MSc GBE Course:

Genes: from sequence to function

Brief Introduction to Systems Biology

Sven Bergmann

Department of Medical Genetics
University of Lausanne
Rue de Bugnon 27 - DGM 328
CH-1005 Lausanne
Switzerland

work: ++41-21-692-5452
cell: ++41-78-663-4980
http://serverdgm.unil.ch/bergmann

Course Overview

- Basics: What is Systems Biology?
- Standard analysis tools for large datasets
- Advanced analysis tools
- Systems approach to "small" networks

What is Systems Biology?

- *To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. Properties of systems, such as robustness, emerge as central issues, and understanding these properties may have an impact on the future of medicine.*

Hiroaki Kitano

SBI The Systems Biology Institute

What is Systems Biology?

- *To me, systems biology seeks to explain biological phenomenon not on a gene by gene basis, but through the interaction of all the cellular and biochemical components in a cell or an organism. Since, biologists have always sought to understand the mechanisms sustaining living systems, solutions arising from systems biology have always been the goal in biology. Previously, however, we did not have the knowledge or the tools.*

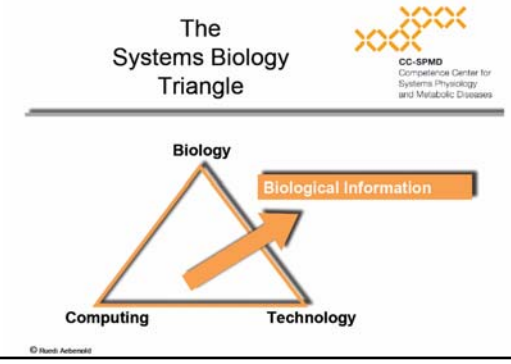
Edison T Liu
Genome Institute of Singapore

What is Systems Biology?

- *addresses the analysis of **entire** biological systems*
- ***interdisciplinary** approach to the investigation of all the components and networks contributing to a biological system*
- *[involves] new **dynamic computer modeling** programs which ultimately might allow us to simulate entire organisms based on their individual cellular components*
- *Strategy of Systems Biology is dependent on **interactive cycles of predictions and experimentation**.*
- *Allow[s Biology] to move from the ranks of a descriptive science to an **exact science**.*

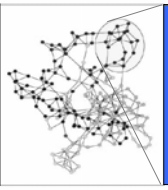
(Quotes from SystemsX.ch website)

What is Systems Biology?



What is Systems Biology?

- **identify *elements* (genes, molecules, cells, …)**
- **ascertain their *relationships* (co-expressed, interacting, …)**
- ***integrate* information to obtain view of system as a *whole***

Large (genomic) systems		Small systems
• many uncharacterized elements • relationships unknown • *computational analysis* should: ▪ improve annotation ▪ reveal relations ▪ reduce complexity		• elements well-known • many relationships established • *quantitative modeling* of systems properties like: ▪ Dynamics ▪ Robustness ▪ Logics

Part 1: Basics

Motivation:

- *What is a "systems biology approach"?*
- *Why to take such an approach?*
- *How can one study systems properties?*

Practical Part:

- *First look at a set of genomic expression data*
- *How to have a global look at such datasets?*
- *Distributions, mean-values, standard deviations, z-scores*
- *T-tests and other statistical tests*
- *Correlations and similarity measures*
- *Simple Clustering*

First look at a set of genomic expression data

Cell, Vol. 102, 109–126, July 7, 2000, Copyright ©2000 by Cell Press

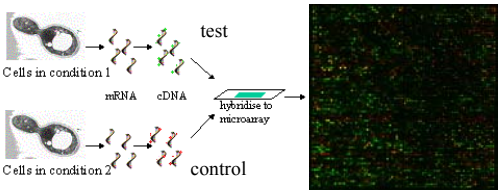
Functional Discovery via a Compendium of Expression Profiles

Timothy R. Hughes,*# Matthew J. Marton,*# Allan R. Jones,* Christopher J. Roberts,* Roland Stoughton,* Christopher D. Armour,* Holly A. Bennett,* Ernest Coffey,* Hongyue Dai,* Yudong D. He,* Matthew J. Kidd,* Amy M. King,* Michael R. Meyer,* David Slade,* Pek Y. Lum,* Sergey B. Stepaniants,* Daniel D. Shoemaker,* Daniel Gachotte,† Kalpana Chakraburtty,‡ Julian Simon,§ Martin Bard,† and Stephen H. Friend*‖
*Rosetta Inpharmatics, Inc. 12040 115th Avenue N.E. Kirkland, Washington 98034

Summary

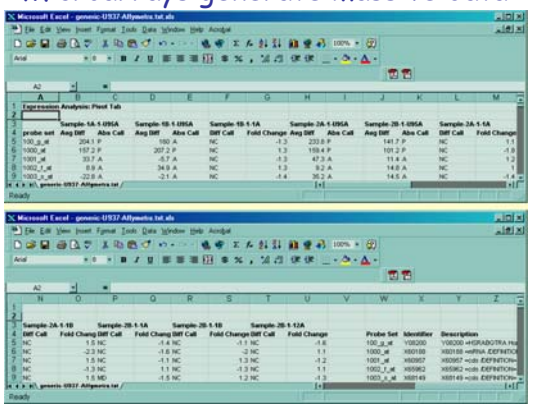
Ascertaining the impact of uncharacterized perturbations on the cell is a fundamental problem in biology. Here, we describe how a single assay can be used to monitor hundreds of different cellular functions simultaneously. We constructed a reference database or "compendium" of expression profiles corresponding to 300 diverse mutations and chemical treatments in *S. cerevisiae*, and we show that the cellular pathways affected can be determined by pattern matching, even among very subtle profiles. The utility of this approach is validated by examining profiles caused by deletions of uncharacterized genes: we identify and experimentally confirm that eight uncharacterized open reading frames encode proteins required for sterol metabolism, cell wall function, mitochondrial respiration, or protein synthesis. We also show that the compendium can be used to characterize pharmacological perturbations by identifying a novel target of the commonly used drug dyclonine.

DNA microarray experiments monitor expression levels of thousands of genes simultaneously:



- allows for studying the *genome-wide* transcriptional response of a cell to interior and exterior changes
- provide us with a first step towards understanding gene function and regulation on a *global* scale

Microarrays generate massive data



Log-ratios of expression values

$$r = \log(E_{test} / E_{control}) = \log(E_{test}) - \log(E_{control})$$

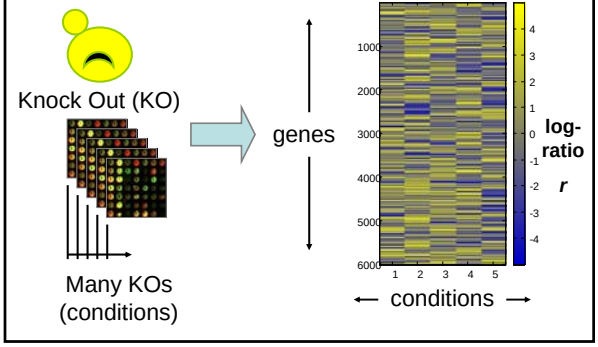
$$E_{test} < E_{control} \quad | \quad E_{test} > E_{control}$$

$$- \longleftarrow \quad 0 \quad \longrightarrow +$$

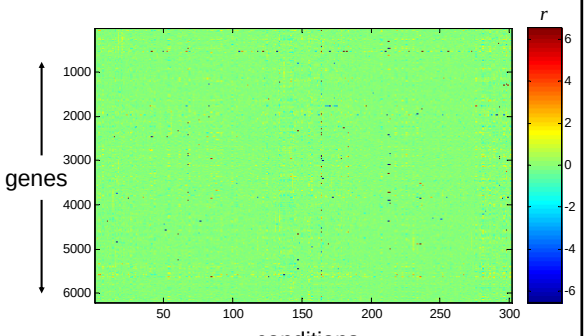
$$E_{test} \simeq E_{control}$$

Log ratios indicate **differential** expression!

Consolidate data from multiple chips into one table and use color-coding

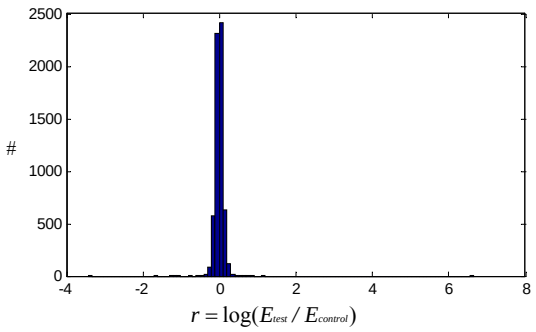


Rosetta data: The real world …



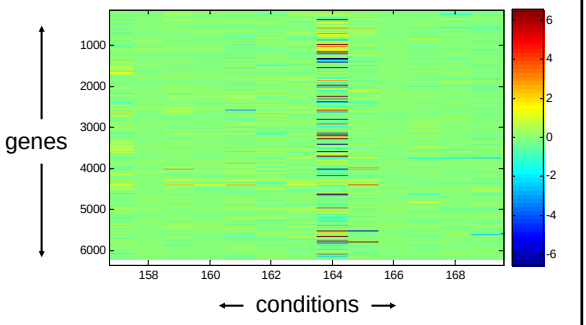
Most genes exhibit little differential expression!

Histogram shows distribution



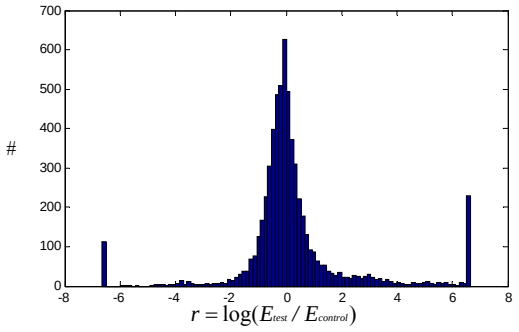
ade1 deletion mutant exhibits small differential expression in most genes!

Rosetta data: Zooming in …



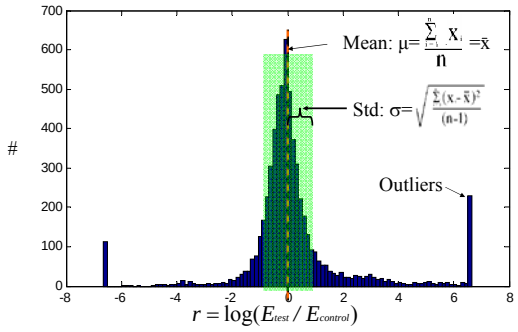
Only few genes exhibit large differential expression!

Histogram shows distribution



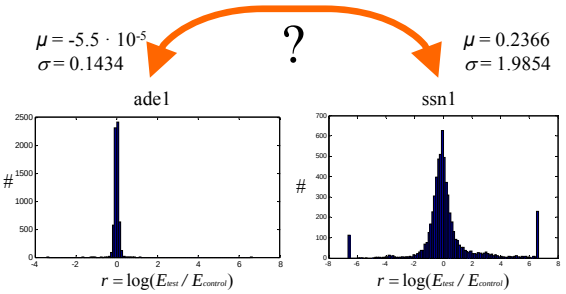
ssn6 deletion mutant exhibits large differential expression in many genes!

Quantification of distribution



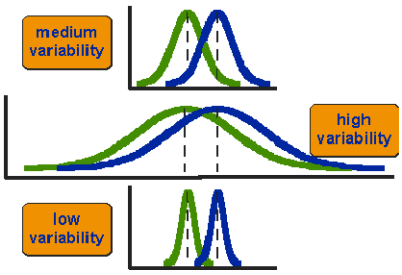
Mean and Standard Deviation (Std) characterize distribution

Comparing distributions

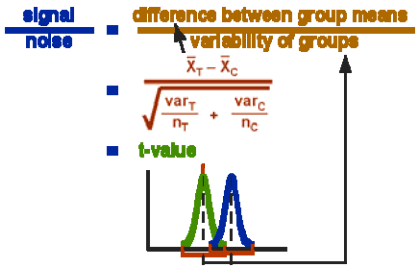


Are the expression values of ade1 different from those of ssn6?

Quantifying Significance



Student's T-test



***t*-statistic: difference between means in units of average error**

Significance can be translated into *p*-value (probability) assuming normal distributions

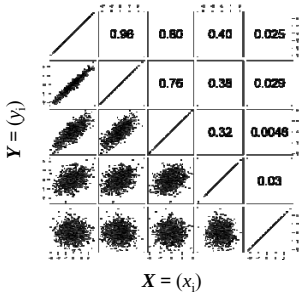
http://www.physics.csbsju.edu/stats/t-test.html

History: W. S. Gossett [1876-1937]

- The t-test was developed by W. S. Gossett, a statistician employed at the Guinness brewery. However, because the brewery did not allow employees to publish their research, Gossett's work on the t-test appears under the name "Student" (and the t-test is sometimes referred to as "Student's t-test.") Gossett was a chemist and was responsible for developing procedures for ensuring the similarity of batches of Guinness. The t-test was developed as a way of measuring how closely the yeast content of a particular batch of beer corresponded to the brewery's standard.

http://ccnmtl.columbia.edu/projects/qmss/t_about.html

Pearson correlations (Graphic)



Each dot is a pair (x_i, y_i)

Comparing profiles *X* and *Y* (not distributions!):

What is the tendency that high/low values in ***X*** match high/low values in ***Y***?

http://davidmlane.com/hyperstat/A34739.html

Pearson correlations: Formulae

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

(complicated version)

$$r = \frac{\sum z_x z_y}{N}$$

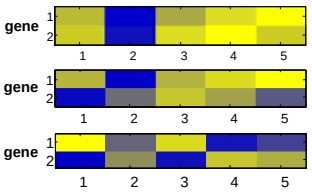
(simple version using *z*-scores)

Pearson correlations: Intuition

$$r = \frac{\sum z_x z_y}{N}$$

Similarity according to *all* conditions ("Democratic vote")

← conditions →



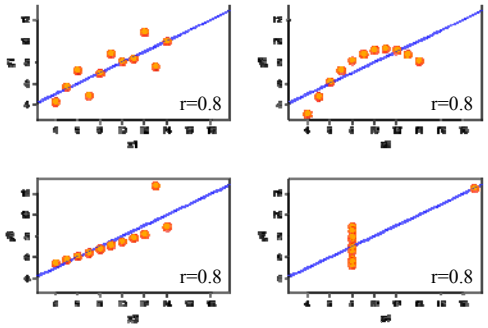
$r_{12} \sim 1$

$r_{12} \sim 0$

$r_{12} \sim -1$

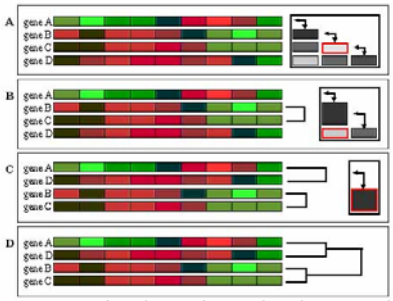
Clustering-coefficient

Pearson correlations: Caution!



High correlation does not necessarily mean co-linearity!

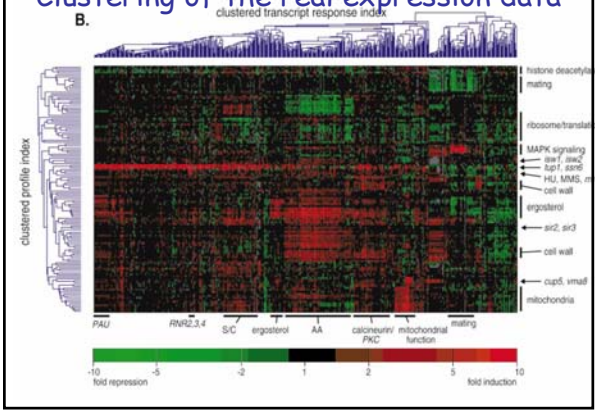
(Hierarchical Agglomerative) Clustering



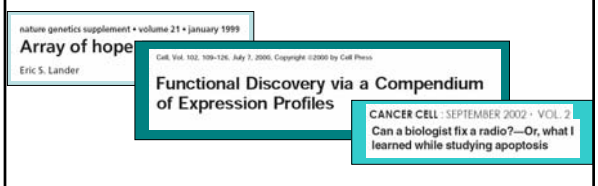
Join most correlated samples and replace correlations to remaining samples by average, then iterate …

http://gepas.bioinfo.cipf.es/cgi-bin/tutoX?c=clustering/clustering.config

Clustering of the real expression data



Further Reading

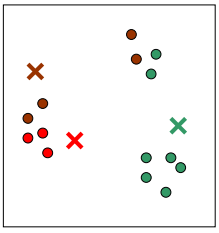


K-means Clustering

"guess" k=3 (# of clusters)

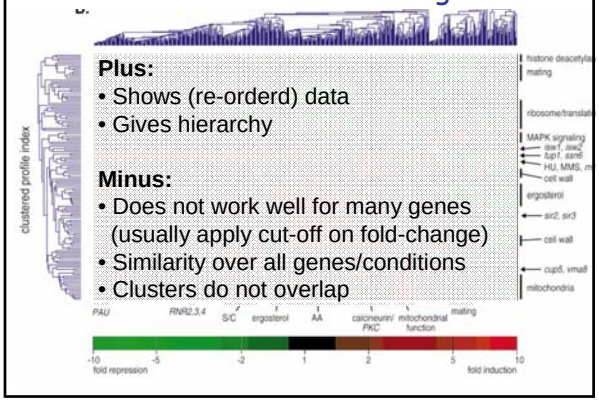
1. Start with random positions of *centroids* (✕)

2. Assign each data point to closest centroid



http://en.wikipedia.org/wiki/K-means_algorithm

Hierachical Clustering



Plus:
- Shows (re-orderd) data
- Gives hierarchy

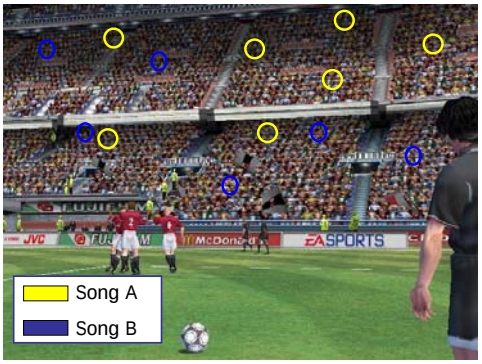
Minus:
- Does not work well for many genes (usually apply cut-off on fold-change)
- Similarity over all genes/conditions
- Clusters do not overlap

Overview of "modular" analysis tools

- Cheng Y and Church GM. ***Biclustering of expression data***. (Proc Int Conf Intell Syst Mol Biol. 2000;8:93-103)
- Getz G, Levine E, Domany E. ***Coupled two-way clustering analysis of gene microarray data***. (Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12079-84)
- Tanay A, Sharan R, Kupiec M, Shamir R. ***Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data***. (Proc Natl Acad Sci U S A. 2004 Mar 2;101(9):2981-6)
- Sheng Q, Moreau Y, De Moor B. ***Biclustering microarray data by Gibbs sampling***. (Bioinformatics. 2003 Oct;19 Suppl 2:ii196-205)
- Gasch AP and Eisen MB. ***Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering***. (Genome Biol. 2002 Oct 10;3(11):RESEARCH0059)
- Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. ***'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns***. (Genome Biol. 2000;1(2):RESEARCH0003.)

… and many more! http://serverdgm.unil.ch/bergmann/Publications/review.pdf

How to "hear" the relevant genes?



Coupled two-way Clustering

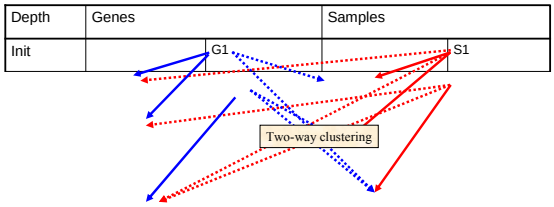
Coupled two-way clustering analysis of gene microarray data

Gad Getz, Erel Levine, and Eytan Domany*

PNAS | October 24, 2000 | vol. 97 | no. 22 | 12079–12084

We present a coupled two-way clustering approach to gene microarray data analysis. The main idea is to identify subsets of the genes and samples, such that when one of these is used to cluster the other, stable and significant partitions emerge. The search for such subsets is a computationally complex task. We present an algorithm, based on iterative clustering, that performs such a search. This analysis is especially suitable for gene microarray data, where the contributions of a variety of biological mechanisms to the gene expression levels are entangled in a large body of experimental data. The method was applied to two gene microar-

Inside CTWC: Iterations



One example in more detail:

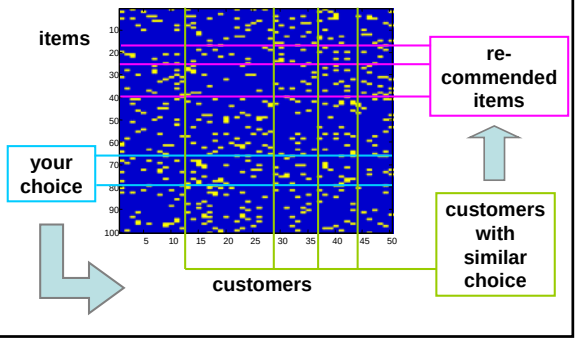
The (Iterative) Signature Algorithm:

- No need for correlations!
- decomposes data into "transcription modules"
  - integrates external information
  - allows for interspecies comparative analysis

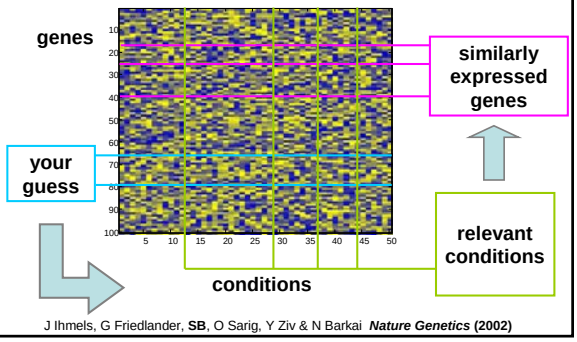
J Ihmels, G Friedlander, **SB**, O Sarig, Y Ziv & N Barkai ***Nature Genetics* (2002)**

Trip to the "Amazon":

How to find related items?

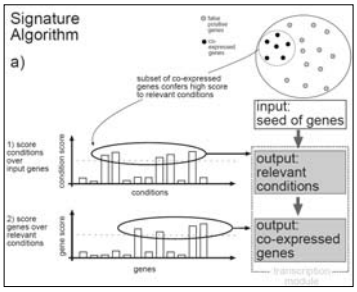


How to find related genes?



J Ihmels, G Friedlander, **SB**, O Sarig, Y Ziv & N Barkai ***Nature Genetics* (2002)**

Signature Algorithm: Score definitions



$$G_I$$

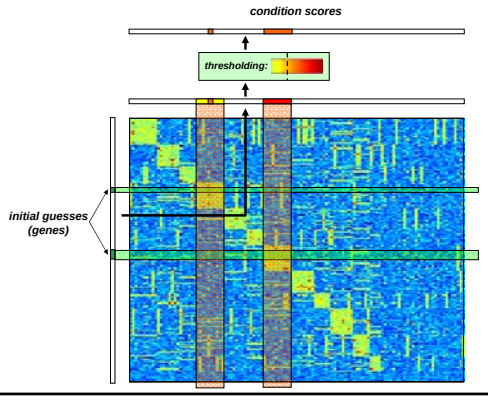
$$s_c = \left\langle E_G^{gc} \right\rangle_{g \in G_I}$$

$$S_C = \{c \in C : |s_c - \langle s_c \rangle_{c \in C}| > t_C \sigma_C\}$$

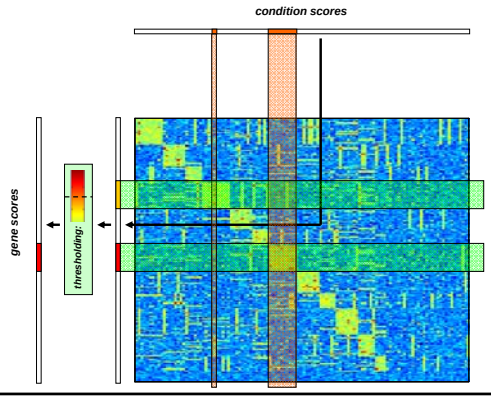
$$s_g = \left\langle s_c E_C^{gc} \right\rangle_{c \in S_c}$$

$$S_G = \{g \in G : s_g - \langle s_g \rangle_{g \in G} > t_G \sigma_G\}$$

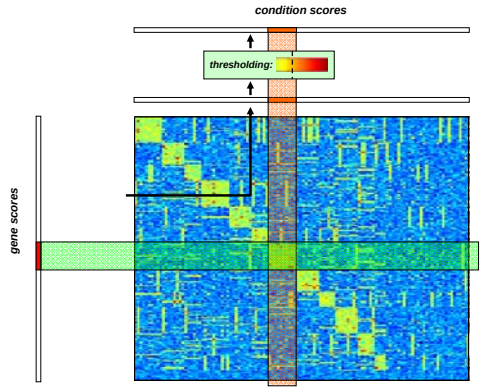
How to find related genes? Scores and thresholds!



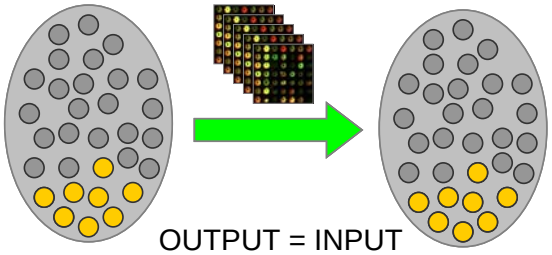
How to find related genes? Scores and thresholds!



How to find related genes? Scores and thresholds!



Iterative Signature Algorithm

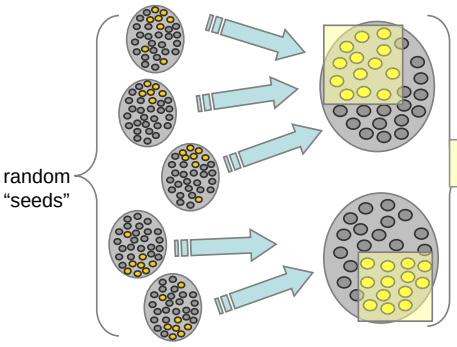


OUTPUT = INPUT

"Transcription Module"

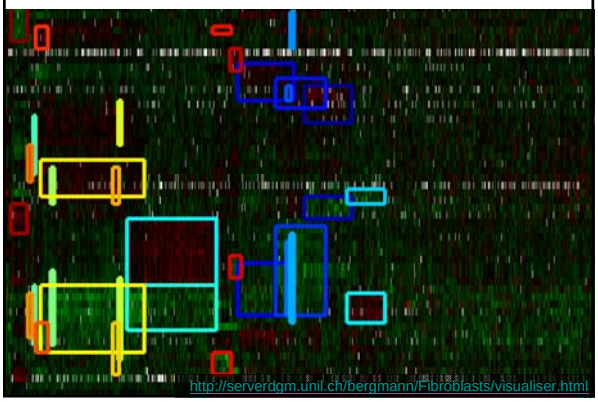
SB, J Ihmels & N Barkai ***Physical Review E (2003)***

Identification of transcription modules using many random "seeds"



Independent identification: *Modules may overlap!*

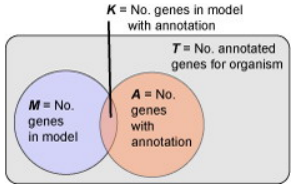
New Tools: Module Visualization



http://serverdgm.unil.ch/bergmann/Fibroblasts/visualiser.html

Gene enrichment analysis

The ***hypergeometric*** distribution *f(M,A,K,T)* gives the probability that *K* out of *A* genes with a particular annotation match with a module having *M* genes if there are *T* genes in total.



http://en.wikipedia.org/wiki/Hypergeometric_distribution

Decomposing expression data into annotated transcriptional modules

identified >100 transcriptional modules in yeast:

high functional consistency!

many *functional links* "waiting" to be verified experimentally

J Ihmels, **SB** & N Barkai ***Bioinformatics 2005***

Higher-order structure