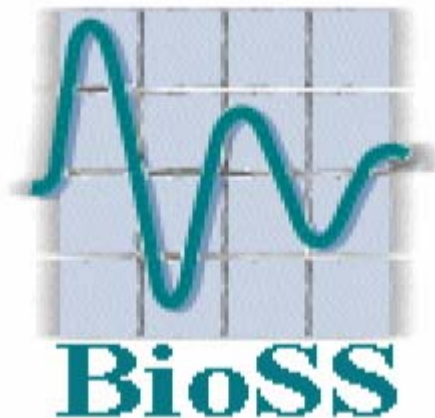
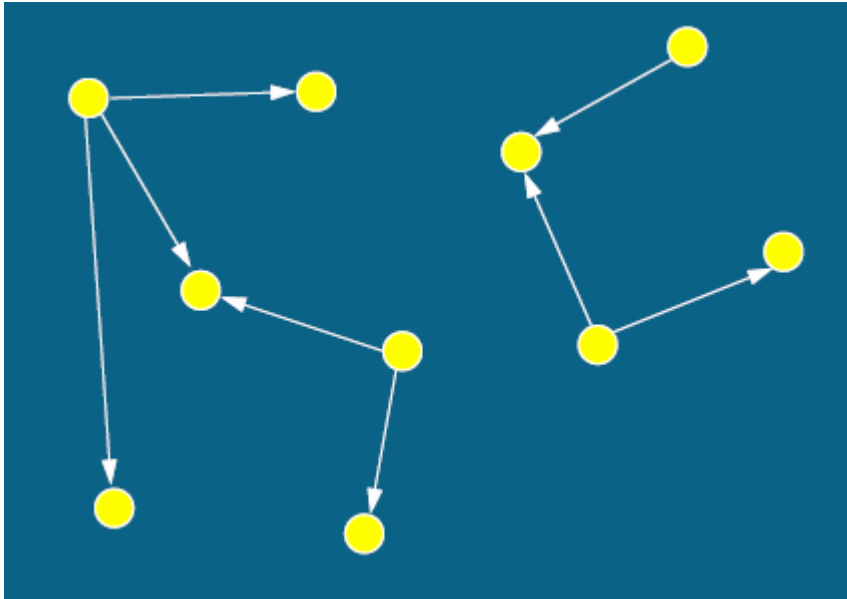


# Learning Bayesian networks from data

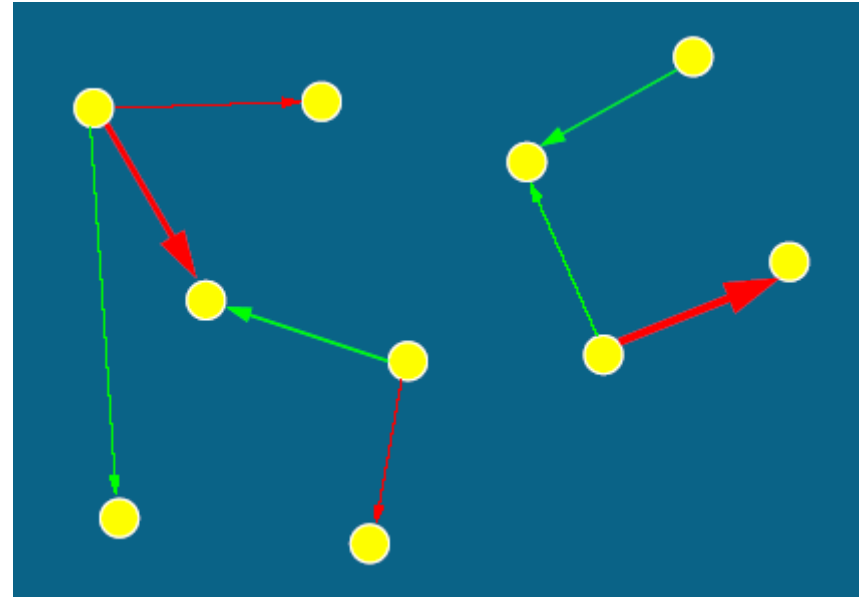
**Dirk Husmeier**



Model  $\mathcal{M}$



Parameters  $\mathbf{q}$



$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\mathbf{q}, \mathcal{M})P(\mathbf{q}|\mathcal{M})d\mathbf{q}$$

Integral **analytically tractable!**

**BDe:**

**Multinomial with a Dirichlet prior**

Heckerman, Geiger, Chickering (1995)

Learning Bayesian Networks:

The Combination of Knowledge and Statistical Data

*Machine learning 20, 245-274*

**BGe:**

**Linear Gaussian with a normal-gamma prior**

Geiger and Heckerman (1994)

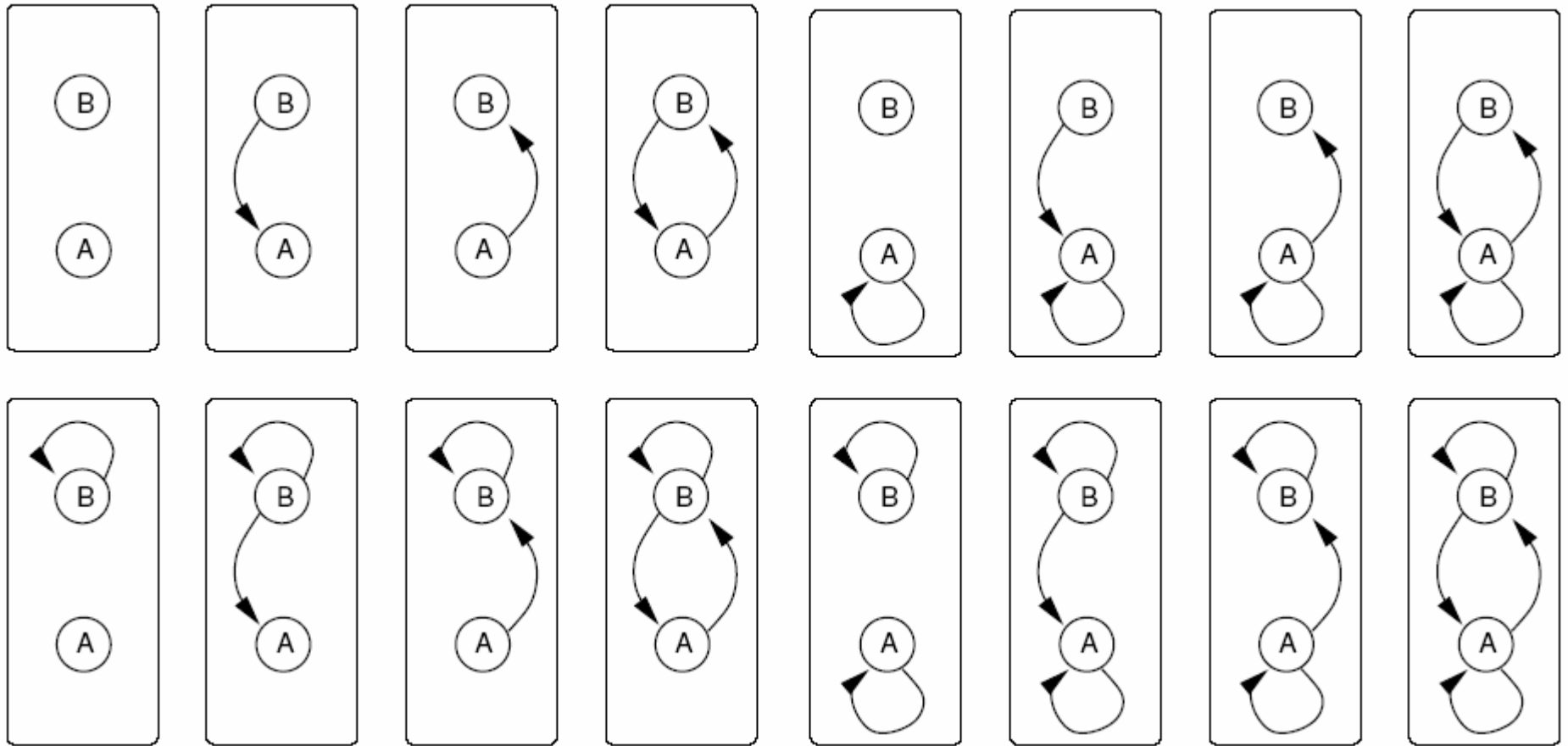
Learning Gaussian networks

*Proceedings of the Tenth Conference on*

*Uncertainty in Artificial Intelligence*

Morgan Kaufmann publisher, San Francisco, 235-243

Example: 2 genes  $\rightarrow$  16 different network structures

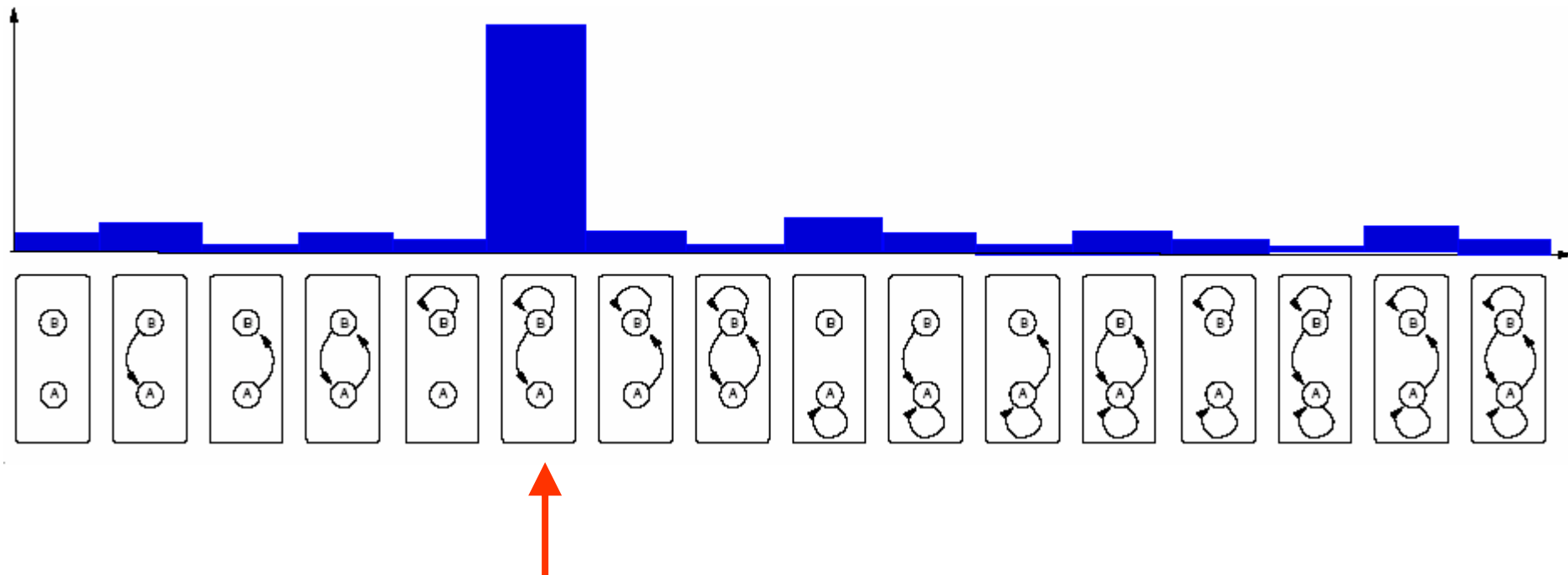


Best network: maximum score  $P(\mathcal{D}|\mathcal{M})$

# Identify the best network structure

Ideal scenario: Large data sets, low noise

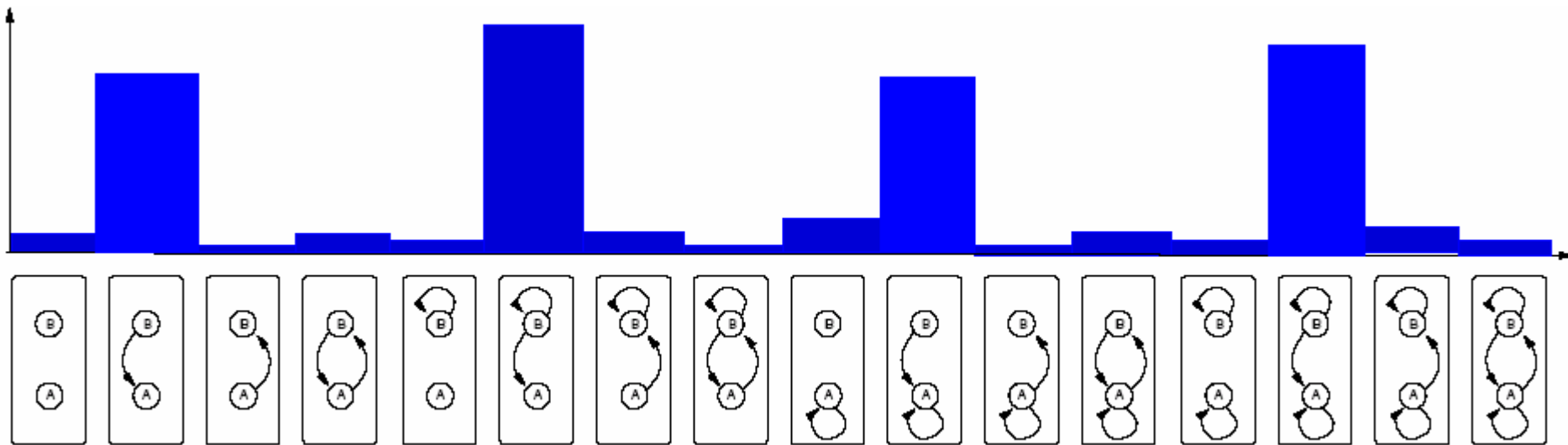
$$P(\mathcal{D}|\mathcal{M})$$



# Uncertainty about the best network structure

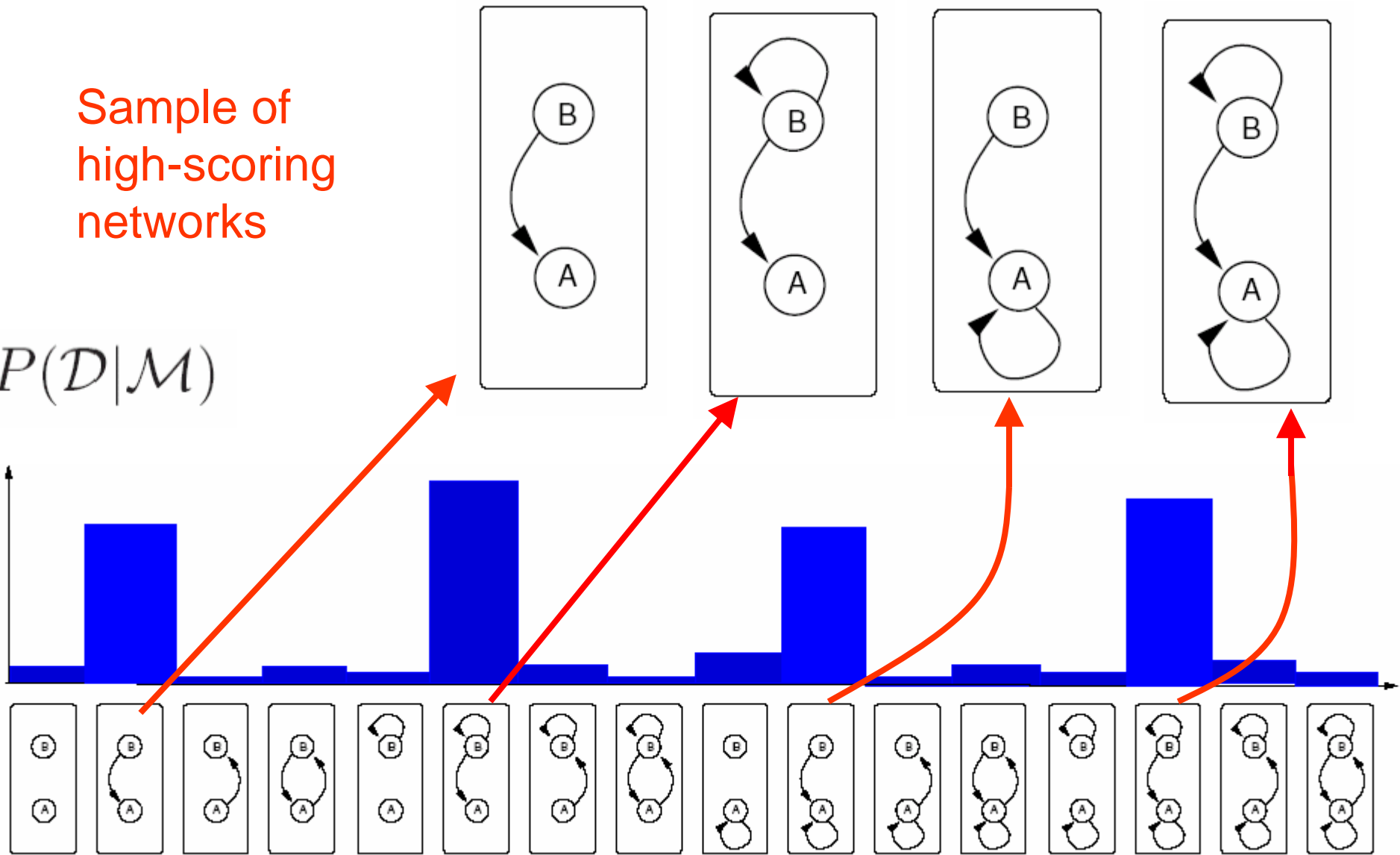
Limited number of experimental replications,  
high noise

$$P(\mathcal{D}|\mathcal{M})$$

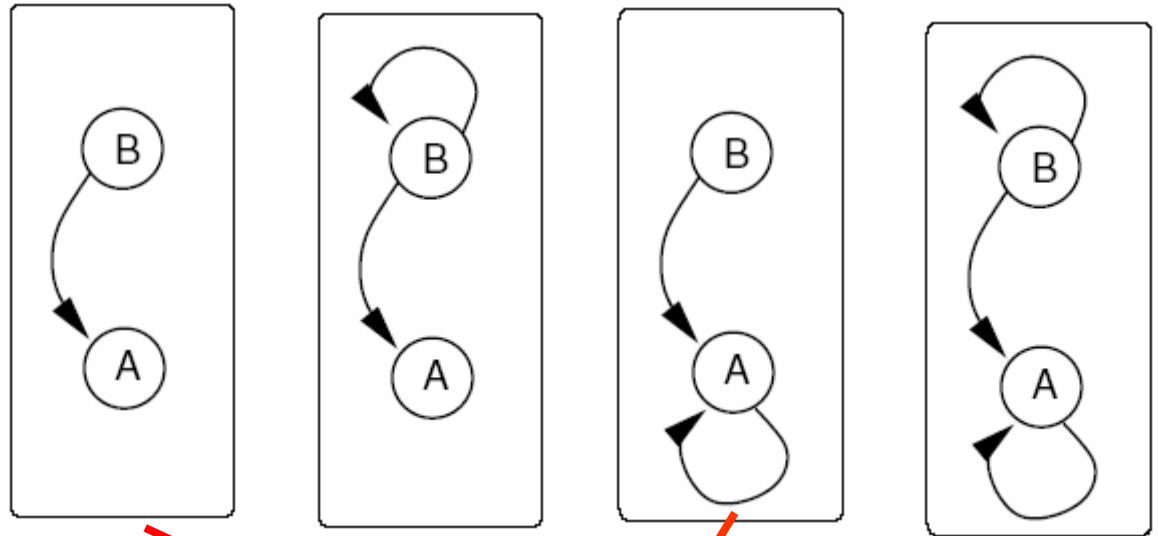


Sample of high-scoring networks

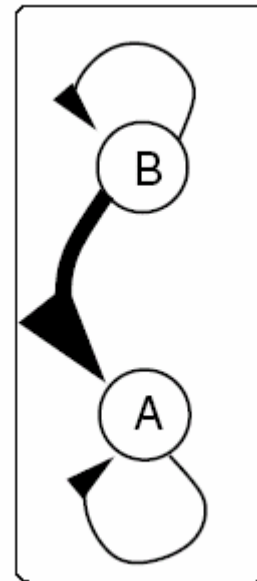
$P(\mathcal{D}|\mathcal{M})$



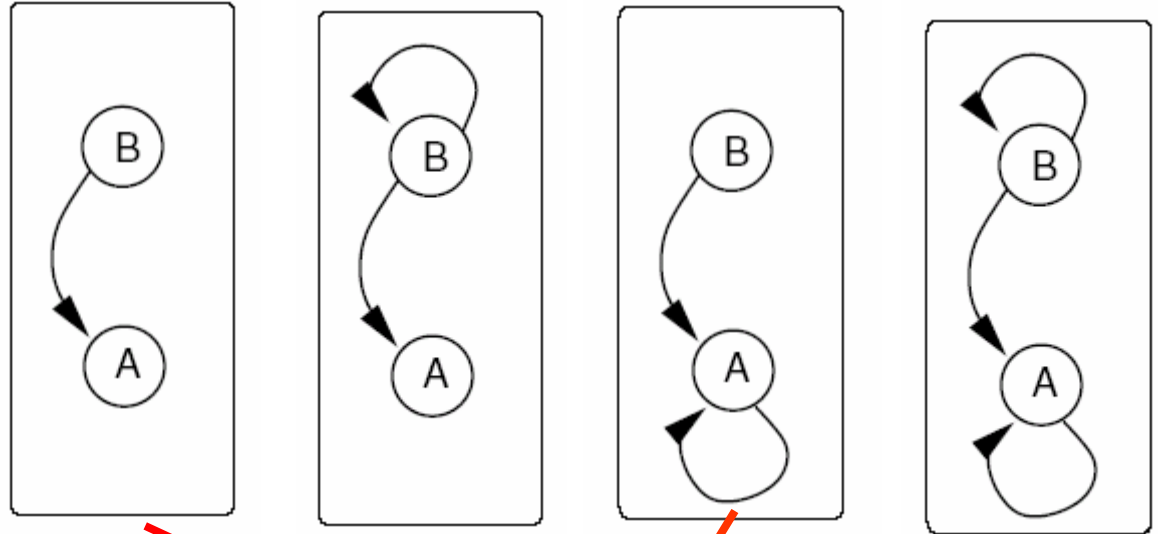
Sample of  
high-scoring  
networks



Feature extraction,  
e.g. marginal posterior  
probabilities of the edges



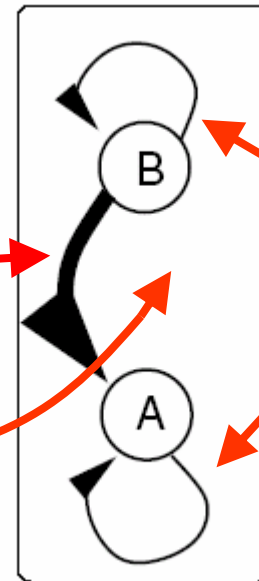
Sample of high-scoring networks



Feature extraction,  
e.g. marginal posterior probabilities of the edges

High-confident edge

High-confident non-edge



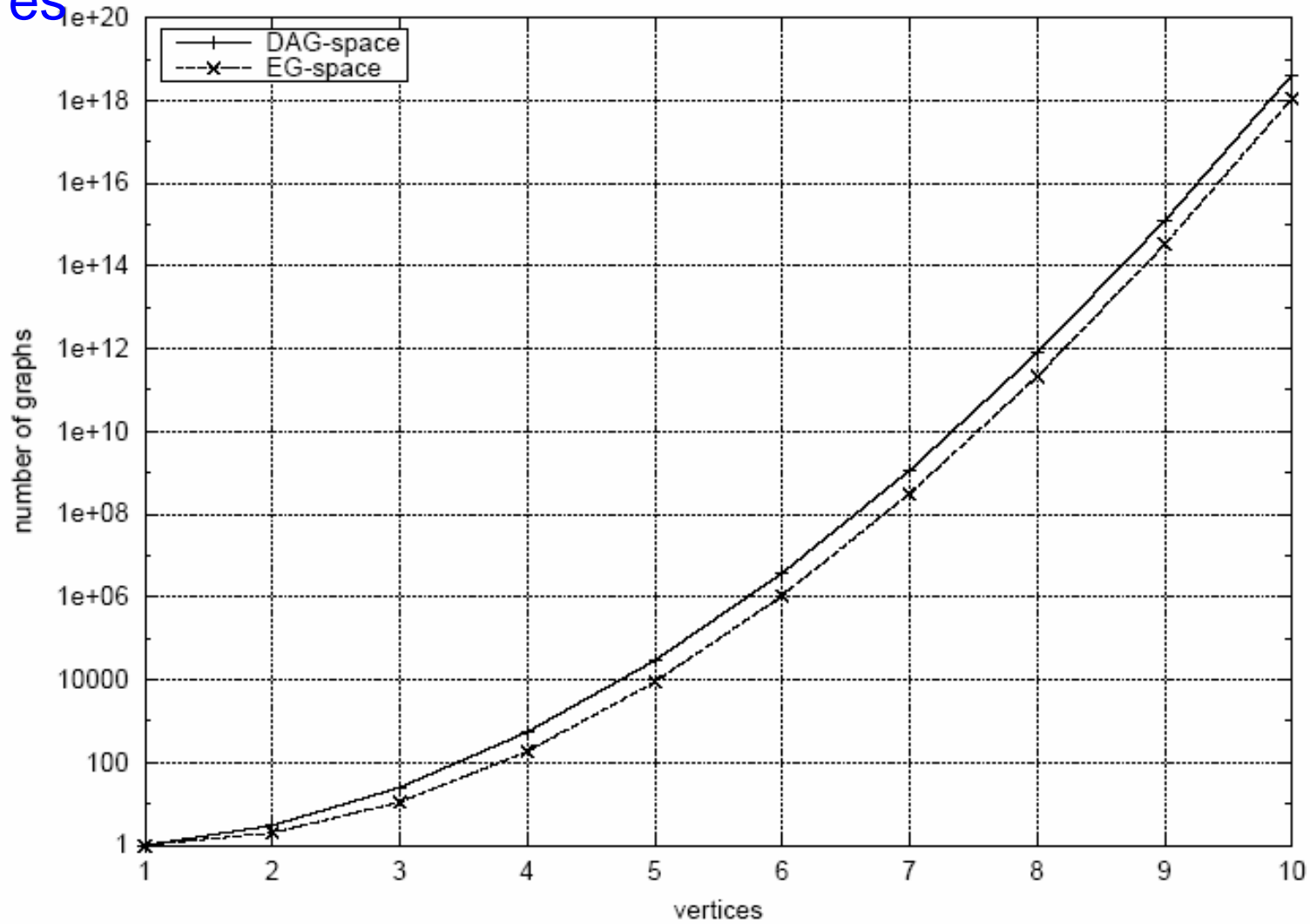
Uncertainty about edges

Can we generalize this scheme  
to more than 2 genes?

In principle yes.

However ...

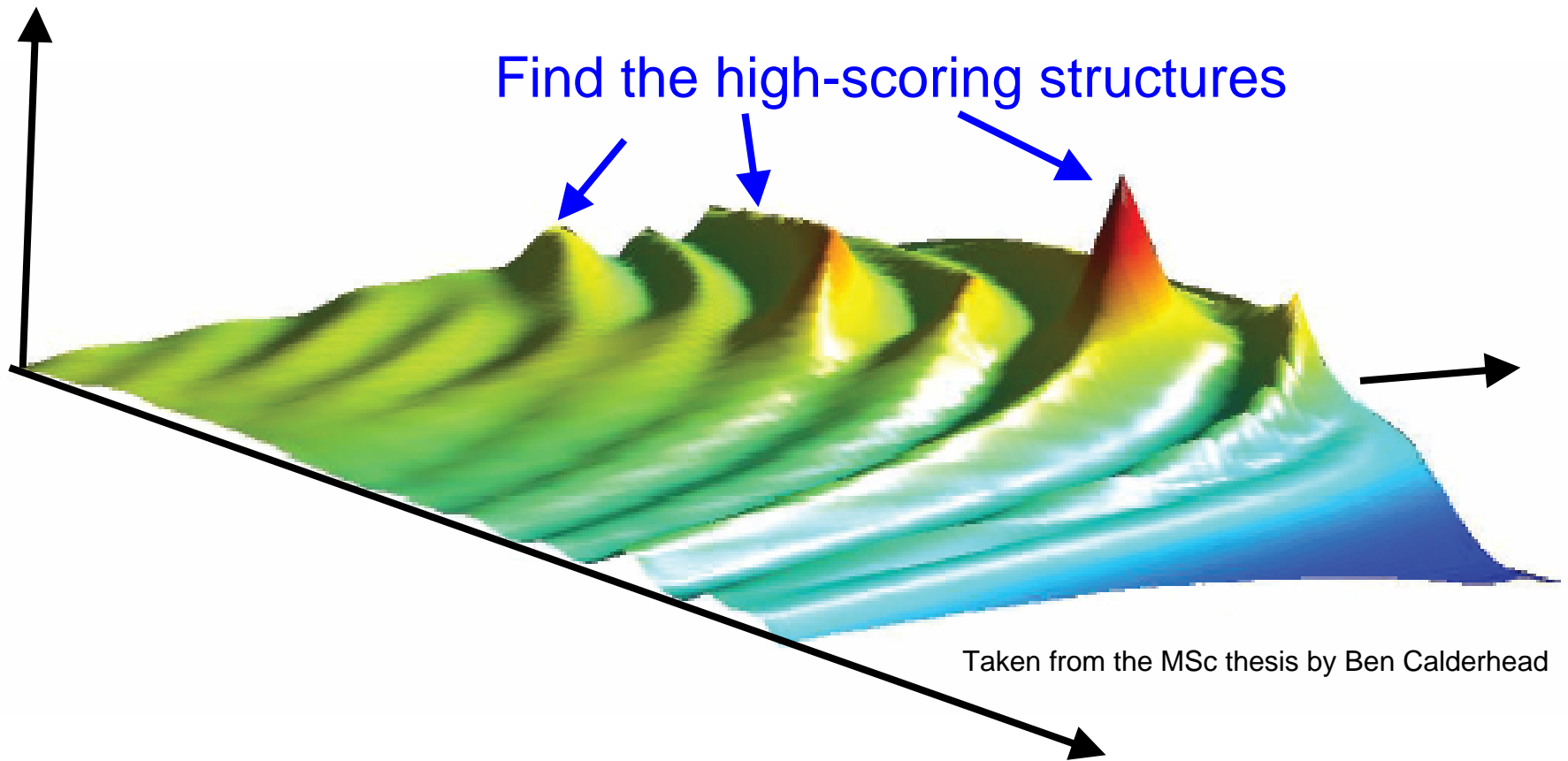
# Number of structures



Number of nodes

# Sampling from the posterior distribution

$$P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{M})P(\mathcal{M})$$



Configuration space of network structures  $\mathcal{M}$

# MCMC

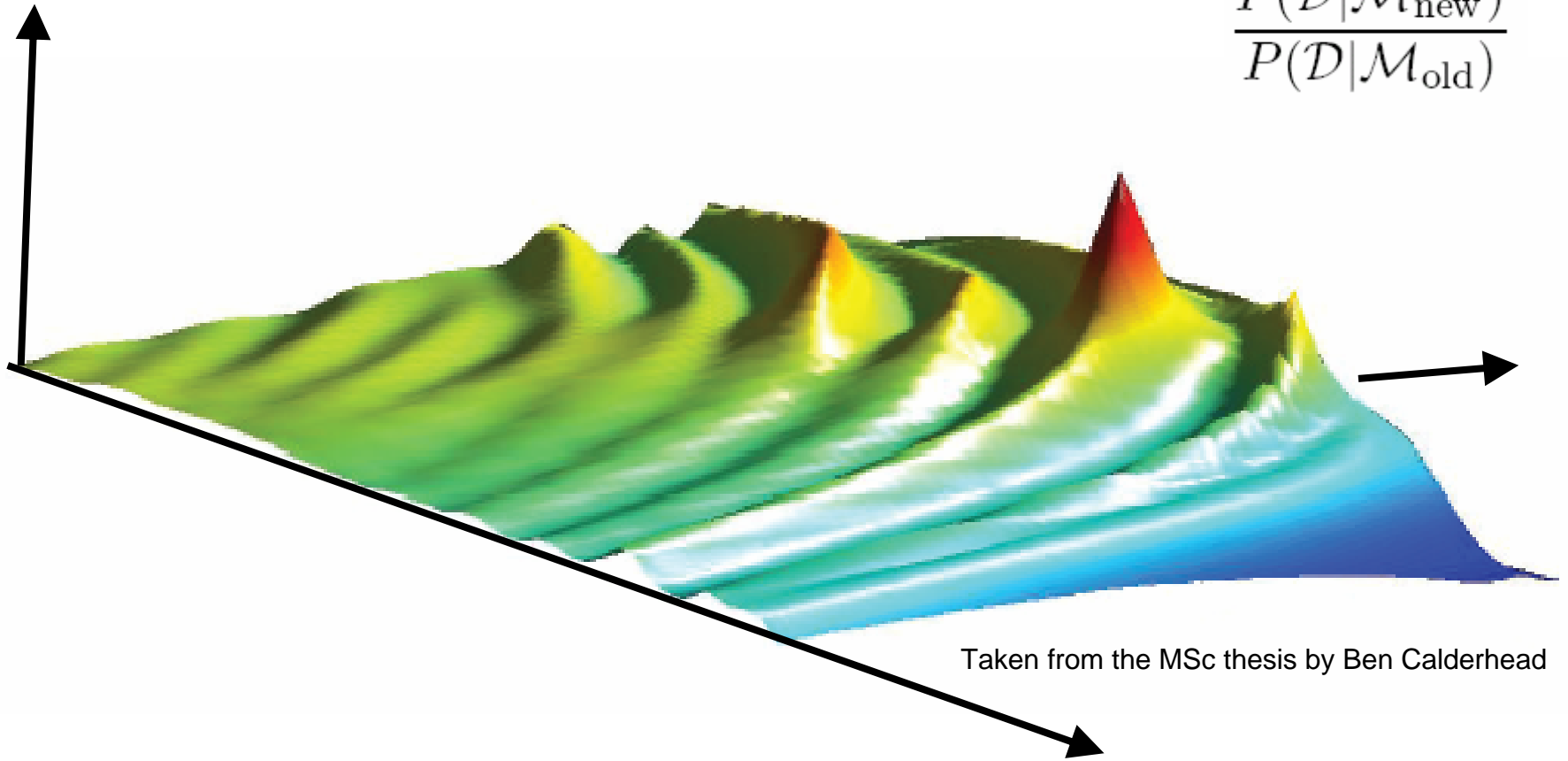
Local change  $\mathcal{M}_{\text{old}} \rightarrow \mathcal{M}_{\text{new}}$

If  $P(\mathcal{D}|\mathcal{M}_{\text{new}}) > P(\mathcal{D}|\mathcal{M}_{\text{old}})$  accept

If  $P(\mathcal{D}|\mathcal{M}_{\text{new}}) < P(\mathcal{D}|\mathcal{M}_{\text{old}})$  accept with probability

$$\frac{P(\mathcal{D}|\mathcal{M}_{\text{new}})}{P(\mathcal{D}|\mathcal{M}_{\text{old}})}$$

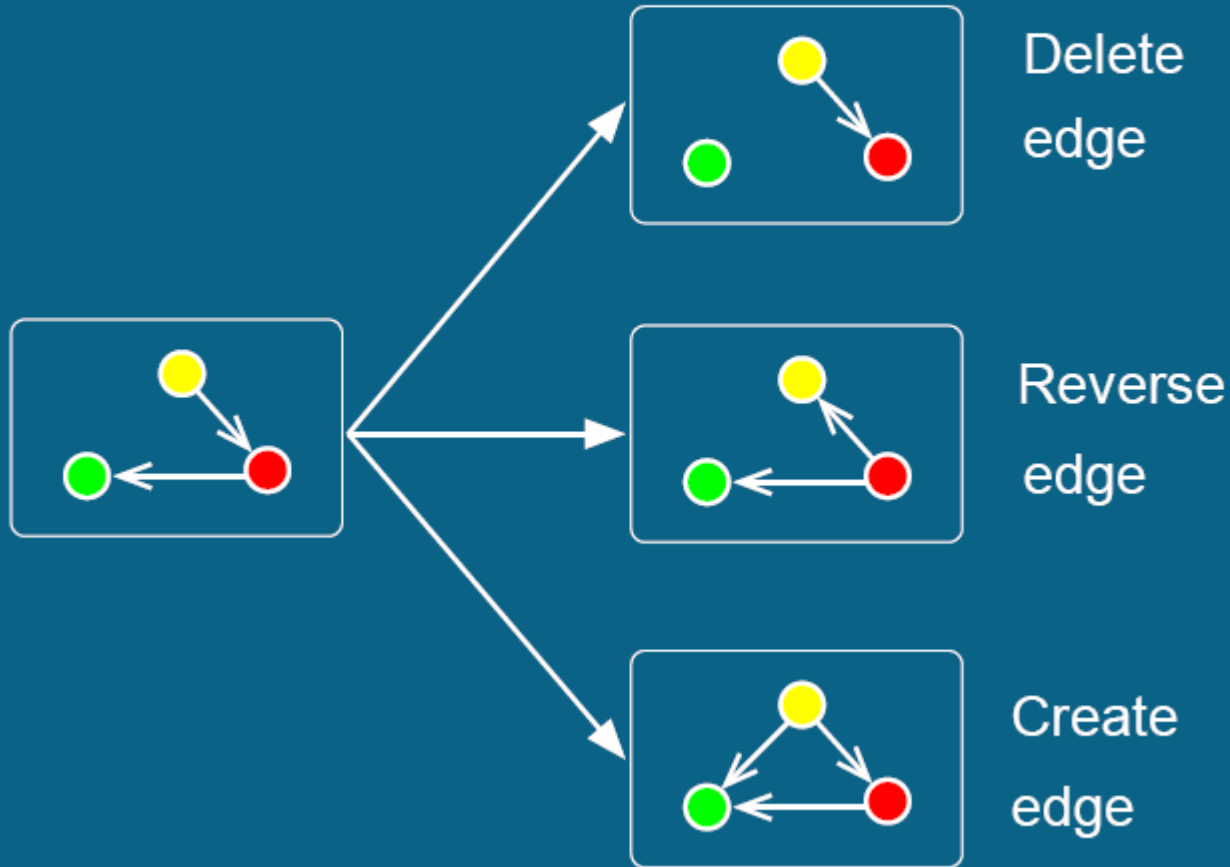
$P(\mathcal{M}|\mathcal{D})$



Taken from the MSc thesis by Ben Calderhead

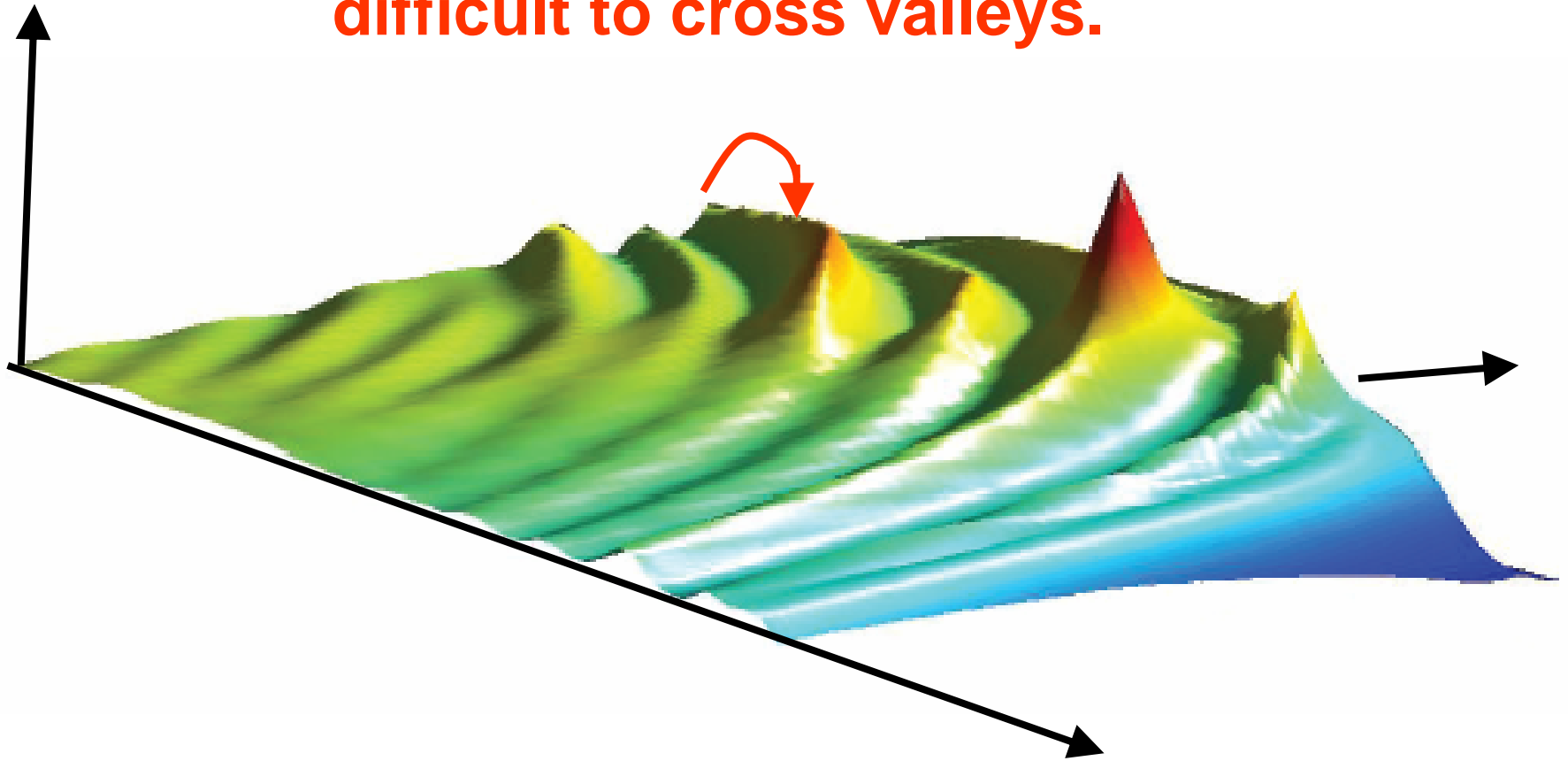
Configuration space of network structures  $\mathcal{M}$

## MCMC moves



**Problem: Local changes  $\rightarrow$  small steps  $\rightarrow$  slow convergence, difficult to cross valleys.**

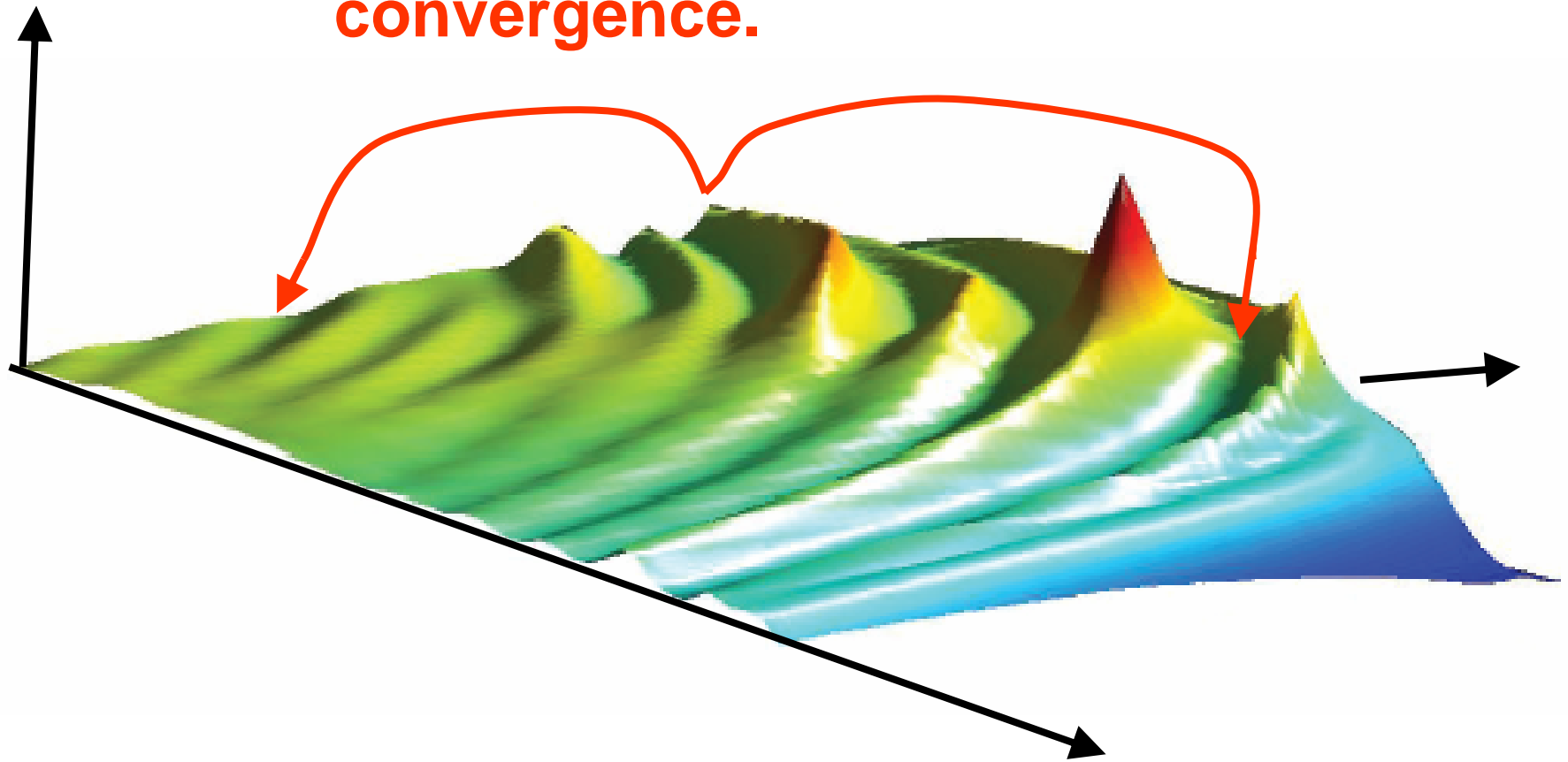
$P(\mathcal{M}|\mathcal{D})$



Configuration space of network structures  $\mathcal{M}$

$P(\mathcal{M}|\mathcal{D})$

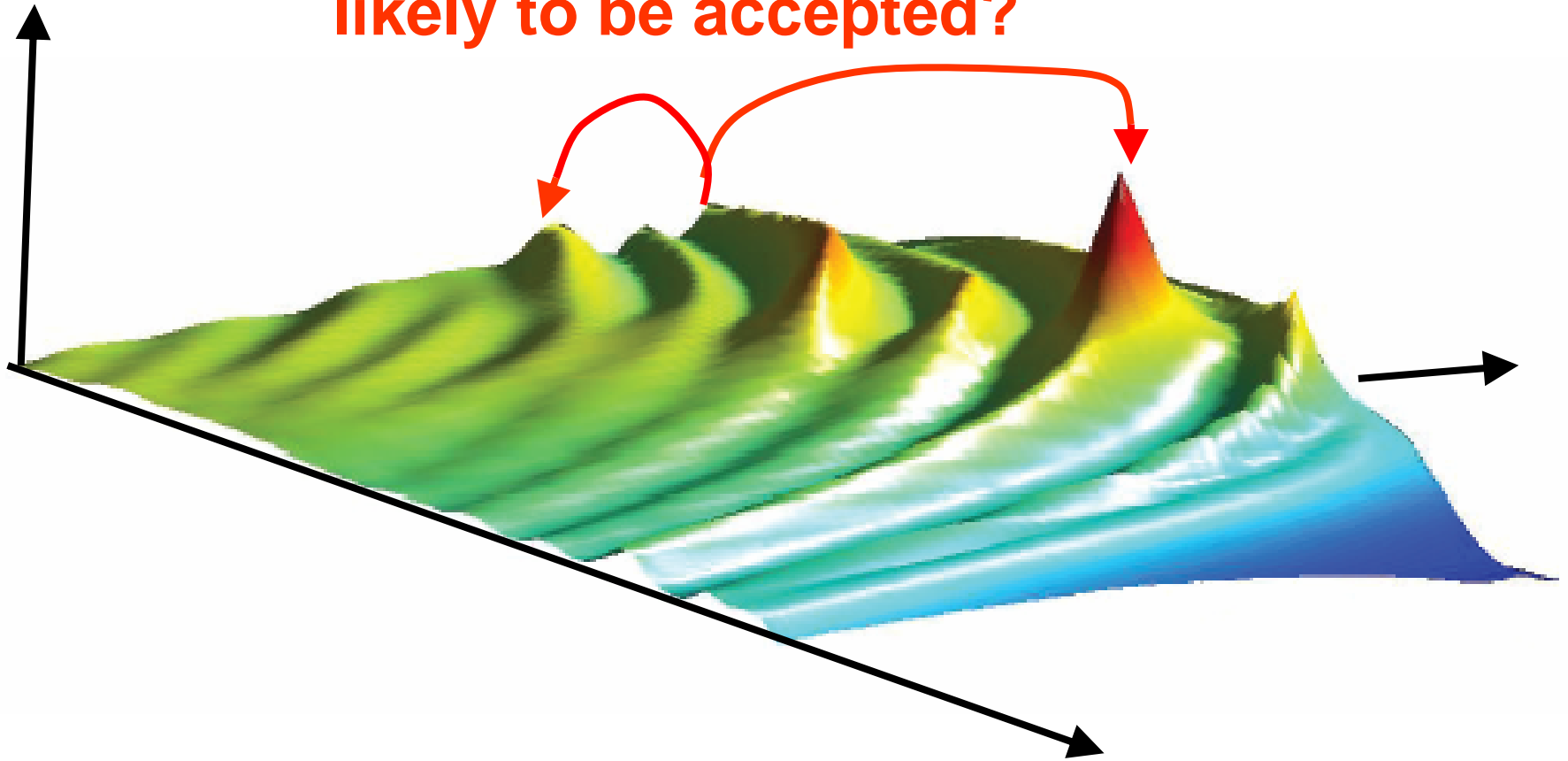
**Problem: Global changes  $\rightarrow$  large steps  $\rightarrow$  low acceptance  $\rightarrow$  slow convergence.**



Configuration space of network structures  $\mathcal{M}$

**Can we make global changes that jump onto other peaks and are likely to be accepted?**

$P(\mathcal{M}|\mathcal{D})$



Configuration space of network structures  $\mathcal{M}$

Mach Learn (2008) 71: 265–305  
DOI [10.1007/s10994-008-5057-7](https://doi.org/10.1007/s10994-008-5057-7)

---

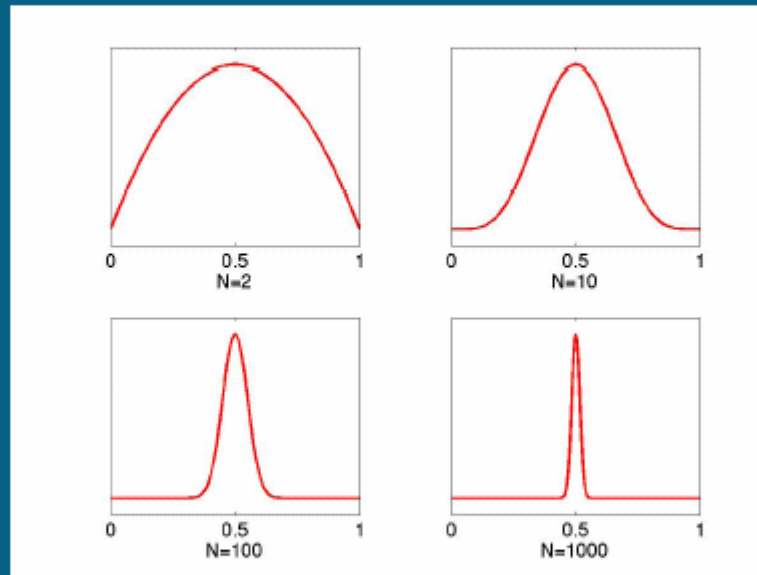
# **Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move**

**Marco Grzegorzcyk · Dirk Husmeier**

Received: 27 March 2007 / Revised: 11 January 2008 / Accepted: 28 March 2008 / Published online: 17 April 2008  
Springer Science+Business Media, LLC 2008

## Problem: Statistical significance of the networks

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



- Posterior probability  $P(M|D)$  **diffused**: We cannot list all the networks that are plausible given the data.

## 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:  $P(f|D) = \sum_M f(M)P(M|D)$

Approximate this sum with MCMC:  $P(f|D) = \frac{1}{T} \sum_{i=1}^T f(M_i)$

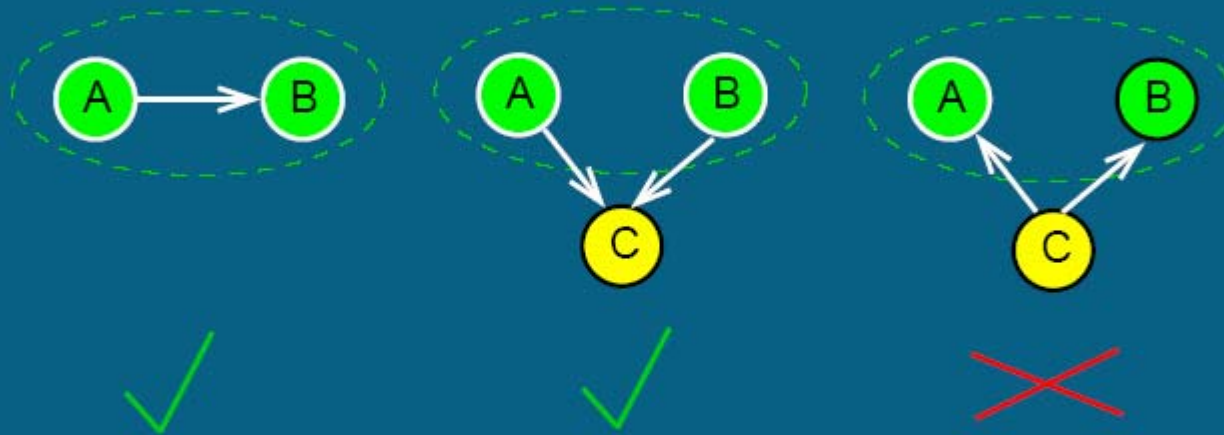
where  $\{M_i\}$  is a sample from the posterior obtained with MCMC.

N. Friedman et al. (2000) and D. Pe'er et al (2001)

- Markov neighbours
- Separators
- Order relations

## Markov neighbours

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



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

# Markov relations

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

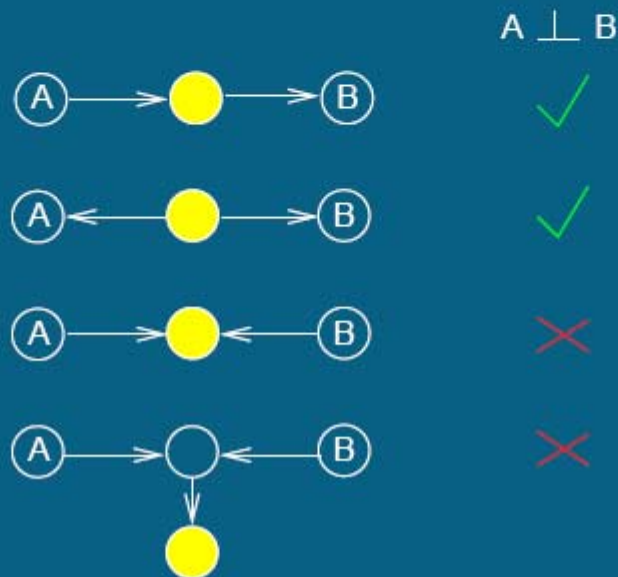
- Most Markov pairs: Intracluster pairings with high correlation in their expression.
- **But:** Genes where  $P(X \leftrightarrow Y|D)$  is high and correlation is low.

<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, holistic

## Separators

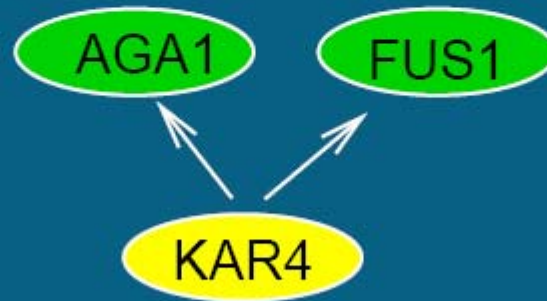
- Given that A and B are indirectly dependent, **what factors *mediate*** this **dependence?**
- Search for a **set of variables Z** such that  $A \perp B | Z$ .
- Z explains all the dependencies between A and B.



# Separator relations

Explain away dependencies.

- **Separators:** Known transcriptional regulators
- **Separated genes:** Share functional roles

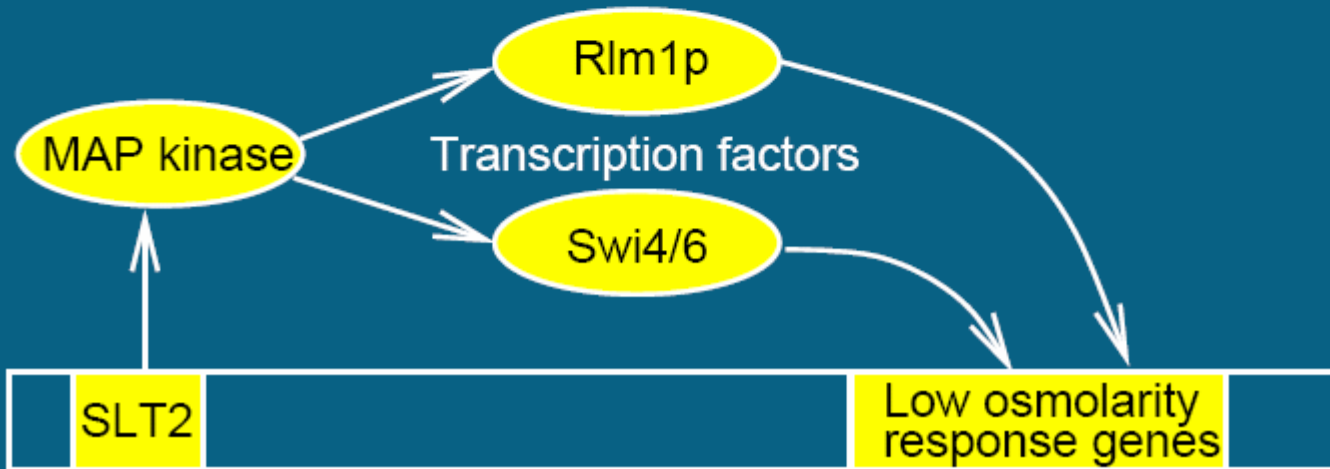
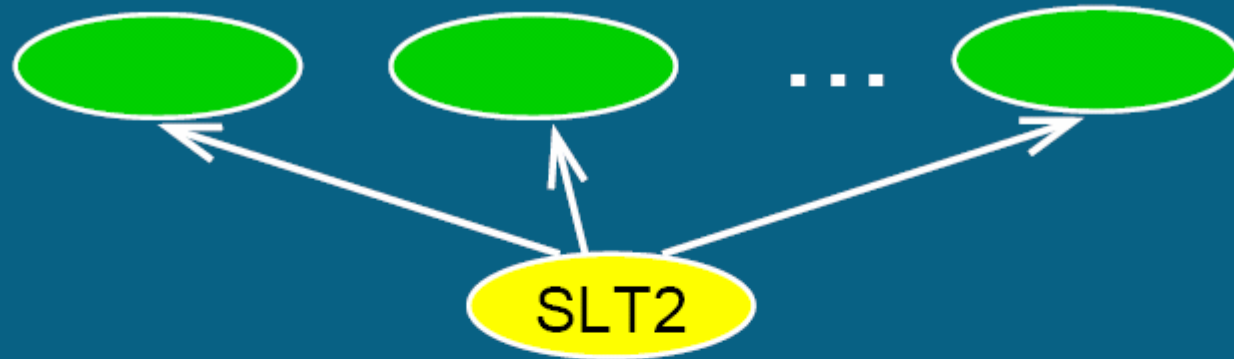


**KAR4**

**AGA1, FUS1**

Transcriptional regulator of nuclear fusion genes.  
Cell fusion genes

# Low osmolarity response genes



## Order relations

- Is  $A$  an **ancestor** of  $B$  in all the networks of a given equivalence class?
- Does the **PDAG** 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  $Y$ :

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

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

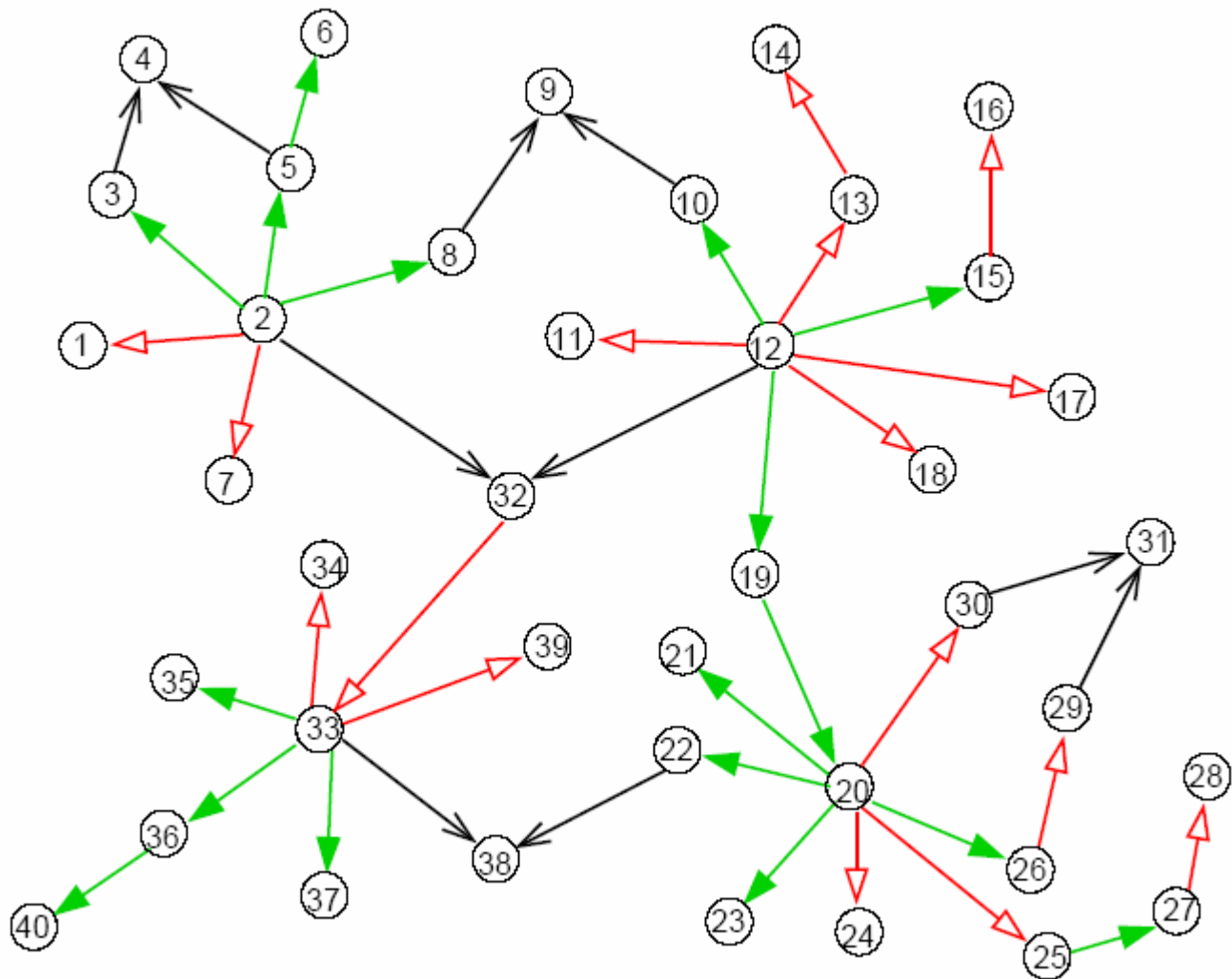
Genes with high dominance scores are **indicative** of potential **causal** sources of the cell cycle process.

## Dominant genes in the ordering relations

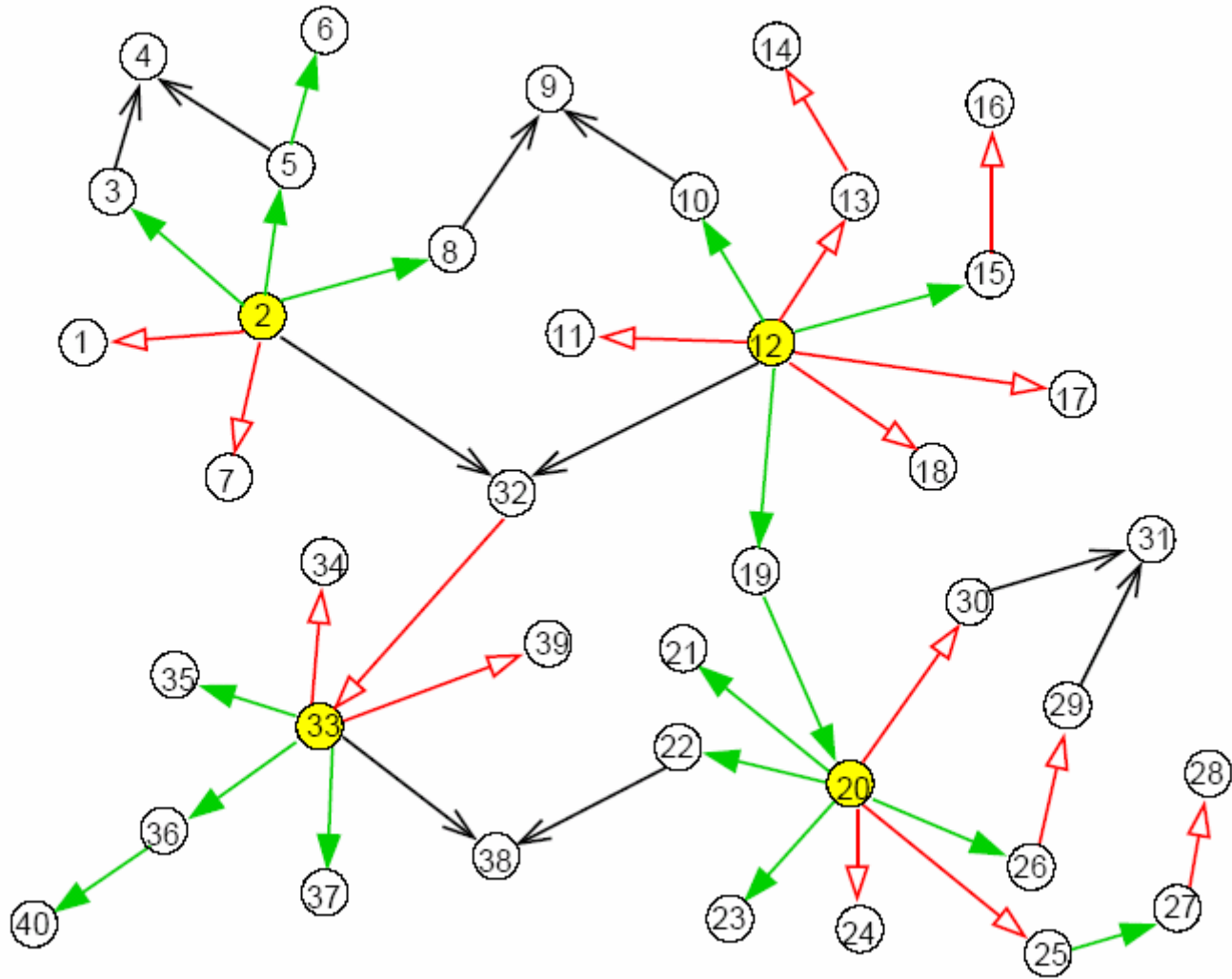
CLN1	Role in cell cycle <b>start</b>
CLN2	Role in cell cycle <b>start</b>
CDC5	Cell cycle <b>control</b> , required for exit from mitosis
RAD53	Cell cycle <b>control</b> : checkpoint function
RFA2	Involved in nucleotide excision <b>repair</b>
PLO30	Required for DNA replication and <b>repair</b>
MSH6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

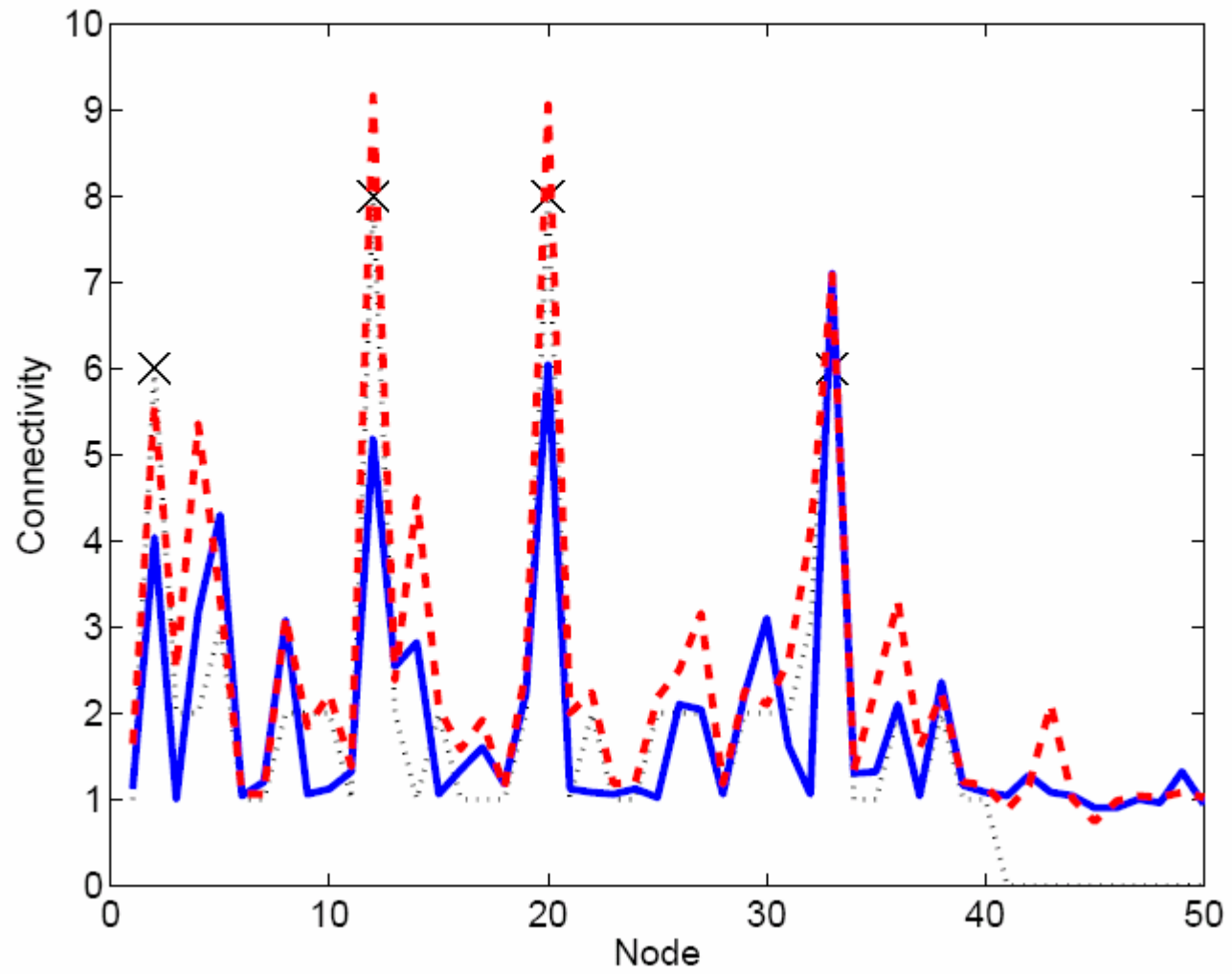
Model network, data set size:  $N = 50$



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

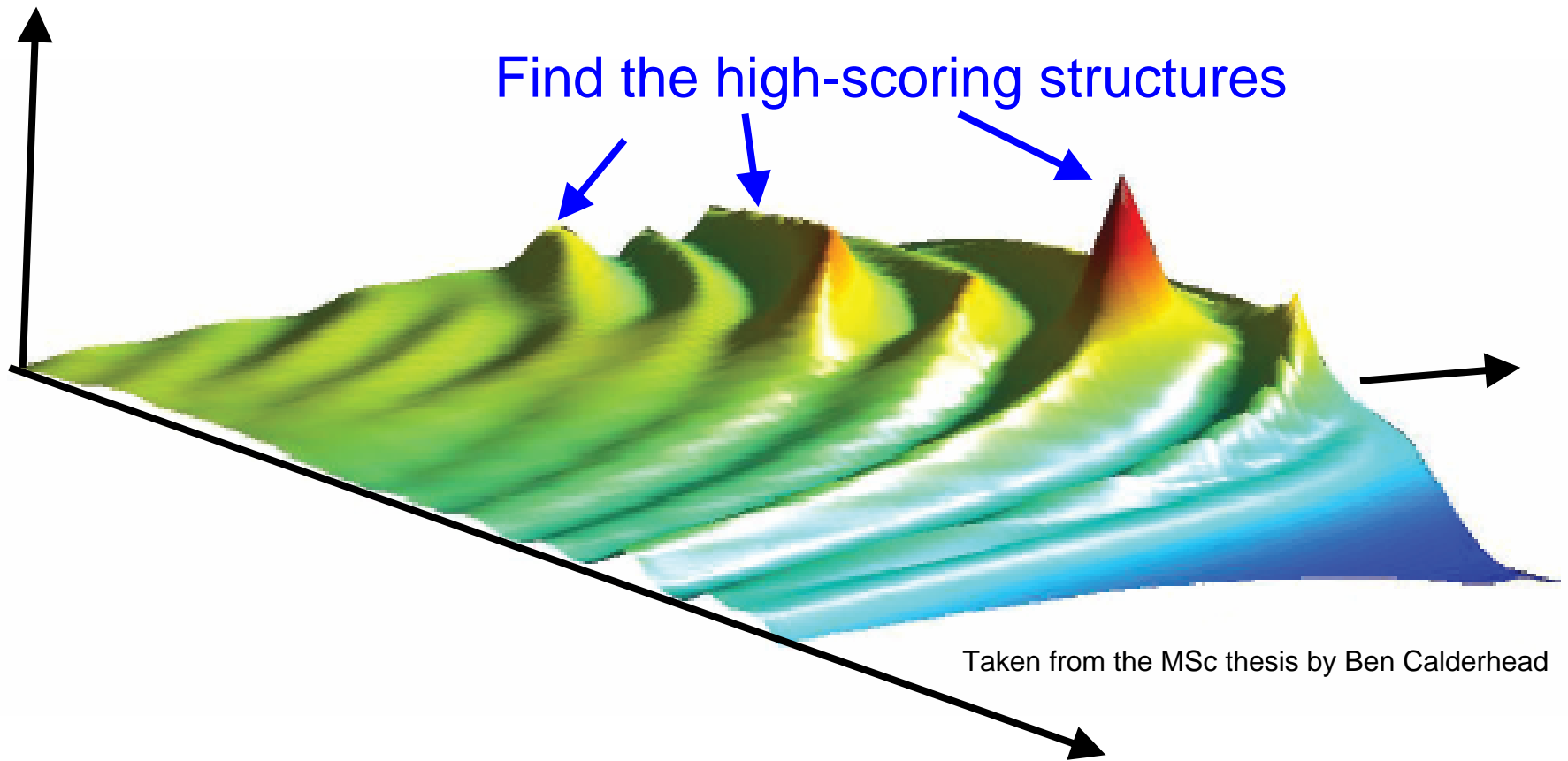


# Predicted connectivity spectrum



# Sampling from the posterior distribution

$$P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{M})P(\mathcal{M})$$



Configuration space of network structures  $\mathcal{M}$