Gene Network Reconstruction:

Fundamental and Current Research at Microarray Center

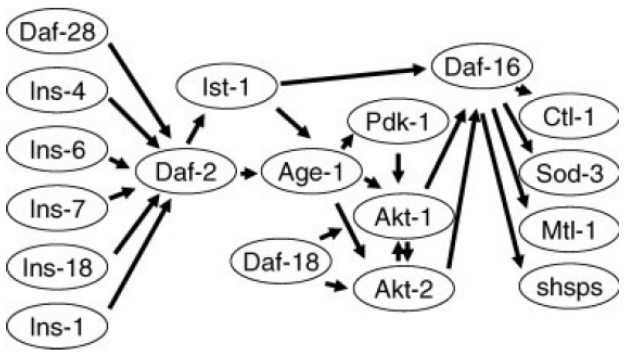
Petr Nazarov
Loïc Couderc

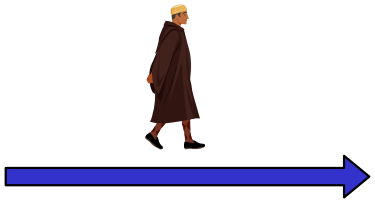
26-06-2009

OUTLINE

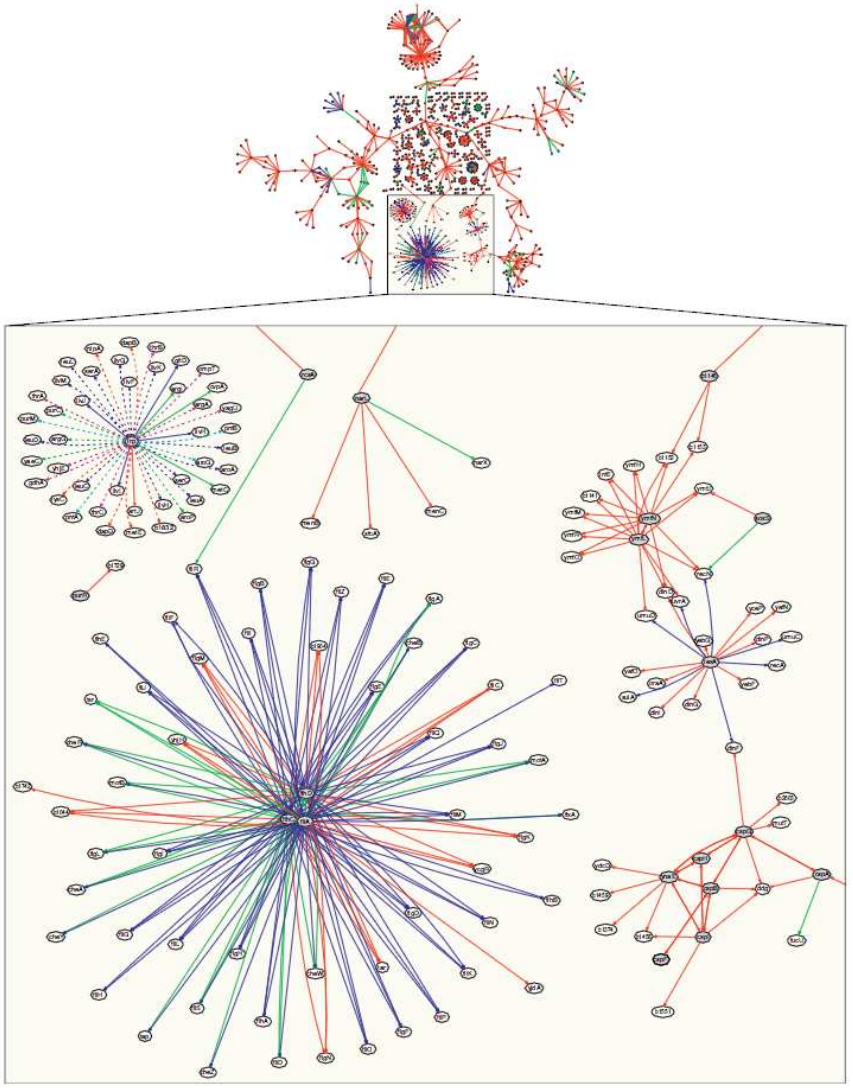
- **Introduction**
- Methods of **gene networks** reconstruction
- Problems and solutions
- **CoExpress** software demonstration (Hypoxia)
- **Public data meta-analysis**
- **Future plans**
  - Exon-level network reconstruction
  - Detection of AS genes
- **Exon-level networks**

Single pathway



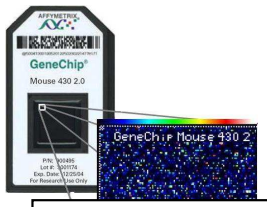


Genome-wide pathways

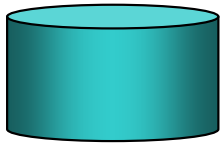


- How to transform experimental data into networks?
- How to validate the hypotheses about networks?
- How to use the knowledge about the networks?

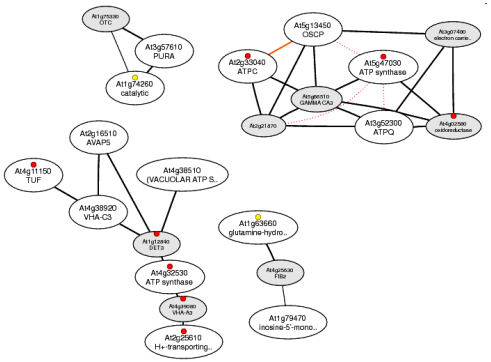
Experiments



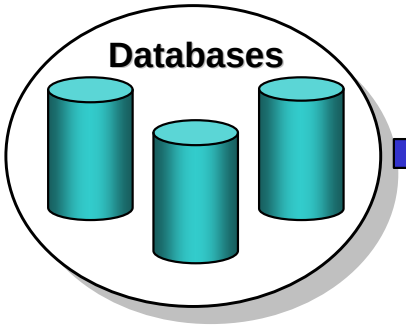
Public data



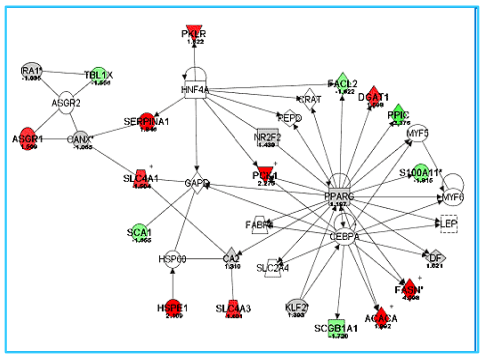
Co-expression network



Databases



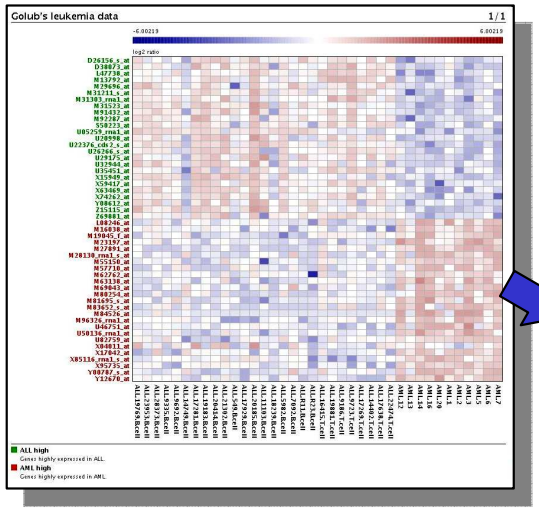
Known gene networks and P-P interactions



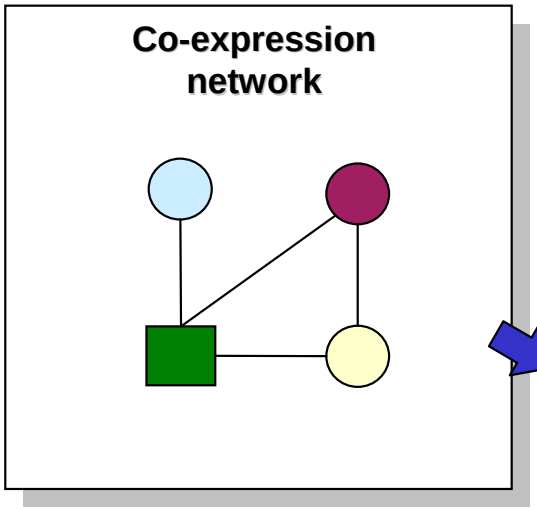
Analysis

- New interactions (pathways)
- Better predictors (markers)
- Disease-related pathways
- Disease-causing genes

GENERAL “PIPELINE”

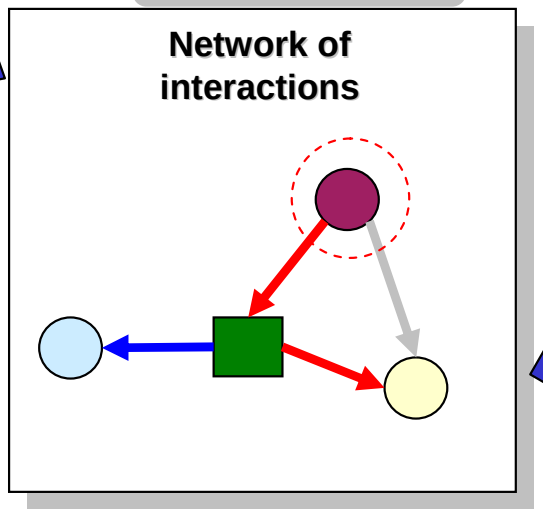


Co-expression analysis

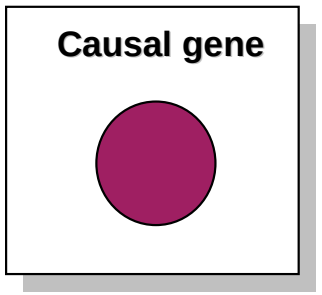


Topology analysis

Databases



Databases: PPI, TF



Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles

Faith, J.J, et al. (2007) PLoS

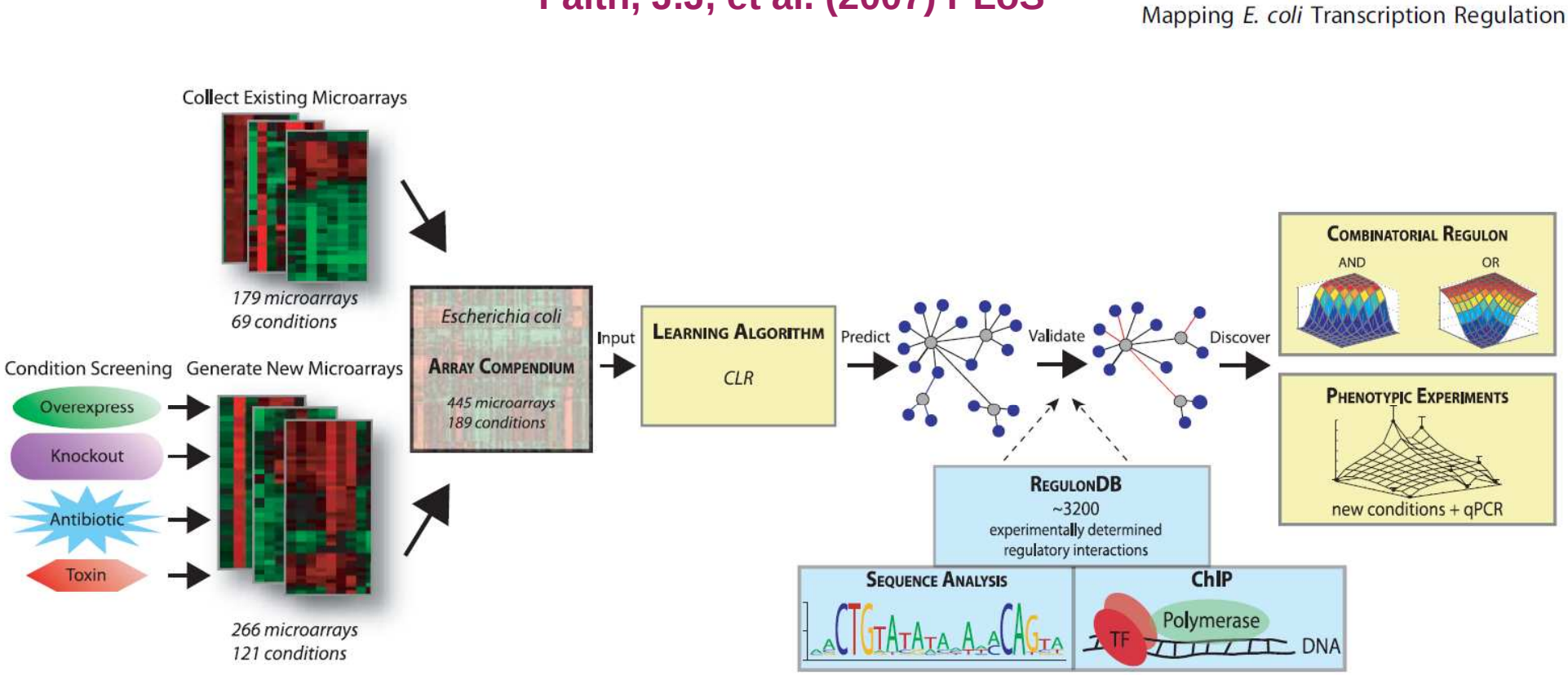
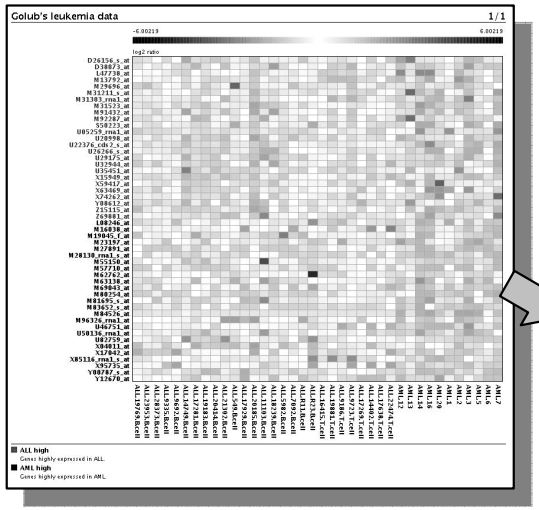


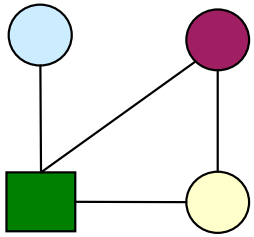
Figure 1. Overview of Our Approach for Mapping the *E. coli* Transcriptional Regulatory Network

Microarray expression profiles were obtained from several investigators. Our laboratory profiled additional conditions, focusing on DNA damage, stress responses, and persistence. These two data sources were combined into one uniformly normalized *E. coli* microarray compendium that was analyzed with the CLR network inference algorithm. The predicted regulatory network was validated using RegulonDB, sequence analysis, and ChIP. The validated network was then examined for cases of combinatorial regulation, one of which was explored with follow-up real-time quantitative PCR experiments.



Co-expression analysis

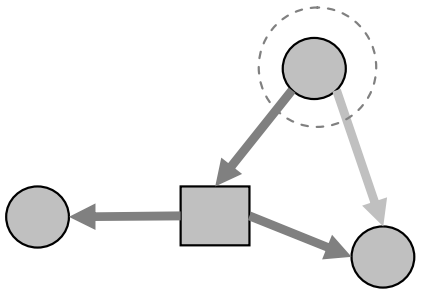
Co-expression network



Topology analysis

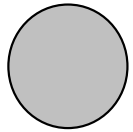
Databases

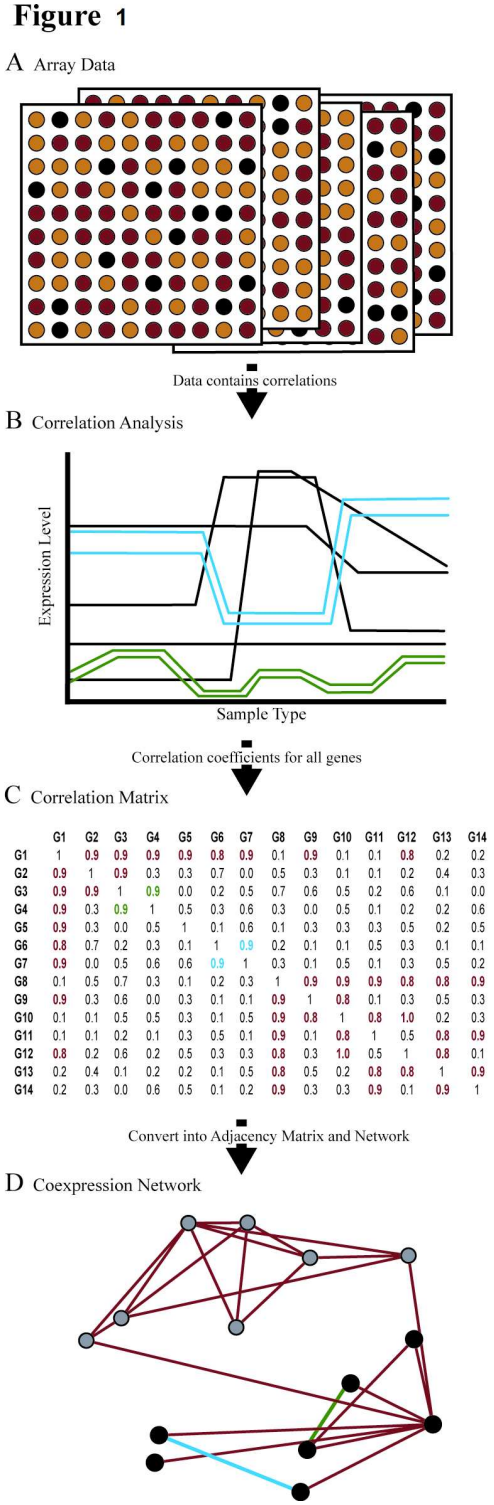
Network of interactions



Databases: PPI, TF

Causal gene





	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

Co-expression Network Reconstruction

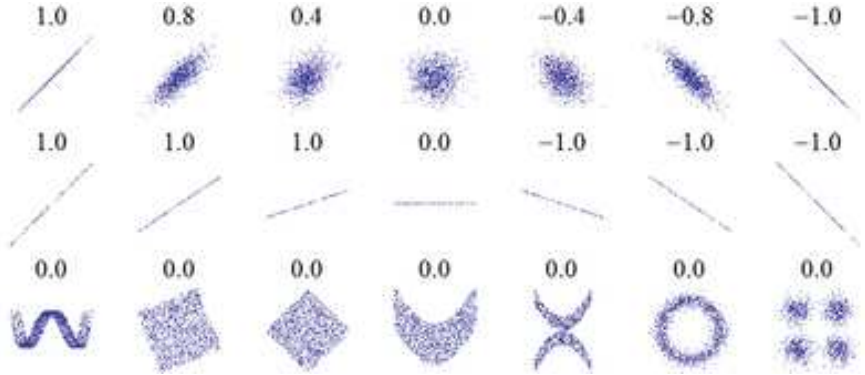
- Microarray transcriptomic data (A)
  - 2-color with a common reference or 1-color data or
  - Data normalization
- Concordance of gene expression (co-expression) (B)
  - Pearson or Spearman correlation (linear interaction)
  - Mutual information (non-linear interaction)
  - Coefficient of determination
  - etc… (> 10 various methods can be found easily)
- Transformation of concordance matrix (C)
  - CM can be dichotomized → unweighted network
  - or transformed continuously → weighted network
- Building and topology analysis of the co-expression network (*math*.: *undirected graph*) (D)
  - modules detection
  - connectivity analysis
  - optimal visualization

Adopted and adapted from Hovath et al.
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/

Correlation

- Detect linear dependency between expression profiles
- Pros and cons:
  - (+) Simple, with a well-defined statistics
  - (-) Only linear interactions can be predicted

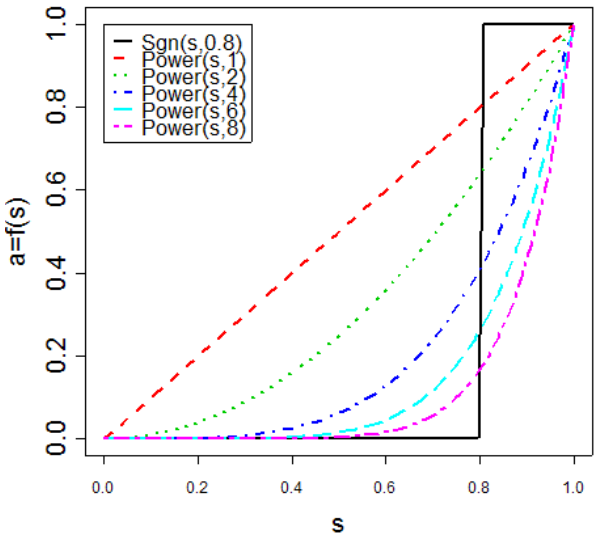
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x \sigma_y}$$



- Modification of the correlation: put power

$$a_{xy} = r_{xy}^{\beta}$$

Often choosing $\beta = 6$ works well but in general the "scale free topology criterion" described in Zhang and Horvath 2005 can be used.



Mutual information

- Detect non-linear patterns between expression profiles
- Pros and cons:
  - (+) Detect nonlinear patterns
  - (-) Expression data should be discretized
  - (-) May lead to overestimation of the interactions

$$MI_{xy} = \sum_{i,j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)\,p(y_j)}$$

Example:

Expression of 2 genes based on 100 microarrays.

Expression discretization:

0-5: no or low expression
5-8: some expression
8-11: significant expression
11-16: high expression

gene 2

gene 1	[0-5]	[5-8]	[8-11]	[11-16]
[0-5]	9	6	5	5
[5-8]	7	11	8	7
[8-11]	2	8	8	3
[11-16]	4	6	7	4

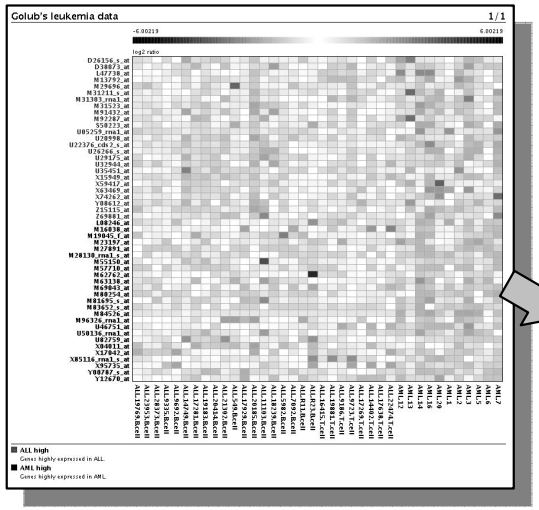
→ No interaction, *MI* = 0.01

gene 2

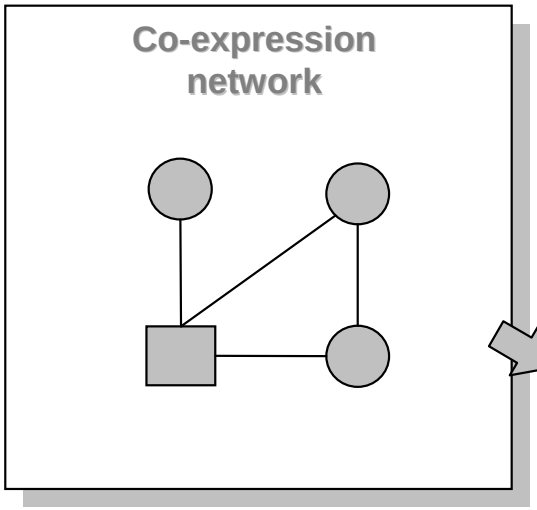
gene 1	[0-5]	[5-8]	[8-11]	[11-16]
[0-5]	0	2	6	3
[5-8]	3	6	**14**	**20**
[8-11]	4	1	11	10
[11-16]	5	7	2	6

→ Interaction, *MI* = 0.05

TOPOLOGY ANALYSIS

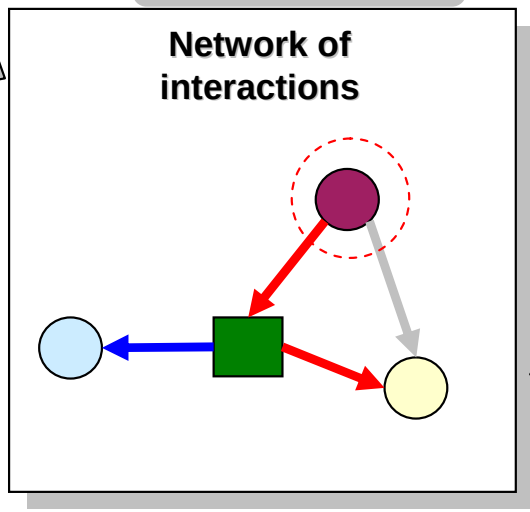


Co-expression analysis

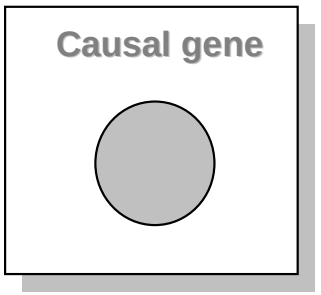


Topology analysis

Databases



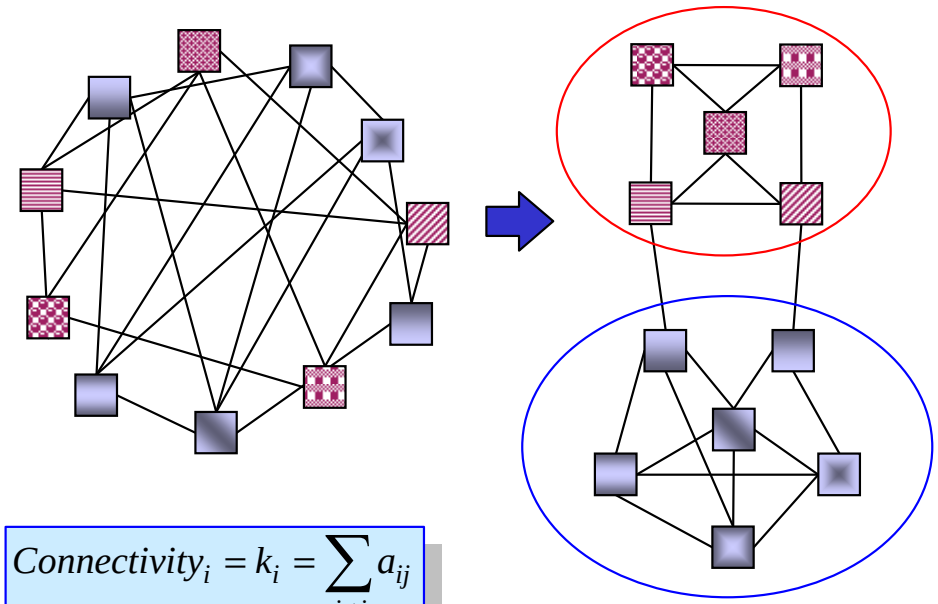
Databases: PPI, TF



TOPOLOGY ANALYSIS

Raw network

Processed network



$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{S_1(k)}{n(n-1)} = \frac{mean(k)}{n-1}$$

where n is the number of network nodes.

$$Centralization = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right) \approx \frac{\max(k)}{n-1} - Density$$

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 - \sum_{l \neq i} a_{il}^2}$$

Main aims of the topology analysis

- Determine the sub-networks or **modules** (usually highly connected)
- Perform the **classification** of the network topologies
  - Distinguish between network of healthy and ill cells
- Optimal **visualization**

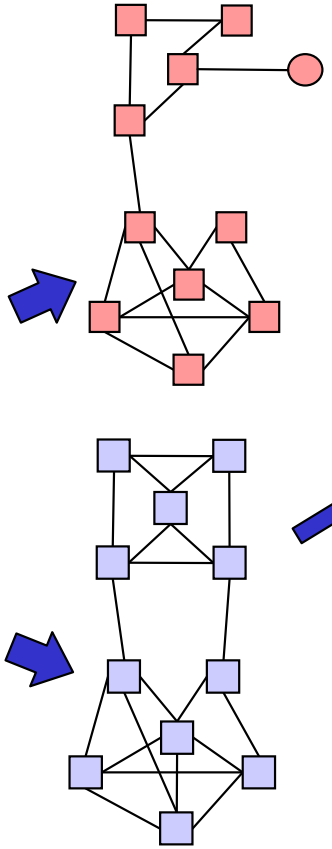
Methods

- Use graph theory (mathematics)
- Introduce network concepts (indices):
  - connectivity (how node is connected)
  - density (mean adjacency)
  - centralization (determines centers)
  - clustering coefficient, etc…
- Use weighted edges when possible → better results

Example 1

MA data from ill patients

MA data from healthy patients

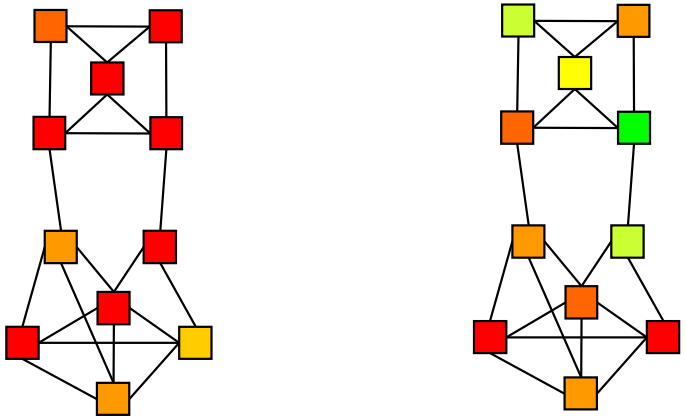


Determine disease-related pathways and define the targets for the treatment

Example 2

healthy patient

ill patient



Classification on the basis of homogeneity

Expertise at CRP

Francisco Azuaje

- **4 books**
- **45 articles**

An integrative approach for causal gene identification and gene regulatory pathway inference

Zhidong Tu, Li Wang, Michelle N. Arbeitman, Ting Chen and Fengzhu Sun*

Molecular and Computational Biology Program, University of Southern California, Los Angeles, USA

Motivation: Gene expression variation can often be linked to certain chromosomal regions and are tightly associated with phenotypic variation such as disease conditions. Inferring the causal genes for the expression variation is of great importance but rather challenging as the linked region generally contains multiple genes. Even when a single candidate gene is proposed, the underlying biological mechanism by which the regulation is enforced remains unknown. Novel approaches are needed to both infer the causal genes and generate hypothesis on the underlying regulatory mechanisms.

Results: We propose a new approach which aims at achieving the above objectives by integrating genotype information, gene expression, protein-protein interaction, protein phosphorylation, and transcription factor (TF)–DNA binding information. A network based stochastic algorithm is designed to infer the causal genes and identify the underlying regulatory pathways. We first quantitatively verified our method by a test using data generated by yeast knock-out experiments. Over 40% of inferred causal genes are correct, which is significantly better than 10% by random guess. We then applied our method to a recent genome-wide expression variation study in yeast. We show that our method can correctly identify the causal genes and effectively output experimentally verified pathways. New potential gene regulatory pathways are generated and presented as a global network.

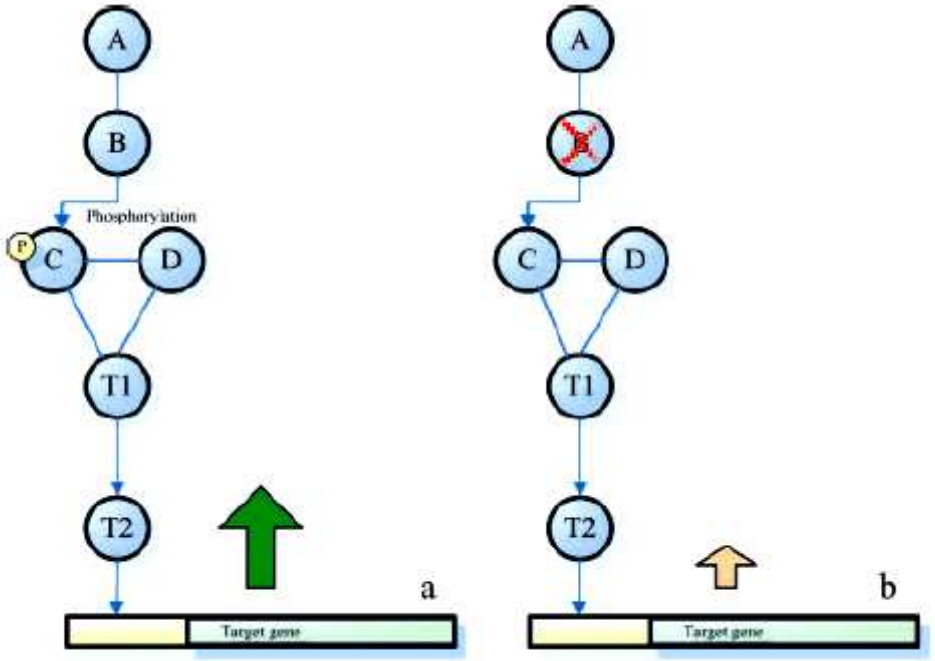


Fig. 1. A conceptual gene regulatory pathway. (a). Genes involved in the pathway are shown as circles (A,B,C,D,T1 and T2). B represents a kinase which activates downstream protein C by phosphorylation. T1 and T2 are transcription factors and T1 positively regulates T2's expression. T2 binds to the promoter region of the target gene and activates its expression. Edges without arrow indicate protein-protein interactions and edges with arrow imply the transcriptional regulations or phosphorylations. (b) Gene B on the pathway is inactivated. The expression of the target gene is down-regulated as consequence. We don't require pathway be strictly linear so that indispensable components of the pathway (e.g., D) can be included.

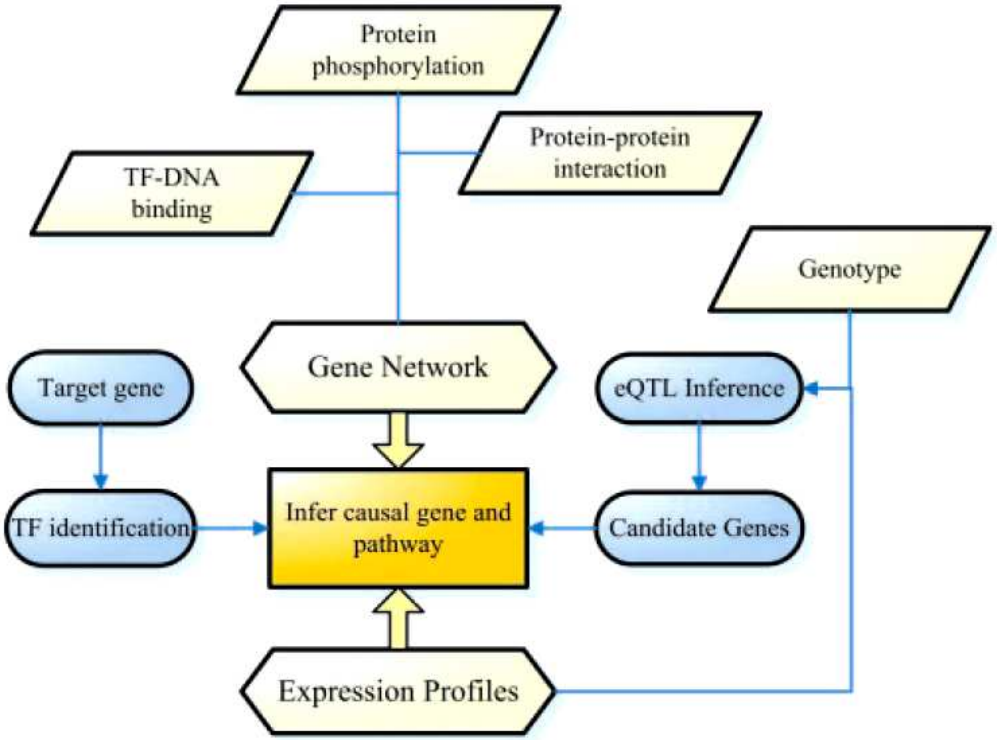


Fig. 2. Overview of our multi-step procedure for causal gene identification and gene regulatory pathway inference.

Goals

- **Enhance analysis we provide on the platform**
  - 2 colors arrays: Agilent, self-spotted
  - Affymetrix (gene and exon level)
- **Develop new tools for the analysis of transcript**
- **Obtain new knowledge concerning cancer-related transcript distortions**
- ***Study exon-level transcriptomics in normal and cancer tissues***

"CoExpress" SOFTWARE TOOL

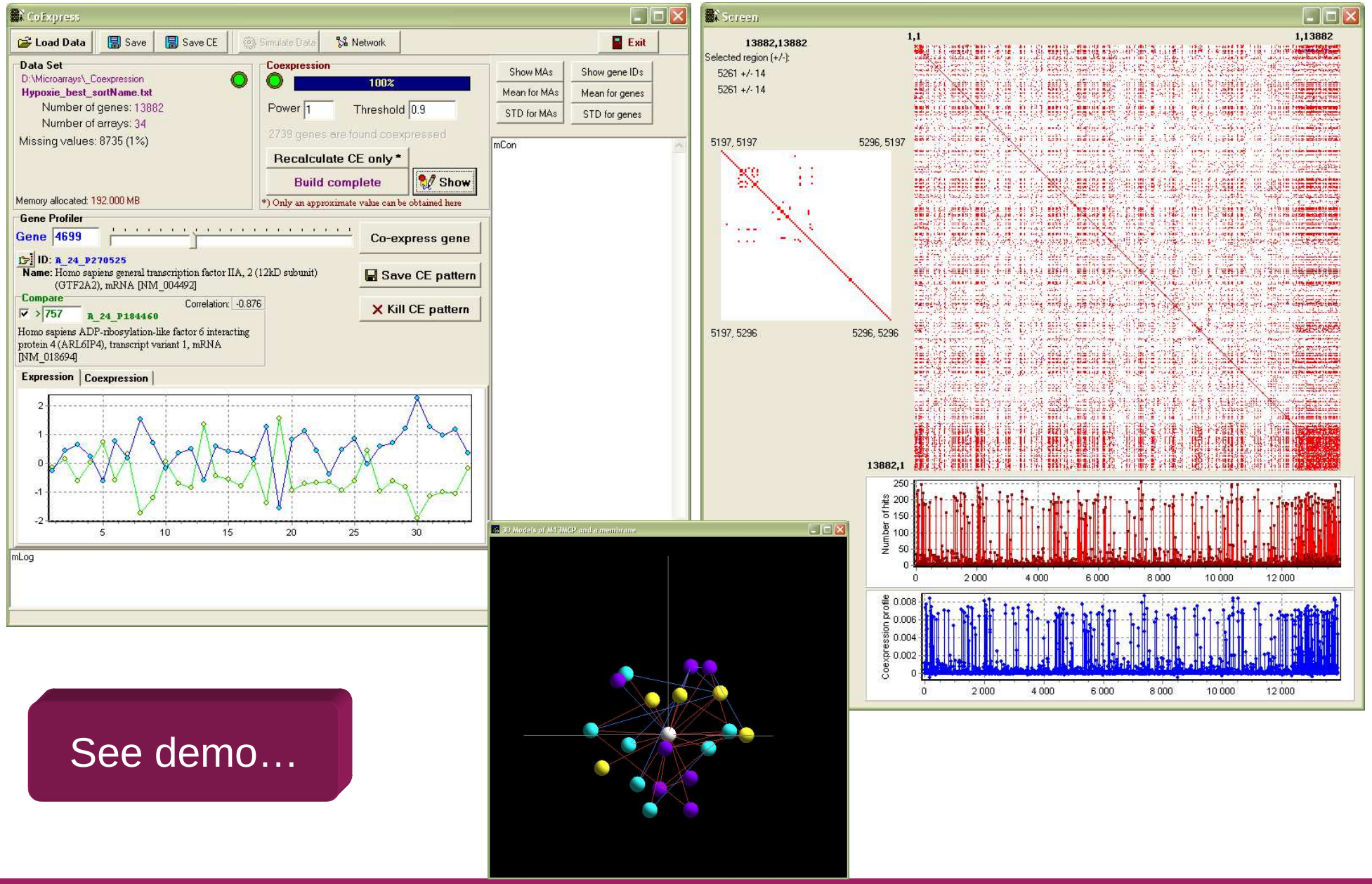
Goal: obtain and study experimentally-derived gene regulatory networks

Features of the current version

- **User friendly**
  - Windows-based, GUI
- **Memory efficient**
  - Simultaneous analysis of up to **40 000 genes** (65 536 is a theoretical limit)
  - 40 000 genes results in **1 600 000 000** values of correlation matrix
- **Fast – coded in C++**
  - 20000x20000 interaction from 34 microarrays are processed in ~5 min
  - GUI programmed in Borland C++ Builder
- Contains some basic tools for ***visualization***, ***export*** and ***suppression*** of the co-expression patterns

CENTRE DE RECHERCHE PUBLIC

“CoExpress” SOFTWARE TOOL



See demo…

Can co-expression pattern be just due to a coincidence?

Experiment. Comparison CE matrixes for experimental and simulated datasets. The filtering criterion selected for this experiment was

$r_{i,j}^2 > 0.8$, i.e. $|r_{i,j}| > 0.894$

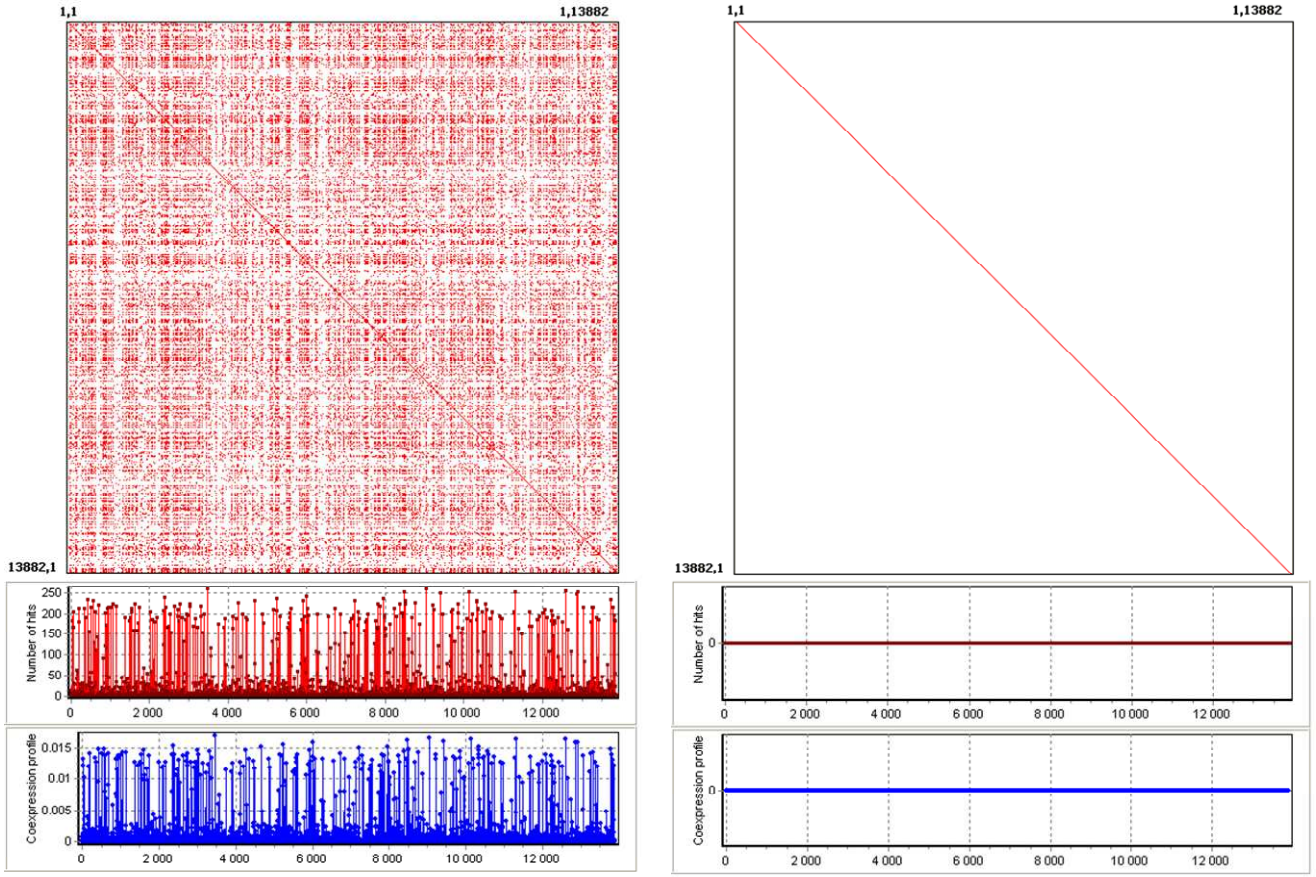


Figure 1. Gene co-expression patterns for the best 13882 genes obtained on 34 experimental (a) and randomized (b) MAs.

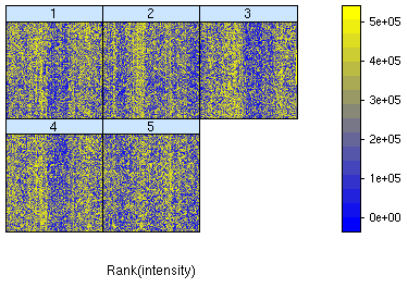
Future of CoExpress

- Introduce additional measures of co-expression
  - Spearman correlation
  - Mutual information
  - Coefficient of determination (check at least)
- Finalize the network visualization
  - Topology analysis
  - Optimal visualization
- Validate the results after application to real data
  - Hypoxia condition in cancer cells (data of Bassam Janji)
  - Cardiovascular research (data from Francisco Azuaje)
- Apply to public data
  - Validation
  - New knowledge discovery
- Use for the analysis of splicing aberration in FNR project of Laurent

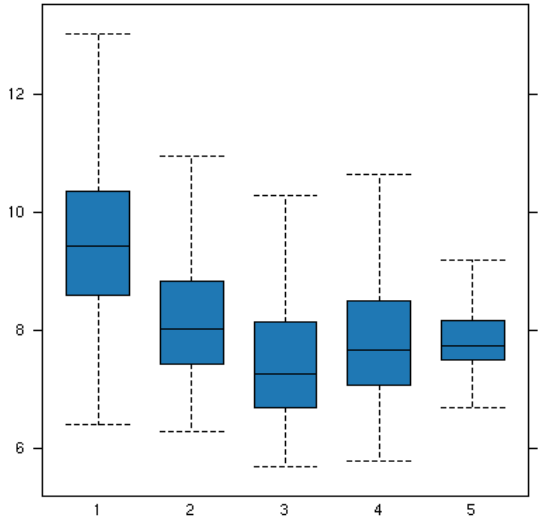
Meta-analysis of the Public Data

- Data sources (public repositories used)
- Workflow
  - Download
  - Metadata annotation and verification
  - QA/QC
  - Normalization/Mapping
  - Merging/Filtering
- Problems and solutions
- Some results: PCA, normalized
- Future plans

- Data an Meta-data come from Geo (**ncbi**) and Array Express (**ebi**)
- Downloading of approximately 10 000 files with 200 pathologies and 300 tissues from 5 five types of platforms (hgu133a,hgu133b,hgu133plu
- **Meta-Data**
  - Manually currated
  - We focus on the localization, pathology and tissues
- **Data**
  - Quality assessment (R scripts)
    - Box plot
    - Spatial distribution
    - ….
  - Quality control
    - We remove all the Data (and Meta-Data) for arrays of poor quality



Spatial distribution



Density plot

Approximatively 7000 files after quality control

Normalization and Mapping

- Normalization
  - Intra-experiment: Perform RMA for each experiment
    - Background correction
    - Normalization
    - Summarization
  - Inter-experiment: Perform a quantile normalization between experiments
  - Mapping
    - Convert probeSet Id into RefSeq id. The mapping is done to be able to combine data from different platformes. Actually, the probeSet from each type of platforme doesn't match necessarely the same RefSeq.

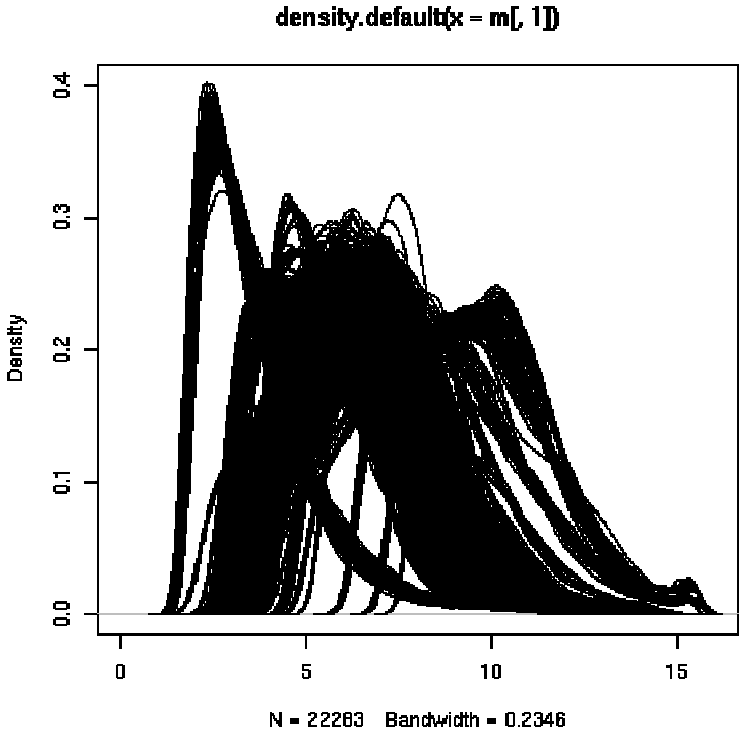
Problems and Solutions

- Problem
  - QA/QC are time consuming and space consuming

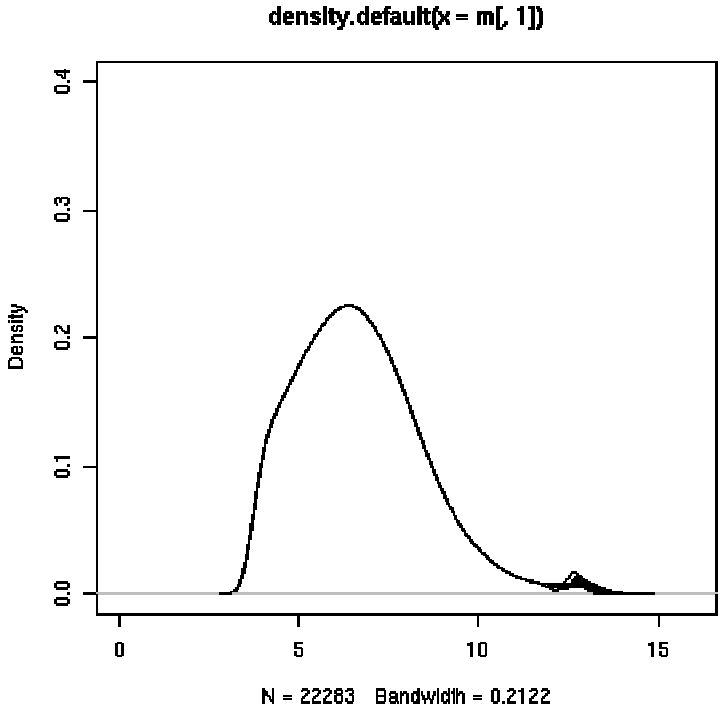
- Solution
  - We have a server with a huge amount of memory (32 GO)
  - Optimization of the source code…
  - Using other programming language (python, C…)

Results of Processing

- Effect of the normalization



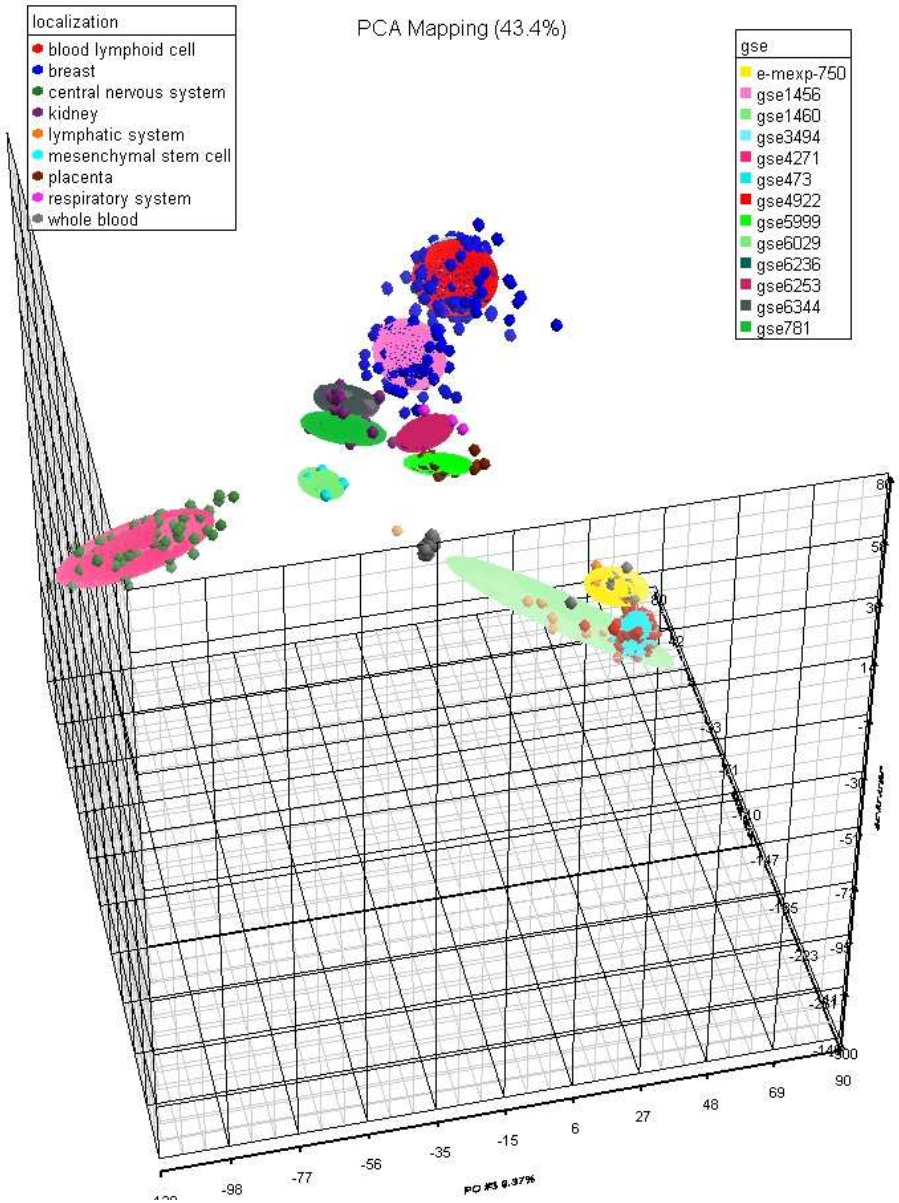
Before normalization



After normalization

Principal Component Analysis (PCA)

As an example, 1000 samples are presented here



Discussion

STORAGE

- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks=number of direct neighbors
  - For weighted networks= sum of connection strengths to other nodes

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

Density

- Density= mean adjacency
- Highly related to mean connectivity

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{S_1(k)}{n(n-1)} = \frac{mean(k)}{n-1}$$

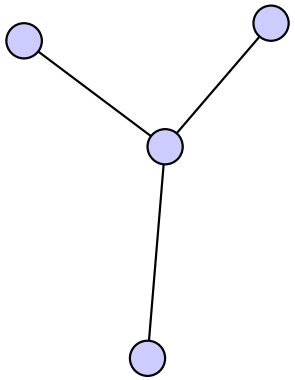
where n is the number of network nodes.

Centralization

$$Centralization = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right) \approx \frac{\max(k)}{n-1} - Density$$

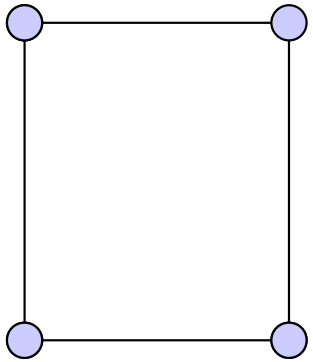
= 1 if the network has a star topology

= 0 if all nodes have the same connectivity



Centralization = 1

because it has a star topology



Centralization = 0

because all nodes have the same connectivity of 2

- Heterogeneity: coefficient of variation of the connectivity
- Highly heterogeneous networks exhibit hubs

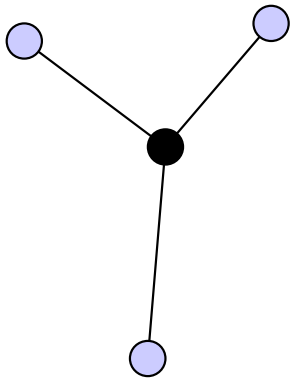
$$Heterogeneity = \frac{\sqrt{variance(k)}}{mean(k)}$$

Clustering Coefficient

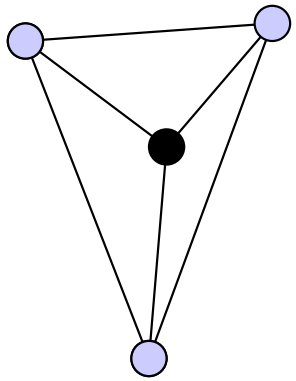
Measures the cliquishness of a particular node
« A node is cliquish if its neighbors know each other »

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il}\right)^2 - \sum_{l \neq i} a_{il}^2}$$

This generalizes directly to weighted networks (Zhang and Horvath 2005)



Clustering Coef of the black node = 0



Clustering Coef = 1

The topological overlap dissimilarity is used as input of hierarchical clustering

$$TOM_{ij} = \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2007) to higher order interactions
- Generalized in Li and Horvath (2006) to multiple nodes