Computational Inference of Cellular Networks from Microarray Data

Achim Tresch

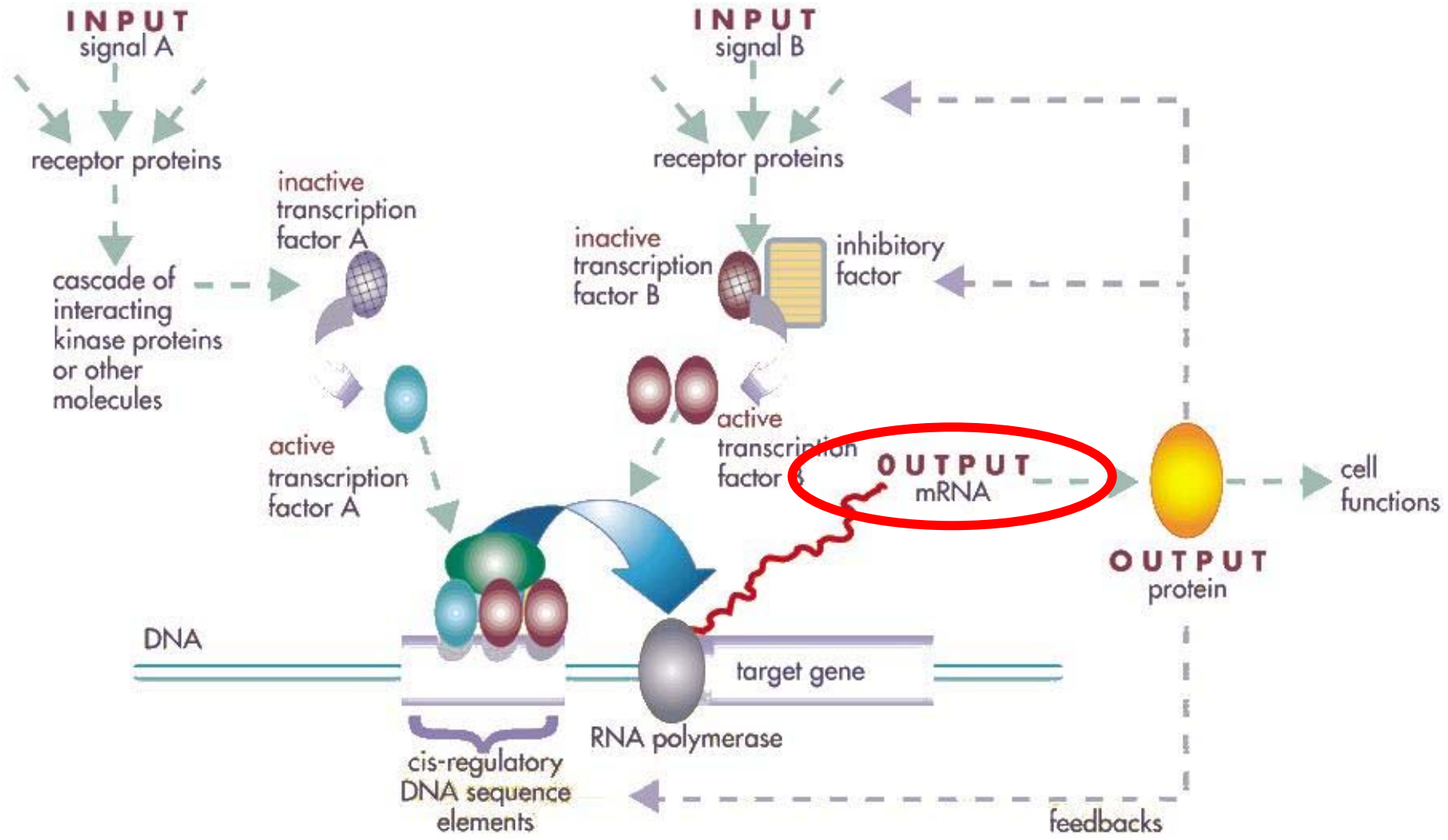
Topics

- **Biological networks vs. Network Models**
- **Learning Networks from non-interventional (=observational) data:**
  - Correlation Graphs
  - Gaussian Graphical Models
  - Bayesian Networks
- **Learning from interventional data:**
  - Pruning
  - Nested Effects Models

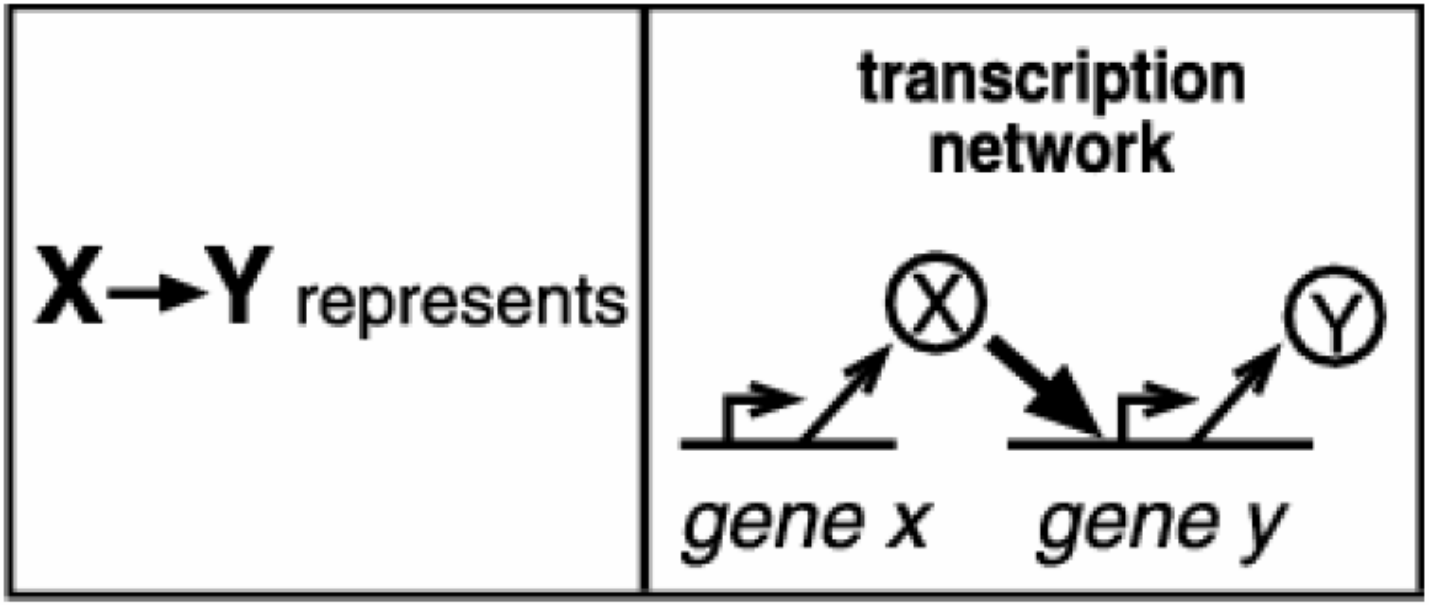
"All models are wrong, some of them are useful"

(Edwards Deming, George Box)

Which biological Network?



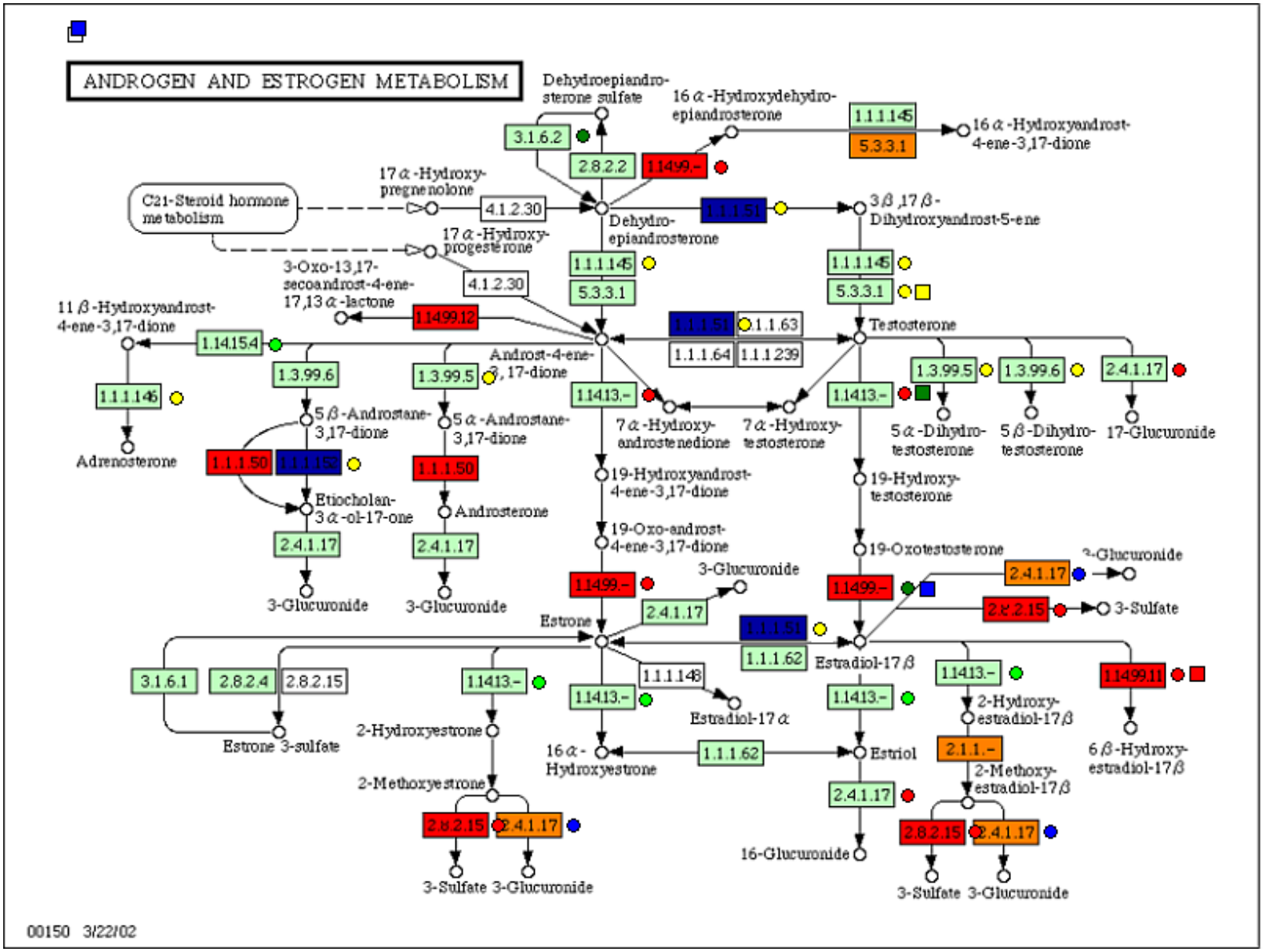
Transcription Networks



Nodes = transcription factors

Directed edge: X regulates transcription of Y

Transcription Networks



LEGEND

- high expression
- medium expression
- no difference
- low expression
- protein not found in array

TRANSCRIPTION FACTORS:

- Oct-1
- p107
- GATA-1
- Sp3
- Rb

DRUG TARGETS

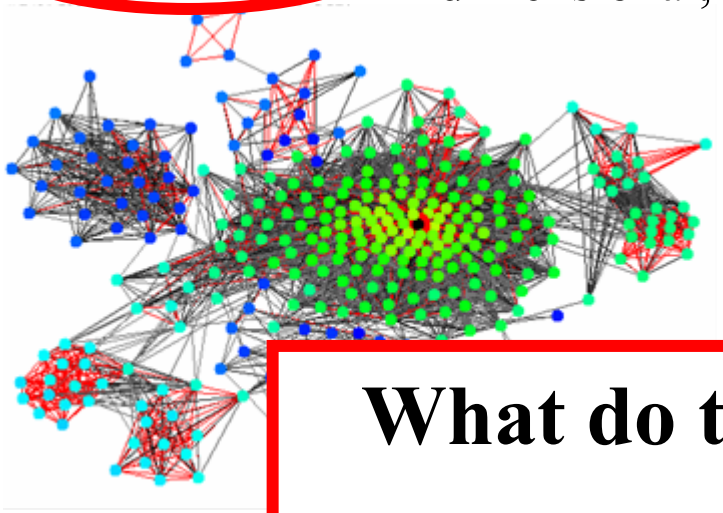
- thamoxifen
- flutamide
- anastrazole
- masterlone

From the KEGG database

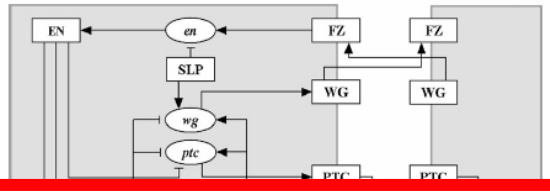
Which Network Model?

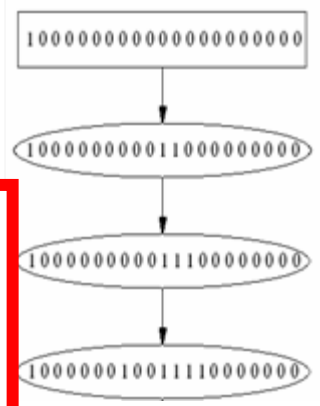
qualitative

Best suited for high dimensional, noisy data



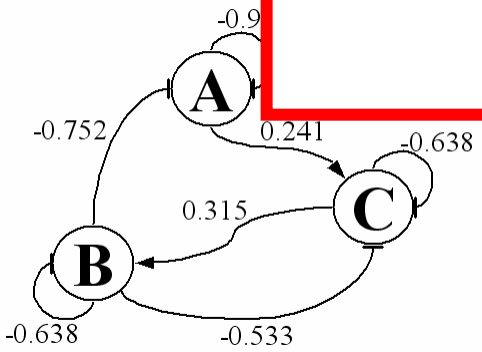
semiquantitative





What do the arrows mean?

Do they have a biological interpretation?

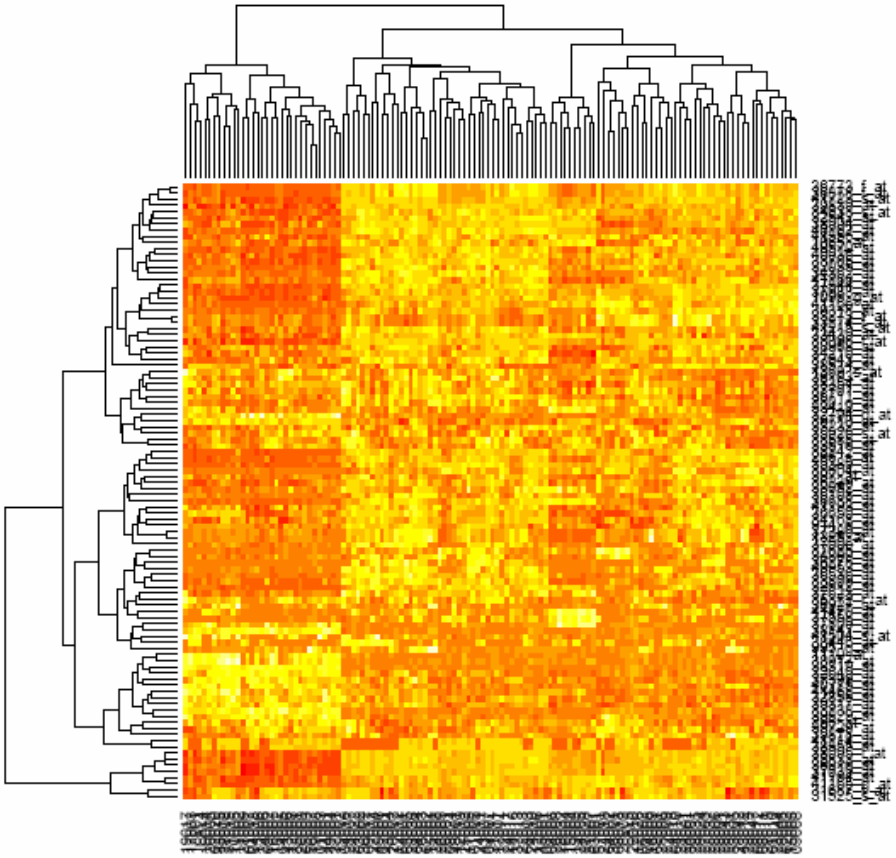


$$\frac{d[B]}{dt} = \frac{V_b}{1+\frac{K_{aC}}{[C]}} - k_b[B]$$

$$\frac{d[C]}{dt} = \frac{V_c}{\left(1+\frac{[B]}{K_{iB'}}\right)\left(1+\frac{K_{aA}}{[A]}\right)} - k_c[C]$$

quantitative

Clustering by coexpression



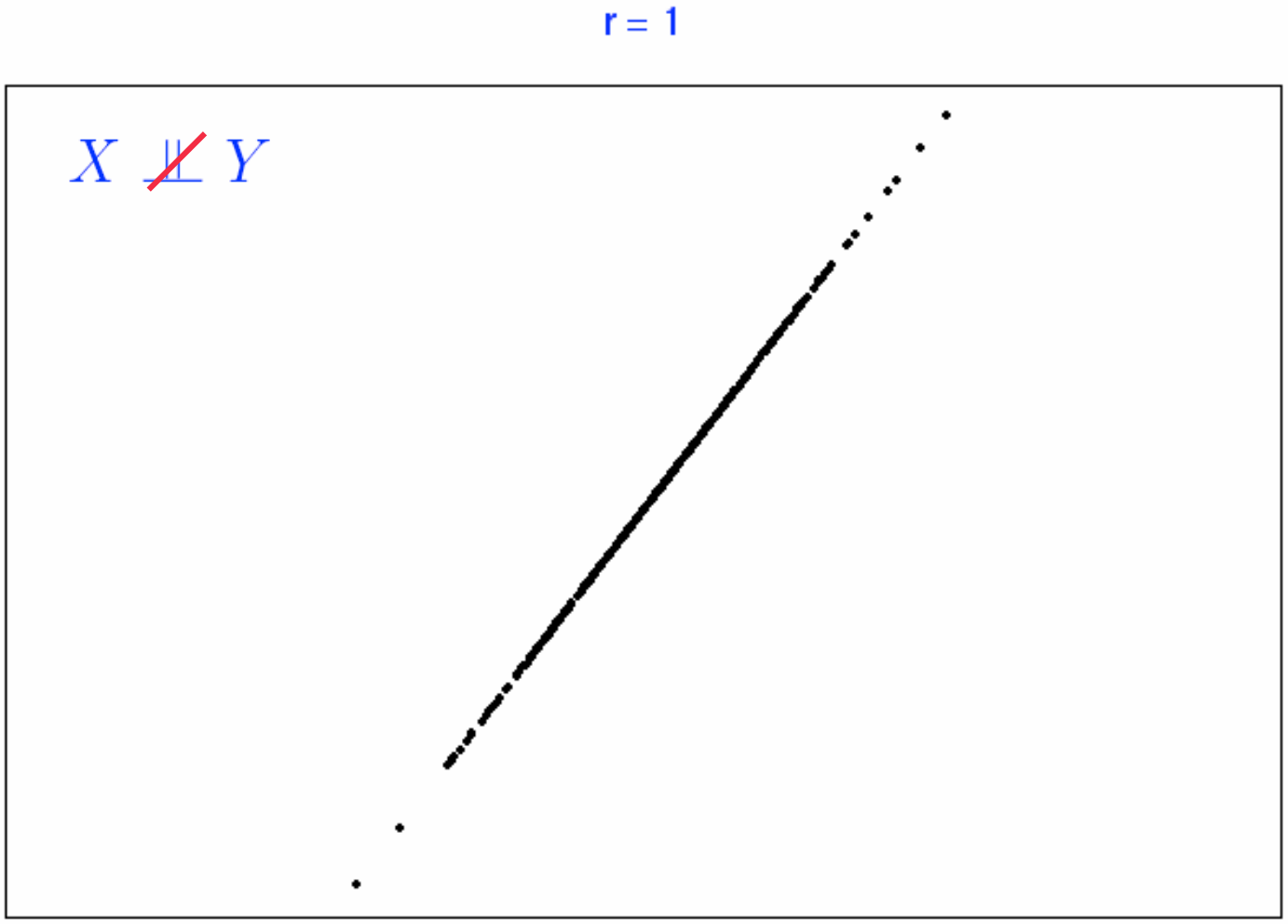
Assumption:
Coexpression $\sim$ coregulation

If genes show the same expression profiles they follow the same regulatory regimes

Coexpression is conveniently measured by correlation.

(Pearson) Correlation

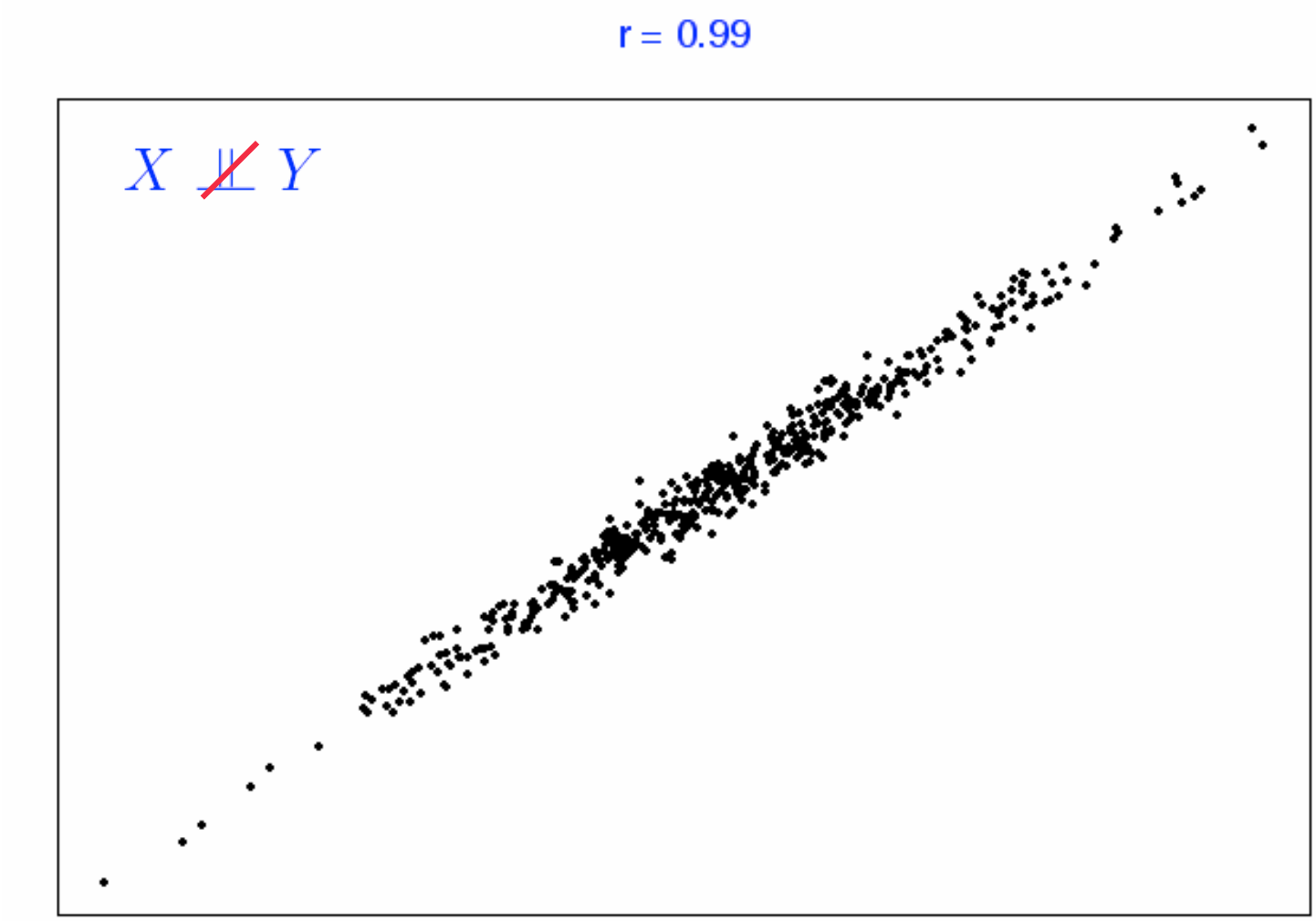
Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

(Pearson) Correlation

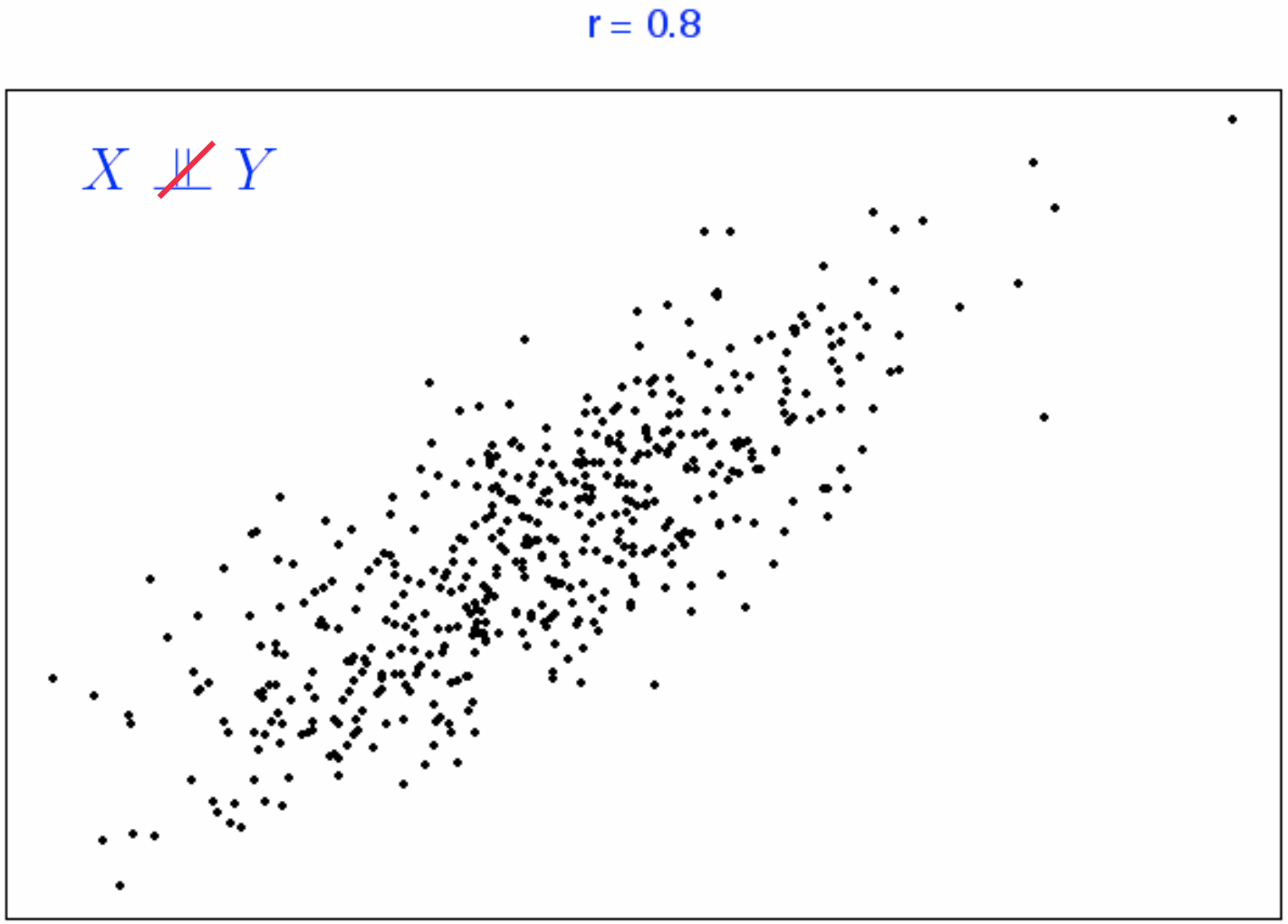
Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

(Pearson) Correlation

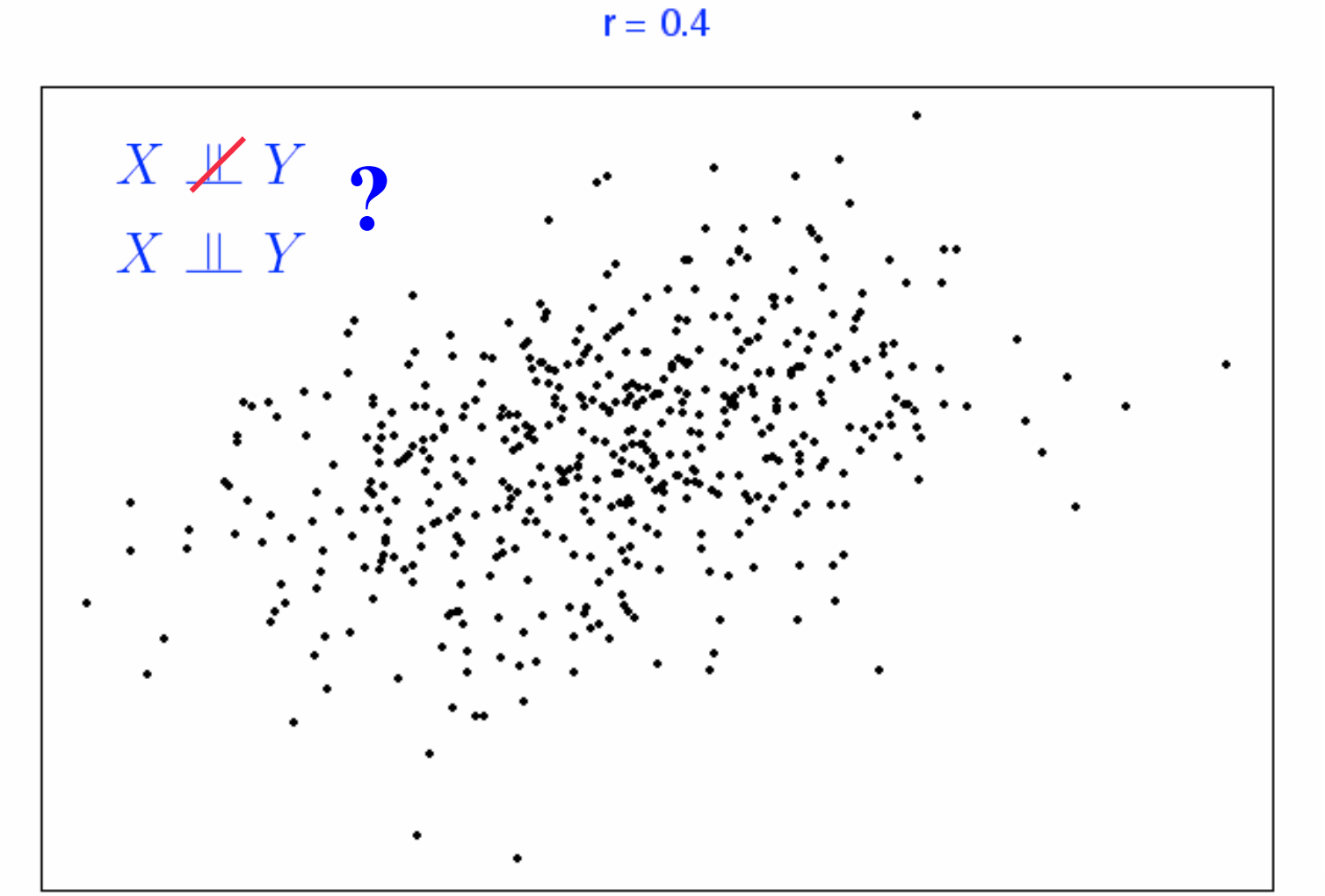
Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

(Pearson) Correlation

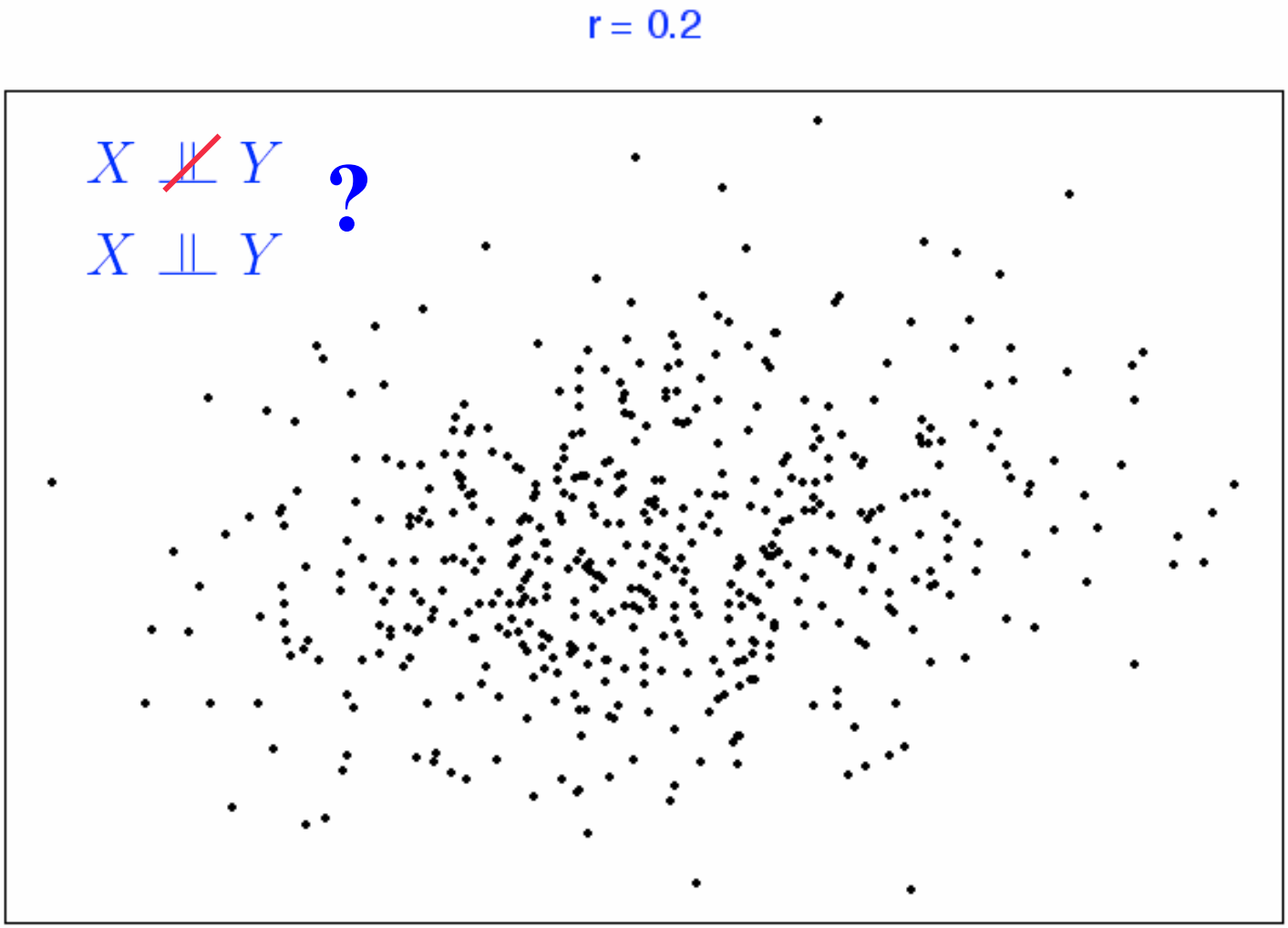
Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

(Pearson) Correlation

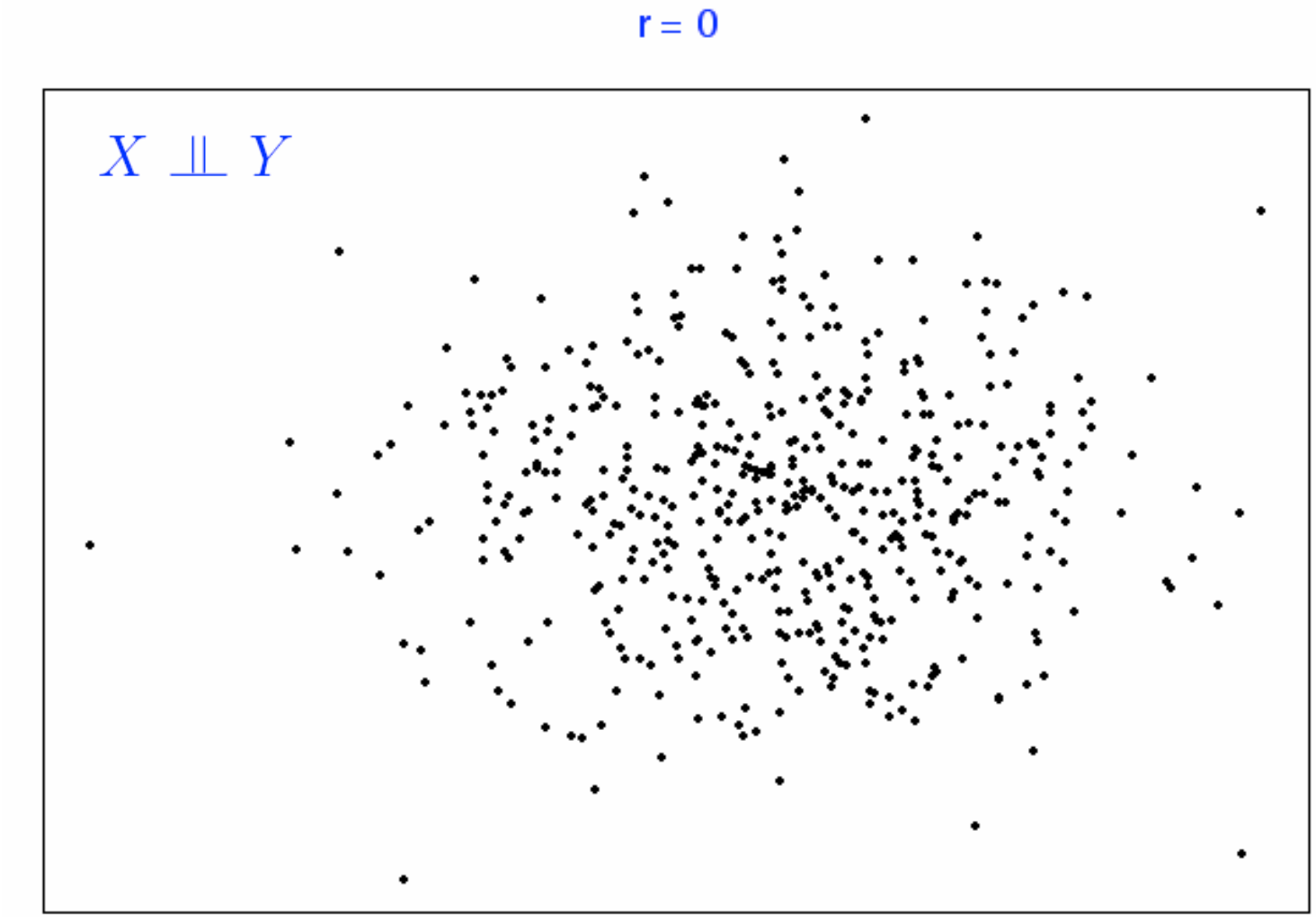
Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

(Pearson) Correlation

Correlation close to 1 or -1 → strong linear dependence
Correlation close to 0 → no or weak linear dependence



Grafiken von A.Wakolbinger

Correlation Graphs

- An expression profile is a collection of expression vectors

$$\{ X_g = (X_{g,s})_{s \in \text{samples}} , g \in \text{Genes} \}$$

- Correlation graph: Genes are the vertices of the graph and an undirected edge (i, j) is drawn if some correlation measure (Pearson correlation, Spearman rank correlation, Kendall's tau) between X_i and X_j is sufficiently different from zero.

- Advantage: This representation of the marginal dependence structure is easy to interpret and can be estimated accurately even if

$$p > N$$

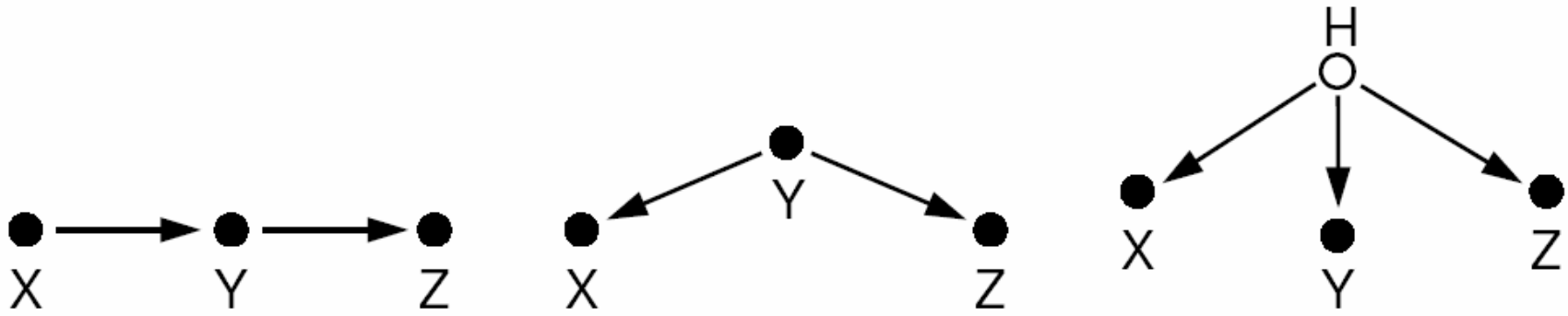
the number of features (genes) > the number of samples

- Application: Stuart et al. (Science, 2003) build a graph from coexpression across multiple organisms.

Problems of correlation based approaches

- **It is impossible to distinguish direct from indirect dependence**

  Three reasons why X, Y , and Z may be highly correlated:



- **A strong correlation is not a strong evidence for regulatory dependence (lots of false positives). But a low correlation is a strong evidence for no regulatory dependence.**

Possible remedies:

- **search for correlations which cannot be explained by other variables.**
- **measure effects of gene perturbations**

Conditional Independence

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

X is conditionally independent of Y given Z

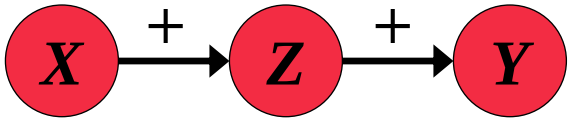
$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow$$

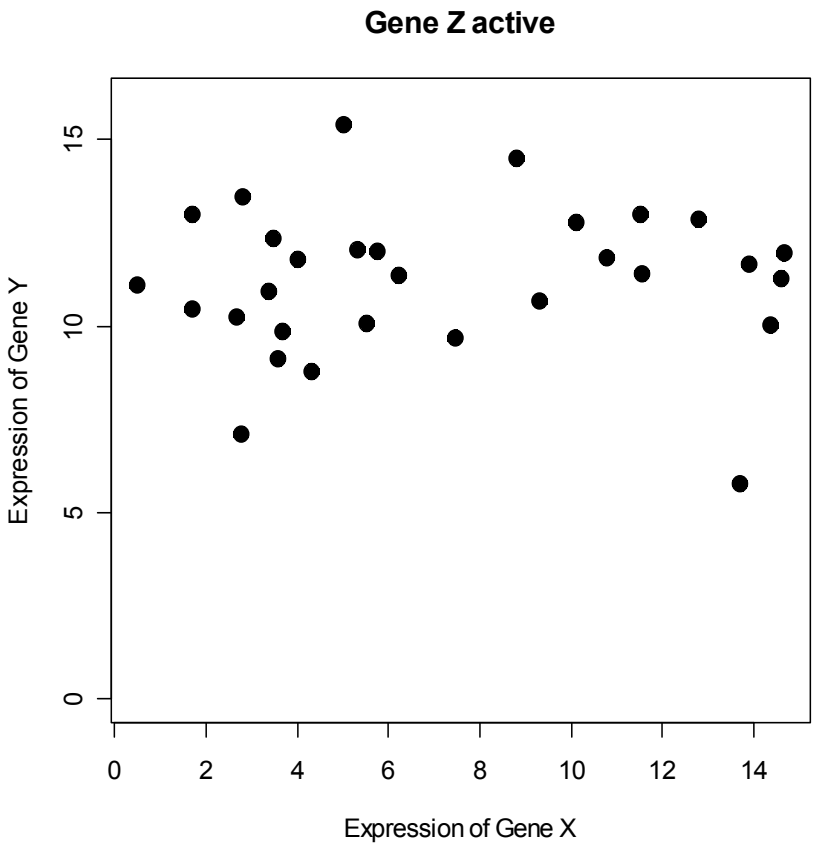
$$P(X = x, Y = y | Z = z) = P(X = x | Z = z) \cdot P(Y = y | Z = z)$$
$$P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

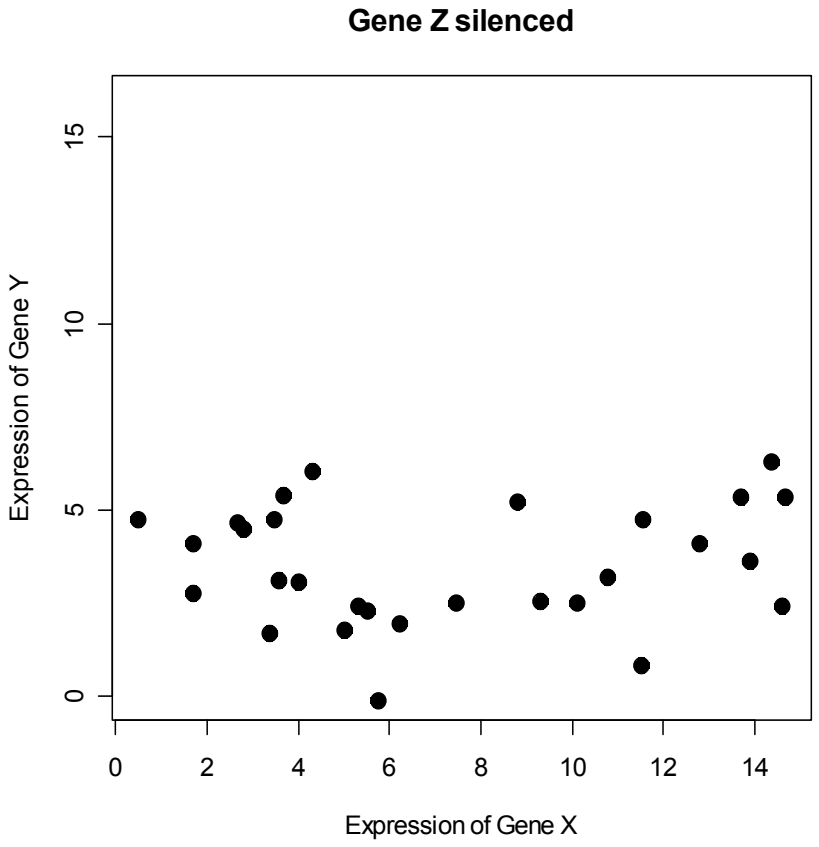
In other words:

- **Knowing Z, knowing Y is irrelevant for knowing X (and vice versa).**
- **Z "explains" any observed dependence between X and Y .**

Conditional Independence

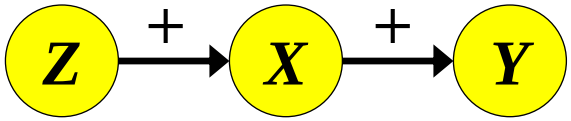


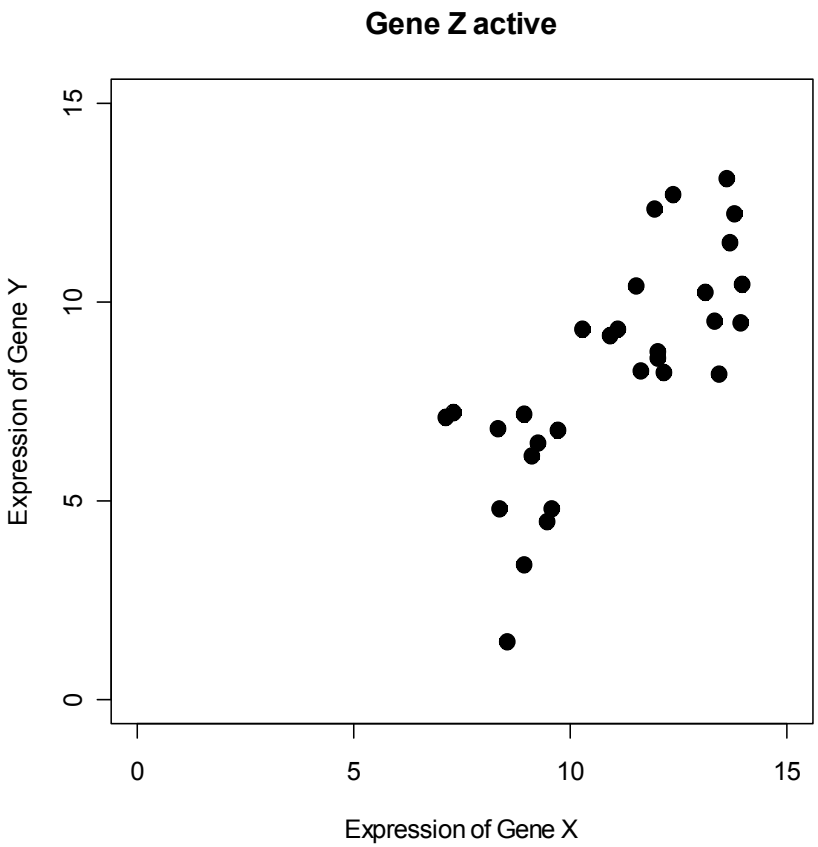


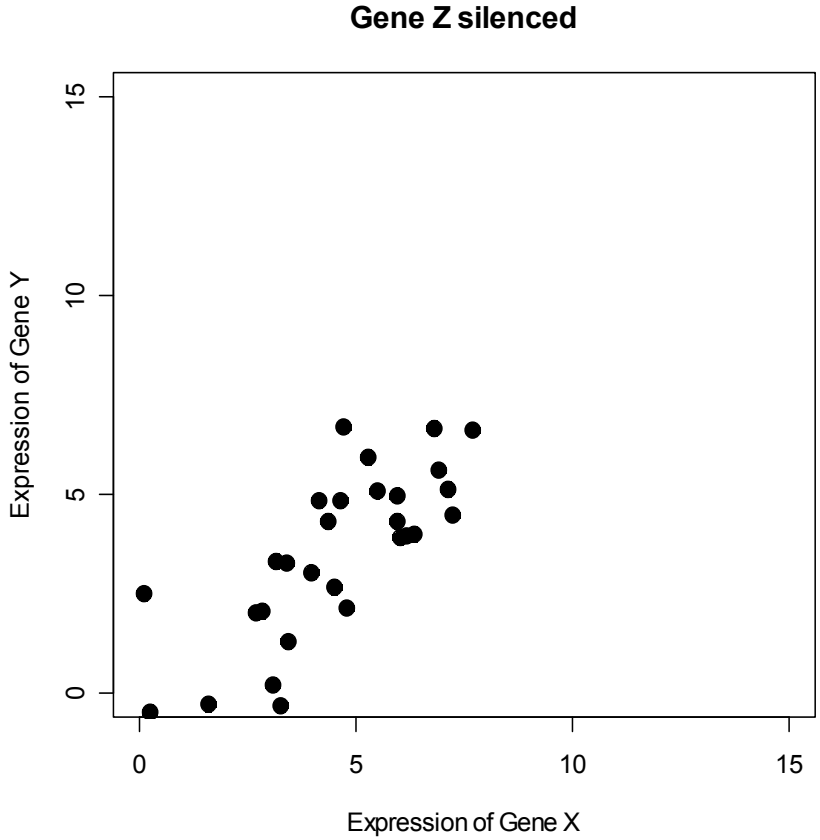


$$X \perp\!\!\!\perp Y \mid Z$$

Conditional Independence







$$X \not\perp\!\!\!\perp Y \mid Z$$

Gaussian Graphical Models (GGM)

Given a random vector $\mathbf{X} = (X_1, \ldots, X_p)$.

A Gaussian graphical model [7, 4] is an **undirected graph** on vertex set V, with $|V| = p$.

To each vertex $i \in V$ corresponds a **random variable** $X_i \in \mathbf{X}$.

Draw an **edge** between vertices i and j if and only if

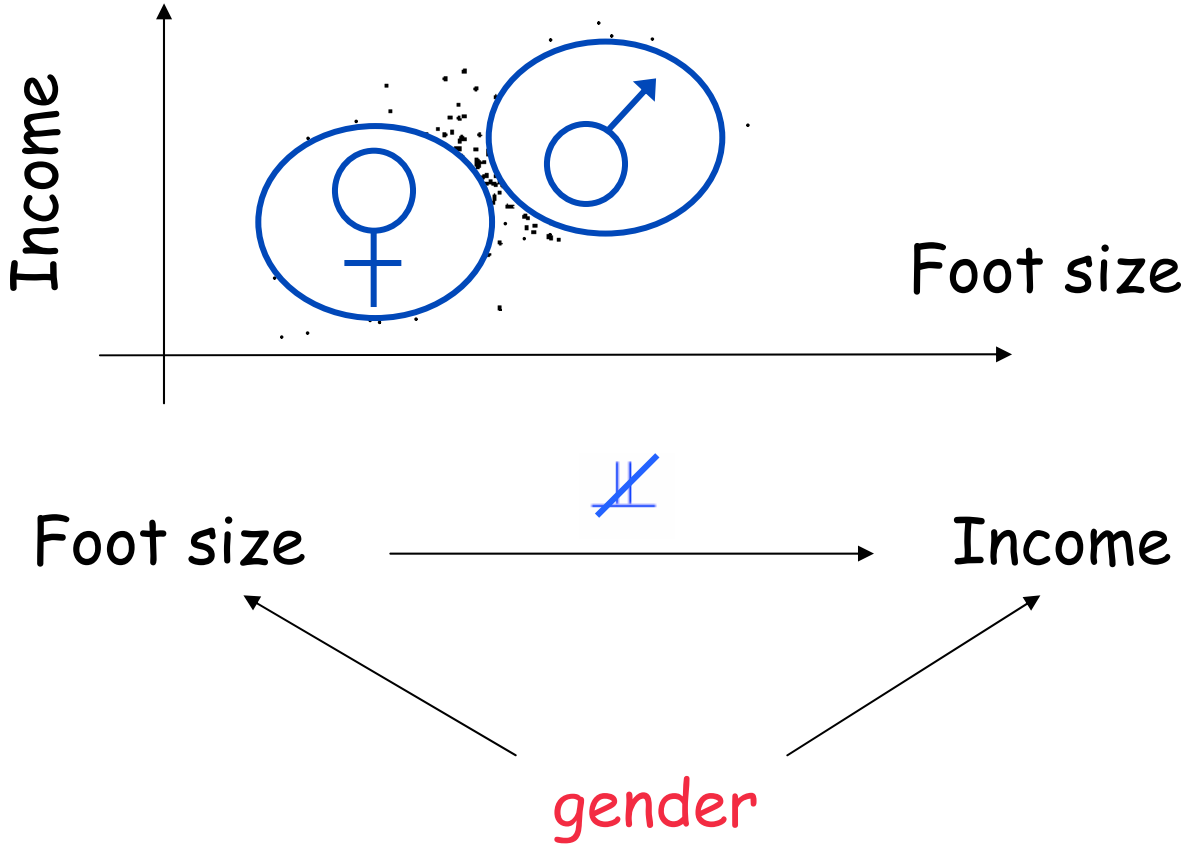
$$X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X}_{\text{rest}}$$

Do **not** draw an edge between vertices *i* and *j* if and only if

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

Gaussian Graphical Models (GGM)

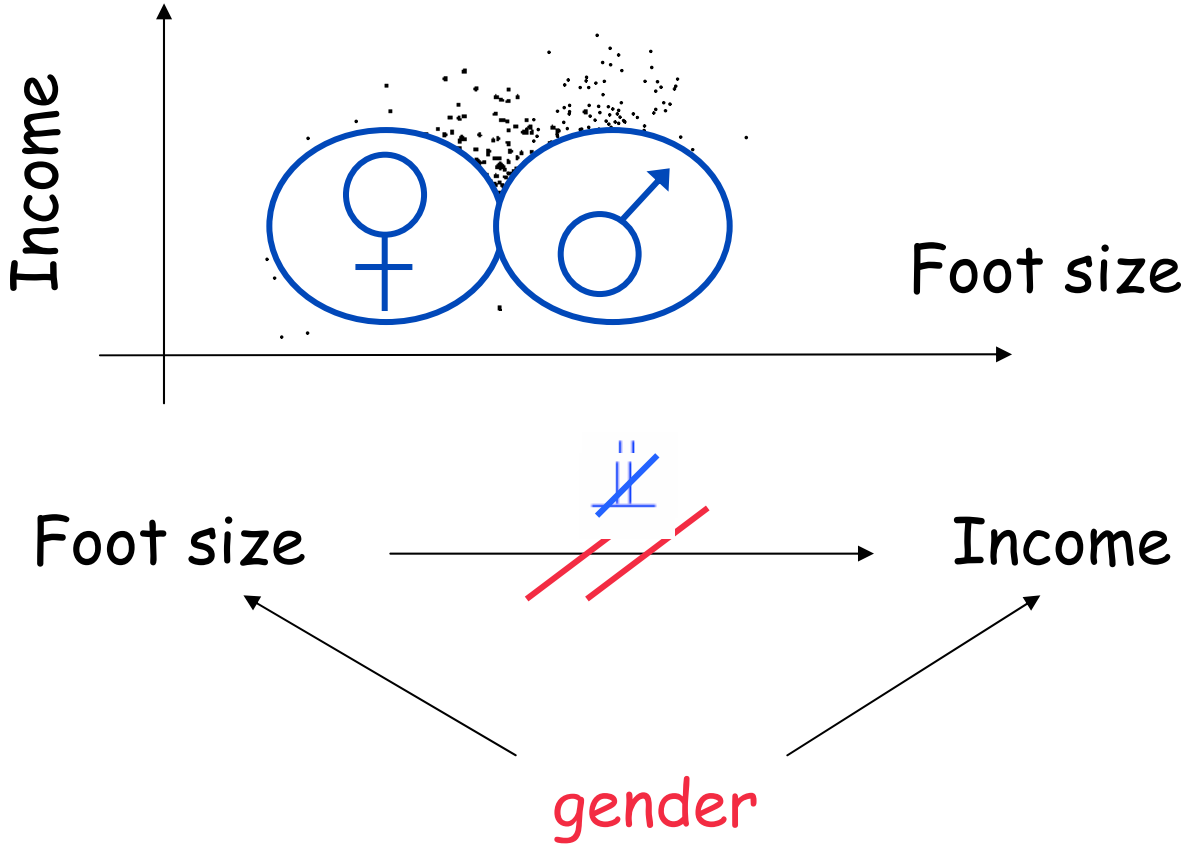
Example:



The variable „gender" explains the correlation of foot size and income.

Gaussian Graphical Models (GGM)

Example:



The variable „gender" explains the correlation of foot size and income.

Gaussian Graphical Models (GGM)

If we assume that the common expression distribution of all genes follows a multivariate Gaussian distribution (which is of course ridiculous), conditional independence can be assessed as follows:

1. First estimate the covariance matrix Σ by the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{N-1}(X - \bar{X})^T(X - \bar{X}).$$

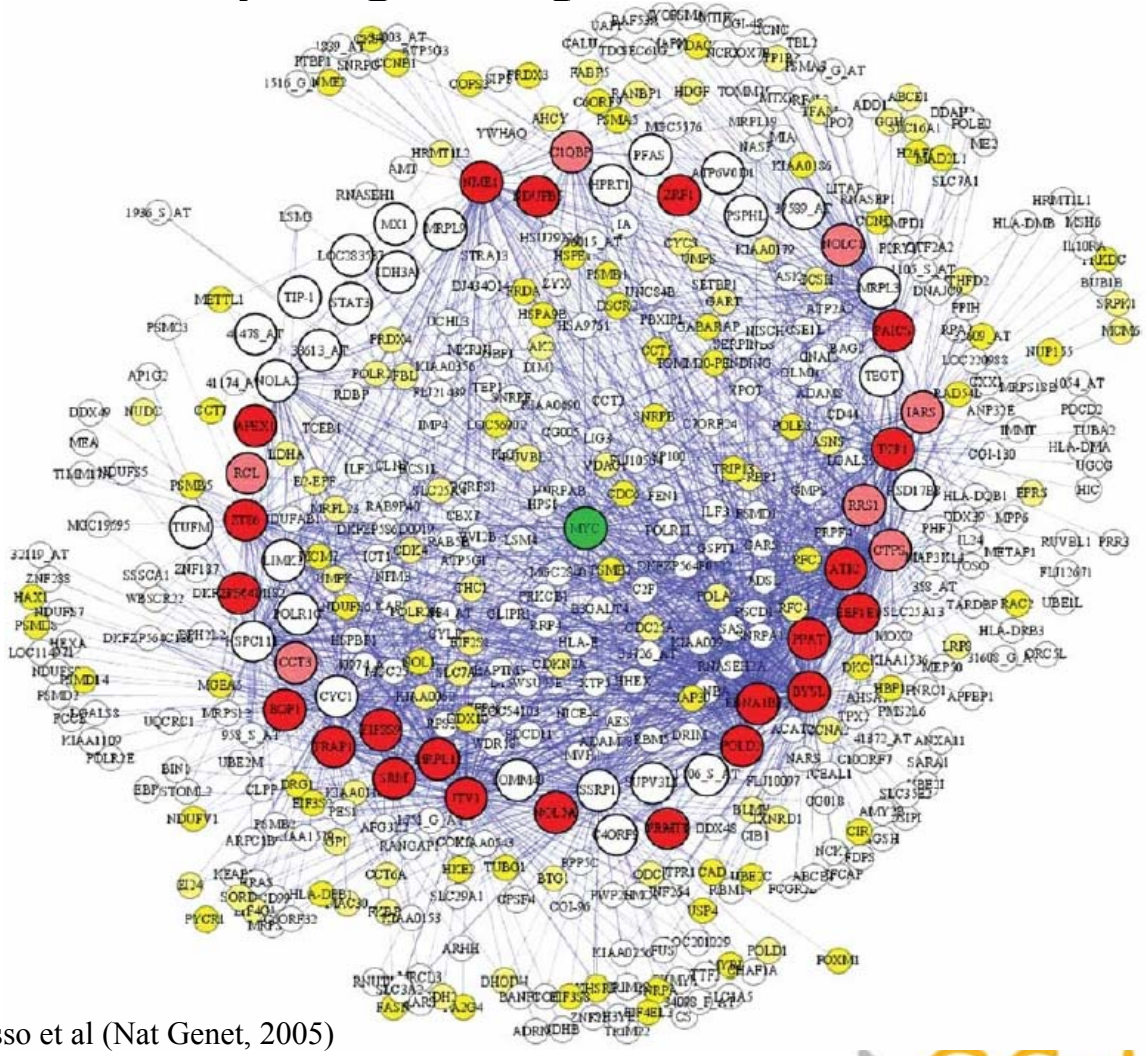
2. Invert $\hat{\Sigma}$ to obtain an estimate $\hat{K}$ of the precision matrix K.
3. Employ statistical tests [56] to decide, which entries in $\hat{K}$ are significantly different from zero.

What if p » 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 » N, you can . . .

- **use the Moore-Penrose pseudoinverse, bootstrap aggregation and shrinkage estimators to stabilize the result (e.g. Schäfer and Strimmer, Bioinformatics´05)**
- **resort to a simpler model that does not rely on full conditional independence**



Graph from Basso et al (Nat Genet, 2005)

Modified Gaussian Graphical Models (GGM)

We have seen methods to build graphs from

1. marginal dependencies

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

2. full conditional dependence

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

3. first order dependencies

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

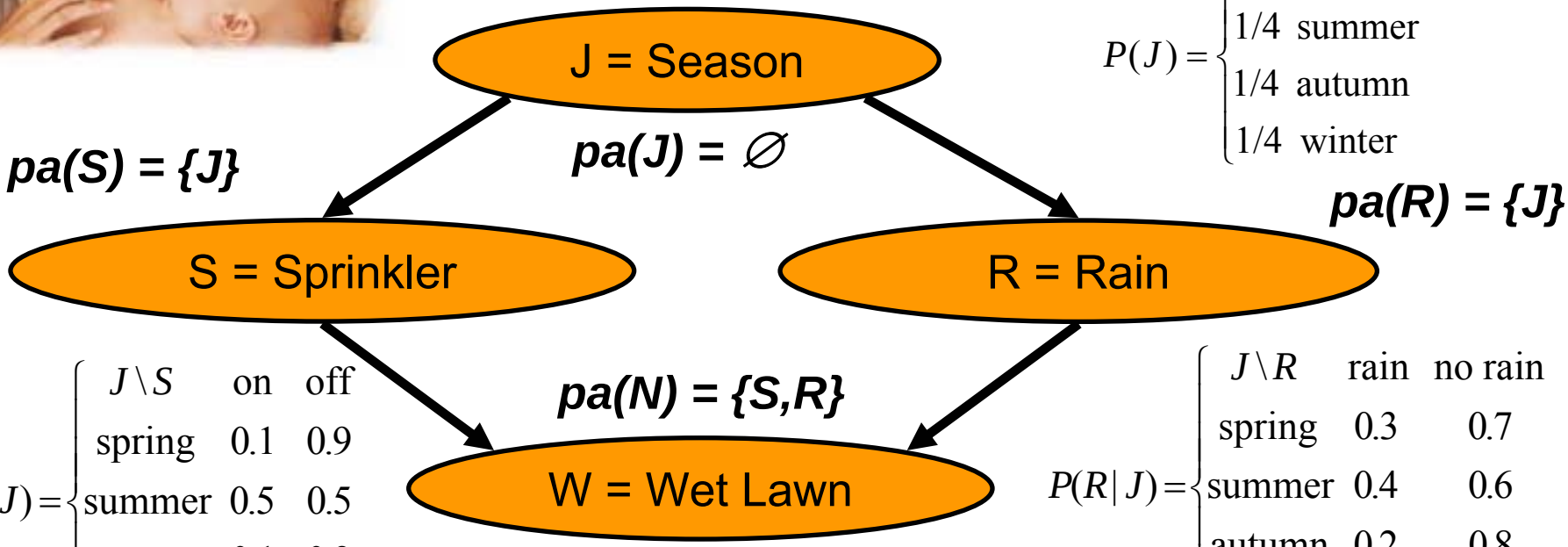
4. This leads use to include **all higher order dependencies**

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

All methods fail to accurately reconstruct networks, even if they are of moderate size (~20)

Bayesian Networks: Children depend on Parents

The common probability distribution of a Bayes Net is the product of its local probability distributions.



$$P(J) = \begin{cases} 1/4 & \text{spring} \\ 1/4 & \text{summer} \\ 1/4 & \text{autumn} \\ 1/4 & \text{winter} \end{cases}$$

$$P(S \mid J) = \begin{cases} J \setminus S & \text{on} & \text{off} \\ \text{spring} & 0.1 & 0.9 \\ \text{summer} & 0.5 & 0.5 \\ \text{autumn} & 0.1 & 0.9 \\ \text{winter} & 0 & 1 \end{cases}$$

$$P(R \mid J) = \begin{cases} J \setminus R & \text{rain} & \text{no rain} \\ \text{spring} & 0.3 & 0.7 \\ \text{summer} & 0.4 & 0.6 \\ \text{autumn} & 0.2 & 0.8 \\ \text{winter} & 0.1 & 0.9 \end{cases}$$

$$P(W \mid S,R) = \begin{cases} (S,R) \setminus W & \text{wet} & \text{dry} \\ (\text{on},\text{rain}) & 1 & 0 \\ (\text{on},\text{no rain}) & 1 & 0 \\ (\text{off},\text{rain}) & 0.9 & 0.1 \\ (\text{off},\text{no rain}) & 0.01 & 0.99 \end{cases}$$

The sprinkler network

Bayesian Networks: The Sprinkler Network

The common distribution of {J,S,R,N} can be coded by the graph topology and 3+4+4+4 =15 real numbers: (instead of 4·2·2·2-1 = 31 real numbers for an arbitrary distribution)

$$pa(.) = \begin{cases} J \mapsto \{\} \\ S \mapsto \{J\} \\ R \mapsto \{J\} \\ N \mapsto \{S, R\} \end{cases}$$

$$P(J) = \begin{cases} 1/4 & \text{spring} \\ 1/4 & \text{summer} \\ 1/4 & \text{autumn} \\ 1/4 & \text{winter} \end{cases}$$

$P(S \mid J)$	on	off
spring	0.1	0.9
summer	0.5	0.5
autumn	0.1	0.9
winter	0	1

$P(R \mid J)$	rain	norain
spring	0.3	0.7
summer	0.4	0.6
autumn	0.2	0.8
winter	0.1	0.9

$P(N \mid S,R)$	wet	dry
(on,rain)	1	0
(on,norain)	1	0
(off,rain)	0.9	0.1
(off,norain)	0.01	0.99

$$P(J = j, S = s, R = r, N = n) = P(N = n \mid S = s, R = r) \cdot P(S = s \mid J = j) \cdot P(R = r \mid J = j) \cdot P(J = j)$$

E.g.

$$\begin{aligned} P(J = \text{summer}, S = \text{off}, R = \text{rain}, N = \text{wet}) &= P(N = \text{wet} \mid S = \text{off}, R = \text{rain}) \cdot P(R = \text{rain} \mid J = \text{summer}) \\ &\quad \cdot P(S = \text{off} \mid J = \text{summer}) \cdot P(J = \text{summer}) \\ &= 0.9 \cdot 0.4 \cdot 0.5 \cdot 0.25 \\ &= 0.045 \end{aligned}$$

Bayesian Networks

Problems:

- **Given a directed acyclic graph (DAG), learn the local probability distributions and score the DAG according to its likelihood („how good does this graph fit the data“?) → Parameter estimation, Bayesian Dirichlet metric (Cooper, Herskovits 1992)**

- **Find the topology(-ies) of the underlying DAG**

The latter point is the crucial problem, since there may be DAGs that are equally likely, and there are in general zillions of DAGs that score comparably well.

Network identification

- **Model Selection:**
  Find a model wíth maximal (or at least exceptionally high) posterior probability P(DAG | Data) and assume that this is the true network topology
- **Model Averaging:**
  Draw a large number of random samples Γ from the distribution P(Γ |Data) and approximate P(edge present | Data) by the sum

$$P(e \mid D) = \sum_{\Gamma \in \text{DAGs}} I(e \in \Gamma) P(\Gamma \mid D) \approx \frac{1}{\#\text{samples}} \sum_{\Gamma \in \text{samples}} I(e \in \Gamma) P(\Gamma \mid D)$$

→ Markov Chain Monte Carlo (MCMC) sampling of directed acyclic graphs

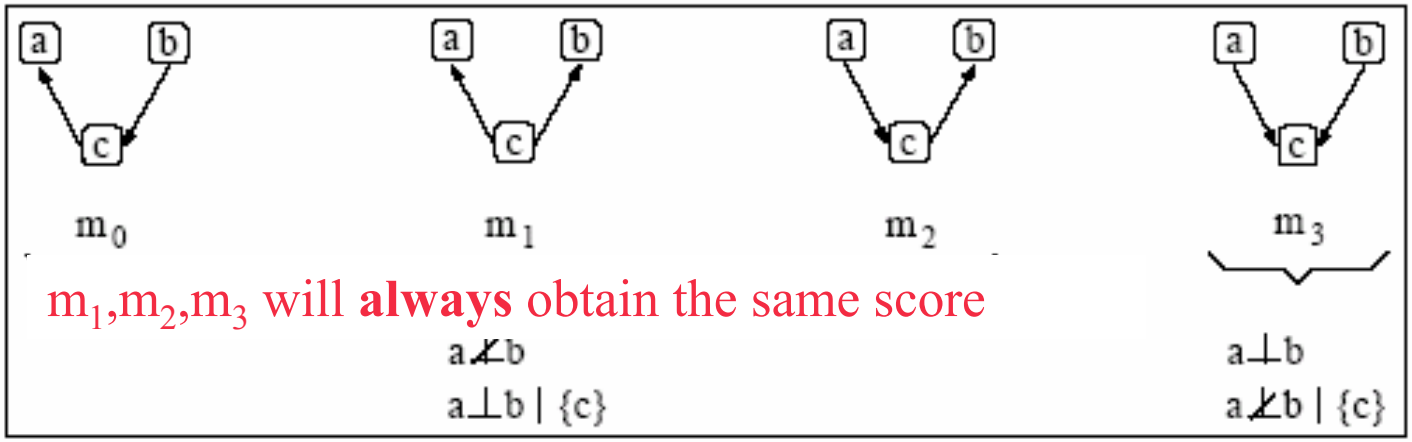
Causality in Bayesian Networks: Likelihood equivalence

Examples of equivalent and non-equivalent graphs

$$P(a,b,c) = \underbrace{P(a\,|\,c)P(c\,|\,b)P(b)}_{m_0} = \underbrace{P(a\,|\,c)P(b\,|\,c)P(c)}_{m_1} = \underbrace{P(c\,|\,a)P(b\,|\,c)\,p(a)}_{m_2}$$

$$P(a,b,c) = \underbrace{P(c\,|\,a,b)P(a)P(b)}_{m_3}$$

Each common distribution *P(a,b,c)*, that can be modelled with BN m_0 can also be modelled with m_1 and m_2 , and vice versa. However there exist distributions P*(a,b,c)*, which can be modeled with BN m_3, but not with m_0 .



Conclusion: Learning from observational data

- **Correct Reconstruction of the complete regulatory network is impossible due to**

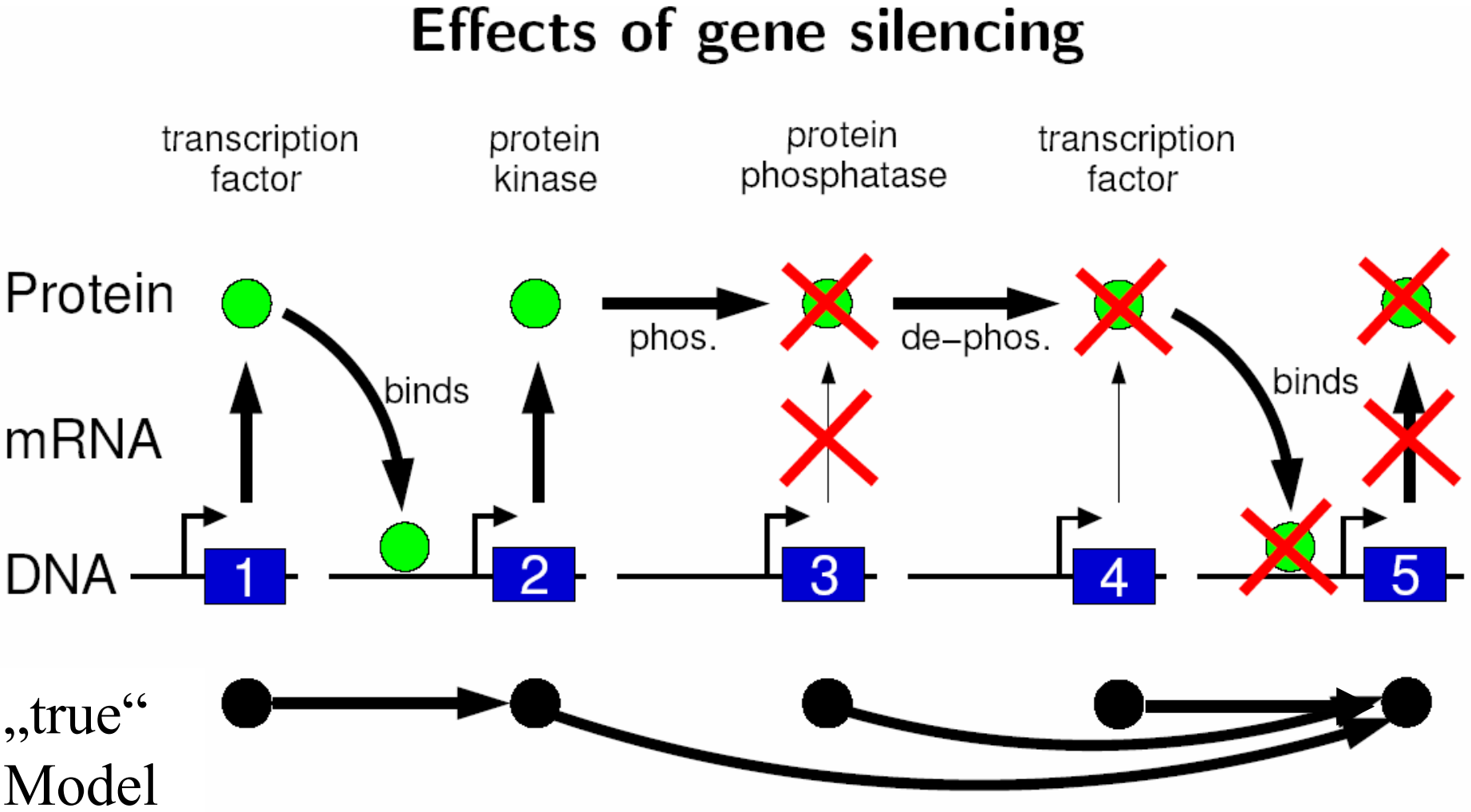
  - Lack of data

  - Measurement error

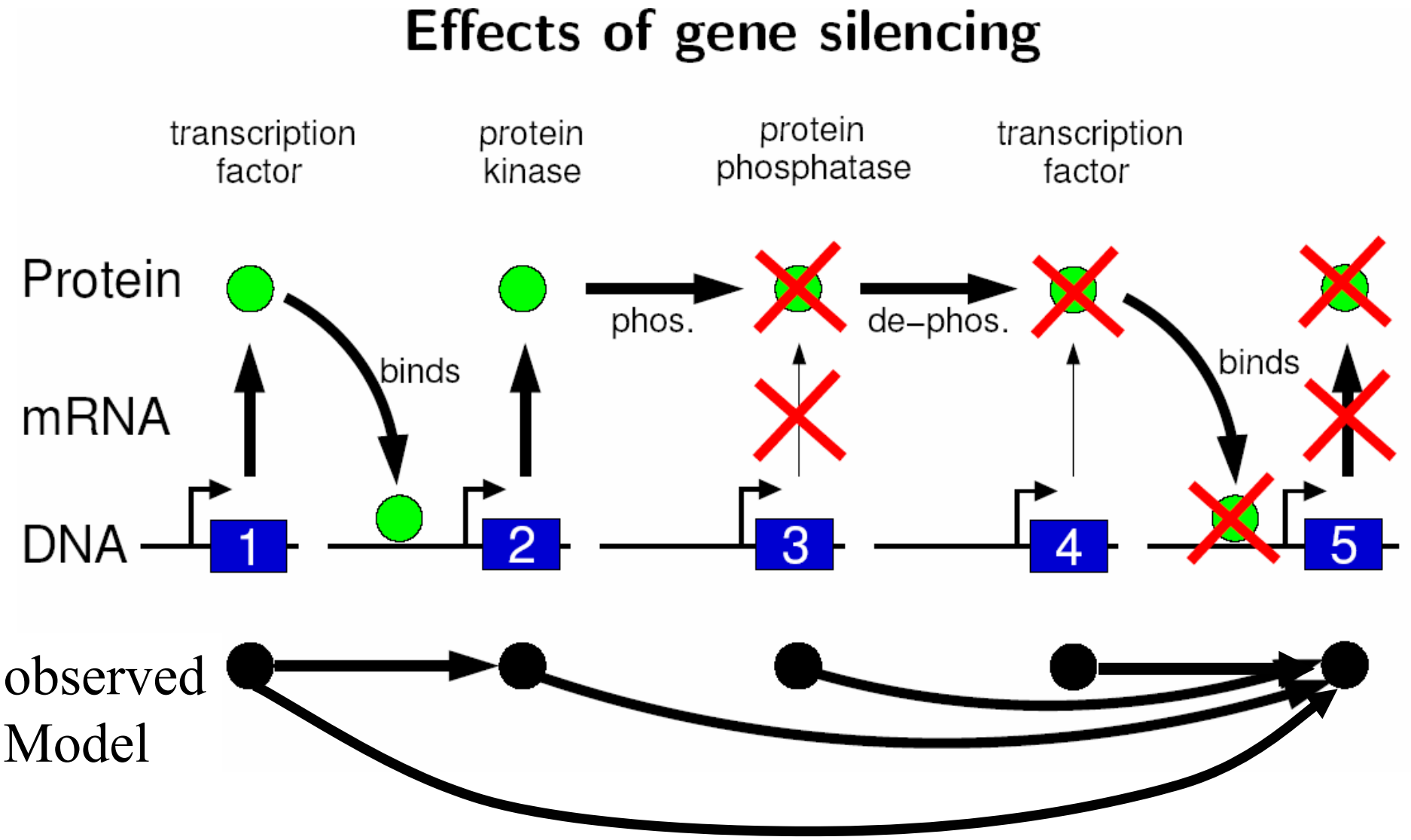
  - Oversimple/wrong model assumptions

- **Reconstruction of regulatory interactions from observational data is merely useful as a screening method.**

Learning from interventional data



Learning from interventional data

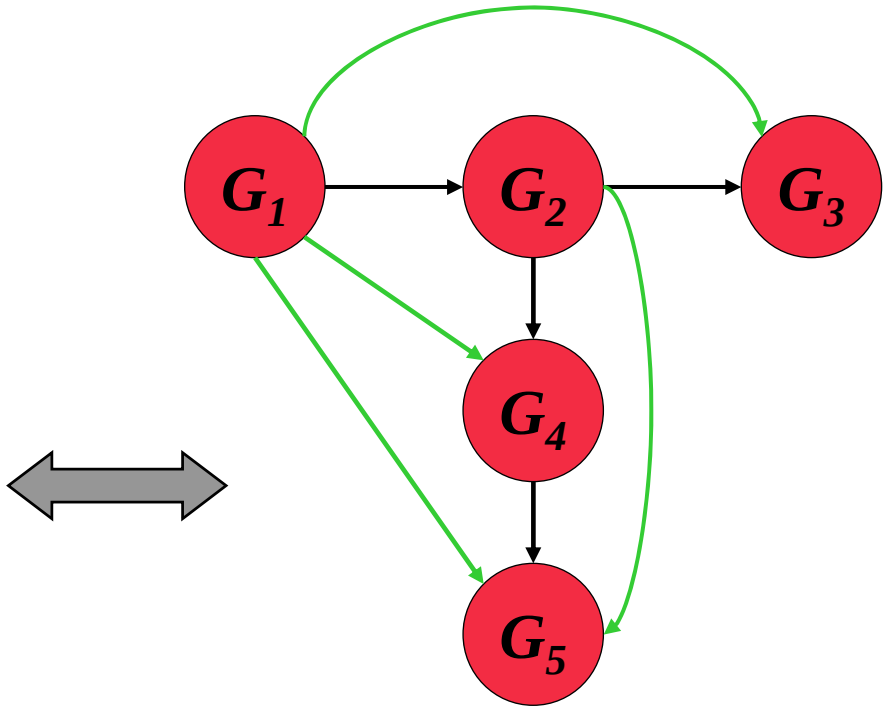


Pruning of Gene interaction Graphs

observations list

Pertur-bation	Effect
G_1	G_2 , G_3 , G_4 , G_5
G_2	G_3 , G_4 , G_5
G_3	-
G_4	G_5
G_5	-

Interaction graph



necessarily direct interactions

optional, possibly indirect interactions

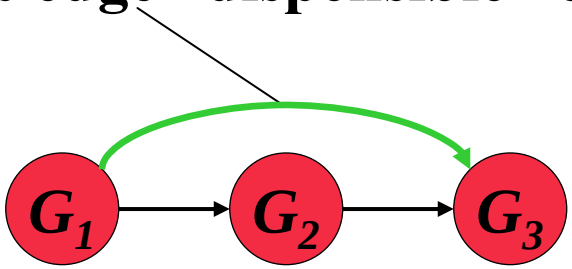
Pruning of Gene interaction Graphs

Given a gene interaction graph, find edges that survive Occam's razor (14th century):

"*non est ponenda pluritas sine necessitate*"

(pluralities ought not to be proposed without necessity)

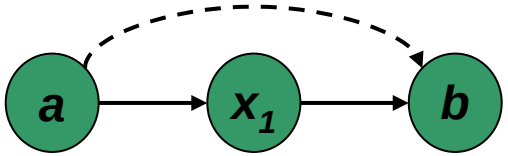
Is this edge "dispensible" or not?



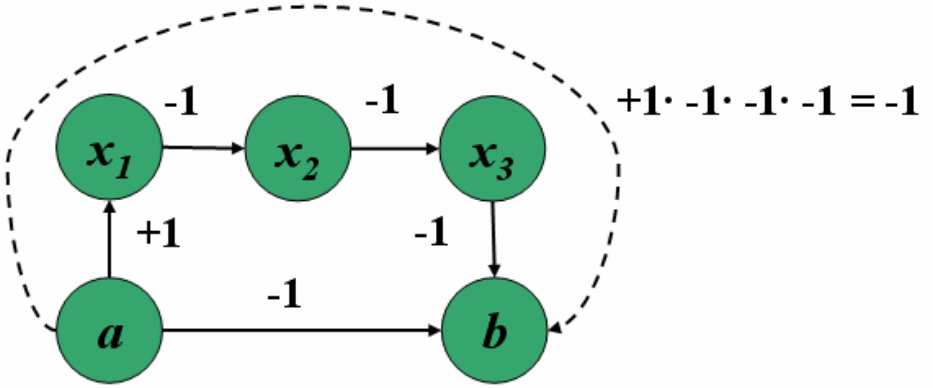
Need for algorithm to define and find
minimal consistent and biologically meaningful graph

Finding non-necessary edges

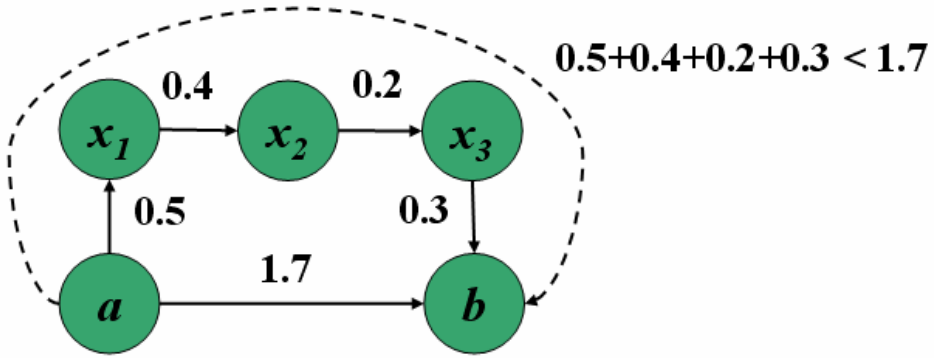
- "Trivial".
  Remove all edges a→b for which there exists a bypass (a longer way from a to b). [Wagner, 2002]



- "Signs".
  Let every edge of the observational graph hava a sign +1 or –1 according to the direction of the regulatory effect. Remove a→b if product of all signs along the path a→... →b equals the sign of the edge a→b [Tringe et al., 2004]

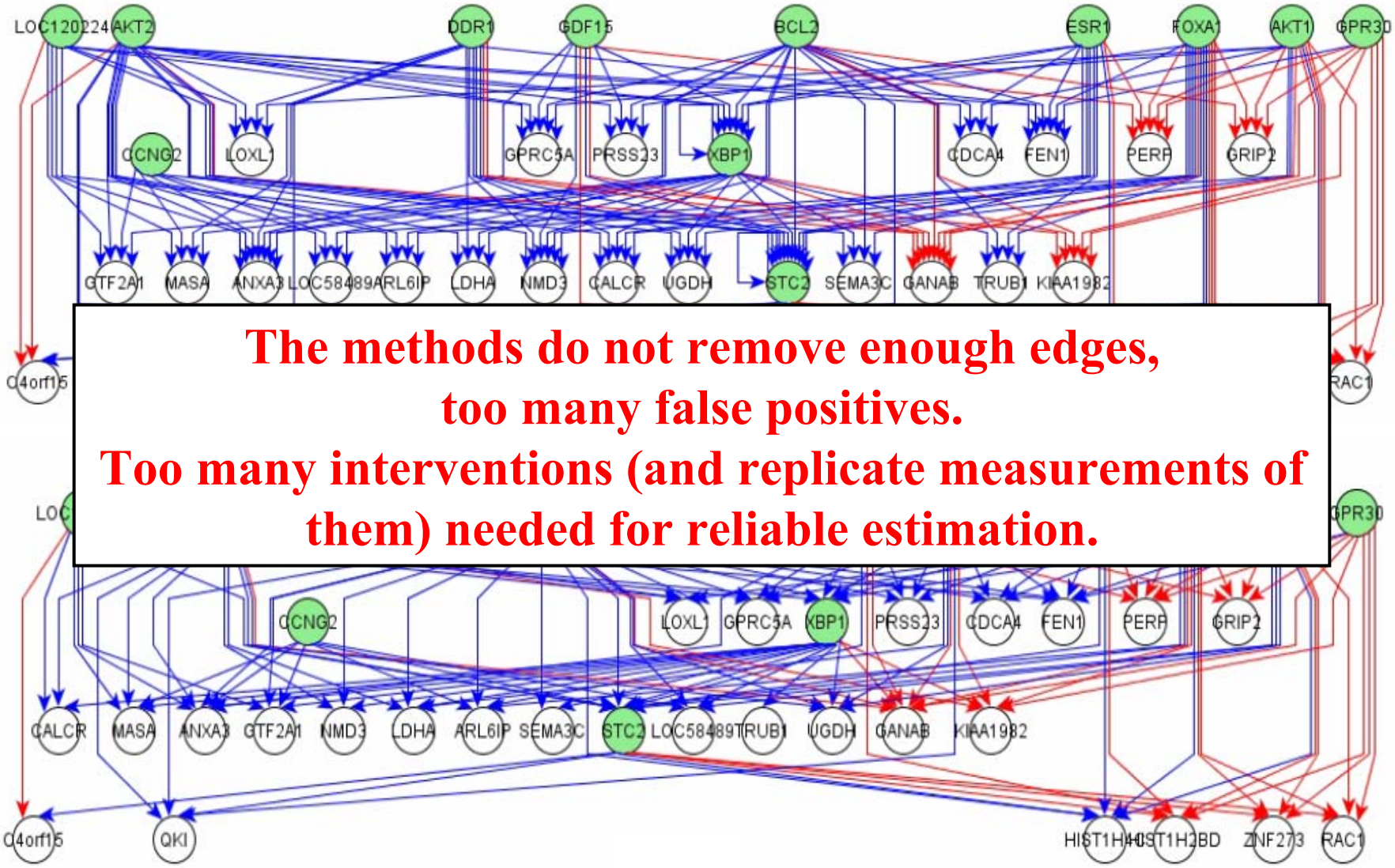


- "Weights".
  Let every edge be weighted with a non-negative number. Edges with low weights are meant to represent edges for which there is strong evidence for a direct regulatory interaction.
  Remove a→b if sum of the weights along the path a→... →b is smaller than the weight of the edge a→b
  [Tresch et al.]



Tresch et al, J.Comp.Biol.

Unpruned vs. Pruned Network

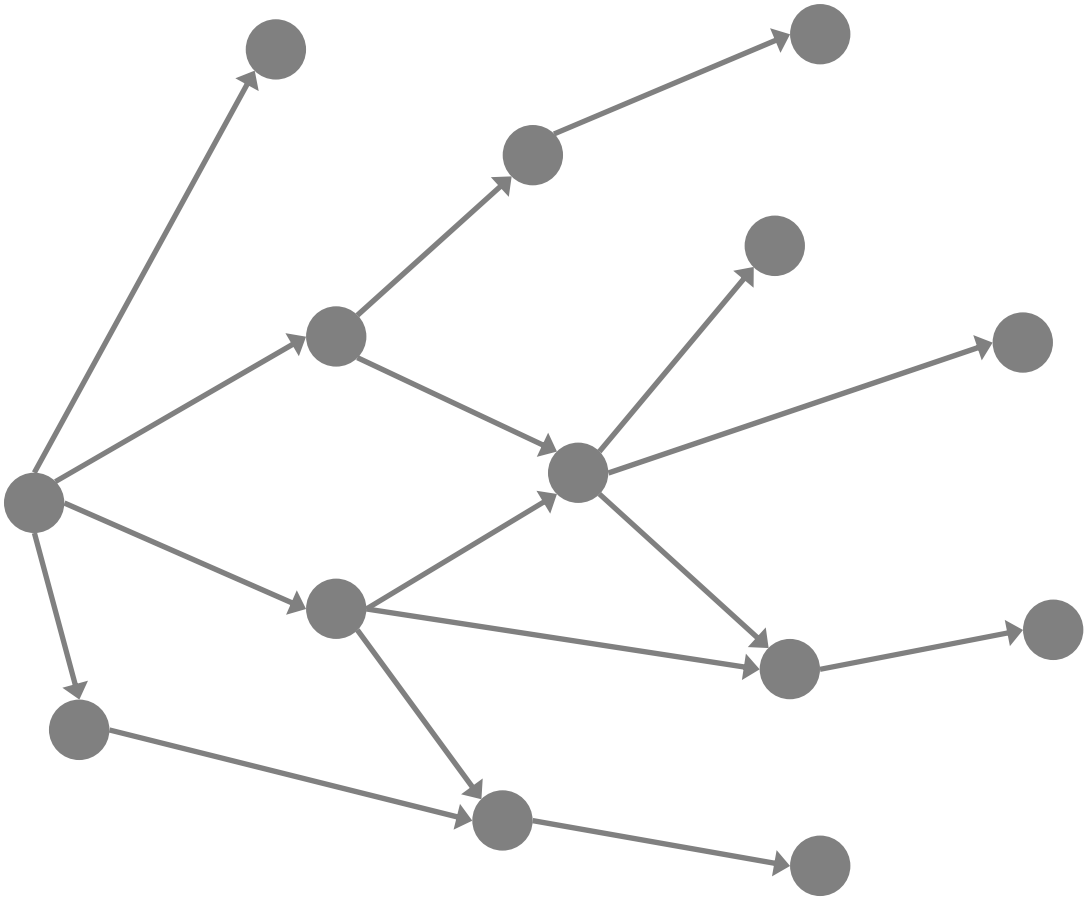


The methods do not remove enough edges, too many false positives.

Too many interventions (and replicate measurements of them) needed for reliable estimation.

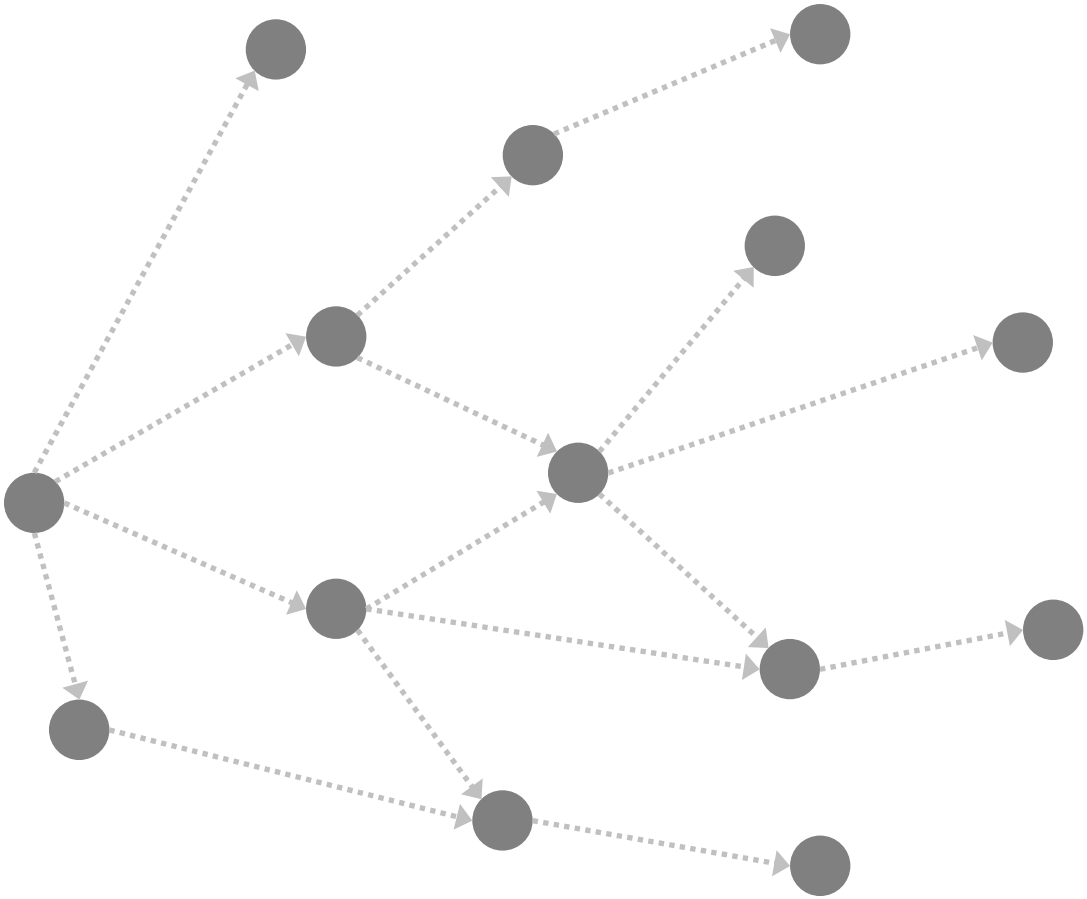
Nested Effects Models

Information flow through a graph of components



Nested Effects Models

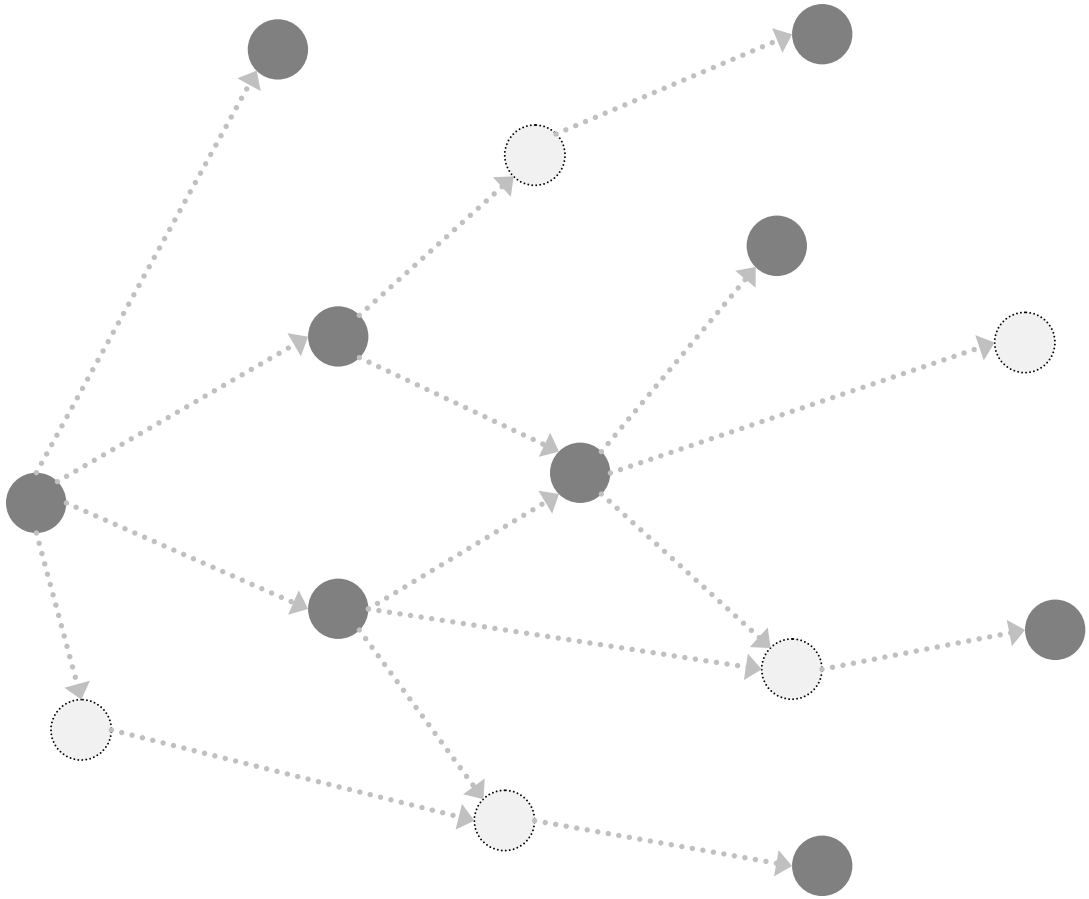
Information flow through a graph of components



Task: Reconstruct the arrows

Nested Effects Models

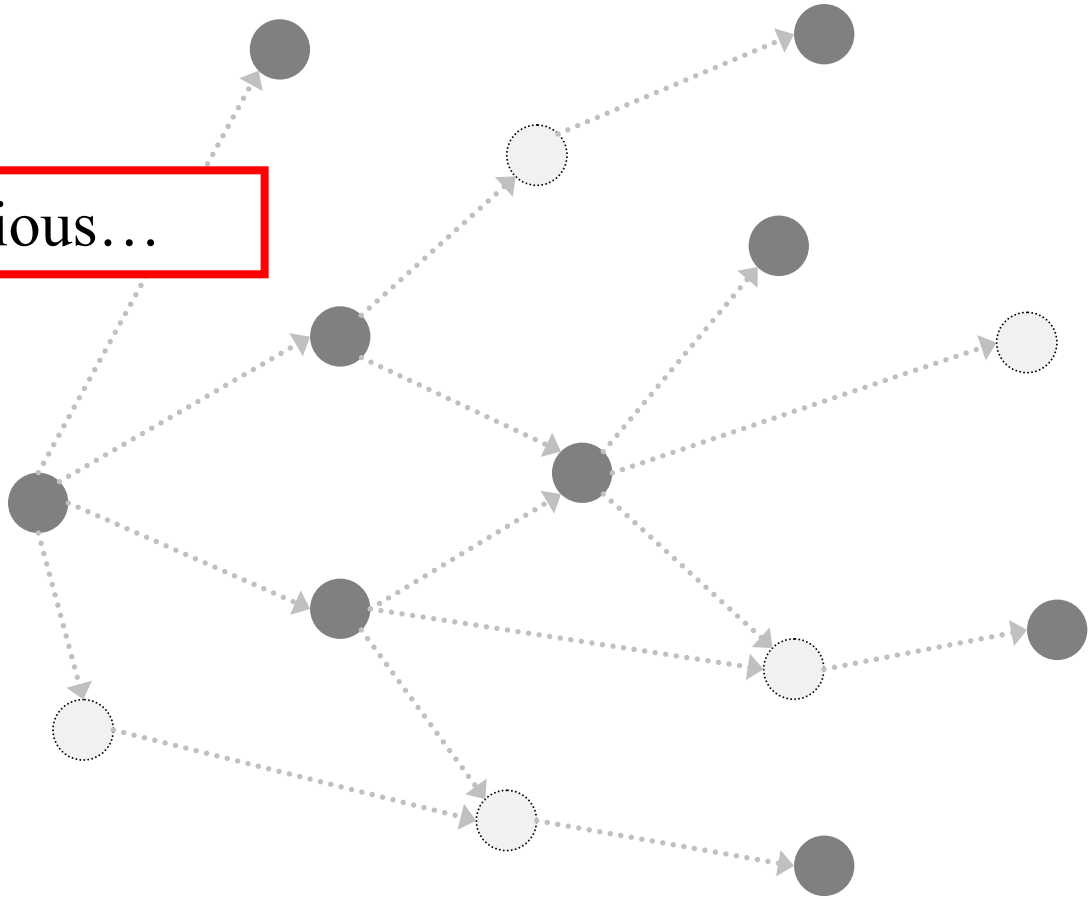
Information flow through a graph of components



Task: Reconstruct the arrows, without measuring all components

Nested Effects Models

Information flow through a graph of components

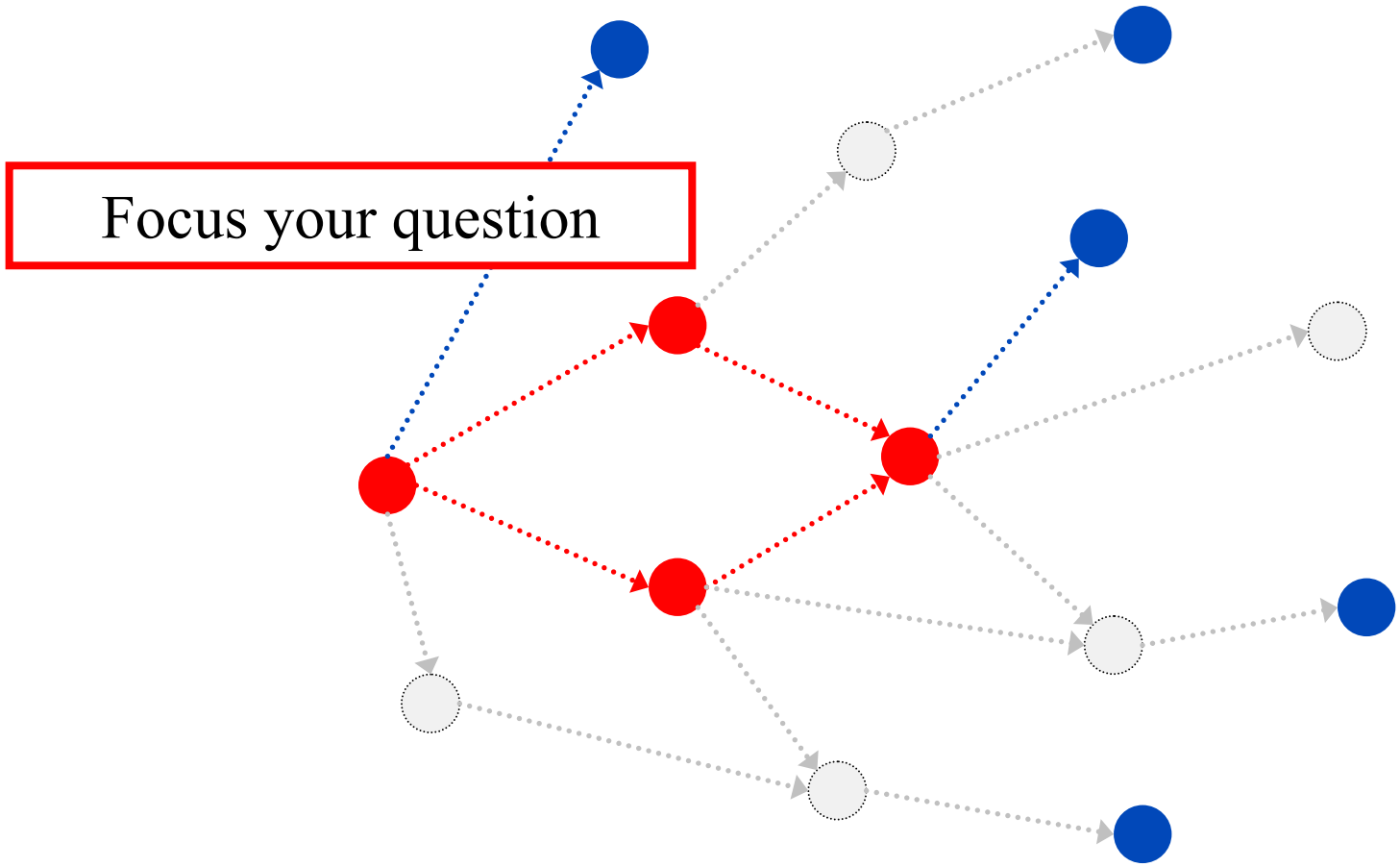


too ambitious…

Task: Reconstruct the arrows, without measuring all components, from noisy observational/interventional data

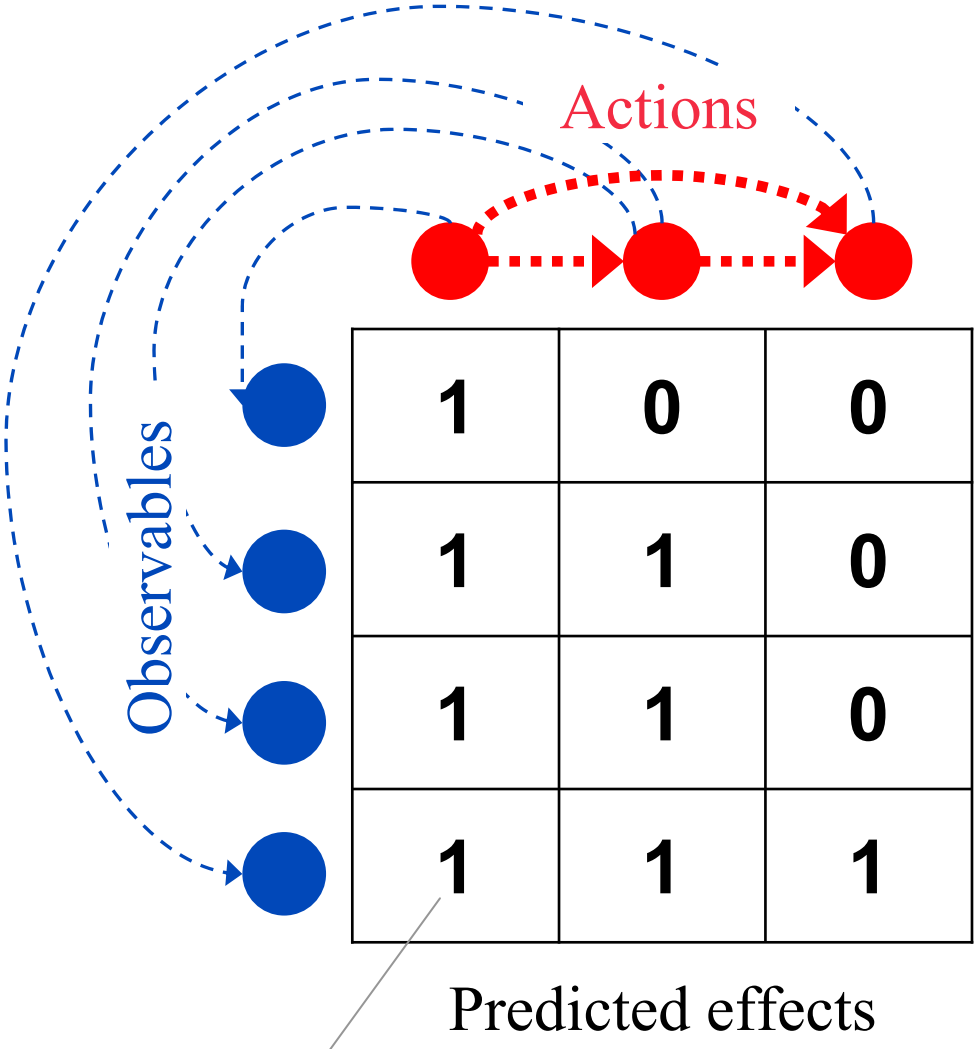
Nested Effects Models

Information flow through a graph of components



Task: Reconstruct the wiring of a **small subset of components**, perform **interventions on these components,** make use of **all observable components**.

Definition of Nested Effects Models

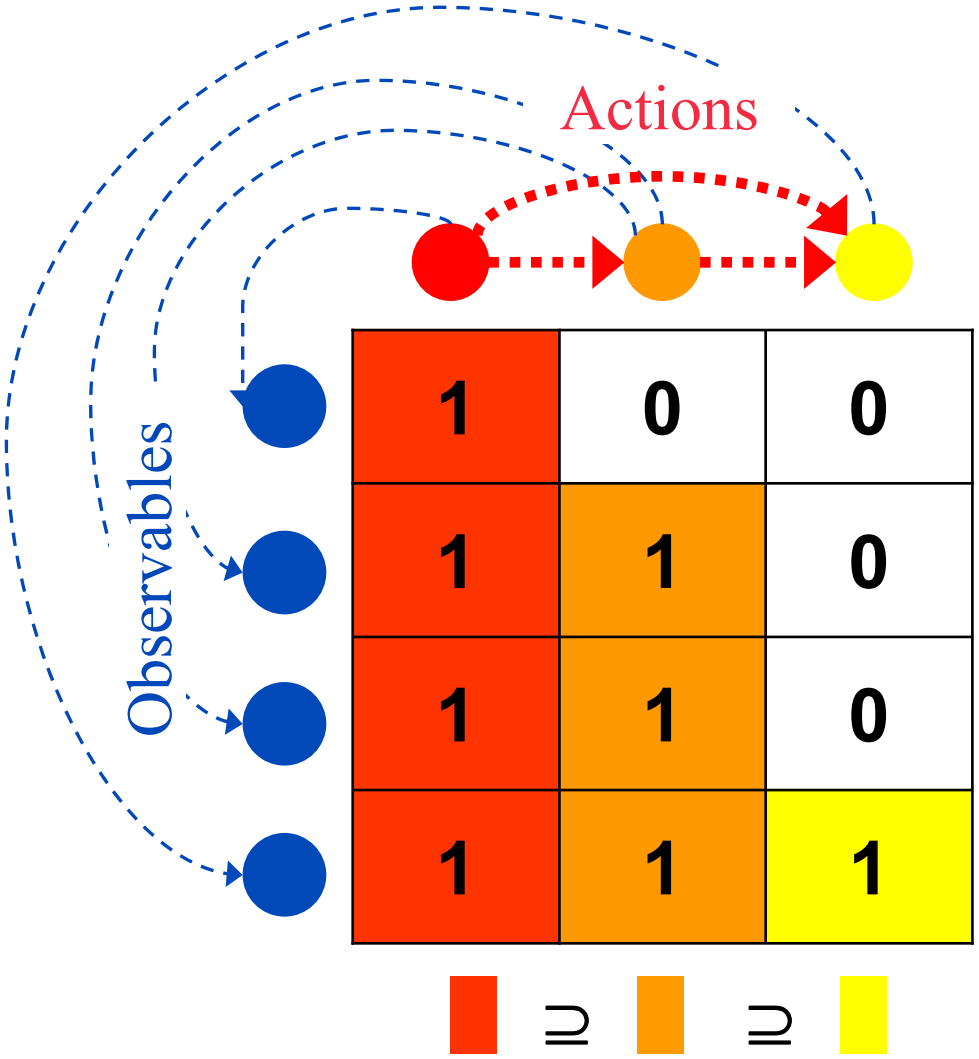


Actions graph:
Adjacency matrix **Γ**

Effects graph:
Adjacency matrix **Θ**

Assumption:
Each observable is linked to exactly one action

Definition of Nested Effects Models



Why „nested“ ?

If the actions graph is transitively closed, then the effects are nested in the sense that

$$a \rightarrow b$$

implies

$$\text{effects}(a) \supseteq \text{effects}(b)$$

Likelihood of Nested Effects Models

$$\log\left[P\left(\begin{array}{c} s \text{ IS differentially expressed} \\ \text{when perturbing } a \end{array} \right) \Big/ P\left(\begin{array}{c} s \text{ is NOT differentially expressed} \\ \text{when perturbing } a \end{array} \right) \right] = R_{s,a}$$

Actions-graph Γ

Effects Graph Θ

1	0	0
1	1	0
1	1	0
1	1	1

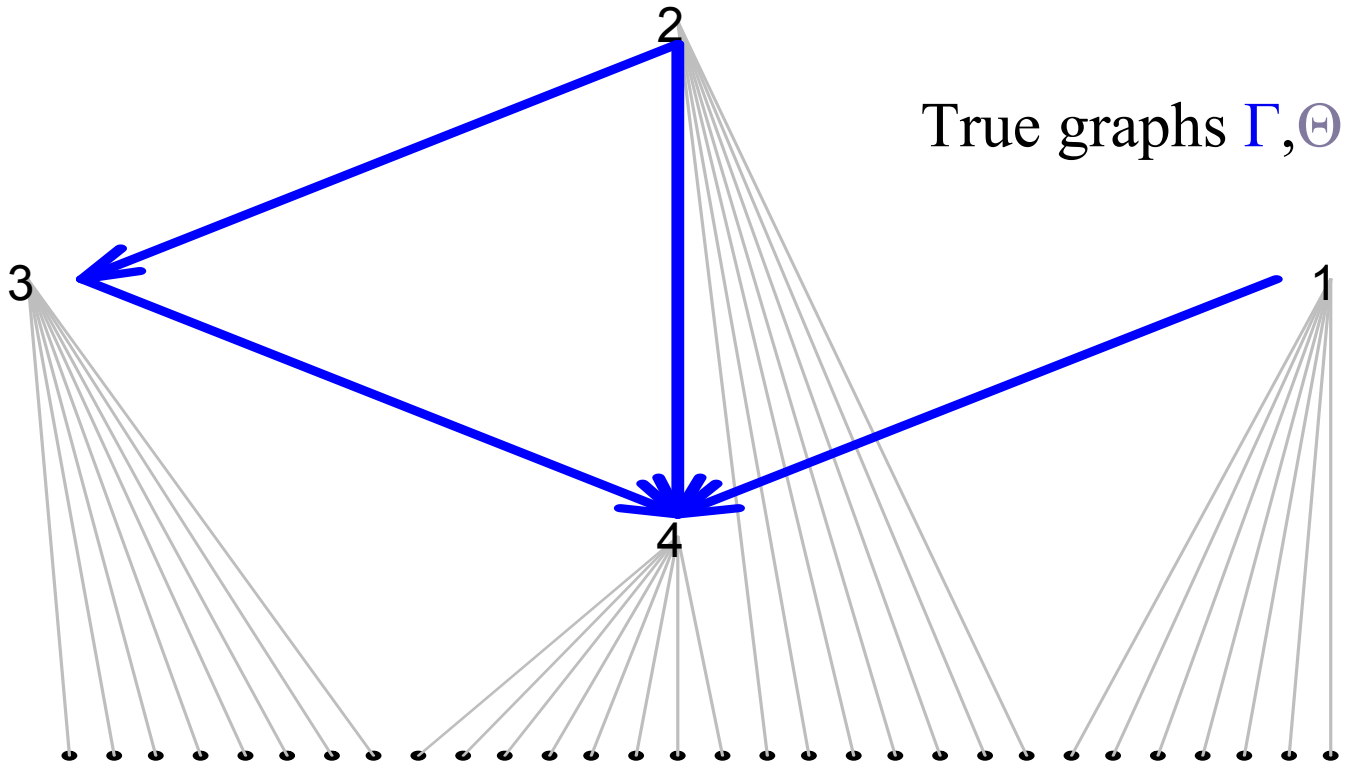
Predicted effects

0.8	-2.5	-0.2
3.1	0.9	-1.3
0.1	2.4	0.1
-0.7	0.4	1.7

Measured effects **R**

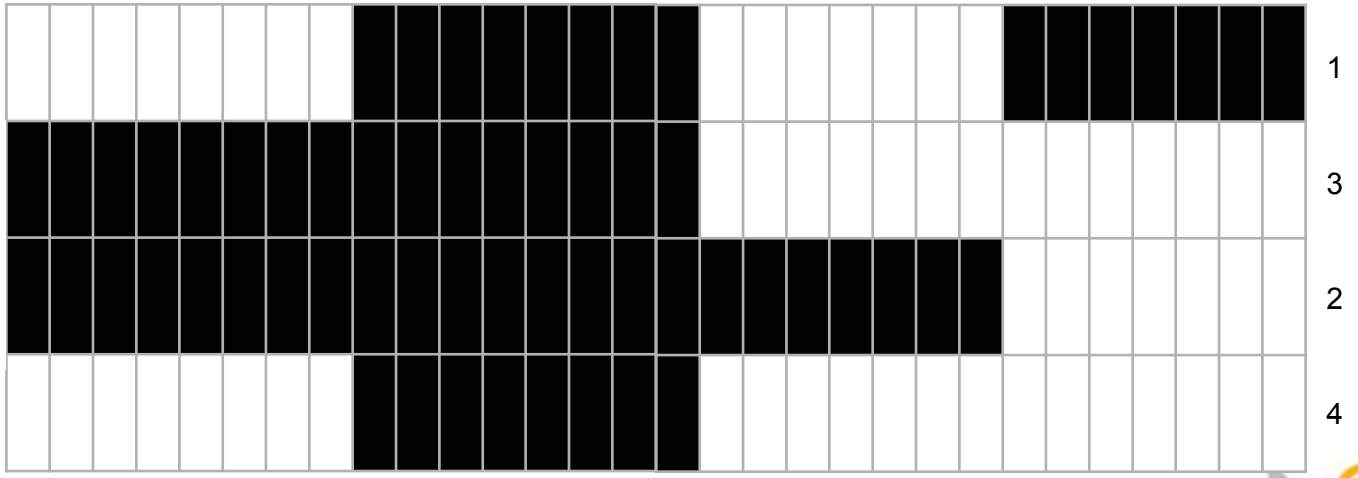
$$\log P(\text{Data} \mid NEM \text{ model}) = tr(\Gamma\Theta R) + const$$

Results on simulated Data



True graphs Γ,Θ

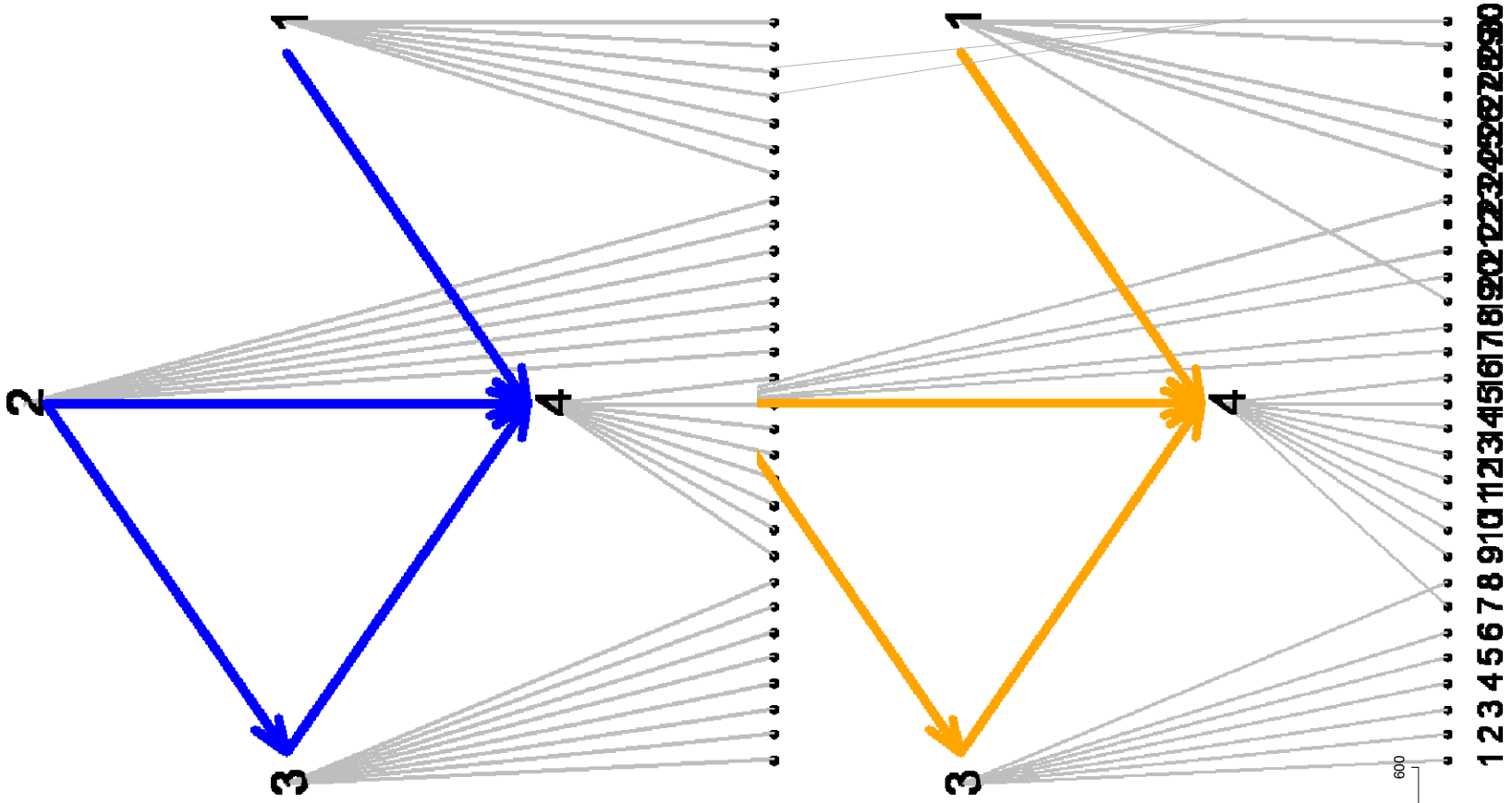
ideal measure-ments (ΓΘ)



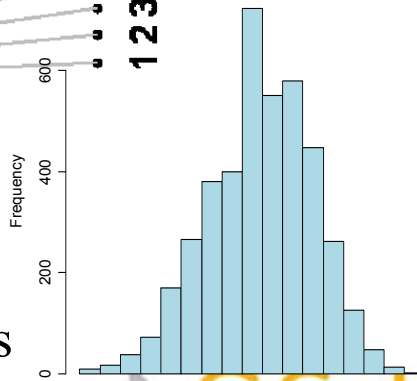
Results on simulated Data

True graph

Estimated graph



12 edges, 2^{12}=4096 action graphs, ~ 4seconds



Distribution of the likelihoods

Pathways from RNAi data – an example

Response to microbial challenge

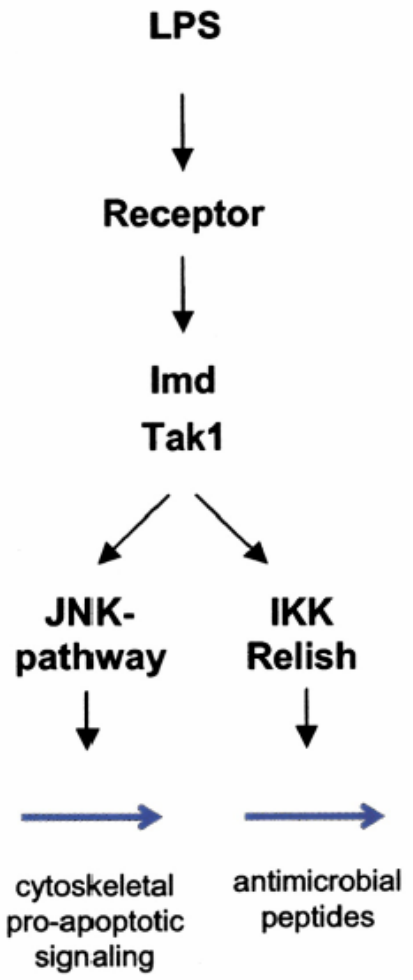
(Boutros *et al.*, Dev Cell, 2002)

Columns: silenced genes.

Rows: effects on other genes.

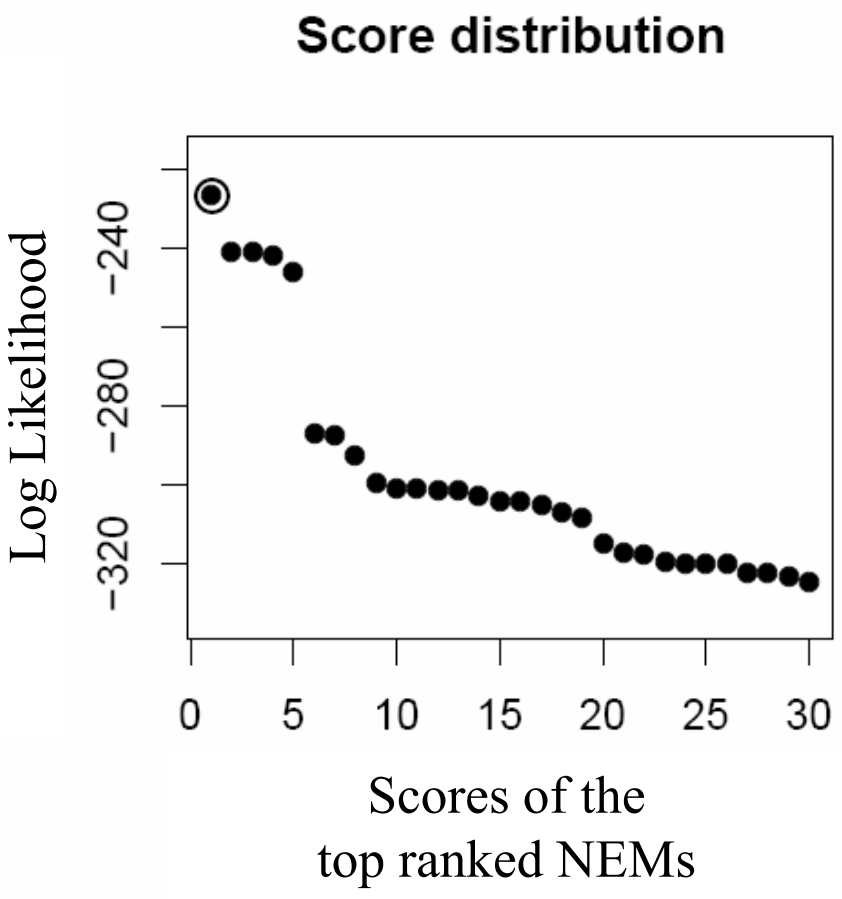
Results:

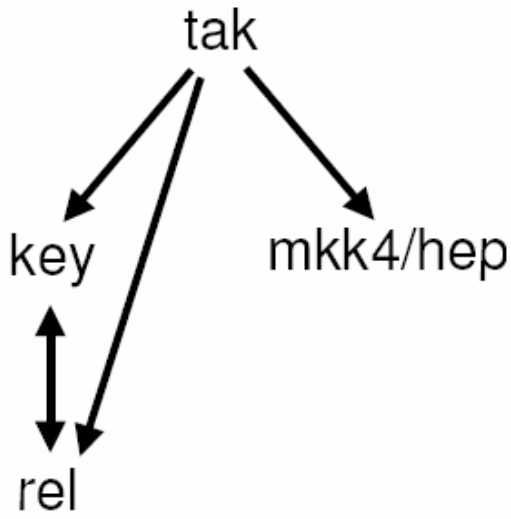
1. Silencing tak1 reduces expression of all LPS-inducible transcripts.
2. Silencing rel (key) or mkk4/hep reduces expression of separate sets of induced transcripts.

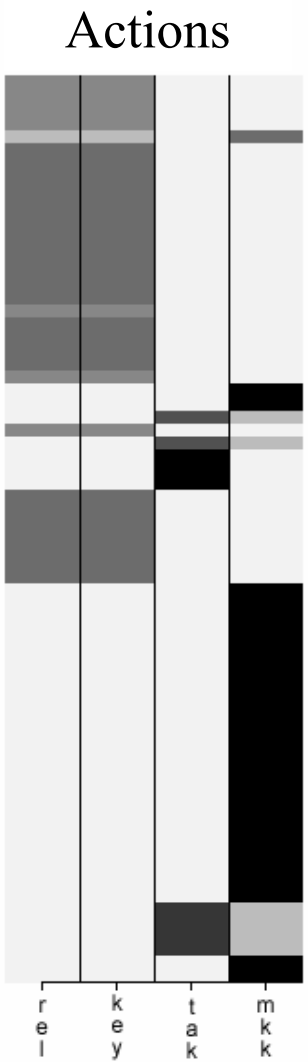


Figures from (Boutros *et al.*, 2002)

Application to Drosophila data







Networks in R/Bioconductor

- graph: basic class definitions and functionality
- RBGL: interface to graph algorithms (e.g. shortest path, connectivity)
- Rgraphviz: Different layout algorithms. Node plotting, line type, colour etc. can be controlled by the user.
- dynamicGraph: visualize interactive Graphs with TclTK.
- GeneTS: Estimate GGMs from Microarray Data.
- Nessy, nem: Implementation and estimation of the Nested Effects Model

Pathways and Visualization Tools

- Some Pathway Databases:
  - KEGG (http://www.genome.jp/kegg)
  - TRANSPATH (http://www.biobase.de)
  - Biocarta (http://www.biocarta.com)
  - Reactome (http://www.reactome.org)
  - HumanCyc (http://humancyc.org)
  - Signal Transduktion Knowledge Environment (http://stke.sciencemag.org)

- Software tools
  - GeneMAPP (www.genemapp.org)
  - GoMiner (http://discover.nci.nih.gov/gominer)
  - Bioconductor/Graphviz (http://www.bioconductor.org)
  - Cytoscape (http://www.cytoscape.org)

Acknowledgements

- Florian Markowetz (many slides, bibliography)
- Tim Beissbarth
  Andreas Buness
  Markus Ruschhaupt
  Holger Fröhlich
- Mark Fellmann
  Holger Sültmann

Thank you!