Differentially Coexpressed Genes

Dennis Kostka & Rainer Spang

Courses in Practical DNA Microarray Analysis

Nationales Genomforschungsnetz

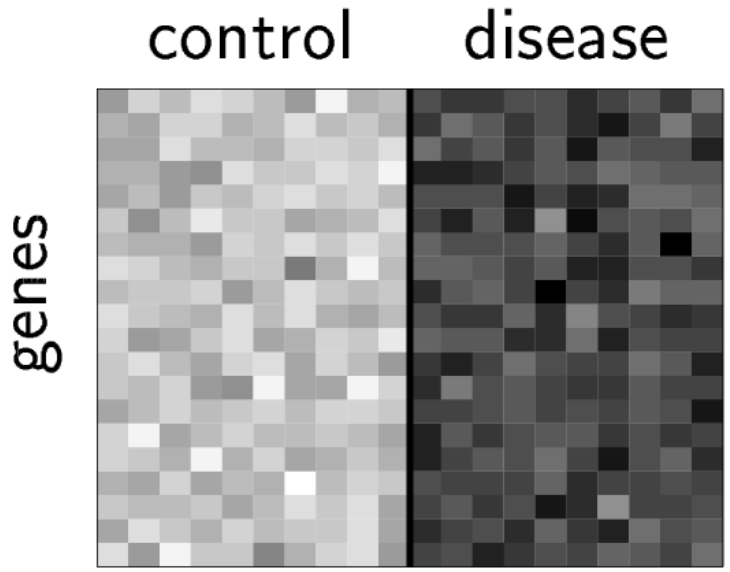
This is joint work with

- Dennis Kostka -

Differential Expression

Molecular disease mechanisms constitute abnormalities in the coregulation of genes

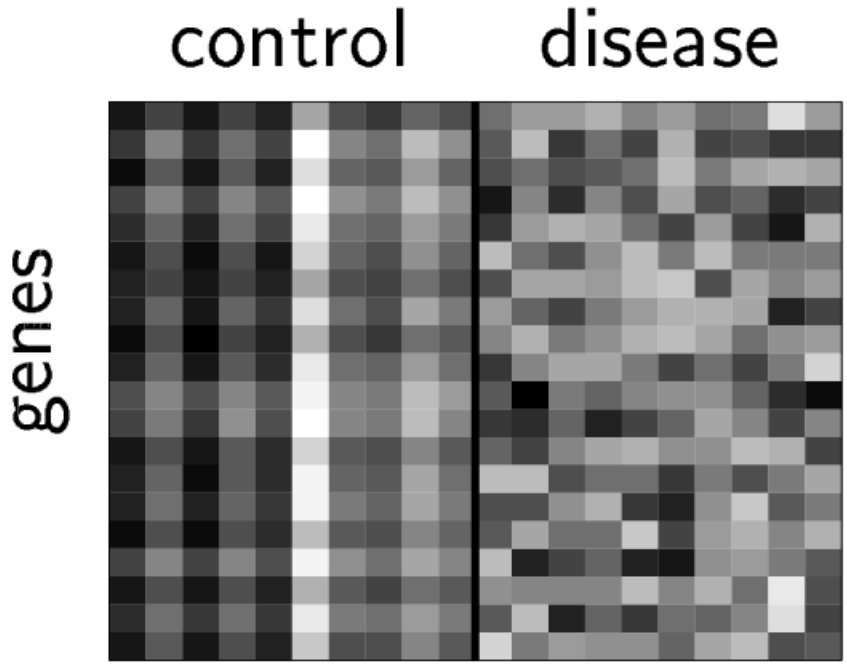
Alterations in gene regulation typically result in up and down regulation of genes



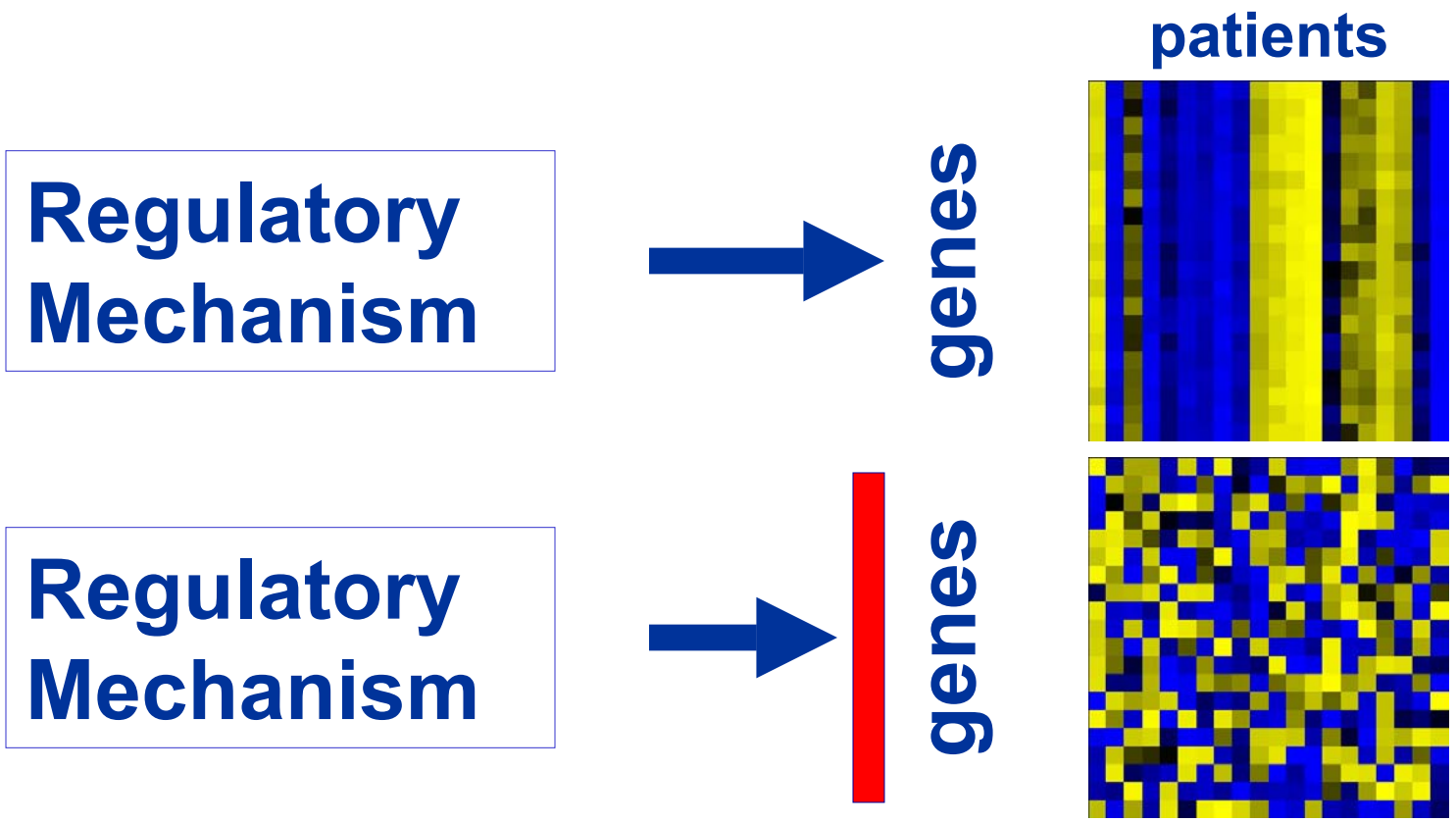
BUT

Not all changes in coregulation show up as patterns of up or down regulated genes

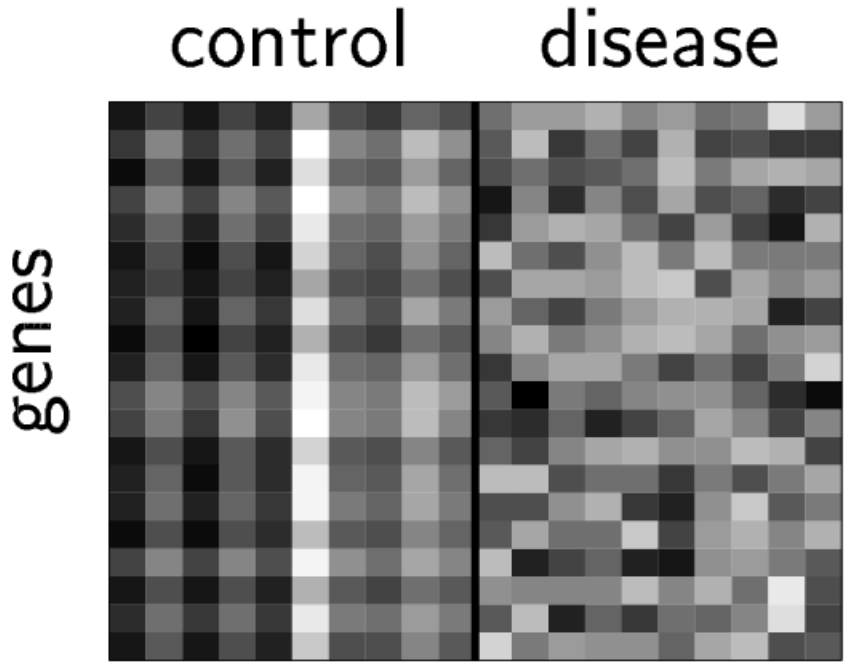
Differential Coexpression



A Regulatory Mechanism is Breaking Down



Questions



How can we find differential coexpression patterns ?

Do these pattern exist in real data ?

Are they biologically meaningful ?

Did we really need a new method to find them ?

How can we find differential coexpression patterns ?

How did we find differential expression patterns ?

By screening one gene after the other

Problem:

Differential expression is a property of a single gene, differential coexpression is a property of a set of genes

... we need to screen all subsets of genes on the chip

... this is hard and can only be done heuristically

The problem of finding differential coexpression is mainly a problem of efficient search

A: Decide on a score for differential coexpression

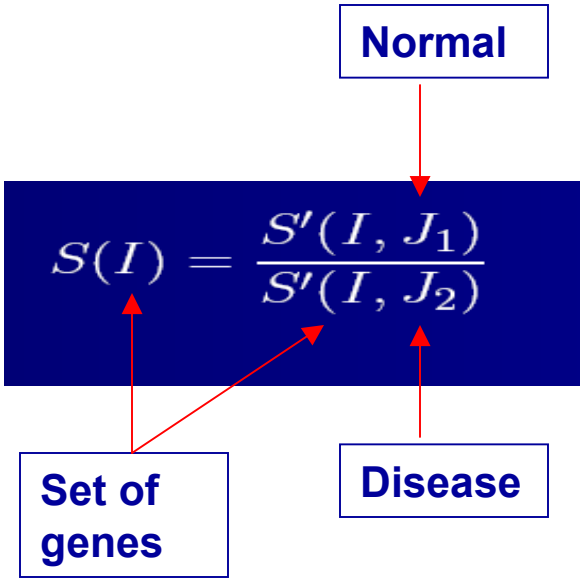
B: Greedy stochastic downhill search

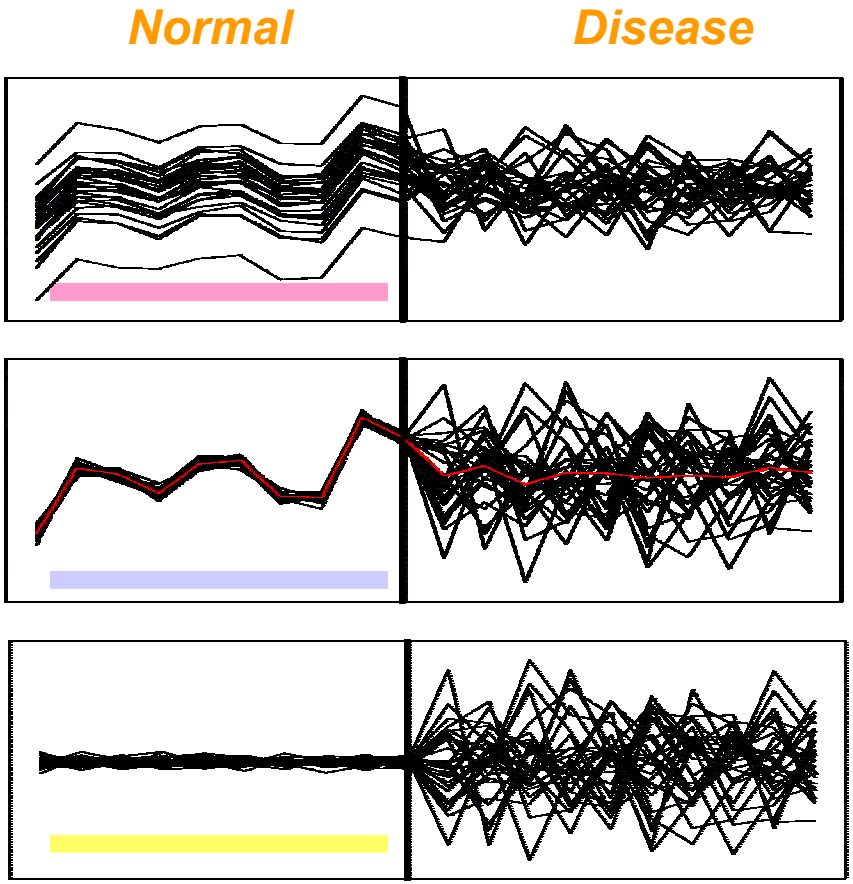
1. Choose a random set of genes and score it
2. Randomly select a neighboring set (no more than k different genes) and calculate its score
3. If the score of the new set is lower, change to the new set otherwise keep the old set
4. Iterate until you find a local minimum of the score

The computational costs for scoring a candidate set are critical for the practicality of the algorithm

The algorithm is stochastic. Restarting it several times can result in different local minima corresponding to different differential coregulation patterns

A score for differential coexpression of several genes





$$S'(I,J) = \frac{1}{(|I|-1)(|J|-1)} \sum_{I,J} (a_{ij} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet})^2$$

The trick *(borrowed from Cheng and Church):*

Calculating S is computationally expensive, but it is very cheap to decide whether adding or replacing genes leads to a higher or lower score, much cheaper than for the correlation coefficient

Neighborhood structure:
Neighboring sets differ only by a single gene.

Given a group of genes I we wish to exclude gene k :

$$S(I) \propto \frac{\text{mean}_{I,G1}(\text{res})}{\text{mean}_{I,G2}(\text{res})} = \frac{A_k^{(1)} + B_k^{(1)}}{A_k^{(2)} + B_k^{(2)}}$$

and modulo refitting of the parameters:

$$S(I \setminus k) < S(I) \quad \text{iff} \quad B_k^{(1)}/B_k^{(2)} > S(I)$$

Some Fine-tuning

We include / exclude a β–fraction of the genes that meet the criterion for a reduced score

To tune the size of the finally found gene sets we introduce a tuning parameter α.

The final criterion for including or excluding a gene now reads:

$$C_k(\alpha) = B_k^{(1)} / B_k^{(2)} \pm \{\alpha \cdot S(I) + (1-\alpha) \cdot 1/|I|\} > 0$$

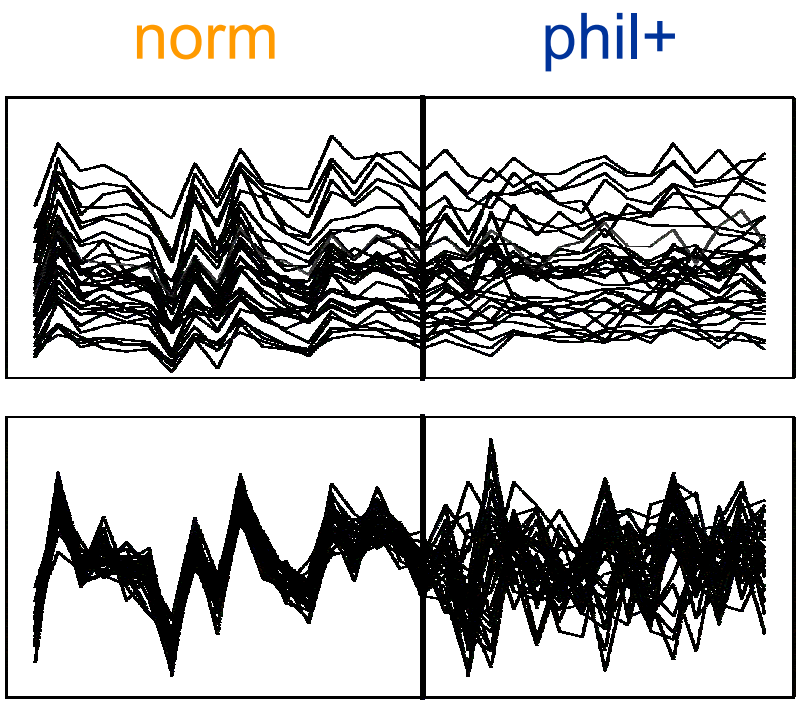
Do these pattern exist in real data ?

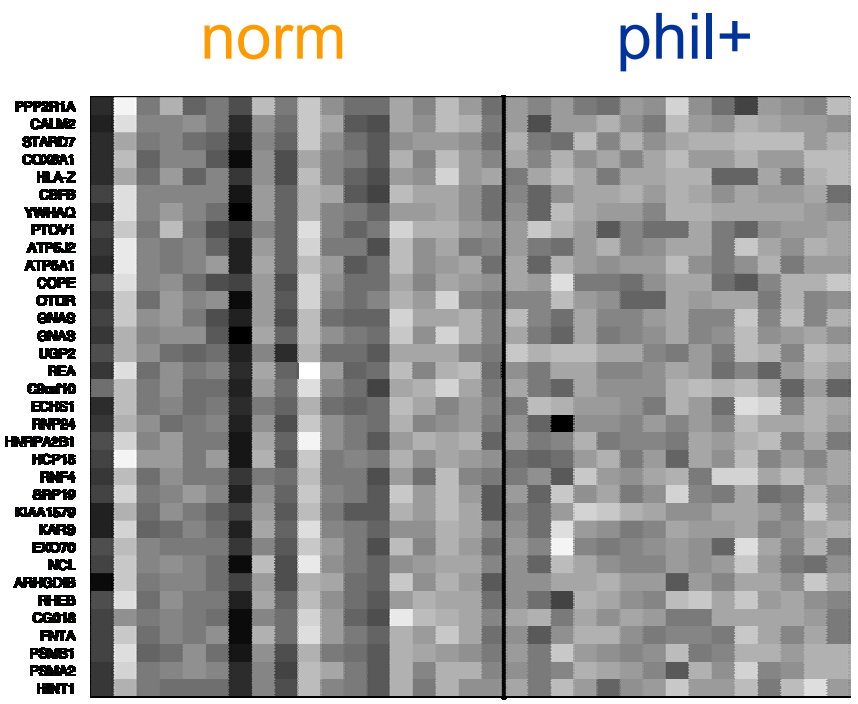
Acute Lymphoblastic Leukemia

- **About 1/3 of all pediatric cancers**
- **Different cytogenetic risk groups (e.g. 70% overall cure rate vs. 30% in phil+)**
- **We compared cytogenetically normal children to those with the phil+ translocation**

*Yeoh EJ, RossMEet al. (2002) Classication, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression proling, Cancer Cell, **1(2)**, 133-43.*

Differential coexpression in phil+ leukemia





Are the patterns biologically meaningful ?

Pattern 1 (34 genes)				
HINT1 (3094)	PSMA2 (5683)	PSMB1 (5689)	FNTA (2339)	CG018 (90634)
RHEB (6009)	ARHGDIB (397)	NCL (4691)	EXO70 (23265	KARS (3735)
KIAA1579 (55225)	SRP19 (6728)	RNF4 (6047)	HCP15 (157317)	HNRPA2B1 (3181)
RNP24 (10959)	ECHS1 (1892)	C9orf10 (23196)	REA (11331)	UGP2 (7360)
GNAS (2778)	GNAS (2778)	OTOR (56914)	COPE (11316)	ATP5A1 (498)
ATP5J2 (9551)	PTOV1 (53635)	YWHAQ (10971)	CBFB (865)	HLA-Z (267017)
COX6A1 (1337)	STARD7 (56910)	CALM2 (805)	PPP2R1A (5518)	

Pattern 2 (21 genes)				
PSMB2 (5690)	UBC (7316)	ARHA (387)	PSMA4 (5685)	ARPC2 (10109)
ACTB (60)	ENG (2022)	GG2-1 (25816)	ACTR3 (10096)	PITPNB (23760)
CAPNS1 (826)	PFN1 (5216)	SF3B2 (10992)	GNAI2 (2771)	ARHA (387)
PPP1CC (5501)	PSMC5 (5705)	ARPC5 (10092)	MRLC2 (103910)	AES (166)
GDI2 (2665)				

→ **Proteasome-Ubiquitin Pathway (for several cancers including CLL, inhibition can induce apoptosis)**

→ **Involved in degradation of p27 (prognostic factor in B-cell lymphoma)**

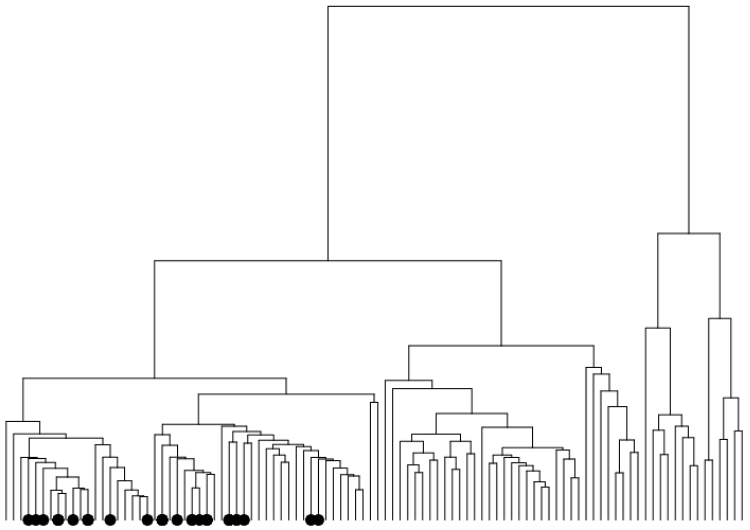
Most others: protein synthesis, protein transport, protein degradation

Did we really need a new method to find the patterns?

Screening for differential expression:

The genes in the two patterns have ranks between 106-6114

Hierarchical clustering:



Thank You