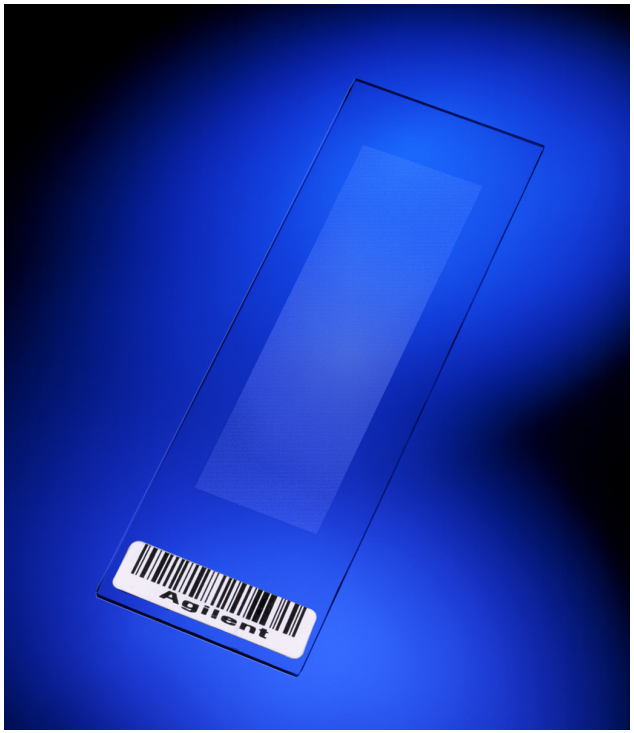
XGENE.ORG: Cross-GENome Cross-ORGanism Expression Data Analysis

Filip Železný, Jiří Kléma, Matěj Holec, Jiří Bělohradský

Czech Grant Agency project 201/09/1665 (2009-2011)
Czech Academy of Sciences project 1ET101210513 (2004-2009)

April 15, 2009

Outline



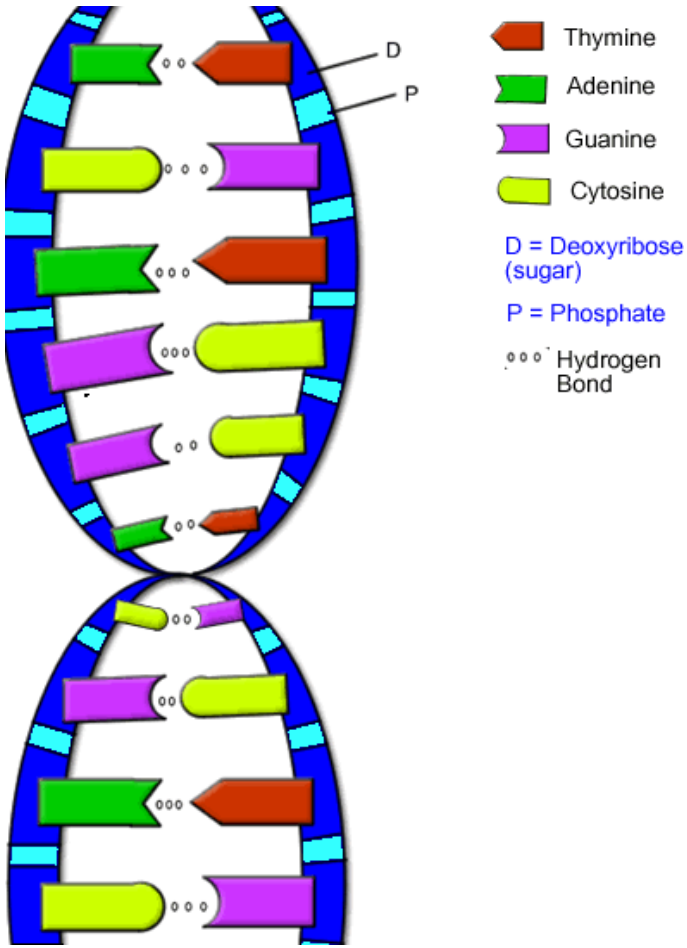
- Microarray chips measure the expression (activity) of genes in a cell
  - Simultaneously for tens of thousands genes
- Data reveal genetic underpinning of diseases, cell differentiation, etc
- Current data analysis methods
  - Statistical (marker gene detection, clustering)
  - Machine learning (predictive classification)

- XGENE.ORG attacks two current challenges in expression data analysis.

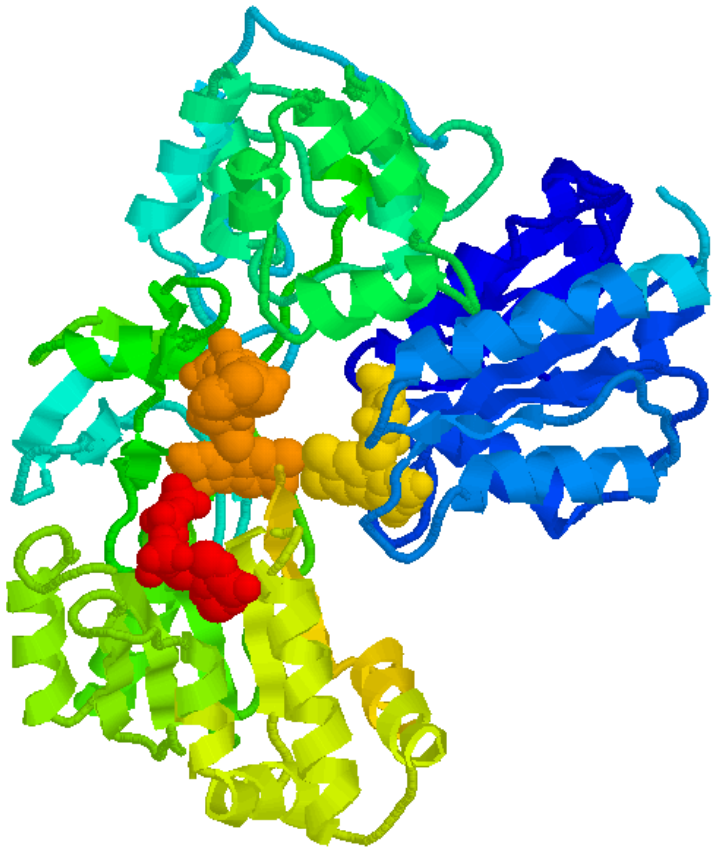
Biological Background

DNA and Protein

- DNA: a sequence on a 4 symbol alphabet

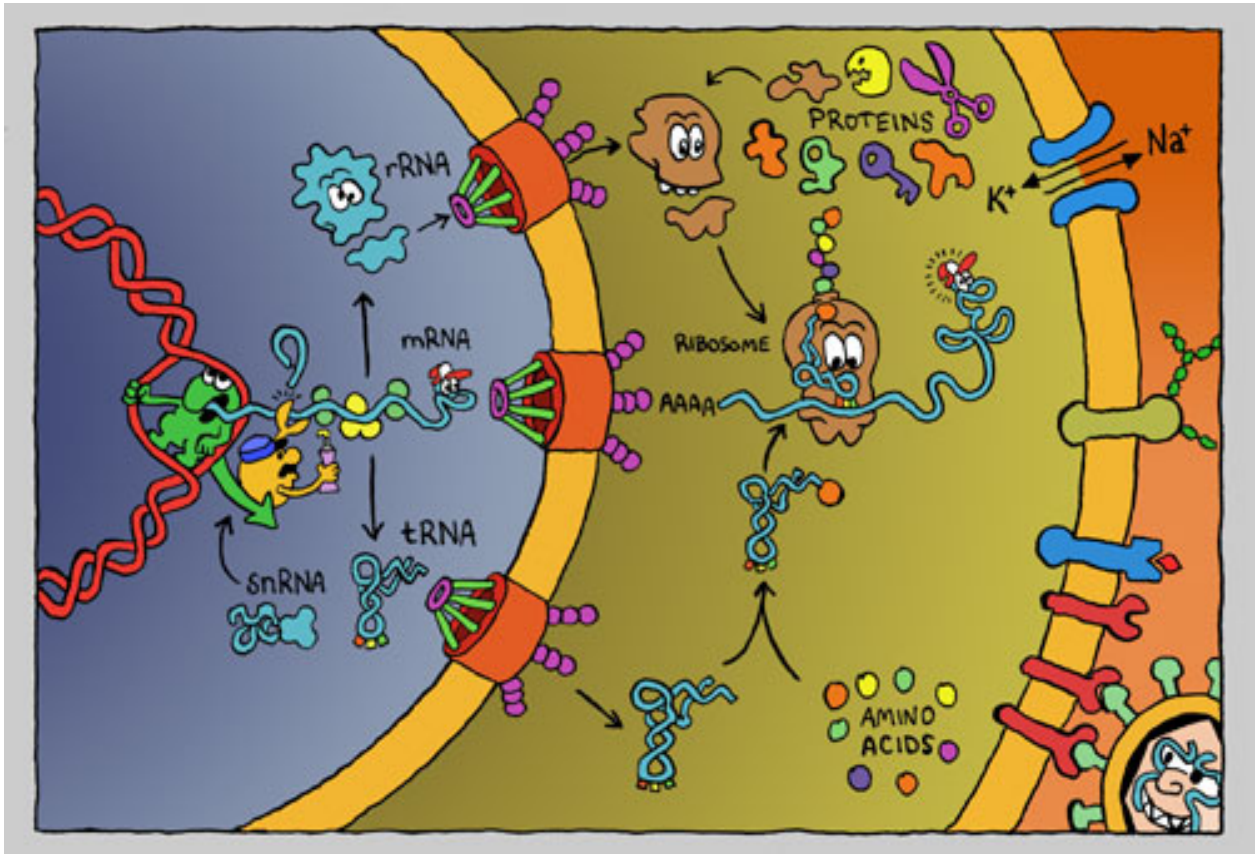


- Protein: a sequence on a cca 25 symbol alphabet



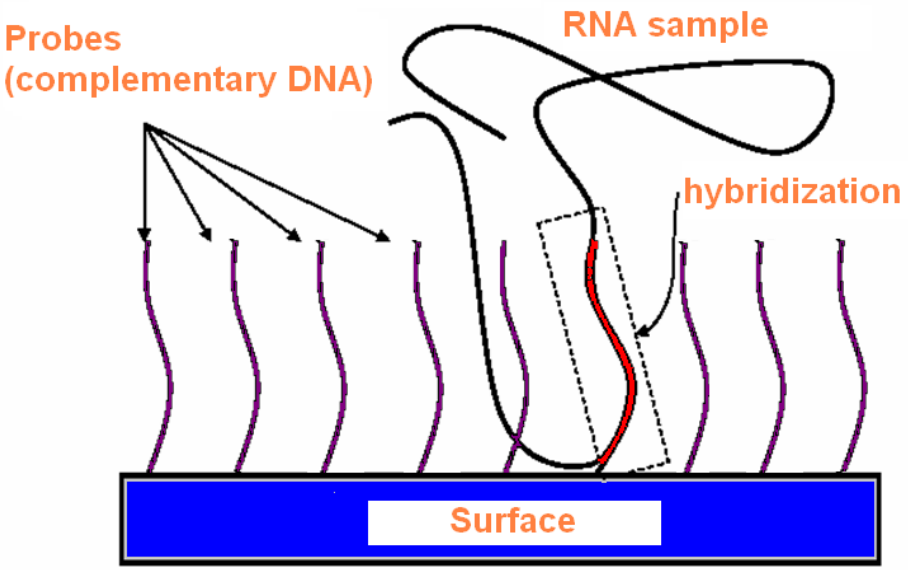
Central Dogma

The Central Dogma of Molecular Biology

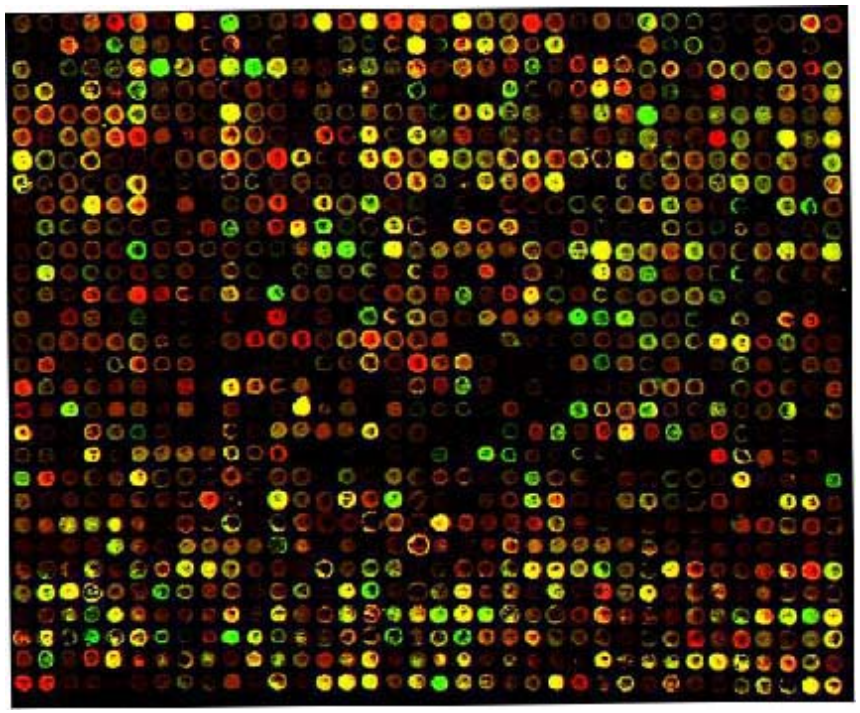


Measuring Gene Expression

- DNA Chip (aka microarray)

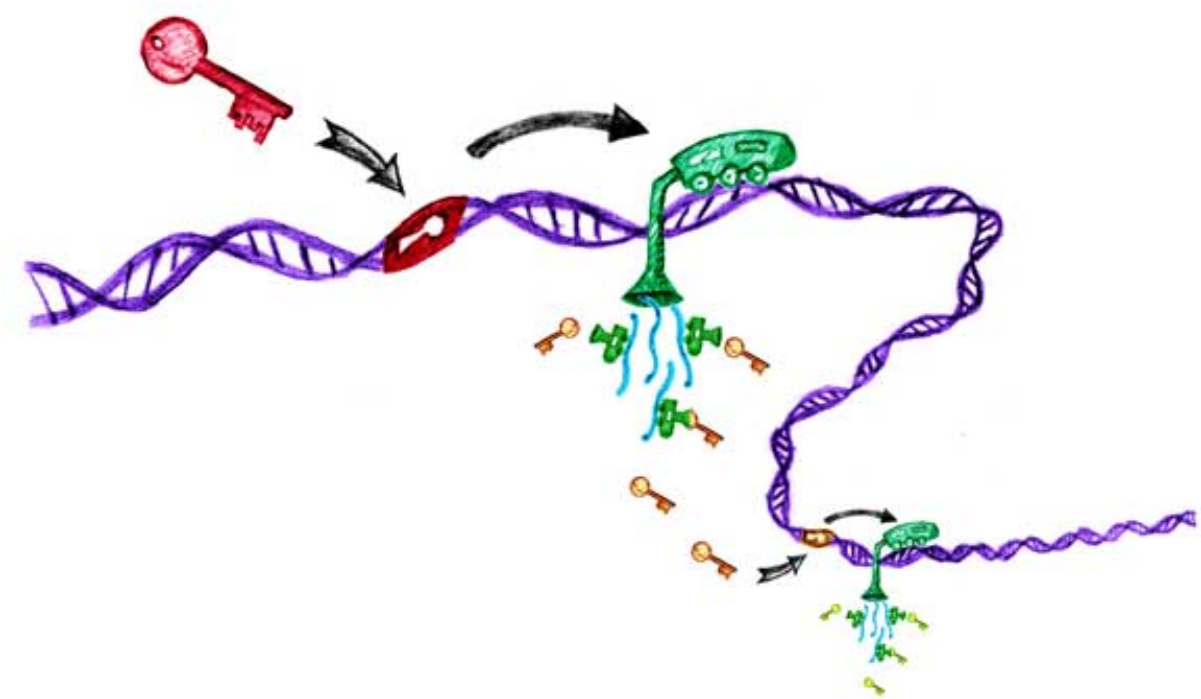


- Scanned



Gene Expression Regulation

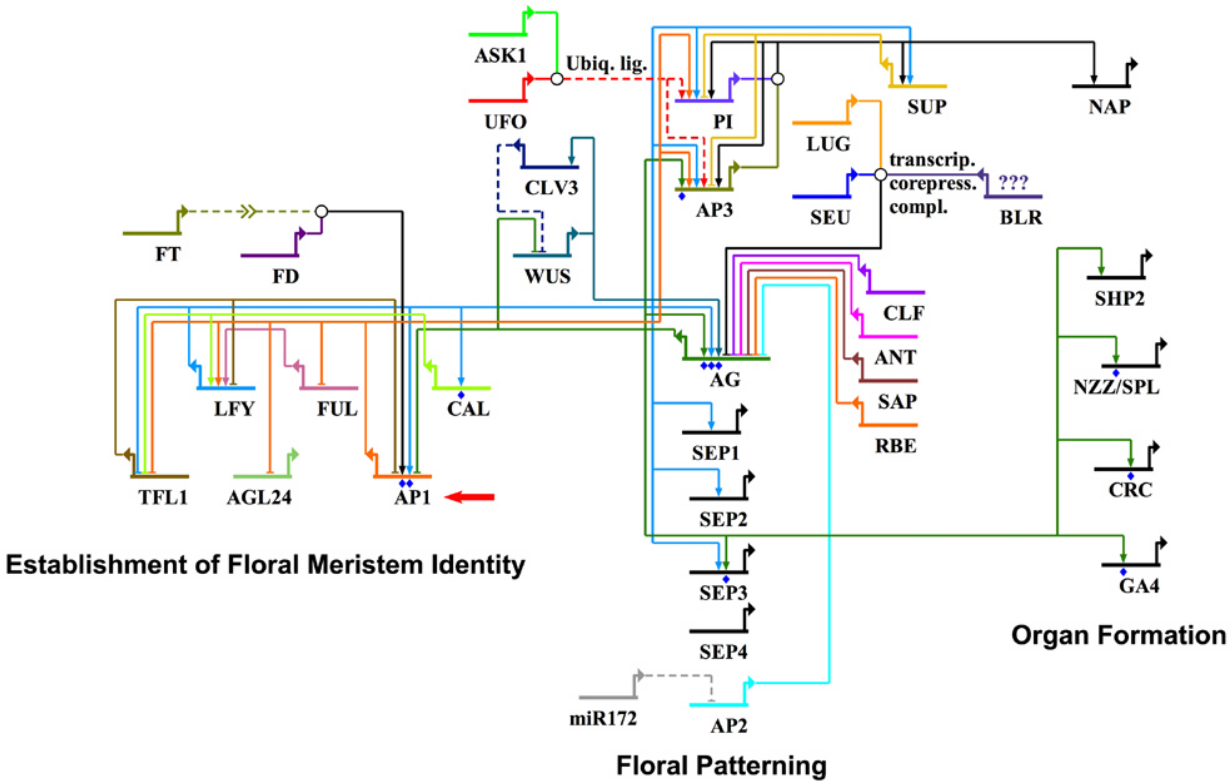
Transcription factor



- One gene regulates the expression of another gene.

Gene Expression Regulation

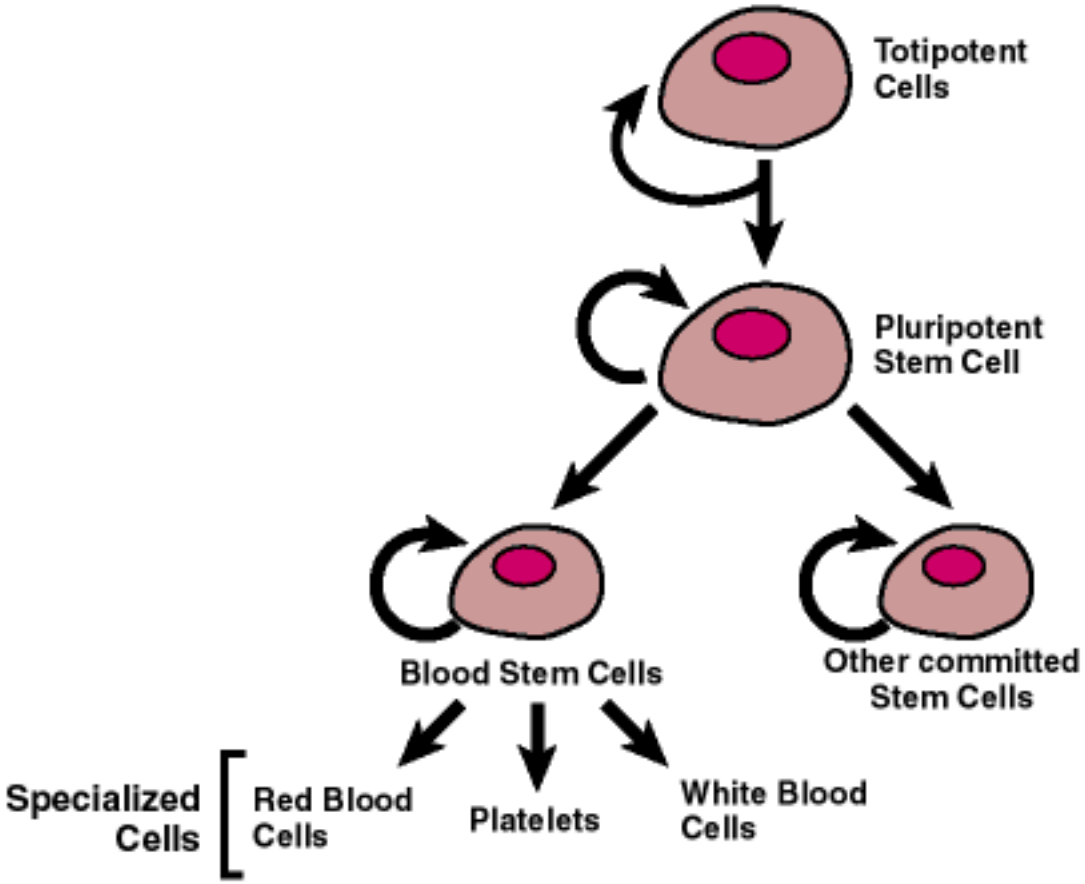
Transcription Network (Pathway)



- A complex dynamic system of mutual gene regulation.

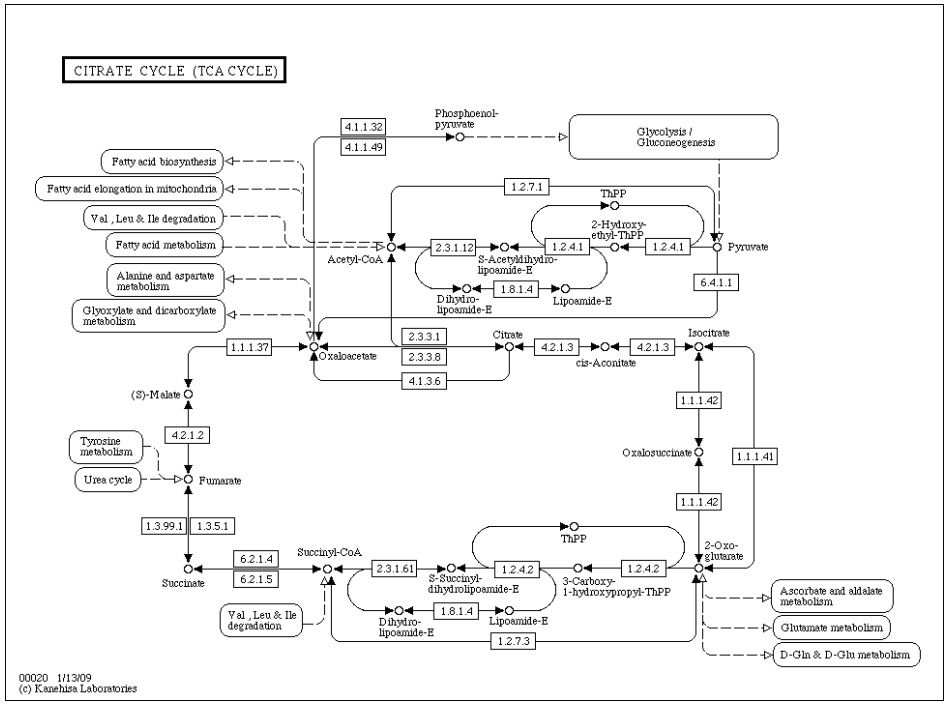
Specialization

Transcription dynamics is the key to cell specialization.



Metabolism and Signalling

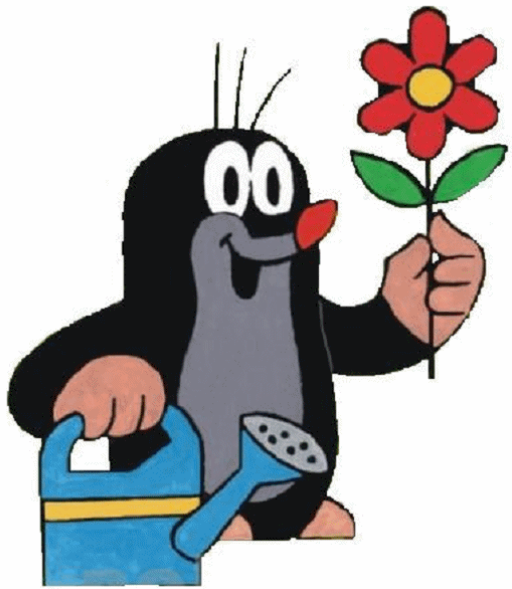
Metabolic and signalling pathways



- Energy processing. Protein act as enzymes.
- Much faster than transcription networks.
-

Metabolism and Signalling

Network orthologs preserved among species.



- Different genes
- Similar (mutually mappable) pathways.

POWERED BY IDA Intelligent Data Analysis RESEARCH GROUP

Our Software

Challenge 1: Heterogeneous Data

Use case: discover features distinguishing two tissue types

Option 1: extract RNA & measure own samples

- Pro: Homogeneous data
- Con: Cost. With 50 samples per type, $100k only for arrays

Perhaps someone did the job before

Option 2: Collect samples from a public database

- Pro: Lots of free-of-charge, relevant samples
- Con: Heterogeneous data

Challenge: create models from heterogeneous data.

Challenge 2: The Gene-List Syndrom

The gene-list syndrom

Example result

Differentially expressed are genes RASSF1, FOXP1, ALOX12, ZNF217, RBL2, ALOX15, CD248, HSPBAP1, EPB41L3, S100A10, SERPINA1, A1BG, UBA1, TNFAIP3, ...

Biologist would prefer

Example result

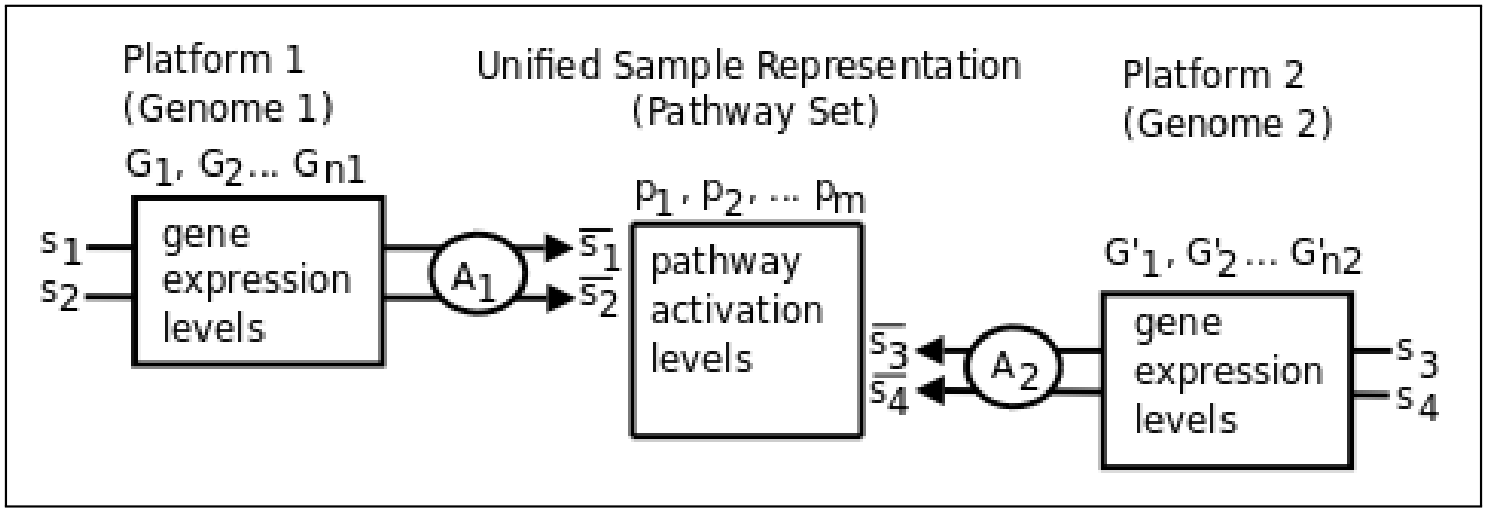
Differentially expressed are genes acting as enzymes in the oxidative phosphorylation metabolic pathway.

Challenge: discover results in terms of complex units (functions, processes)
Needs **background knowledge**

State of the Art

- The gene list syndrom (challenge 1) partially solved
- **Gene Set Enrichment Analysis** (GSEA)
  - GSEA takes apriori defined gene sets
  - Detects the overexpressed
- Integrating multiple-species expression data (challenge 2) not addressed before
  - To our best knowledge
  - But called for

XGENE.ORG Strategy

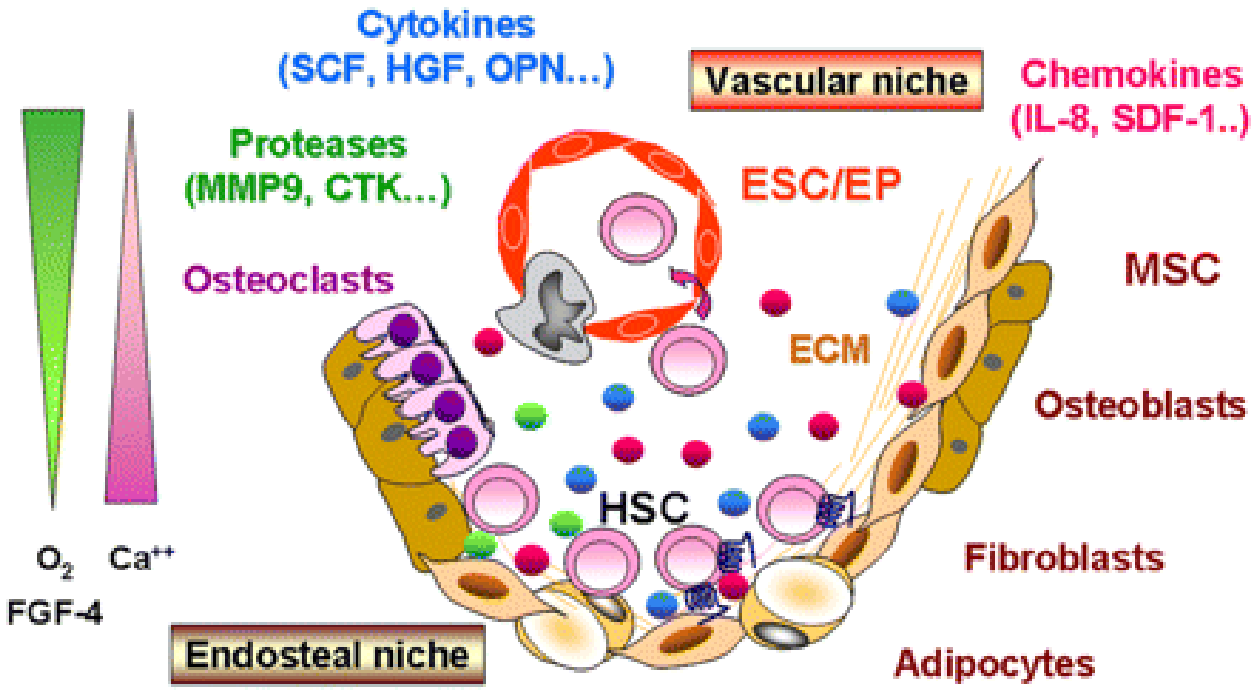


- Background knowledge sources
  - The Gene Ontology, KEGG Pathways, Probeset annotations (Affymetrix, Bioconductor)
- Meshup technology

XGENE.ORG Ingredients

- **Statistics:**
  - Bioconductor package in R (normalization, ANOVA, PCA)
- **Artificial Intelligence:**
  - Machine Learning (Weka)
  - Prolog-based graph analysis for metabolic flux extraction
- **Software Technology**
  - Server-side application, computation on IDA's grid
  - Web-based user interface
  - JAVA: import of expression data
  - RUBY, Apache: web environment

Use case: stromal vs. hematopoietic stem cells



- Explain differences in terms of gene activity
- Large sample sets of gene expression needed (at zero budget)
- Can be found only at a *multi-platform* level (various genomes, organisms, DNA chips), from public repositories
- Multi-platform analysis also provides more general insights

Starring

Matej Holec:

- Multi-platform expression data integration,
- Comparative survey
- New gene-set types

Karel Moulik:

- How can pathway activity be best estimated from expression data?

Jiri Belohradsky:

- XGENE.ORG implementation

Filip Zelezny, Jiri Klema

Publications

- Holec M., Zelezny F., Klema J., Tolar J.: *Integrating Multiple-Platform Expression Data through Gene Set Features*. ISBRA 2009: the 5th International Symposium on Bioinformatics Research and Applications, Springer 2009
  - Interesting findings question state-of-the-art beliefs on gene-set based analysis
  - pdf on Filip's website
- XGENE.ORG whitepaper Accepted to Int. Conf. on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC 2009)