Structured Analysis of Microarrays & Differential Coexpression

Claudio Lottaz, Dennis Kostka

& Rainer Spang

Courses in Practical DNA Microarray Analysis

Structured Analysis of Microarrays

Tumor Diagnosis/Prognosis with expression profiles

Data:

Tens of thousands of genes

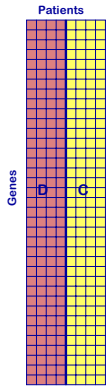
Tens to hundreds of patients

Patients are labeled

E.g. Disease (D) / Control (C)

Problem:

Predict the label of a new patient given his/her expression profile



Statistical Context

The high dimensionality of the expression data causes *overfitting problems*, when fitting multivariate models to the data

The solution are additional constraints in the models:

Regularized Multivariate Models

Classification Algorithms

What are these constraints:

- A small maximal number of genes allowed in the model *(Variable Selection, Sparse Models)*
- Likelihood penalties *(Ridge Regression)*
- Informative priors *(Bayesian Regression)*
- Large Margins *(Support Vector Machines)*

The models identify genes that are informative for the class distinction. We call such a model a *molecular signature*

Typical frustrations

Falsely predicted patients:

The predictive model seems to work very well for 80% of the patients and for the remaining 20% of patients you get wrong predictions although the pattern in the expression profiles seem to be quite obvious.

Genes that make no sense in this context:

If you look at the list of genes that drive the model the list does not tell you a unique biological story. You observe some genes that are expected to be there and many more genes that look like being randomly collected.

Implicit assumptions of standard approaches

The models aim to identify characteristic expression pattern for the whole patient groups (global molecular signatures)

The patient groups are seen as molecular homogenous groups

This is often not the case

There might be many different molecular phenotypes with a poor treatment outcome

Genes are anonymous variables: $x_1,...x_n$

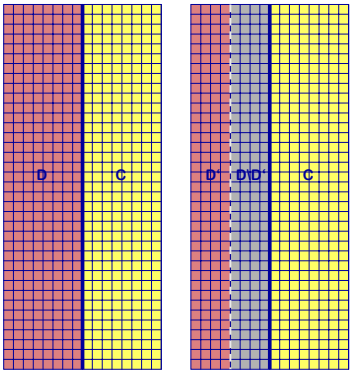
This ignores that for many genes we already know a lot about their function and the biological processes they are involved in

Our approach:

1. Sub-class finding instead of global class prediction
2. Use of functional annotations of genes

Molecular Symptoms

Global class prediction vs. Subclass finding



Global class prediction:

Find a molecular signature that separates D from C and generalizes to new patients

Subclass finding:

Find a Subclass D' ⊂ D and a molecular signature that separates D' from C

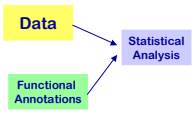
We call this signature a *molecular symptom* associated to D

Exploiting functional annotations of genes

A posteriori use of functional annotations

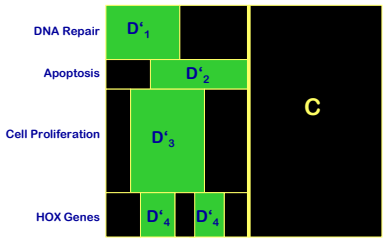
A priori use of functional annotations

(suggested here)

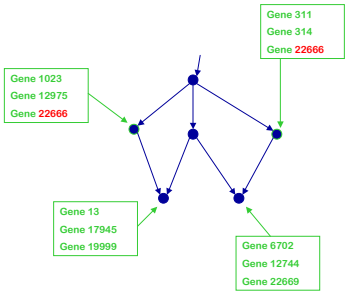


What are we looking for?

We are looking for groups of genes involved in a common biological process with a clearly changed expression profile in part of patients having a certain disease but not in all patients with this disease



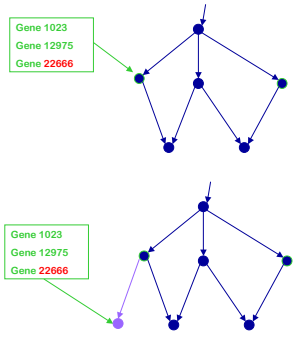
The Annotation



Genes are annotated to both leave and inner nodes

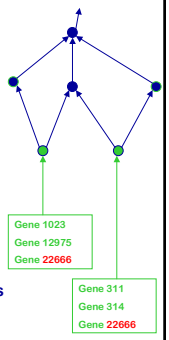
Genes can have multiple annotations

Augmented Ontology



Structured Analysis of Microarrays (StAM)

Modular Grid of three components



1. *Classification in Leave Nodes:*

   Train a classifier for every leave node using some regularized multivariate model …this gives us local signatures in the leave nodes
2. *Diagnosis Propagation:*

   Combine the diagnostic results in the children to a diagnosis for an inner node … this gives us signatures in the inner nodes
3. *Regularization:*

   Get rid of noisy but non informative branches

1. Classification in the leave nodes by Shrunken centroid classification: PAM

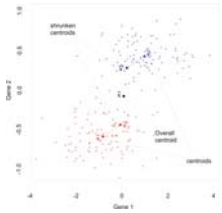
(Tibshirani et al 2002)

DLDA-like Discrimination

Discriminant function via the distance to the shrunken class centroids d(C), d(D)

Regularization:

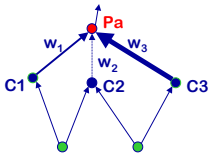
Variable selection via centroid shrinkage



Class Probabilities:

$$P[\text{Group} = C] = \frac{e^{-\frac{1}{2}d(C)}}{e^{-\frac{1}{2}d(C)} + e^{-\frac{1}{2}d(D)}}$$

2. Propagation of Diagnosis by weighted averages



$$P_{Pa} = \sum_{C_i} w_{C_i} p_{C_i}$$

Weights are proportional to CV-Performance measured by deviance

$$d_{C_i} = -\frac{2}{N} \sum_{\text{Sample}_x} I_{\text{Group}=D} \log p_{C_i}(x) + I_{\text{Group}=C} \log(1 - p_{C_i}(x))$$

$$w_{C_i} = d_{C_i} (\textstyle\sum_j d_{C_j})^{-1}$$

3. Regularization by graph shrinkage

To get rid of uninformative branches of the Gene Ontology, we shrink the weights in the progression step by a constant Δ

$$w_i' = (w_i - \Delta)^+$$

Δ is chosen by crossvalidation

Expression data from a leukemia study

Yeoh et al., Cancer Cell 2002

Study on acute lymphoblastic leukaemia (ALL) 327 *patients*
12625 *genes* (Affymetrix HG-U95Av2)

My focus in this talk:
MLL – ( ) vs. Others

Objective:
Diagnosis of cytogenetic subtypes of ALLs:

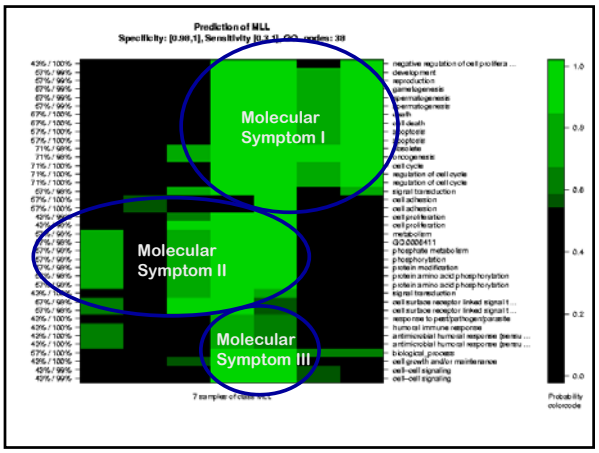
20 MLL - ( )
27 E2A-PBX1
15 BCR-ABL
79 TEL-AML1
87 Hyperdiploid
7 Hypodiploid
29 Pseudodiploid
18 nomal (B-cell ALL)
43 T-ALL

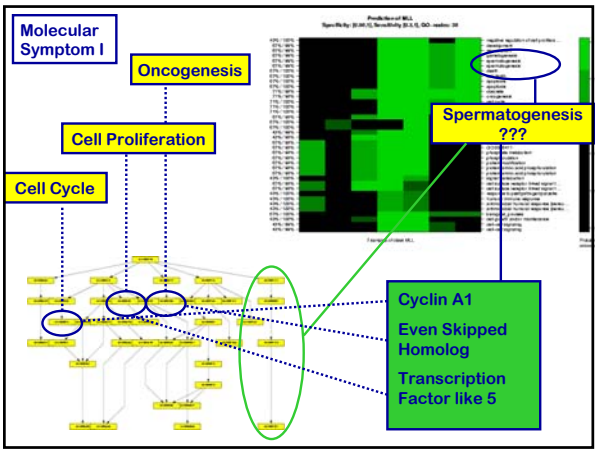
Selectivity/ Specificity

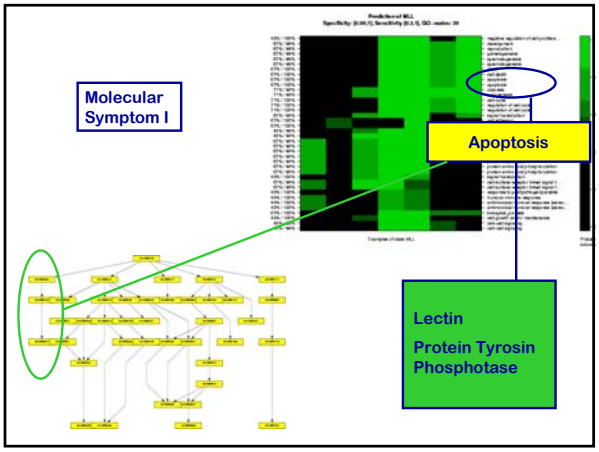
Color: MLL-Score

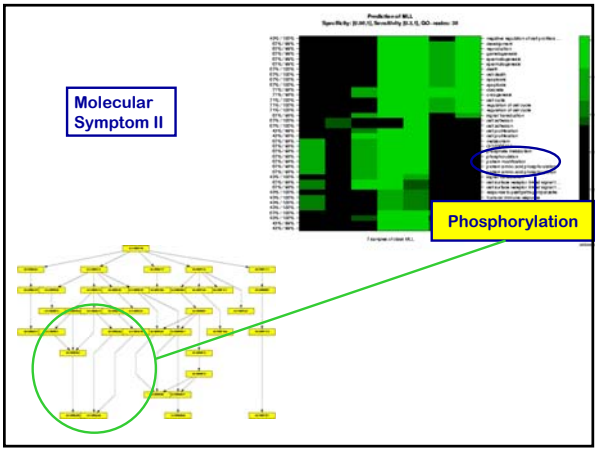
Columns: 7 MLL-Patients (Testset)

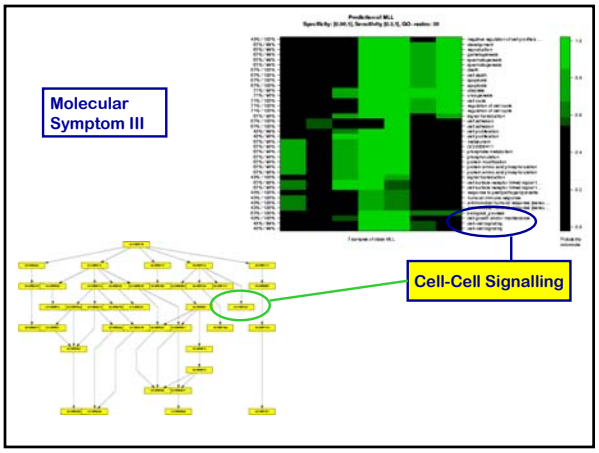
Rows: GO-Nodes

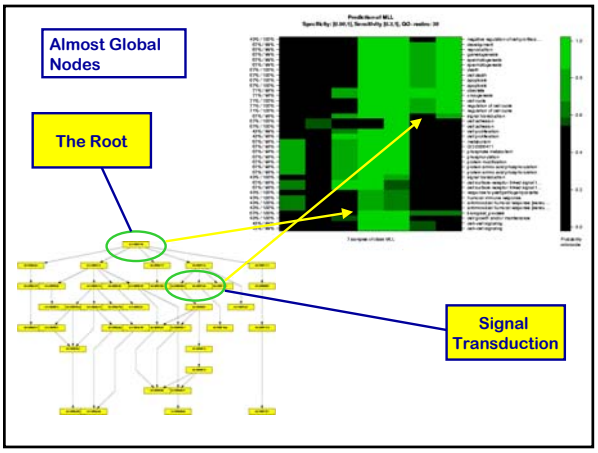

Molecular Symptom I
Molecular Symptom II
Molecular Symptom III


Molecular Symptom I
Oncogenesis
Cell Proliferation
Cell Cycle
Spermatogenesis ???
Cyclin A1
Even Skipped Homolog
Transcription Factor like 5


Molecular Symptom I
Apoptosis
Lectin
Protein Tyrosin Phosphotase


Molecular Symptom II
Phosphorylation


Molecular Symptom III
Cell-Cell Signalling

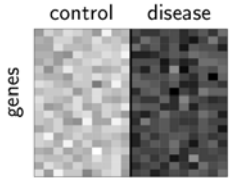

Almost Global Nodes
The Root
Signal Transduction

Differential Coexpression

Differential Expression

Molecular disease mechanisms constitute abnormalities in the coregulation of genes

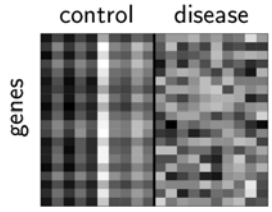
Alterations in gene regulation typically result in up and down regulation of genes



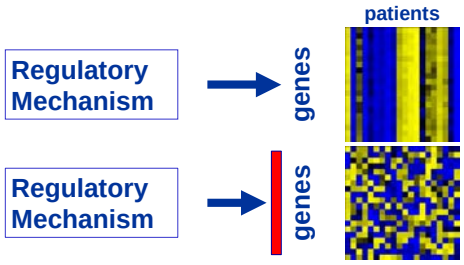
BUT

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

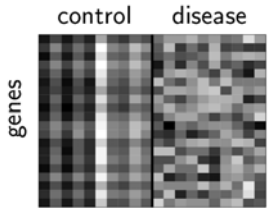
Differential Coexpression



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

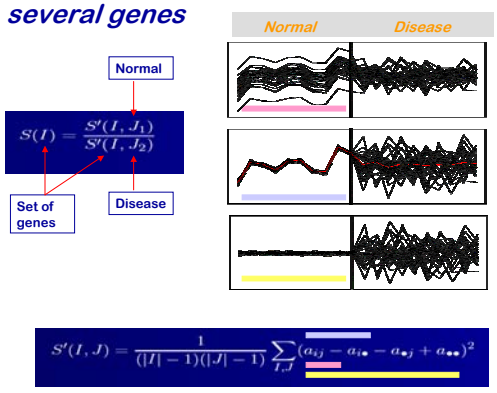
B: Greedy stochastic downhill search

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:

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

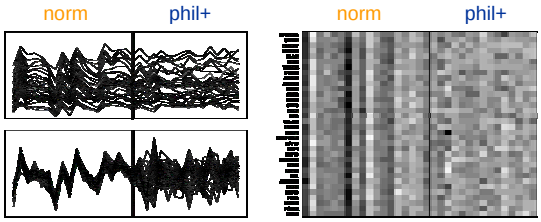
Do these pattern exist in real data ?

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)

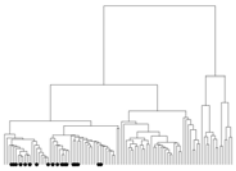
Most others: protein synthesis, protein transport, protein degradation

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