

---

# A leisurely look at probabilistic modelling in systems biology

Dirk Husmeier

Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ  
<http://www.bioss.ac.uk/~dirk>

---

# A leisurely look at probabilistic modelling in **systems biology**

Dirk Husmeier

Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ  
<http://www.bioss.ac.uk/~dirk>

---

# A leisurely look at **probabilistic** modelling in **systems biology**

Dirk Husmeier

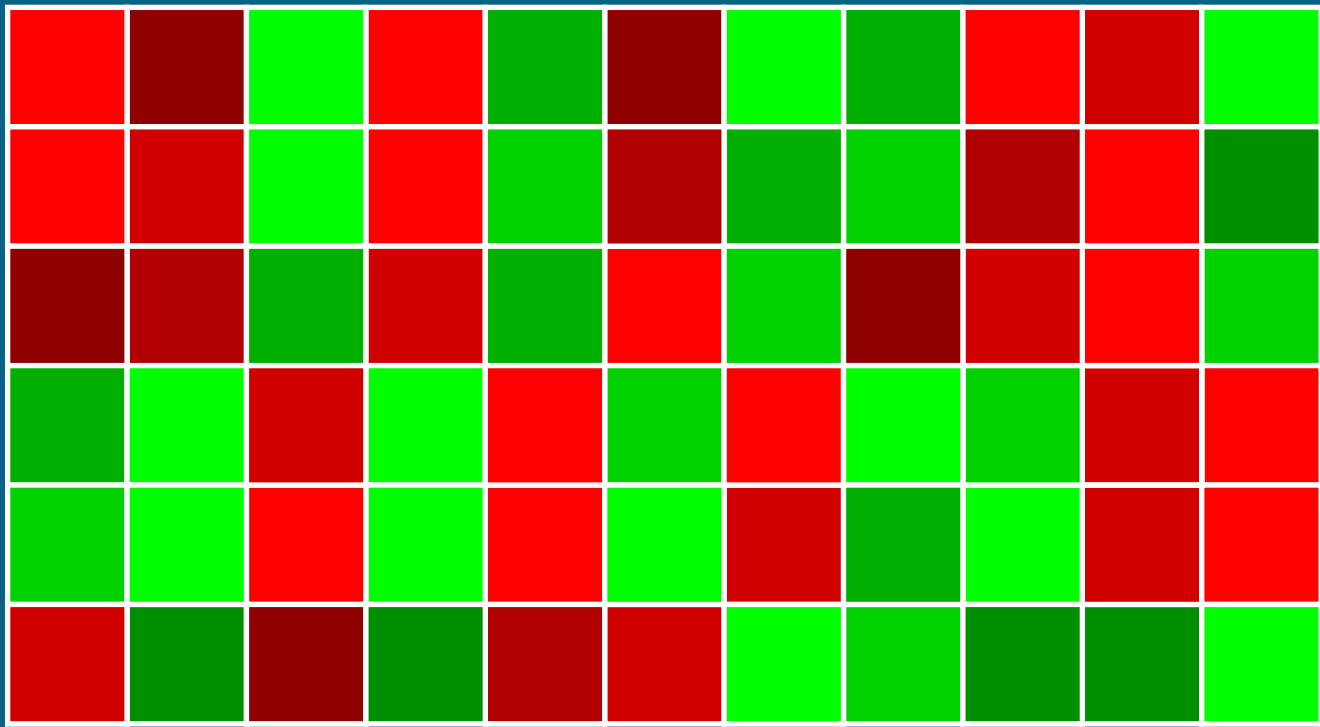
Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ  
<http://www.bioss.ac.uk/~dirk>

---

# A **leisurely** look at **probabilistic** modelling in **systems biology**

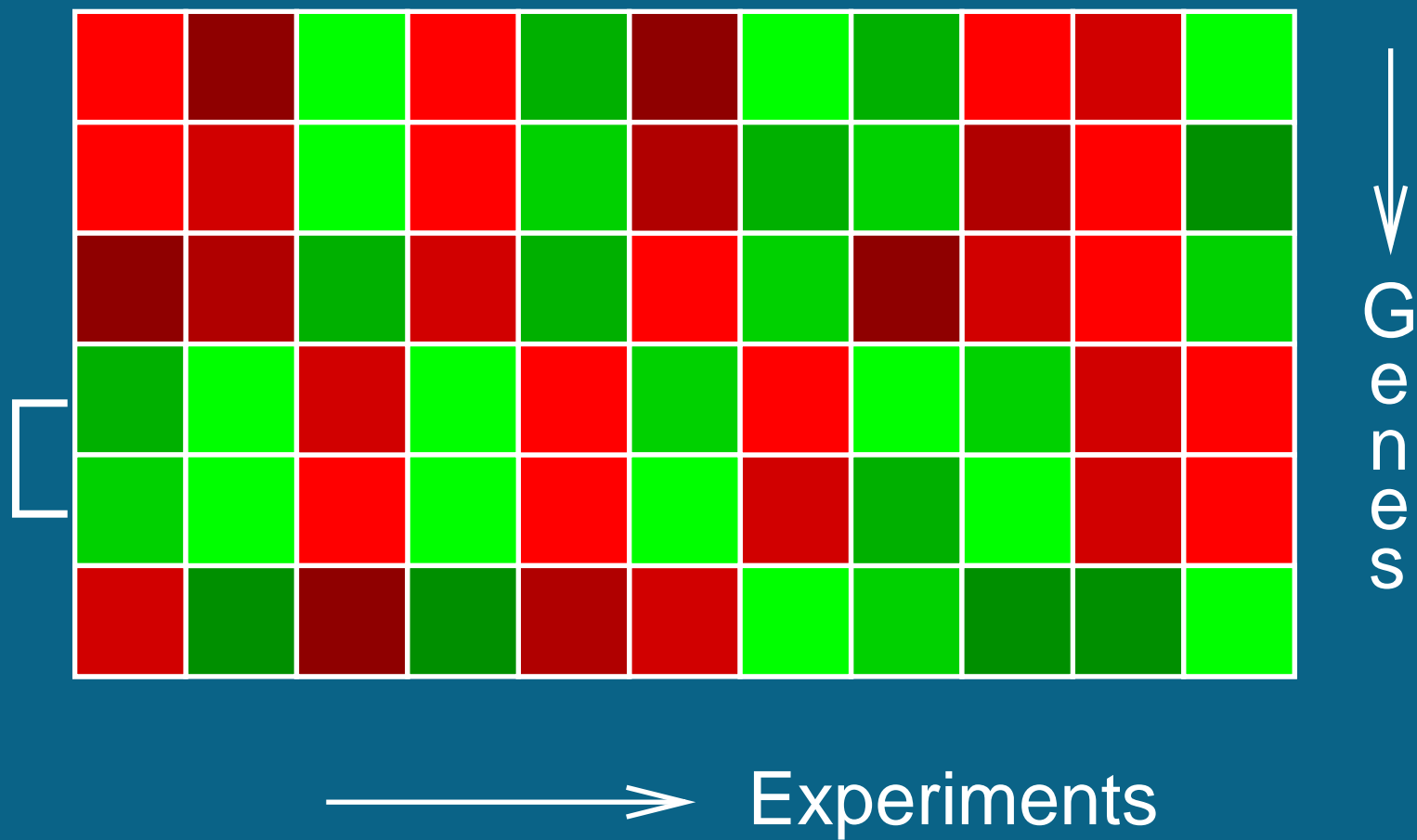
Dirk Husmeier

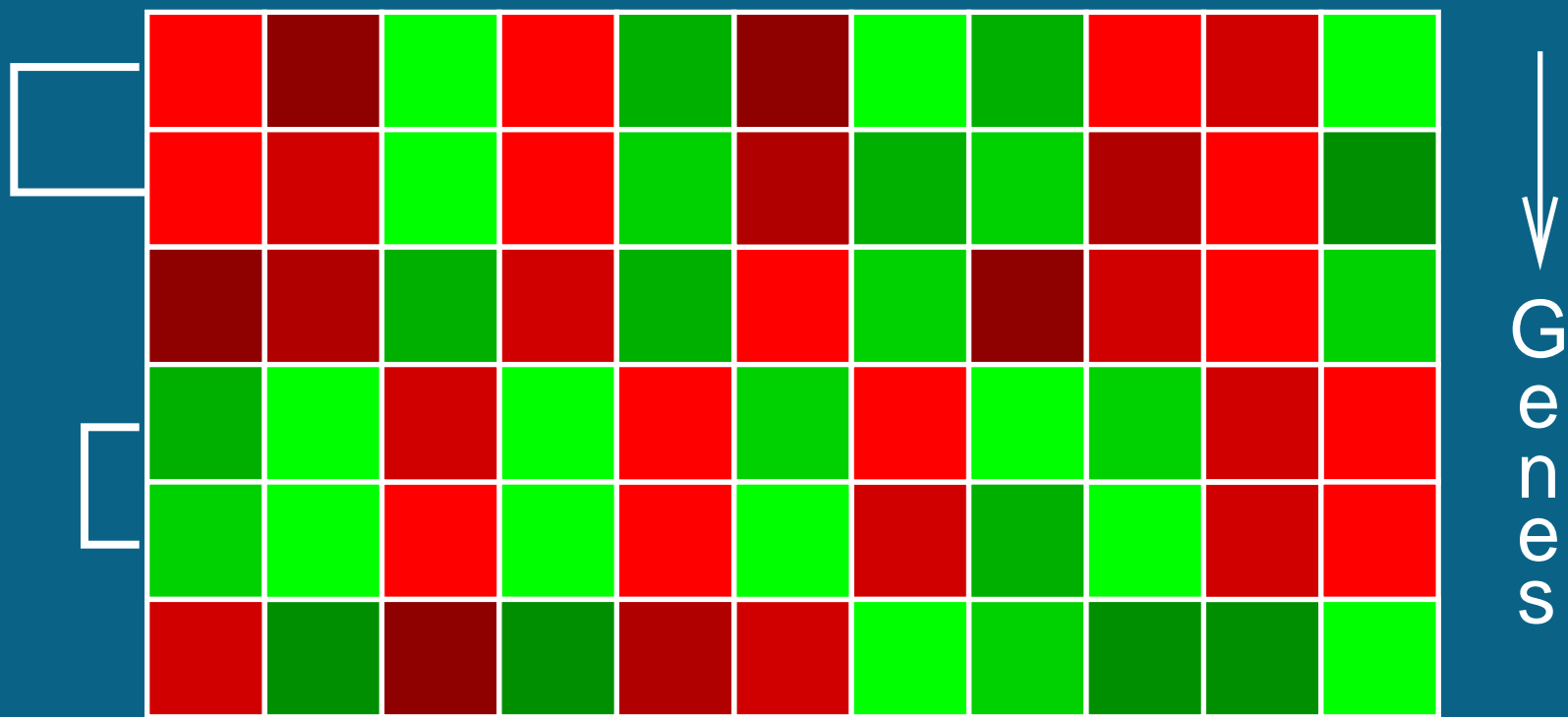
Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ  
<http://www.bioss.ac.uk/~dirk>



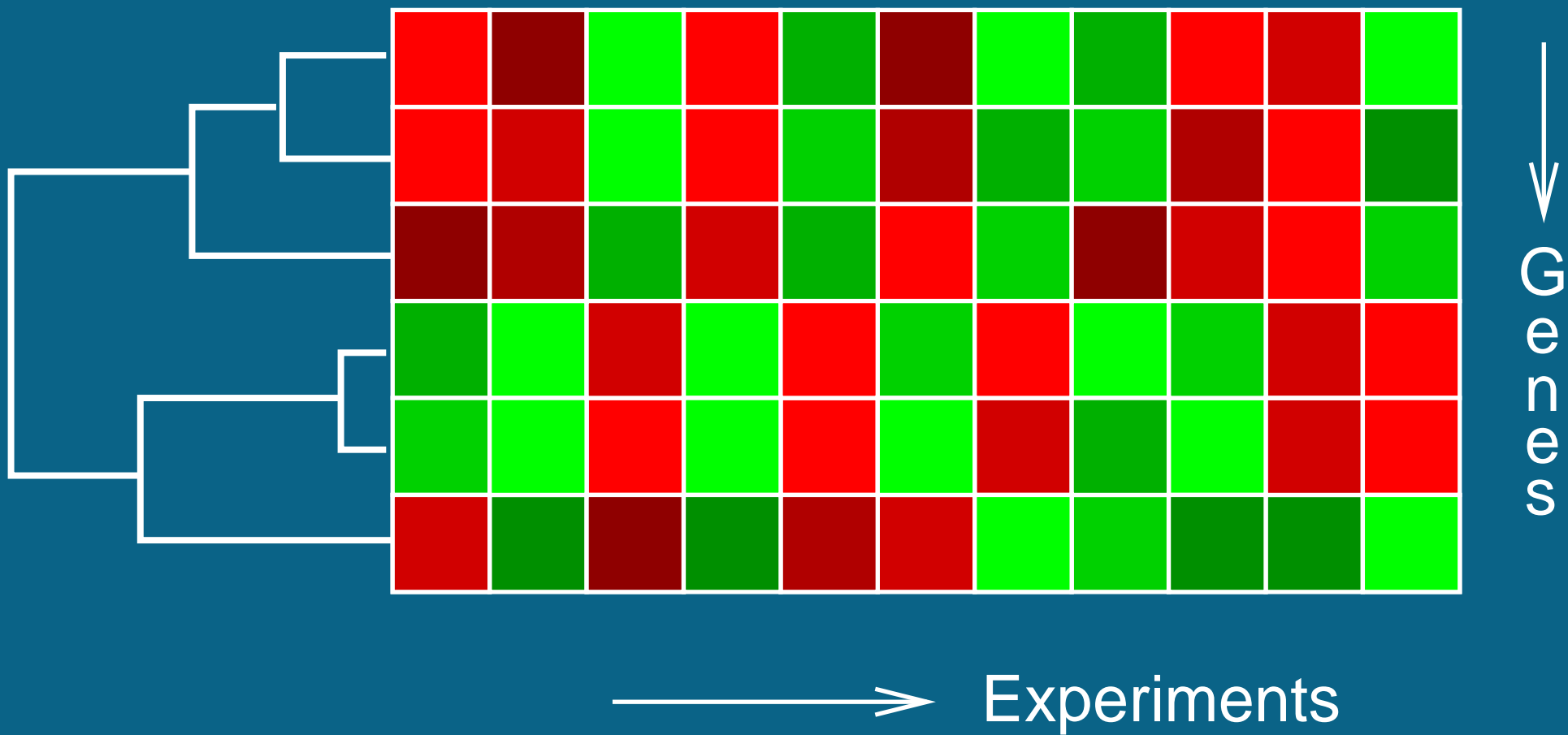
↓  
Genes

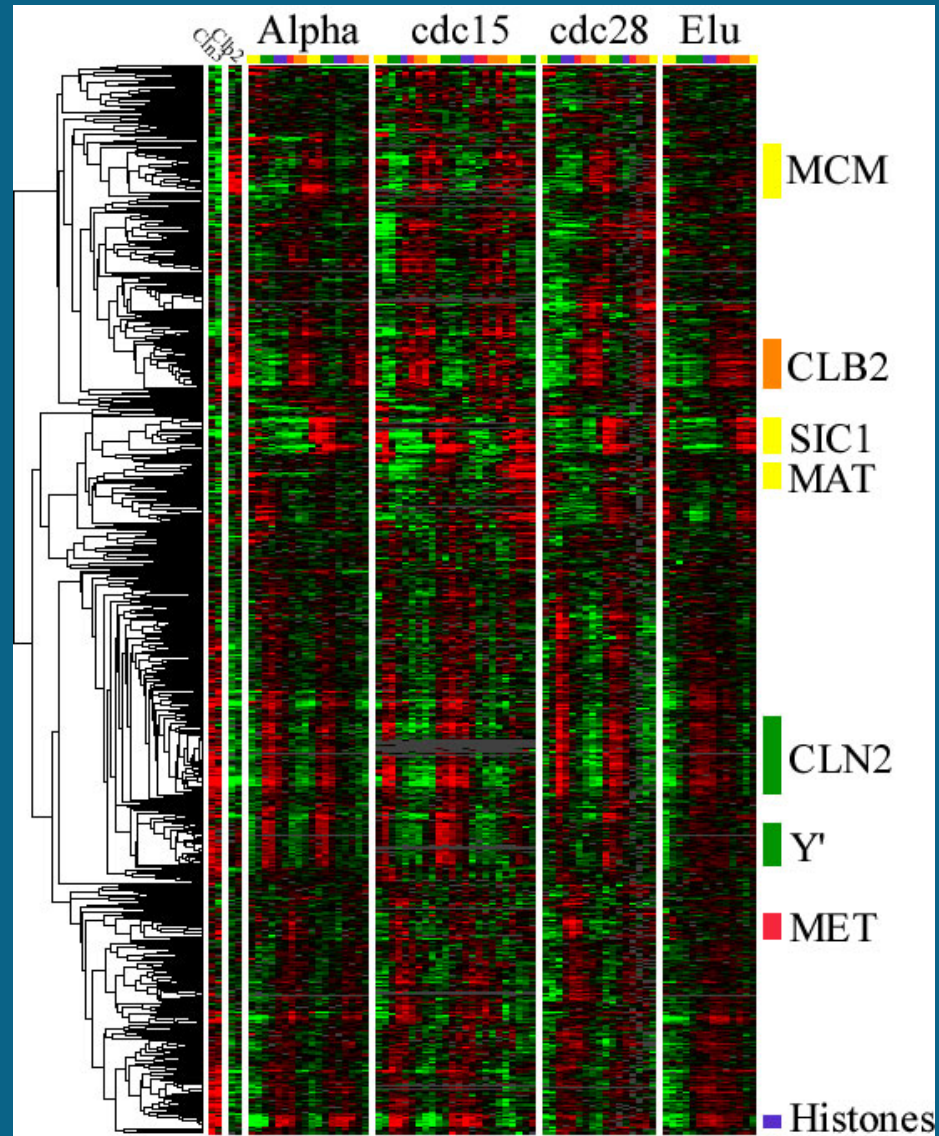
→ Experiments





Experiments





From Spellman et al., <http://cellcycle-www.stanford.edu/>

# Advantage of clustering

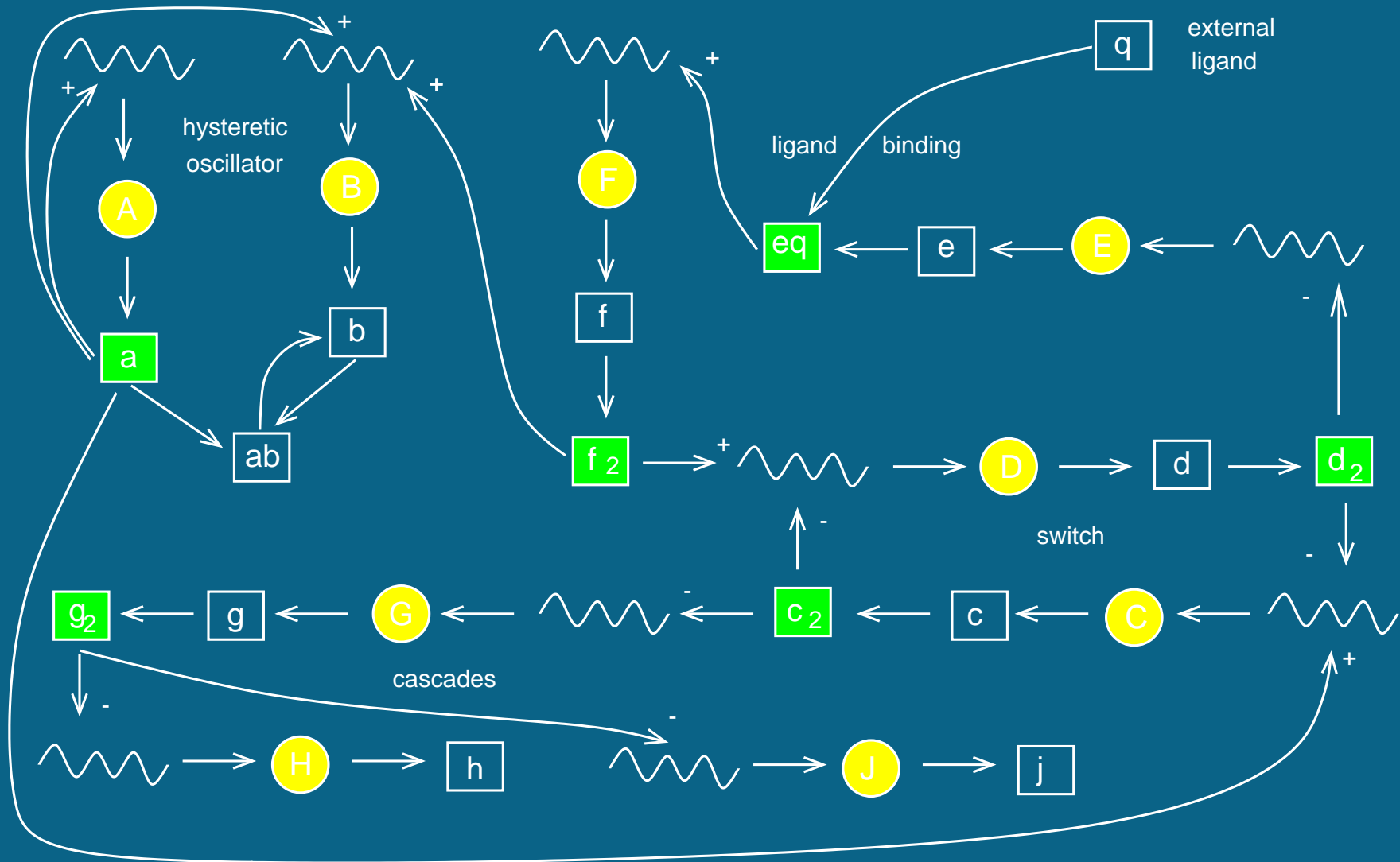
Fast, computationally cheap

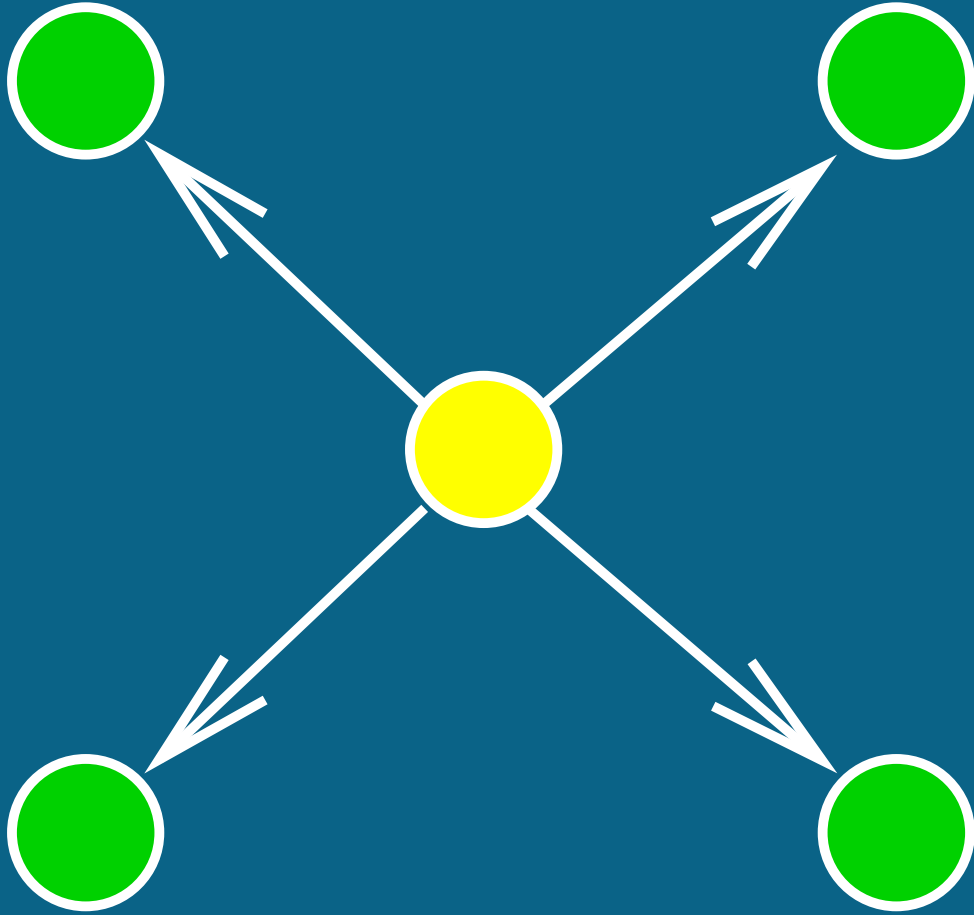
# Advantage of clustering

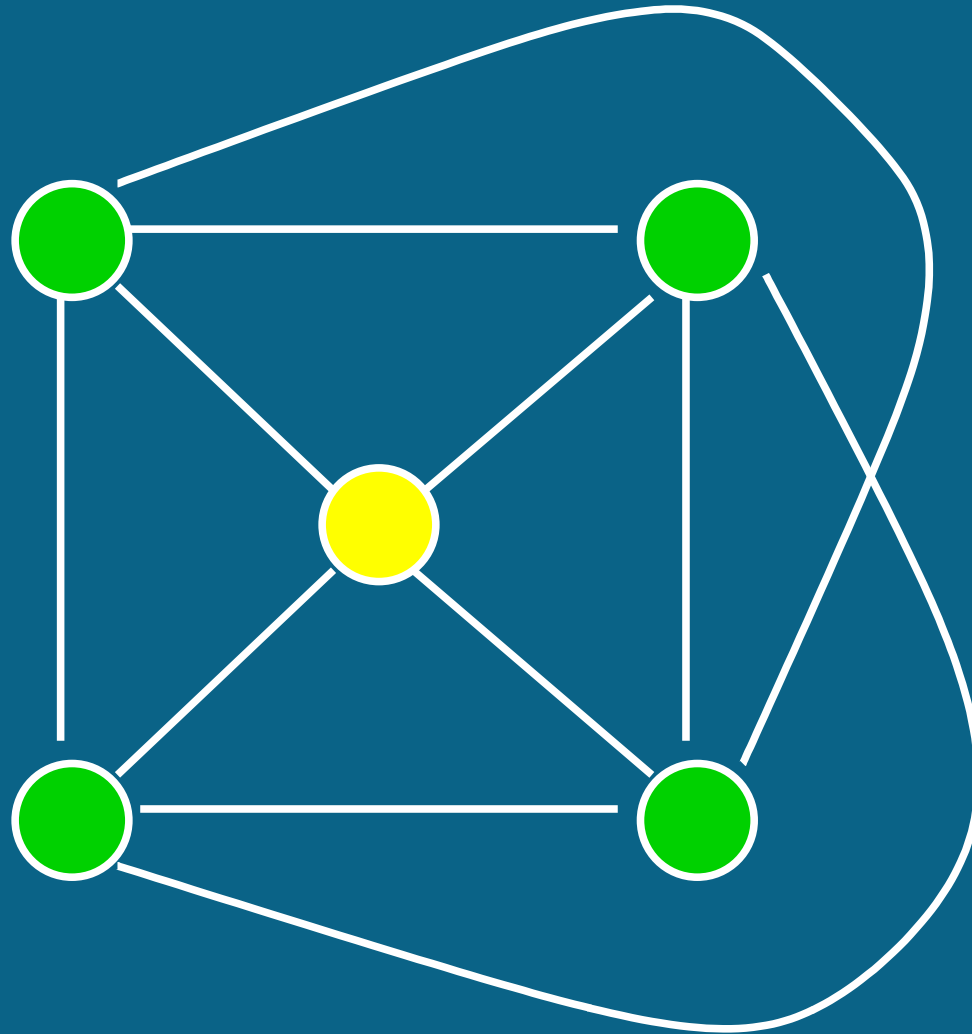
Fast, computationally cheap

# Shortcoming of clustering

It is **NOT** reverse engineering







# Reverse engineering

Learn the network structure from gene expression data.

**Problem:** Noise, sparse data

# Bayesian networks

Probabilistic framework for  
robust inference of interactions  
in the presence of noise

Nir Friedman et al. (2000)  
Journal of Computational Biology 7: 601-620

# Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering:  
Learning networks from data
- Application to the yeast cell cycle
  
- Probabilistic models for  
postgenomic data integration

- **Recapitulation: Bayesian networks**
- Reverse engineering:  
Learning networks from data
- Application to the yeast cell cycle
- Probabilistic models for  
postgenomic data integration

A

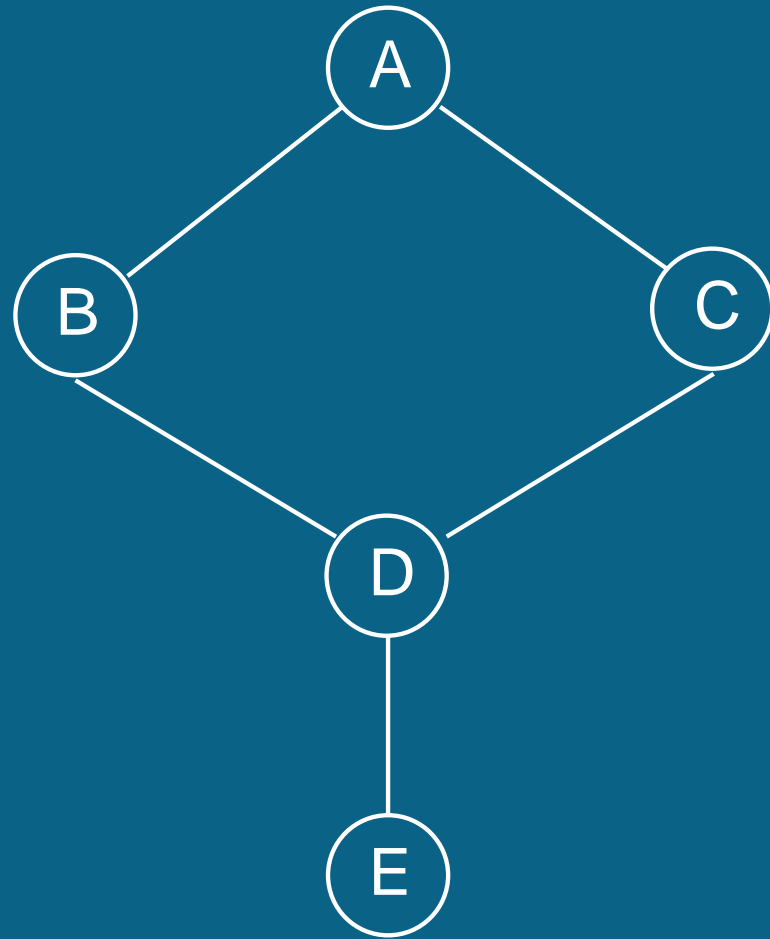
B

C

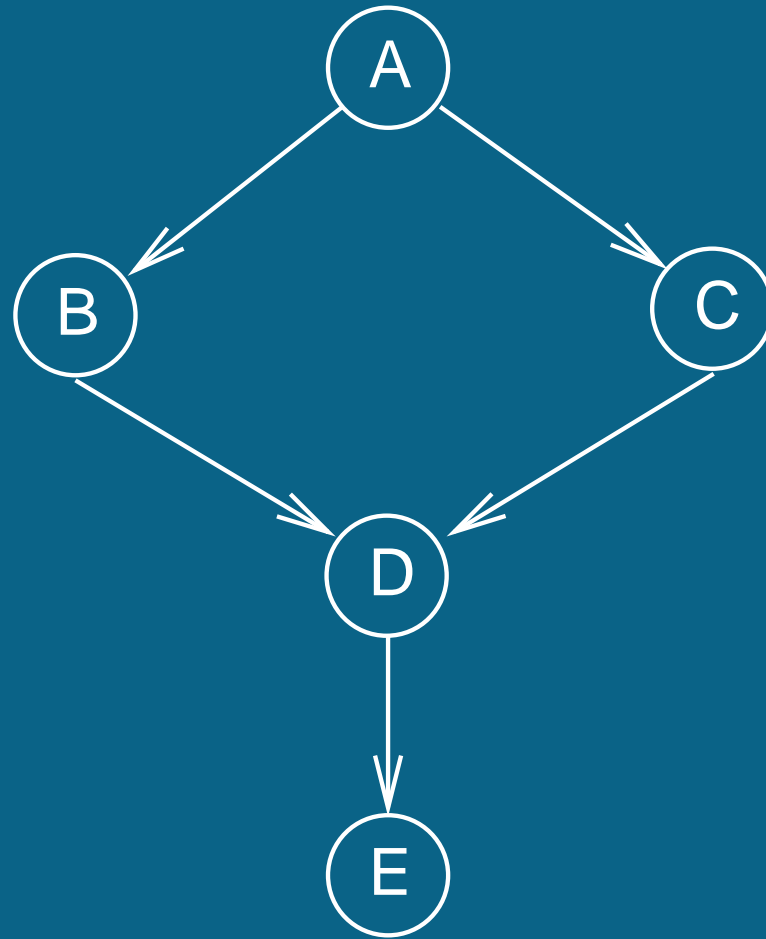
D

E

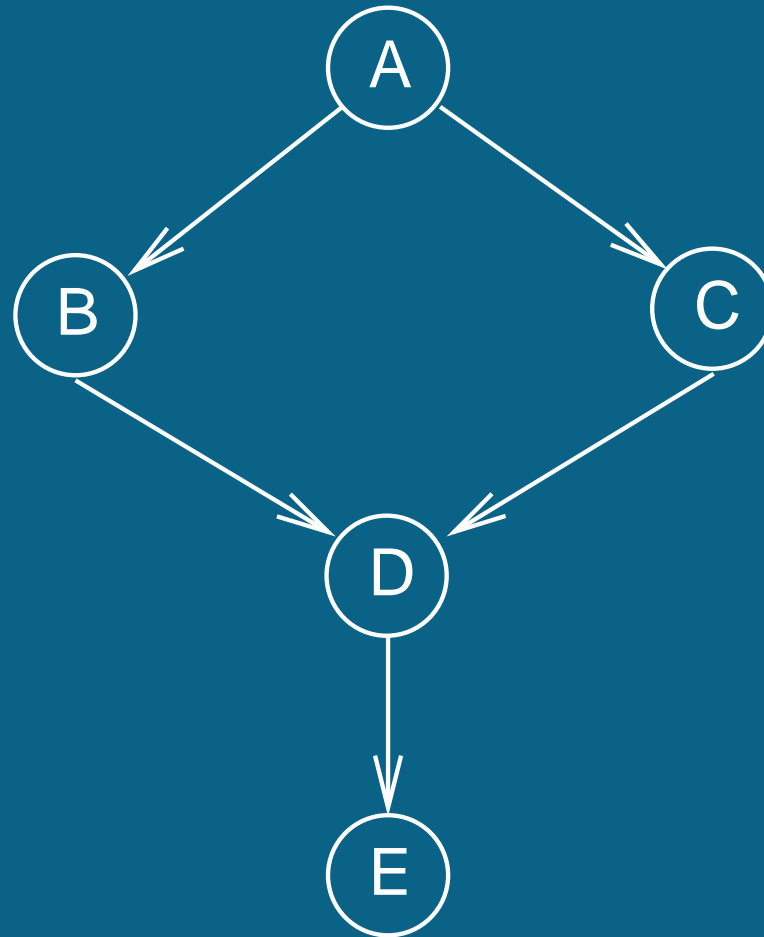
Nodes



Edges

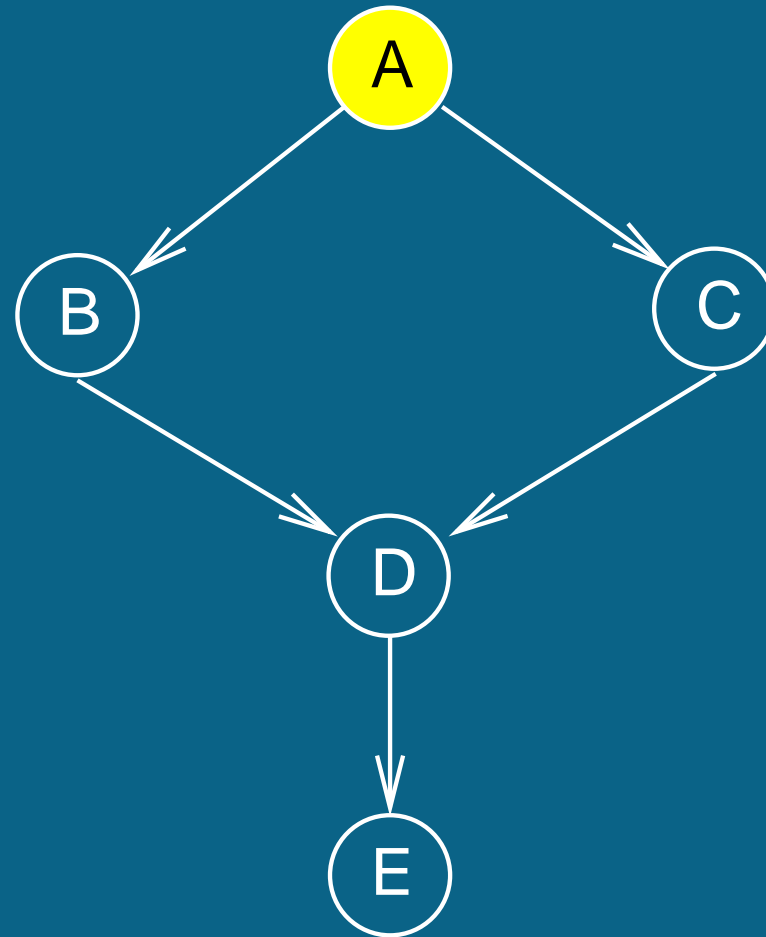


Edges = directed

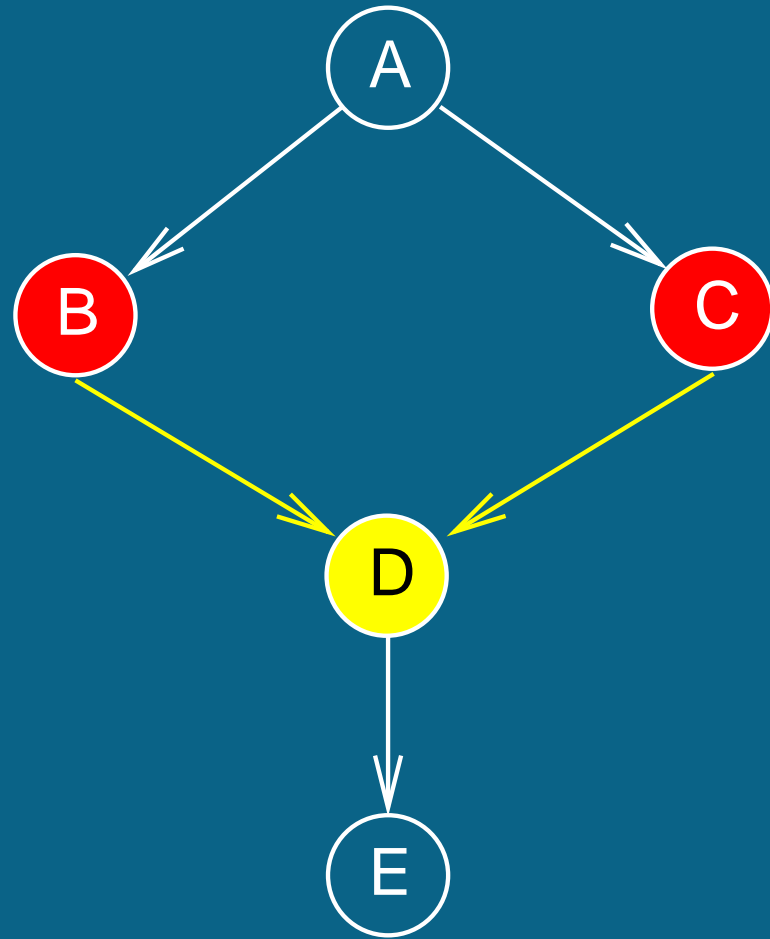


$$P(A, B, C, D, E) = \prod_i P(\text{node}_i | \text{parents}_i)$$

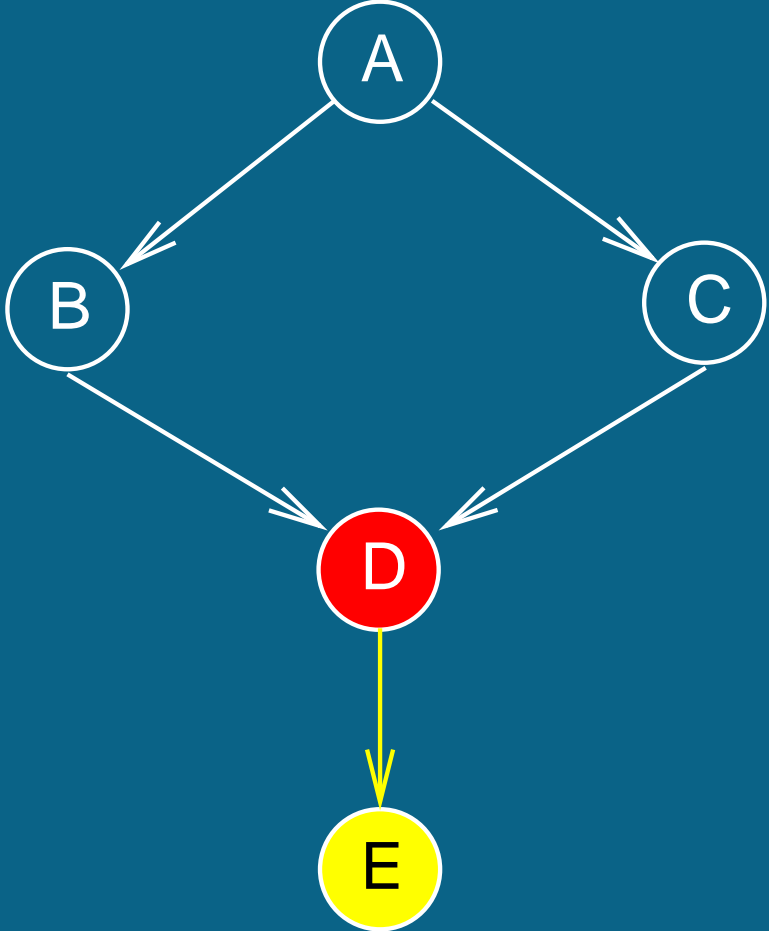
# Biological interpretation



Initiation of cell (sub-)cycle

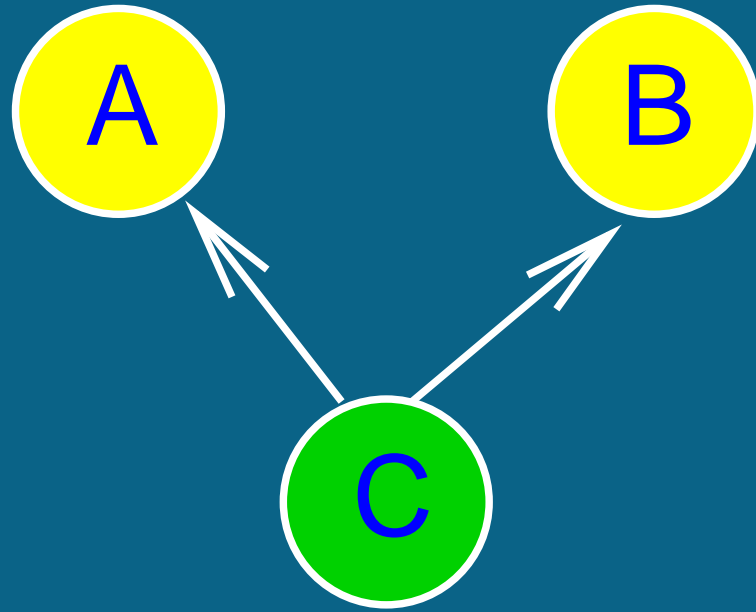


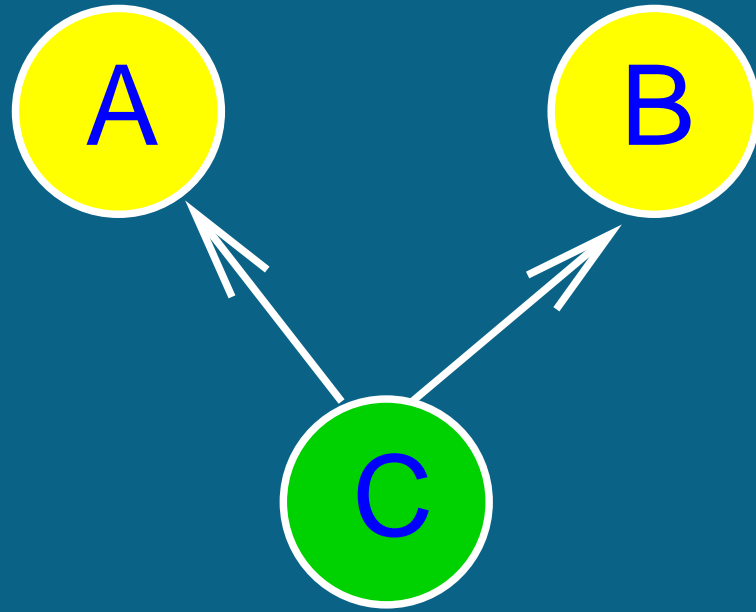
Co-regulation



Mediation

# Conditional independence relations





$$P(A, B, C) = P(A|C)P(B|C)P(C)$$

$$P(A, B|C) = \frac{P(A, B, C)}{P(C)} = P(A|C)P(B|C)$$

But:  $P(A, B) \neq P(A)P(B)$

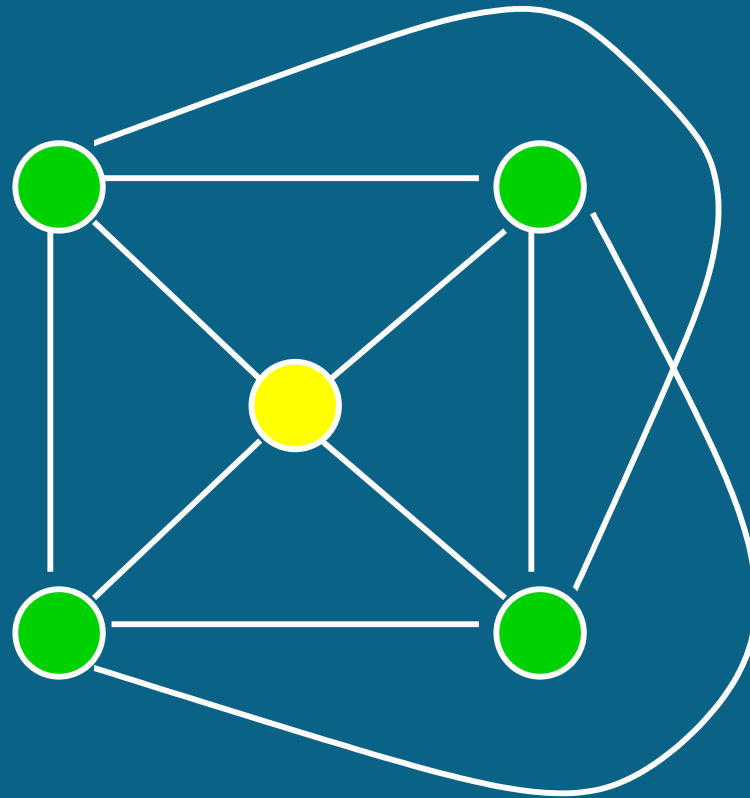
# Biological example

Yeast cell cycle

Clustering

Spellman et al. 1998

Molecular Biology of the Cell 9 (12) :3273-97



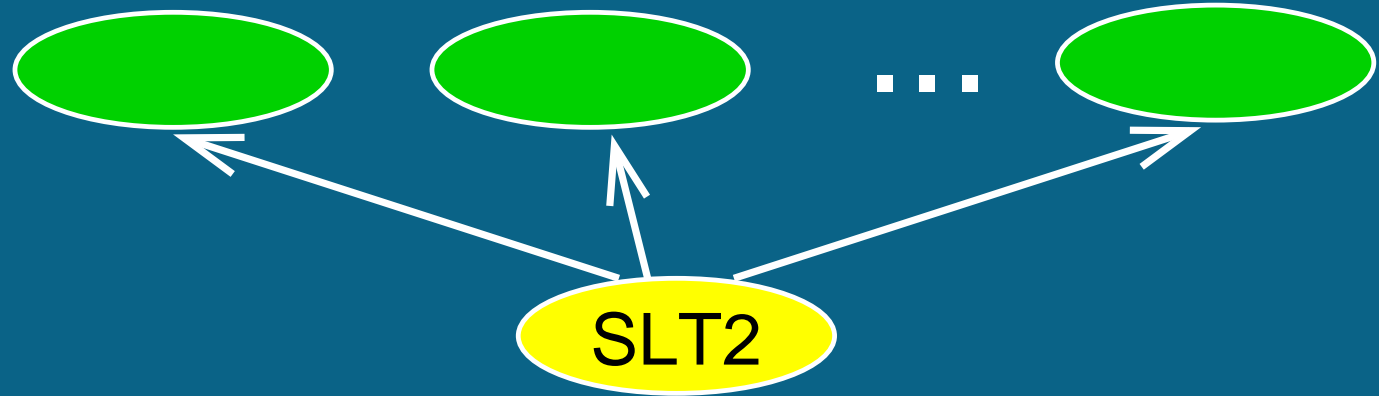
SLT2 clusters with low-osmolarity response genes

# Bayesian networks

Nir Friedman et al. (2000)

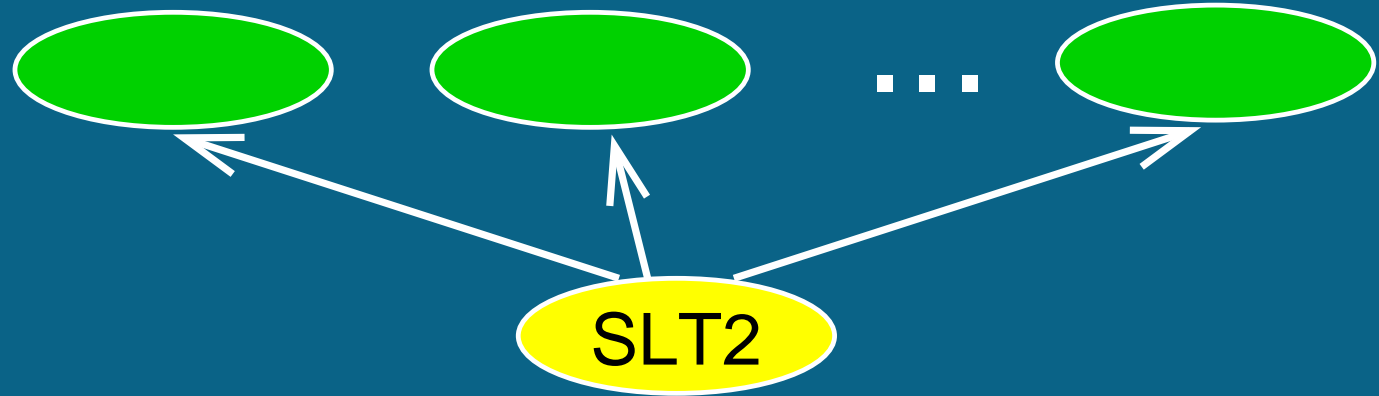
Journal of Computational Biology 7: 601-620

# Low osmolarity response genes

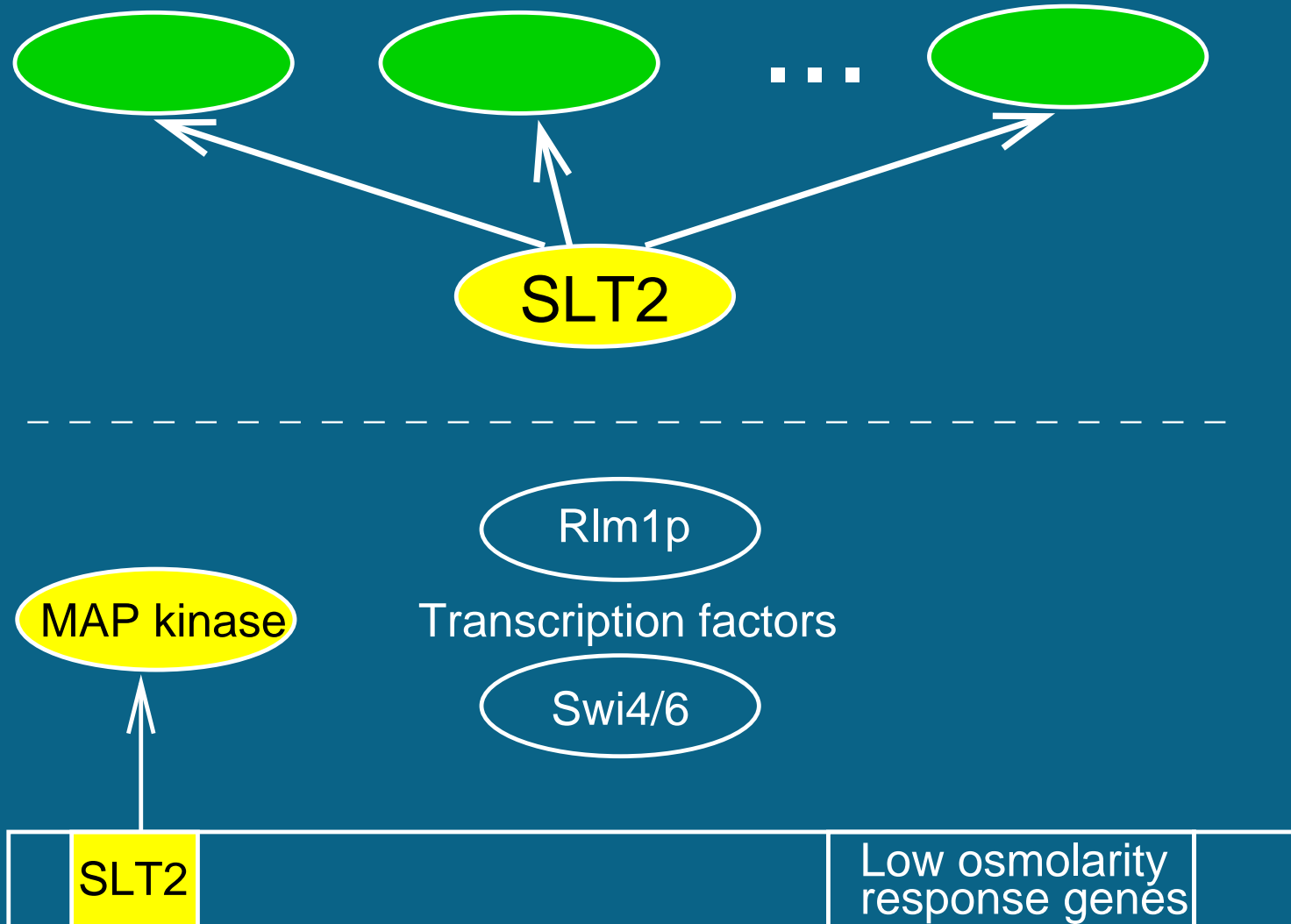


	SLT2		Low osmolarity response genes	
--	------	--	-------------------------------	--

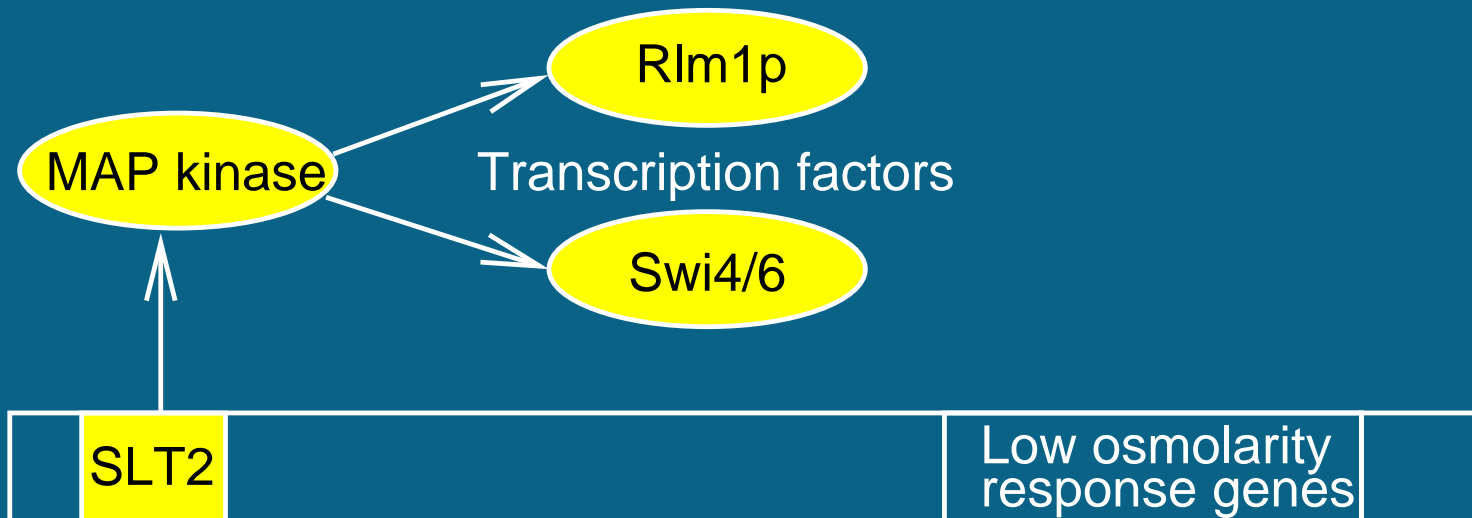
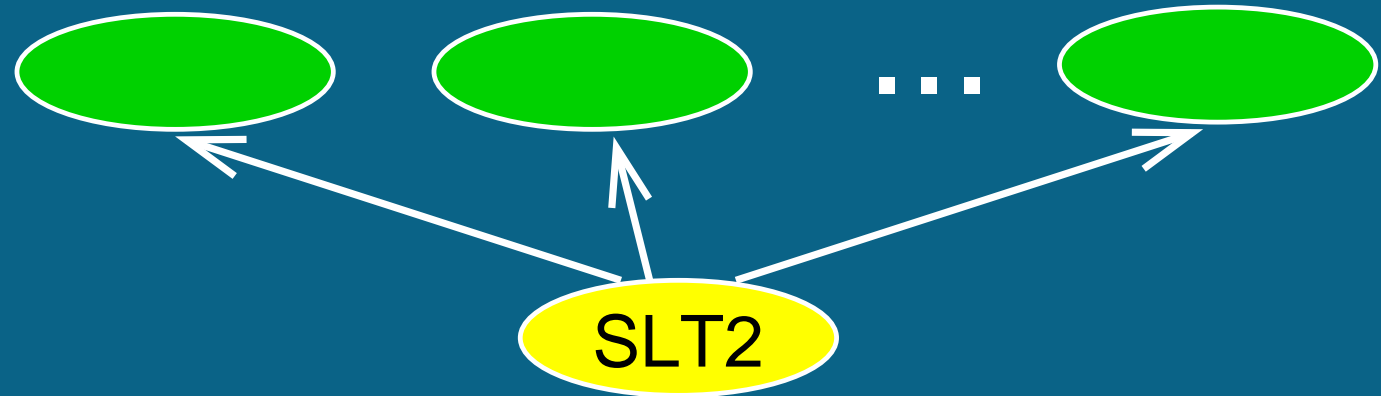
# Low osmolarity response genes



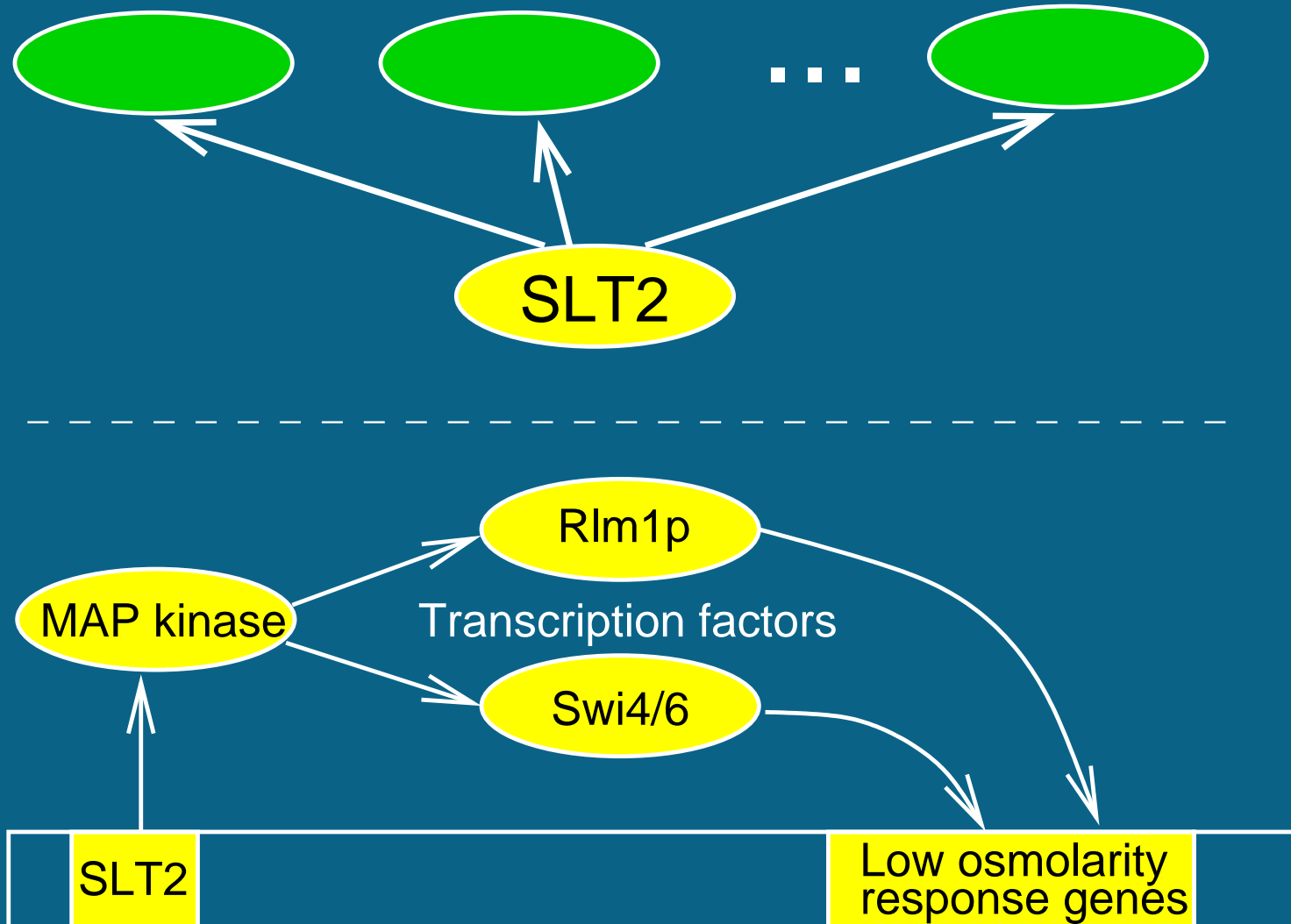
# Low osmolarity response genes



# Low osmolarity response genes



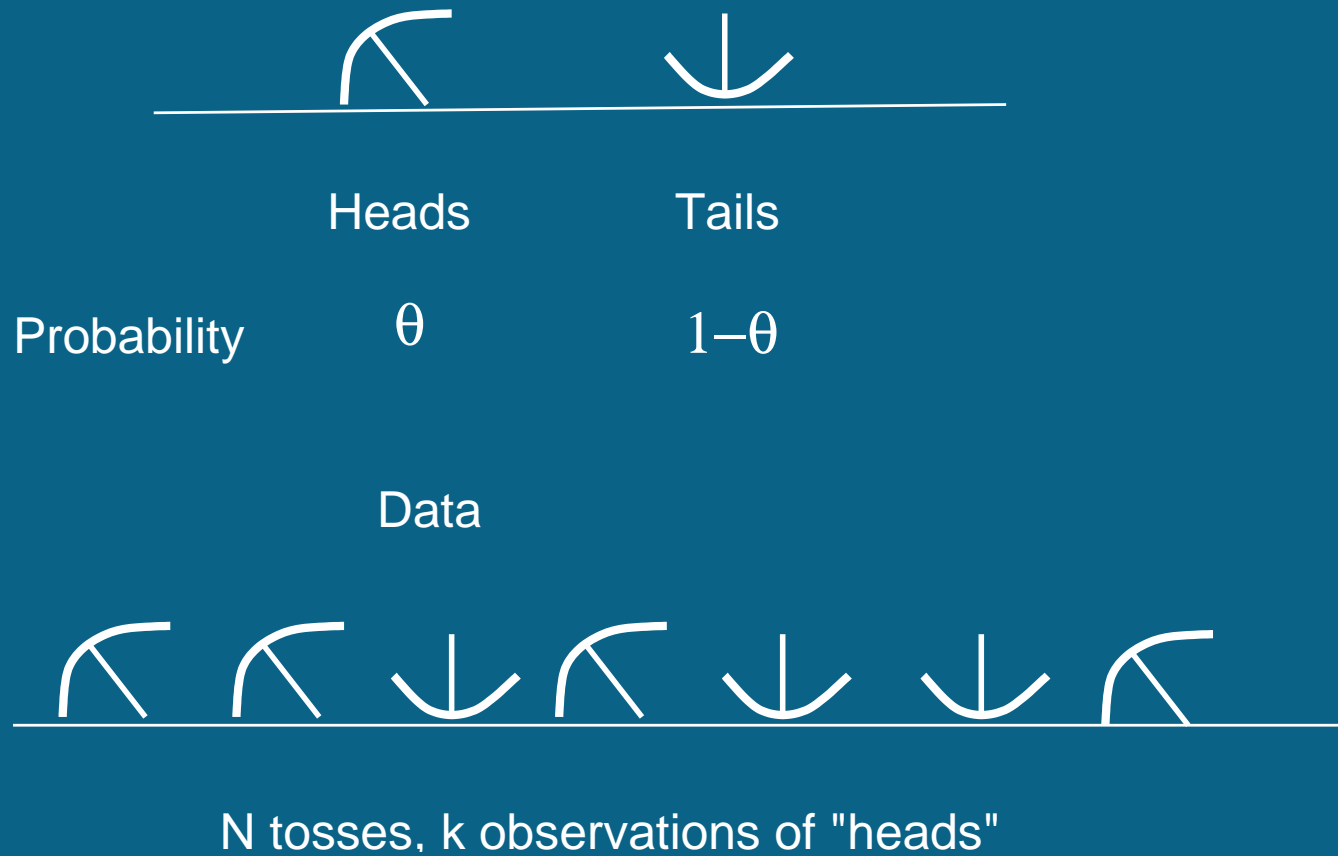
# Low osmolarity response genes



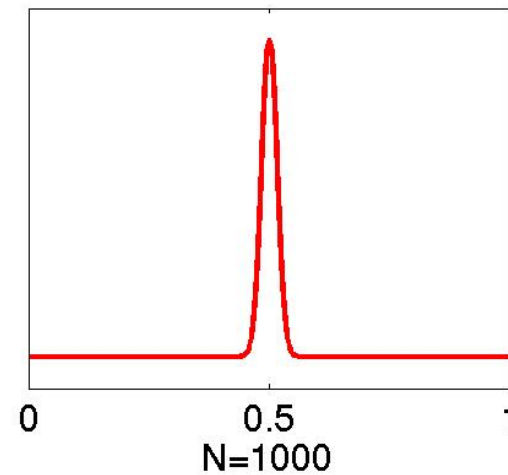
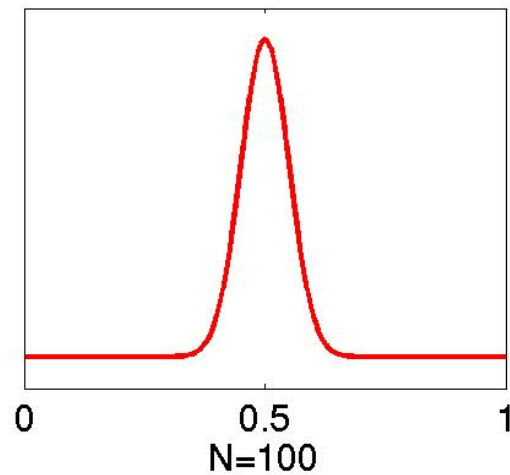
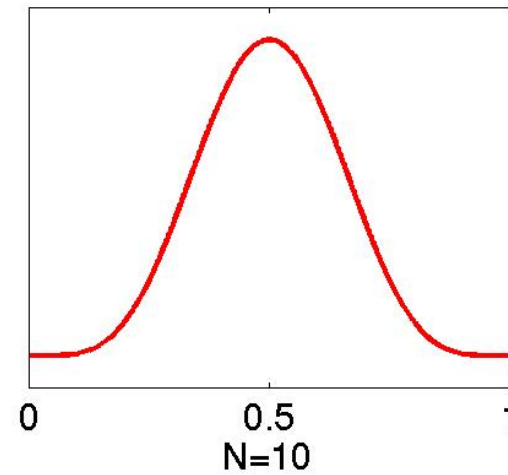
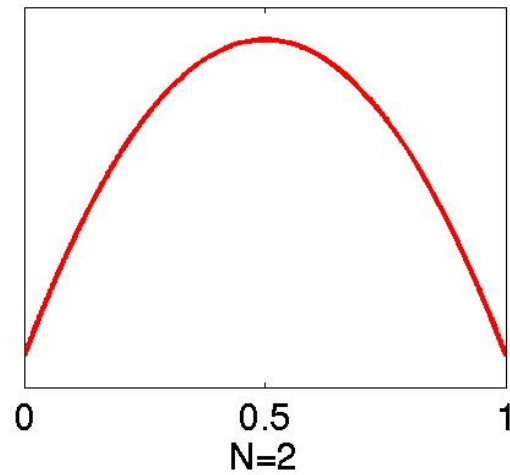
- Recapitulation: Bayesian networks
- **Reverse engineering:  
Learning networks from data**
- Application to the yeast cell cycle
- Probabilistic models for  
postgenomic data integration

# Learning from data

# Learning from data



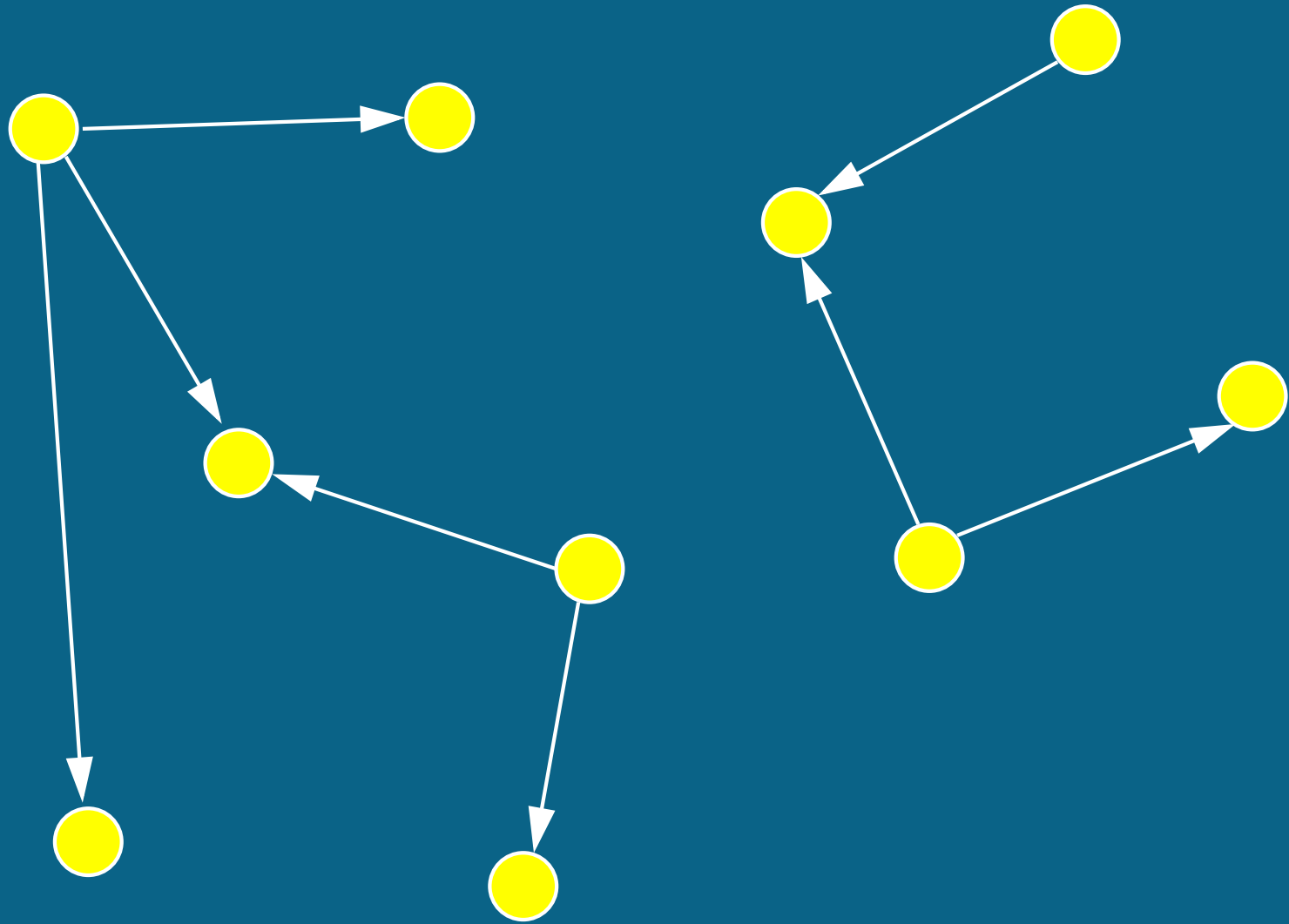
Example:  $P(\theta|D)$  for equal numbers of heads and tails



# Genetic networks:

We want to learn the  
network structure ( $M$ )





## Naive approach

- Compute  $P(M|D)$  for all possible network structures  $M$  .
- Select network structure  $M^*$  that maximizes  $P(M|D)$

## Naive approach

- Compute  $P(M|D)$  for all possible network structures  $M$ .
- Select network structure  $M^*$  that maximizes  $P(M|D)$

### Problem 1:

Number of different network structures increases super-exponentially with the number of nodes.

N of nodes	2	4	6	8	10
N of structures	3	543	$3.7 \times 10^6$	$7.8 \times 10^{11}$	$4.2 \times 10^{18}$

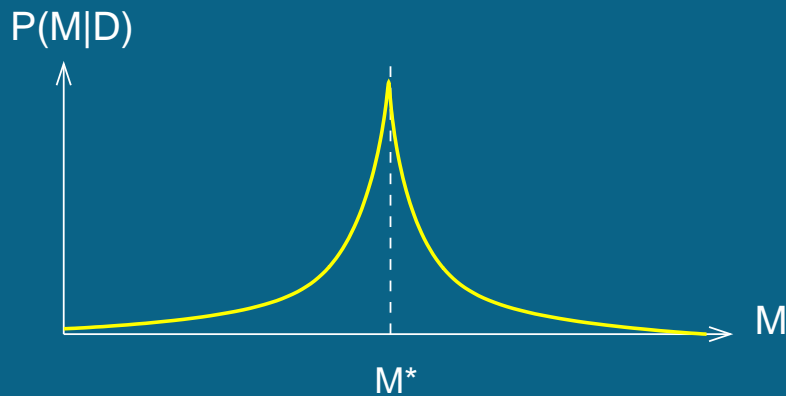
→ Optimization problem intractable for large N of nodes

## Naive approach

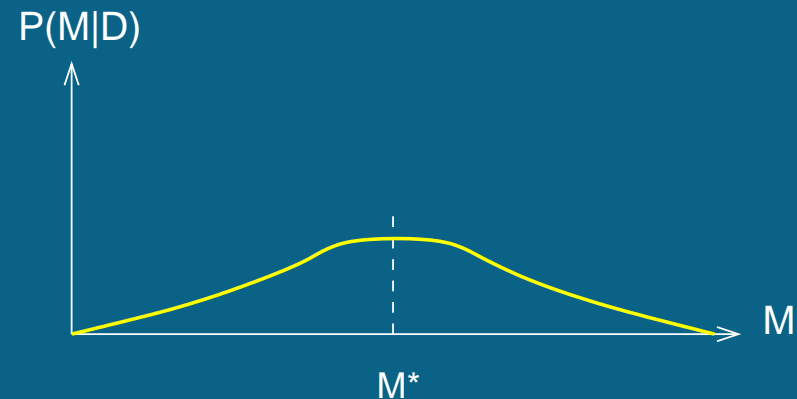
- Compute  $P(M|D)$  for all possible network structures  $M$ .
- Select network structure  $M^*$  that maximizes  $P(M|D)$

### Problem 2:

Data are sparse  $\rightarrow$  Intrinsic uncertainty of inference

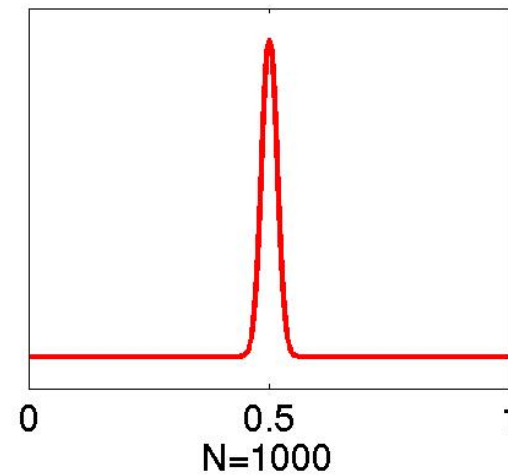
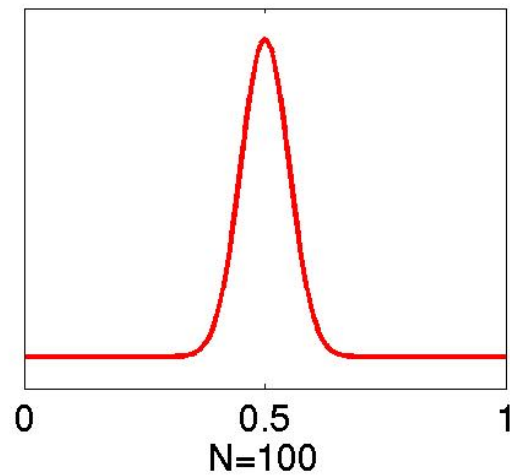
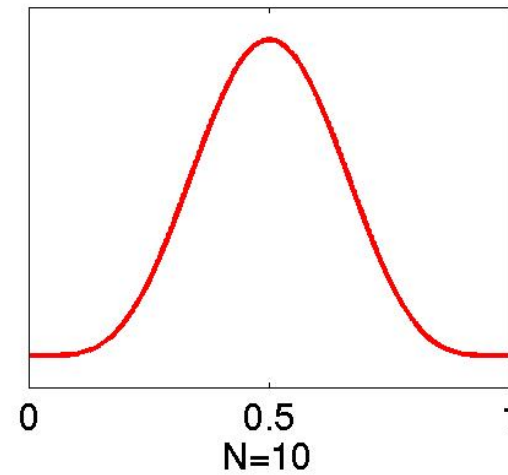
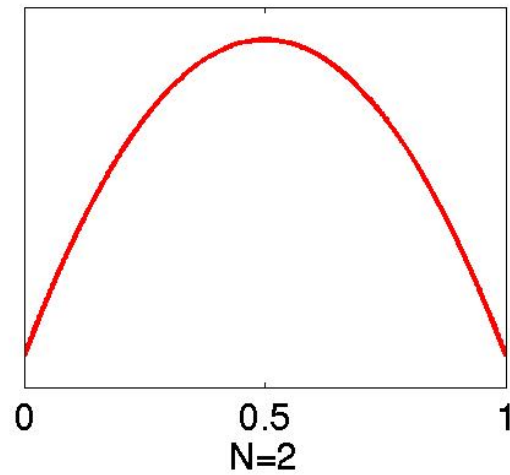


Large data set  $D$ :  
Best network structure  $M^*$  well defined



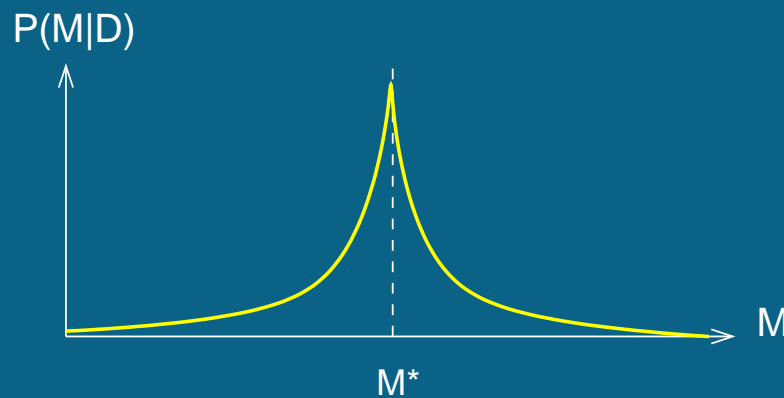
Small data set  $D$ :  
Intrinsic uncertainty about  $M^*$

Example:  $P(\theta|D)$  for equal numbers of heads and tails

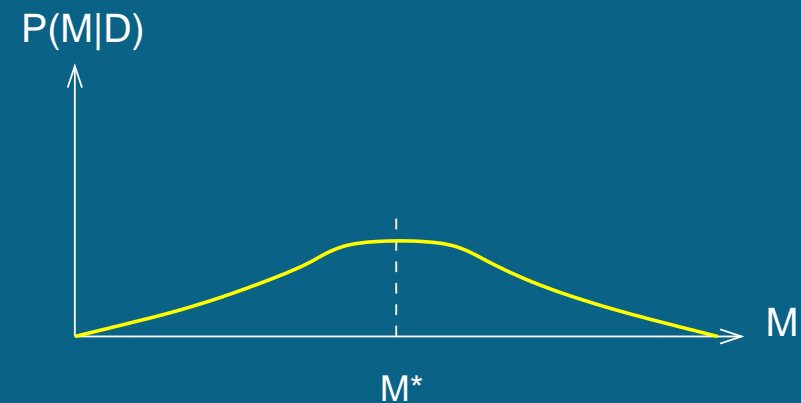


## Problem: Statistical significance of the networks

- **Complex models:** Transcript levels of thousands of genes.
- **Sparse data:** Typically a few dozen samples.



Large data set D:  
Best network structure  $M^*$  well defined



Small data set D:  
Intrinsic uncertainty about  $M^*$

- Posterior probability  $P(M|D)$  diffuse: **Global network** inference is **meaningless**.

**Solution:** Focus on **features** and **subnetworks**

**Feature:** Indicator variable for a property of interest,  
e.g.: Are  $X$  and  $Y$  close neighbours in the network?

$$f(M) = \begin{cases} 1 & \text{if } M \text{ satisfies the feature} \\ 0 & \text{otherwise} \end{cases}$$

**Solution:** Focus on **features** and **subnetworks**

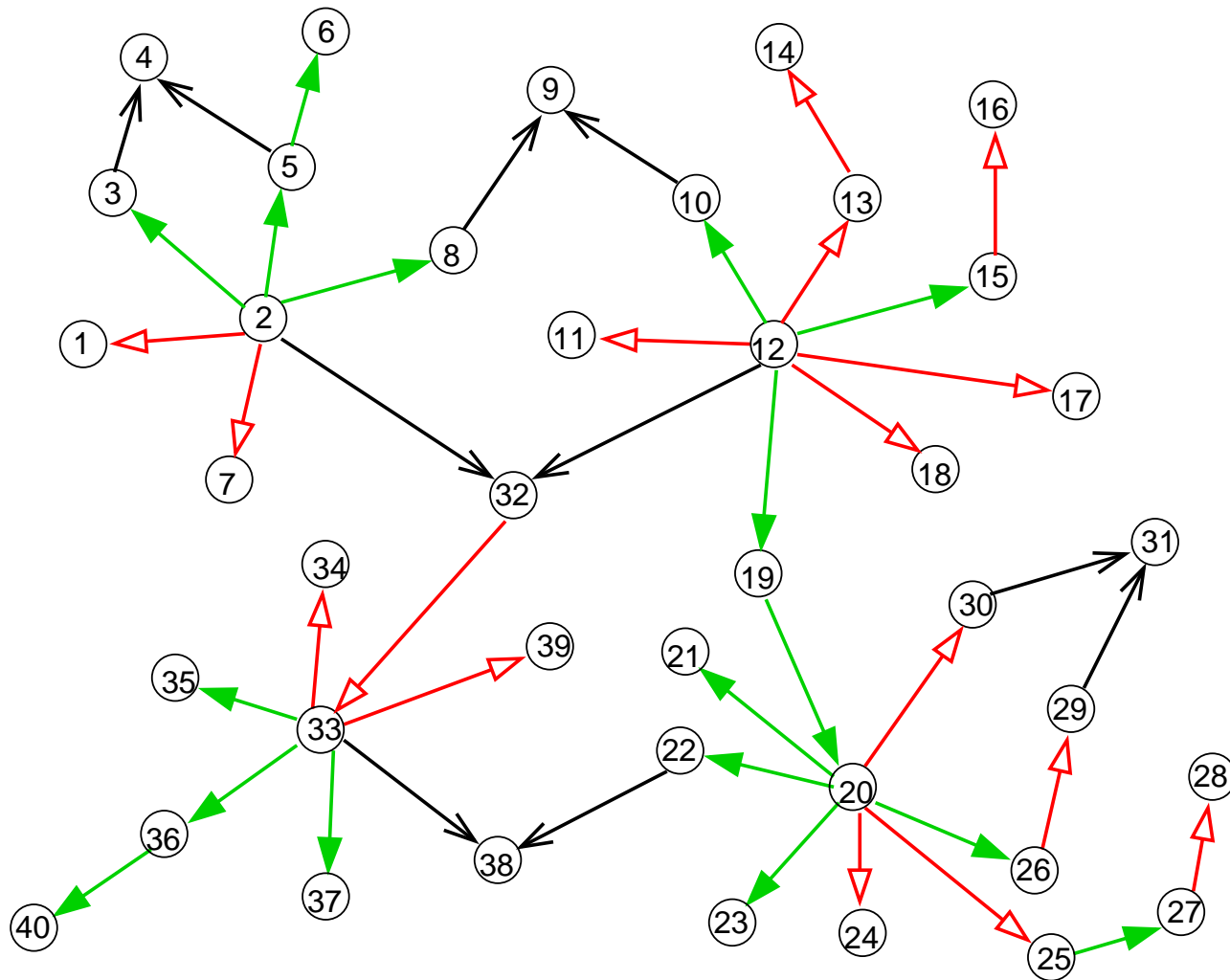
**Feature:** Indicator variable for a property of interest,  
e.g.: Are X and Y close neighbours in the network?

$$f(M) = \begin{cases} 1 & \text{if } M \text{ satisfies the feature} \\ 0 & \text{otherwise} \end{cases}$$

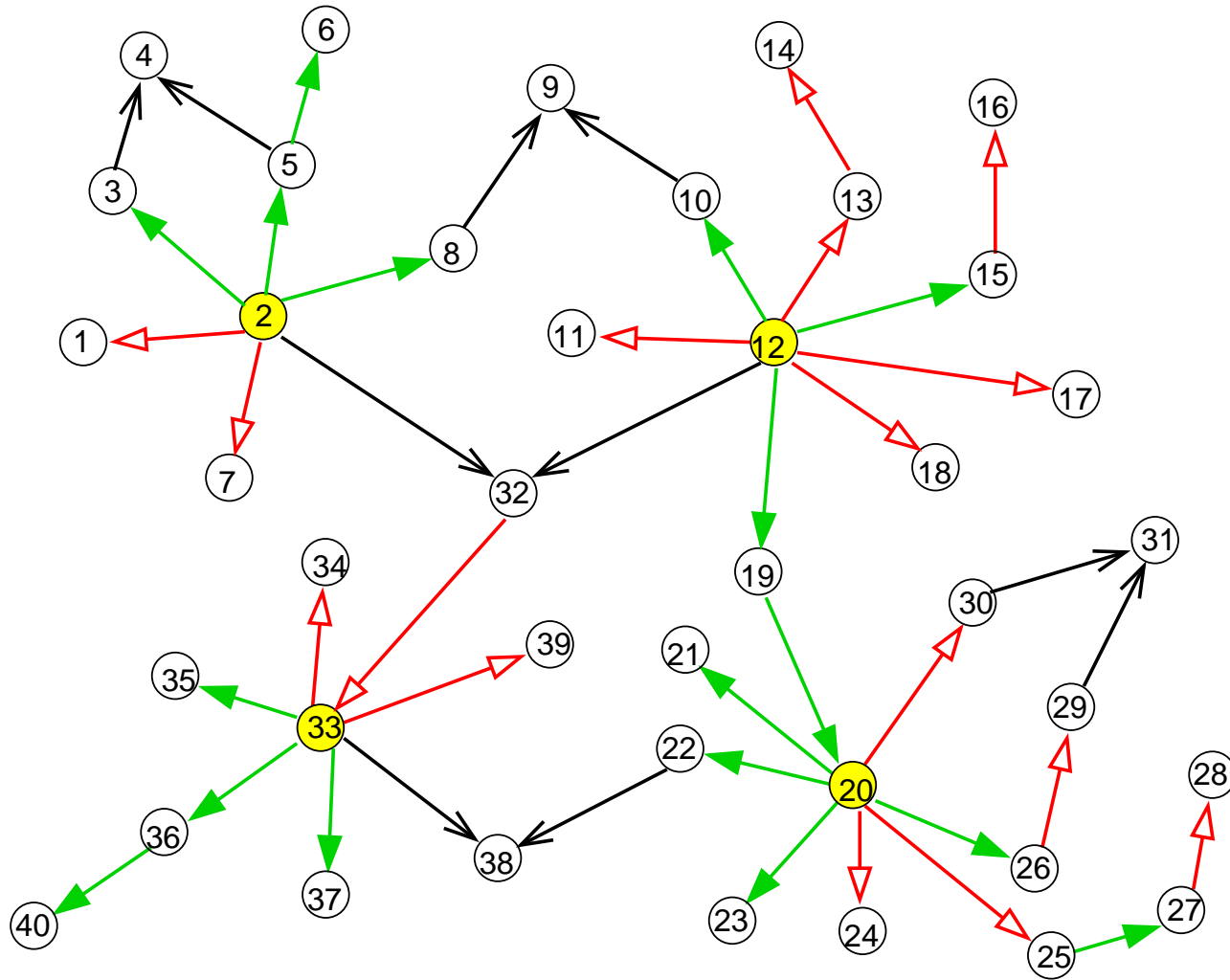
Posterior probability of features:  $P(f|D) = \sum_M f(M)P(M|D)$

assumed to be sufficiently informative.

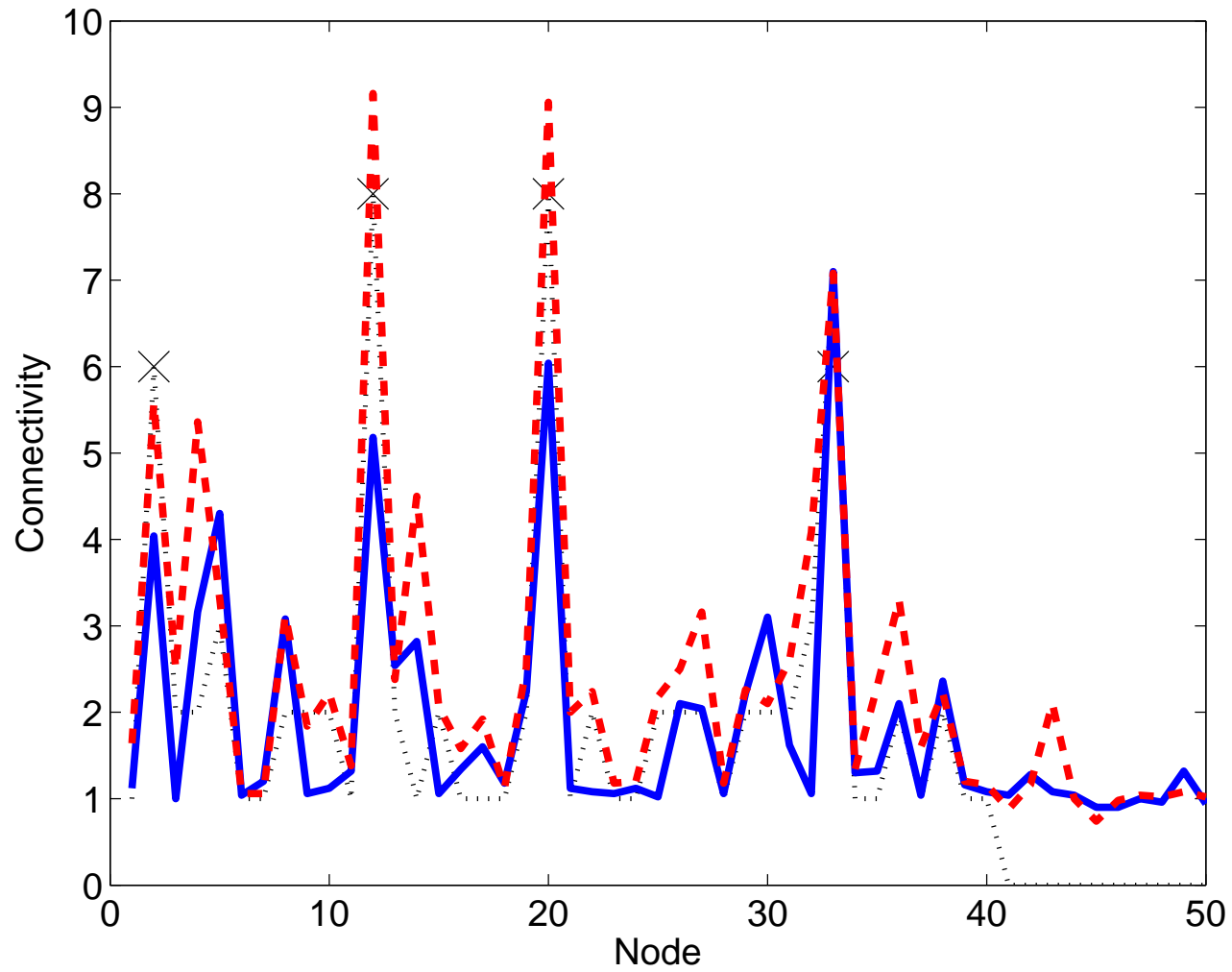
Model network, data set size:  $N = 50$



Model network, data set size:  $N = 50$



# Predicted connectivity spectrum



- Recapitulation: Bayesian networks
- Reverse engineering:  
Learning networks from data
- **Application to the yeast cell cycle**
- Probabilistic models for  
postgenomic data integration

## Experimental Results

- Friedman, Linial, Nachman, Pe'er (2000)  
Journal of Computational Biology 7 (3/4): 601-620  
<http://www.cs.huji.ac.il/labs/compbio/expression/#papers>
- Pe'er, Regev, Elidan, Friedman (2001)  
Bioinformatics S1: 215-224  
<http://www.cs.huji.ac.il/labs/compbio/ismb01/>
- Spellman, Sherlock, Zhang, Iyer, Anders, Eisen, Brown, Botstein, Futcher (1998)  
Molecular Biology of the Cell 9 (12) :3273-97  
<http://cellcycle-www.stanford.edu/>

- **Yeast** cell cycle (*S. cerevisiae*).
- Six time series under different experimental conditions, altogether **76 gene expression measurements**.
- **800 genes**.
- No biological **prior knowledge**.
- Do not take into account the **temporal aspect** of the measurements. Introduce an additional root node representing the cell cycle phase.
- **Discretization**: Underexpressed (-1), normal (0), overexpressed (1).

## Order relations

- Is  $A$  an **ancestor** of  $B$  in all the networks of a given equivalence class?
- Does the **network** contain a **directed path** from  $A$  to  $B$ ?  
Indication that  $A$  might be a **causal ancestor** of  $B$ .

## Order relations

Confidence in  $X$  being an ancestor of  $Y$ :

$$P(X \rightarrow Y | D)$$

**Dominance score** of  $X$ :  $\sum_Y P(X \rightarrow Y | D)$

Genes with high dominance scores are **indicative** of potential **causal** sources of the cell cycle process.

## Order relations

Confidence in  $X$  being an ancestor of  $Y$ :

$$P(X \rightarrow Y | D)$$

**Dominance score** of  $X$ :  $\sum_Y P(X \rightarrow Y | D)$

Genes with high dominance scores are **indicative** of potential **causal** sources of the cell cycle process.

**Finding:** Only a few genes dominate the order.

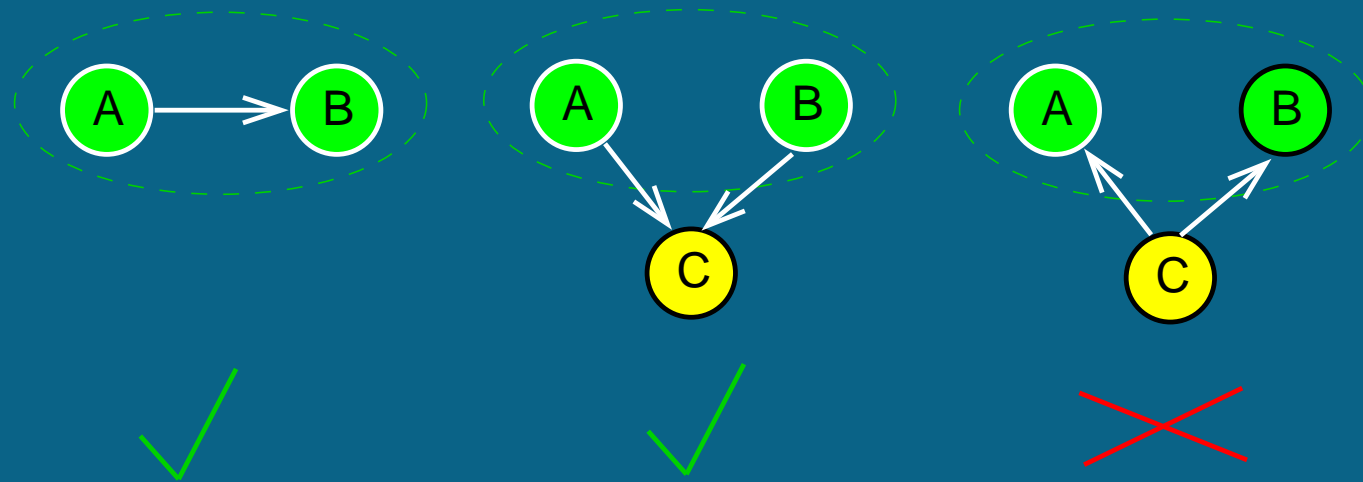
## Dominant genes in the ordering relations

CLN1	Role in cell cycle <b>start</b>
CLN2	Role in cell cycle <b>start</b>
CDC5	Cell cycle <b>control</b> , required for exit from mitosis
RAD53	Cell cycle <b>control</b> : checkpoint function
RFA2	Involved in nucleotide excision <b>repair</b>
PLO30	Required for DNA replication and <b>repair</b>
MSH6	Required for mismatch <b>repair</b> in mitosis and meiosis

DNA repair is associated with **transcription initiation**: DNA areas which are more active in transcription are also repaired more frequently.

## Markov neighbours

- Variables that are not separated by any other measured variable in the domain.



- Indication that two genes are related in some **joint biological interaction or process**.
- Parent-child**: One gene regulating another.
- Spouse relations**: Two genes co-regulating another.

## Markov relations

$P(X \leftrightarrow Y|D)$ : Indication that genes are functionally related.

## Markov relations

$P(X \leftrightarrow Y|D)$ : Indication that genes are **functionally related**.

- Most Markov pairs: **Intracluster pairings** with high correlation in their expression.
- **But:** Genes where  $P(X \leftrightarrow Y|D)$  is high and correlation is low.

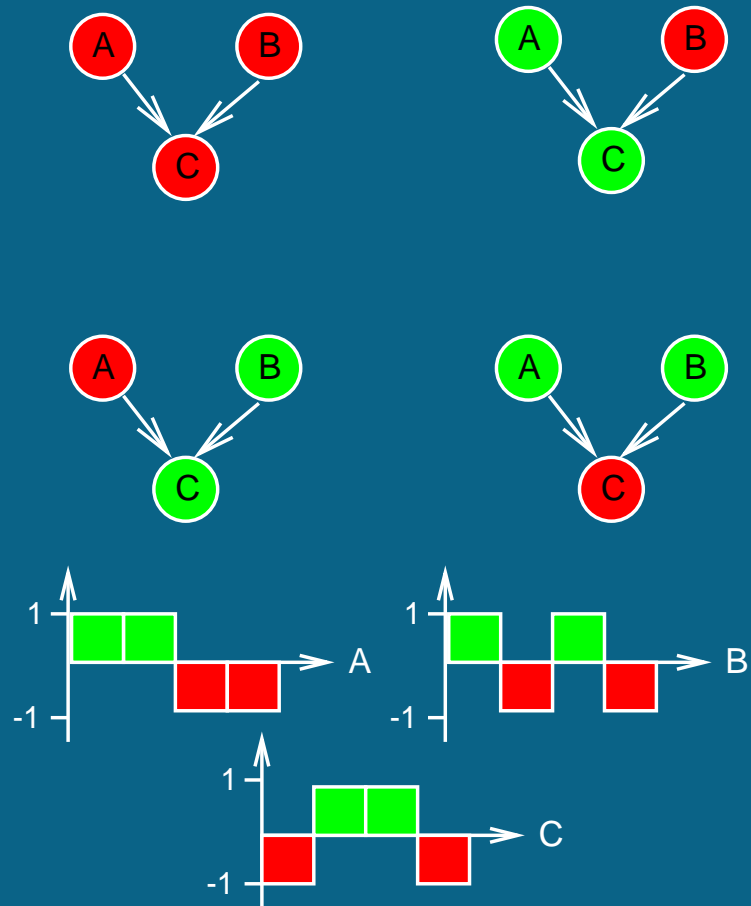
## Markov relations

$P(X \leftrightarrow Y|D)$ : Indication that genes are **functionally related**.

- Most Markov pairs: **Intracluster pairings** with high correlation in their expression.
- **But:** Genes where  $P(X \leftrightarrow Y|D)$  is high and correlation is low.

<b>FAR1</b>	Role in a mating type switch
<b>ASH1</b>	Role in a mating type switch
<b>LAC1</b>	GPI transport protein
<b>YNL300W</b>	Modified by GPI
<b>SAG1</b>	Induces the mating process
<b>MF-ALPHA-1</b>	Participates in the mating process

Advantage of Bayesian networks: **context-specific** and **non-linear**.

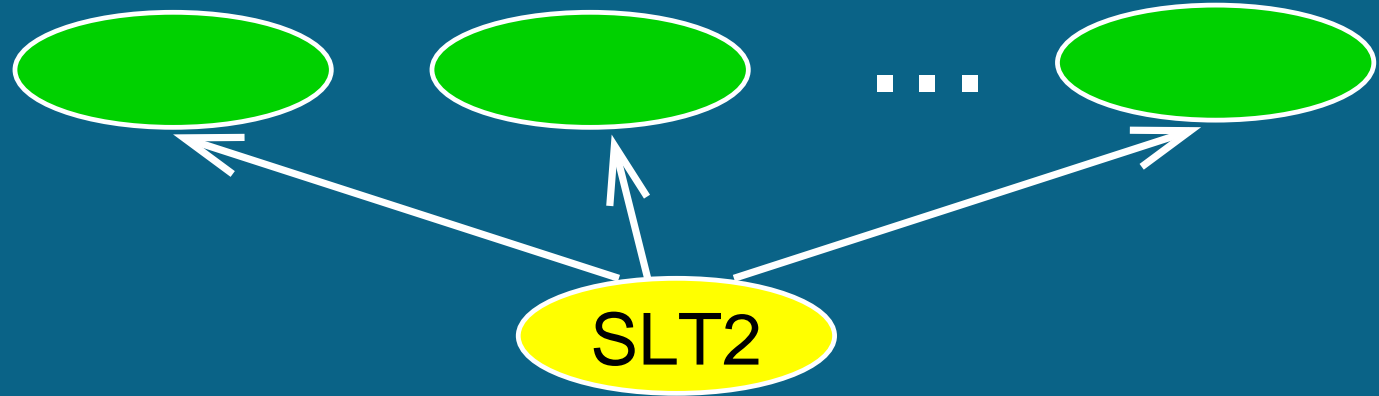


Separator relations

and

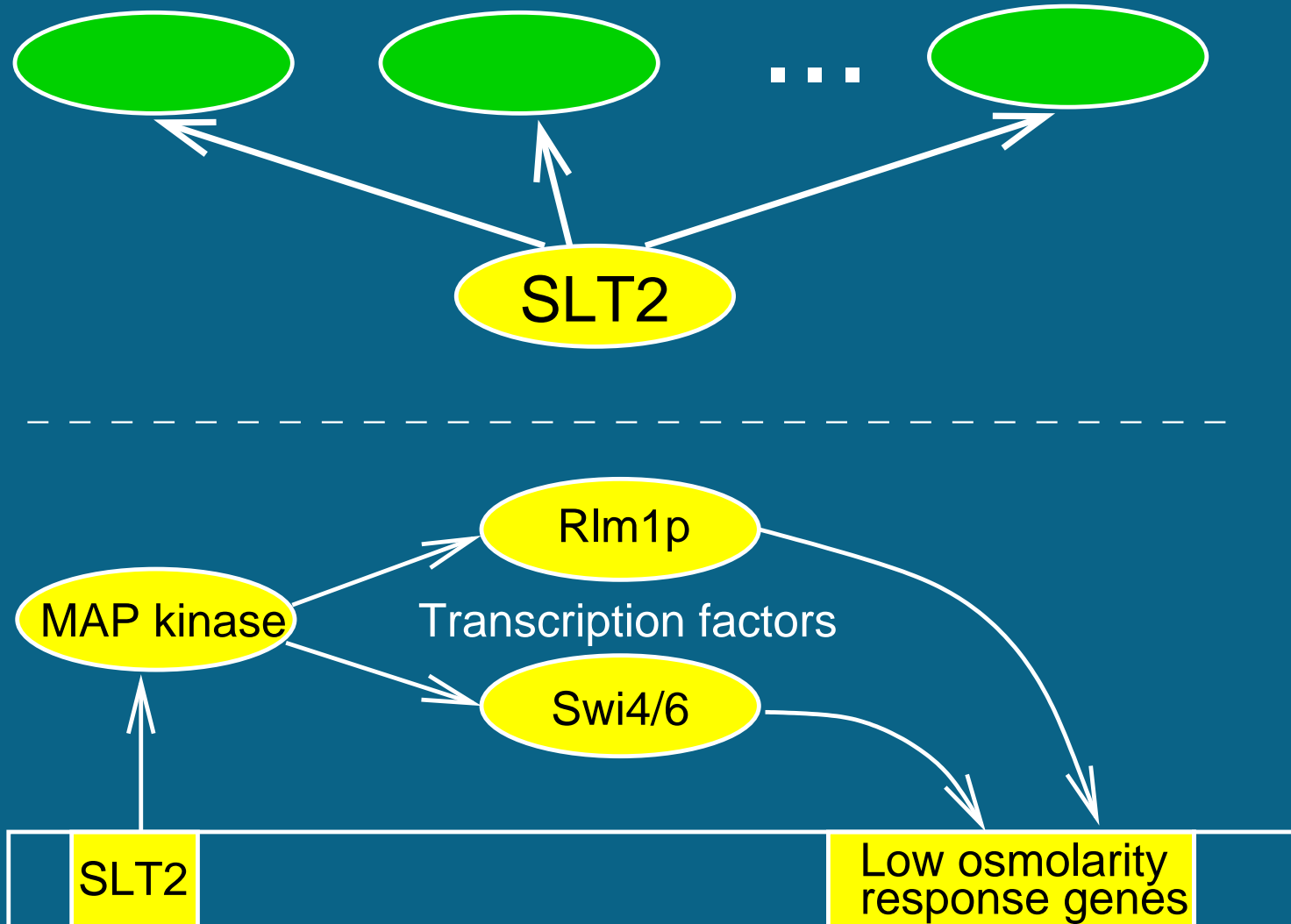
subnetworks

# Low osmolarity response genes



	SLT2		Low osmolarity response genes	
--	------	--	-------------------------------	--

# Low osmolarity response genes



# Conclusions

# Conclusions

- Learning the **global network** → impossible.
- Intrinsic **uncertainty** due to lack of data.

# Conclusions

- Learning the **global network** → impossible.
- Intrinsic **uncertainty** due to lack of data.
- Inference of **local substructures** possible.

- Recapitulation: Bayesian networks
- Reverse engineering:  
Learning networks from data
- Application to the yeast cell cycle
- **Probabilistic models for  
postgenomic data integration**

Integrated analysis  
of  
regulatory networks

# Integrated analysis of regulatory networks

- Expression data alone are not sufficient.
- Combining multiple sources of information yields complementary constraints.

# Combining promoter sequences and gene expression data

# Combining promoter sequences and gene expression data

Conventional approach:

- Find clusters of co-expressed genes.
- Identify regulatory elements by searching for common over-represented motifs in the promoter regions of these genes.

# Shortcomings of the conventional algorithm

Microarray  
data

Model

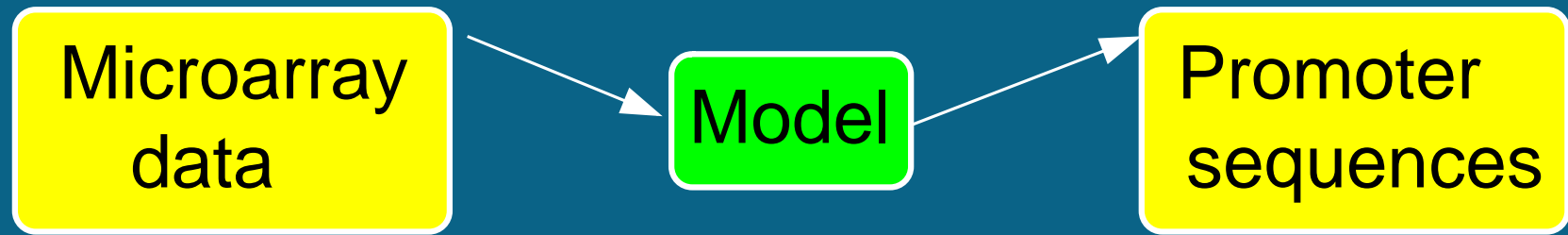
Promoter  
sequences

Microarray  
data

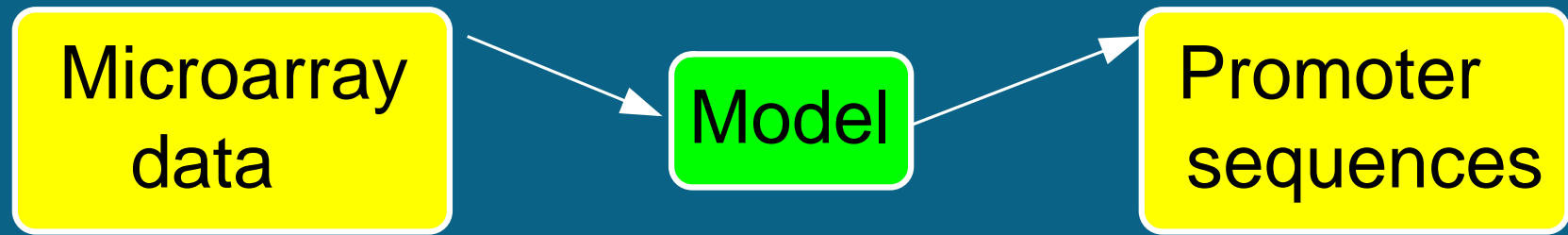


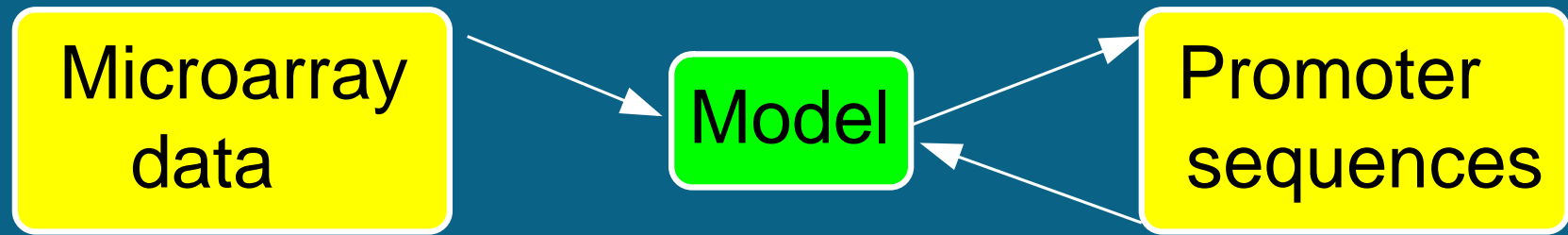
Model

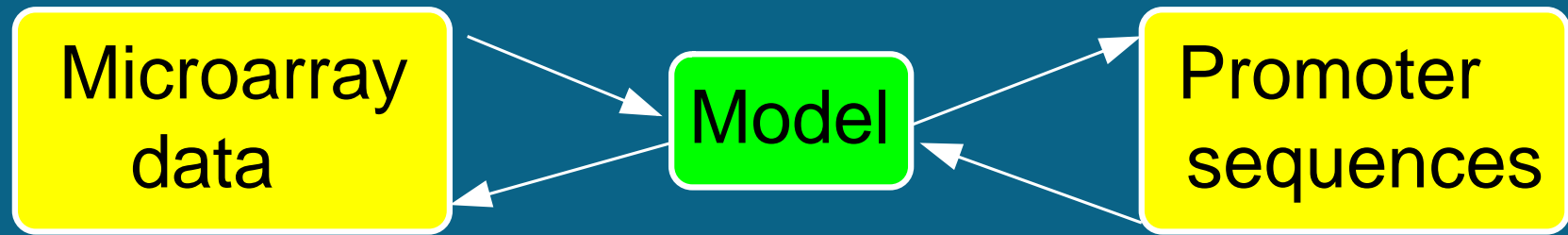
Promoter  
sequences



# Segal's unifying probabilistic model







Segal, Yelensky, Koller (2003)

Bioinformatics 19

*Saccharomyces cerevisiae*

# Experiment 1

173 microarrays, measuring responses to various stress conditions (Gasch et al. 2000)

- **Unified probabilistic model:** 45% of the predicted motifs are known.
- **Conventional algorithms:** 20% of the predicted motifs are known.

## Experiment 2

77 microarrays, expression during the cell cycle  
(Spellman et al. 1998)

- **Unified probabilistic model:** 56% of the predicted motifs are known.
- **Conventional algorithms:** 30% of the predicted motifs are known.

# Literature

- Application of Bayesian networks to microarray data analysis.  
Friedman et al., J. Comp. Biol. 2000
- Probabilistic models for post-genomic data integration.  
Segal et al. (PhD Stanford, August 2004)

Dirk Husmeier  
Richard Dybowski  
Stephen Roberts (Eds.)

# Probabilistic Modeling in Bioinformatics and Medical Informatics

 Springer