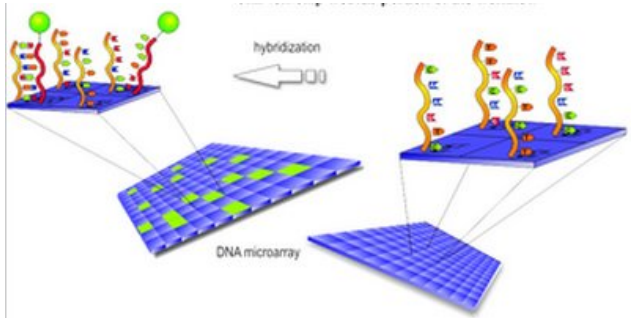
High-throughput data analysis

- Gene expression data are called *high-troughput* since lots of measurements (thousands of genes) are produced in a single experiment
- Puts biologists in a new, difficult situation: how to interpret such data?
- Example problems:
  - Too many suspects (genes), multiple hypothesis testing
  - How to spot functional patterns among so many variables?
  - How to construct multi-factorial predictive models?
- Wide opportunities for novel data analysis methods, incl. machine learning

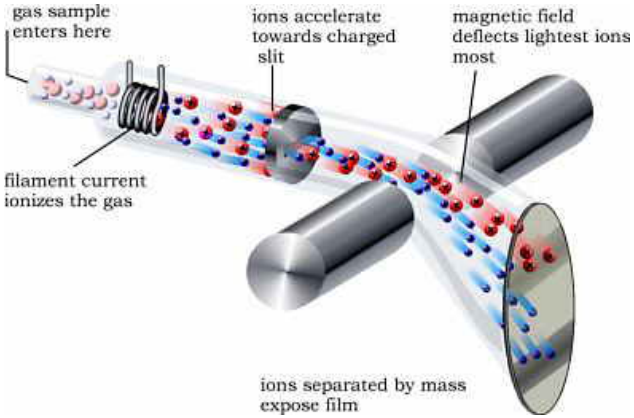
Other high-throughput technologies



Methylation arrays
(epigenetics)



Chip-on-chip
(protein X DNA interactions)

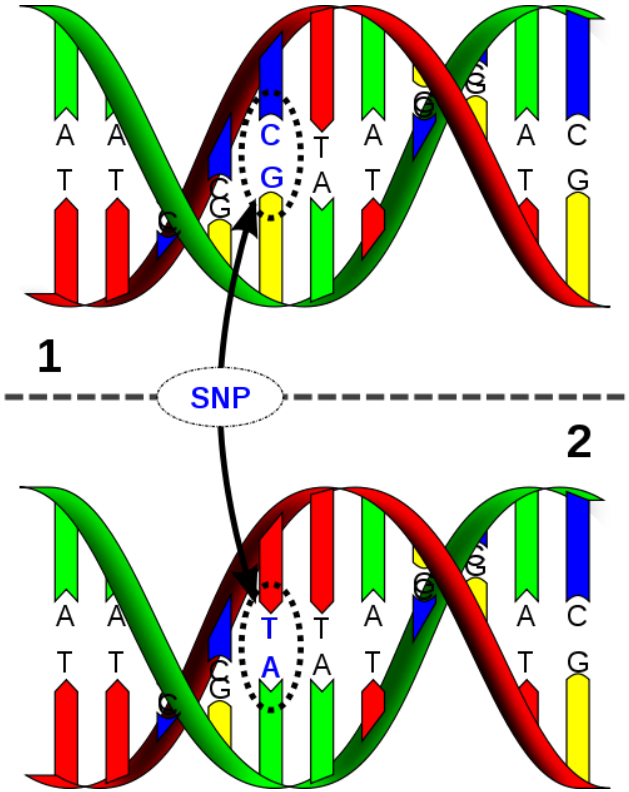


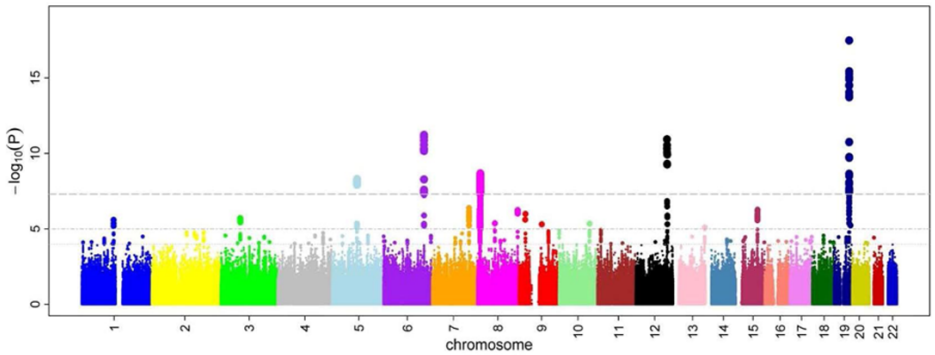
mass spectrometry
(presence of proteins)

..and more

Genome-wide association studies

- Correlates traits (e.g. susceptibility to disease) to genetic variations
- "variations": single nucleotide polymorphisms (SNP) in DNA sequence
- involves a *population* of people

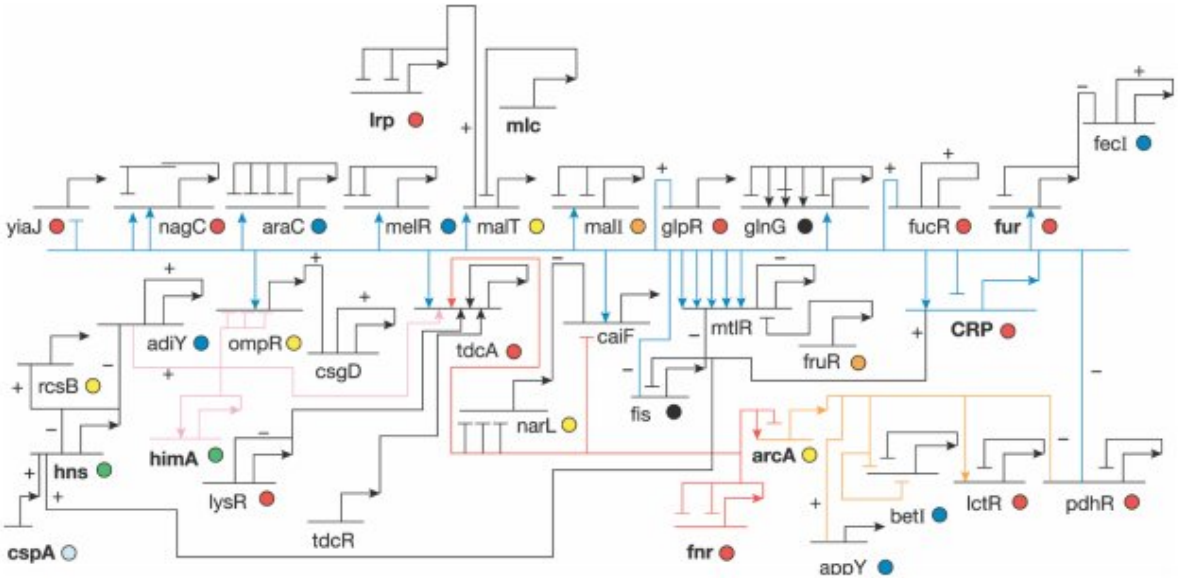




X: SNP's, Y: level of association

Gene Regulatory Networks

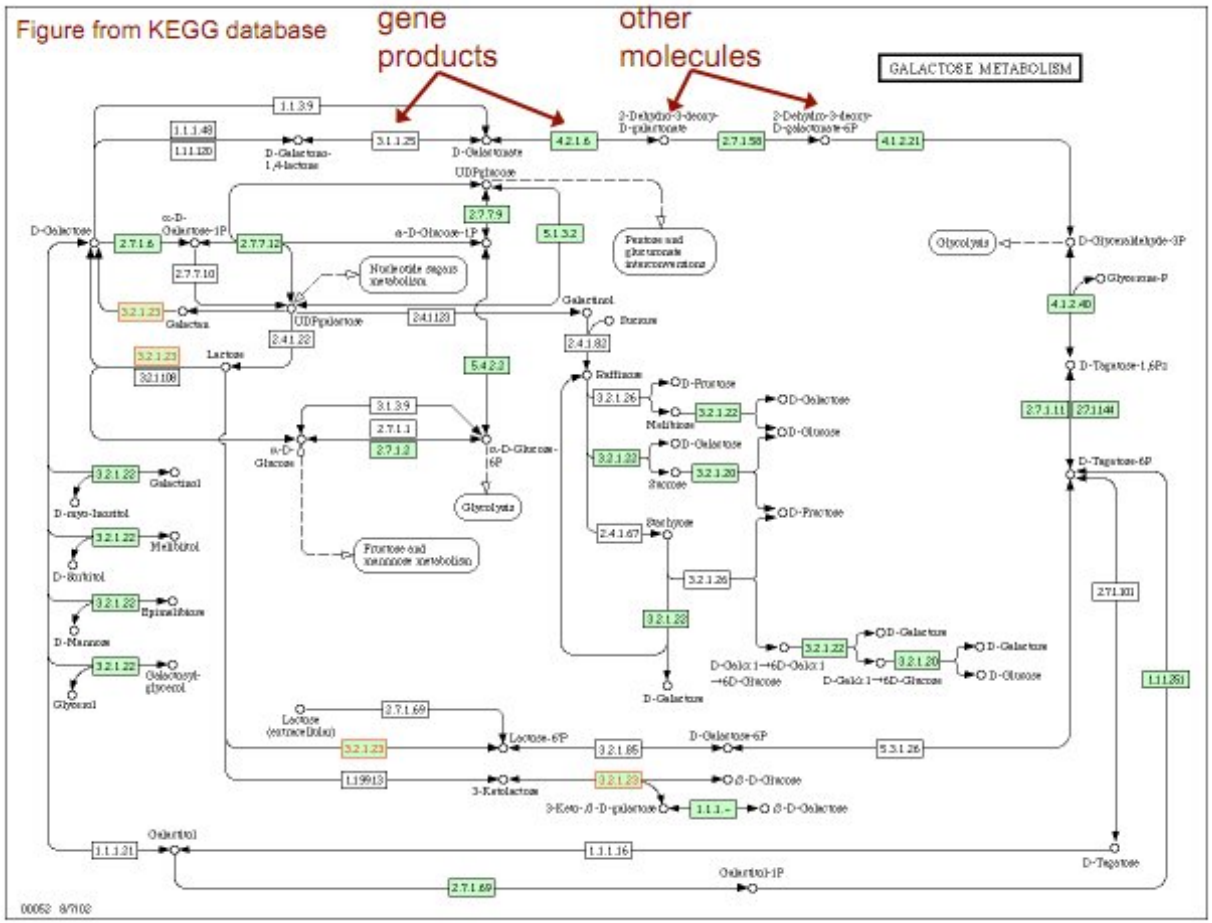
- Feedback loops in expression:
  - (a protein coded by) a gene influences the expression of another gene
  - positively (transcription factor) or negatively (inhibitor)
- Results in extremly complex networks with intricate dynamics



- Most of regulatory networks are unknown or only partially known.
- Can we *infer* such networks from time-stamped gene expression data?

Metabolic Networks

- Capture metabolism (energy processing) in cells
- Involves gene/proteins but also other molecules
- Computational problems similar as in gene regulation networks



Exploiting Background Knowledge

- The bioinformatics tasks exemplified so far followed the pattern

$$\text{Data} \rightarrow \text{Genomic knowledge}$$

- A lot of relevant formal (computer-understandable) knowledge available so the equation should be

$$\text{Data} + \text{Current Genomic Knowledge} \rightarrow \text{New Genomic Knowledge}$$

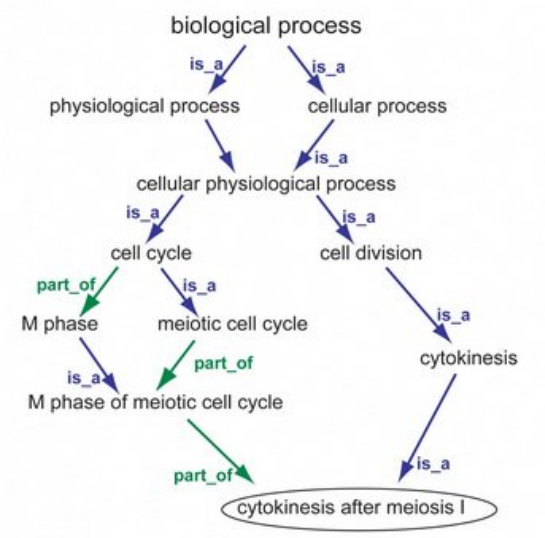
for example:

$$\text{Gene expression data} + \text{Known functions of genes} \rightarrow \text{Phenotype linked to a gene function}$$

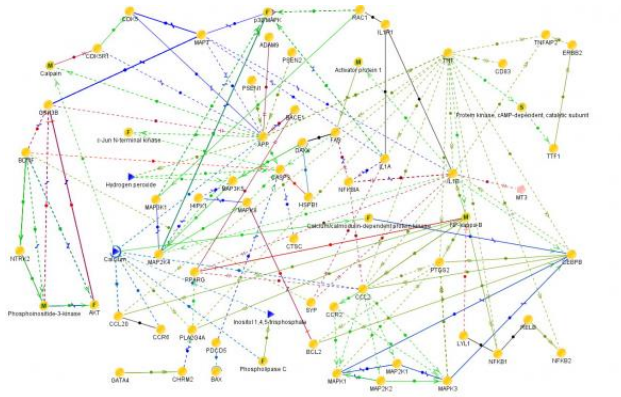
- But how to represent backround knowledge and use it systematically in data analysis?
- Important bioinformatics problem

Examples of Genomic Background Knowledge

scientific abstracts



gene ontology



interaction networks

- and many other kinds

Bioinformatics: impact in scientific literature

Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades [Wren, Bioinformatics '16]

Table 1. Most cited non-review articles from the approximate start of the Internet Age (~1994) to 2013 according to the Institute for Scientific Information (ISI) Web of Knowledge

Most highly cited paper	Year published	Citations	# bioinf in Top 20	Avg bioinf JIF	Avg non-bioinf JIF
MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0	**2013**	**4531**	5	9.3	26.5
Observation of a new particle in the search for the Higgs boson	2012	3163	5	14.8	28.4
MEGA5: Molecular Evolutionary Genetics Analysis	**2011**	**19 098**	5	18.6	35.5
Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008	2010	5676	10	8.2	24.1
Systematic and integrative analysis of large gene lists using DAVID	**2009**	**6242**	7	7.5	23.3
A short history of SHELX	**2008**	**47 516**	8	10.2	29.5
MEGA4: Molecular evolutionary genetics analysis	**2007**	**20 470**	8	6.9	33.6
Induction of pluripotent stem cells from mouse embryonic cultures	2006	8503	5	10.8	23.9
Two-dimensional gas of massless Dirac fermions in graphene	2005	9091	5	5.8	25.5
Electric field effect in atomically thin carbon films	2004	20 395	11	5.4	30.5
MrBayes 3: Bayesian phylogenetic inference under mixed models	**2003**	**14 638**	11	8.6	21.1
The Cambridge Structural Database	**2002**	**8982**	6	4.1	26.4
Analysis of relative gene expression data using real-time quantitative PCR	2001	38 893	7	6.9	32.3
The Protein Data Bank	**2000**	**14 420**	4	6.8	23.1
From ultrasoft pseudopotentials to the projector augmented-wave method	1999	18 566	5	11.2	16.6
Crystallography & NMR system: A new software suite	**1998**	**15 269**	5	6.3	24.1
Gapped BLAST and PSI-BLAST	**1997**	**40 205**	10	5.8	32.8
Generalized gradient approximation made simple	1996	47 033	7	3.2	16.8
Controlling the false discovery rate	1995	21 224	7	3.2	27.1
CLUSTAL-W - improving sensitivity of multiple sequence alignment	**1994**	**42 995**	5	7.1	19.1

Citation data was compiled March 21, 2016 and data for all papers analyzed can be found in Supplementary Tables S1 and S2. Bioinformatics papers are **bolded**, and general methods papers frequently used in bioinformatics programs are *italicized*. Shown are the titles of the most cited papers each year (sometimes shortened to fit), the number of citations accrued at the time of this study (dataset citations from ISI's Data Citation Index not included), the number of bioinformatics (including methods) papers in the top 20 for each year, and the average JIF for the bioinformatics papers and non-bioinformatics papers for each year.

IDA methods in journal papers

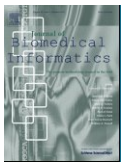
BioMed Central

BMC Genomics

BMC Bioinformatics

IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

In Silico Biology
An International Journal on Computational Molecular Biology



Semantic biclustering for finding local, interpretable and predictive expression patterns

Jiří Kléma*, František Malinka and Filip Železný

Network-constrained forest for regularized classification of omics data

Michael Anděl[a], Jiří Kléma[a], Zdeněk Krejčík[b]

Comparative Evaluation of Set-Level Techniques in Predictive Classification of Gene Expression Samples

Matěj Holec[1], Jiří Kléma*[1], Filip Železný[1], Jakub Tolar[2]

Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification

Miloš Krejník and Jiří Kléma

Learning Relational Descriptions of Differentially Expressed Gene Groups

Igor Trajkovski, Filip Železný, Nada Lavrač, and Jakub Tolar

Constraint-based knowledge discovery from SAGE data

Jiří Kléma1[1,3], Sylvain Blachon[2], Arnaud Soulet[4], Bruno Crémilleux[1] and Olivier Gandrillon[2]*

Induction of comprehensible models for gene expression datasets by subgroup discovery methodology

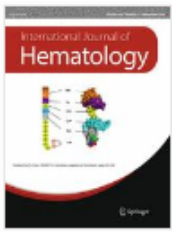
Dragan Gamberger[a,*], Nada Lavrač[b,c], Filip Železný[d,e], Jakub Tolar[f]

IDA applications in medical studies

PLOS ONE | TENTH ANNIVERSARY

Replication of SNP associations with keratoconus in a Czech cohort

Petra Liskova, Lubica Dudakova, Anna Krepelova, Jiri Klema, Pirro G. Hysi



Up-regulation of ribosomal genes is associated with a poor response to azacitidine in myelodysplasia and related neoplasms

M. Monika Belickova, Michaela Dostalova Merkerova, Hana Votavova, Jan Valka, Jitka Vesela, Barbora Pejsova, Hana Hajkova, Jiri Klema, Jaroslav Cermak, Anna Jonasova

Differential Regulation of the Nuclear Factor-κB Pathway by Rabbit Antithymocyte Globulins in Kidney Transplantation

Mariana Urbanova,[1,2] Irena Brabcova,[2] Eva Girmanova,[2] Filip Zelezny,[3] and Ondrej Viklicky[1,2,4]

pft PARTICLE AND FIBRE TOXICOLOGY

RESEARCH Open Access

Global gene expression changes in human embryonic lung fibroblasts induced by organic extracts from respirable air particles

Helena Líbalová[1,2], Kateřina Uhlířová[1], Jiří Kléma[3], Miroslav Machala[4], Radim J Šrám[1], Miroslav Ciganek[4] and Jan Topinka[1*]

Bioinformatics at the IDA lab

If you find this course interesting, you can take part in IDA's research!