

Bioinformatics: course introduction

Filip Železný

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

- 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/Prerequisites

- ▶ mainly: probability and statistics, algorithms (complexity classes), programming skills
- ▶ also: discrete math topics (graphs, automata), relational databases

- Lectures by FZ may be held in English (pending your consensus)

- Purpose of this lecture

Sneak informal preview of the major bioinformatics topics

Teachers



Doc. Filip Železný
CTU Prague, Dept. of Cybernetics
zelezny@fel.cvut.cz



Dr. Jiří Kléma
CTU Prague, Dept. of Cybernetics
klema@labe.felk.cvut.cz



Prof. Zdeněk Sedláček
Charles Univ., Dept. of Biology and Medical Genetics



Ing. Ondřej Kuželka
CTU Prague, Dept. of Cybernetics
kuzelon2@fel.cvut.cz

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

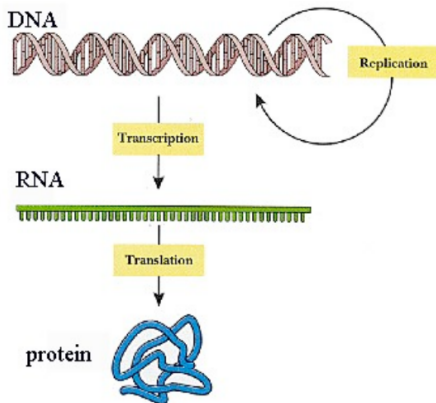
- ▶ representation
- ▶ storage
- ▶ retrieval
- ▶ **analysis**

of gene- and protein-centric biological data

- Not just bio databases!
- Also: computational biology
- Related: systems biology, structural biology

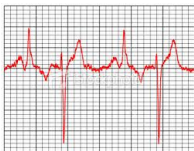
Bioinformatics: Main sources of data

- 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

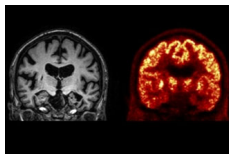


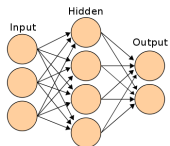
image analysis



healthcare informatics

not usually associated with bioinformatics.

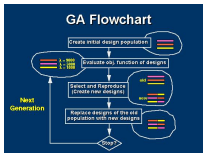
Bioinformatics vs. Bio-Inspired Computing



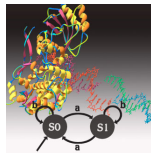
Artificial neural networks



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

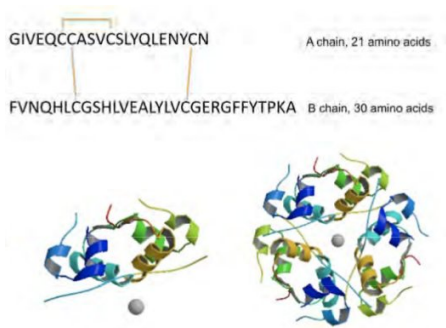
- Basic biological research
- Personalized health care
- Gene-therapy
- Drug discovery
- etc.

Czech landscape

- Small community (FEL, VSCHT, MMF, ...)
- High demand (IKEM, IEM, IMB,)
- 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

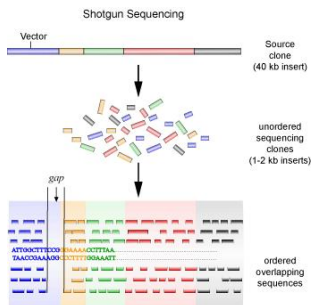
- Sequencing: reading the letters in the macromolecules of interest

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 ϕ X 174	5,375
1978			Bacteriophage ϕ X 174	5,386
1981			Mitochondria	16,569
1982			Bacteriophage λ	48,502
1984			Epstein-Barr virus	172,282
2004			<i>Homo sapiens</i>	2.85 billion

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

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

CA--GATTCGAAT
CGCCGATT---AT

mismatch

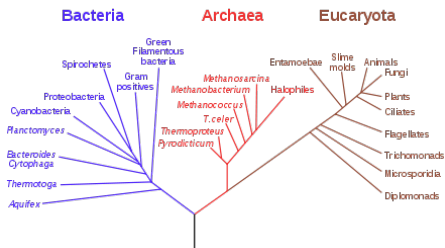
gap

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

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

Phylogenetic Tree of Life



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

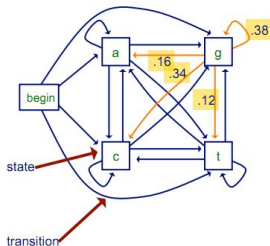
```
tctgaasatgagctgttgacaattaatcatcogaactagttaaactagtagcgaagtcc  
acgggaagaaaaacgctgacattttaacacgcttggttacaaggtaaaaggcagcggc  
aaattaaaaattttatgacttaggtcactaaataactttaacaaatataagcattagcg  
ttgtcataastcgacttgaacocaaattgaaaagatttaggtttacaagctcaccc  
catcctcgcaccagctcagcagcggtttacgctttacgtatagtgggacaaattttt  
tccagataaatttggcataaattaaagtagcagcagataaaaattacataacctggccg  
acagttatccactattcctgtgataaccatgtgtattagagttagaaaaacagag
```

these sequences are not promoters

```
atagctcagagctcttgacctactacgcccagcattttggcgggtgaagtaaccatt  
aaactcaaggctgatacggcgagacttggagccttggctcctggcgtagcacagcagcg  
tactgtgaacattattcgtctccggactacgatgagatgocctgagtgcttccgtt  
tattctcaacaagattaacgcagacagattcaactctcgtggatggcagcttcaacattga  
aacgagtaaatcagaccgctttgactctggattactgtgaacattattcgtctccg  
aagtgcttagcttcaaggtcacggatcagaccgaagcagcagcctcctcctcaatggcc  
gaagaccacgctcggccacggagtagacccttagagagcagatgtagcctcagcctcgaacat
```

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

```
ccatcaaaaaaaaaattctcaacataaaaaaaaaactttgtgtaatacttgtaacgctacat
```

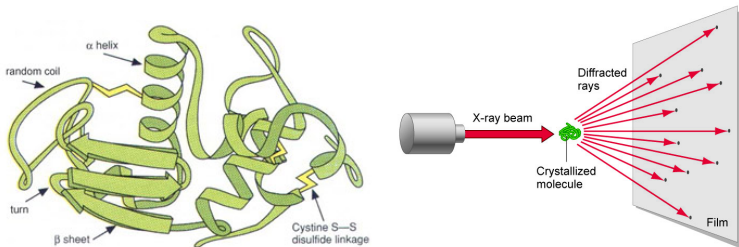


Markov Chain Model

- Each type of site has a different probabilistic model

Protein Spatial Structure

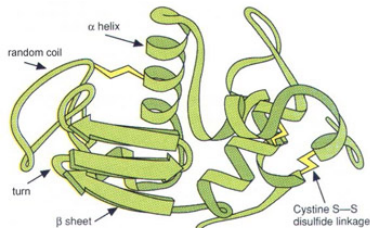
- From the DNA nucleic-acid sequence, the protein amino-acid sequence is constructed by cell machinery
- The protein folds into a complex spatial conformation



- 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

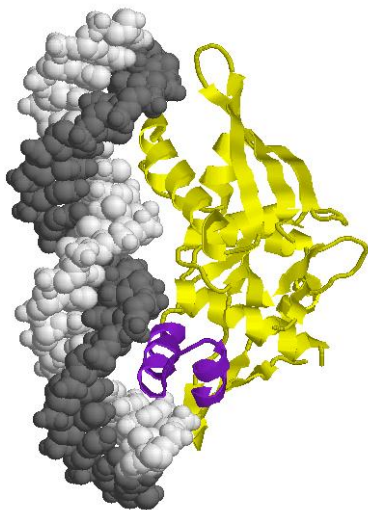
- Can we compute protein structure from sequence?
- At least distinguish α -helices from β -sheets



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

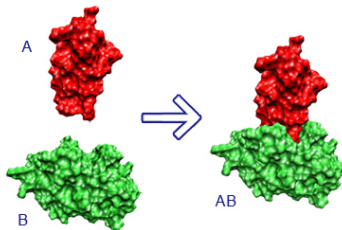
Protein Function Prediction

- 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

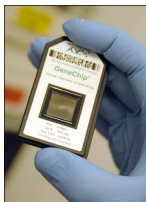
- Proteins interact by *docking*



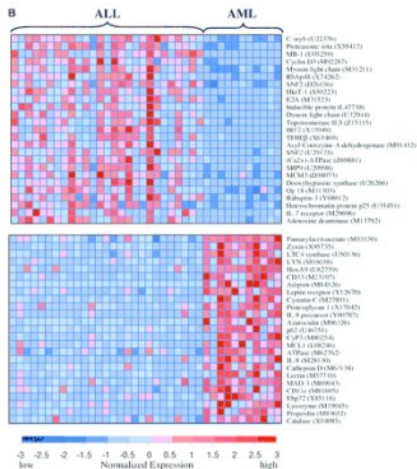
- Will a protein dock into another protein?
- Optimization problem in a geometrical setting
- Important for novel drug discovery
 - ▶ e.g: green - receptor, red - drug
 - ▶ the trouble is, the protein may dock also in many unwanted receptors
 - ▶ immensely hard computational problems under uncertainty

Gene Expression Analysis

- A gene is *expressed* if 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?



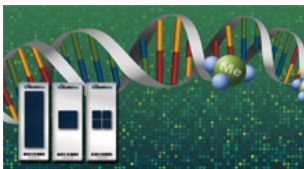
© David Howe



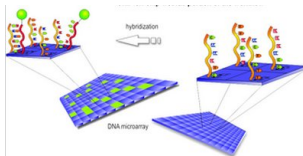
High-throughput data analysis

- Gene expression data are called *high-throughput* 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

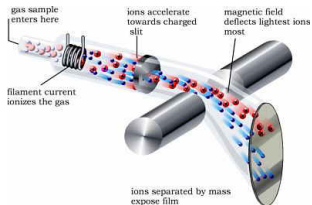
Other high-throughput technologies



Methylation arrays
(epigenetics)



Chip-on-chip
(protein X DNA interactions)

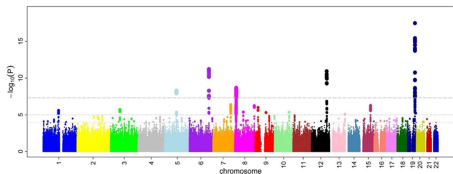
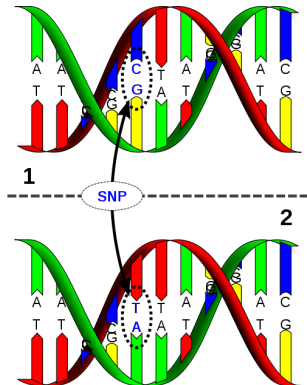


mass spectrometry
(presence of proteins)

..and more

Genome-wide association studies

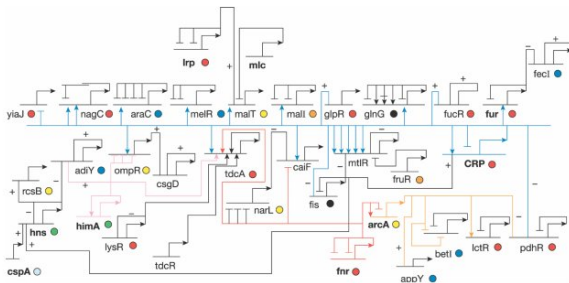
- Correlates traits (e.g. susceptibility to disease) to genetic variations
- “variations”: single nucleotide polymorphisms (SNP) in DNA sequence
- involves a *population* of people



X: SNP's, Y: level of association

Gene Regulatory Networks

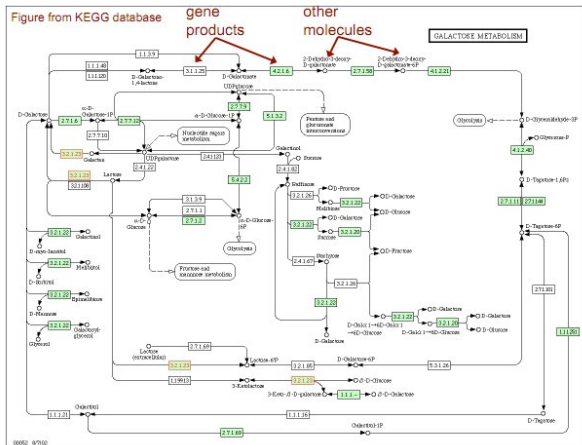
- Feedback loops in expression:
 - ▶ (a protein coded by) a gene influences the expression of another gene
 - ▶ positively (transcription factor) or negatively (inhibitor)
- Results in extremely complex networks with intricate dynamics



- 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

Data \rightarrow Genomic knowledge

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

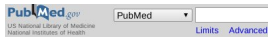
Data + Current Genomic Knowledge \rightarrow New Genomic Knowledge

for example:

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

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

Examples of Genomic Background Knowledge



Display Settings: Abstract

J Pathol 2008 Oct;216(2):141-50.

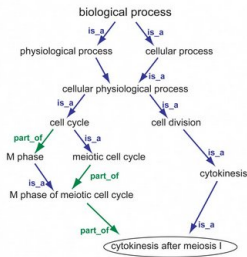
Refinement of breast cancer classification by types.

Weigelt B, Horlings HM, Kreike B, Hayes MM, Hauptmann M, Wessels LF. Division of Experimental Therapy, The Netherlands Cancer Institute, Amsterdam.

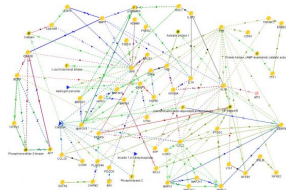
Abstract

Most invasive breast cancers are classified as invasive ductal carcinoma histological 'special types'. These special-type breast cancers are also constitute discrete molecular entities remains to be determined. classification of breast cancer (luminal, basal-like, HER2+). The mol this classification applies to all histological subtypes. We aimed to re histological special types (invasive lobular carcinoma (ILC), tubular, cells, micropapillary, adenoid cystic, metaplastic, and medullary carcinoma). Hierarchical clustering analysis confirmed that some histologic carcinoma, but also revealed that others, including tubular and lobule expression profiling. IDC NOS and ILC contain all molecular breast c

scientific abstracts



gene ontology



interaction networks

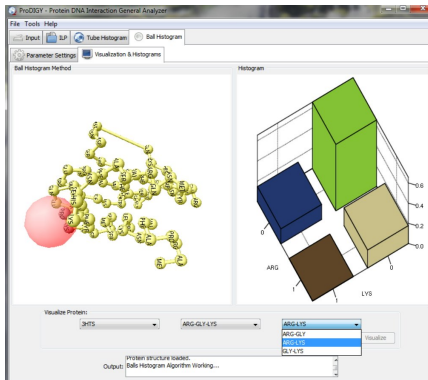
● and many other kinds

Bioinformatics at the IDA lab

Protein structure analysis with machine learning

Prediction of DNA-binding Propensity of Proteins by the Ball-Histogram Method using Automatic Template Search

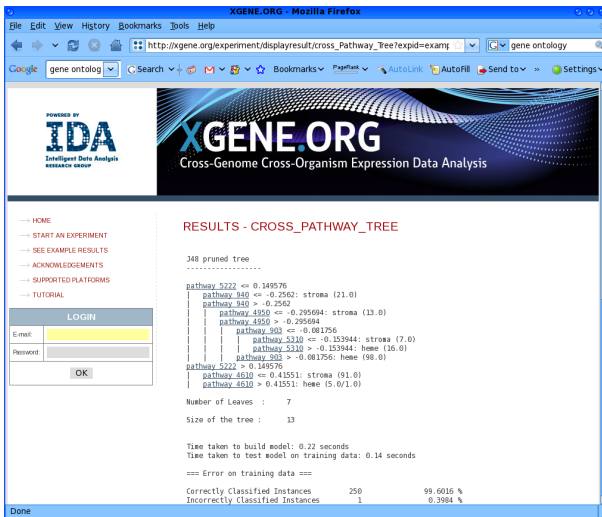
Andrea Szabóová^{*1}, Ondřej Kuželka¹, Filip Železný¹ and Jakub Tolar²



Prodigy Software

Bioinformatics at the IDA lab

Gene expression analysis with machine learning



The screenshot shows a Mozilla Firefox browser window displaying the XGENE.ORG website. The address bar shows the URL: `http://xgene.org/experiment/displayresult/cross_Pathway_Tree?expid=exam...`. The search bar contains "gene ontolog". The website header includes the IDA logo (Intelligent Data Analysis Research Group) and the XGENE.ORG logo (Cross-Genome Cross-Organism Expression Data Analysis).

Navigation links on the left include: HOME, START AN EXPERIMENT, SEE EXAMPLE RESULTS, ACKNOWLEDGEMENTS, SUPPORTED PLATFORMS, and TUTORIAL. Below these is a LOGIN form with fields for Email and Password, and an OK button.

The main content area is titled "RESULTS - CROSS_PATHWAY_TREE" and displays a "348 pruned tree" with the following structure:

```
pathway_5222 <= 0.349576
| pathway_940 <= -0.2562: stroma (21.0)
| | pathway_940 > -0.2562
| | | pathway_4950 <= -0.295694: stroma (13.0)
| | | | pathway_4950 > -0.295694
| | | | | pathway_503 <= -0.081756
| | | | | | pathway_5310 <= -0.153944: stroma (7.0)
| | | | | | | pathway_5310 > -0.153944: heme (16.0)
| | | | | | | | pathway_503 > -0.081756: heme (98.0)
pathway_5222 > 0.149576
| pathway_4610 <= 0.41551: stroma (91.0)
| | pathway_4610 > 0.41551: heme (5.0/1.0)
```

Summary statistics:

- Number of Leaves : 7
- Size of the tree : 13
- Time taken to build model: 0.22 seconds
- Time taken to test model on training data: 0.14 seconds
- === Error on training data ===
- Correctly Classified Instances: 250 (99.6016 %)
- Incorrectly Classified Instances: 1 (0.3984 %)

The status bar at the bottom of the browser window shows "Done".

Bioinformatics at the IDA lab

Gene expression analysis with machine learning



In Silico Biology
An International Journal on
Computational Molecular Biology



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

Matěj Holec¹, Jiří Kléma*¹, Filip Železný¹, Jakub Tolar²

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

Miloš Krejčík and Jiří Kléma

Learning Relational Descriptions of Differentially Expressed Gene Groups

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

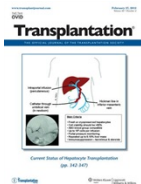
Constraint-based knowledge discovery from SAGE data

Jiří Kléma^{1,3}, Sylvain Blachon², Arnaud Soulet⁴, Bruno Crémilleux¹ 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

Applications in medical studies



Differential Regulation of the Nuclear Factor- κ B 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}



RESEARCH

Open Access

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

Helena Libalová^{1,2}, Kateřina Uhlířová¹, Jiří Kléma³, Miroslav Machala⁴, Radim J. Šrám¹, Miroslav Ciganek⁴ and Jan Topinka^{1*}

Bioinformatics at the IDA lab



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