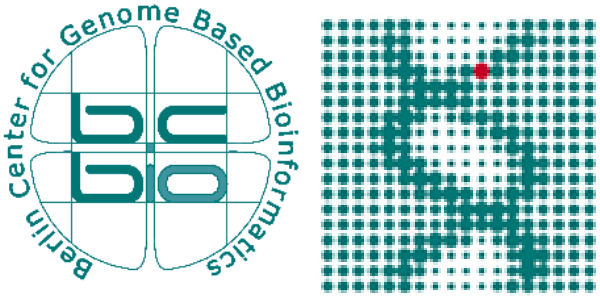
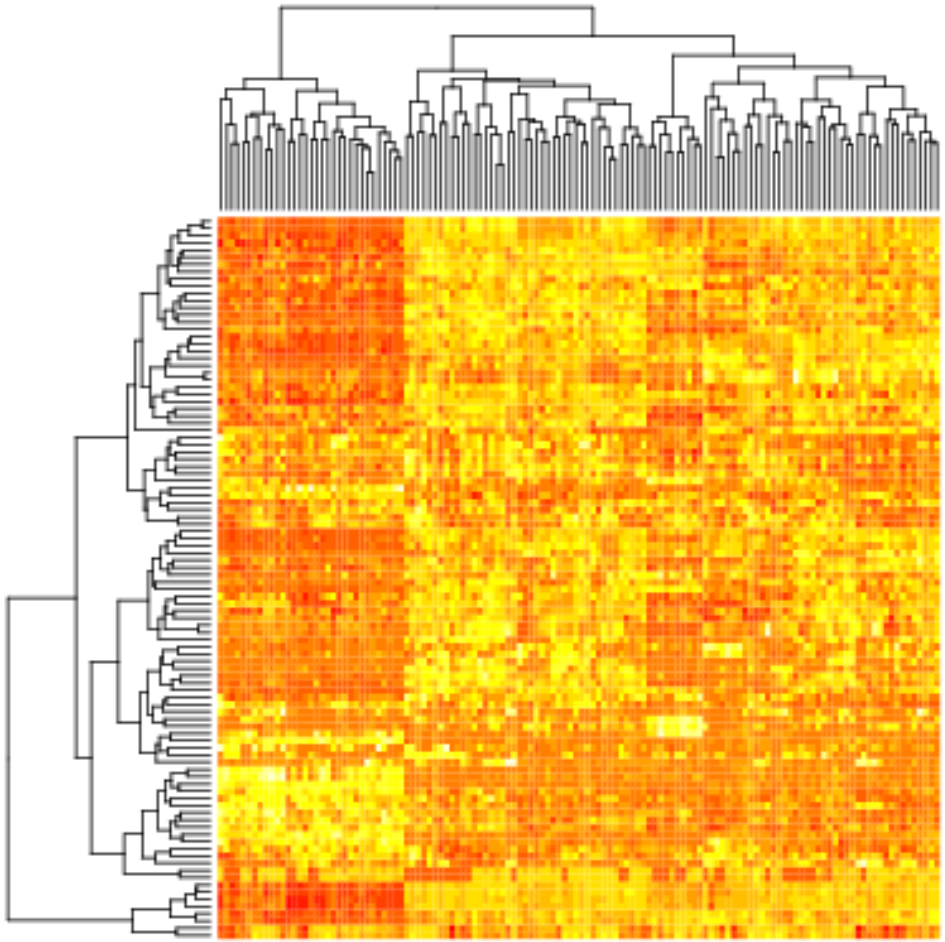
Computational Inference of Cellular Pathways

Florian Markowetz
florian.markowetz@molgen.mpg.de
Max Planck Institute for Molecular Genetics
Computational Diagnostics Group
Berlin, Germany



Practical Microarray Analysis 2006

Coexpression



Coexpression hints to

- **coregulation**
- **gene function**

If genes show the same expression profiles they follow the same regulatory regimes [4, 10].

Co-expression graphs

An expression profile is a random vector $\mathbf{X} = (X_1, \ldots, X_p)$.

Correlation graph: Depict genes as vertices of a graph and draw an edge (i, j) iff the correlation coefficient $\rho_{ij} \neq 0$.

Advantage: This representation of the marginal dependence structure is easy to interpret and can be accurately estimated even if $p \gg N$.

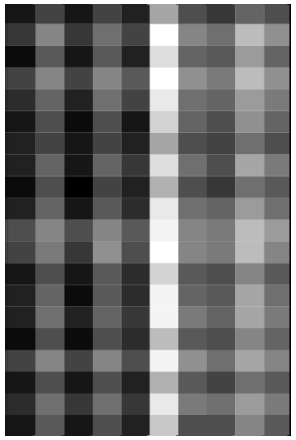
Application: Stuart *et. al* [11] build a graph from coexpression across multiple organisms.

Differential Co-expression

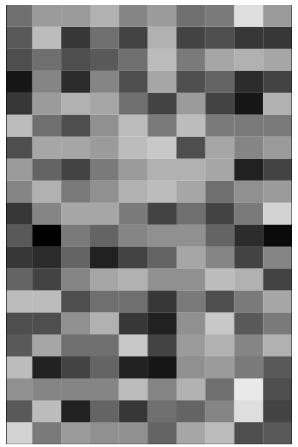
Kostka and Spang [6] find sets of genes, which are correlated in one environment and lose correlation in the second environment.

Interpretation: loss (or gain) of regulatory mechanism.

Genes are controlled by common regulatory mechanism:



Genes in chaos! No regulatory mechanism to organize expression:

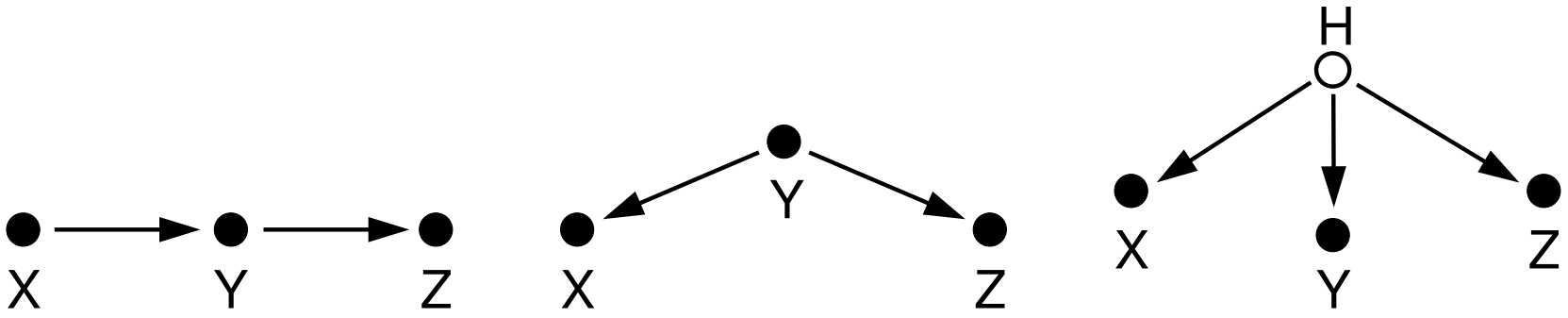


R-package `dcoex` in preparation.

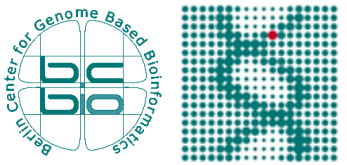
Problems of co-expression based approaches

We cannot distinguish direct from indirect dependencies!

Three reasons, why X, Y, and Z are highly correlated:



As a cure: search for correlations which cannot be explained by other variables.



Part I.

Conditional independence models

Conditional independence

Be X, Y, Z random variables with joint distribution P.

X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow$$

$$P(X = x, Y = y | Z = z) = P(X = x | Z = z) \cdot P(Y = y | Z = z)$$

If I already know Z,
then Y offers me no new information
to understand X.

Conditional independence in Gaussian models

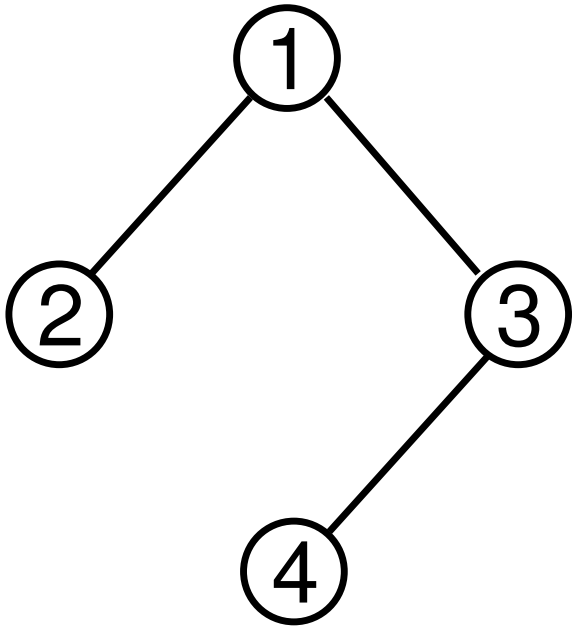
Assume that genes are multivariate normal distributed with covariance matrix Σ. We call $K = \Sigma^{-1}$ the *concentration matrix* of the distribution.

Then it holds for two genes i and j:

$$X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{\text{rest}} \Leftrightarrow k_{ij} = 0$$

Coexpression networks model via the correlation matrix Σ, Gaussian Graphical Models (GGMs) use the inverse $K = \Sigma^{-1}$.

Gaussian Graphical Model



Missing edges indicate independencies:

$$X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{\text{rest}}$$

$$X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}$$
$$X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_4\}$$
$$X_2 \perp\!\!\!\perp X_4 \mid \{X_1, X_3\}$$

What if $p \gg N$?

Full conditional relationships can only be accurately estimated if the number of samples N is relatively large compared to the number of variables p.

Thus, if $p \gg N$, you can ...

1. **either** improve your estimators of partial correlations (*e.g.* Schäfer and Strimmer [9] use the Moore-Penrose pseudoinverse and bootstrap aggregation (bagging) to stabilize the estimator.)
2. **or** resort to a simpler model.

Sparse graphical modeling

Idea: Do not condition on the complete rest as in GGMs. Instead explore dependency of two variables given a single third one.

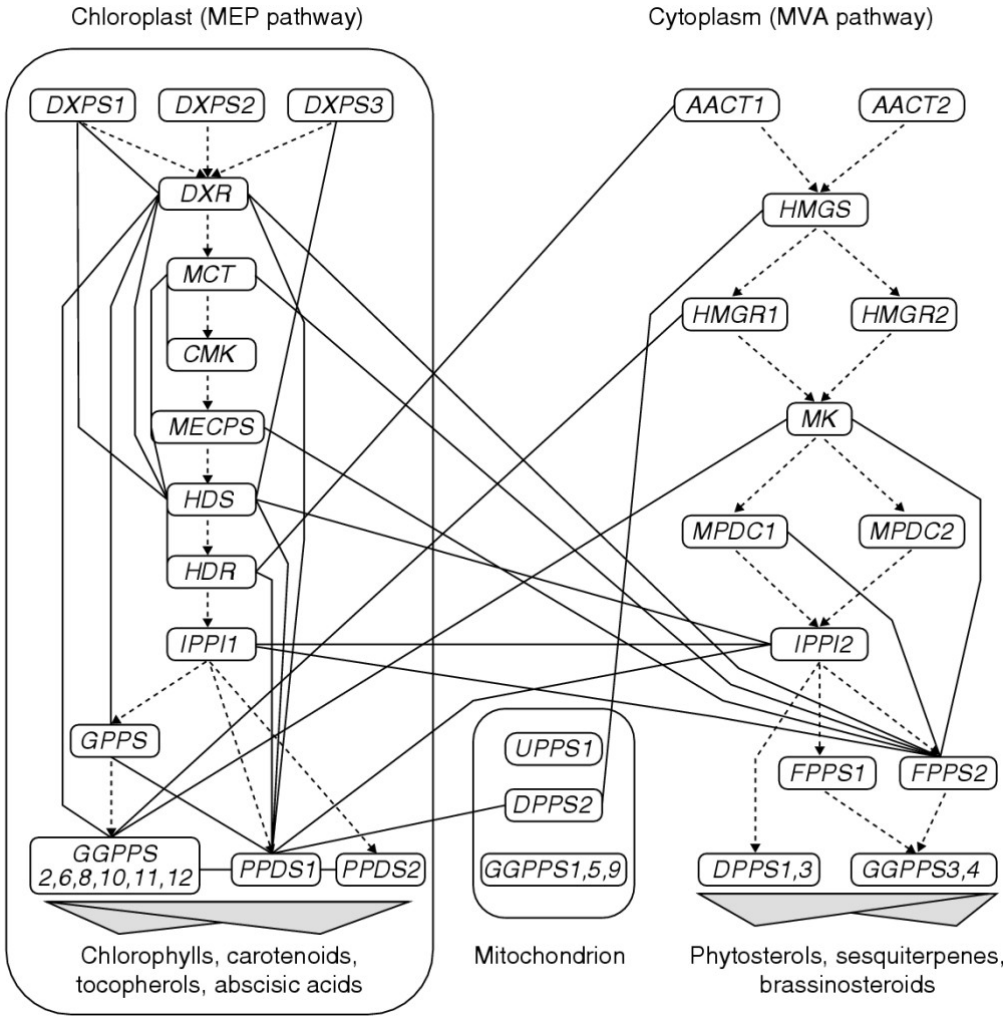
Draw an edge between genes i and j if they are correlated and no third variable can explain the correlation:

$$X_i \not\perp\!\!\!\perp X_j \mid X_k \quad \text{for all } k \in \text{rest.}$$

Implementations

sparse GGMs [2, 7, 12, 13]; mutual information: ARACNE [1]

Isoprenoid gene network in Arabidopsis thaliana



Modules of closely related genes and candidate genes for **cross-talk** between pathways.

Figure from [13]. The **solid undirected edges** connecting individual genes (in boxes) represent the GGM. **Dotted directed edges** mark the metabolic network, and are not part of the GGM.

Where are we?

We have seen methods to build graphs from

1. marginal dependencies

$$X_i \not\perp\!\!\!\perp X_j \mid \emptyset$$

2. full conditional dependence

$$X_i \not\perp\!\!\!\perp X_j \mid X_{\text{rest}}$$

3. first order dependencies

$$X_i \not\perp\!\!\!\perp X_j \mid X_k \quad \forall k \in \text{rest}$$

4. This leads use to include **all higher order dependencies**

$$X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X}_S \quad \text{for all } S \subseteq \text{rest}$$

Bayesian network

A Bayesian Network for a random vector $\mathbf{X}$ consists of

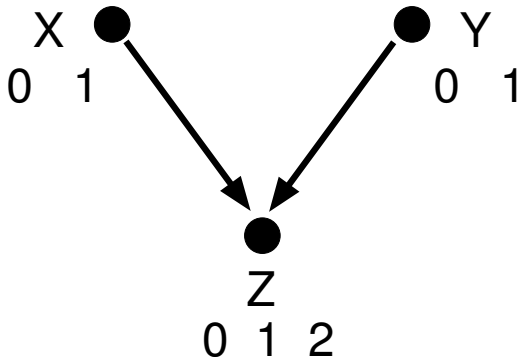
1. a **network structure**
   - directed acyclic graph (DAG) on vertex set V,
   - node v corresponds to variable X_v,
2. a set of **local probability distributions**
   - conditional distribution of a gene given its parents.

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v \mid \mathbf{x}_{pa(v)}, \theta_v)$$

Children depend on parents

The DAG defines families.

Relationships are further characterized by local probability distributions:



$$p(x) = (0.6 \quad 0.4)$$

$$p(y) = (0.2 \quad 0.8)$$

$$p(z|x,y) = \begin{cases} (0.8 \quad 0.1 \quad 0.1) & \text{if } (X,Y) = (0,0) \\ (0.1 \quad 0.8 \quad 0.1) & \text{if } (X,Y) = (0,1) \\ (0.1 \quad 0.8 \quad 0.1) & \text{if } (X,Y) = (1,0) \\ (0.1 \quad 0.1 \quad 0.8) & \text{if } (X,Y) = (1,1) \end{cases}$$

A caveat [5]

If the expression of gene **A** is regulated by proteins **B** and **C**, then **A**'s expression level is a function of the activity levels of **B** and **C**.

Problem 1: In most current biological data sets, however, we do not have access to measurements of protein activity levels.

Resort: Expression levels of genes as a proxy for the activity level of the proteins they encode.

Problem 2: There are numerous examples where an activation or silencing of a regulator is carried out by posttranscriptional protein modifications.

A first summary

1. Conditional independence is the central concept of statistical network models;
2. Graphical models ask: “Can the correlation between two genes be attributed to other genes?”
3. Increasing order of resolution:
   Clustering, Graphical Gaussian models, Bayesian networks;
4. Models don’t capture signaling on protein level.

Part II.

Learning from interventions

Motivation

Response to microbial challenge

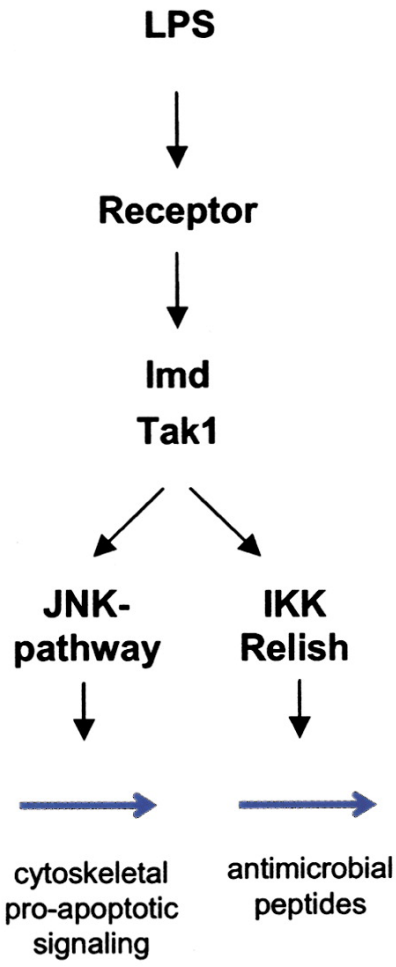
(Boutros *et al.*, Dev Cell, 2002)

Columns: silenced genes.

Rows: effects on other genes.

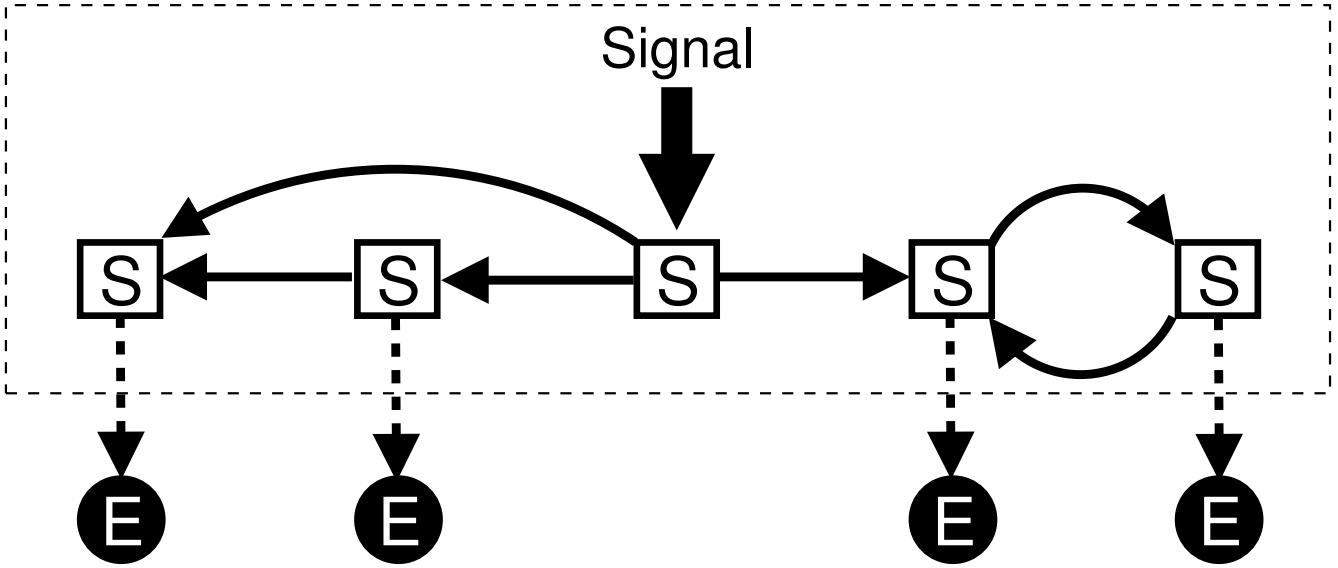
Results:

1. Silencing tak1 reduces expression of all LPS-inducible transcripts.
2. Silencing rel (key) or mkk4/hep reduces expression of separate sets of induced transcripts.



Figures from (Boutros *et al.*, 2002)

The model [8]



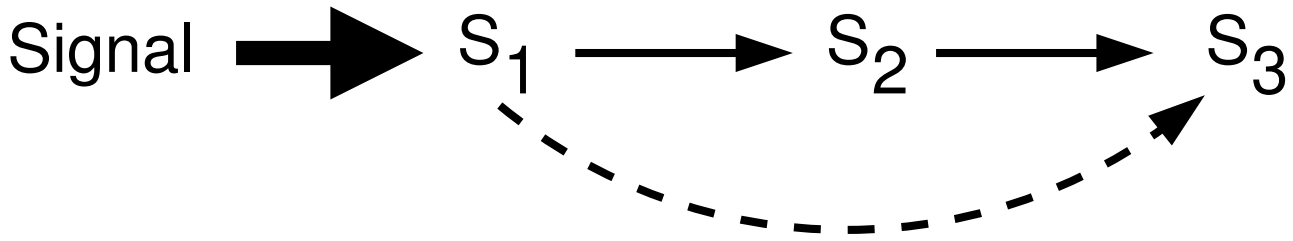
S-genes (for “signaling” or “silenced”): candidate pathway genes.

E-genes (for “effects”): reporters for S-gene activity.

Silencing schemes

A pathway topology allows prediction of intervention effects.

We summarize predictions in a **silencing scheme** Φ: also a directed graph on S-genes, but transitively closed.



Framework flexible to include **epistatic effects** by local logics.

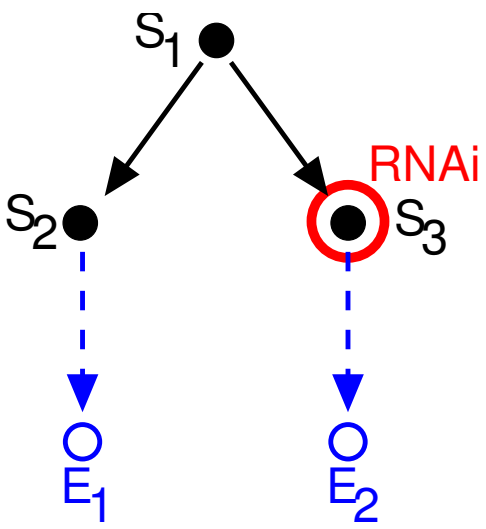
Experiments and Data

Do microarrays for:

1. **Negative controls** — no signal, no interventions
2. **Positive controls** — pathway activated by signal, no interventions
3. **Interventions** — while signal is on!

Data: binary matrix $D = (e_{ik})$,
where $e_{ik} = 1$ if E-gene E_i shows in experiment k the same expression as in the negative controls.

Likelihood



The silencing scheme Φ allows **prediction** of E-gene states (when position is known).

We expect a number of **false positive and false negative** observations.

The **likelihood** $P(D|\Phi,\Theta)$ is a product over atomic terms:

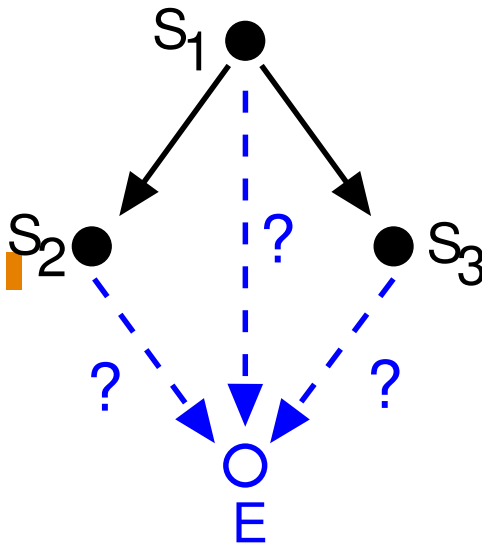
$$P(e_{ik}|\Phi,\theta_i=j)=\left\{\begin{array}{cc|l} e_{ik}=1 & e_{ik}=0 & \\ \hline \alpha & 1-\alpha & \text{if } \Phi \text{ predicts } \textbf{no effect} \\ 1-\beta & \beta & \text{if } \Phi \text{ predicts } \textbf{effect} \end{array}\right.$$

Marginal likelihood

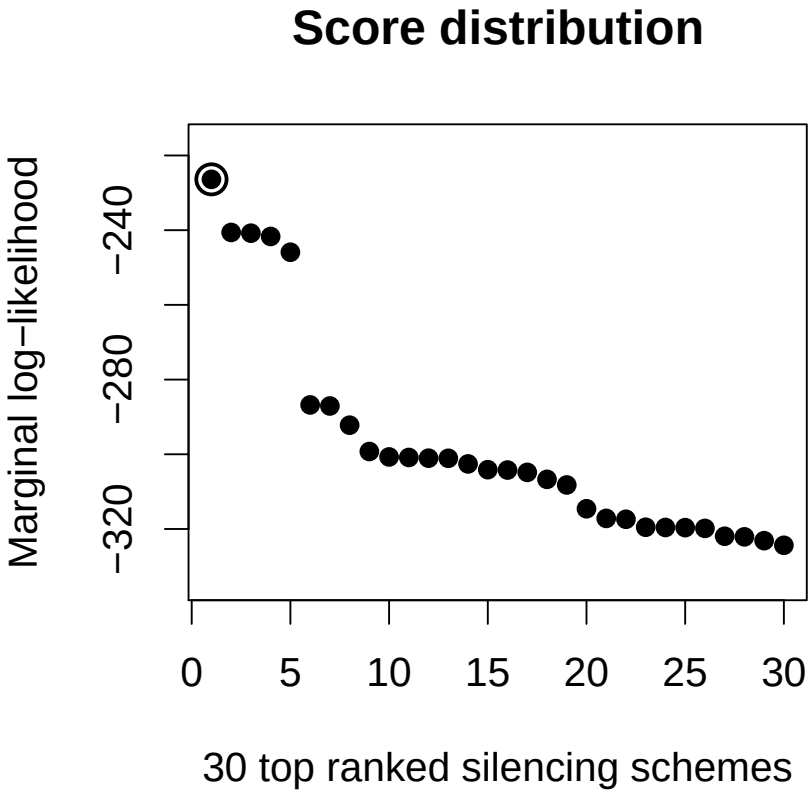
Computation of likelihood requires that E-gene positions are known. In reality this is not true.

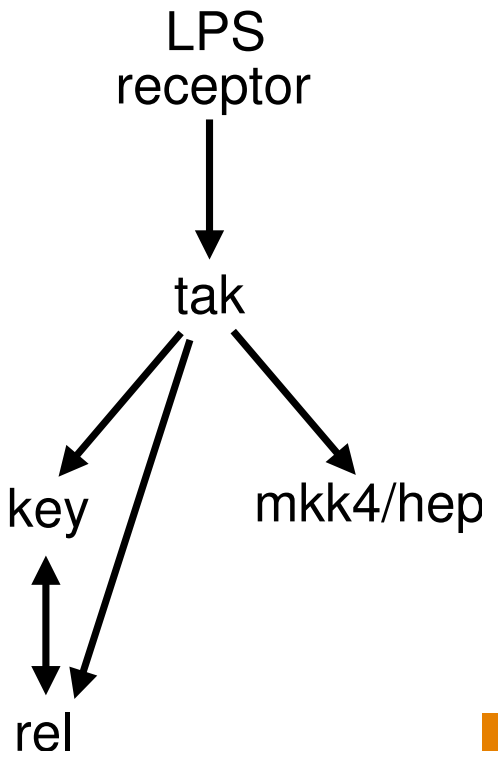
$$P(D|\Phi) = \int P(D|\Phi,\Theta)P(\Theta|\Phi)\,\mathrm{d}\Theta$$

$$= \frac{1}{\eta^m}\prod_{i=1}^{m}\sum_{j=1}^{n}\prod_{k=1}^{l}P(e_{ik}|\Phi,\theta_i=j)$$



Application to *Drosophila* data

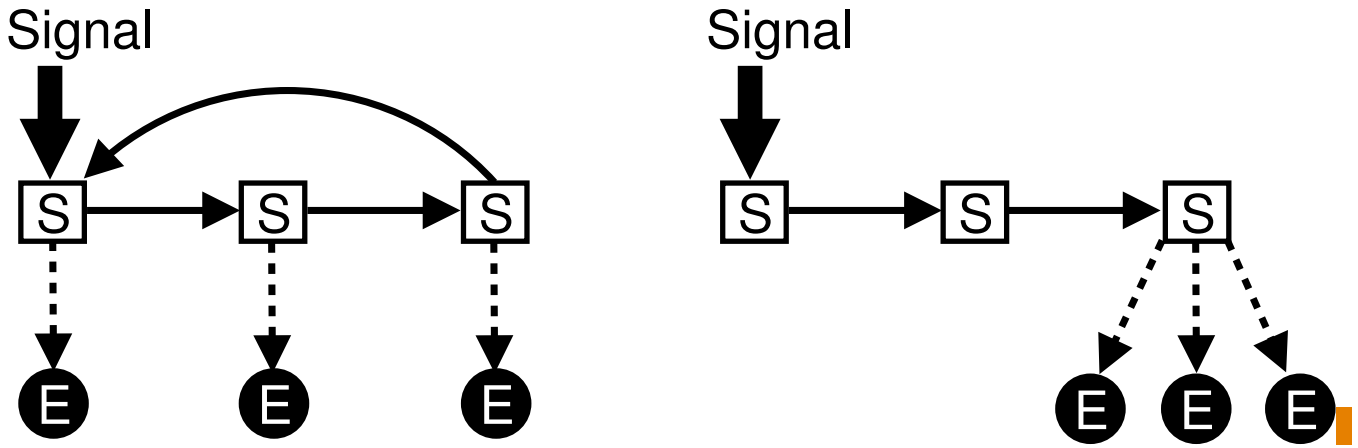




Limits of identification

Prediction equivalence — Multiple pathway topologies result in the same silencing scheme, if they only differ in transitive edges.

Likelihood equivalence — Two hypotheses with different silencing schemes can produce identical data:



Conclusion

- The algorithm reconstructs pathway features from the nested structure of affected down-stream genes.
- Pathway features are encoded as silencing schemes. They contain all information to predict a cell's behaviour to an external intervention.
- **Not shown:** in simulation studies we confirmed small sample size requirements and high accuracy.
- Limitations only result from the information content of indirect observations.

On true models

A quote from Edwards [3]:

"*Any method (or statistician) that takes a complex multivariate dataset and, from it, claims to identify one true model, is both naive and misleading.*"

What we have found is just

a simple model consistent with the data.

Graphical models in R

www.r-project.org/gR

ggm: Gaussian Graphical Models

deal: Bayesian networks with mixed variables

www.bioconductor.org

GeneTS: large GGMs

compdiag.molgen.mpg.de/software

dcoex: finding groups of differentially coexpressed genes

References

[1] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, *et al.* Reverse engineering of regulatory networks in human B cells. *Nat Genet*, Mar 2005.

[2] Alberto de la Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[3] David Edwards. *Introduction to Graphical Modelling*. Springer, 2000.

[4] MB Eisen, PT Spellman, PO Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, Dec 1998.

[5] Nir Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659):799–805, 2004.

[6] Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20 Suppl 1:I194–I199, Aug 2004.

[7] Paul M Magwene and Junhyong Kim. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 5(12):R100, 2004.

[8] Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 2005.

[9] Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–64, Mar 2005.

[10] PT Spellman, G Sherlock, MQ Zhang, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, Dec 1998.

[11] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, Oct 2003.

[12] Anja Wille and Peter Bühlmann. Tri-graph: a novel graphical model with application to genetic regulatory networks. Technical report, Seminar for Statistics, ETH ZÃrich, 2004.

[13] Anja Wille, Philip Zimmermann, Eva Vranová, *et al.* Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol*, 5(11):R92, 2004.