

Finding disease specific alterations in the coexpression of genes

Dennis Kostka

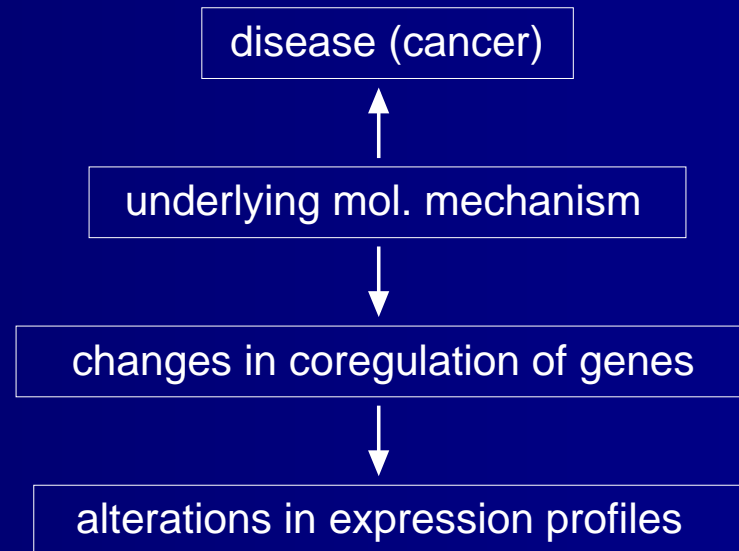
`dennis.kostka@molgen.mpg.de`

Max Planck Institute for Molecular Genetics

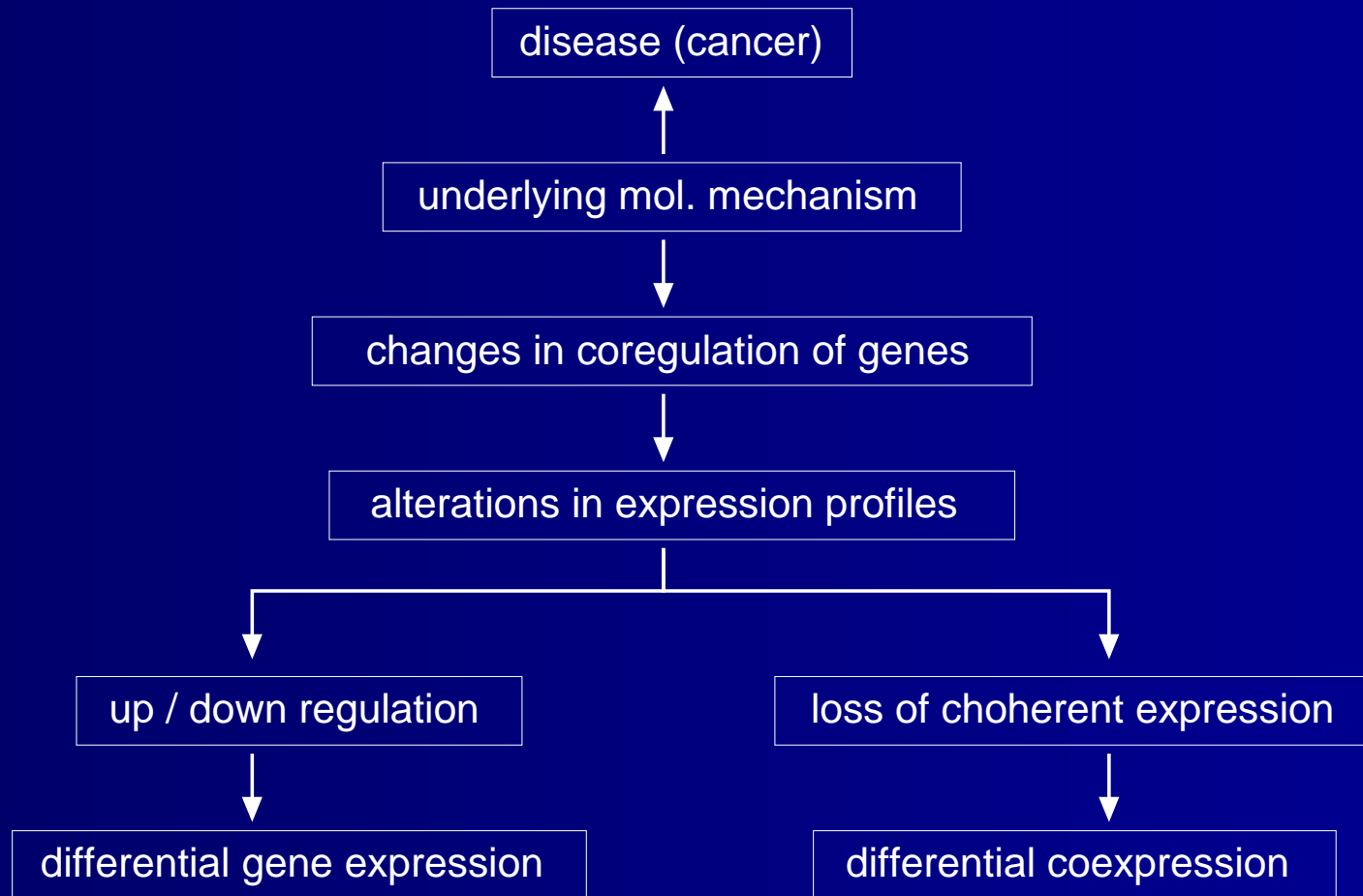
differential coexpression

disease (cancer)

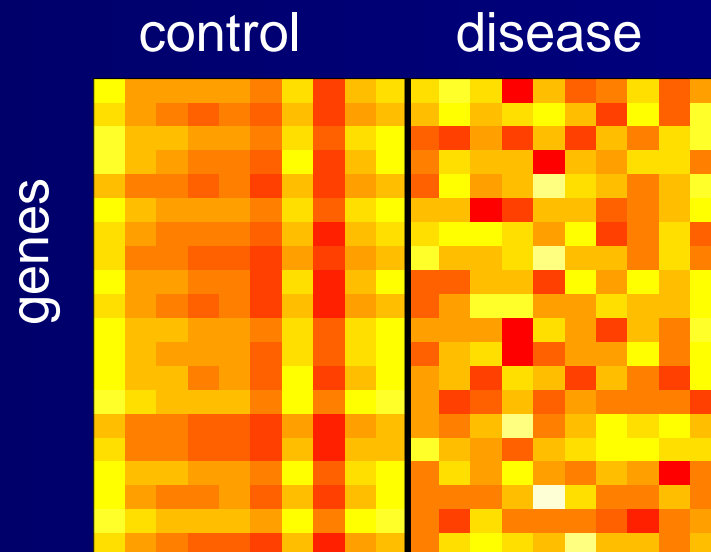
differential coexpression



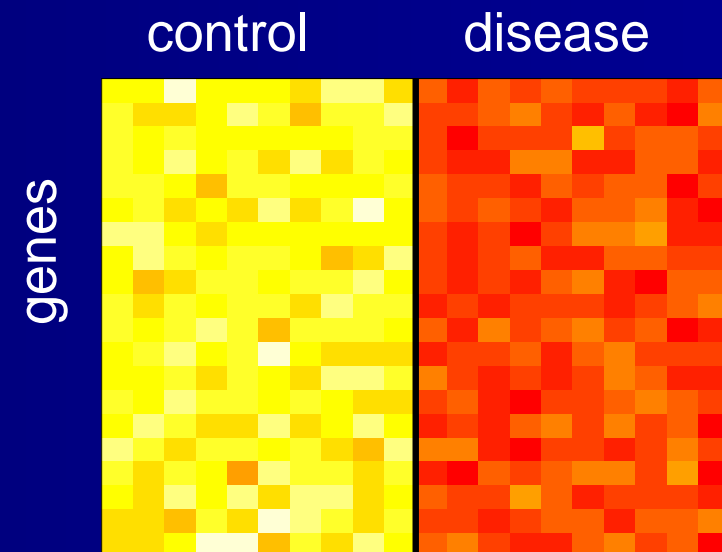
differential coexpression



differential coexpression



(a) diff. coexpression



(b) diff. expression

finding coexpression patterns

- differential expression cannot be analyzed gene by gene
- we need to take into account all the possible subsets of genes

finding coexpression patterns

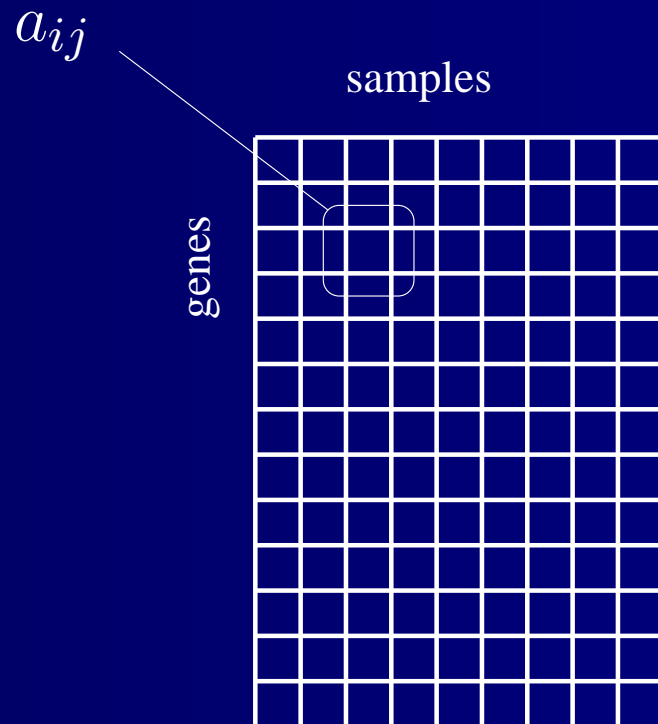
- differential expression cannot be analyzed gene by gene
- we need to take into account all the possible subsets of genes
- therefore we need an efficient screening / scoring method:
 - we propose an additive model for scoring differential coexpression
 - this model allows for a fast search heuristic

outline

- ✓ introduction
 - a search algorithm for differentially coexpressed groups of genes
 - application of the method to
 - simulated data (proof of concept)
 - real data from a clinical study
 - significance and comparison
 - summary

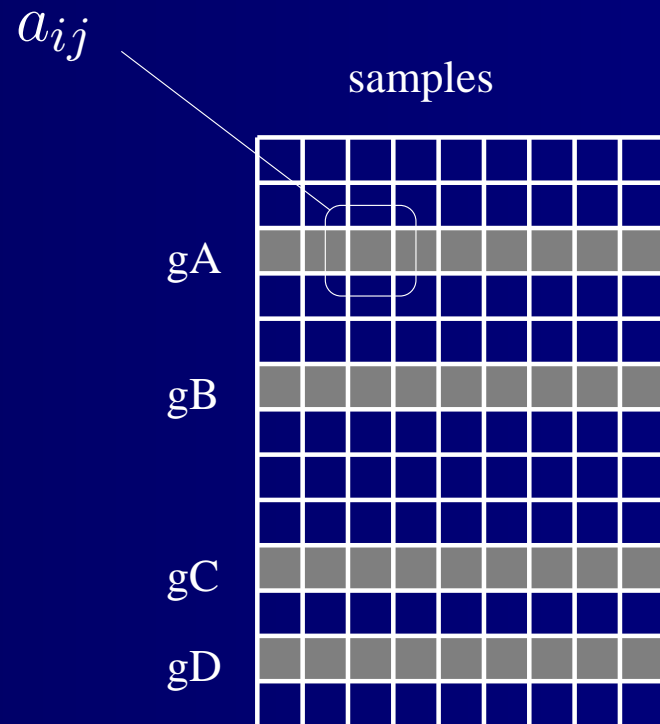
coexpression patterns

Assume $A = \{a_{ij}\}$ is the usual expression matrix:



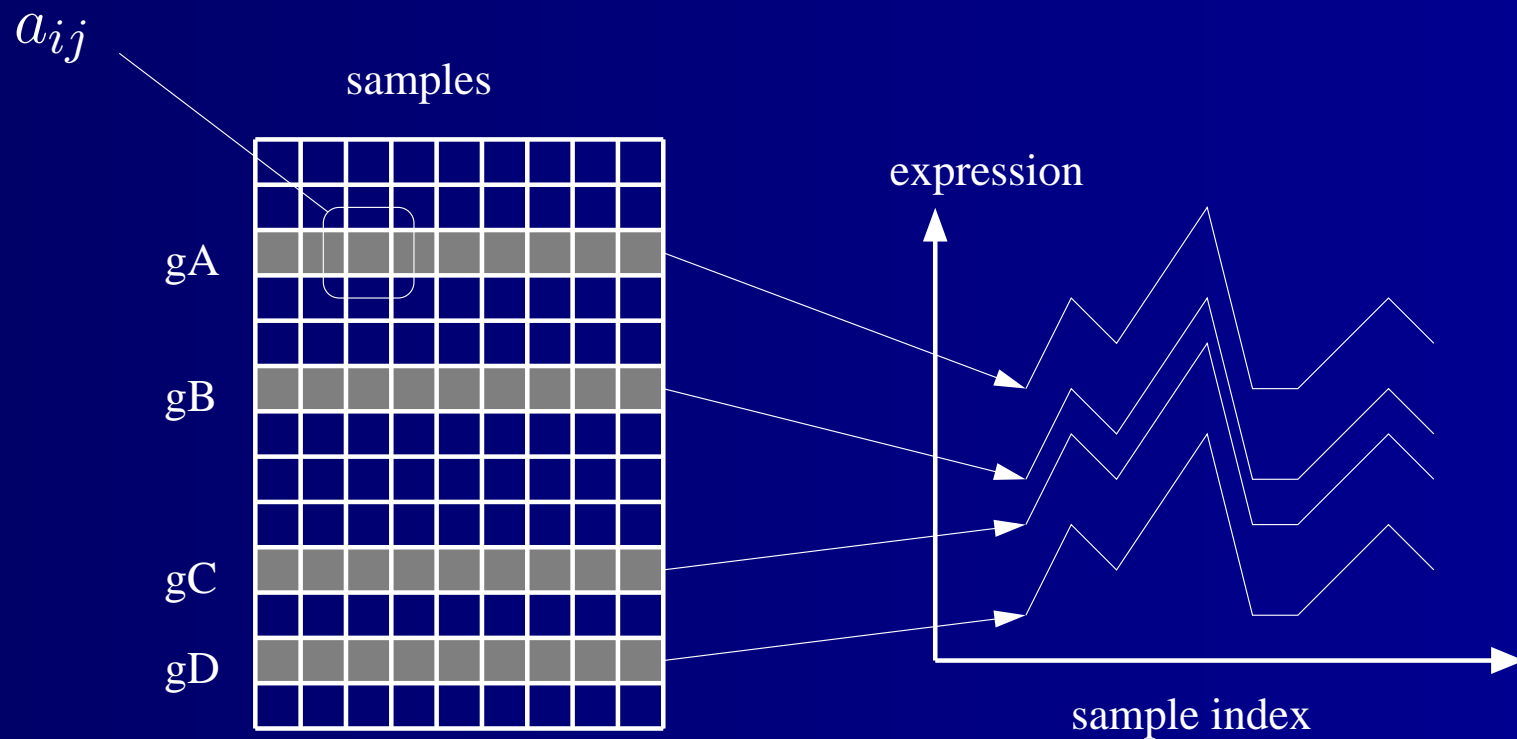
coexpression patterns

Assume $A = \{a_{ij}\}$ is the usual expression matrix:



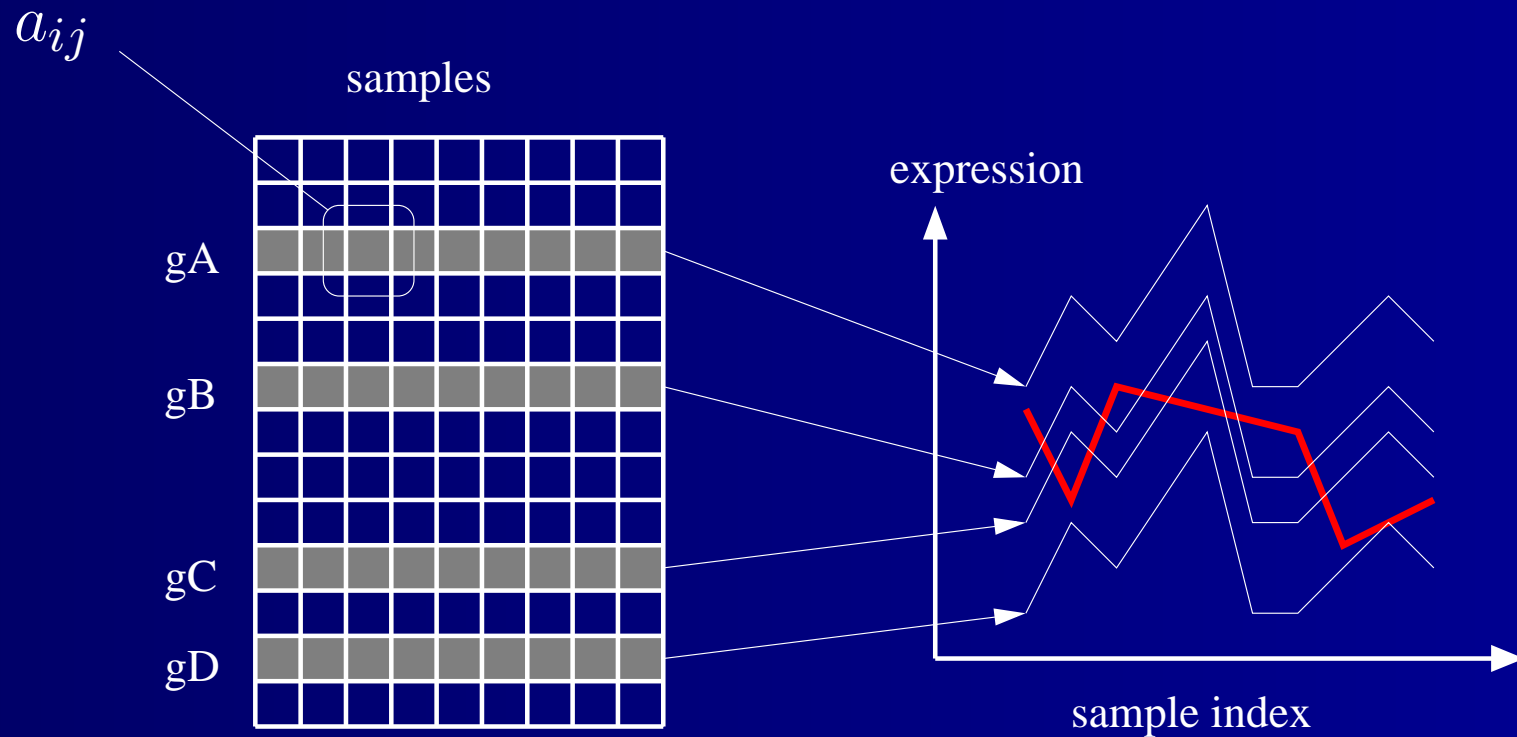
coexpression patterns

Assume $A = \{a_{ij}\}$ is the usual expression matrix:



coexpression patterns

Assume $A = \{a_{ij}\}$ is the usual expression matrix:



additive model

An additive model assumes the expression matrix $\{a_{ij}\}$ composed of *row effects* b_i , *column effects* c_j and of an overall contribution d :

$$a_{ij} = b_i + c_j + d + \epsilon_{ij}$$

additive model

An additive model assumes the expression matrix $\{a_{ij}\}$ composed of *row effects* b_i , *column effects* c_j and of an overall contribution d :

$$a_{ij} = b_i + c_j + d + \epsilon_{ij}$$

We estimate the parameters as follows:

$$\begin{aligned} b_i &\longleftarrow a_{i\bullet} && \text{mean expression of gene } i \\ c_j &\longleftarrow a_{\bullet j} && \text{mean expression of patient } j \\ d &\longleftarrow a_{\bullet\bullet} && \text{overall mean of expression} \end{aligned}$$

scoring coexpression

Score a group of genes I by their mean squared residuals. Say, we focus on $|J|$ patients:

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

scoring coexpression

Score a group of genes I by their mean squared residuals. Say, we focus on $|J|$ patients:

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

In terms of coexpression that means the following:

high coexpression \longrightarrow *low* S'
low coexpression \longrightarrow *high* S'

differential coexpression

To score differential coexpression with respect to two groups (G_1 and G_2) of patients take the quotient of the two coexpression scores:

$$S(I) = \frac{S'(I, J_1)}{S'(I, J_2)}$$

differential coexpression

To score differential coexpression with respect to two groups (G_1 and G_2) of patients take the quotient of the two coexpression scores:

$$S(I) = \frac{S'(I, J_1)}{S'(I, J_2)}$$

$S(I)$ is *low*, if the genes in I are *more* coexpressed in group G_1 than in group G_2 .

This attribute renders a group of genes interesting

search for low scoring gene sets

- We need to identify low scoring sets of genes
- The number of all possible subsets of genes too high for exhaustive search

search for low scoring gene sets

- We need to identify low scoring sets of genes
- The number of all possible subsets of genes too high for exhaustive search
- We resort to a heuristic:
 - take a random starting point
 - greedy stochastic downhill search
 - S lets you efficiently calculate downhill directions

search heuristic

- Neighborhood structure:
Neighboring sets differ only by a single gene.
- Given a group of genes I we wish to exclude gene k :

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

search heuristic

- 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:

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

search heuristic

- Given a random set I we screen $\mathcal{N}(I)$ via the B_k
- We include / exclude a β -fraction of the genes that meet the criterion for a reduced score

search heuristic

- Given a random set I we screen $\mathcal{N}(I)$ via the B_k
- 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:

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

algorithm

```
initialize  $I$  randomly
 $G \leftarrow \emptyset$ 
while counter < maxiter do
  for all  $I' \in \mathcal{N}(I)$  do
     $k \leftarrow I' \Delta I$ 
    if  $C_k(\alpha) > 0$  then
       $G \leftarrow G \cup \{k\}$ 
  if  $G \neq \emptyset$  then
     $n \leftarrow \max\{\lfloor \beta \cdot |G| \rfloor, 1\}$ 
     $g \leftarrow$  uniform sample of size  $n$  from  $G$ 
     $I \leftarrow I \Delta g$ 
  else
    return  $I$ 
  counter  $\leftarrow$  counter + 1
return  $I$ 
```

outline

- ✓ introduction
- ✓ a search algorithm for differentially coexpressed groups of genes
 - application of the method to
 - simulated data (proof of concept)
 - real data from a clinical study
 - significance and comparison
 - summary

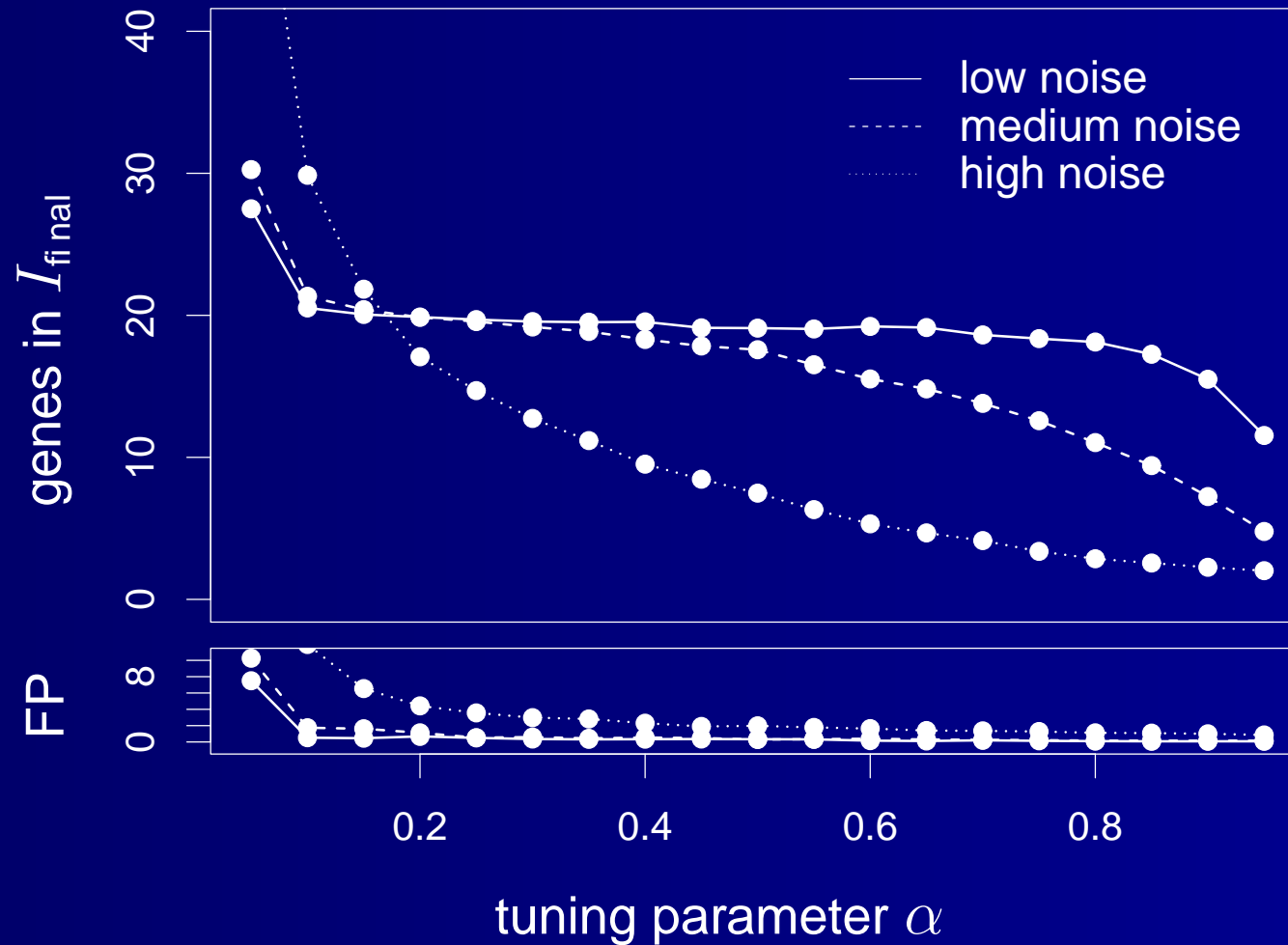
simulated data

- Two groups of 10 samples each
- 120 genes:
 - 20 genes drawn according to the additive model with $\epsilon \sim \mathcal{N}(0, \sigma)$
 - 100 genes drawn independently $\sim \mathcal{N}(0, 1)$

simulated data

- Two groups of 10 samples each
- 120 genes:
 - 20 genes drawn according to the additive model with $\epsilon \sim \mathcal{N}(0, \sigma)$
 - 100 genes drawn independently $\sim \mathcal{N}(0, 1)$
- strength of signal relative to noise
 - $\sigma = 1/10$ low noise
 - $\sigma = 1/4$ medium noise
 - $\sigma = 1$ high noise

simulated data – results



clinical data

- expression levels in bone marrow from children with acute leukemia
- 327 samples divided into subgroups according to characteristic cytogenetic aberrations, including one *normal* group
- we compare all subgroups against the normal group

clinical data

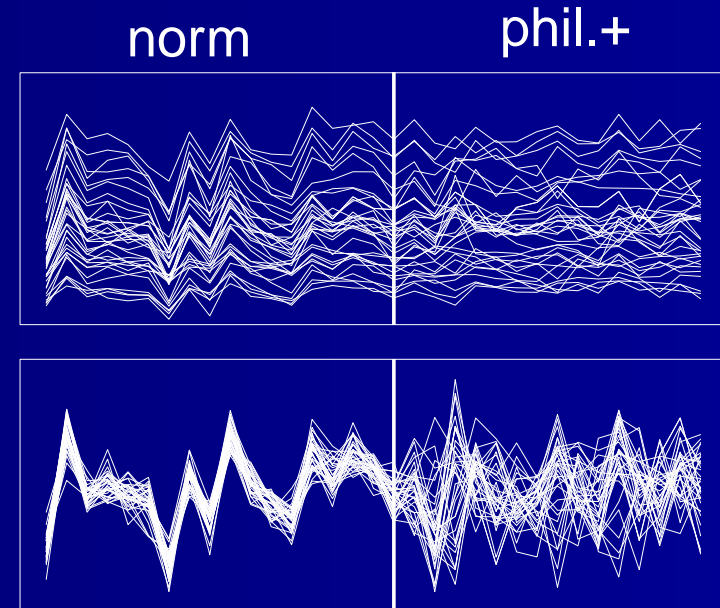
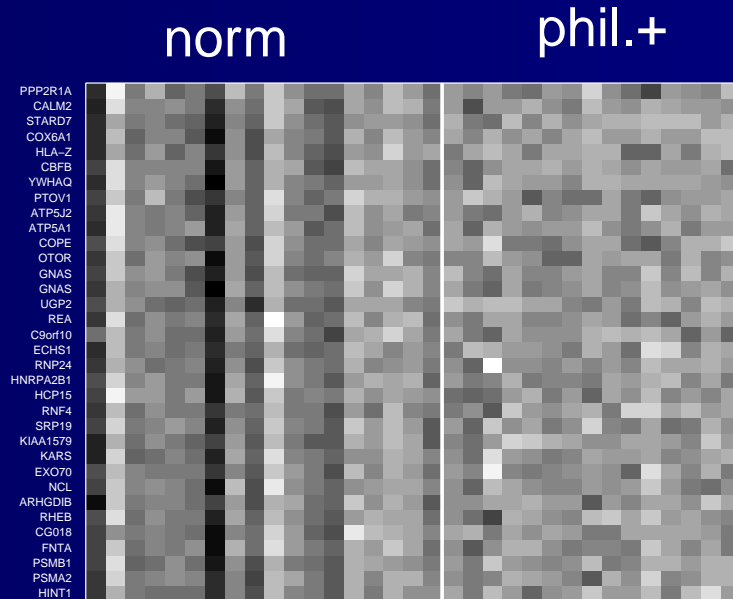
- expression levels in bone marrow from children with acute leukemia
- 327 samples divided into subgroups according to characteristic cytogenetic aberrations, including one *normal* group
- we compare all subgroups against the normal group
- scaling of the groups is necessary, otherwise the algorithm does not discover coexpression patterns

clinical data

- expression levels in bone marrow from children with acute leukemia
- 327 samples divided into subgroups according to characteristic cytogenetic aberrations, including one *normal* group
- we compare all subgroups against the normal group
- scaling of the groups is necessary, otherwise the algorithm does not discover coexpression patterns
- as an example we compare the philadelphia positive (t(9;22)+, BCR-ABL+) to the cytogenetically normal leukemias

clinical data – results

a set of genes displaying differential coexpression:
in the *norm* group the genes display a coherence they
lose in the *phil+* group.



outline

- ✓ introduction
- ✓ a search algorithm for differentially coexpressed groups of genes
- ✓ application of the method to
 - simulated data (proof of concept)
 - real data from a clinical study
- significance and comparison
- summary

significance

- are those patterns artifacts of the high dimensionality of the data?

significance

- are those patterns artifacts of the high dimensionality of the data?
- permutation procedure:
 - assume that coexpressed genes do not exist, i.e. take all genes are independent
 - we sample from this null hypothesis by (group wise) shuffling the expression values for each gene
 - empirical p-value is 0.001 for 1000 draws

significance

- are those patterns artifacts of the high dimensionality of the data?
- permutation procedure:
 - assume that coexpressed genes do not exist, i.e. take all genes are independent
 - we sample from this null hypothesis by (group wise) shuffling the expression values for each gene
 - empirical p-value is 0.001 for 1000 draws
- it's unlikely we are seeing a chance artifact

comparison

- We illustrate that two widespread approaches
 - ranking genes by *t*-scores
 - *hierarchical clustering*would not identify the same gene pattern we found

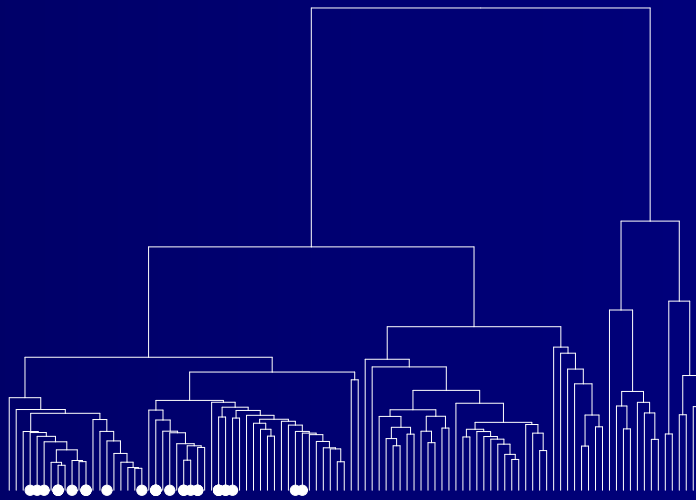
comparison

- We illustrate that two widespread approaches
 - ranking genes by *t*-scores
 - *hierarchical clustering*would not identify the same gene pattern we found
- for the *t*-score, the ranks of 'our' genes are from 106 to 6114, with a mean of 2340.

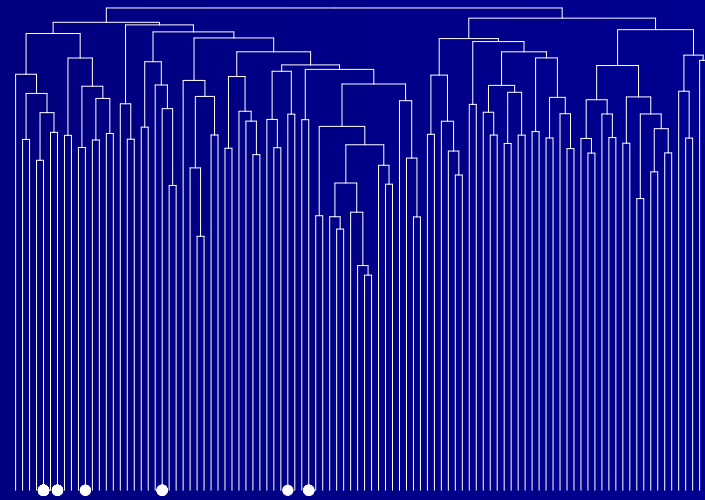
comparison

- We illustrate that two widespread approaches
 - ranking genes by *t*-scores
 - *hierarchical clustering*would not identify the same gene pattern we found
- for the *t*-score, the ranks of 'our' genes are from 106 to 6114, with a mean of 2340.
- for clustering we present two dendrograms.
 - we form 100 representative clusters in a first aggregation step
 - we use average linkage and euclidean distance

comparison



unscaled data



scaled data

outline

- ✓ introduction
- ✓ a search algorithm for differentially coexpressed groups of genes
- ✓ application of the method to
 - simulated data (proof of concept)
 - real data from a clinical study
- ✓ significance and comparison
 - summary

wrap up

we have ...

- addressed the problem of detecting *sets of differentially coexpressed genes*
- described a heuristic algorithm to find them
- demonstrated they exist in real data
- illustrated that our method can be used to complement other exploratory analysis tools

biological meaning ?

- any interpretation of exploratory analyses is speculative
- two most prominent diff. coexpression patterns contain several genes of the proteasome–ubiquitin pathway
- for some cancer types it has been shown that inhibition of proteasome activity results in apoptosis
- further investigation necessary

thank you