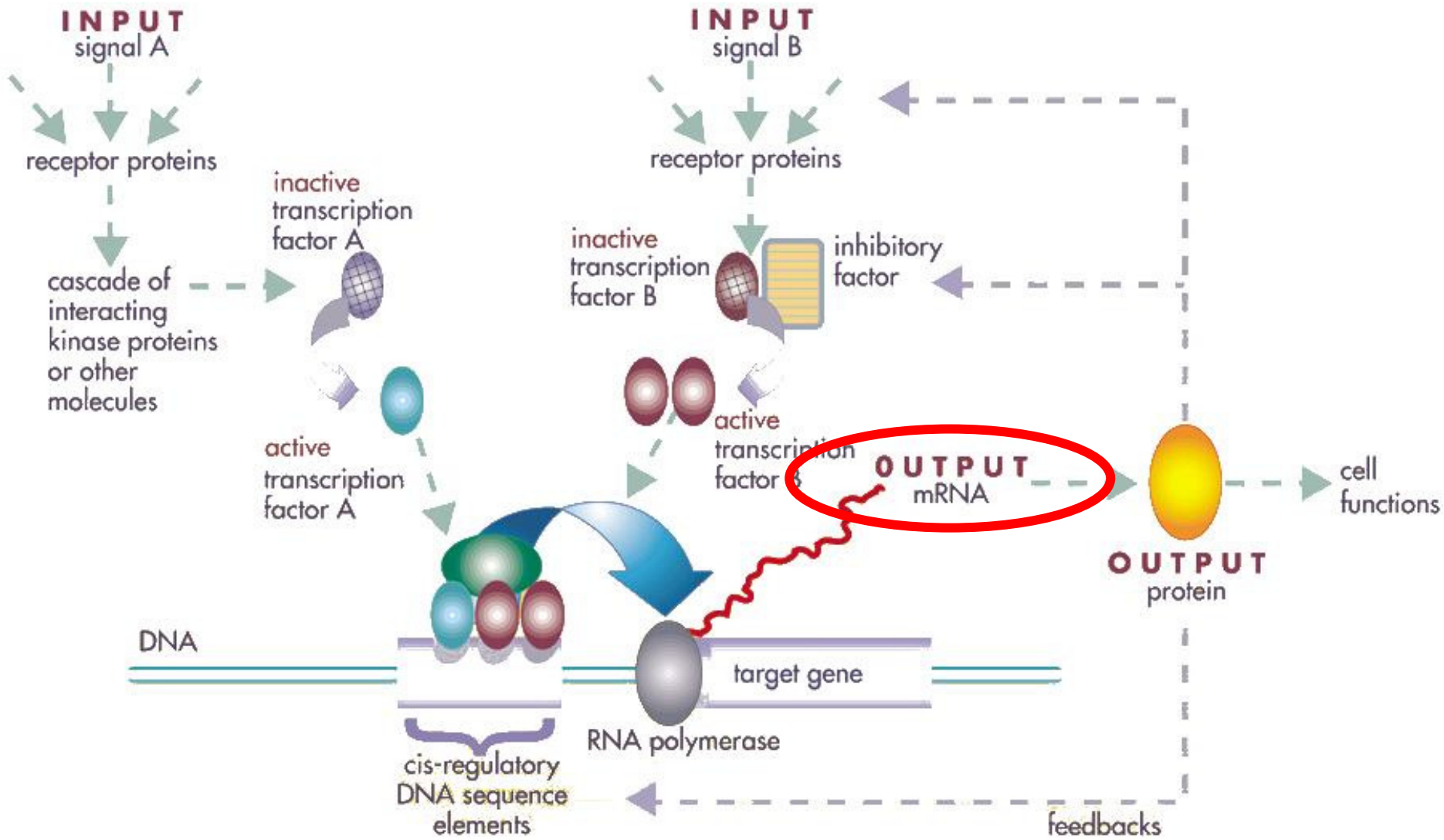
Computational Inference of Cellular Networks from Microarray Data

Achim Tresch

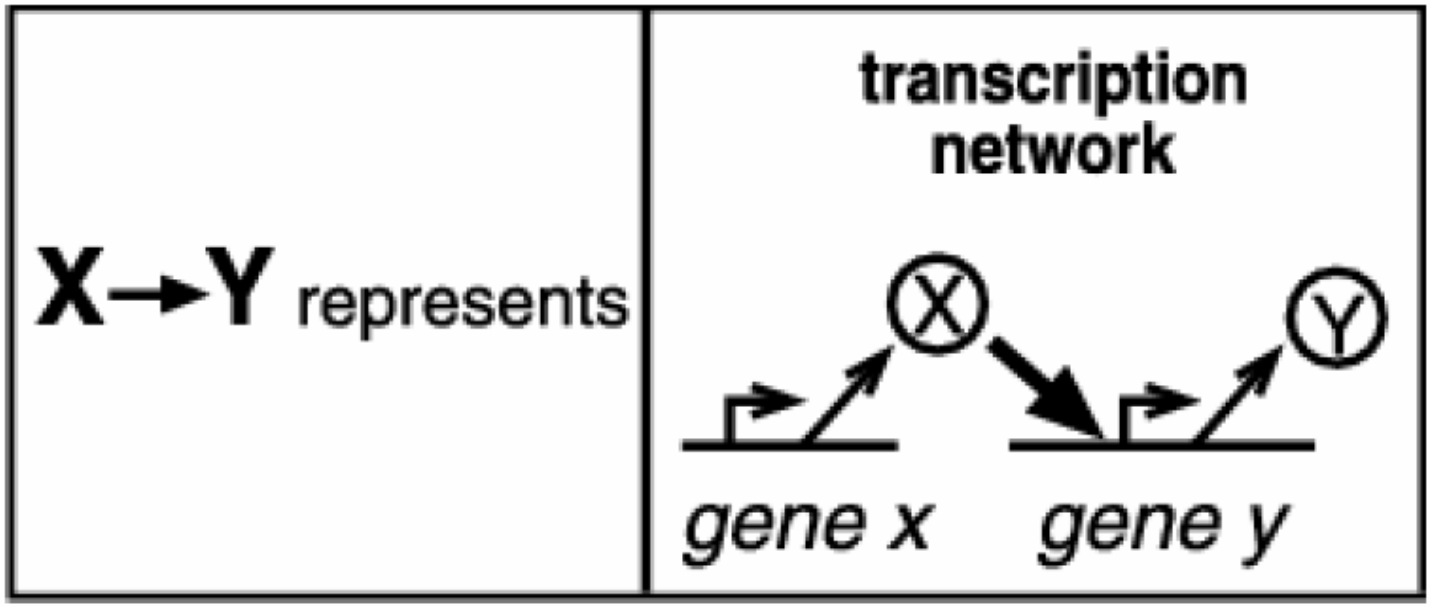
Topics

- **Biological networks vs. Network Models**
- **Learning Networks from non-interventional data:**
  - Gaussian Graphical Models
  - Bayesian Networks
- **Learning from interventional data:**
  - Pruning
  - Signal-Effects Model

Which biological Network?



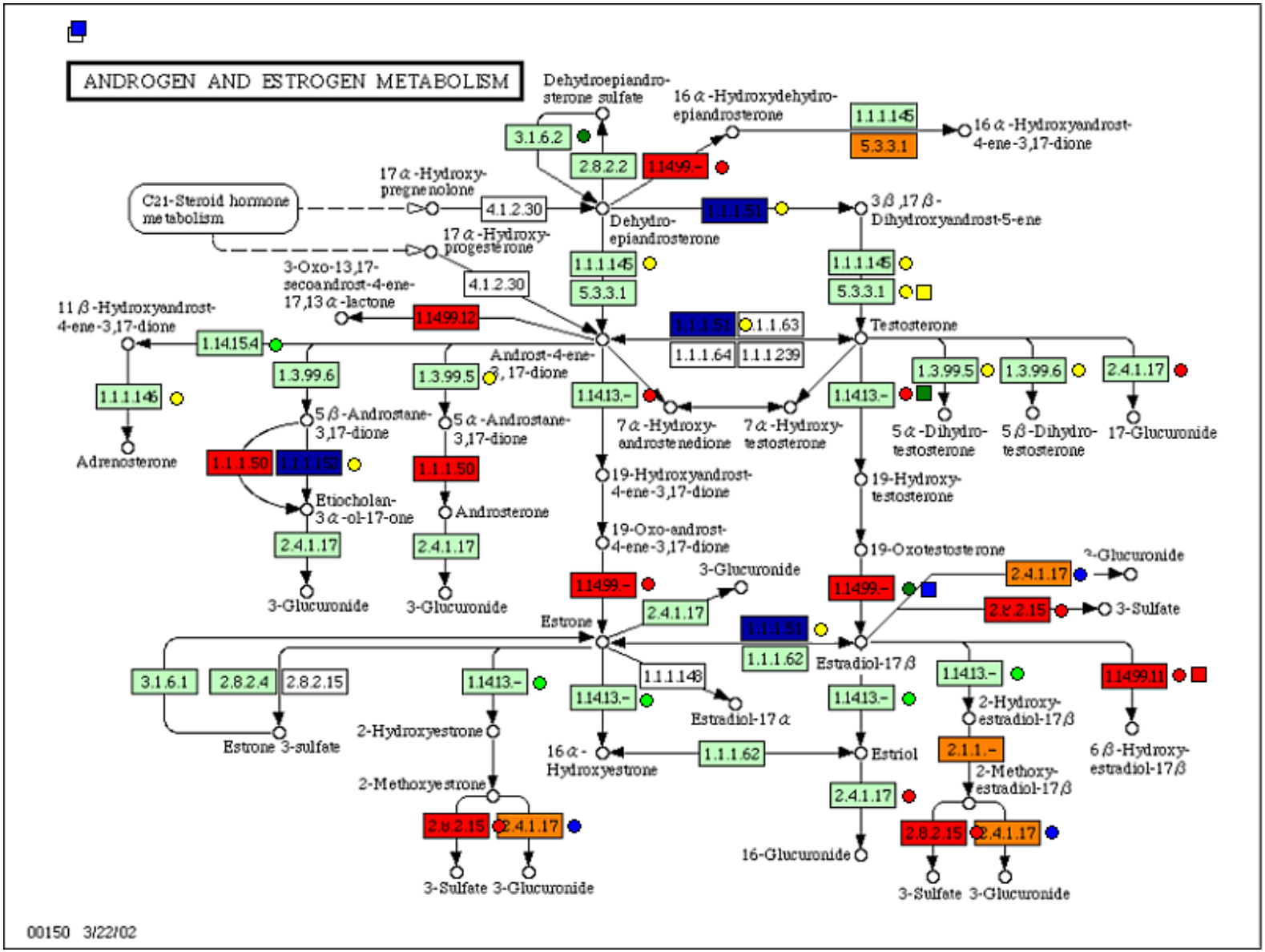
Transcription Networks



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

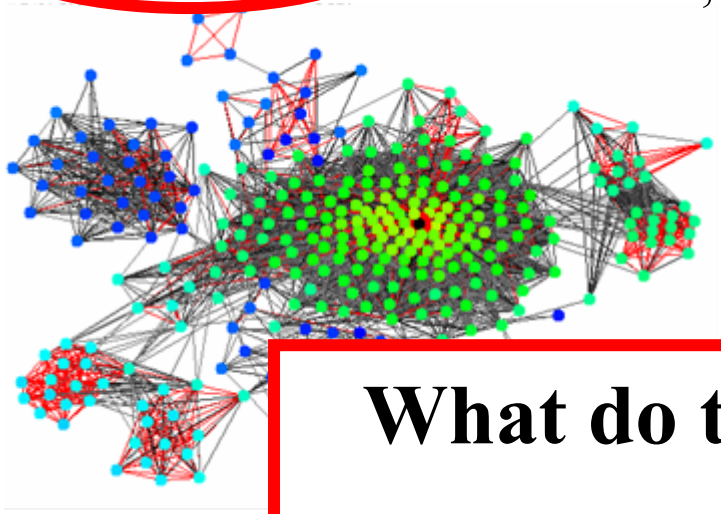
- thamoxifen
- flutamide
- anastrazole
- masterlone

From KEGG database

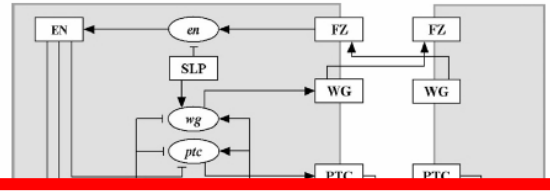
Which Network Model?

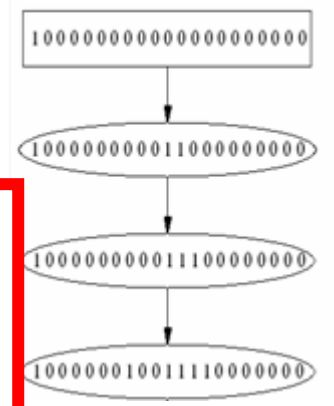
qualitative

Best suited for high dimensional, noisy data



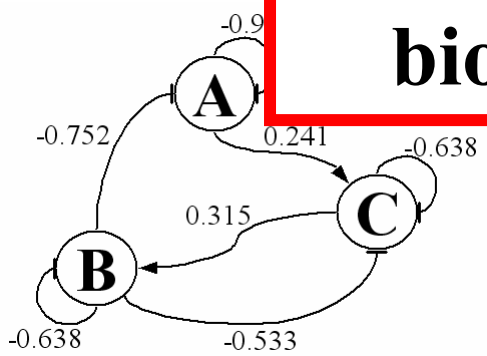
semiquantitative





What do the arrows mean?

Can they be ascribed 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

Qulitative Network Models

Possible Models include

- **Correlation Graphs**
- **Gaussian Graphical Models**
- **Bayesian Networks**

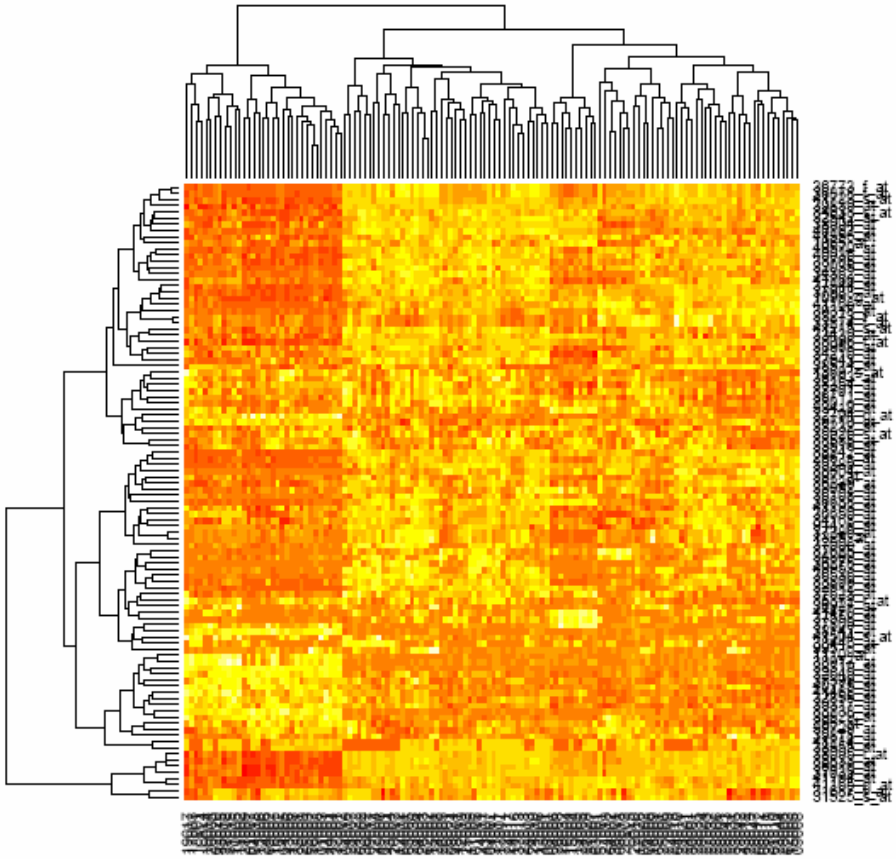
However: Correct Reconstruction of the complete regulatory network is impossible due to

- **Lack of data**
- **Measurement error**
- **Oversimple/wrong model assumptions**

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

(Edwards Deming, George Box)

Clustering by coexpression



Assumption:

Coexpression $\sim$ coregulation

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

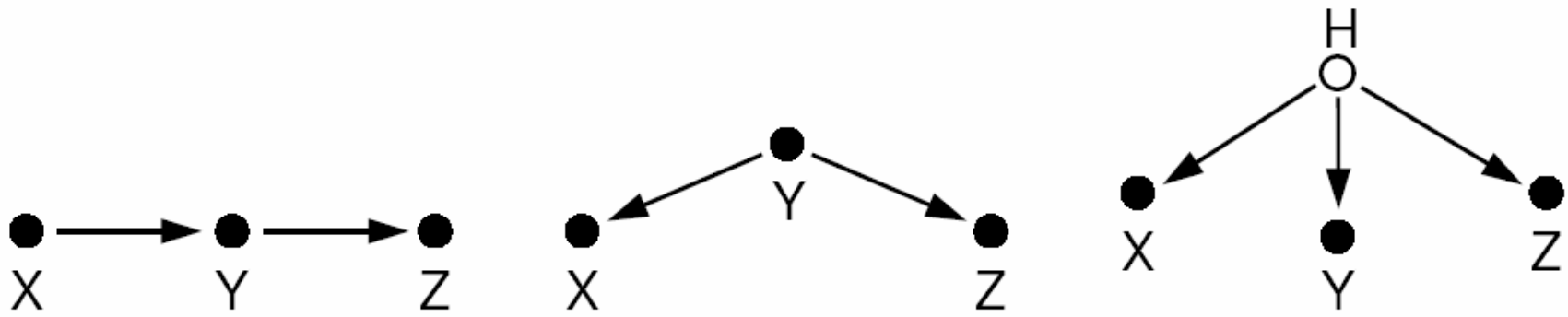
Correlation Graphs

- **An expression profile is a collection of expression vectors**
  $$\{ X_g = (X_{g,s})_{s \in \text{samples}} , g \in \text{Genes} \}$$
- **Correlation graph: Depict genes as vertices of a graph and draw an undirected edge (i, j) 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 accurately estimated even if p » N.**
- **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 dependencies**

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



Possible remedies:

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

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

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 .**

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

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

where $\mathbf{X}_{\text{rest}} = \mathbf{X} \backslash \{X_i, X_j\}$.

Gaussian Graphical Models (GGM)

If we assume that the common expression distribution of all genes follows a multivariate Gaussian distribution (which is of course never the case), conditional independence can be assessed as follows:

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

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

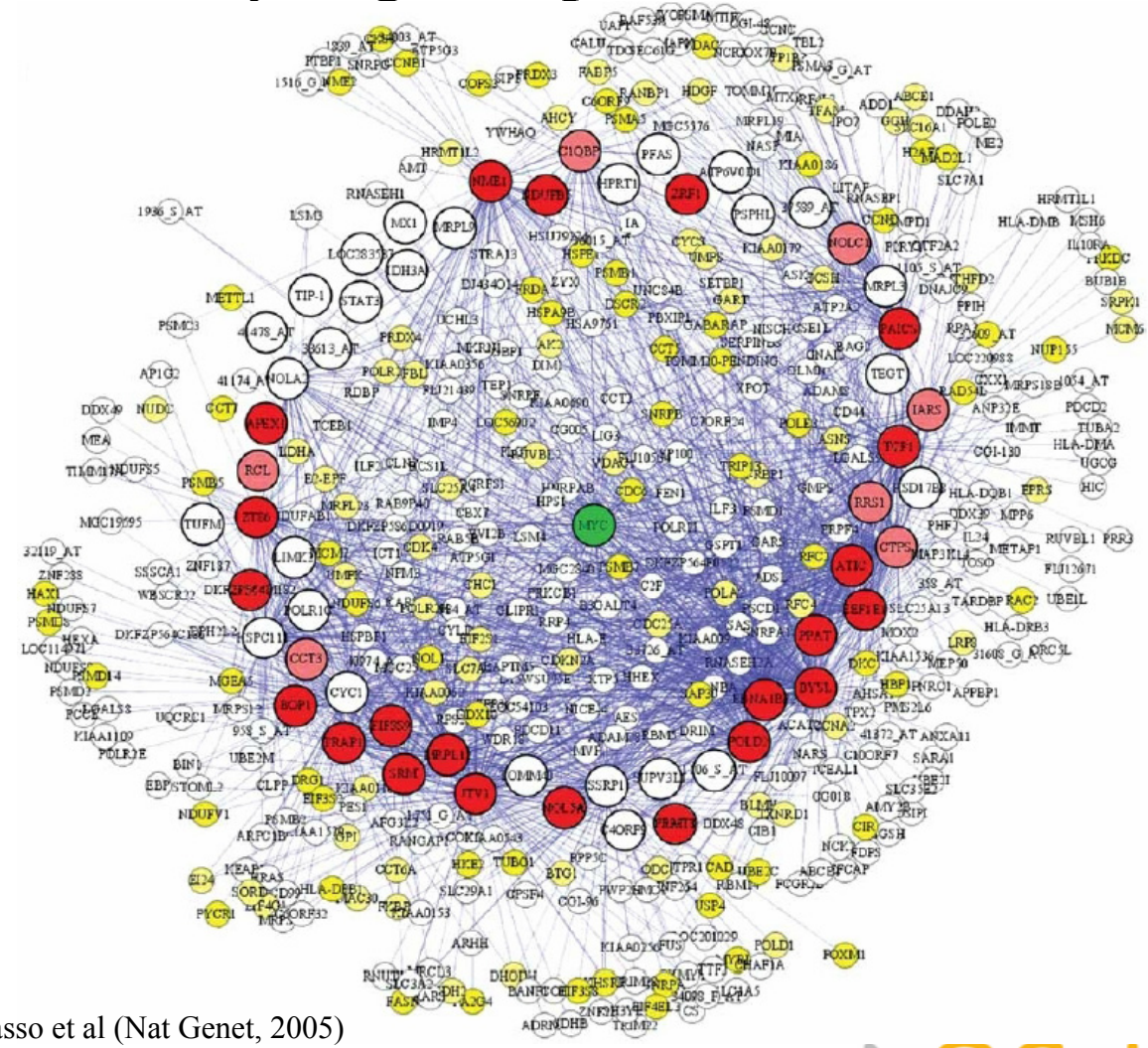
2. Invert $\hat{\Sigma}$ to obtain an estimate $\hat{K}$ of the precision matrix K.
3. Employ statistical tests [56] to decide, which entries in $\hat{K}$ are significantly different from zero.

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 failed to accurately reconstruct networks, even if they were of very moderate size (~20)

Children depend on parents

The DAG defines families.

Relationships are further characterized by local probability distributions:

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

J = Season

pa(J) = ∅

pa(S) = {J}

pa(R) = {J}

S = Sprinkler

R = Rain

pa(N) = {S,R}

W = Wet Lawn

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

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

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

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)

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

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

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

$$P(R|J) = \begin{cases} J\backslash R & \text{rain} & \text{norain} \\ \text{spring} & 0.3 & 0.7 \\ \text{summer} & 0.4 & 0.6 \\ \text{autumn} & 0.2 & 0.8 \\ \text{winter} & 0.1 & 0.9 \end{cases}$$

$$P(N|S,R) = \begin{cases} (S,R)\backslash N & \text{wet} & \text{dry} \\ (\text{on,rain}) & 1 & 0 \\ (\text{on,norain}) & 1 & 0 \\ (\text{off,rain}) & 0.9 & 0.1 \\ (\text{off,norain}) & 0.01 & 0.99 \end{cases}$$

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

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) for the underlying DAG**

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

Network identification

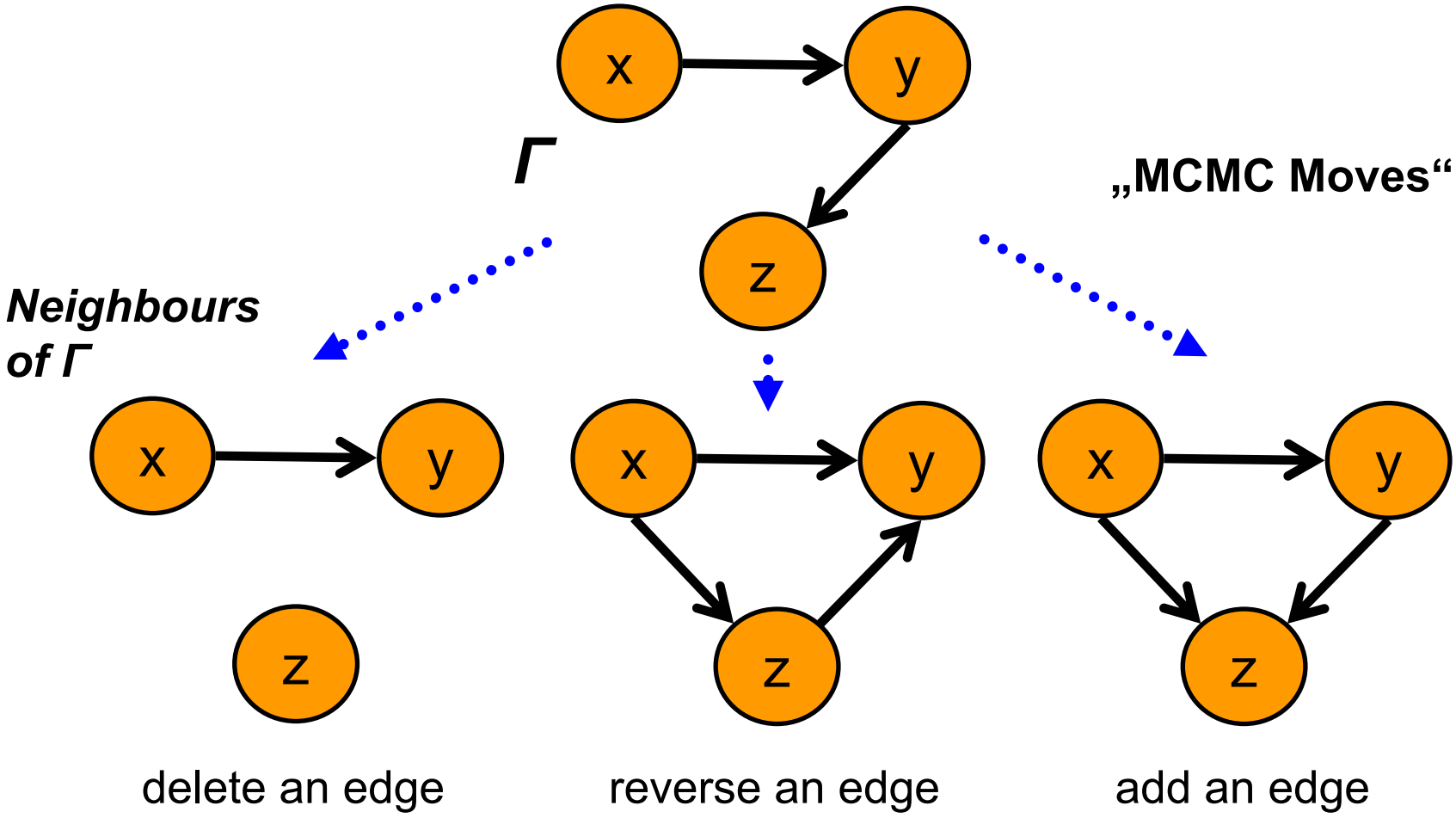
- **Model Selection:**
  Find a model wíth maximal (or at least exceptionally high) posterior probability P(Γ|Data) and assume that this ist the true network topology

- **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

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

MCMC sampling of directed acyclic graphs

Wander through the space of all possible topologies



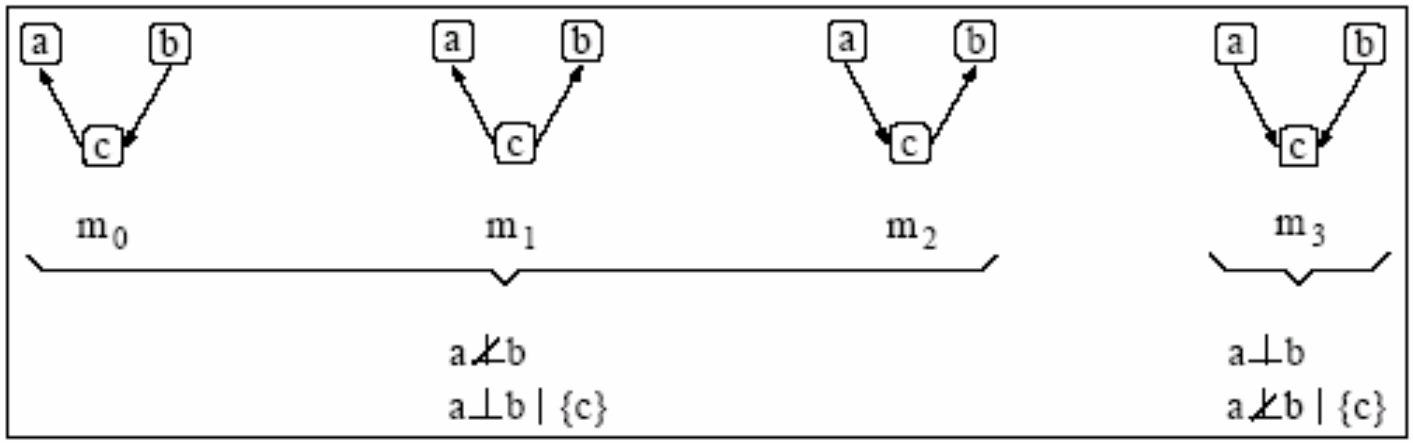
Causality in Bayesian Networks: Likelihood equivalence

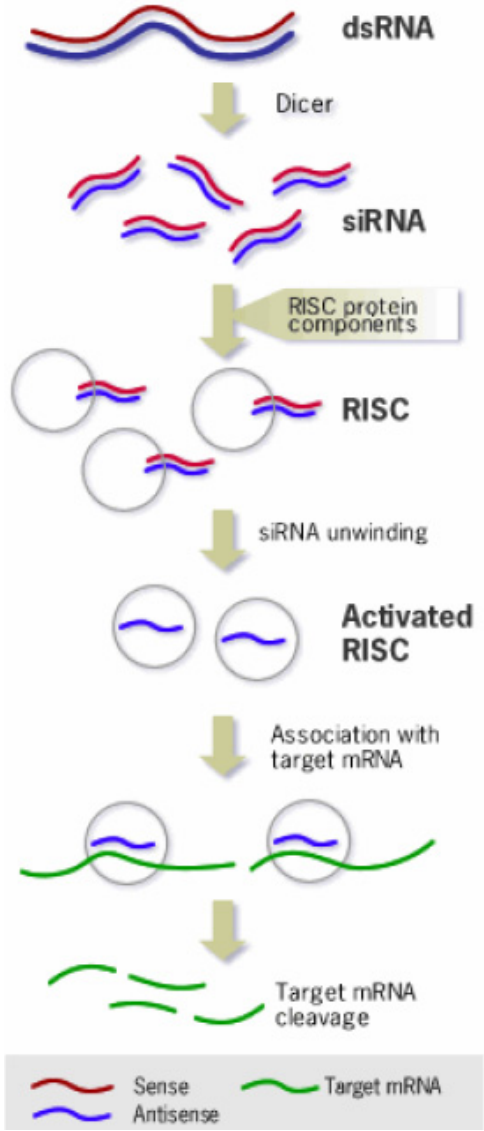
Examples of equivalent and non-equivalent graphs

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

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

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 .





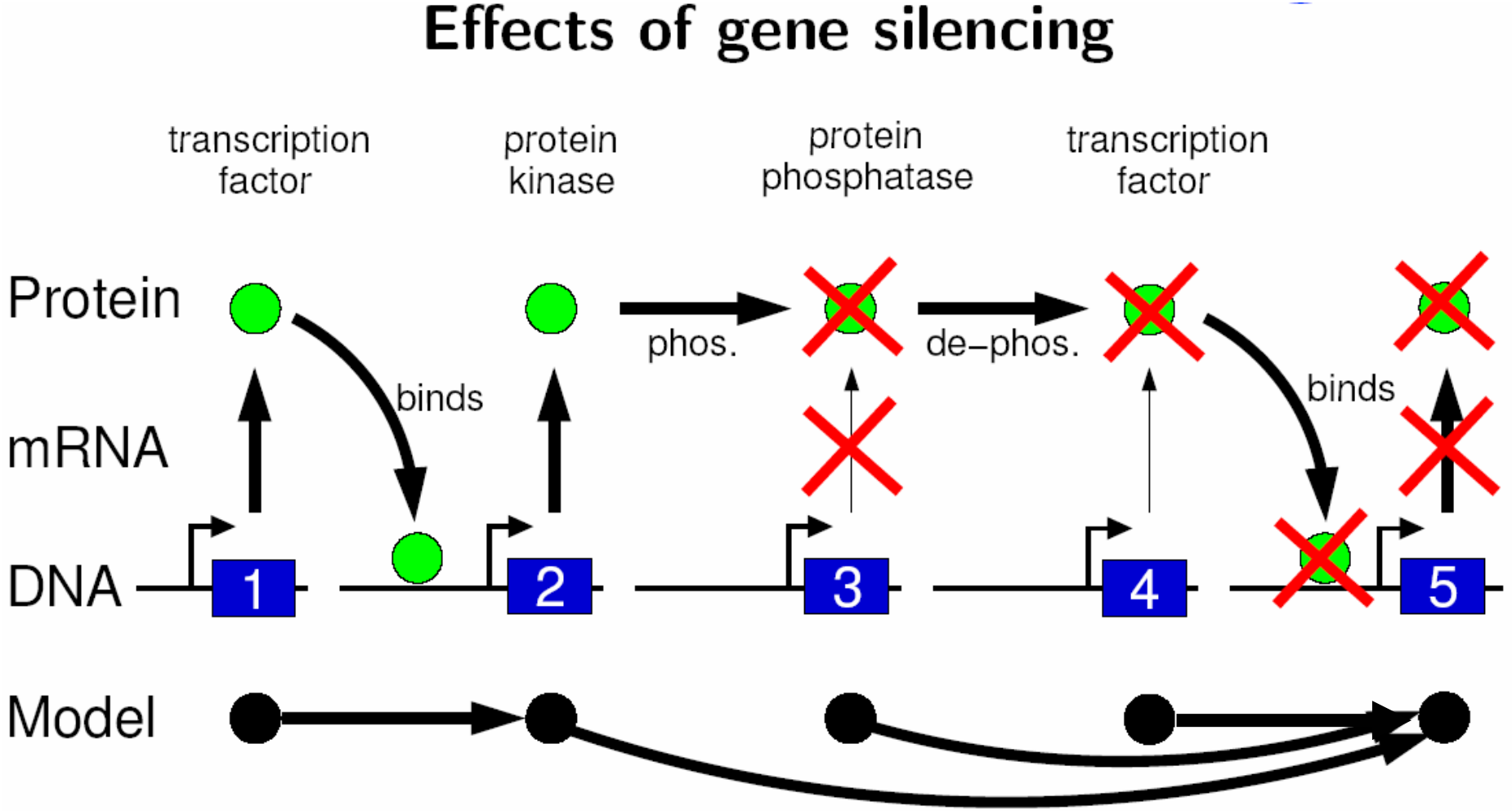
RNAi kills the messenger

(A. Fire, C. Mello, Nobel prize 2006)

1. **Double-stranded RNAs** (dsRNAs) get processed into **small interfering RNAs** (siRNAs).
2. siRNAs assemble into **RNA-induced silencing complexes** (RISCs).
3. The siRNA strands guide the RISCs to **complementary RNA molecules**, . . .
4. . . . where they **cleave and destroy** the cognate RNA.

Figure from http://www.ambion.com

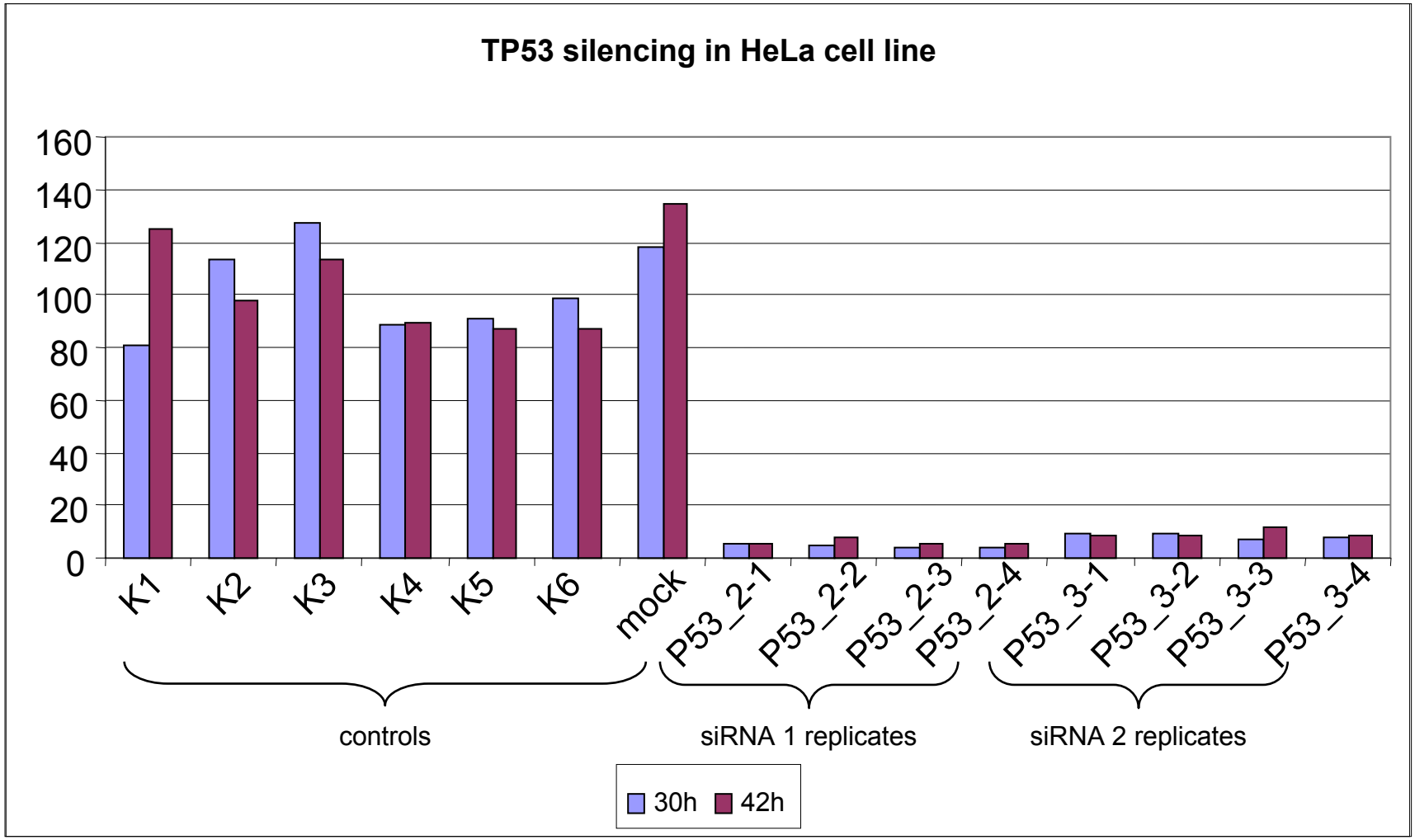
Learning from interventional data



Proof-of-principle Experiment

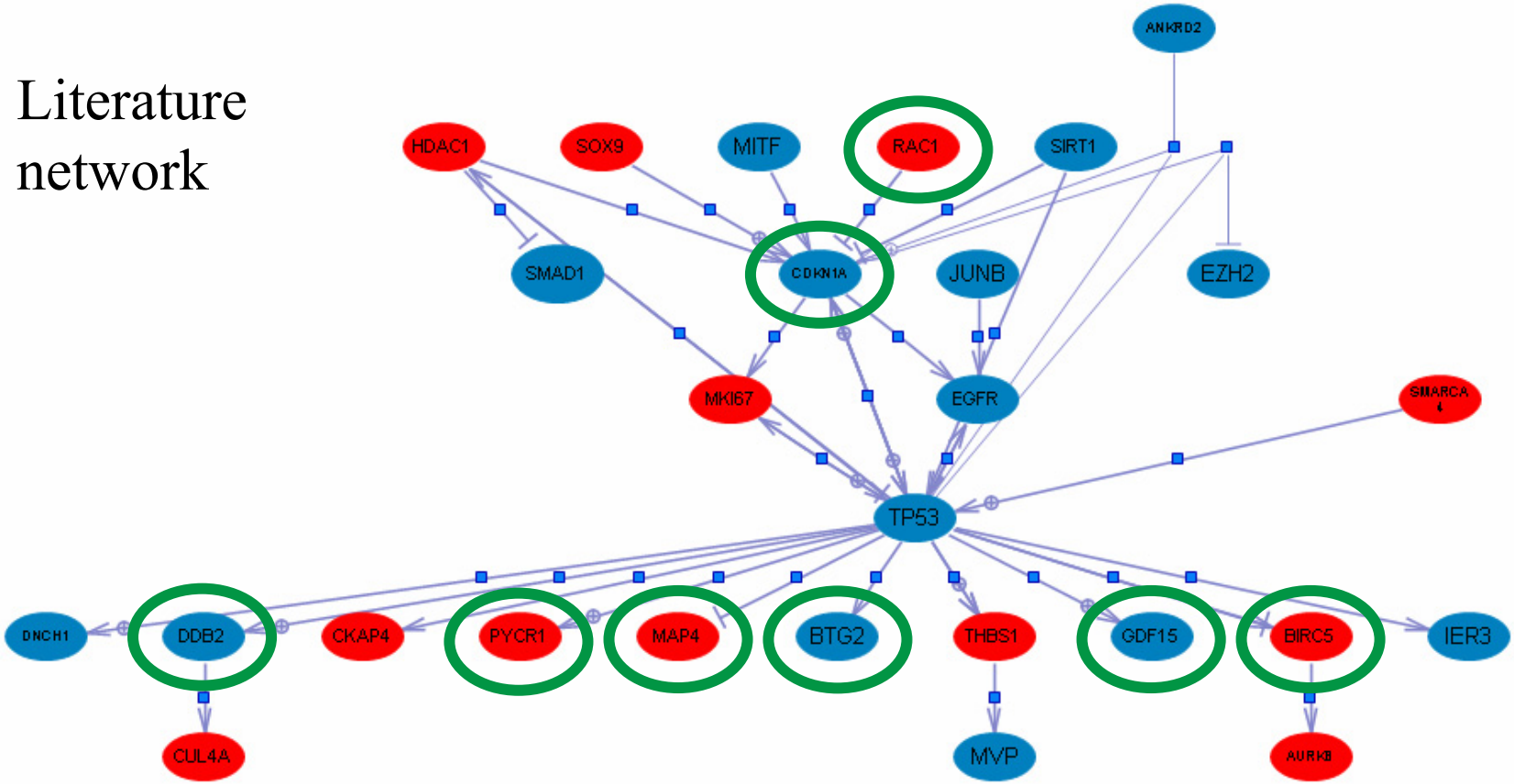
- TP53 chosen as model-gene for siRNA mediated knock-down
- nuclear protein, DNA-binding, postulated to bind as tetramer, activates expression of downstream genes that inhibit growth and/or invasion, thus function as a tumor suppressor
- very well known in literature, ease to find interaction partners
- Transfection of siRNA into HeLa cell line using HiPerFect transfection-reagent (HeLa cell line known to show very good transfection efficiencies)
- Minimal concentration of siRNA used to reduce possible off-target effects
- Measurement of silencing efficiency regarding mRNA of TP53 using qRT-PCR

Proof-of-principle Experiment



Proof of principle: The TP53 network

Literature network



increase

decrease

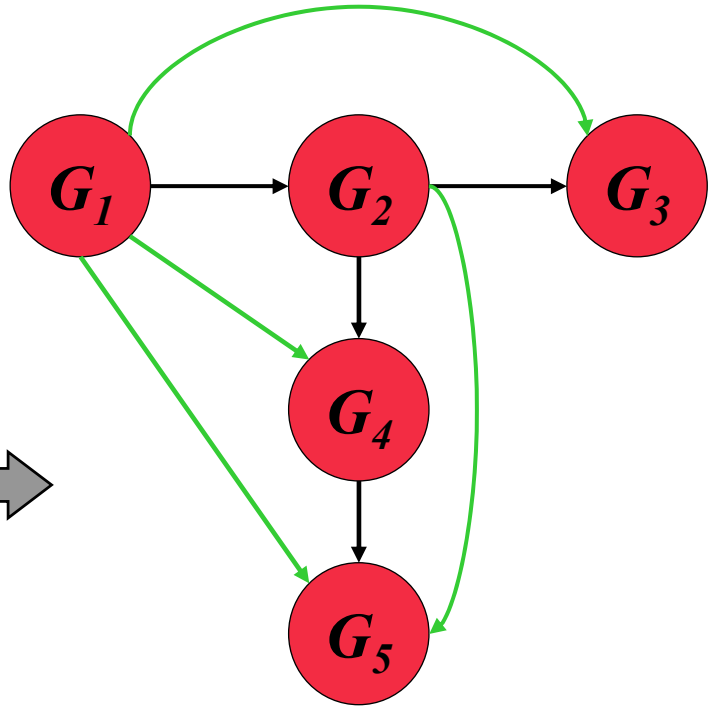
found by siRNA experiment

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

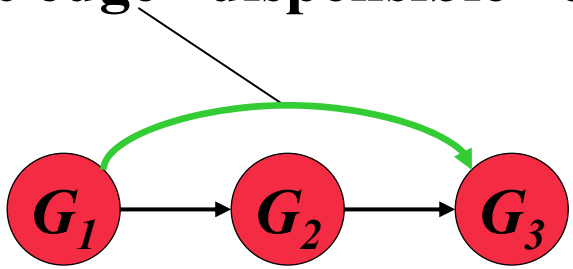
Pruning of Gene interaction Graphs

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)

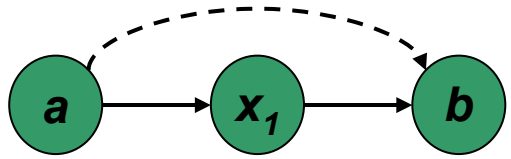
Is this edge "dispensible" or not?



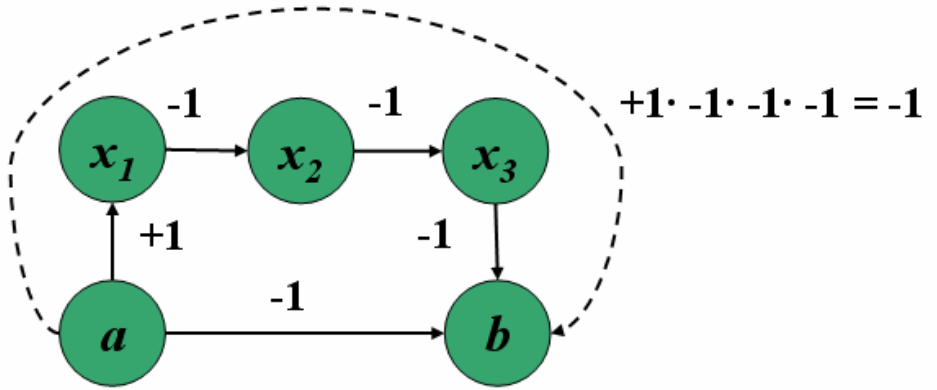
Need for algorithm to define and find
minimal consistent and biologically meaningful graph

Finding non-necessary edges

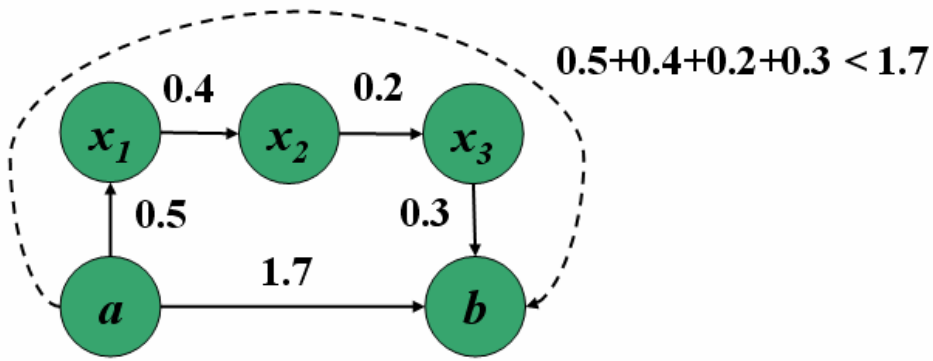
- "Trivial".
  Remove all edges a→b for which there exists a bypass (a longer way from a to b). [Wagner, 2002]



- "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]

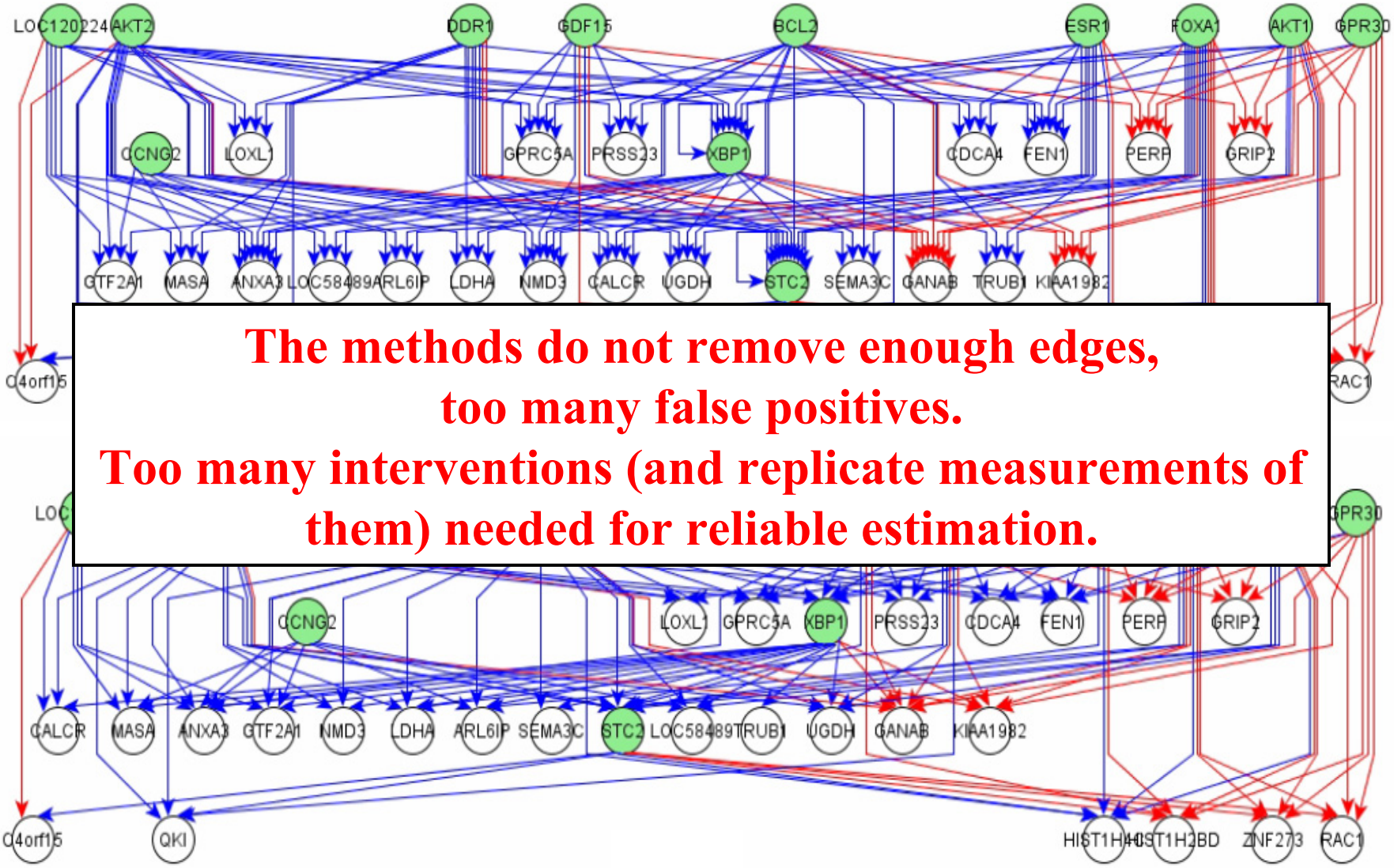


- "Weights".
  Let every edge be weighted with a non-negative number. Edges with low weights are meant to represent edges for which there is strong evidence for a direct regulatory interaction.
  Remove a→b if sum of the weights along the path a→... →b is smaller than the weight of the edge a→b
  [Tresch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



Tresch et al, unpublished

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.

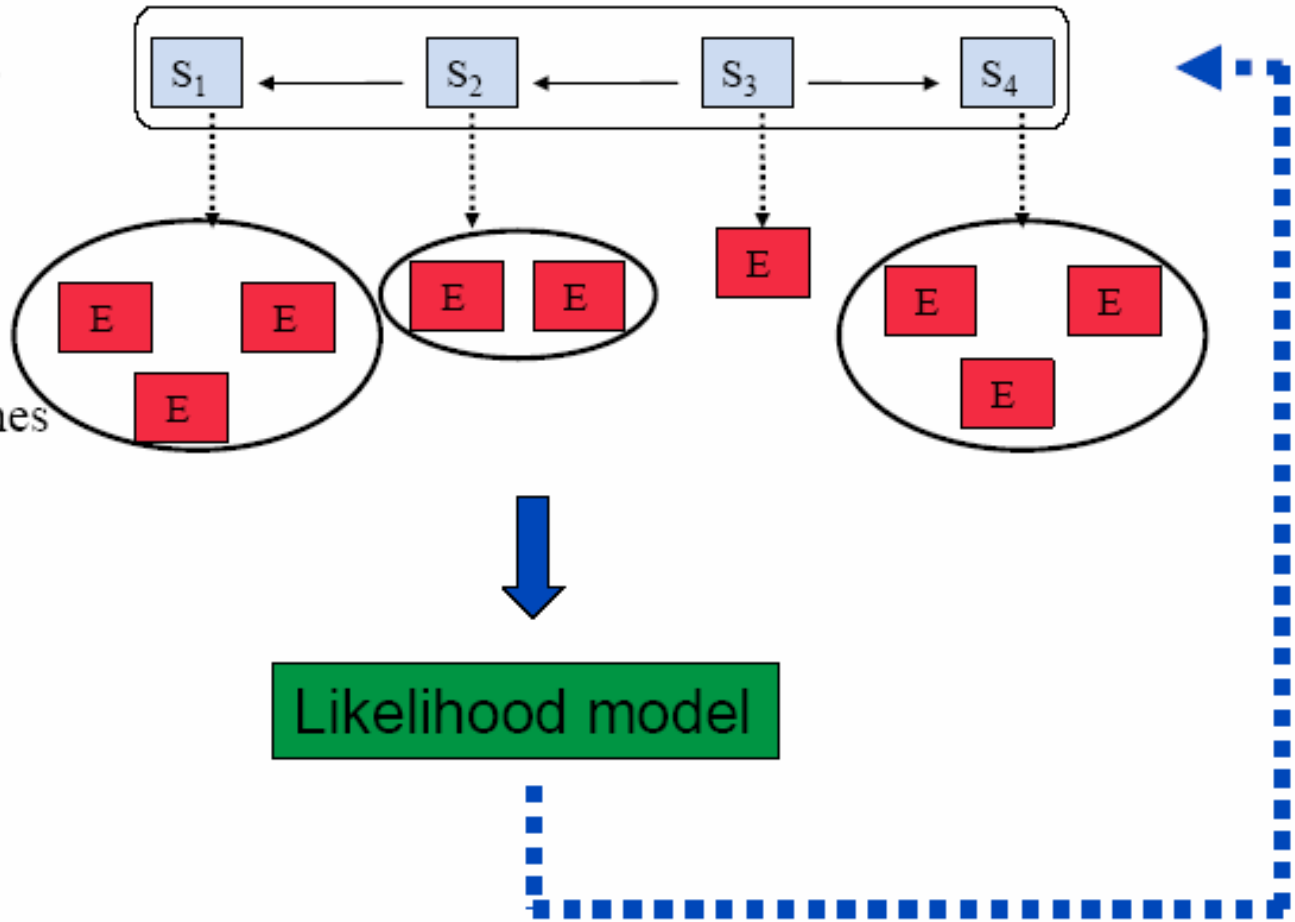
The Signal-Effects Model

1. Choose **candidate network topology** of silenced genes (S-genes)

2. Attach effected genes (E-genes)

3. Calculate score using Bayesian statistics

4. Propose different topology



Markowetz et al., Bioinformatics, 2005

Pathways from RNAi data – an example

Response to microbial challenge

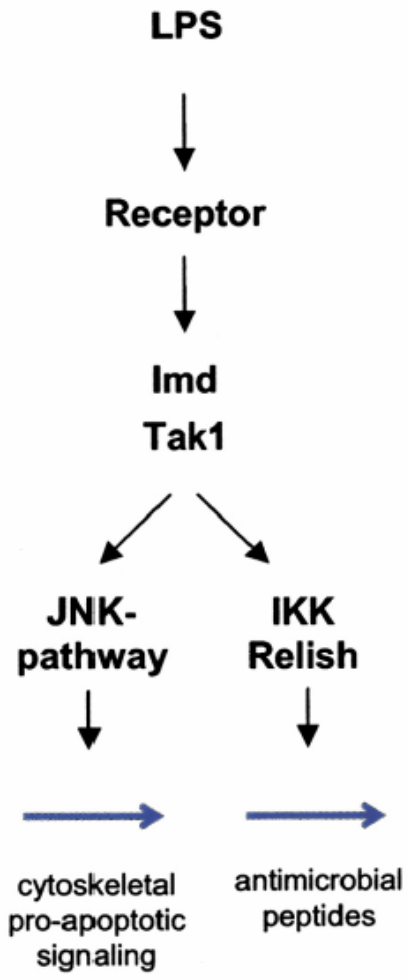
(Boutros *et al.*, Dev Cell, 2002)

Columns: silenced genes.

Rows: effects on other genes.

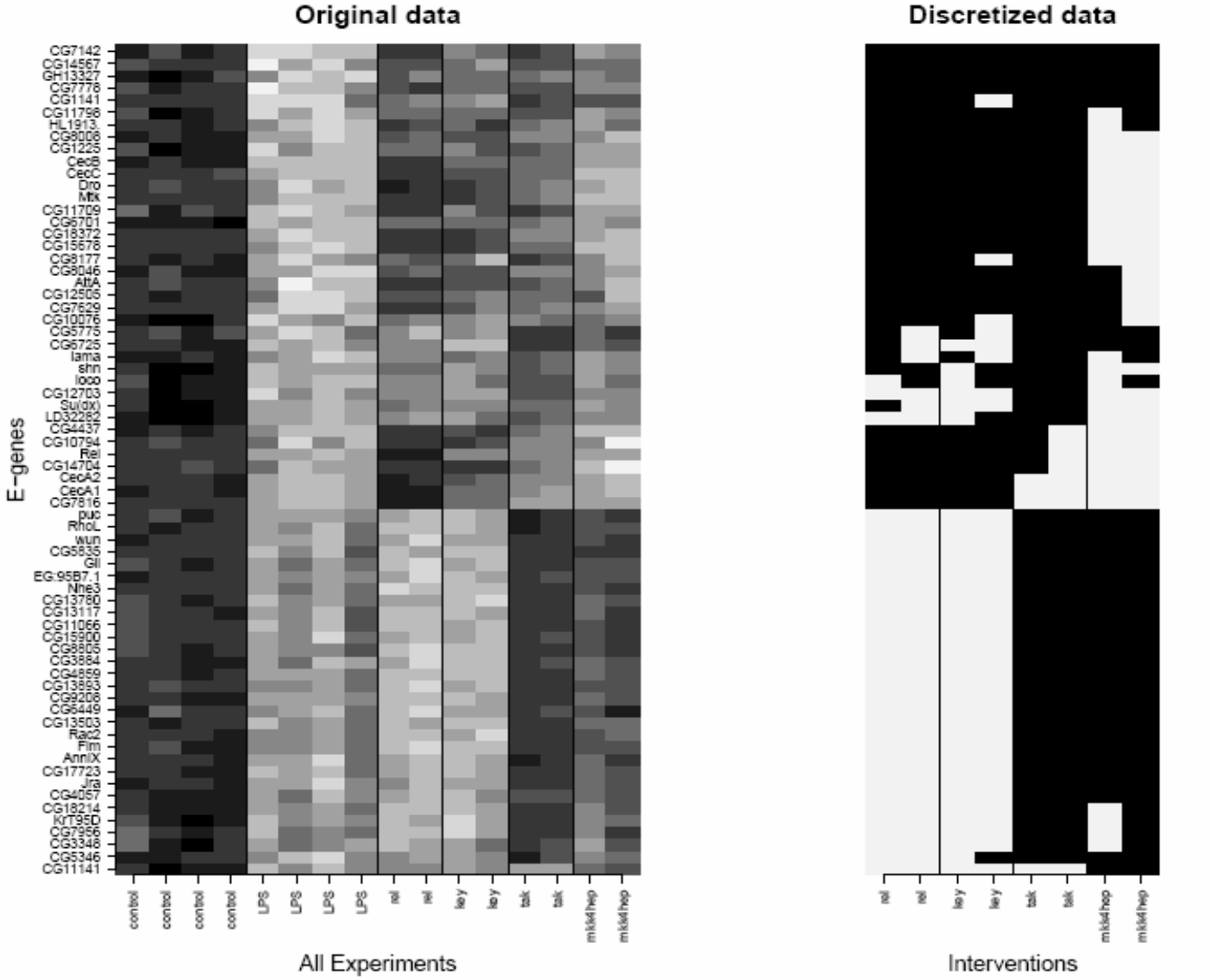
Results:

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.

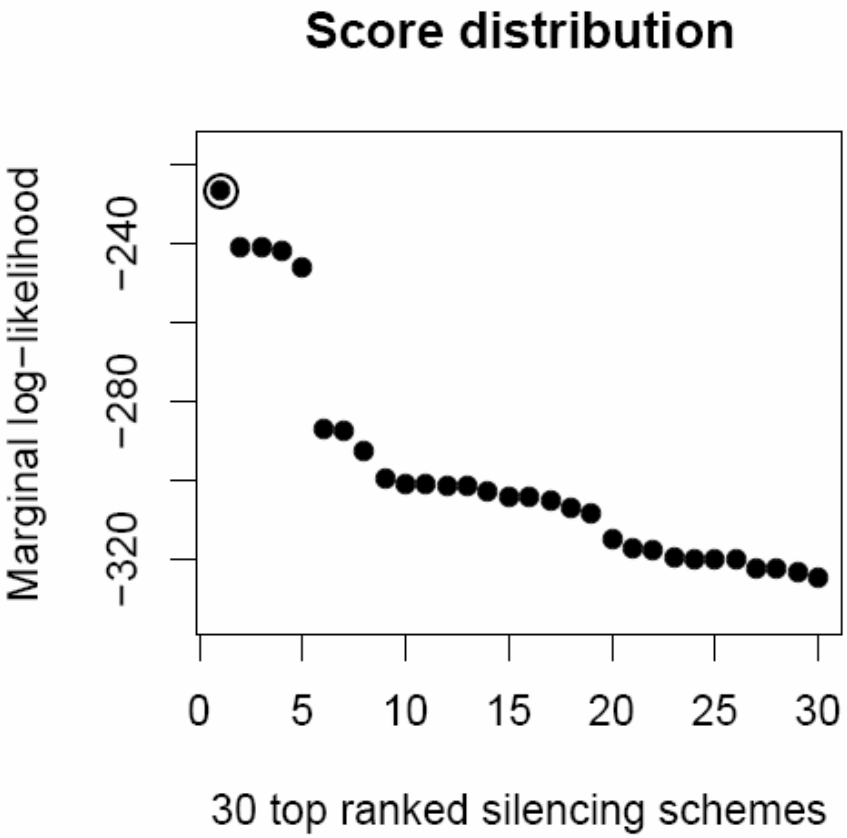


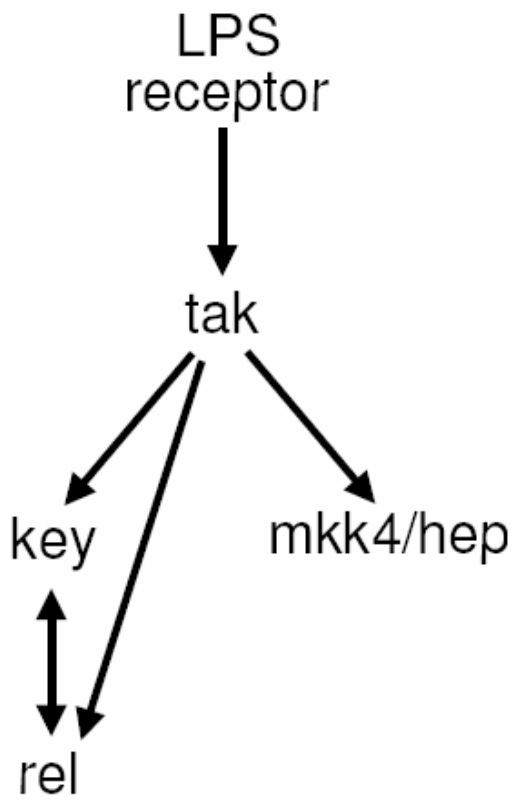
Figures from (Boutros *et al.*, 2002)

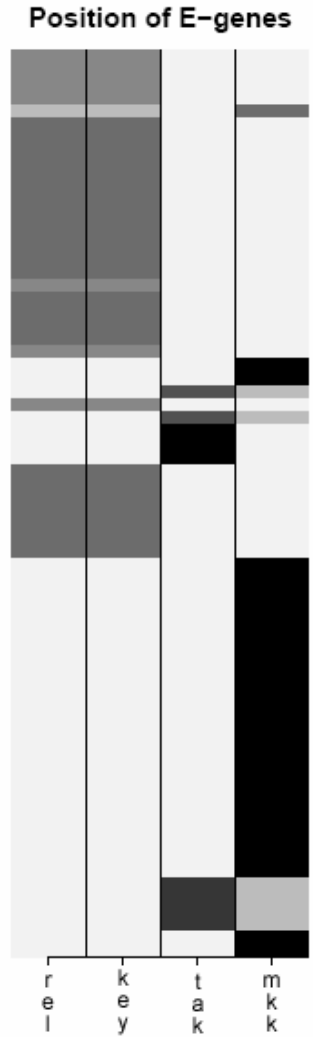
Application to Drosophila data



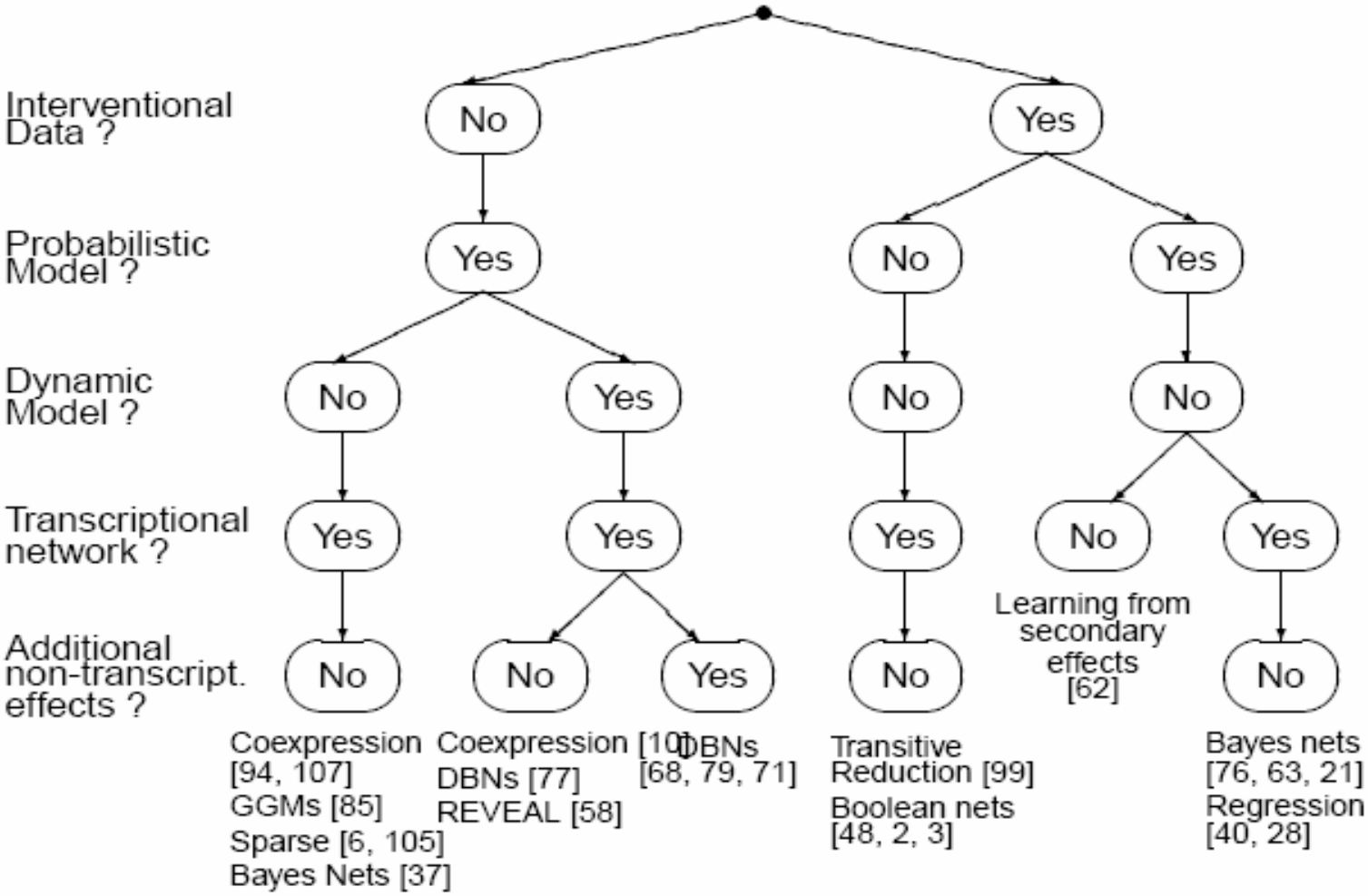
Application to Drosophila data







Survey of network models



Florian Markowetz, *Computational Inference of Cellular Pathways*, 2005 Oct 5

Networks in R/Bioconductor

- graph: basic class definitions and functionality
- RBGL: interface to graph algorithms (e.g. shortest path, connectivity)
- Rgraphviz: rendering functionality Different layout algorithms. Node plotting, line type, colour etc. can be controlled by the user.
- dynamicGraph: visualize interactive Graphs with TclTK.
- KEGGSOAP: Connecting to the KEGG database.
- GeneTS: Estimate GGMs from Microarray Data.

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
- Micheal Boutros

Thank you!

References

[1] Alfred V. Aho, M.R. Garey, and Jeffrey D. Ullman. The transitive reduction of a directed graph. *SIAM J. Comput.*, 1(2):131 – 137, 1972.

[2] Tatsuya Akutsu, Satoru Kuhara, Osamu Maruyama, and Satoru Miyano. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 695–702, 1998.

[3] Tatsuya Akutsu, Satoru Kuhara, Osamu Maruyama, and Satoru Miyano. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. In Satoru Miyano and T. Takagi, editors, *Genome Informatics 9*, pages 151–160, Tokyo, 1998. Universal Academy Press.

[4] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47), 2002.

[5] L Avery and S Wasserman. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet*, 8(9):312–6, Sep 1992.

[6] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nat Genet*, Mar 2005.

[7] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2005.

[8] Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–98, Apr 2004.

[9] Allister Bernard and Alexander J Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput*, pages 459–70, 2005.

[10] David R Bickel. Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics*, 21(7):1121–8, Apr 2005.

[11] Susanne Gammelgaard Bøttcher. *Learning Bayesian Networks with Mixed Variables*. PhD thesis, Aalborg University, Denmark, 2004.

[12] Susanne Gammelgaard Bøttcher and Claus Dethlefsen. deal: A package for learning bayesian networks. *Journal of Statistical Software*, 8(20), 2003.

[13] Anne-Laure Boulesteix and Korbinian Strimmer. Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, 2(23), 2005.

[14] C.T. Brown, A.G. Rust, P.J.C. Clarke, Z. Pan, M.J. Schilstra, T.D. Buysscher, G. Griffin, B.J. Wold, R.A. Cameron, E.H. Davidson, and H. Bolouri. New computational approaches for analysis of cis-regulatory networks. *Developmental Biology*, (246):86–102, 2002.

[15] Svetlana Bulashevska and Roland Eils. Inferring genetic regulatory logic from expression data. *Bioinformatics*, Mar 2005.

[16] Wray L. Buntine. Theory refinement of Bayesian networks. In *Uncertainty in Artificial Intelligence*, 1991.

[17] HJ Bussemaker, H Li, and ED Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, Feb 2001.

[18] AJ Butte and IS Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29, 2000.

[19] D.M. Chickering. Learning equivalence classes of Bayesian network structures. In *Proceedings of Twelfth Conference on Uncertainty in Artificial Intelligence, Portland, OR*, pages 150–157. Morgan Kaufmann, August 1996.

[20] Gregory F. Cooper and Edward Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9:309–347, 1992.

[21] Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In K. Laskey and H. Prade, editors, *Proc. Fifthteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pages 116–125, San Francisco, Calif., 1999. Morgan Kaufman.

[22] Eric H. Davidson, Jonathan P. Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, Ochan Otim, C. Titus Brown, Carolina B. Livi, Pei Yun Lee, Roger Revilla, Alistair G. Rust, Zheng jun Pan, Maria J. Schilstra, Peter J. C. Clarke, Maria I. Arnone, Lee Rowen, R. Andrew Cameron, David R. McClay, Leroy Hood, and Hamid Bolouri. A Genomic Regulatory Network for Development. *Science*, 295(5560):1669–1678, 2002.

[23] Hidde De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, January 2002.

[24] 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.

[25] Lloyd Demetrius and Thomas Manke. Robustness and network evolution–an entropic principle. *Physica A: Statistical and Theoretical Physics*, 346(3-4):682–696, 2005.

[26] Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[27] Diego di Bernardo, Gardner Timothy S, and James J Collins. Robust identification of large genetic networks. *Pac Symp Biocomput*, pages 486–97, 2004.

[28] Diego di Bernardo, Michael J Thompson, Timothy S Gardner, Sarah E Chobot, Erin L Eastwood, Andrew P Wojtovich, Sean J Elliott, Scott E Schaus, and James J Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol*, 23(3):377–83, Mar 2005.

[29] Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, July 2004.

[30] Nancy Van Driessche, Janez Demsar, Ezgi O Booth, Paul Hill, Peter Juvan, Blaz Zupan, Adam Kuspa, and Gad Shaulsky. Epistasis analysis with global transcriptional phenotypes. *Nat Genet*, 37(5):471–7, May 2005.

[31] Mathias Drton and Michael D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91(3), 2004.

[32] Mathias Drton and Michael D. Perlman. A SINful approach to gaussian graphical model selection. *Submitted to special issue of Statistical Science*, 2004.

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

[34] 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.

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

[36] Nir Friedman and Daphne Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–126, 2003.

[37] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3):601–620, August 2000.

[38] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 139–147, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[39] Nir Friedman, Iftach Nachman, and Dana Peer. Learning bayesian network structures from massive datasets: The sparse candidate algorithm. In *Proc. of Uncertainty in Artificial Intelligence*, 1999.

[40] Timothy S. Gardner, Diego di Bernardo, David Lorenz, and James J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, 2003.

[41] Dan Geiger and David Heckerman. Learning Gaussian Networks. In Ramon López de Mántaras and David Poole, editors, *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 235–243, Seattle, Washington, USA, July 29-31 1994. Morgan Kaufmann.

[42] A. Gelman, J. B. Carlin, H.S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall-CRC, 1996.

[43] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Proceedings of Pacific Symposium on Biocomputing 7:437-449*, 2002.

[44] Alexander J Hartemink. Reverse engineering gene regulatory networks. *Nat Biotechnol*, 23(5):554–5, May 2005.

[45] David Heckerman. A Tutorial on Learning Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, March 1995.

[46] David Heckerman, Dan Geiger, and David Maxwell Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, Sep. 1995.

[47] Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[48] Trey Ideker, Vesteinn Thorsson, and Richard M. Karp. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Proc. of the Pacic Symp. on Biocomputing*, volume 5, pages 302–313, 2000.

[49] Seiya Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. In *Pacific Symposium on Biocomputing*, volume 7, pages 175–186, 2002.

[50] Seiya Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In *Proc. 2nd Computational Systems Bioinformatics*, pages 104–113, 2003.

[51] Seiya Imoto, Sunyong Kim, Takao Goto, Satoru Miyano, Sachiyo Aburatani, Kousuke Tashiro, and Satoru Kuhara. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J Bioinform Comput Biol*, 1(2):231–52, Jul 2003.

[52] Hirohisa Kishino and Peter J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. In A.K. Dunker, A. Konagaya, S. Miyano, and T.Takagi, editors, *Genome Informatics*, Tokyo, 2000. Universal Academy Press.

[53] Edda Klipp, Bodil Nordlander, Roland Krger, Peter Gennemark, and Stefan Hohmann. Integrative model of the response of yeast to osmotic shock. *Nat Biotechnol*, 23(8):975–82, Aug 2005.

[54] Ina Koch, Björn H Junker, and Monika Heiner. Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21(7):1219–26, Apr 2005.

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

[56] Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[57] Jan van Leeuwen. Graph algorithms, 1990. in: Handbook of Theoretical Computer Science, Elsevier, 525–632.

[58] Shoudan Liang, Stefanie Fuhrmann, and Roland Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proc. of Pacific Symposium on Biocomputing*, number 3, pages 18–29, 1998.

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

[60] Johannes Mandel, Niall M Palfreyman, Jesus A Lopez, and Werner Dubitzky. Representing bioinformatics causality. *Brief Bioinform*, 5(3):270–83, Sep 2004.

[61] Florian Markowetz. A bibliography on learning causal networks of gene interactions, 2005. URL http://www.molgen.mpg.de/~markowet/docs/network-bib.pdf.

[62] Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics (accepted)*, ?(?):?, 2005.

[63] Florian Markowetz, Steffen Grossmann, and Rainer Spang. Probabilistic soft interventions in conditional gaussian networks. In Robert Cowell and Zoubin Ghahramani, editors, *Proc. Tenth International Workshop on Artificial Intelligence and Statistics*, Jan 2005.

[64] Florian Markowetz and Rainer Spang. Evaluating the effect of perturbations in reconstructing network topologies. In Kurt Hornik, Friedrich Leisch, and Achim Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003.

[65] K. Murphy and S. Mian. Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[66] Kevin Murphy. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33, 2001.

[67] Kevin P. Murphy. Active Learning of Causal Bayes Net Structure, 2001.

[68] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl.1):i248–256, 2004.

[69] Iftach Nachman, Gal Elidan, and Nir Friedman. "ideal parent" structure learning for continuous variable networks. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 400–409, Arlington, Virginia, United States, 2004. AUAI Press.

[70] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

[71] Irene M Ong, Jeremy D Glasner, and David Page. Modelling regulatory pathways in E. coli from time series expression profiles. *Bioinformatics*, 18 Suppl 1:S241–8, 2002.

[72] Jason A. Papin, Tony Hunter, Bernhard O. Palsson, and Shankar Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol*, 6(2):99–111, 2005.

[73] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[74] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2000.

[75] Dana Pe'er. Bayesian Network Analysis of Signaling Networks: A Primer. *Science's STKE*, 2005(281):pl4–, 2005.

[76] Dana Pe'er, Aviv Regev, Gal Elidan, and Nir Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(90001):S215–S224, 2001.

[77] Bruno-Edouard Perrin, Liva Ralaivola, Aurelien Mazurie, Samuele Bottani, Jacques Mallet, and Florence d'Alche Buc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19(90002):138ii–148, 2003.

[78] Iosifina Pournara and Lorenz Wernisch. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, 20(17):2934–2942, 2004.

[79] Claudia Rangel, John Angus, Zoubin Ghahramani, Maria Lioumi, Elizabeth Sotheran, Alessia Gaiba, David L. Wild, and Francesco Falciani. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

[80] Claudia Rangel, David L. Wild, Francesco Falciani, Zoubin Ghahramani, and Alessia Gaiba. Modeling biological responses using gene expression profiling and linear dynamical systems. In *Proceedings of the 2nd International Conference on Systems Biology*, pages 248–256, Madison, WI, 2001. Omnipress.

[81] Robert W. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New Directions in the Theory of Graphs*, pages 239–273. Academic Press, New York, 1973.

[82] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18(90002):202S–210, 2002.

[83] Karen Sachs, David Gifford, Tommi Jaakkola, Peter Sorger, and Douglas A Lauffenburger. Bayesian network approach to cell signaling pathway modeling. *Sci STKE*, 2002(148):PE38, Sep 2002.

[84] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9, Apr 2005.

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

[86] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels*. The MIT Press, Cambridge, MA, 2002.

[87] Eran Segal, Nir Friedman, Naftali Kaminski, Aviv Regev, and Daphne Koller. From signatures to models: understanding cancer using microarrays. *Nat Genet*, 37 Suppl:S38–45, Jun 2005.

[88] Peter W. F. Smith and Joe Whittaker. Edge exclusion tests for graphical Gaussian models. In Michael Jordan, editor, *Learning in Graphical Models*, pages 555 – 574. MIT Press, 1999.

[89] PT Spellman, G Sherlock, MQ Zhang, VR Iyer, K Anders, MB Eisen, PO Brown, D Botstein, and B Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, Dec 1998.

[90] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

[91] Harald Steck and Tommi Jaakkola. On the dirichlet prior and Bayesian regularization. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2002. MIT Press.

[92] Harald Steck and Tommi Jaakkola. (semi-)predictive discretization during model selection. Technical Report AI Memo AIM-2003-002, MIT, 2003.

[93] Harald Steck and Tommi S. Jaakkola. Bias-corrected bootstrap and model uncertainty. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[94] 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.

[95] Yoshinori Tamada, SunYong Kim, Hideo Bannai, Seiya Imoto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19(90002):227ii–236, 2003.

[96] Jesper Tegner, M K Stephen Yeung, Jeff Hasty, and James J Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A*, 100(10):5944–9, May 2003.

[97] Simon Tong and Daphne Koller. Active Learning for Structure in Bayesian Networks. In *Proceedings of the Seventeenth International Joint Conference on Artifical Intelligence (IJCAI)*, Seattle, Washington, August 2001.

[98] E. van Someren, L. Wessels, and M. Reinders. Genetic network models: A comparative study. In *Proc. of SPIE, 2001.*, 2001.

[99] Andreas Wagner. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, 17(12):1183–1197, 2001.

[100] Andreas Wagner. Estimating Coarse Gene Network Structure from Large-Scale Gene Perturbation Data. *Genome Res.*, 12(2):309–315, 2002.

[101] Andreas Wagner. Reconstructing pathways in large genetic networks from genetic perturbations. *Journal of Computational Biology*, 11(1):53–60, 2004.

[102] Junbai Wang, Ola Myklebost, and Eivind Hovig. Mgraph: graphical models for microarray data analysis. *Bioinformatics*, 19(17):2210–2211, 2003.

[103] Wei Wang and Gregory F. Cooper. An bayesian method for biological pathway discovery from high-throughput experimental data. In *Proc. 3rd International IEEE Computer Society Computational Systems Bioinformatics Conference (CSB 2004)*, pages 645–646, Stanford, CA, USA, 2004. IEEE Computer Society.

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

[105] Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelic, Peter von Rohr, Lothar Thiele, Eckart Zitzler, Wilhelm Gruissem, and Peter Bühlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol*, 5(11):R92, 2004.

[106] F. Wimberly, T. Heiman, J. Ramsey, and C. Glymour. Experiments on the accuracy of algorithms for inferring the structure of genetic regulatory networks from microarray expression levels. In *Proc. IJCAI 2003 Bioinformatics Workshop*, 2003.

[107] Cecily J. Wolfe, Isaac S. Kohane, and Atul J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics 2005*, 6:227, 2005.

[108] Biao Xing and Mark J van der Laan. A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. *J Comput Biol*, 12(2):229–46, 2005.

[109] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl.1):i363–370, 2004.

[110] Chen-Hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of Computational Biology*, 11(2):243 – 262, 2004.

[111] Chen-Hsiang Yeang, H Craig Mak, Scott McCuine, Christopher Workman, Tommi Jaakkola, and Trey Ideker. Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biology*, 6(R62), 2005.

[112] M K Stephen Yeung, Jesper Tegnr, and James J Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–8, Apr 2002.

[113] C. Yoo and G. F. Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Journal Artificial Intelligence in Medicine*, 31(2):169–182, 2004.

[114] Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, Jul 2004.

[115] D. E. Zak, F. J. Doyle, G. E. Gonye, and J. S. Schwaber. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In *Proceedings of the Second International Conference on Systems Biology*, pages 231–238, 2001.

[116] Daniel E. Zak, Gregory E. Gonye, James S. Schwaber, and Francis J. Doyle. Importance of Input Perturbations and Stochastic Gene Expression in the Reverse Engineering of Genetic Regulatory Networks: Insights From an Identifiability Analysis of an In Silico Network. *Genome Res.*, 13(11):2396–2405, 2003.

[117] Min Zou and Suzanne D. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2005.