

---

# Probabilistic methods for post-genomic data integration

Dirk Husmeier

Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ  
United Kingdom

<http://www.bioss.ac.uk/~dirk>

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

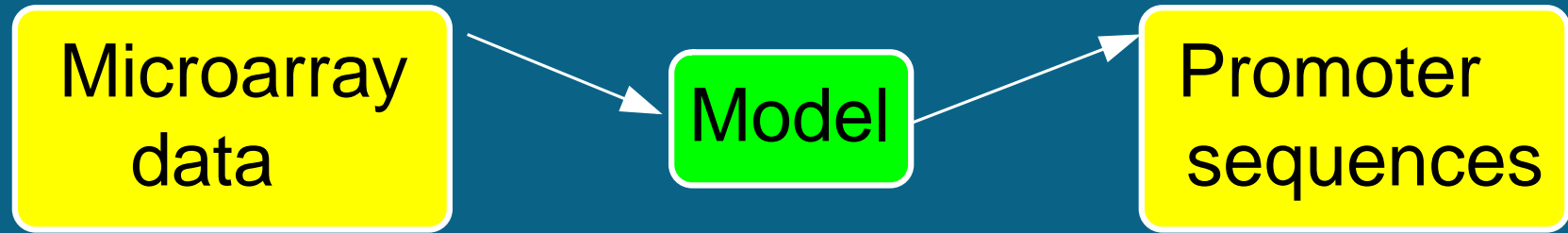
Promoter  
sequences

Microarray  
data

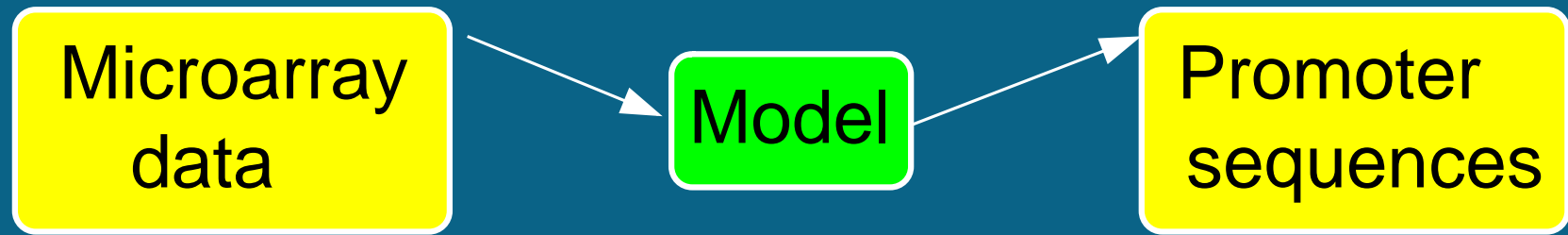


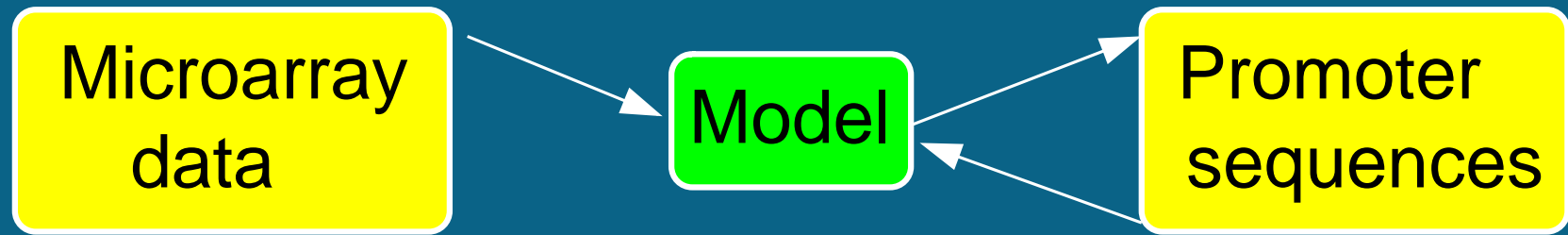
Model

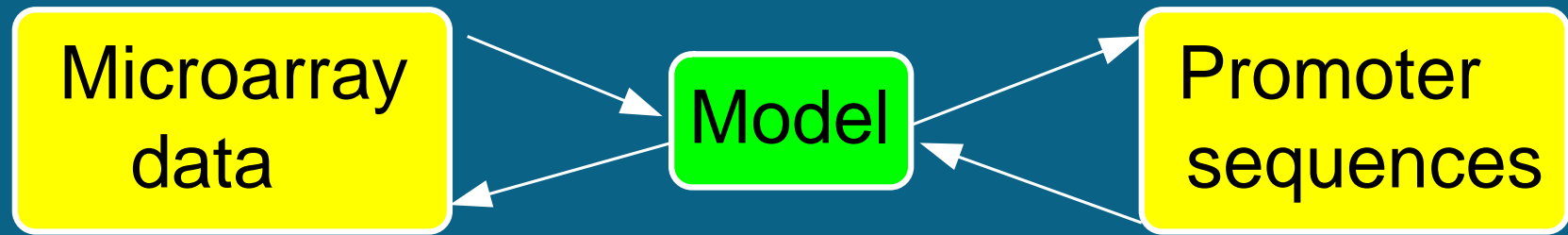
Promoter  
sequences



# Segal's unifying probabilistic model







Segal, Yelensky, Koller (2003)

Bioinformatics 19

Segal, Yelensky, Koller (2003)

Bioinformatics 19

Revision:

Motif finding

Motif: T A T A

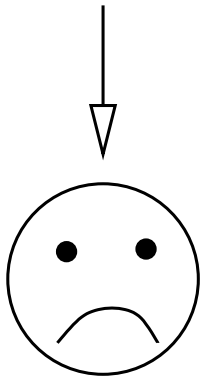
T C G A A T T C T A T A G C C A C

Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C

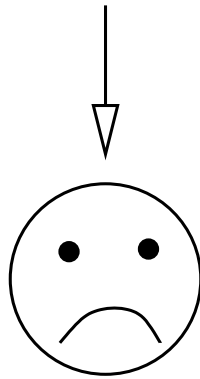
Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C



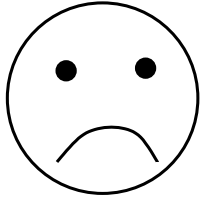
Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C



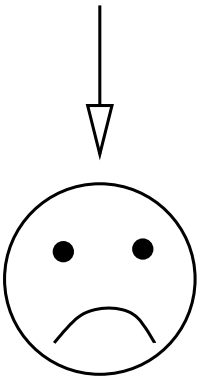
Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C



Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C



Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C



## Position Specific Scoring Matrix (PSSM)

Search for a **motif** of length  $W$  in **binding sequences**.

## Position Specific Scoring Matrix (PSSM)

Search for a **motif** of length  $W$  in **binding sequences**.

$W \times 4$  matrix  $\psi_k(l)$ :

Probability that the nucleotide in the  $k$ th position,  
 $k \in [1, \dots, W]$ , is an  $l \in \{A, C, G, T\}$ .

## Position Specific Scoring Matrix (PSSM)

Search for a **motif** of length  $W$  in **binding sequences**.

$W \times 4$  matrix  $\psi_k(l)$ :

Probability that the nucleotide in the  $k$ th position,  
 $k \in [1, \dots, W]$ , is an  $l \in \{A, C, G, T\}$ .

## Background model

for **non-binding sequences**

4-dim vector  $\theta_0(l)$ :

Probability of nucleotide  $l$ ; this distribution is **position-independent**.

Sequence  $S_1, S_2, \dots, S_N$

Sequence  $S_1, S_2, \dots, S_N$

Non-binding sequence:  $R=0$

$$P(S_1, S_2, \dots, S_N | R = 0) = \prod_{t=1}^N \theta_0(S_t)$$

Sequence  $S_1, S_2, \dots, S_N$

Non-binding sequence:  $R=0$

$$P(S_1, S_2, \dots, S_N | R = 0) = \prod_{t=1}^N \theta_0(S_t)$$

Binding sequence:  $R=1$ , motif starting at position  $m+1$

$$\begin{aligned} &P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) \\ &= \prod_{t=1}^m \theta_0(S_t) \prod_{k=1}^W \psi_k(S_{m+k}) \prod_{t=m+W+1}^N \theta_0(S_t) \\ &= \prod_{t=1}^N \theta_0(S_t) \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Binding sequence:  $R=1$ , motif starting at position  $m+1$

$$P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) = \prod_{t=1}^N \theta_0(S_t) \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence:  $R=1$ , motif starting at position  $m+1$

$$P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) = \prod_{t=1}^N \theta_0(S_t) \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence:  $R=1$ , motif starting anywhere

$$\begin{aligned} &P(S_1, S_2, \dots, S_N | R = 1) \\ &= \sum_{m=0}^{N-W} P(\text{start} = m + 1) P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) \\ &= \prod_{t=1}^N \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Binding sequence:  $R=1$ , motif starting at position  $m+1$

$$P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) = \prod_{t=1}^N \theta_0(S_t) \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})}$$

Binding sequence:  $R=1$ , motif starting anywhere

$$\begin{aligned} P(S_1, S_2, \dots, S_N | R = 1) &= \sum_{m=0}^{N-W} P(\text{start} = m + 1) P(S_1, S_2, \dots, S_N | R = 1, \text{start} = m + 1) \\ &= \prod_{t=1}^N \theta_0(S_t) \frac{1}{N - W + 1} \sum_{m=0}^{N-W} \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \end{aligned}$$

Objective: Prediction of binding activity from sequence:  
 $P(R = 1 | S_1, S_2, \dots, S_N)$

Apply Bayes rule:

$$\begin{aligned} & P(R = 1 | S_1, S_2, \dots, S_N) \\ &= \frac{P(S_1, S_2, \dots, S_N | R = 1)P(R = 1)}{P(S_1, S_2, \dots, S_N | R = 0)P(R = 0) + P(S_1, S_2, \dots, S_N | R = 1)P(R = 1)} \\ &= \left( 1 + \frac{P(R = 0)P(S_1, S_2, \dots, S_N | R = 0)}{P(R = 1)P(S_1, S_2, \dots, S_N | R = 1)} \right)^{-1} \\ &= \left( 1 + \left[ \frac{P(R = 1)}{P(R = 0)} \frac{1}{(N - W + 1)} \sum_{m=0}^{N-W} \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \right]^{-1} \right)^{-1} \end{aligned}$$

Apply Bayes rule:

$$\begin{aligned} & P(R = 1 | S_1, S_2, \dots, S_N) \\ &= \frac{P(S_1, S_2, \dots, S_N | R = 1)P(R = 1)}{P(S_1, S_2, \dots, S_N | R = 0)P(R = 0) + P(S_1, S_2, \dots, S_N | R = 1)P(R = 1)} \\ &= \left( 1 + \frac{P(R = 0)P(S_1, S_2, \dots, S_N | R = 0)}{P(R = 1)P(S_1, S_2, \dots, S_N | R = 1)} \right)^{-1} \\ &= \left( 1 + \left[ \frac{P(R = 1)}{P(R = 0)} \frac{1}{(N - W + 1)} \sum_{m=0}^{N-W} \prod_{k=1}^W \frac{\psi_k(S_{m+k})}{\theta_0(S_{m+k})} \right]^{-1} \right)^{-1} \end{aligned}$$

Define:

$$w_k(l) = \log \frac{\psi_k(l)}{\theta_0(l)}, \quad w_0 = \log \frac{P(R=1)}{P(R=0)}, \quad \text{logit}(z) = \frac{1}{1 + \exp(-z)}$$

$$P(R = 1 | S_1, S_2, \dots, S_N)$$
$$= \text{logit} \left( \log \left[ \frac{w_0}{N - W + 1} \sum_{m=0}^{N-W} \exp \left( \sum_{k=1}^W w_k(S_{t+k}) \right) \right] \right)$$

$4 \times W + 1$  parameters:  $w_k(l)$ ,  $w_0$

Motif: T A<sup>C</sup> T A<sup>C</sup> G

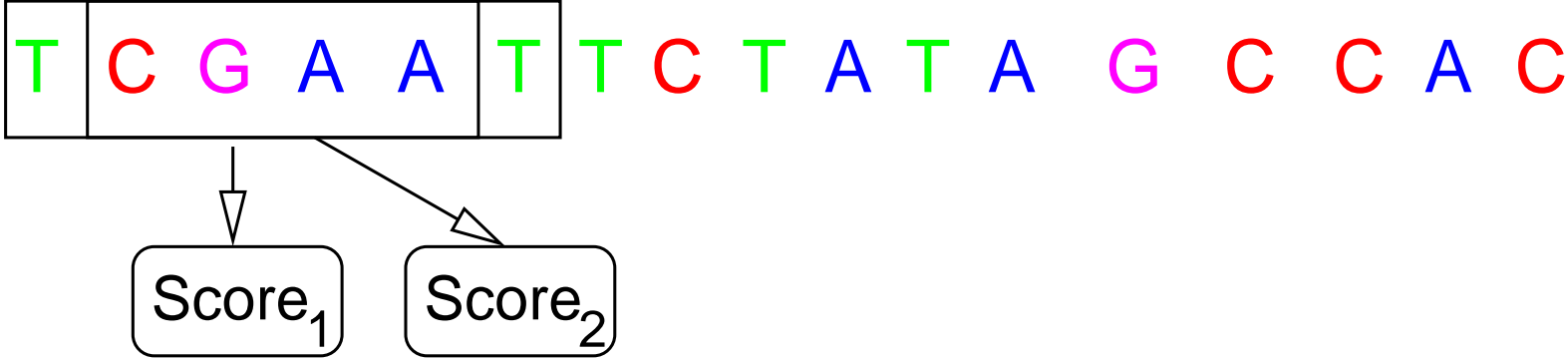
T C G A A T T C T A T A G C C A C

Motif: T A<sup>C</sup> T A<sup>C</sup> G

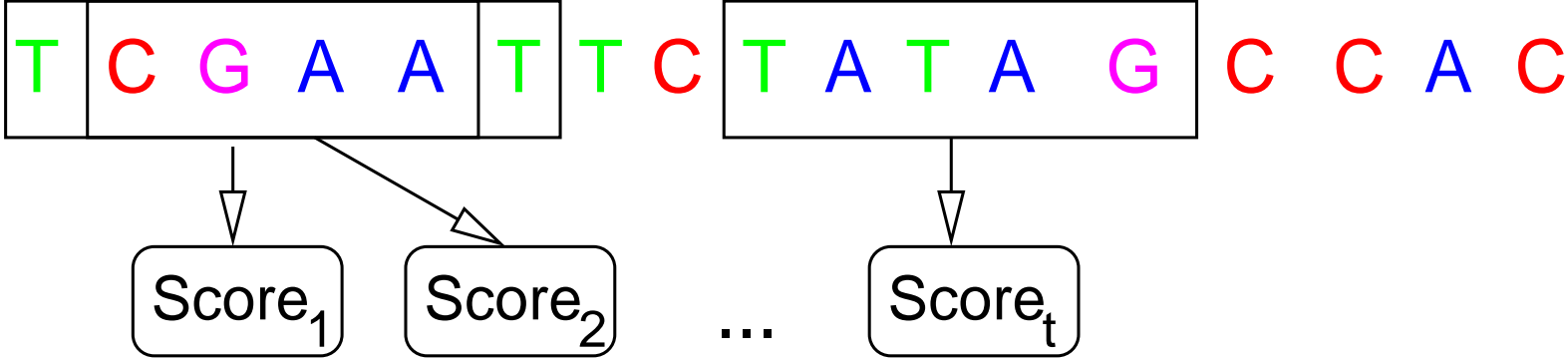
T C G A A T T C T A T A G C C A C

↓  
Score<sub>1</sub>

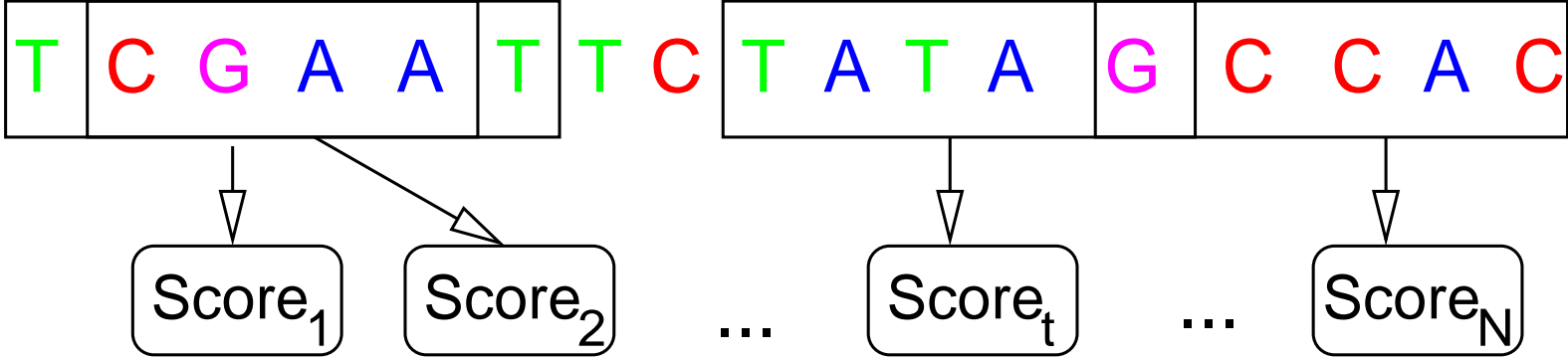
Motif: T A<sup>C</sup> T A<sup>C</sup> G



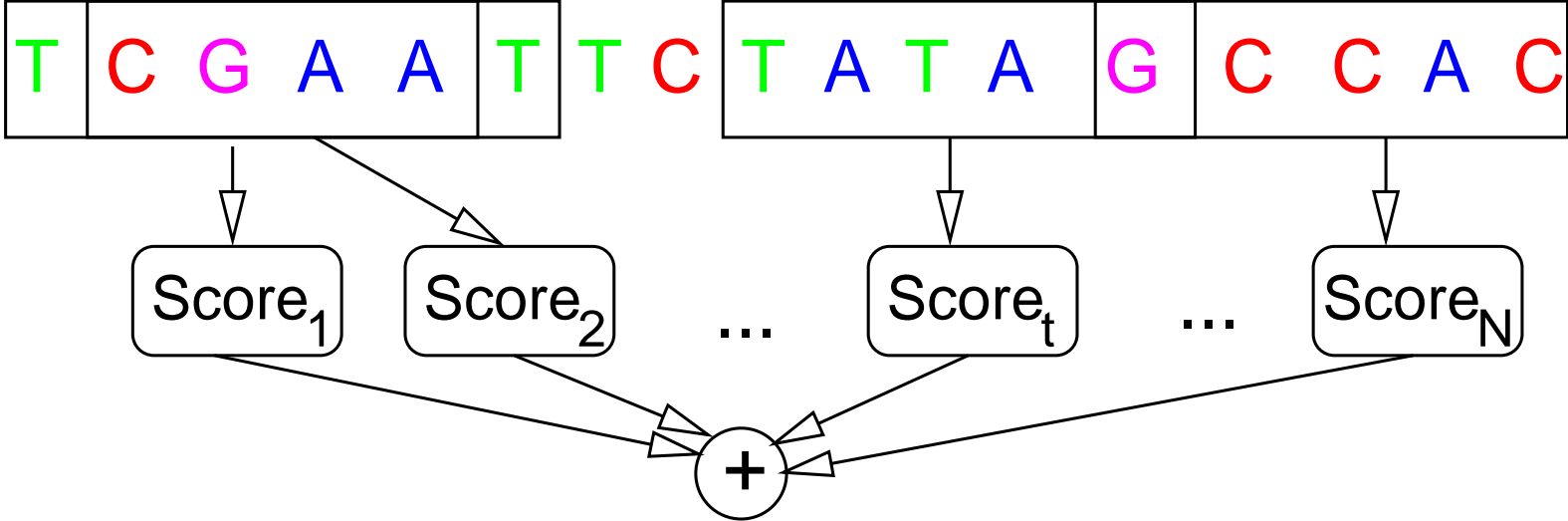
Motif: T A<sup>C</sup> T A<sup>C</sup> G



Motif: T A<sup>C</sup> T A<sup>C</sup> G



Motif: T A<sup>C</sup> T A<sup>C</sup> G



Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C

Score<sub>1</sub>

Score<sub>2</sub>

...

Score<sub>t</sub>

...

Score<sub>N</sub>

+

Nonlinear transfer function

Motif: T A<sup>C</sup> T A<sup>C</sup> G

T C G A A T T C T A T A G C C A C

Score<sub>1</sub>

Score<sub>2</sub>

...

Score<sub>t</sub>

...

Score<sub>N</sub>

+

Nonlinear transfer function

P(R=1|sequence)

$$P(R = 1 | S_1, S_2, \dots, S_N)$$
$$= \text{logit} \left( \log \left[ \frac{w_0}{N - W + 1} \sum_{m=0}^{N-W} \exp \left( \sum_{k=1}^W w_k(S_{t+k}) \right) \right] \right)$$

$4 \times W + 1$  parameters:  $w_k(l)$ ,  $w_0$

Wolfgang Lehrach

Biomathematics & Statistics Scotland

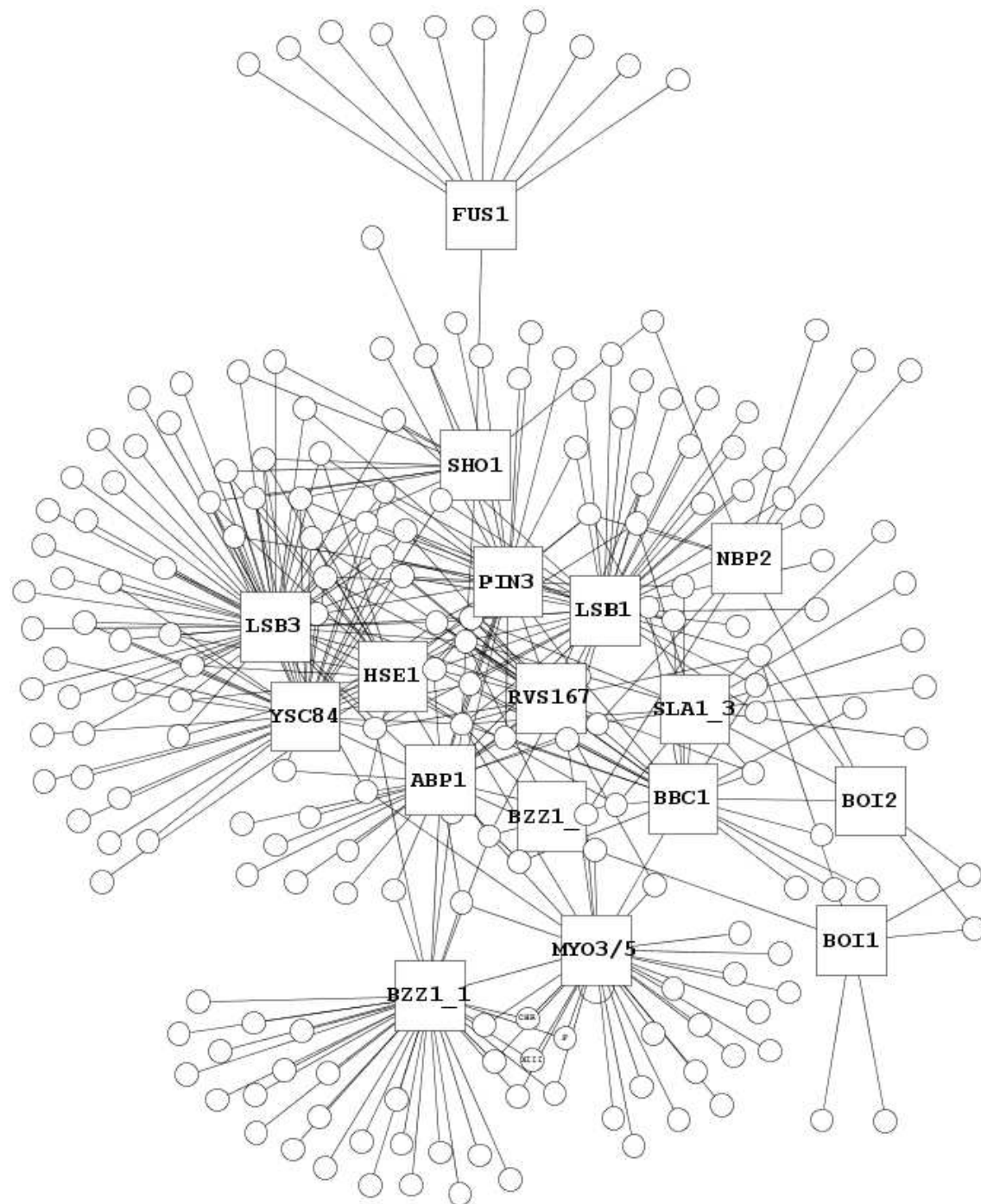
Ab initio prediction of protein interaction

## SH3 yeast two-hybrid interaction network

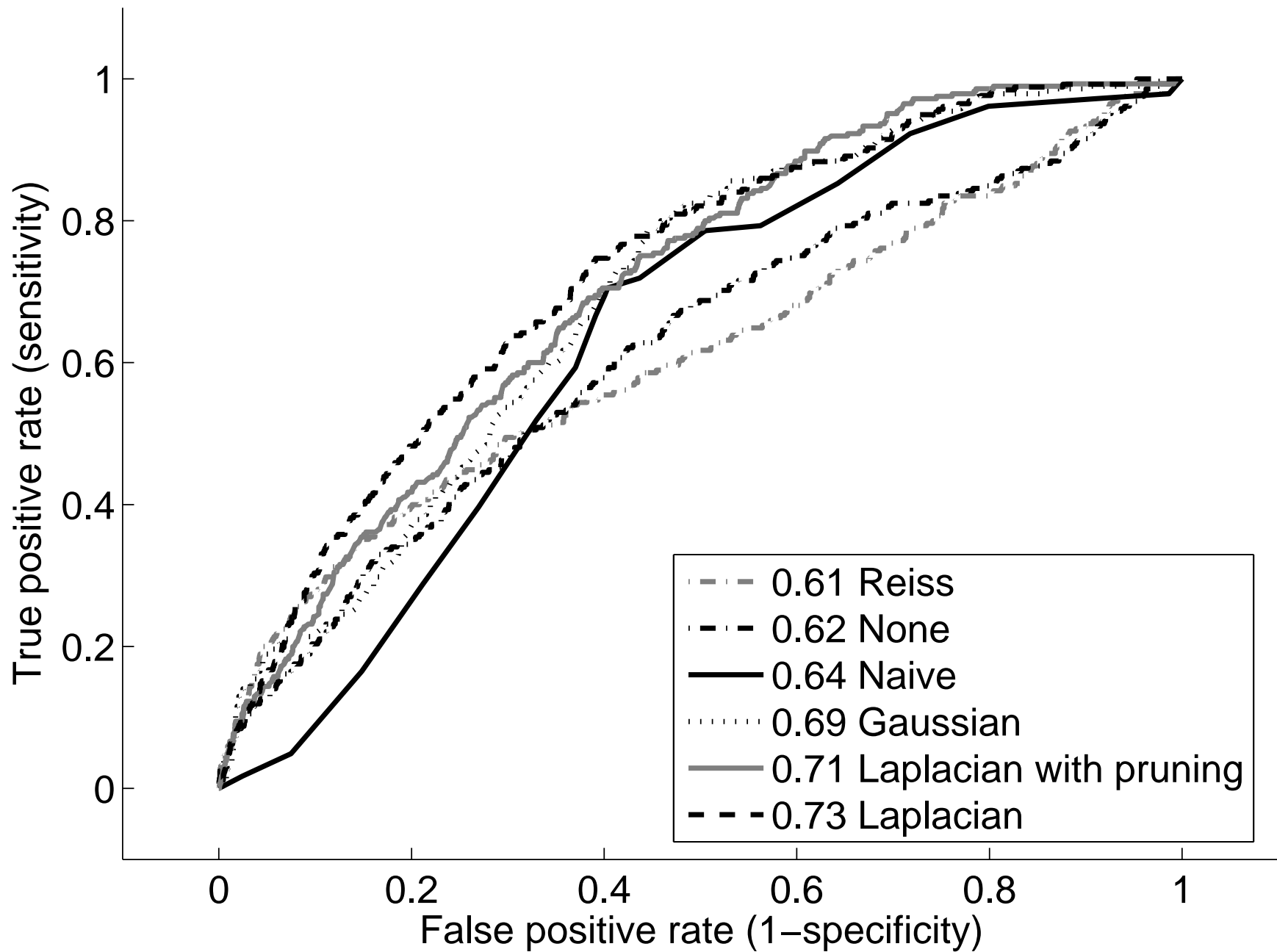
Tong et al. (2002), Science 295, 321-324

285 interactions between 28 SH3 proteins  
and 143 binding peptides

9 binding partners per SH3 domain on average

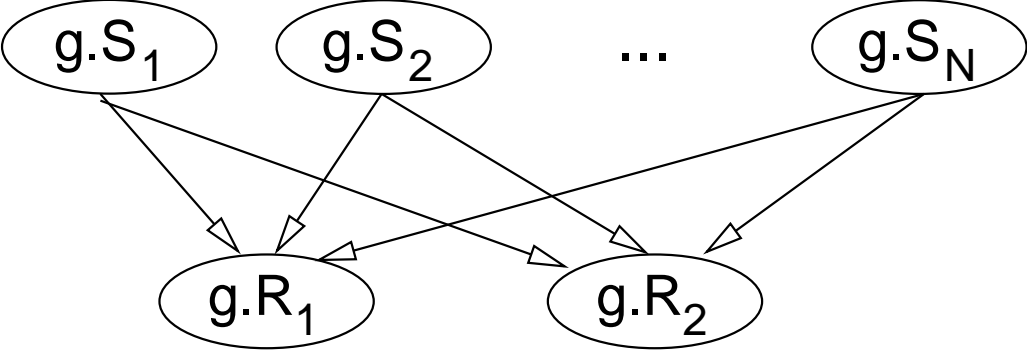


Final Test Set Performance

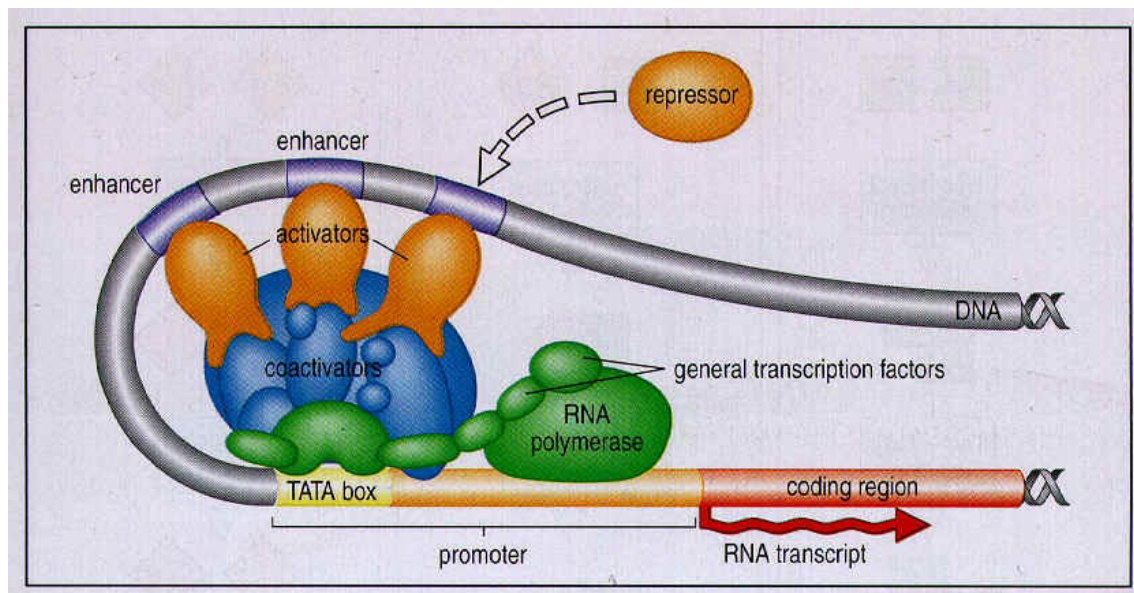


The model  
of  
Segal, Yelensky and Koller  
Bioinformatics 19, 2003

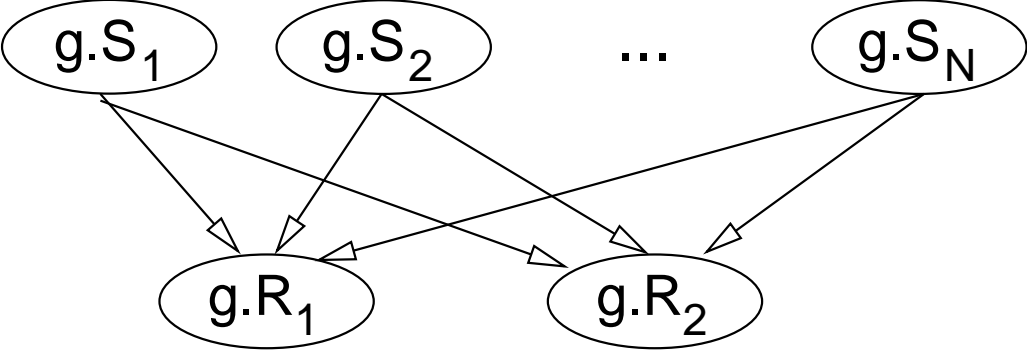
$$P(g.R_2 | g.S) \quad T A T \quad A C$$



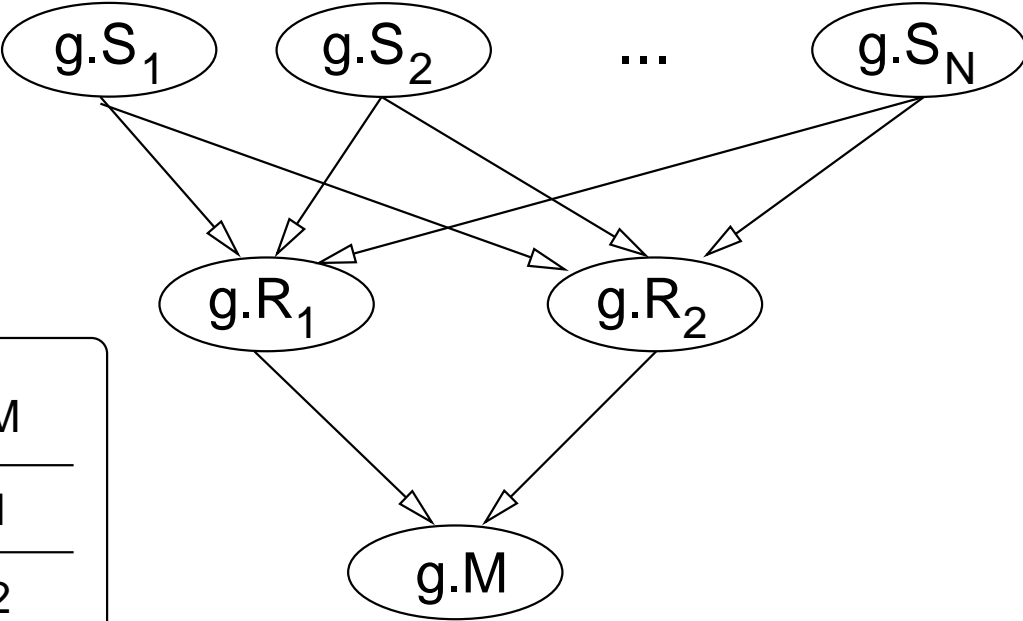
# Transcriptional Regulation



$$P(g.R_2 | g.S) \quad T A T^c A^c$$

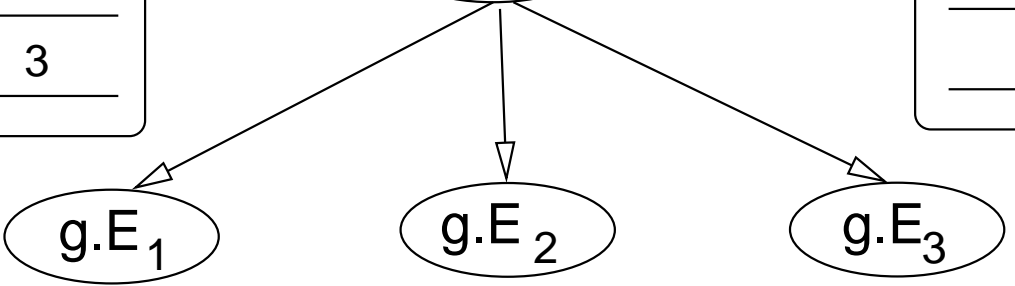
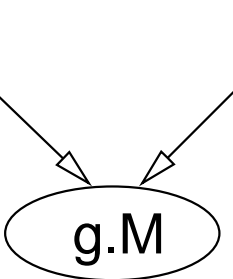
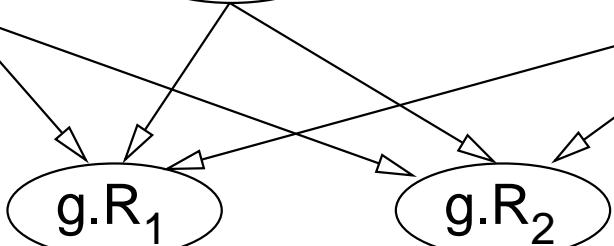
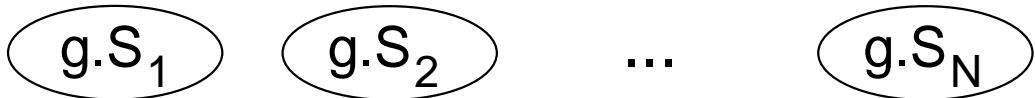


$$P(g.R_2 | g.S) \quad T A T^c A^c$$



g.R <sub>1</sub>	g.R <sub>2</sub>	g.M
█		1
	█	2
█	█	3

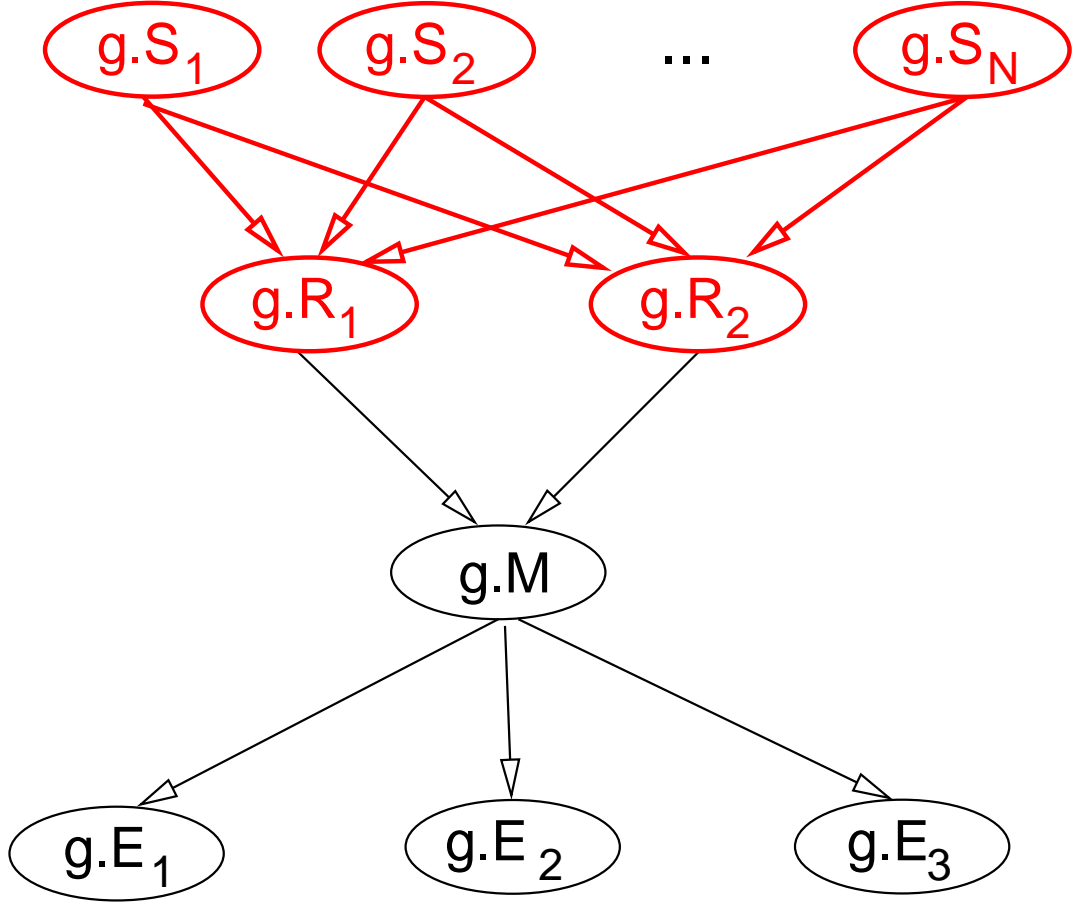
$$P(g.R_2 | g.S) \quad T A T^c \quad A^c G$$



$g.R_1$	$g.R_2$	$g.M$
		1
		2
		3

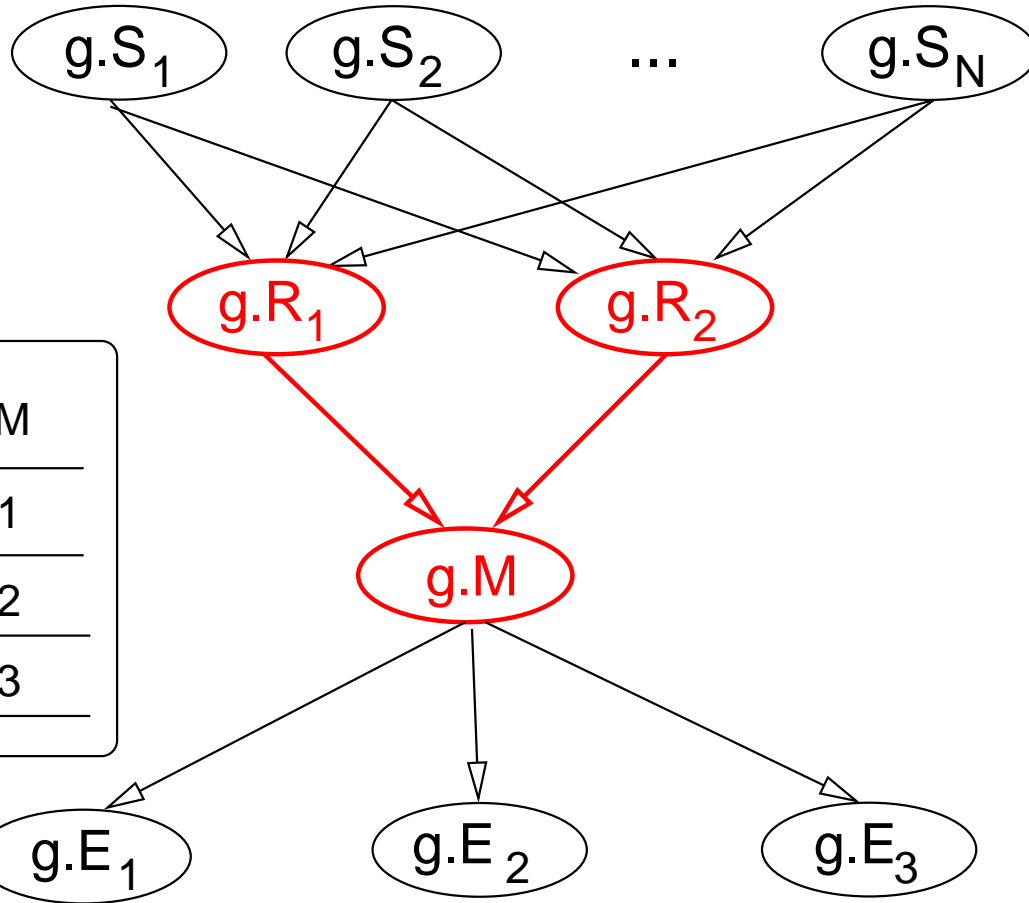
$P(g.E_3   g.M)$	$g.M$
	1
	2
	3





$$P(g.R_2 | g.S) \quad T A T^c \quad A^c G$$



$$\begin{aligned}
& P(g.R_i = 1 | g.S_1, g.S_2, \dots, g.S_N) \\
&= \text{logit} \left( \log \left[ \frac{w_0}{N - W + 1} \sum_{m=0}^{N-W} \exp \left( \sum_{k=1}^W w_k(g.S_{t+k}) \right) \right] \right)
\end{aligned}$$

$4 \times W + 1$  parameters per binding motif  $R_i$ :  $w_k^i(l)$ ,  $w_0^i$



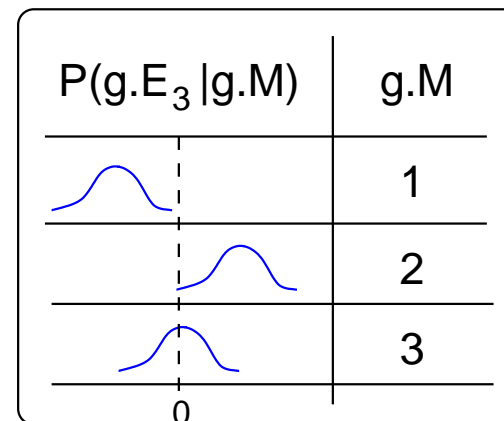
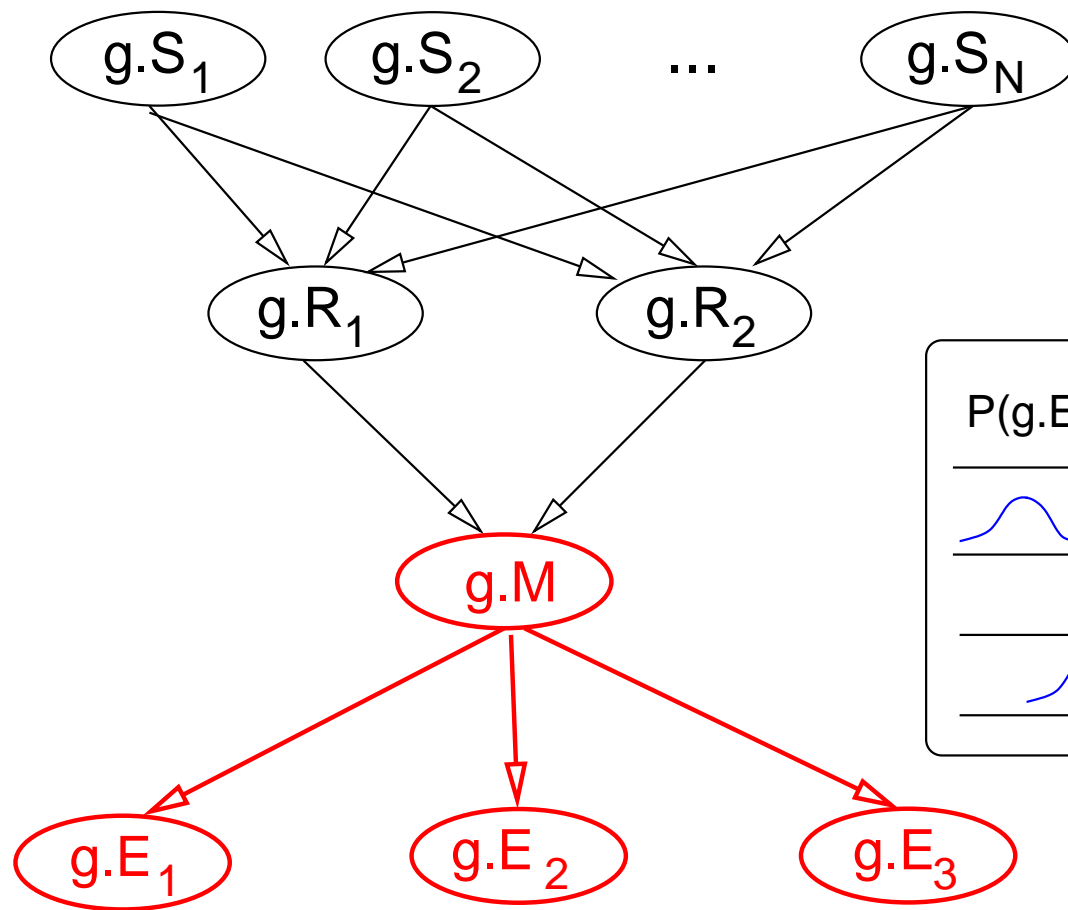
$g.R_1$	$g.R_2$	$g.M$
		1
		2
		3

## Softmax function

$$P(g.M = m | g.R_1 = r_1, g.R_2 = r_2, \dots, g.R_N = r_N) \\ = \frac{\exp\left(\sum_{i=1}^L u_{mi} r_i\right)}{\sum_{\tilde{m}} \exp\left(\sum_{i=1}^L u_{\tilde{m}i} r_i\right)}$$

Parameter matrix:

Number of motifs/regulators  $\times$  number of modules



## Independent Gaussian distributions

$$P(g.E_1, g.E_2, \dots, g.E_L | g.M = m) = \prod_j P(g.E_j | g.M = m)$$

$$P(g.E_j | g.M = m) = N(\mu_{j,m}, \sigma_{j,m})$$

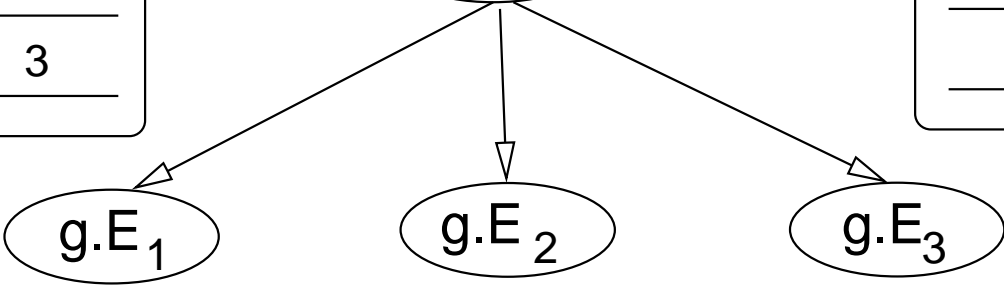
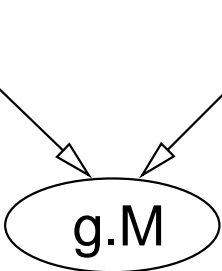
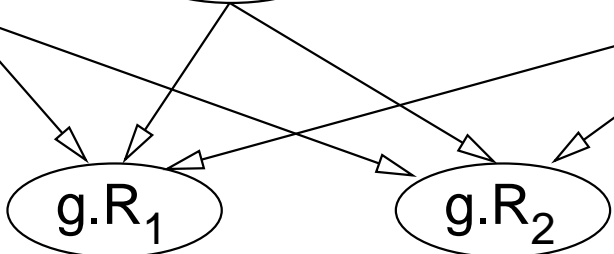
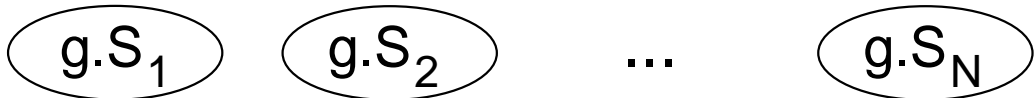
For each module  $m$  and each condition  $j$ :

Mean:  $\mu_{j,m}$

Standard deviation:  $\sigma_{j,m}$

# Parameter estimation

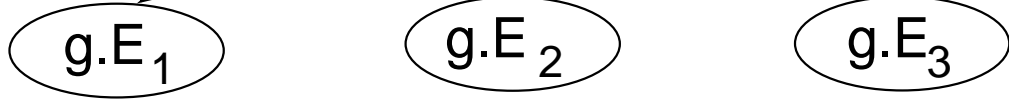
$$P(g.R_2 | g.S) \quad \text{T A T}^c \text{ A}^c \text{ G}^c$$



$g.R_1$	$g.R_2$	$g.M$
		1
		2
		3

$P(g.E_3   g.M)$	$g.M$
	1
	2
	3

$$P(g.R_2 | g.S) \quad T A T^c \quad A^c G$$



$g.R_1$	$g.R_2$	$g.M$
		1
		2
		3

$P(g.E_3   g.M)$	$g.M$
	1
	2
	3

## Bayesian approach

$$P(\text{parameters} \mid \text{data}) = \sum P(\text{parameters, latent variables} \mid \text{data})$$

## Bayesian approach

$$P(\text{parameters} \mid \text{data}) = \sum P(\text{parameters, latent variables} \mid \text{data})$$

Intractable!

## Bayesian approach

$$P(\text{parameters} \mid \text{data}) = \sum P(\text{parameters, latent variables} \mid \text{data})$$

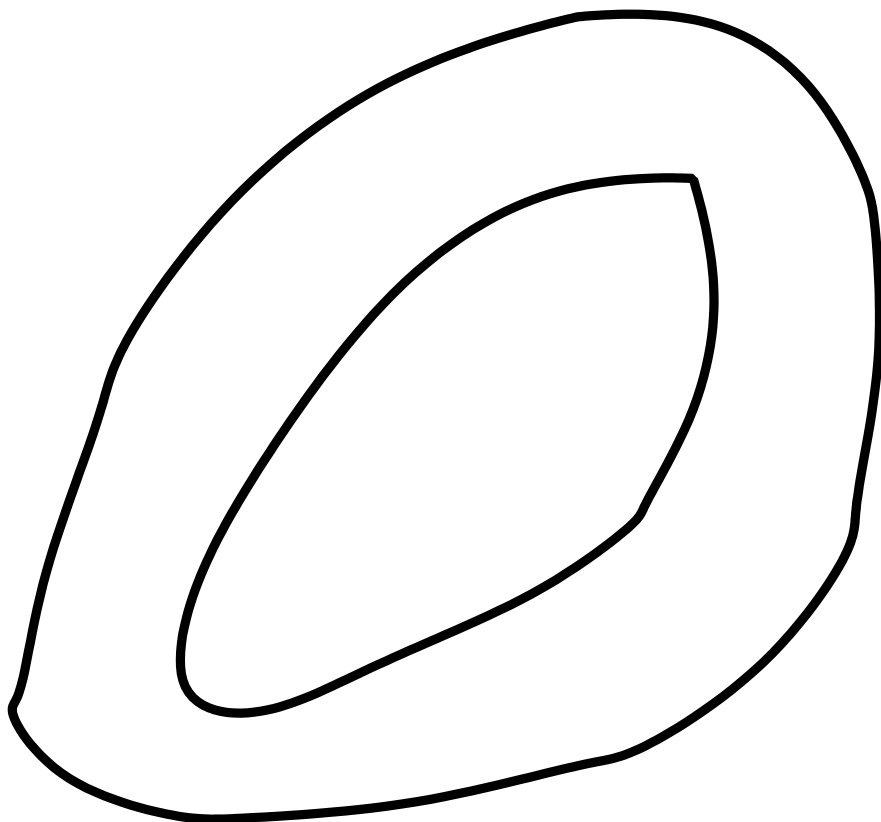
Intractable!

## Gibbs sampling

$\text{parameters} \sim P(\text{parameters} \mid \text{latent variables, data})$

$\text{latent variables} \sim P(\text{latent variables} \mid \text{parameters, data})$

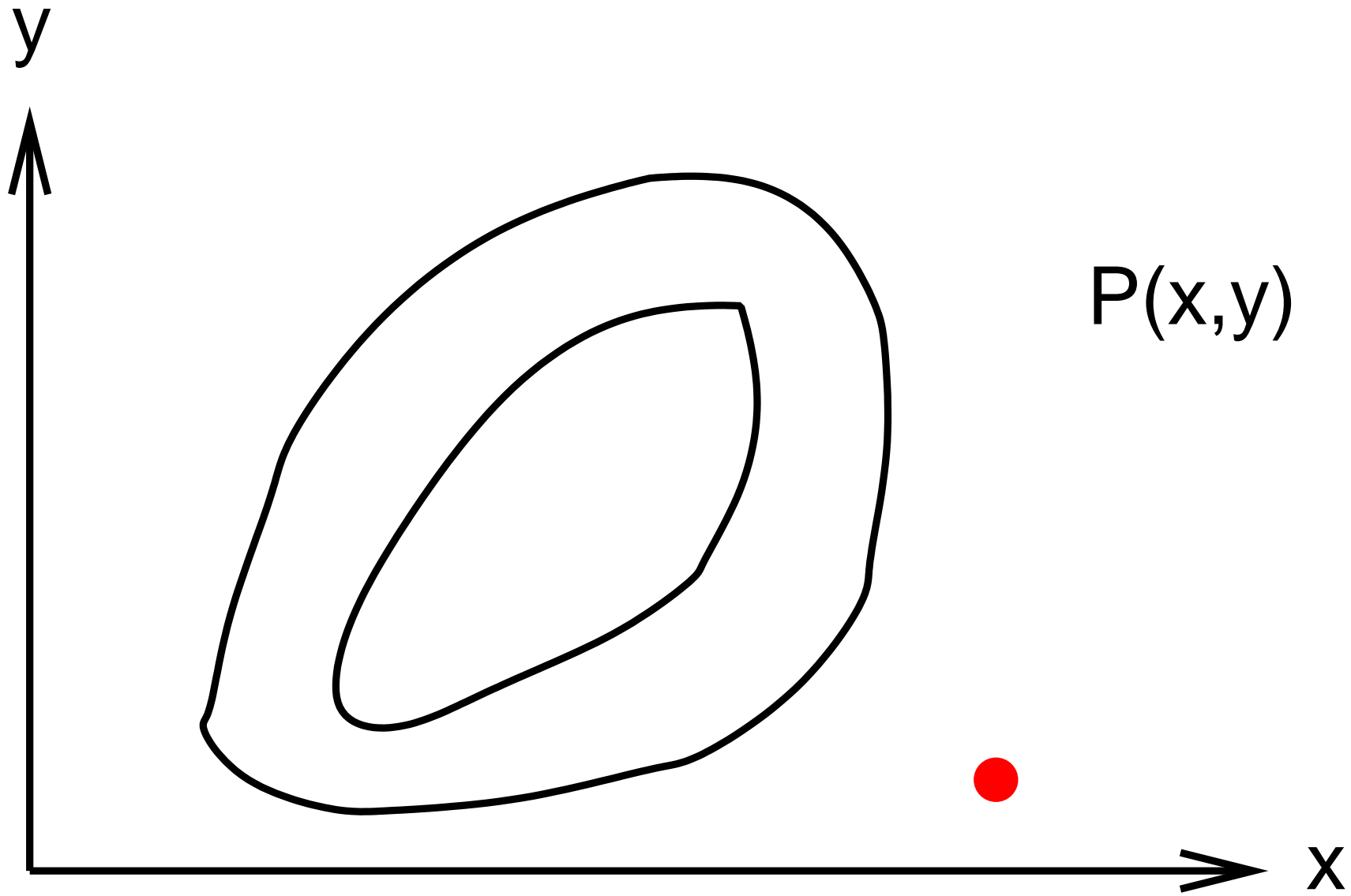
y

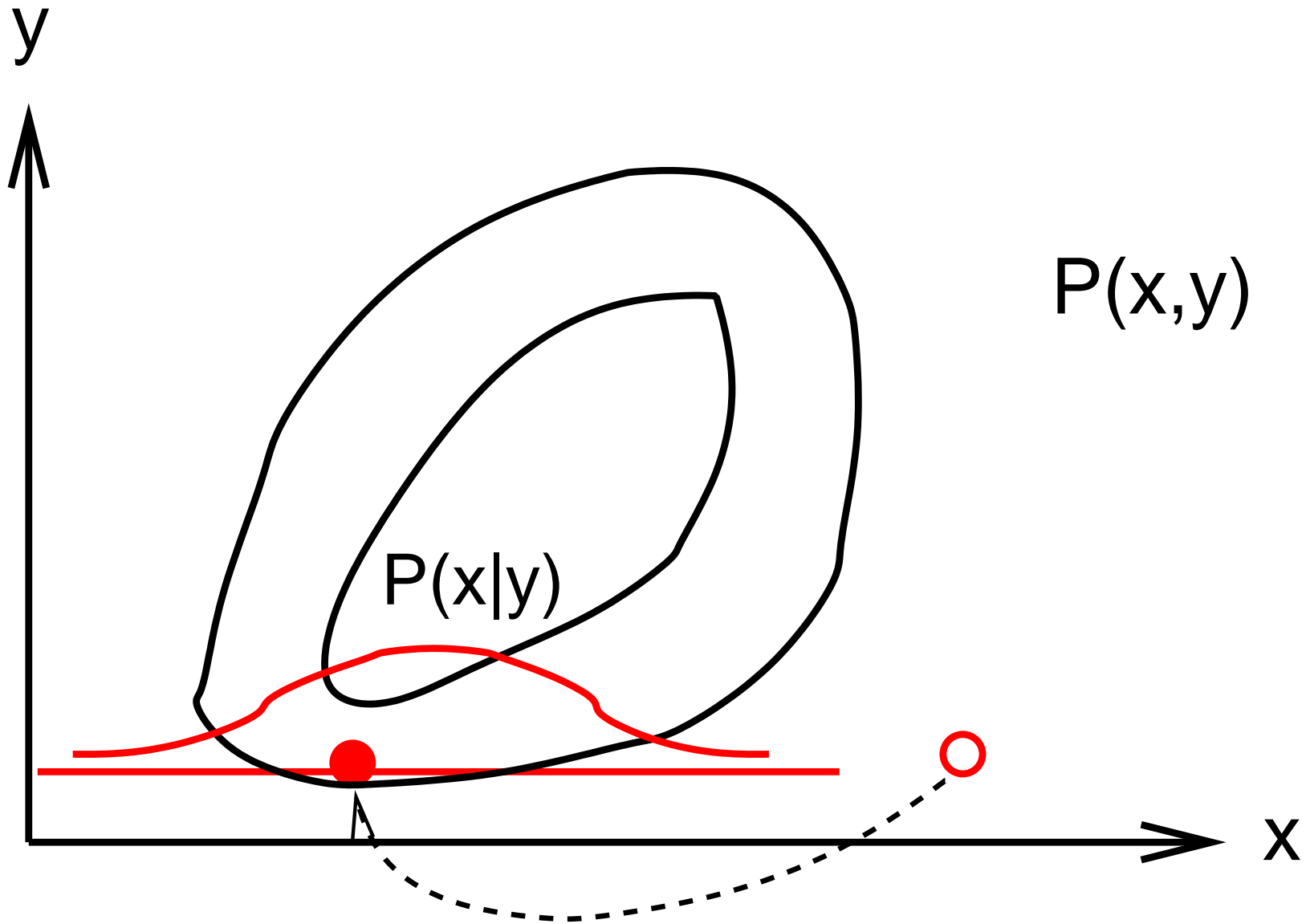


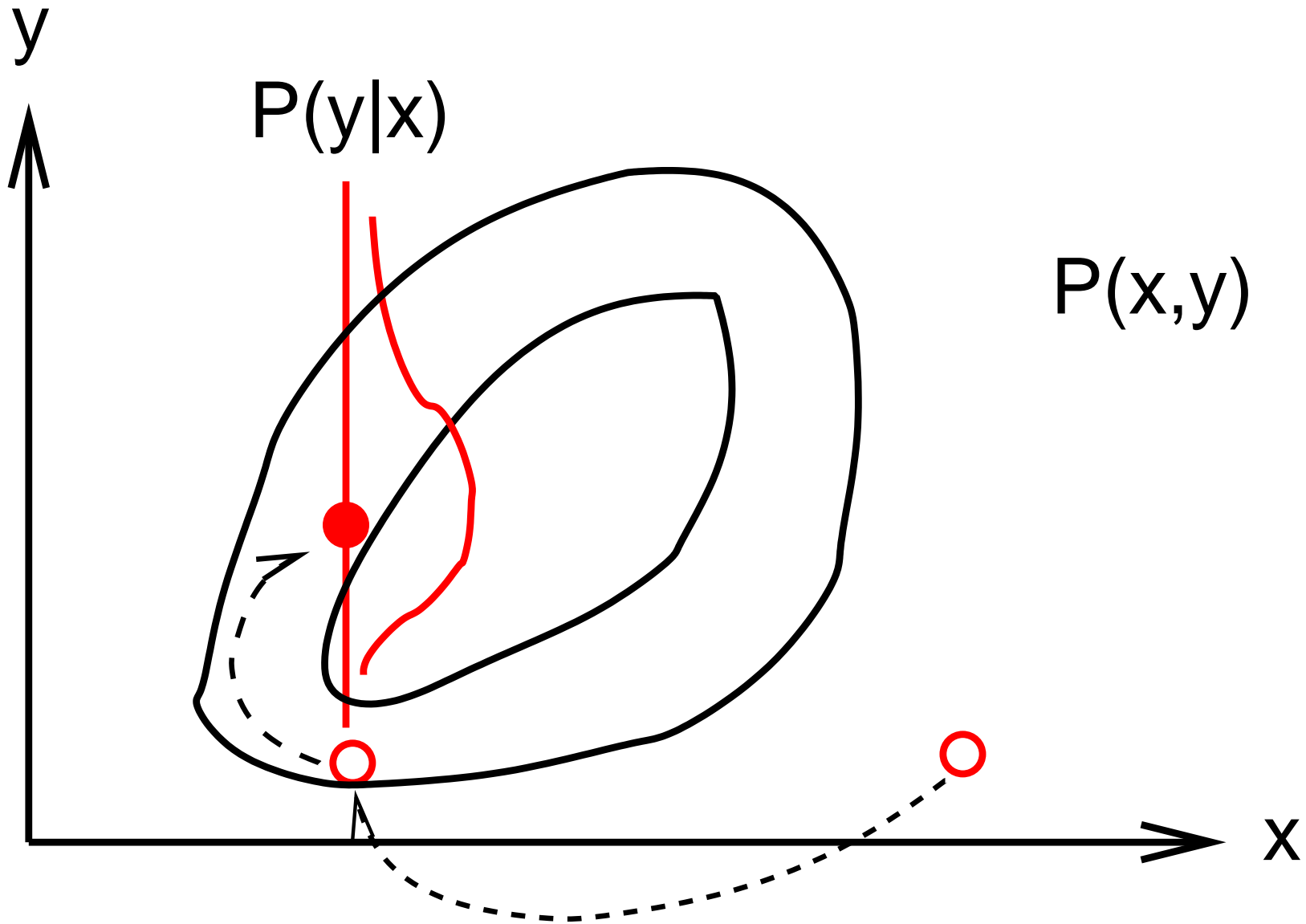
$P(x,y)$

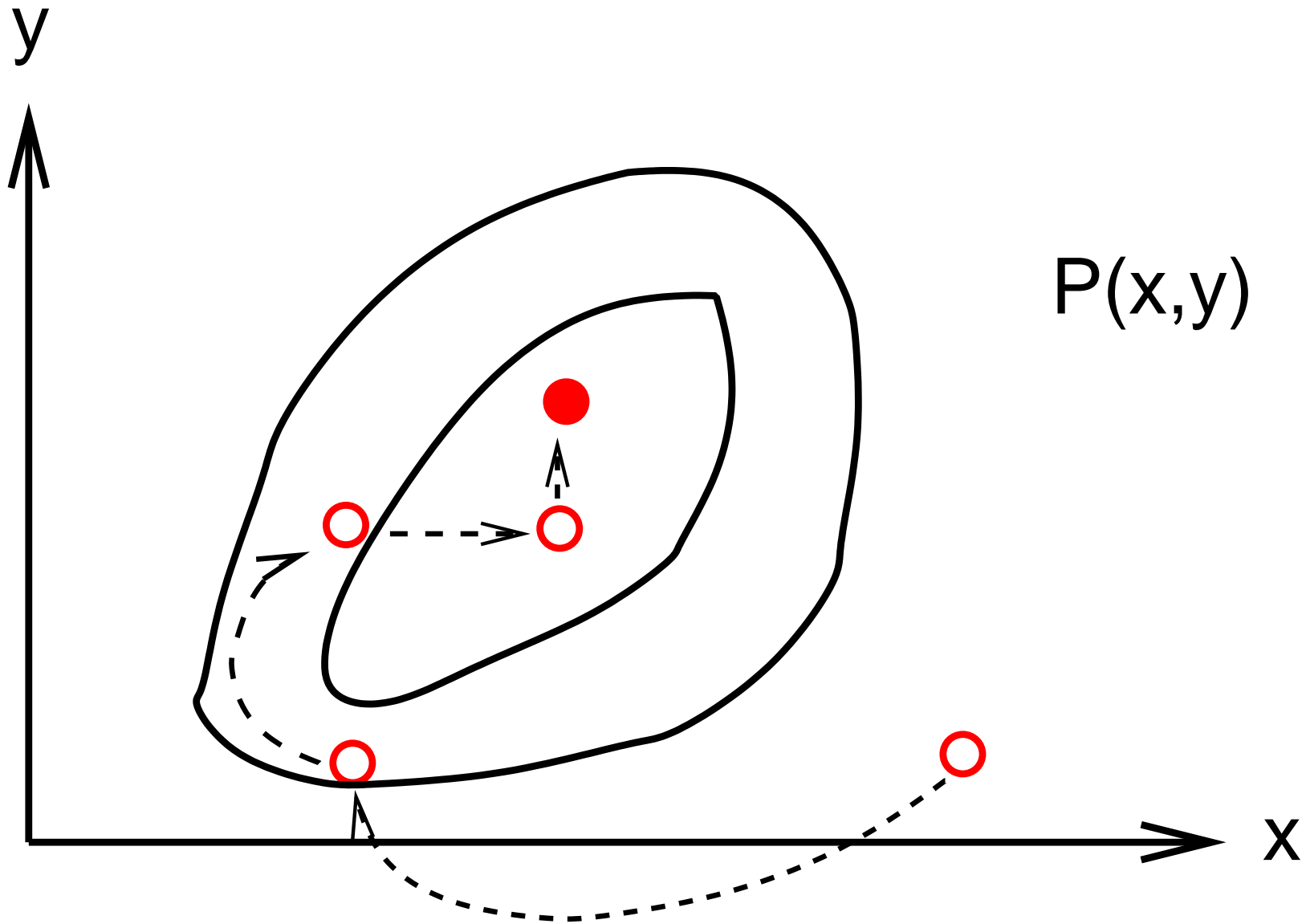


x









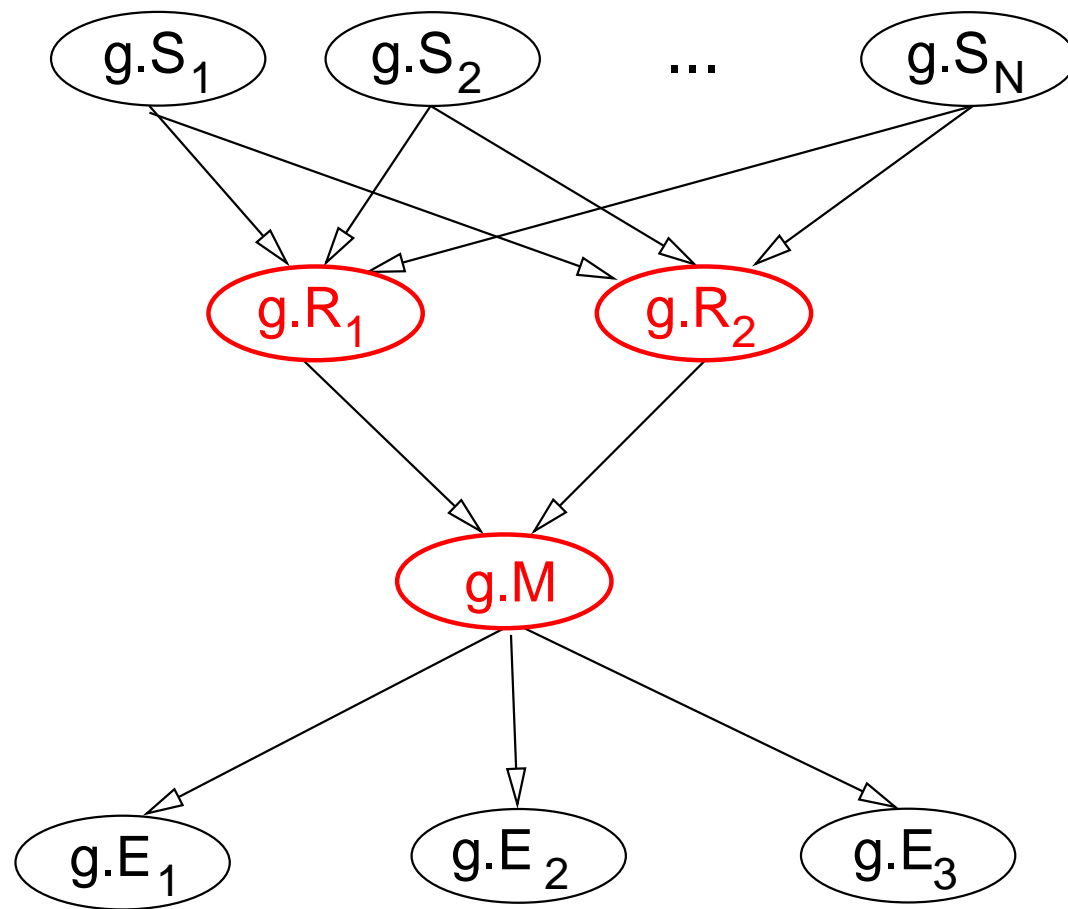
Still too expensive

Find one “good” set of parameters rather than a whole sample from the posterior distribution.

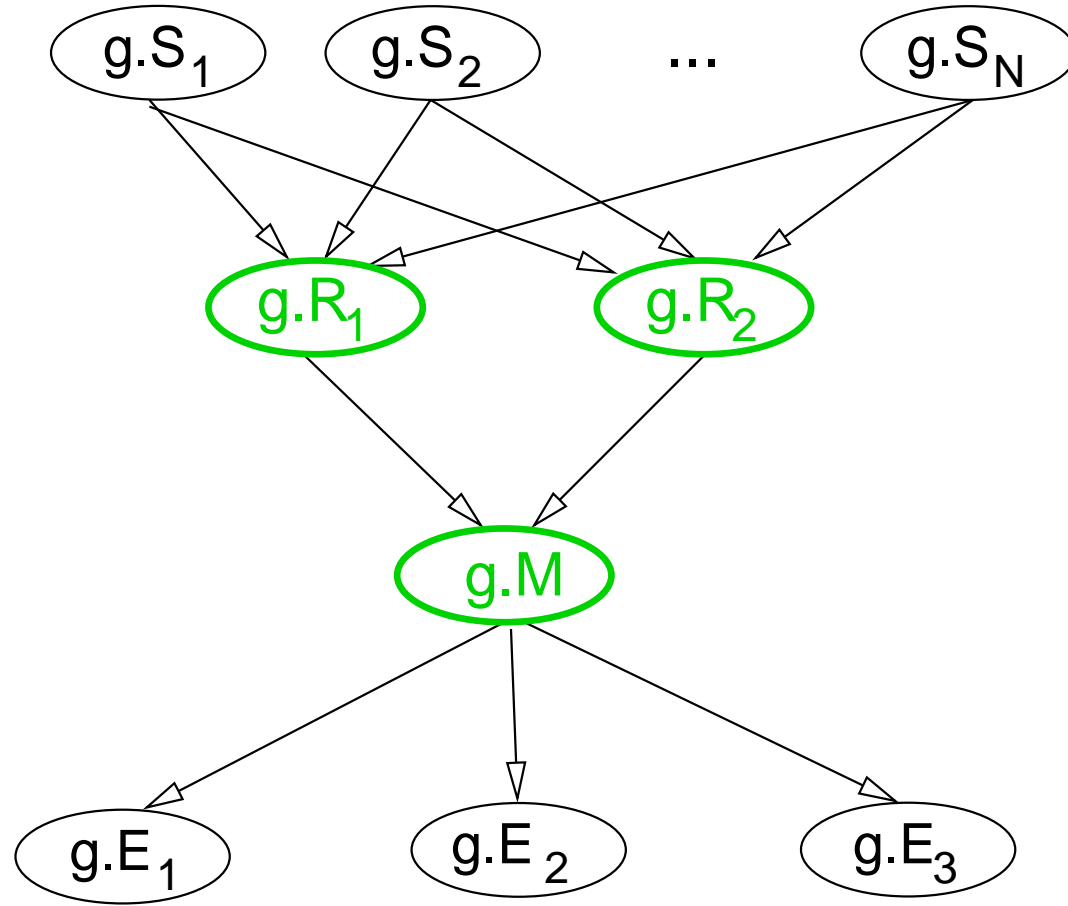
Hard-assignment EM algorithm.

Various heuristic simplifications.

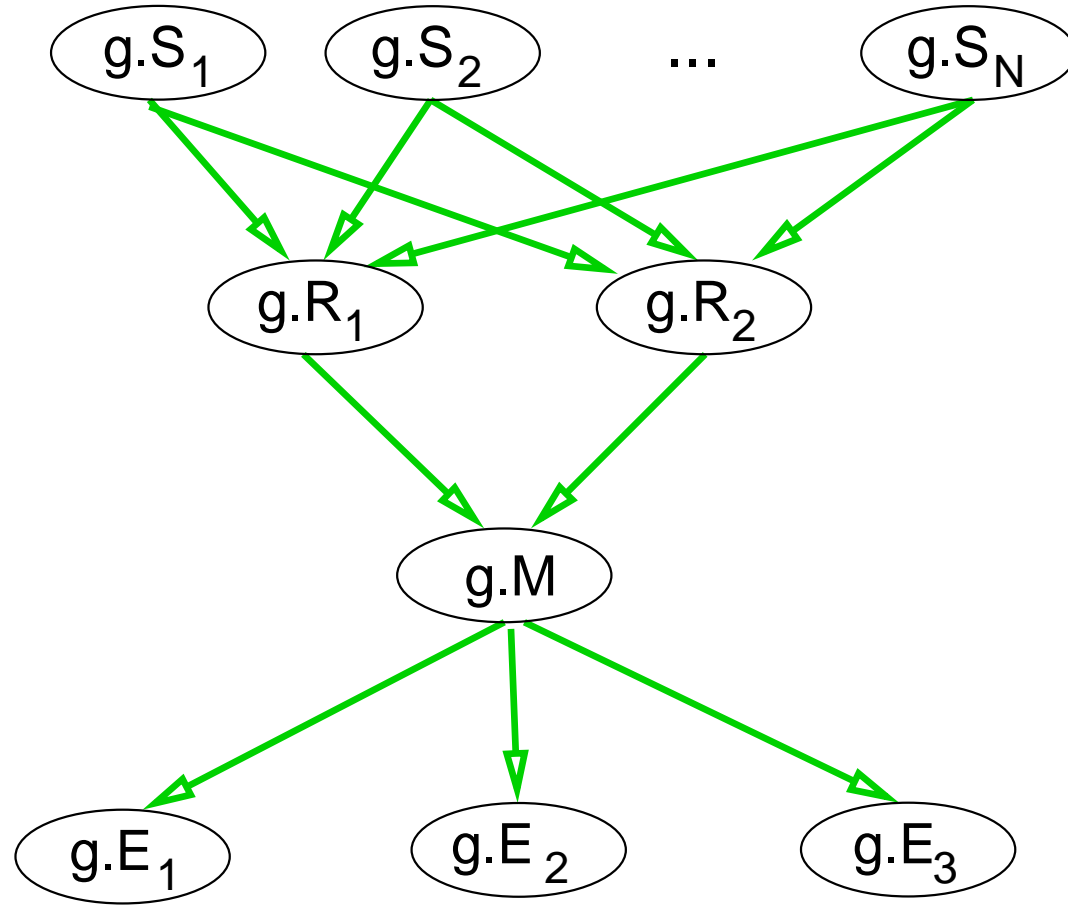
See [Bioinformatics 19, 2003](#) for details.

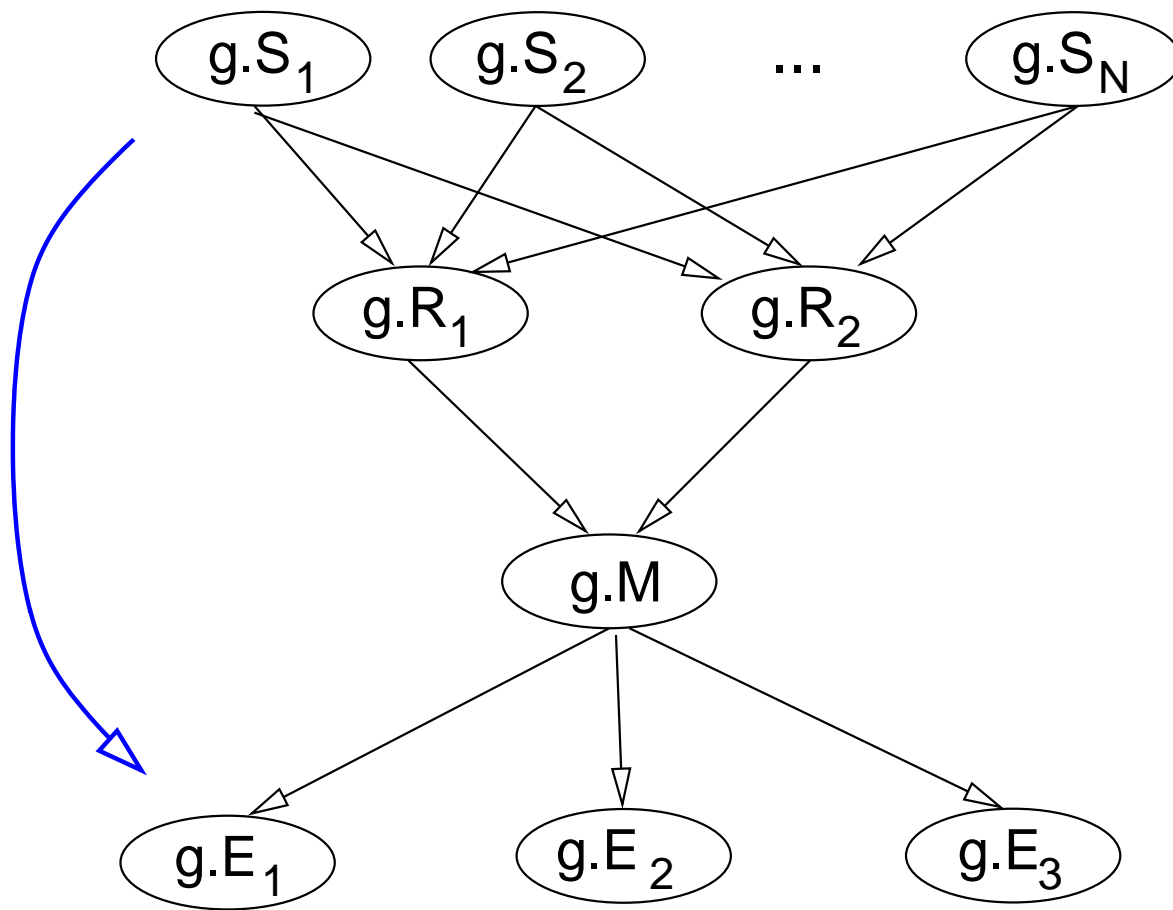


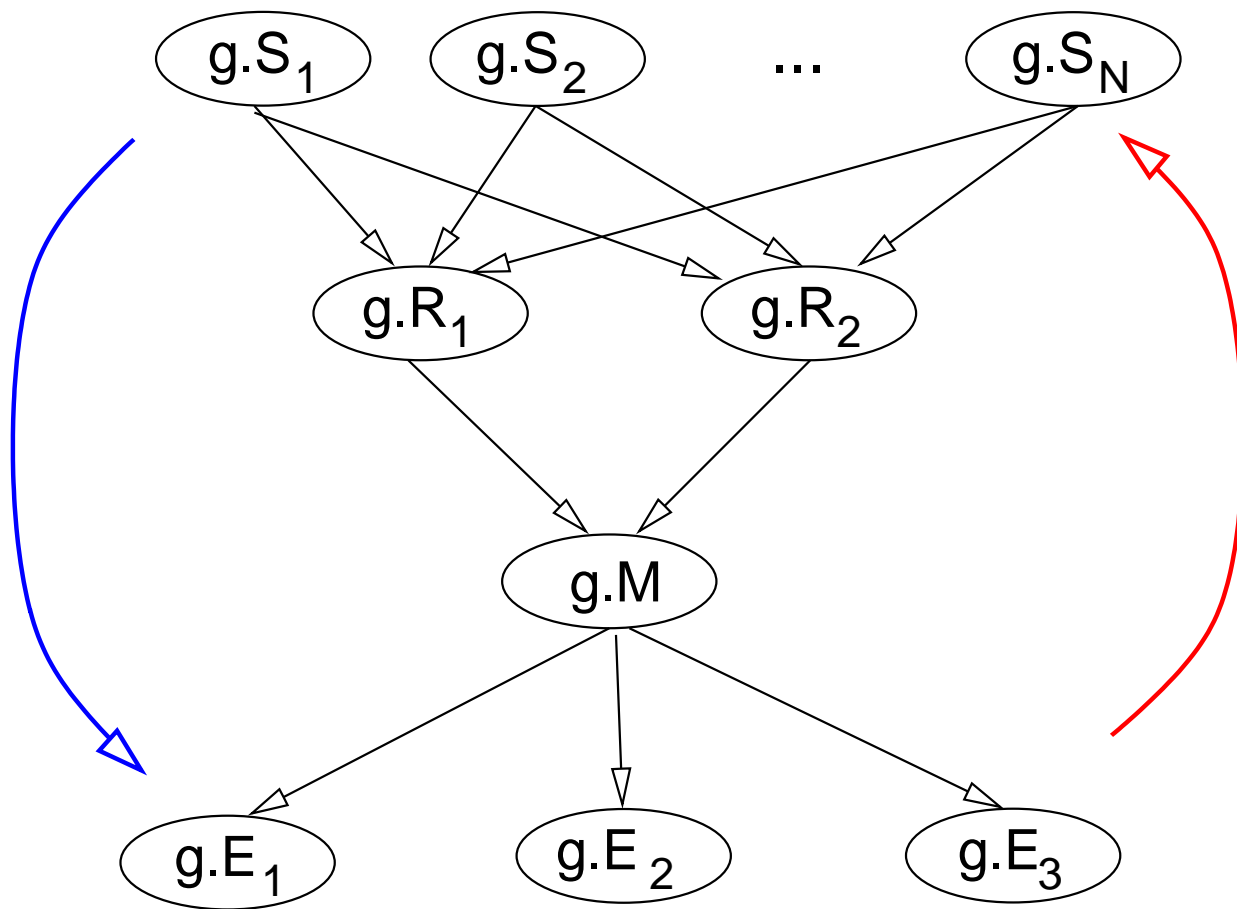
# E-step



M-step



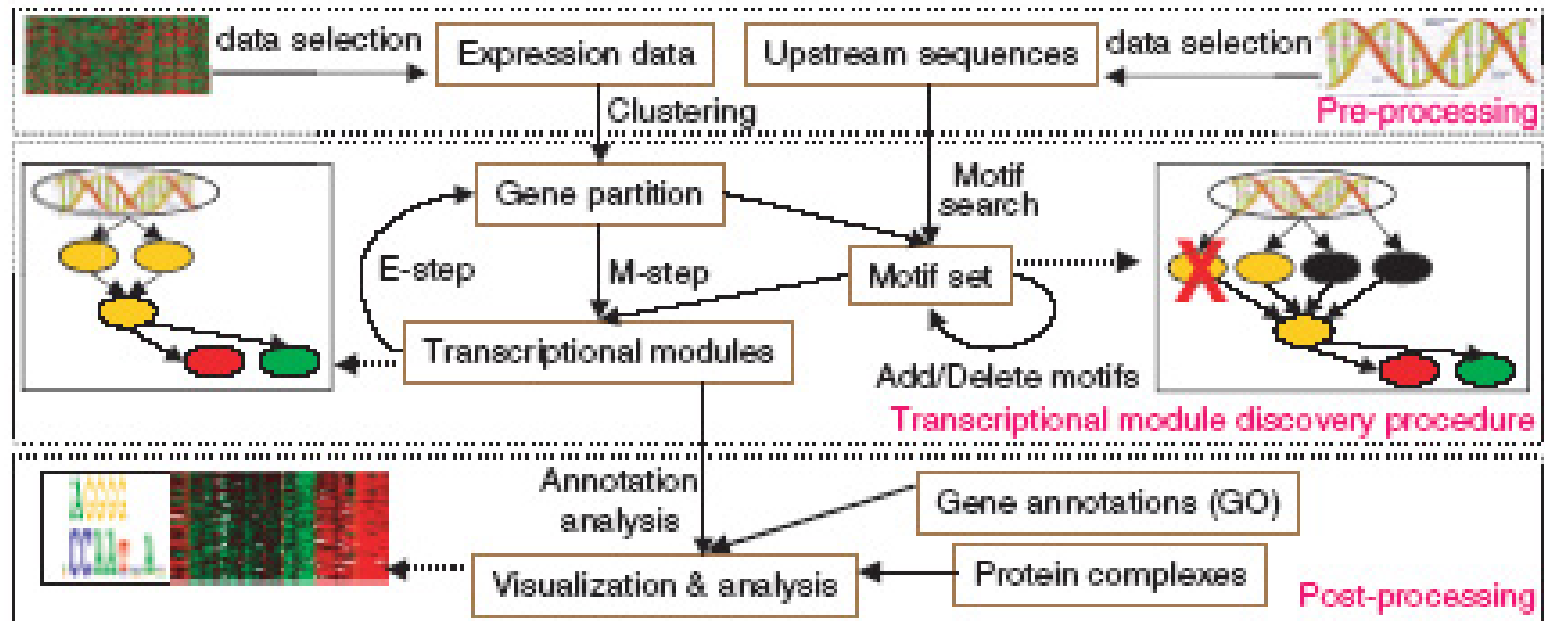




Segal, Yelensky, Koller (2003)

Bioinformatics 19

*Saccharomyces cerevisiae*



From Segal et al., Bioinformatics 2003

# Experiment 1

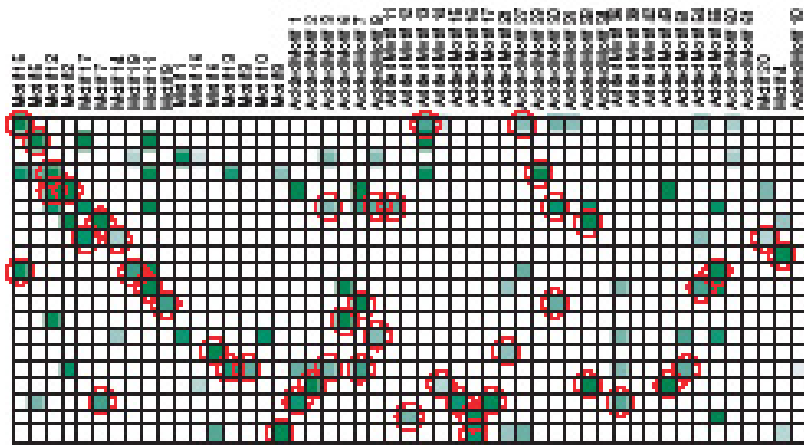
173 microarrays, measuring responses to various stress conditions (Gasch et al. 2000)

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

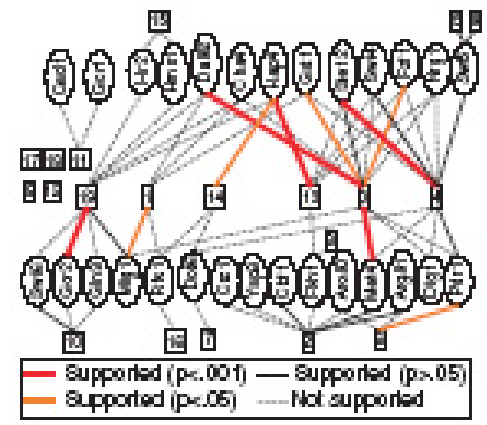
## Experiment 2

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

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



- 28 - Glycolysis (68)
- 29 - Redox regulation (118)
- 17 - Alcohol (22)
- 16 - Protein folding (42)
- 15 - Chromatin remodeling (8)
- 14 - Energy and TCA cycle (66)
- 13 - Cellulose phosphorylation (21)
- 12 - Unknown (17)
- 11 - Nitrogen metabolism (36)
- 10 - Cell cycle (44)
- 9 - Protein folding (47)
- 8 - Sporulation/biosynthesis (10)
- 7 - Lipid metabolism (54)
- 6 - Transcription (6)
- 5 - Iron uptake (61)
- 4 - Phosphorus response (32)
- 3 - Transport (30)
- 2 - Aldehyde metabolism (34)
- 1 - Galactose metabolism (16)
- 0 - Amino acid metabolism (32)



From Segal et al., Bioinformatics 2003

Dirk Husmeier  
Richard Dybowski  
Stephen Roberts (Eds.)

# Probabilistic Modeling in Bioinformatics and Medical Informatics

 Springer