Bioinformatics: course introduction

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

IDA
Intelligent Data Analysis
RESEARCH GROUP

http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

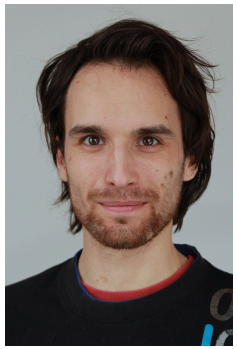
BAM36BIN – Medical Electronics and Bioinformatics
B4M36BIN – Open Informatics, Bioinformatics

- 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
Intelligent data analysis group
klema@fel.cvut.cz

Ing. Petr Ryšavý
CTU Prague, Dept. of Computer Science
Intelligent data analysis group
rysavpe1@fel.cvut.cz



Ing. Jáchym Barvínek
CTU Prague, Dept. of Computer Science
Intelligent data analysis group
barvijac@fel.cvut.cz

Other courses

- B4M36MBG – Molecular biology and genetics
  - 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.

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

Course materials

- Main page: find BIN on department's courseware
  - http://cw.felk.cvut.cz/wiki/courses/bin
- Course largely based on Mark Craven's class at University of Wisconsin
- Contains a lot of links to useful materials in English
- The main books
  - Durbin et al.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge University Press, 1998),
  - Jones, Pevzner: An Introduction to Bioinformatics Algorithms (The MIT Press, 2004),
  - more recent papers linked pn the course webpage.
- The only Czech bioinformatics book
  - Fatima Cvrčková: Úvod do praktické bioinformatiky (Academia, 2006)
  - user-oriented, for biologists/medics, not informaticians.

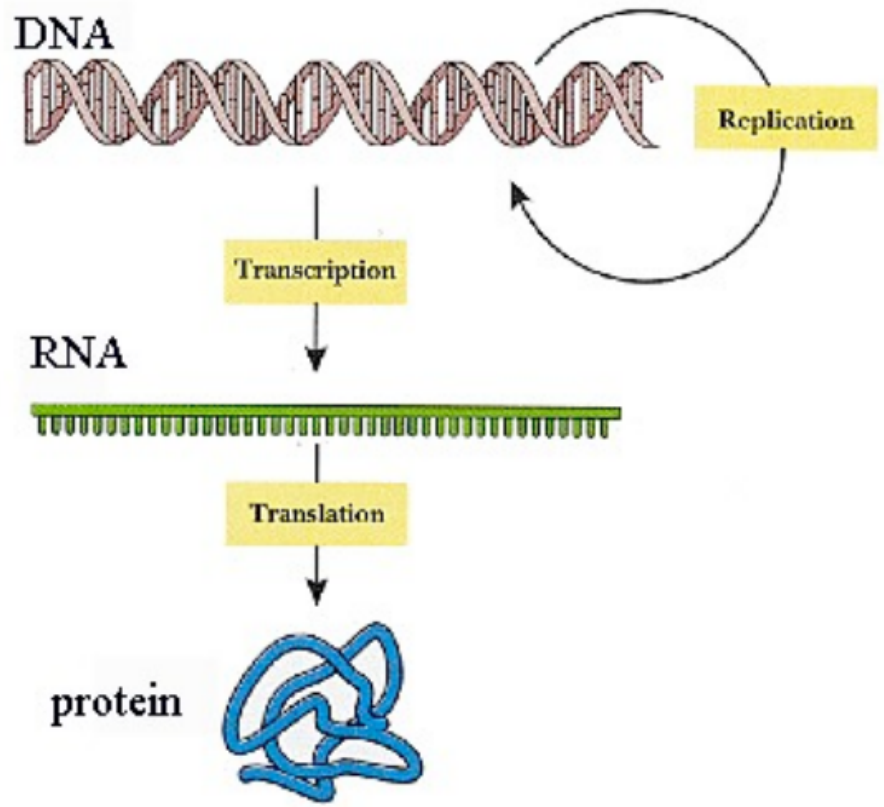
Bioinformatics

- 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

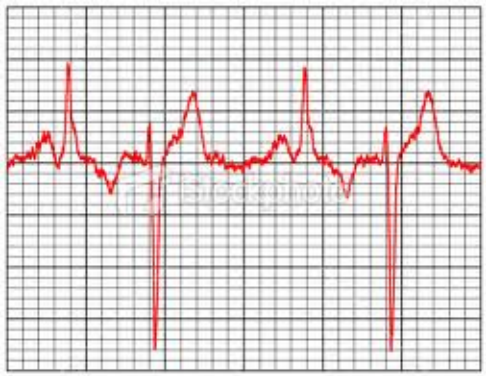
Bioinformatics: Main sources of data

- 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

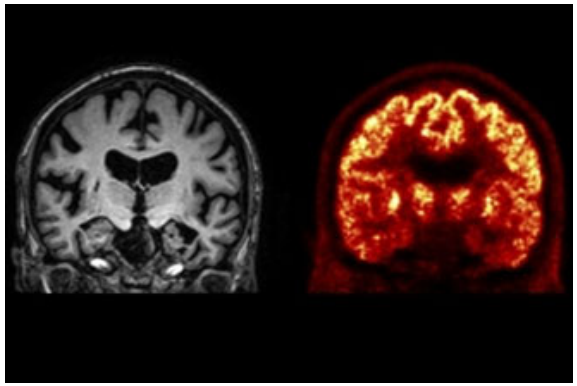
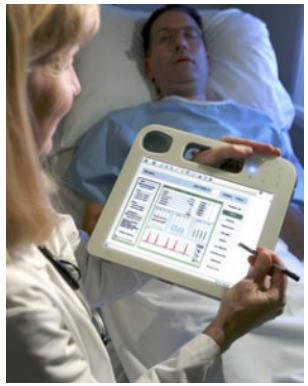


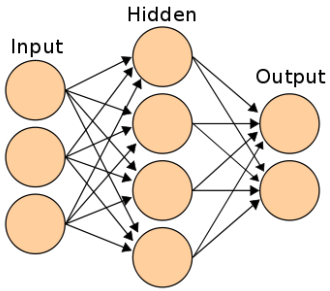
image analysis



healthcare informatics

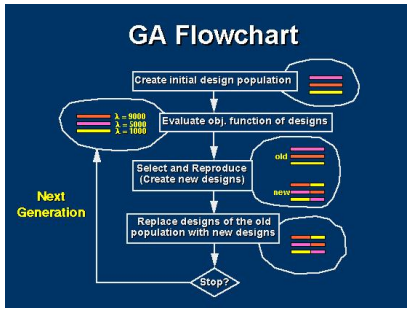
not usually associated with bioinformatics.

Bioinformatics vs. Bio-Inspired Computing

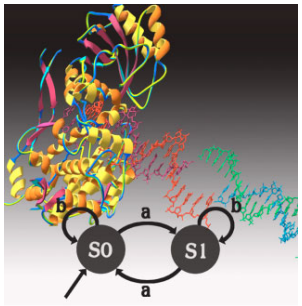


Artificial neural networks

Swarm intelligence



Genetic algorithms



DNA computing

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

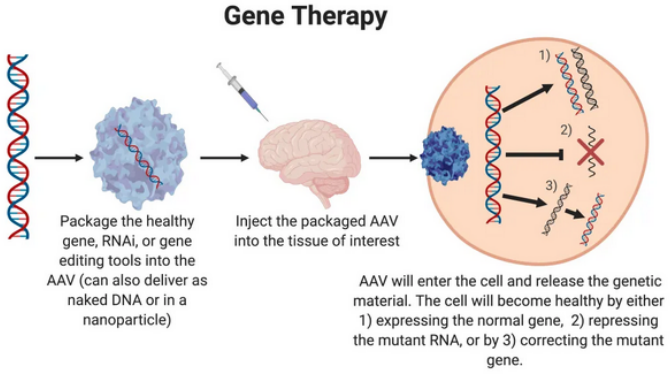
Bioinformatics: Impact

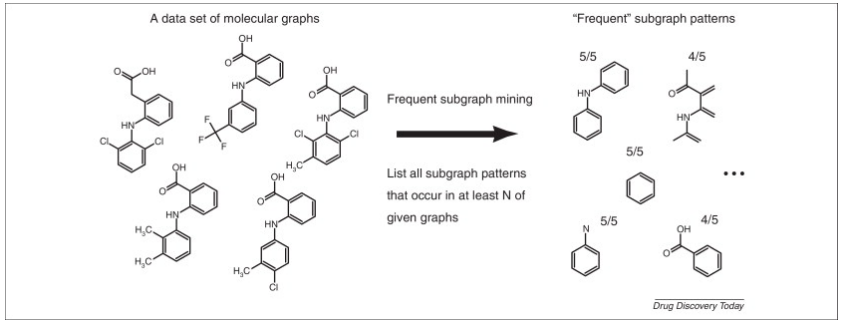
Worldwide

- understanding of raw genomic data,
- facilitates basic biological research,
- improves personalized health care,
- helps in gene-therapy,
- reduces the price of drug discovery,

Czech landscape

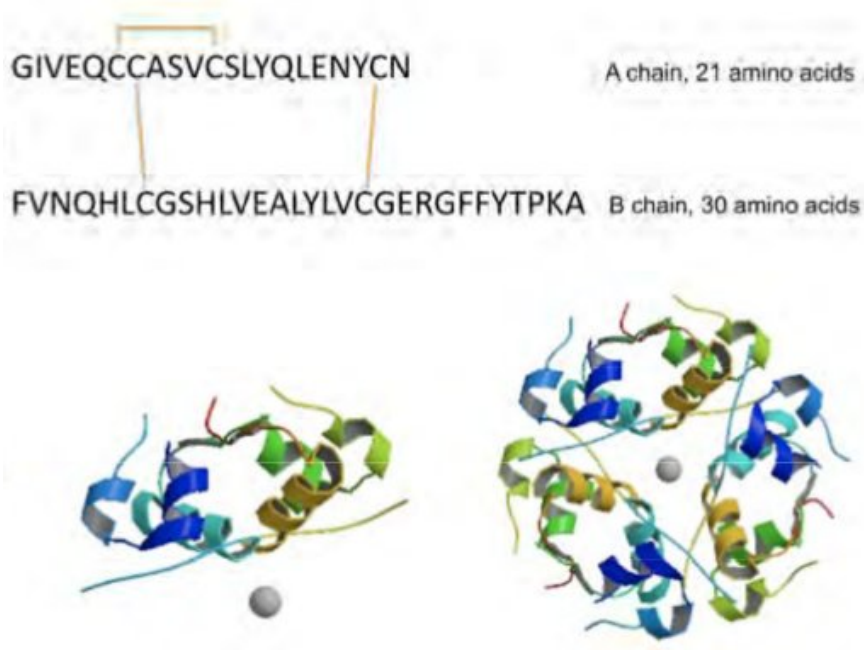
- Small community
  (FEL, VSCHT, MFF, FI MU, . . .)
- High demand
  (IKEM, IEM, IMB, UHKT, . . .)
- come to see our IDA projects
  (some of them mentioned later).





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)

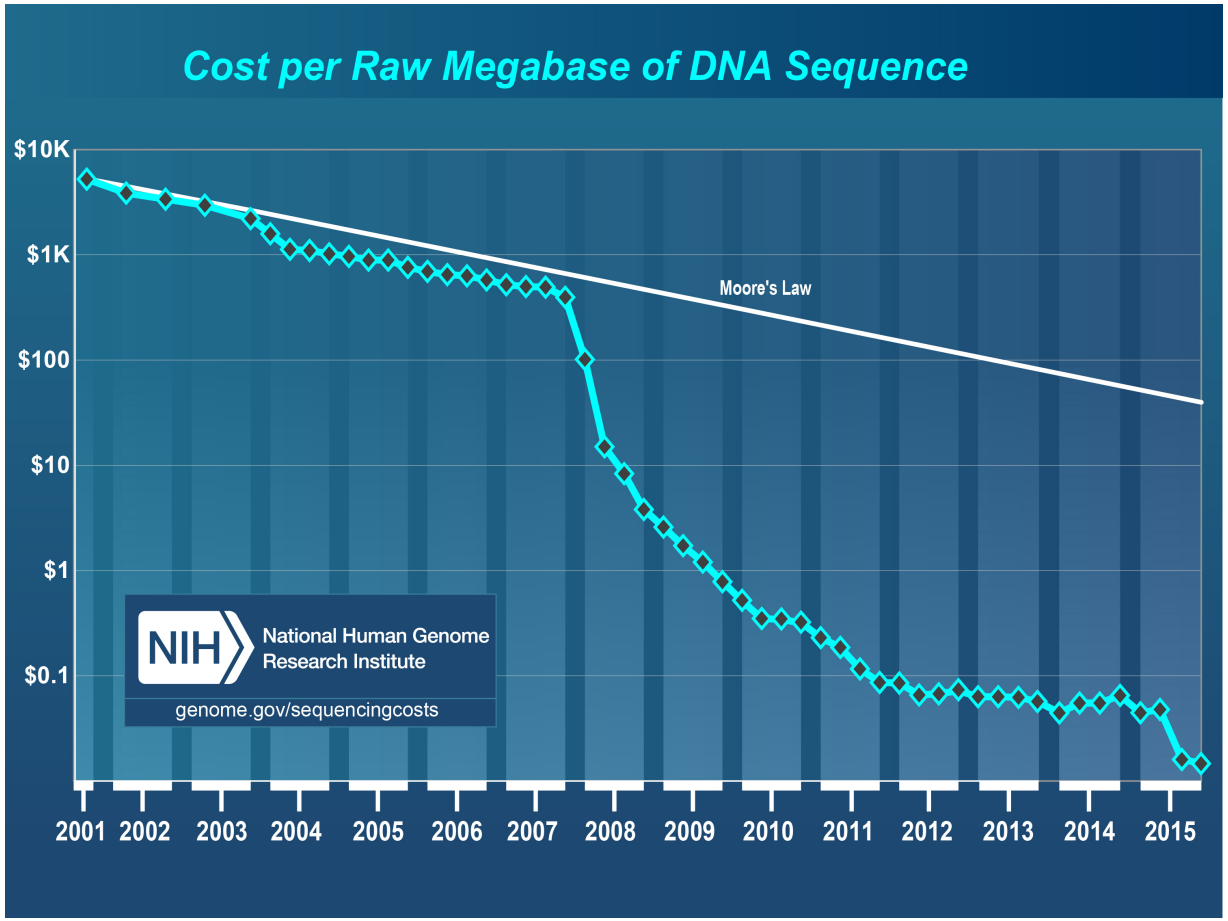
Progress in sequencing

- Sequencing: reading the letters in the macromolecules of interest

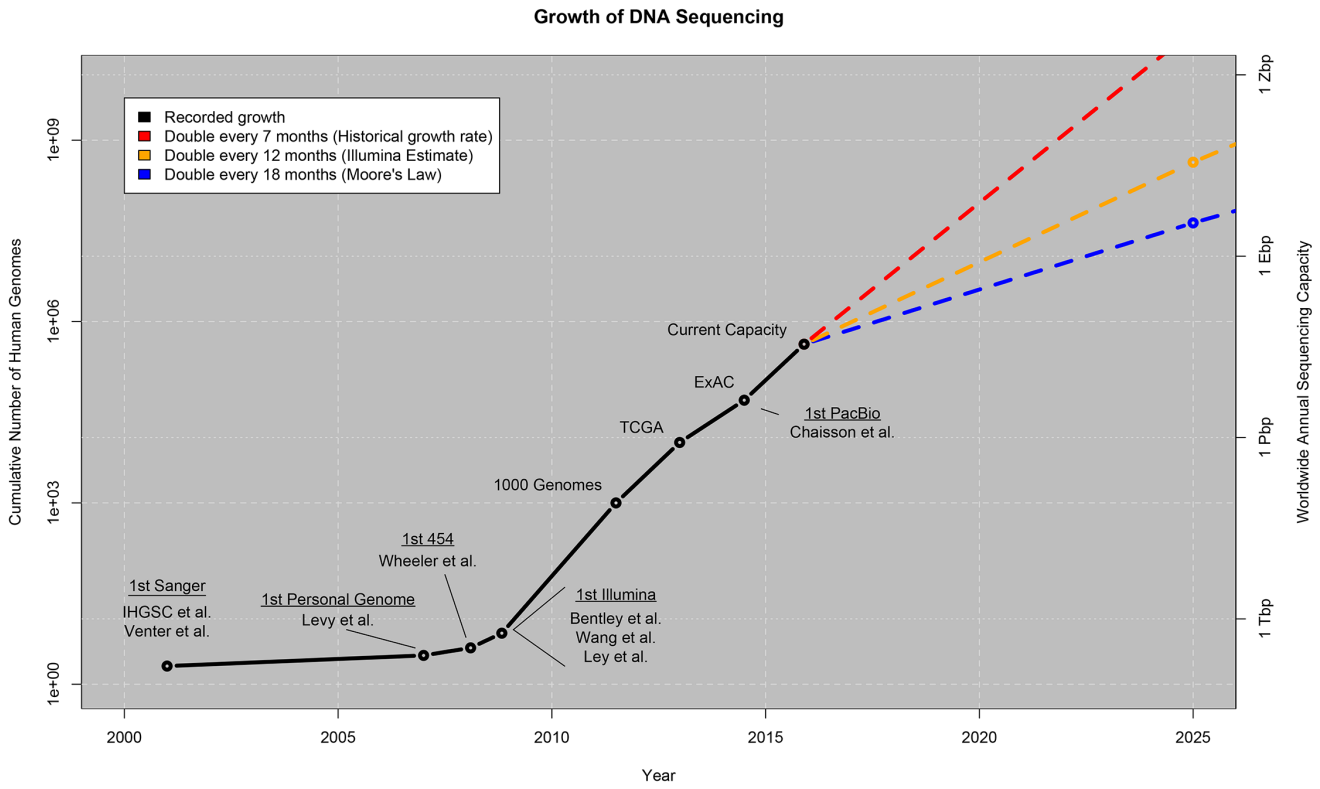
Year	Protein	RNA	DNA	No. of residues
1935	Insulin			1
1945	Insulin			2
1947	Gramicidin S			5
1949	Insulin			9
1955	Insulin			51
1960	Ribonuclease			120
1965		$tRNA_{Ala}$		75
1967		5S RNA		120
1968			Bacteriophage λ	12
1977			Bacteriophage ϕX 174	5,375
1978			Bacteriophage ϕX 174	5,386
1981			Mitochondria	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),
  - ...
  - sequencing based surveillance helps fight COVID-19 (2021).

DNA sequencing cost



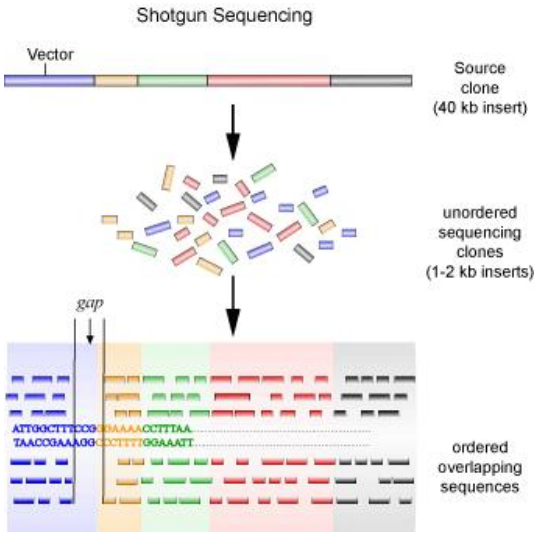
DNA, one of the biggest contemporary data sources



- by 2025, genomics will be the largest digital data source with 10^{21} bases/year,
- see Big Data: Astronomical or Genomical?, Stephens 2015,
- it will overcome other big data generators: astronomy, YouTube, Twitter.

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

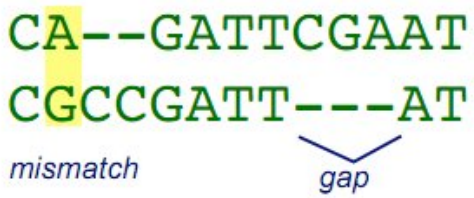
European Bioinformatics Institute (EBI)

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

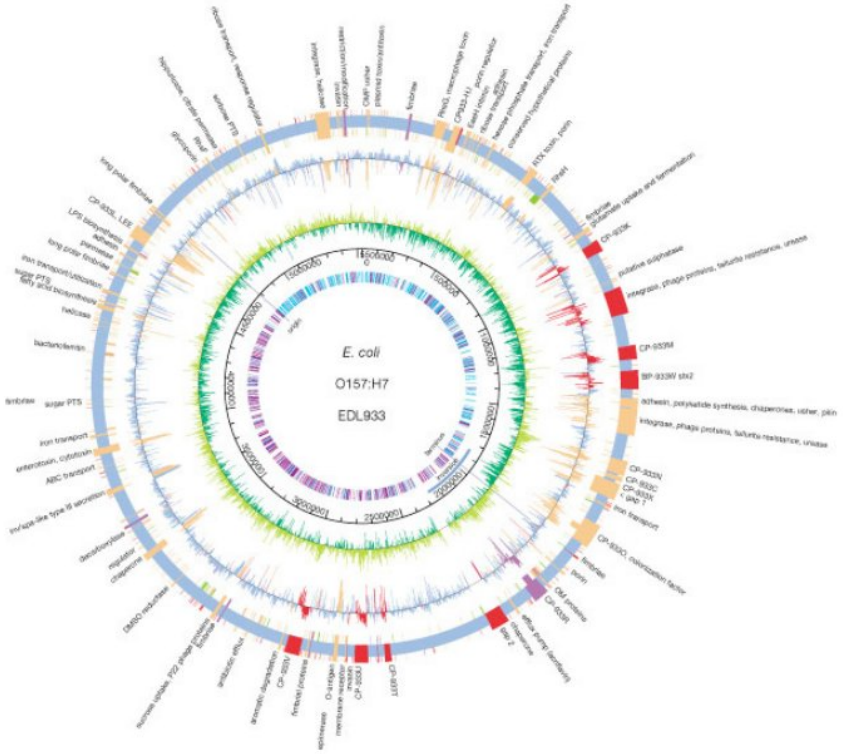
- Typical biologist's problem: retrieve sequences similar to one I have (protein, DNA fragment, ..)
- Sequence similarity may imply homology (descent from a common ancestor) and similar functions
- "Similarity" is tricky: insertions and deletions must be considered



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

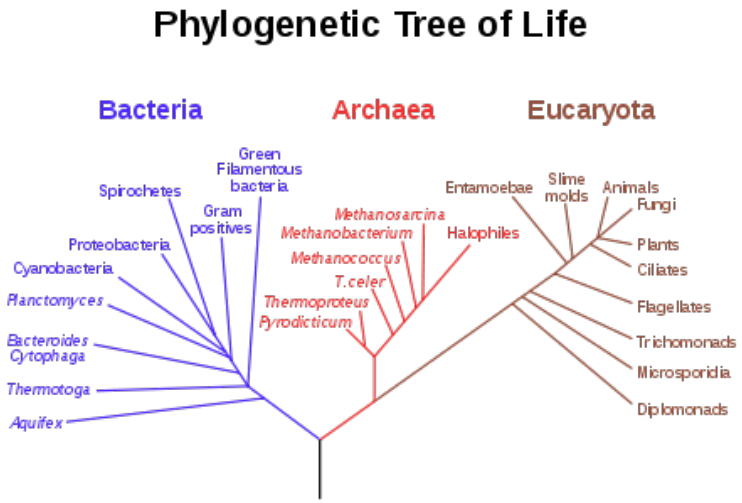
Whole genome similarity

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



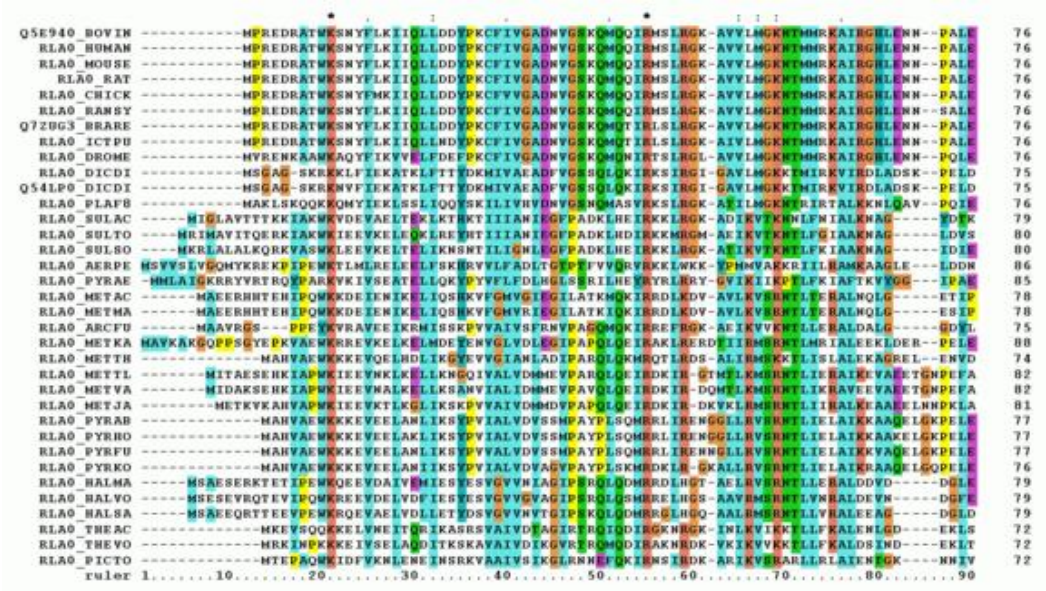
Inference of phylogenetic trees

- 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.



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

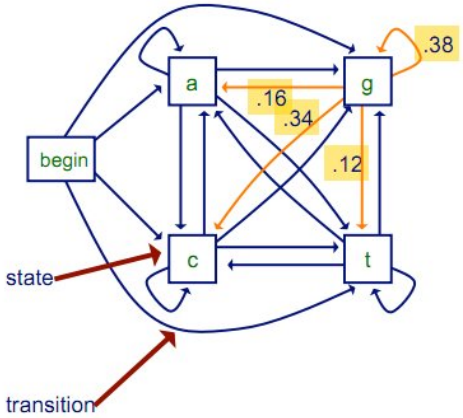
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca
accggaagaaaaccgtgacattttaacacgtttgttacaaggtaaaggcgacgccgc
aaattaaaattttattgacttaggtcactaaatactttaaccaatataggcatagcg
ttgtcataatcgacttgtaaaccaaattgaaaagatttaggtttacaagtctacacc
catcctcgcaccagtcgacgacggtttacgctttacgtatagtggcgacaatttttt
tccagtataatttgttggcataattaagtacgacgagtaaaattacatacctgcccg
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

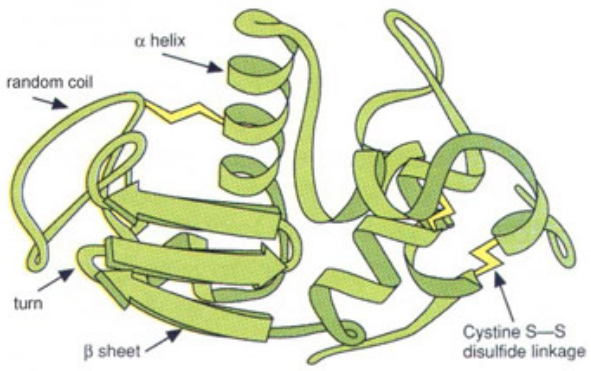


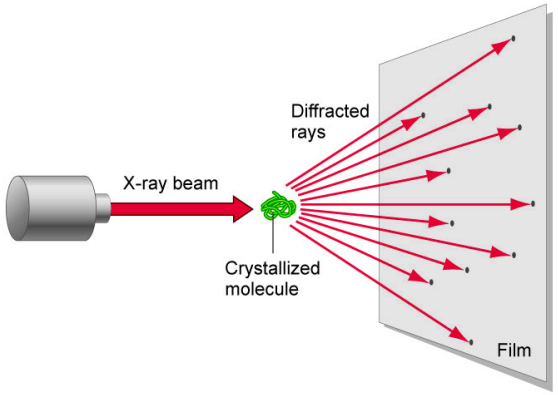
Markov Chain Model

- 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

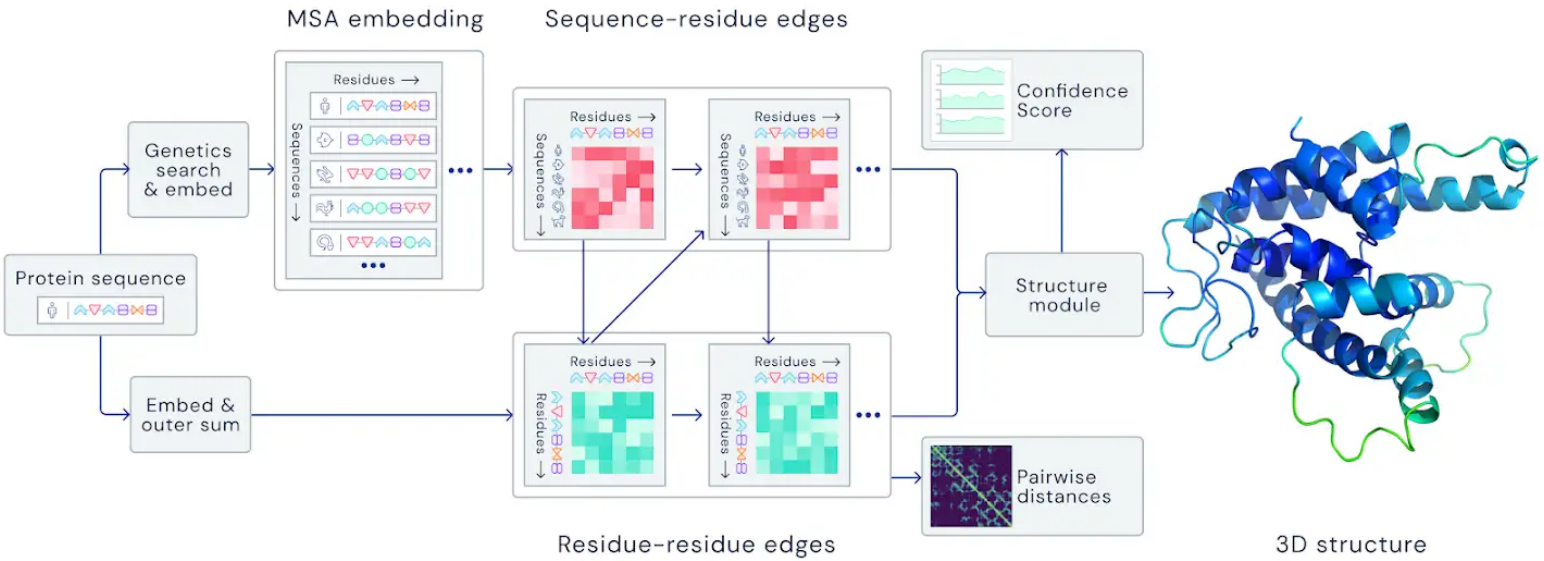




- Spatial conformation can be determined at high cost
- e.g. X-ray crystallography
- Determined structures are deposited in public protein data bases

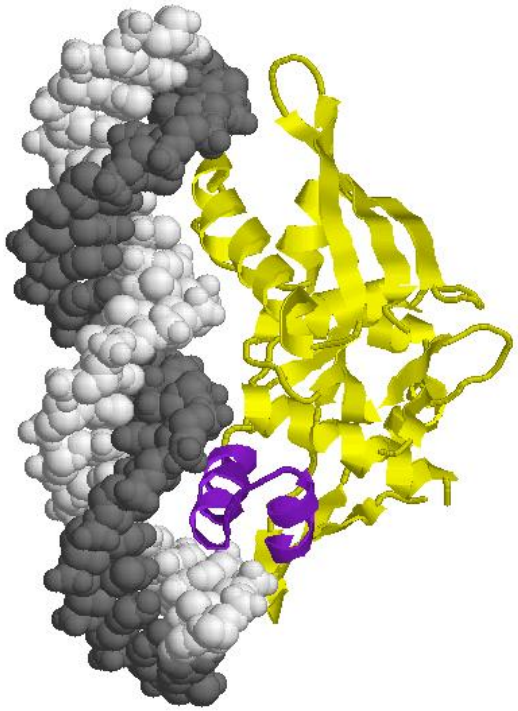
Protein structure prediction

- Can we compute protein structure from sequence?
- A couple of years ago, at least distinguish α-helices from β-sheets
- In 2020, a breakthrough reached by AI and machine learning
  - AlphaFold2 (Deepmind) has a level of accuracy comparable to that achieved with expensive and time-consuming lab experiments.



Protein function prediction

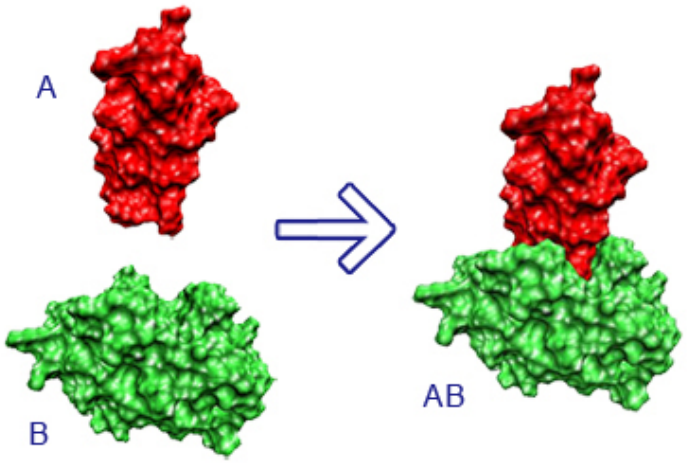
- Protein function
  - is given by protein's geometrical conformation,
- E.g., ability to bind to DNA or to other proteins
- The *active site* is most important
- Important machine-learning tasks:
  - prediction of function from structure,
  - detection of active sites within structure.



purple - *active site*

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.

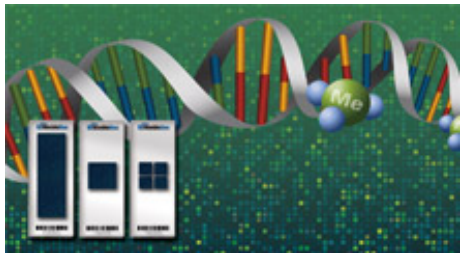
Gene Expression Analysis

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

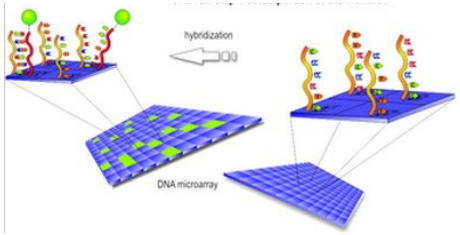
High-throughput data analysis

- 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

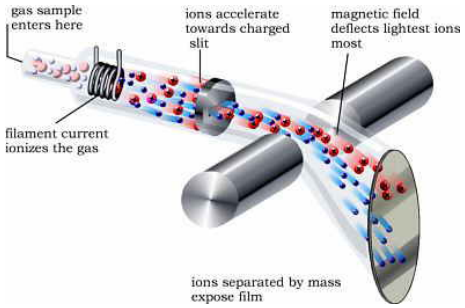
Other high-throughput technologies



Methylation arrays
(epigenetics)



Chip-on-chip
(protein X DNA interactions)

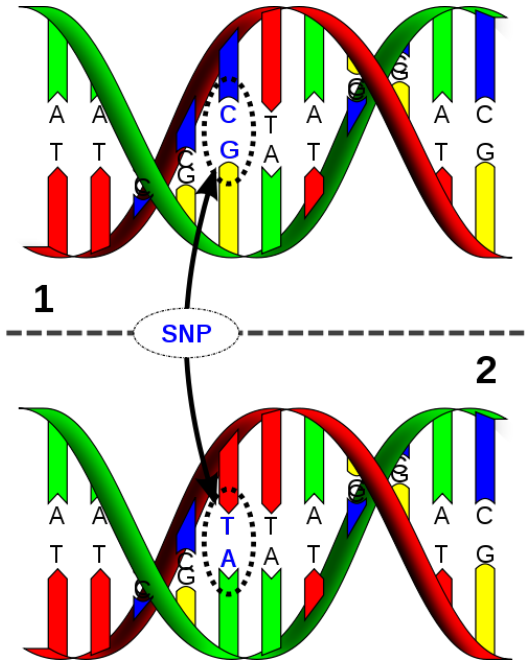


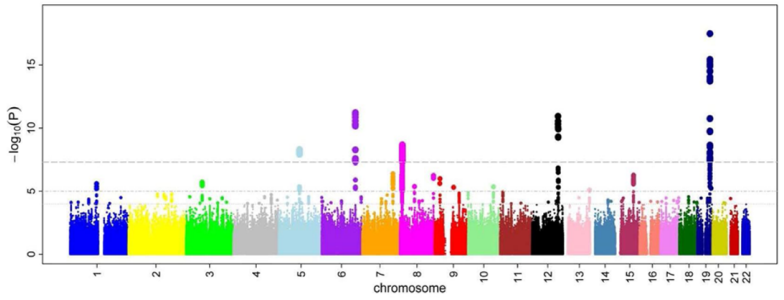
mass spectrometry
(presence of proteins)

... 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

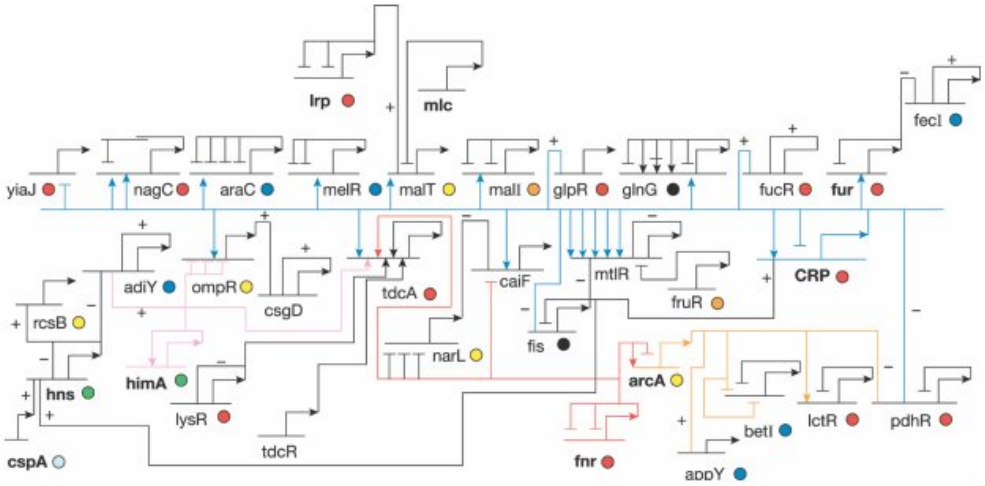




X: SNP's, Y: level of association

Gene regulatory networks

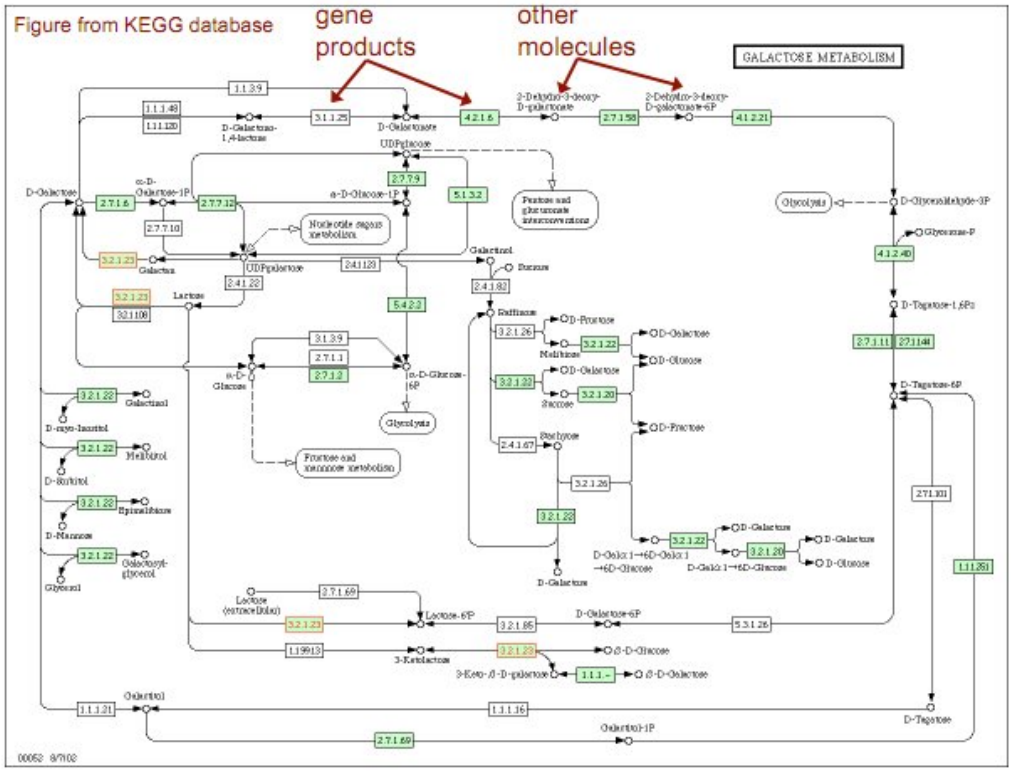
- Feedback loops in expression:
  - (a protein coded by) a gene influences the expression of another gene
  - positively (transcription factor) or negatively (inhibitor)
- Results in extremly complex networks with intricate dynamics



- 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

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

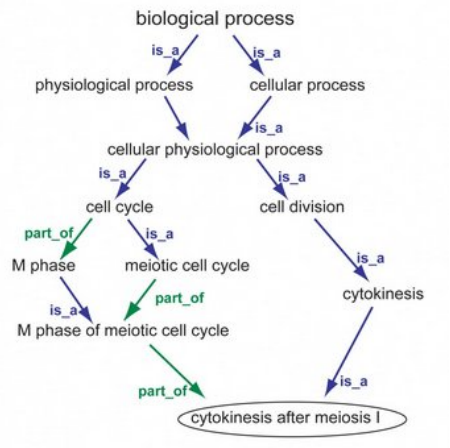
  for example:

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

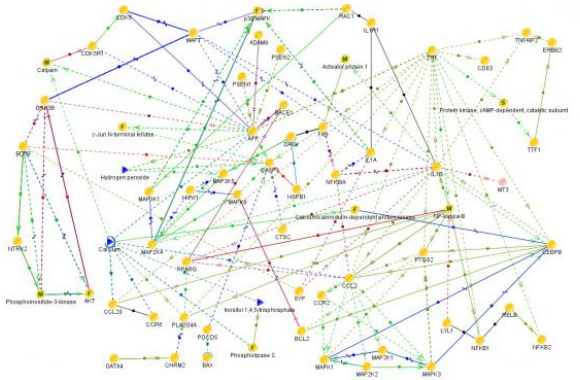
- But how to represent backround knowledge and use it systematically in data analysis?
- Important bioinformatics problem

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]

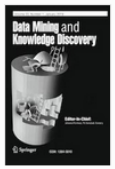
Table 1. Most cited non-review articles from the approximate start of the Internet Age (~1994) to 2013 according to the Institute for Scientific Information (ISI) Web of Knowledge

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.

Bioinformatics at the IDA lab

We regularly publish in bioinformatics and medical journals



Data Mining and Knowledge Discovery

January 2019, Volume 33, Issue 1, pp 1–23 | Cite as

Estimating sequence similarity from read sets for clustering next-generation sequencing data

Petr Ryšavý, Filip Železný

BioMed Central

BMC Genomics

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[a]

BMC Bioinformatics

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]

IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

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

Bioinformatics at the IDA lab

Recent theses and student projects

- **Example thesis:** fungal genome detection on decomposing leaf litter.
- Partners: Institute of Microbiology (Czech Academy of Sciences), Faculty of Science (Charles University).
- Description: metagenome analysis, the task is to understand its structure.
- Our task: automatic intron detection in fungal genomes.
- Methods: probabilistic models (HMMs), classification (SVM + string kernels).

Bioinformatics at the IDA lab

Motivation: metagenome analysis

- Metagenome: genetic material directly from environmental samples.
- Goals: evaluation of genetic diversity, detection of organisms, understand its relationship with other features (climate, soil).
- Obstacles: detection of prokaryotes (bacteria, no introns) much easier than for eukaryotes (fungi, introns present).
- Result: obvious fungal underestimation (60% expected, 5% actually detected).

Bioinformatics at the IDA lab

Technical task: automatic intron detection in fungi

- Input: ∼1,000 annotated fungal genomes (exons/introns known),
  ∼60GB of raw data (fasta, gff annotations),
  other 90,000 fungi named and around 1,000,000 expected.
- Output: a tool that annotates an unknown fungal sequence,
  efficient, accurate, able to generalize across fungal genomes,
  (perfect positions, otherwise reading frame problems)
- Utilization: fungal genomes could be recognized from protein
  (amino acid) sequences (via alignment/approximate match).