
Can we infer genetic networks from gene expression data with **Bayesian networks?**

Dirk Husmeier

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

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

Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering:
Learning networks from data
- Estimating the reliability of inference

Outline of the talk

- **Recapitulation: Bayesian networks**
- Reverse engineering:
Learning networks from data
- Estimating the reliability of inference

A

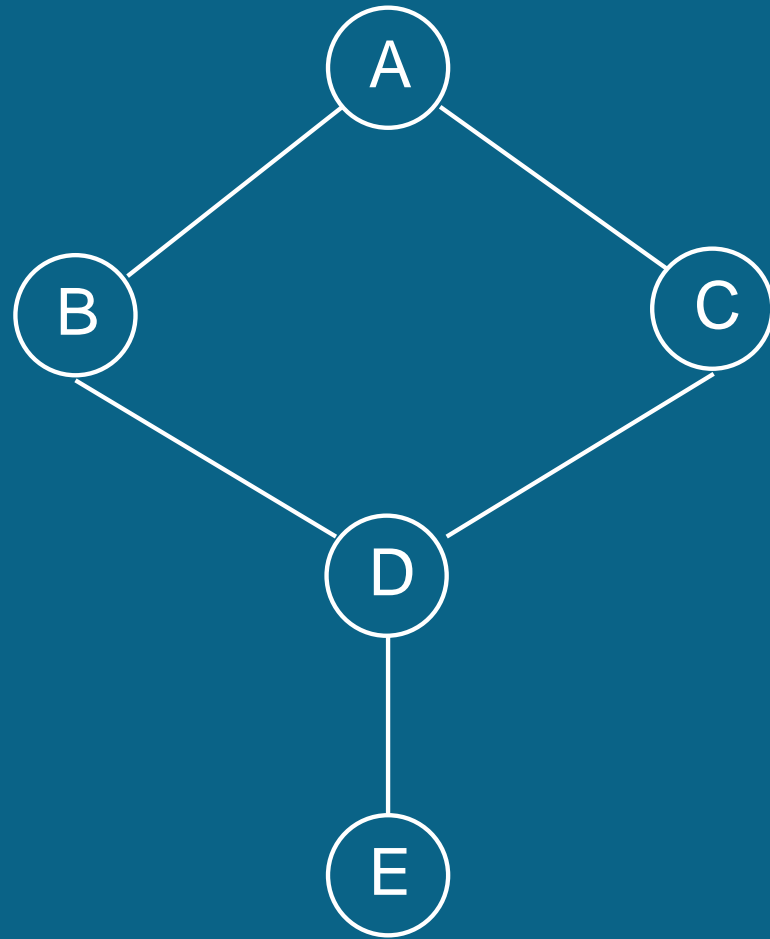
B

C

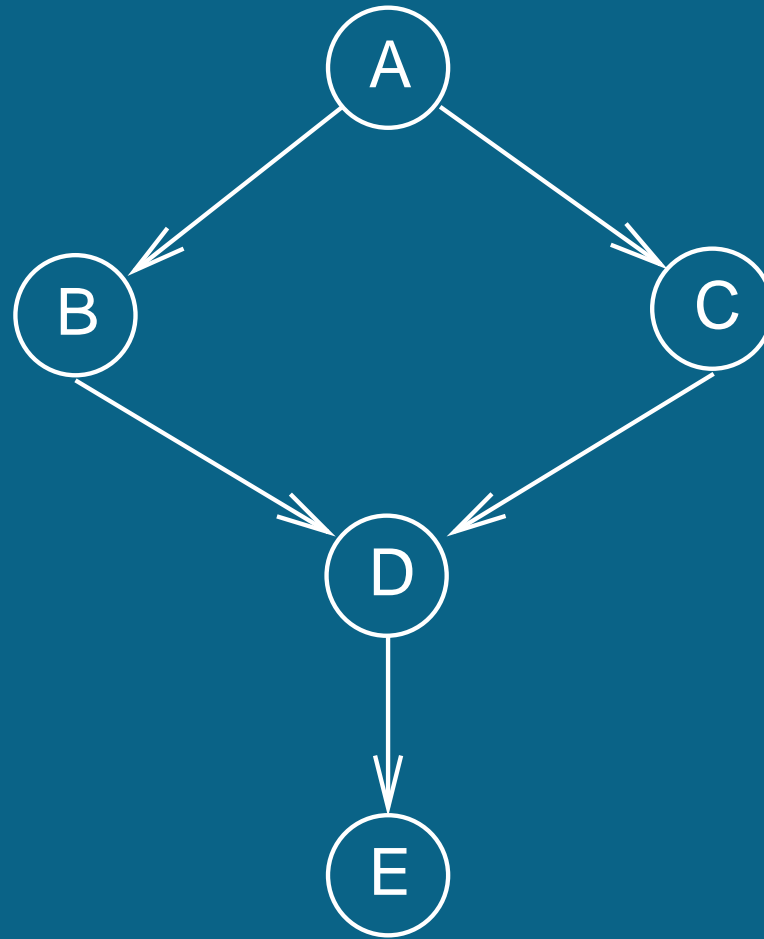
D

E

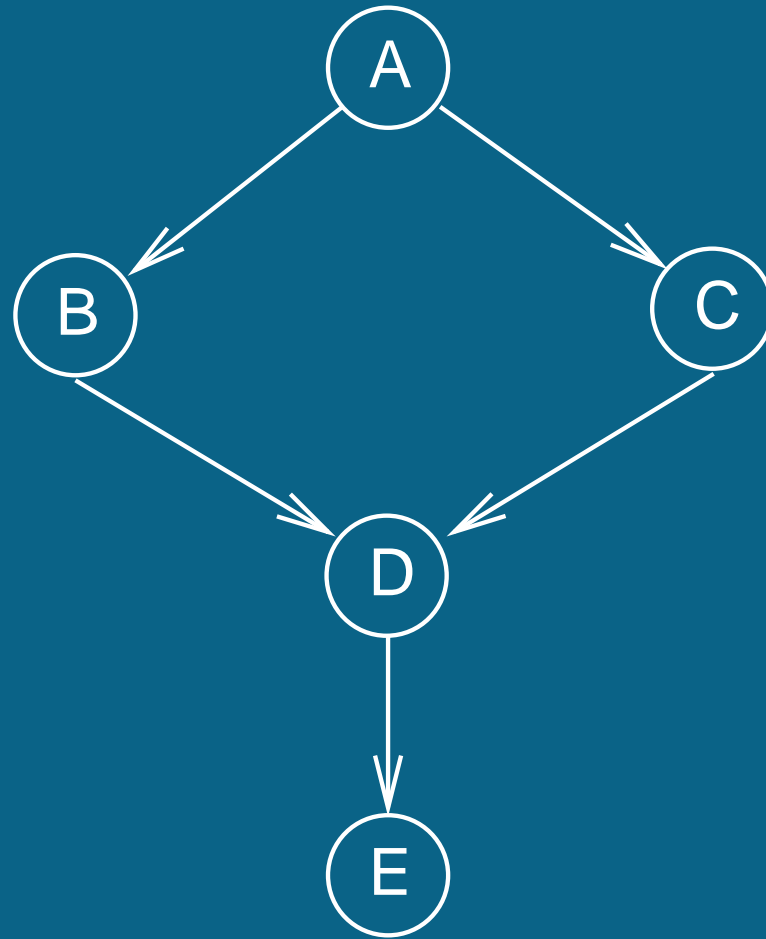
Nodes



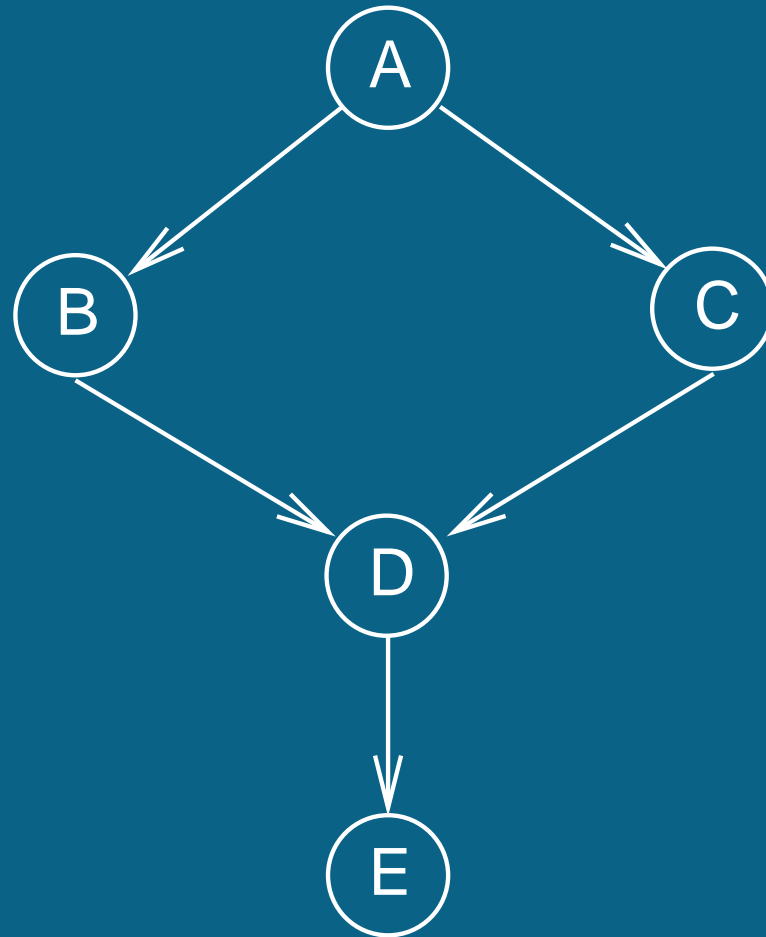
Edges



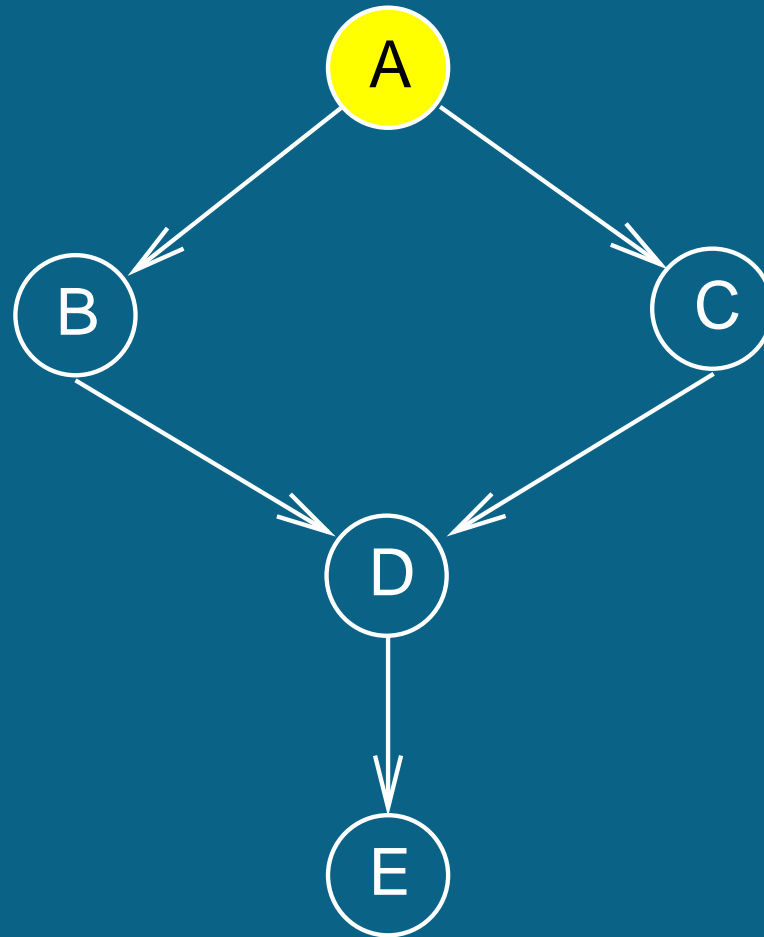
Edges = directed



No directed cycles !

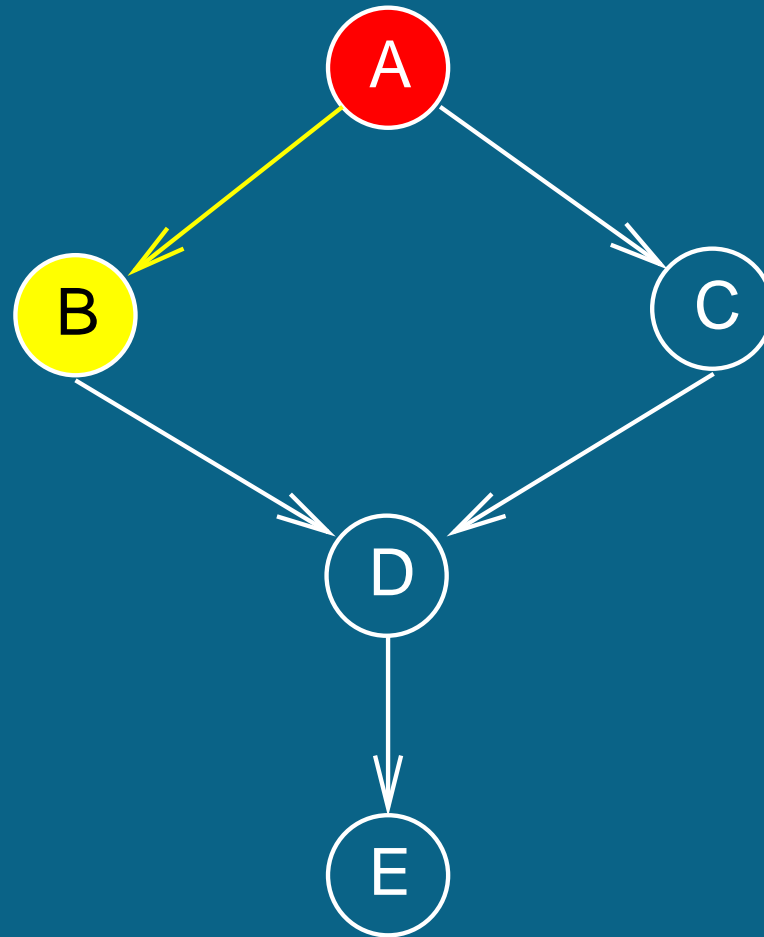


$$P(A, B, C, D, E) = \prod_i P(\text{node}_i | \text{parents}_i)$$

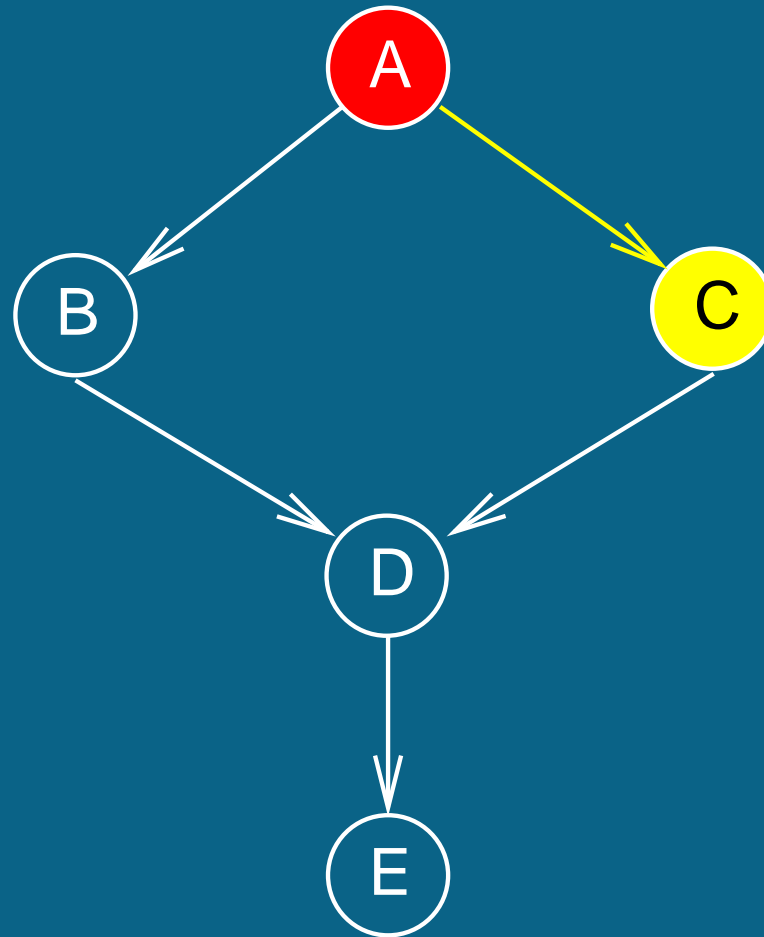


$$P(A, B, C, D, E) =$$

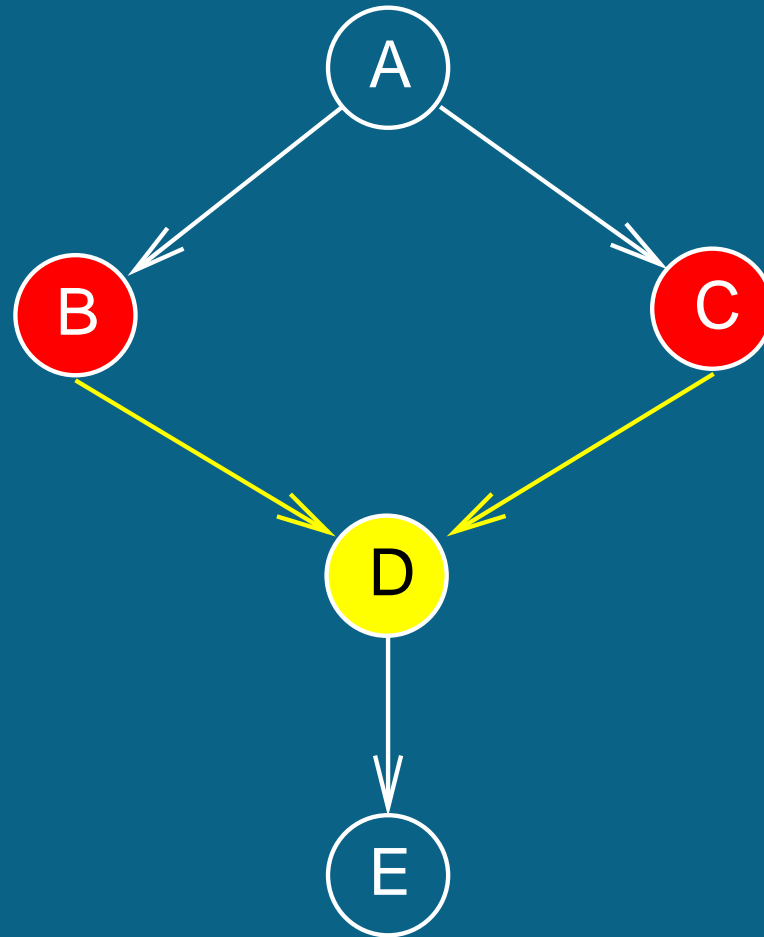
$$P(A)$$



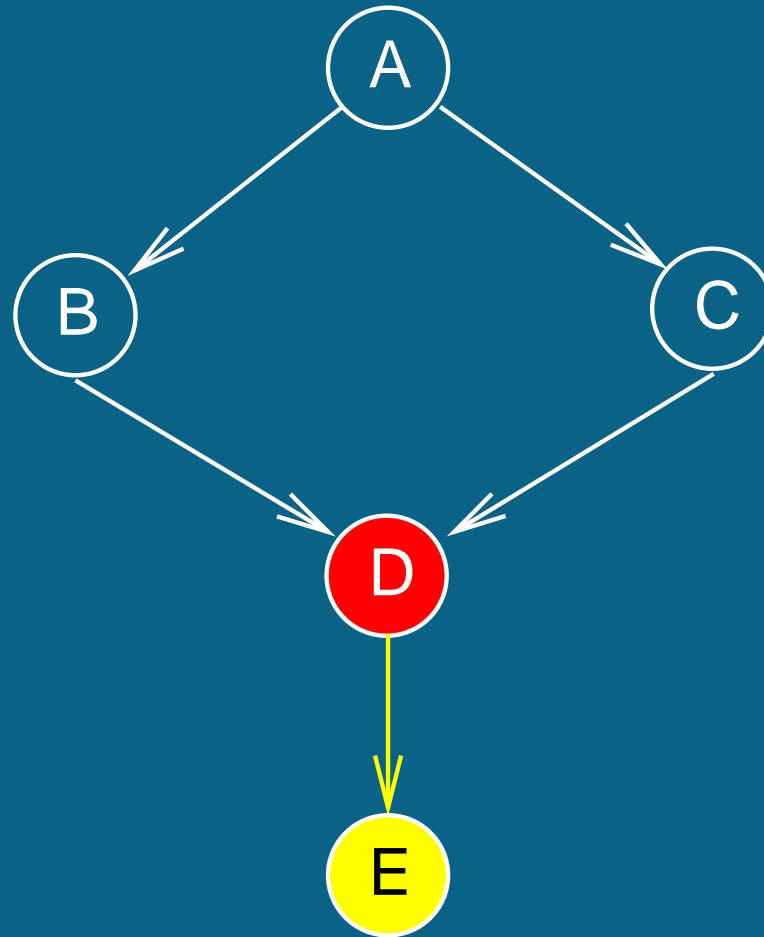
$$P(A, B, C, D, E) = P(A)P(B|A)$$



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

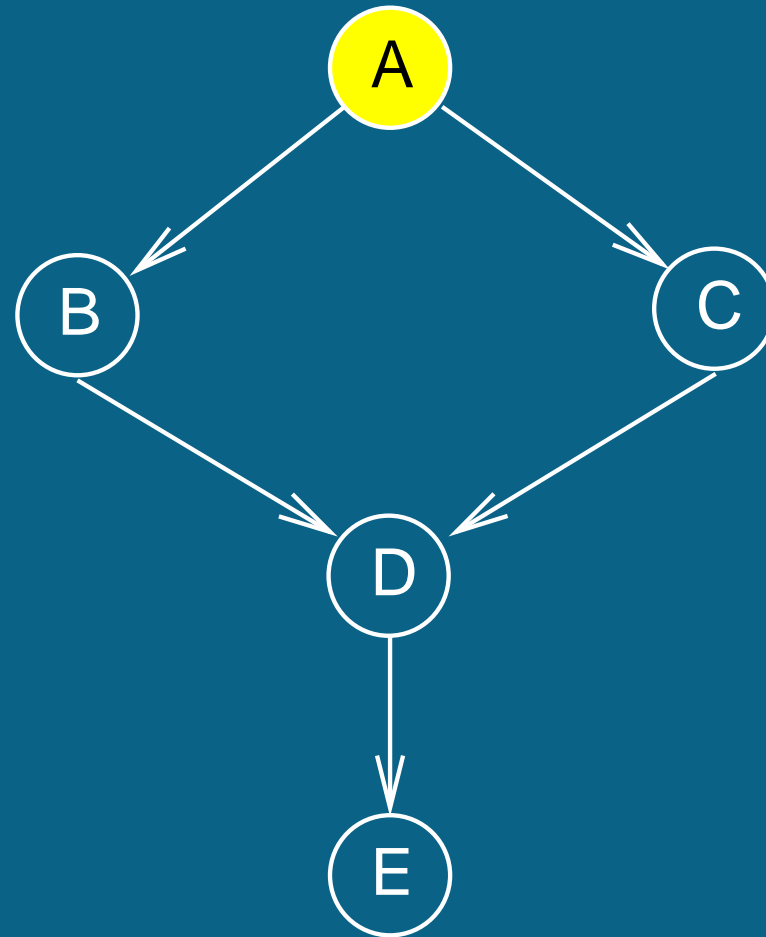


$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)$$

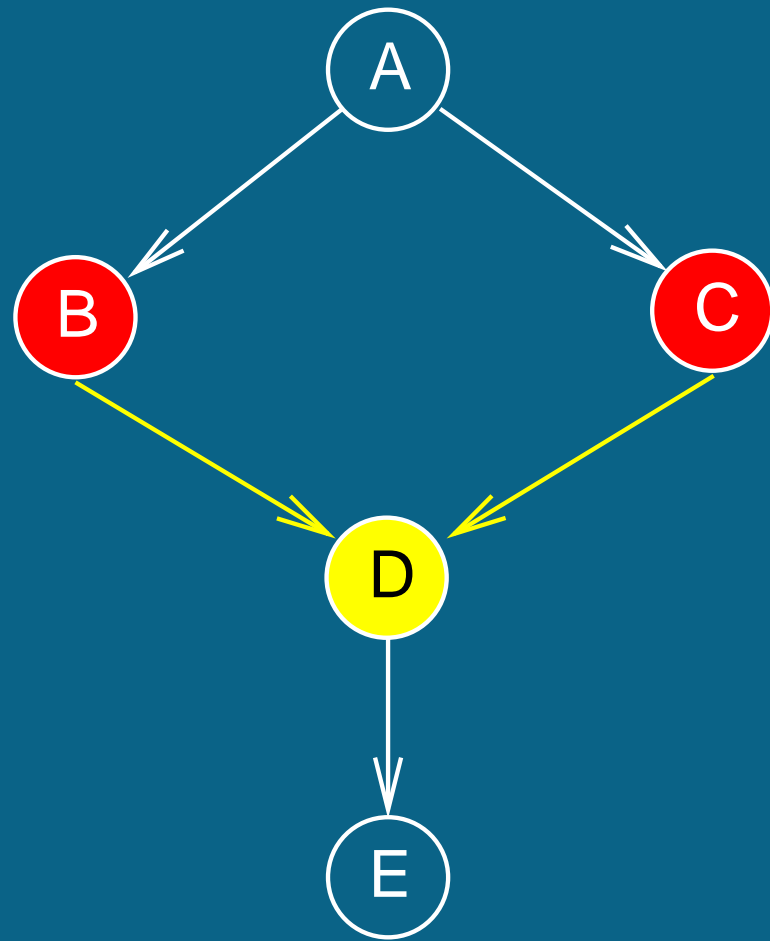


$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)P(E|D)$$

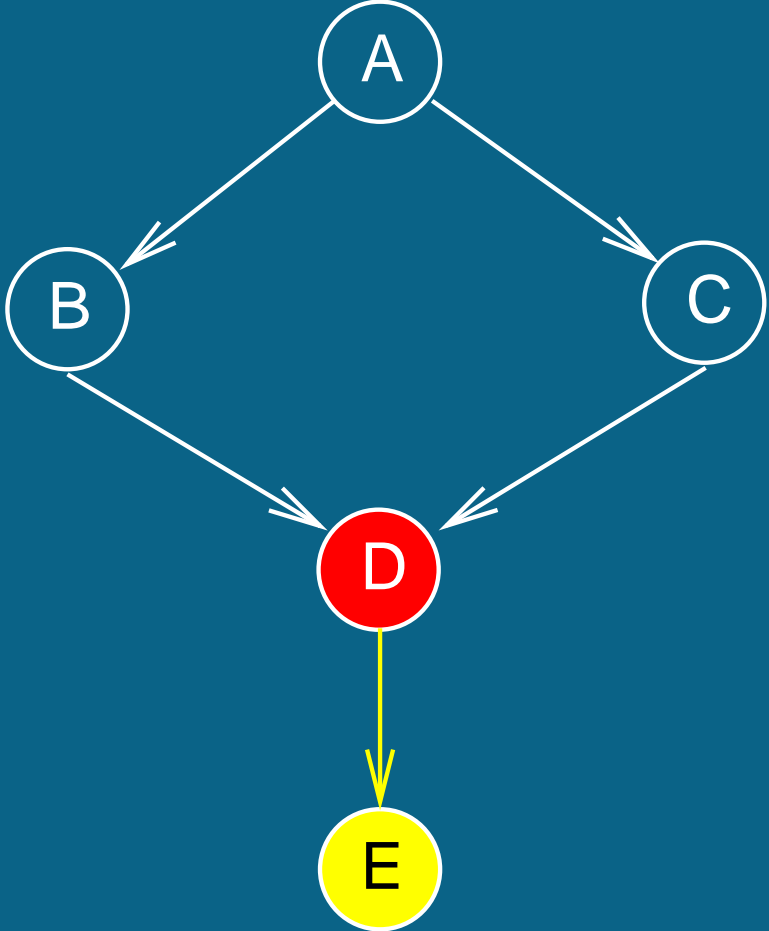
Biological interpretation



Initiation of cell (sub-)cycle

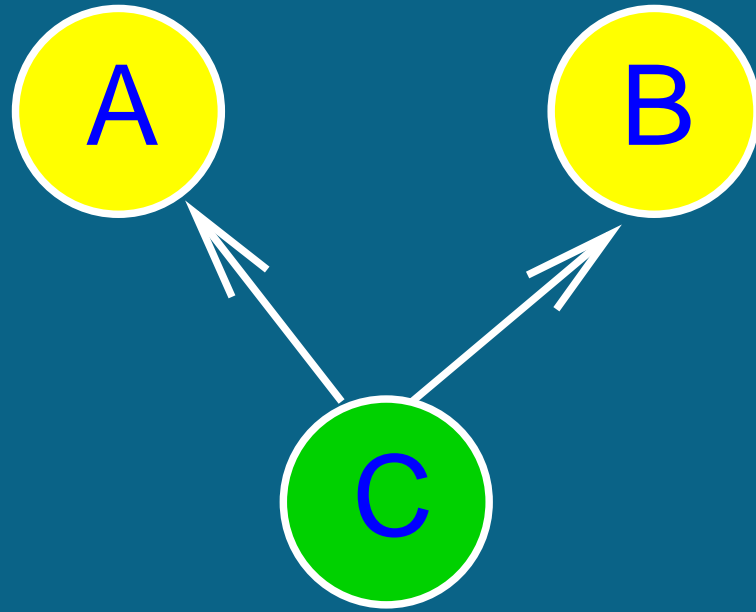


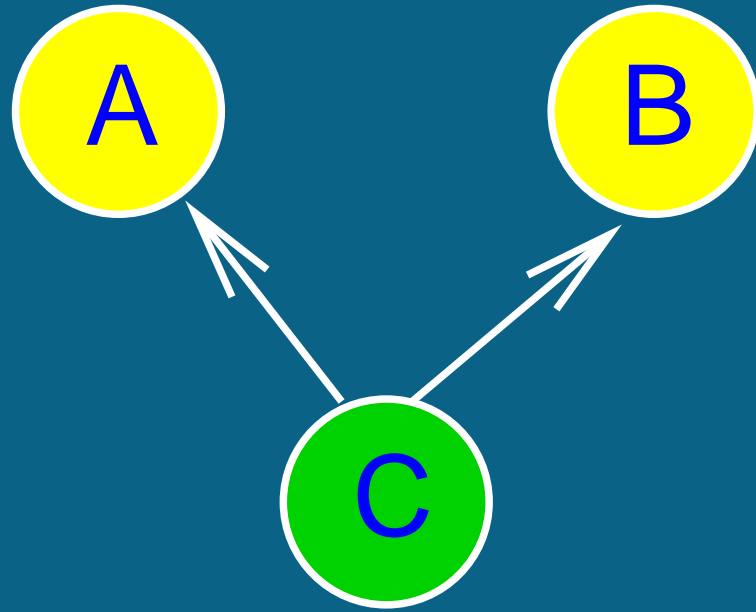
Co-regulation



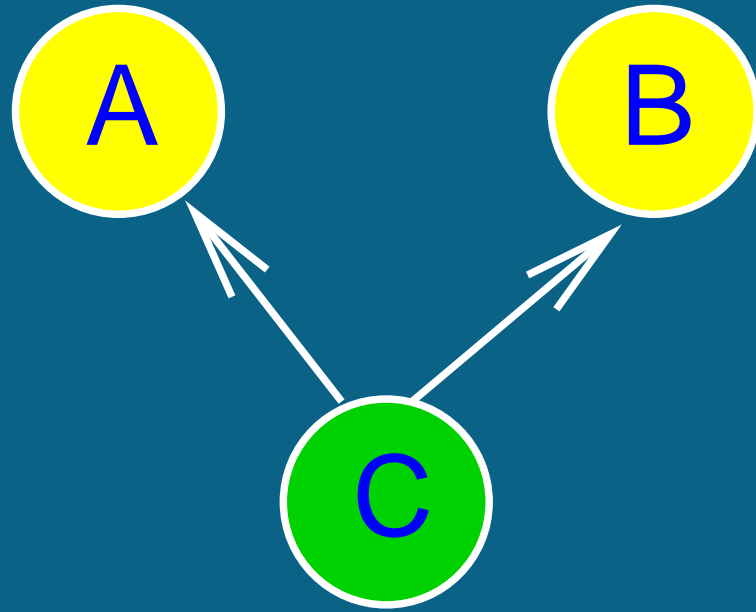
Mediation

Conditional independence relations



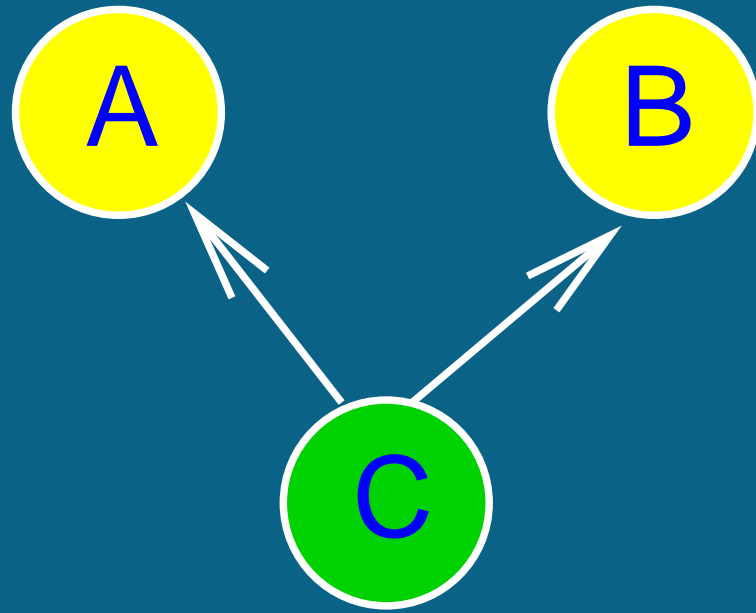


$$P(A, B, C) = P(A|C)P(B|C)P(C)$$



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

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



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

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

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

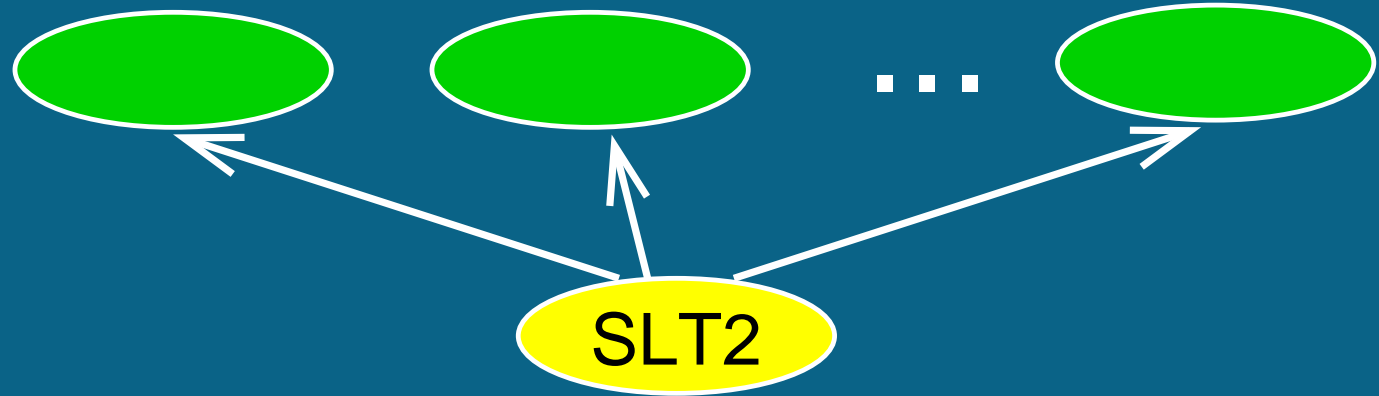
Biological example

Yeast cell cycle

Nir Friedman et al. (2000)

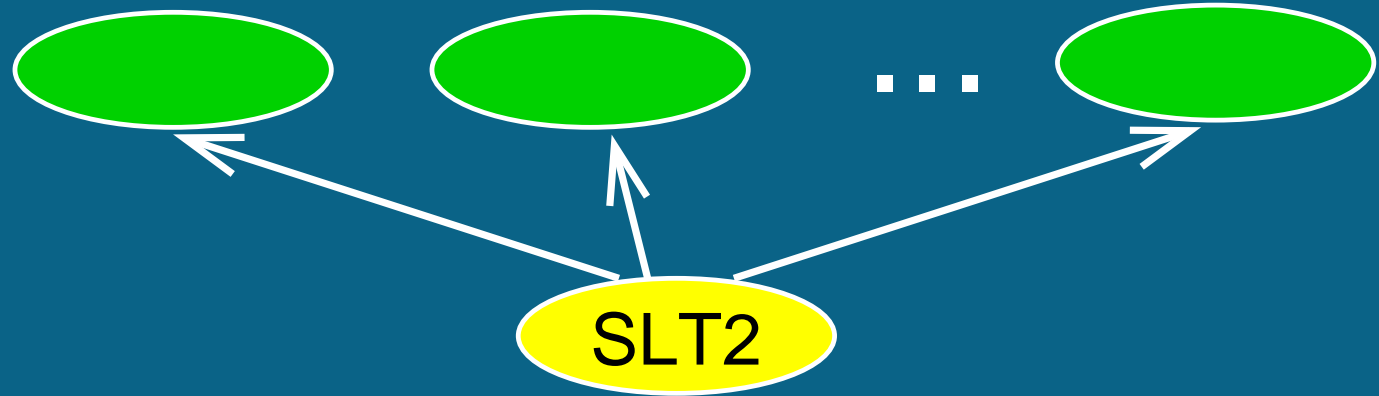
Journal of Computational Biology 7: 601-620

Low osmolarity response genes

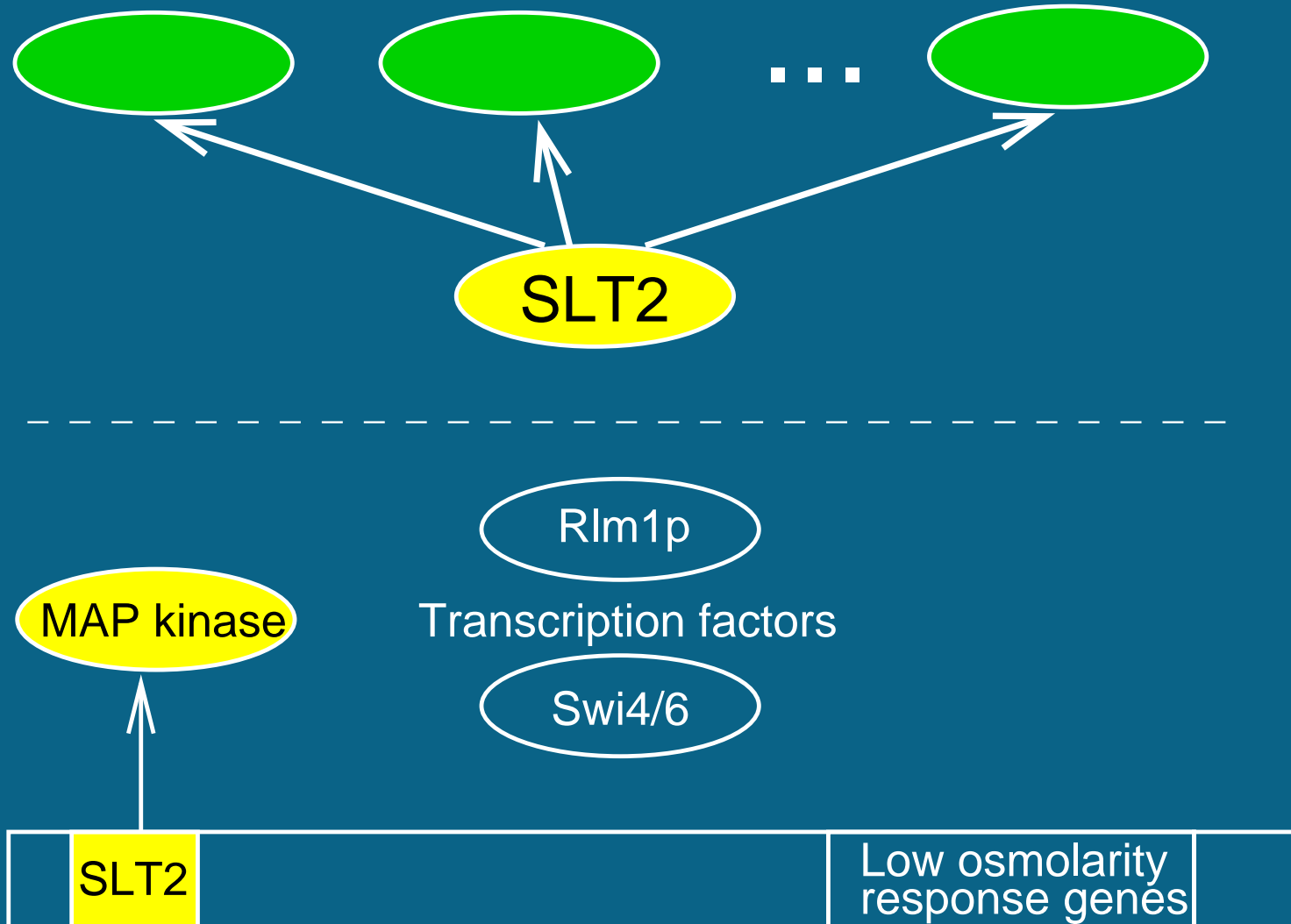


	SLT2		Low osmolarity response genes	
--	------	--	-------------------------------	--

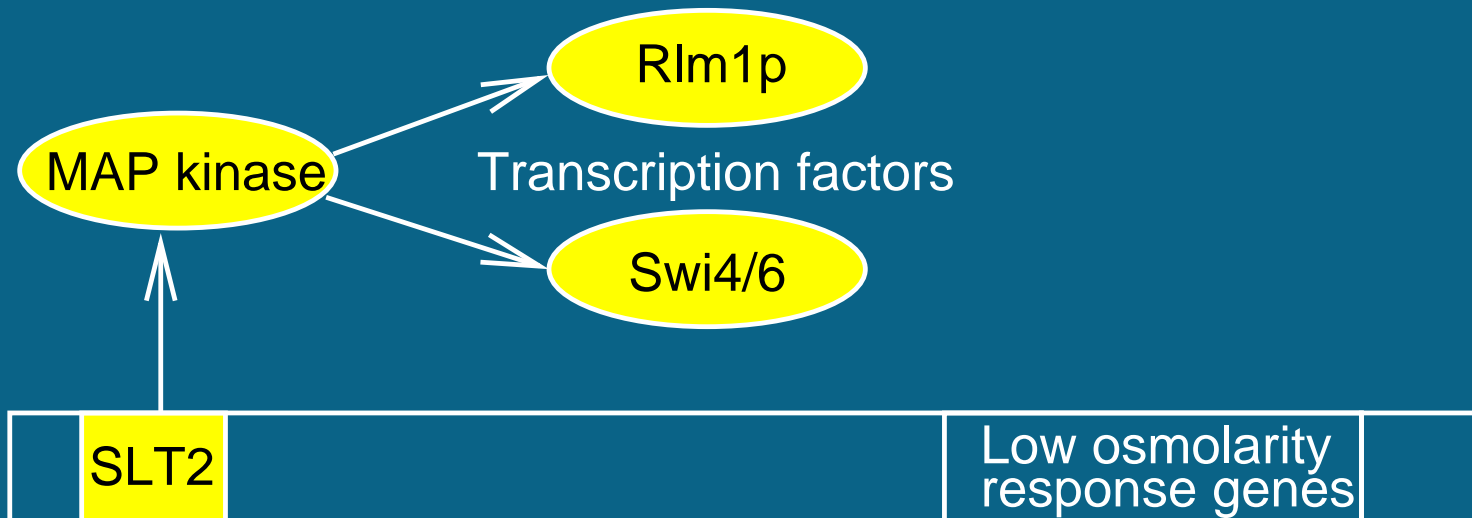
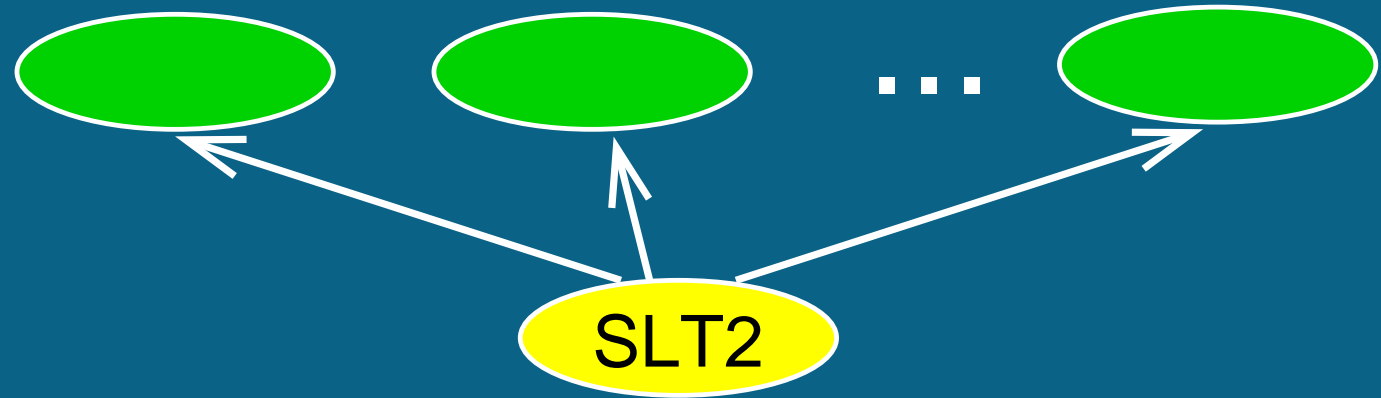
Low osmolarity response genes



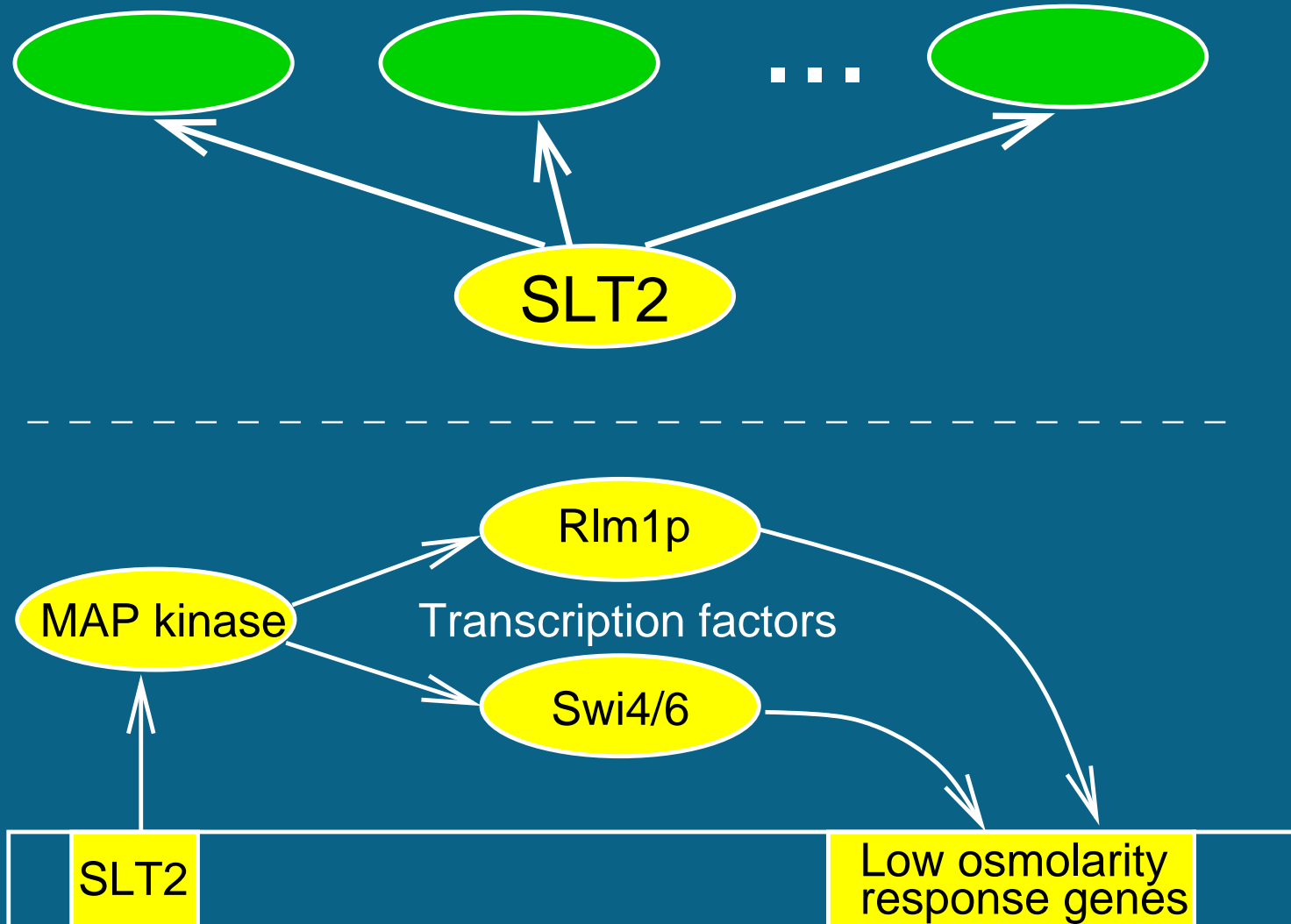
Low osmolarity response genes



Low osmolarity response genes

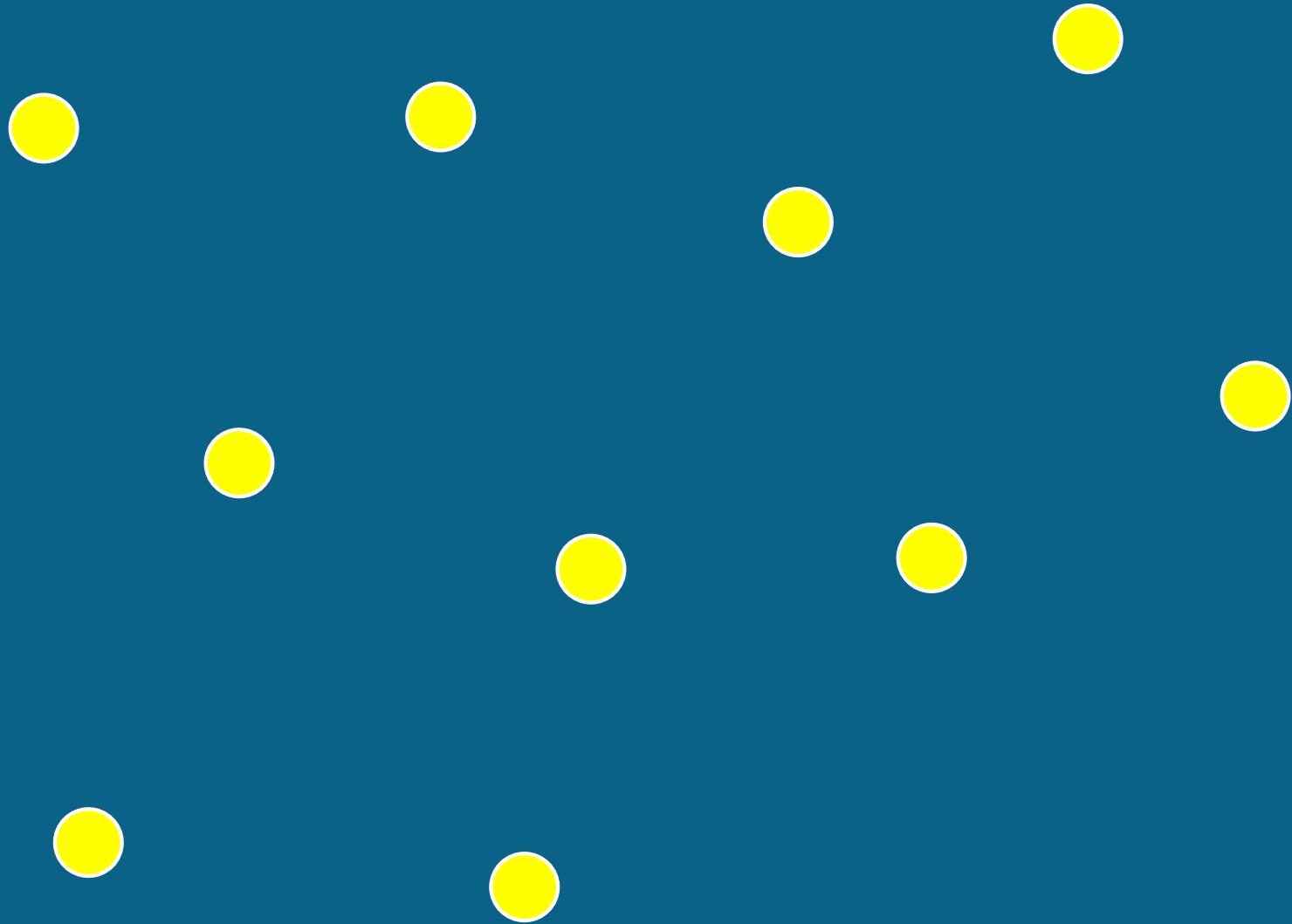


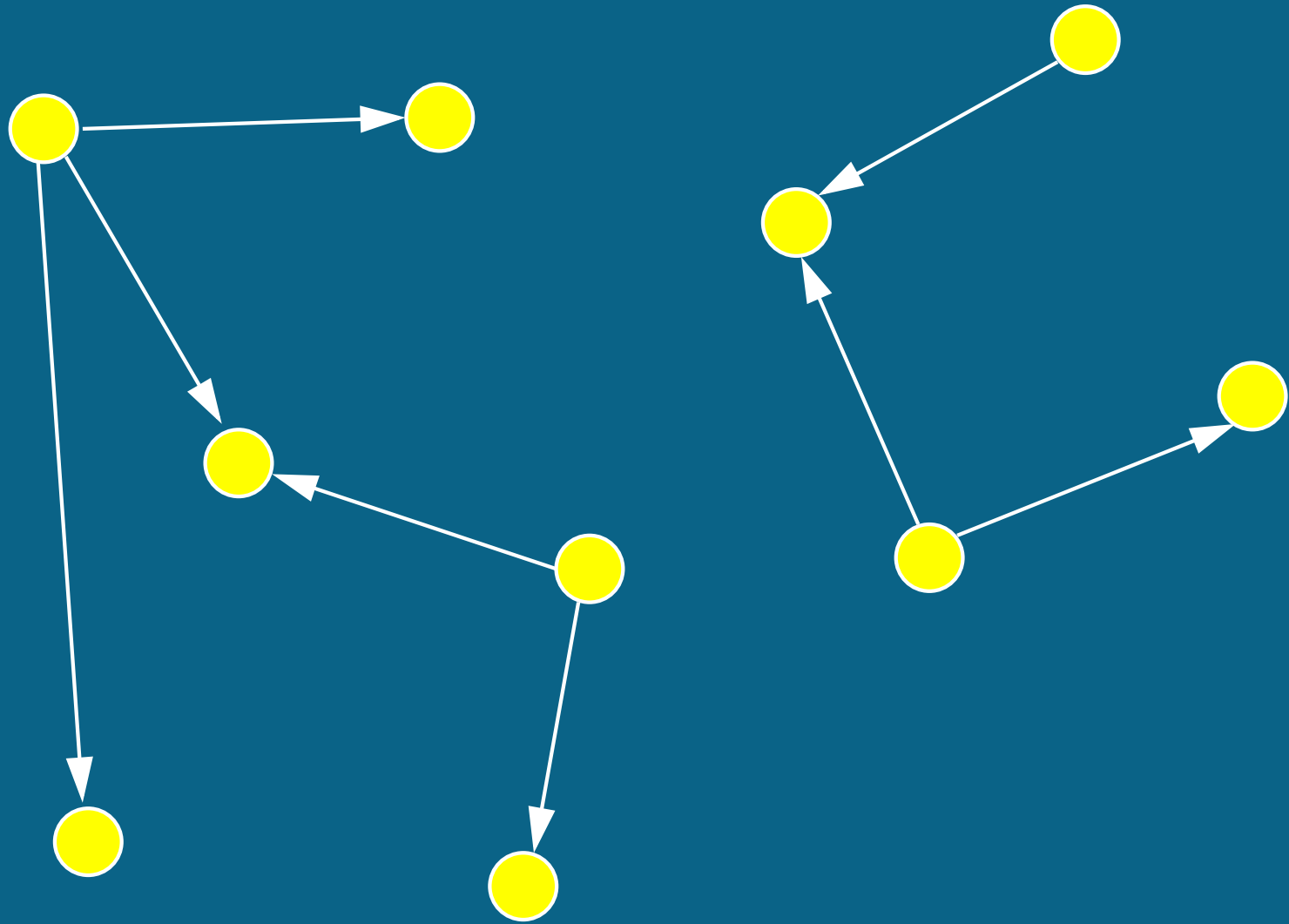
Low osmolarity response genes

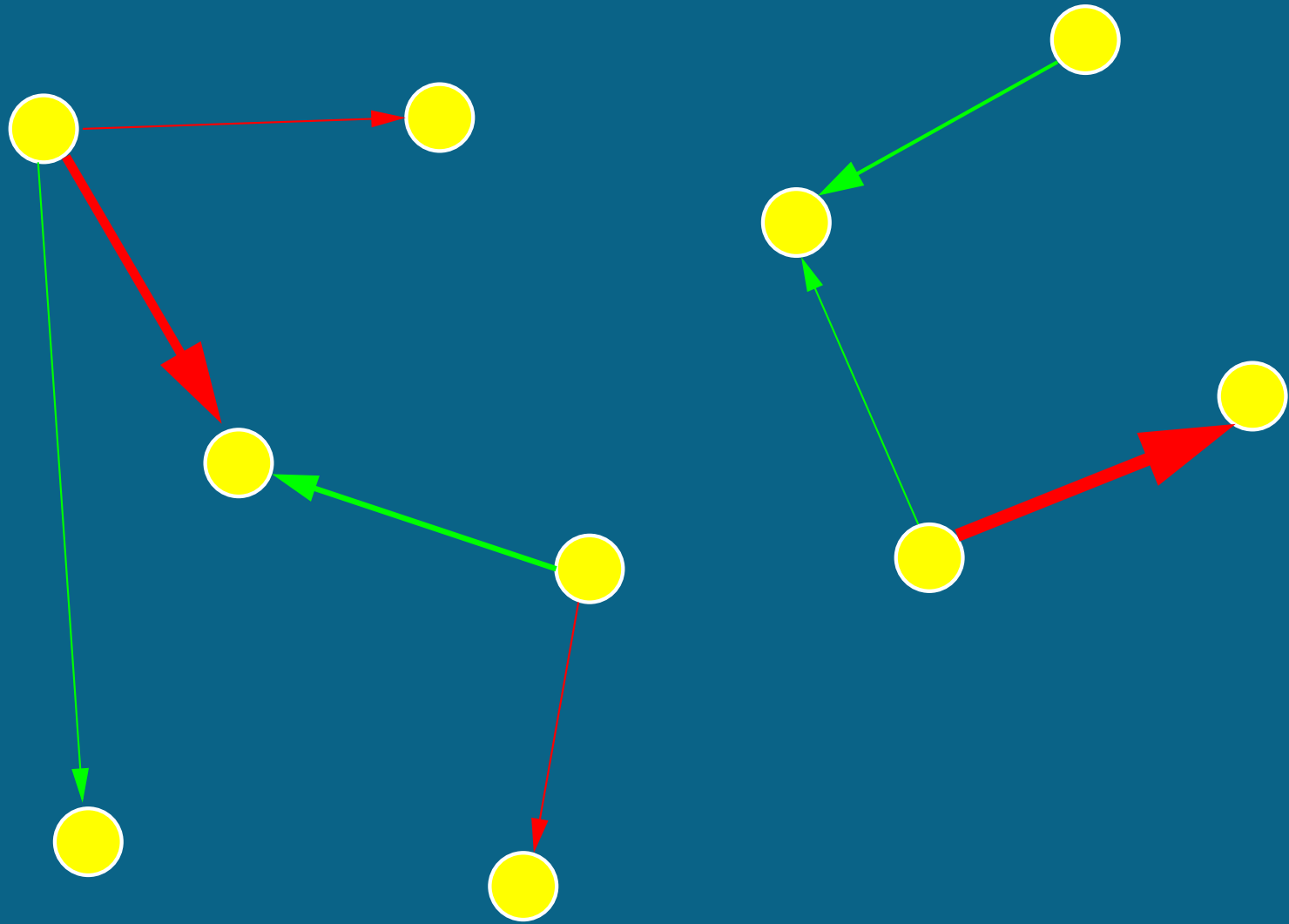


Outline of the talk

- Recapitulation: Bayesian networks
- **Reverse engineering:**
Learning networks from data
- Estimating the reliability of inference







Classical learning paradigm

Classical learning paradigm

Find the best network structure M :

$$M^* = \operatorname{argmax}\{P(M|D)\}$$

Classical learning paradigm

Find the **best network structure** M :

$$M^* = \operatorname{argmax}\{P(M|D)\}$$

Find the **best parameters** θ^*

$$\theta^* = \operatorname{argmax}\{P(\theta|D, M^*)\}$$

Find the best model M , that is, the best network

$$P(M|D) \propto P(D|M)P(M)$$

Find the best model M , that is, the best network

$$P(M|D) \propto P(D|M)P(M)$$

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

When is the integral **analytically tractable**?

Find the best model M , that is, the **best network**

$$P(M|D) \propto P(D|M)P(M)$$

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

When is the integral **analytically tractable**?

- Complete observation: **No missing values.**
- $P(D|\theta, M)$ and $P(\theta|M)$ must satisfy certain regularity conditions.
- Examples: **Multinomial** with a Dirichlet prior, **linear Gaussian** with a normal-gamma prior.

Find the best model M , that is, the **best network**

$$P(M|D) \propto P(D|M)P(M)$$

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

When is the integral **analytically tractable**?

- Complete observation: **No missing values.**
- $P(D|\theta, M)$ and $P(\theta|M)$ must satisfy certain regularity conditions.
- **Multinomial** distribution \longrightarrow **Discretization** of the data.

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	3.7×10^6	7.8×10^{11}	4.2×10^{18}

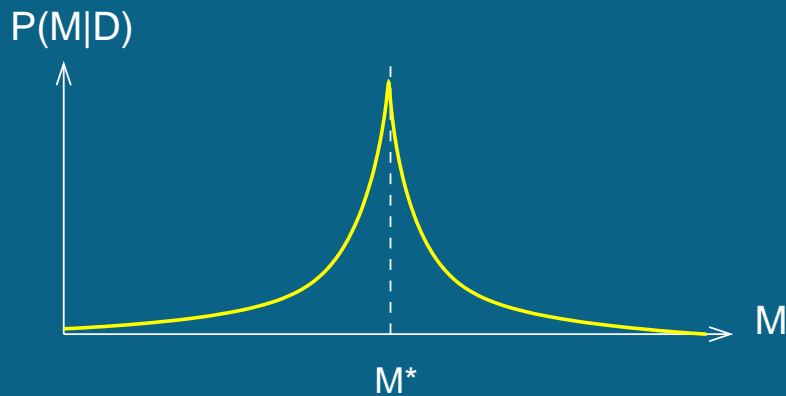
→ Optimization problem intractable for large N of nodes

Naive approach

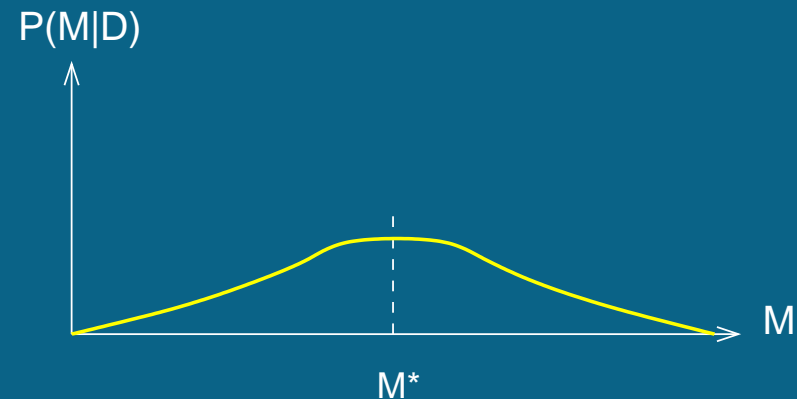
- Compute $P(M|D)$ for all possible network structures M .
- Select network structure M^* that maximizes $P(M|D)$

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^*

Objective: Sample from the posterior distribution

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_i P(D|M_i)P(M_i)}$$

Direct approach intractable due to $\sum_i P(D|M_i)P(M_i)$

Objective: Sample from the posterior distribution

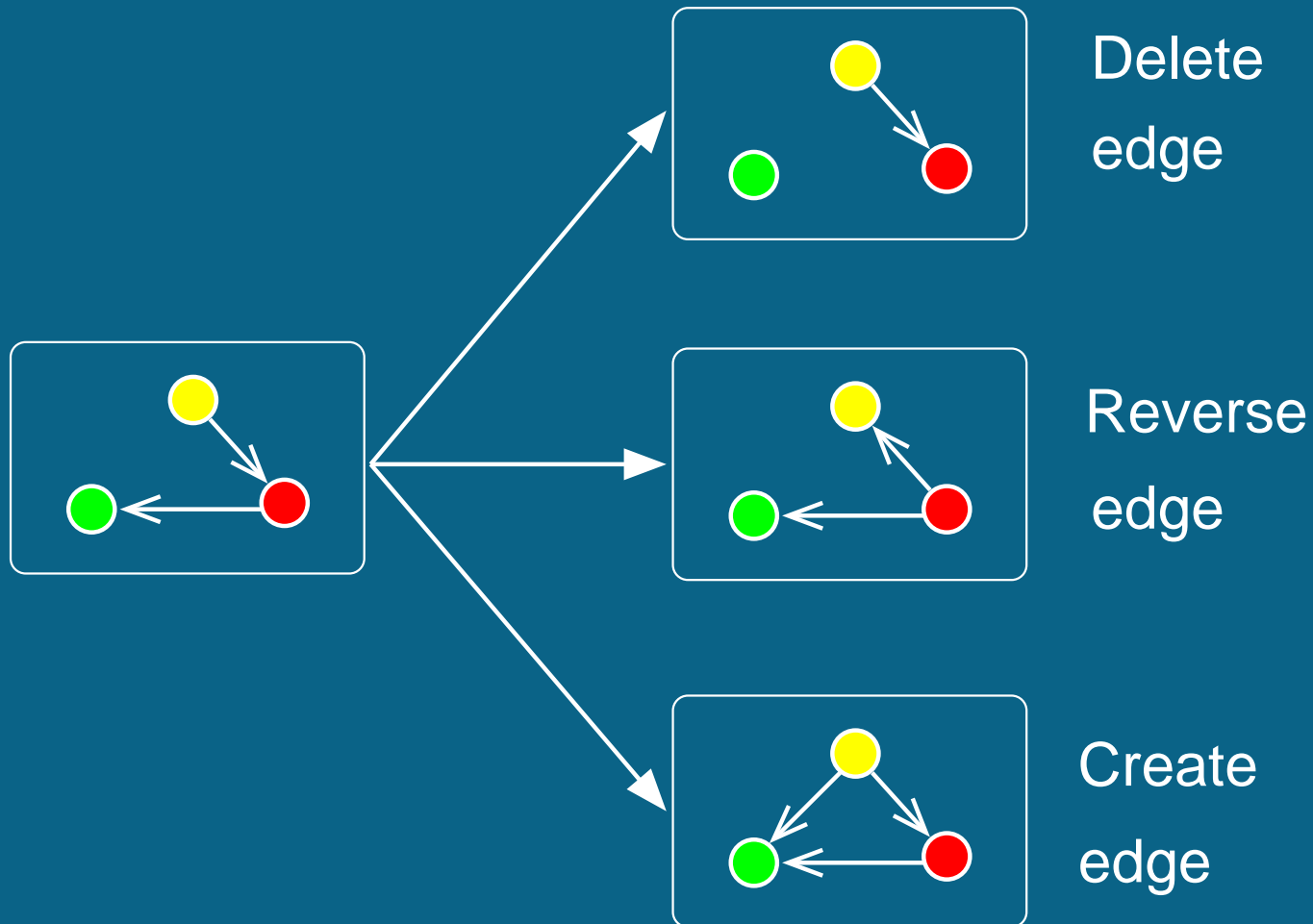
$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_i P(D|M_i)P(M_i)}$$

Direct approach intractable due to $\sum_i P(D|M_i)P(M_i)$

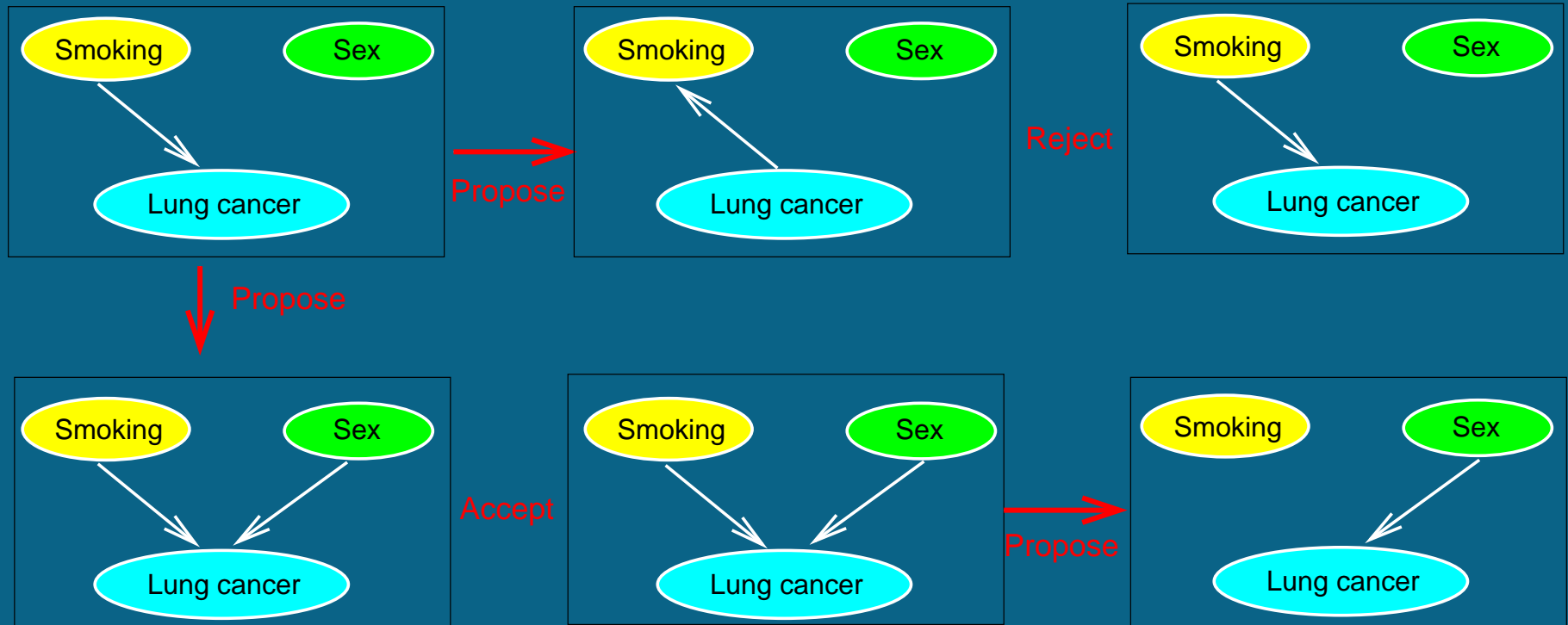
Markov chain Monte Carlo (MCMC):

- **Proposal move:** Given network M_{old} , propose a new network M_{new} with probability $Q(M_{new}|M_{old})$.
- **Acceptance/Rejection:** Accept this new network with probability $\min \left\{ 1, \frac{P(D|M_{new})P(M_{new})}{P(D|M_{old})P(M_{old})} \times \frac{Q(M_{old}|M_{new})}{Q(M_{new}|M_{old})} \right\}$

MCMC moves



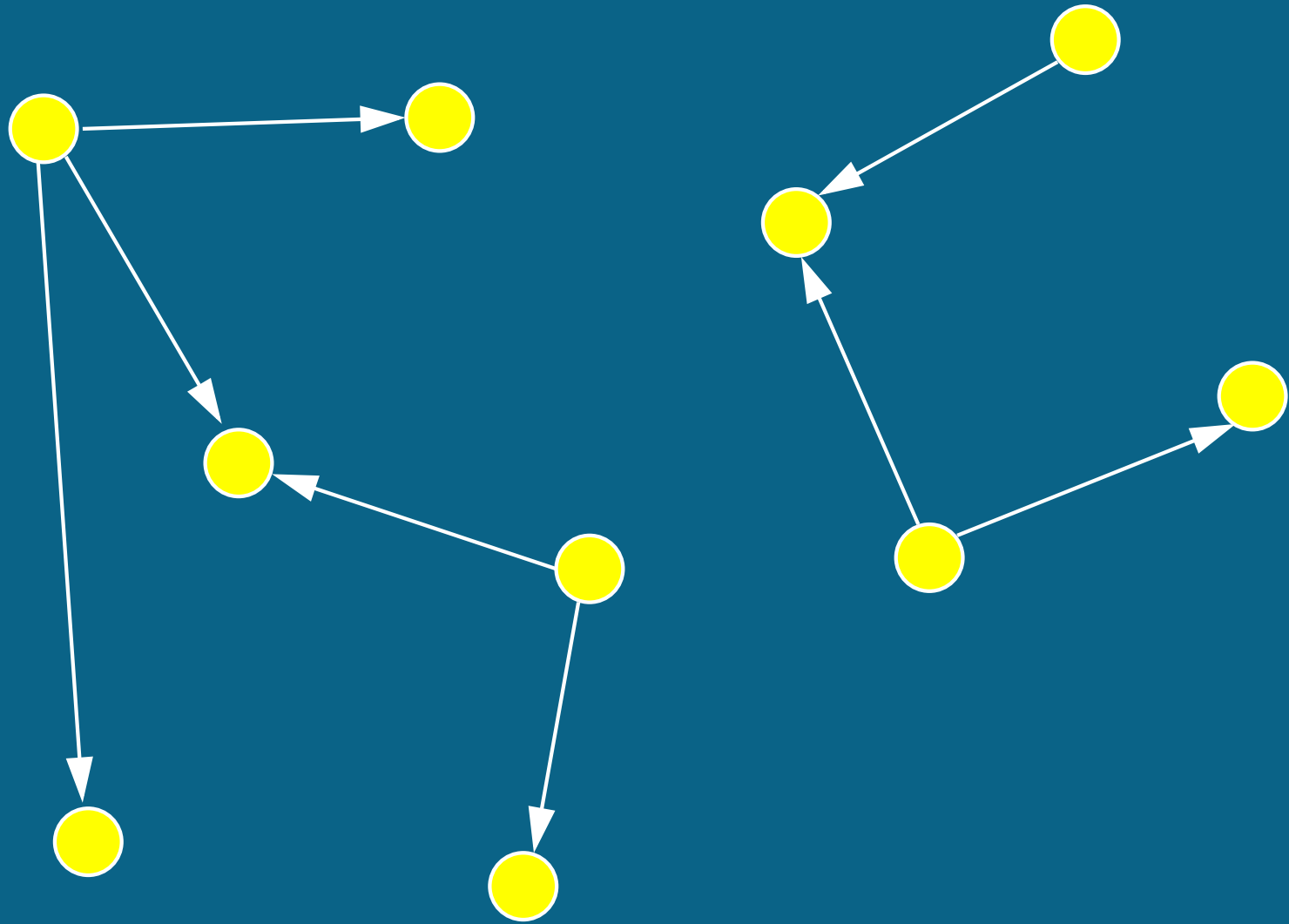
Markov chain Monte Carlo (MCMC)

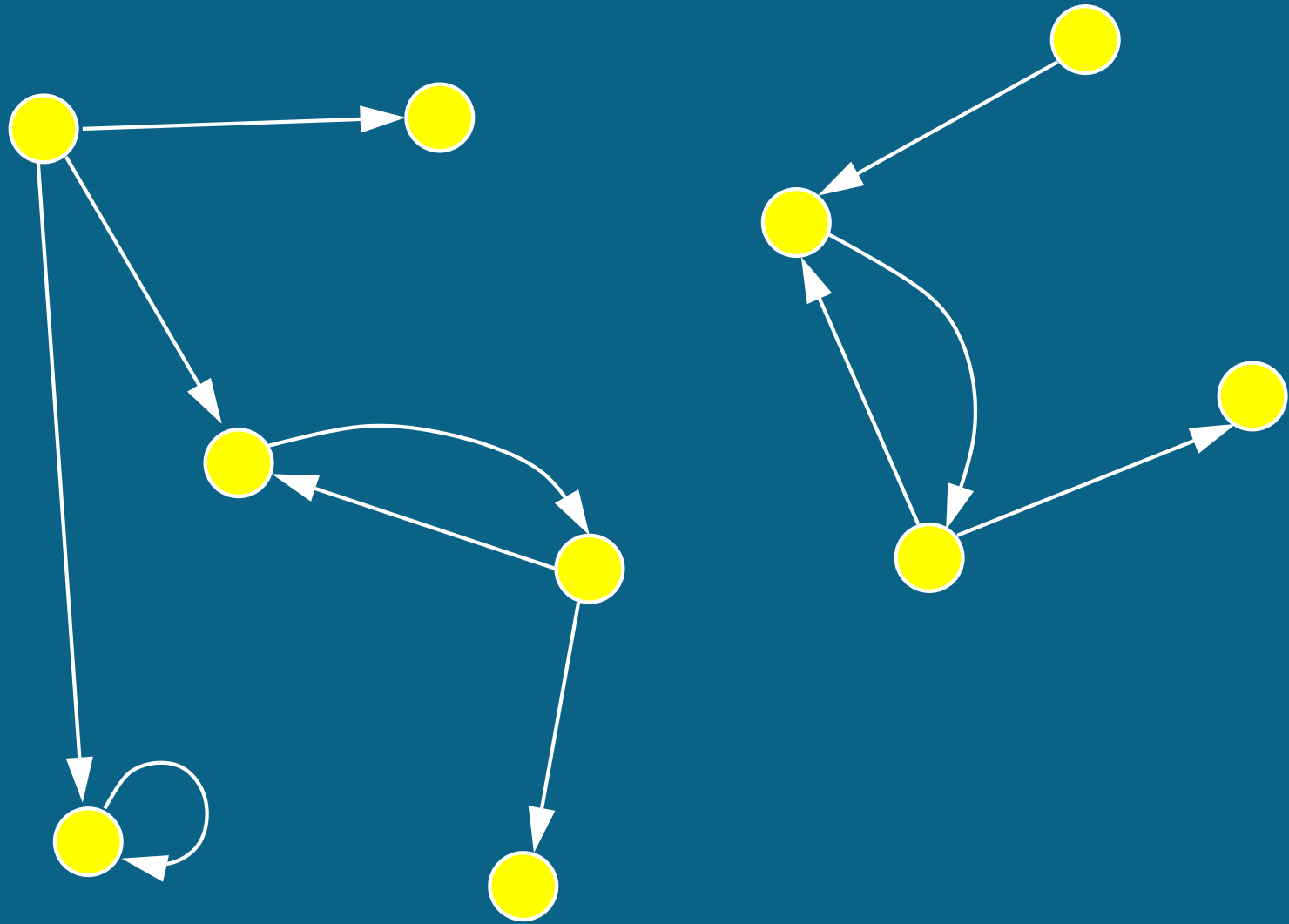


Accept move with probability: $\min \left\{ 1, \frac{P(M_{new}|D)}{P(M_{old}|D)} \times \frac{Q(M_{old}|M_{new})}{Q(M_{new}|M_{old})} \right\}$

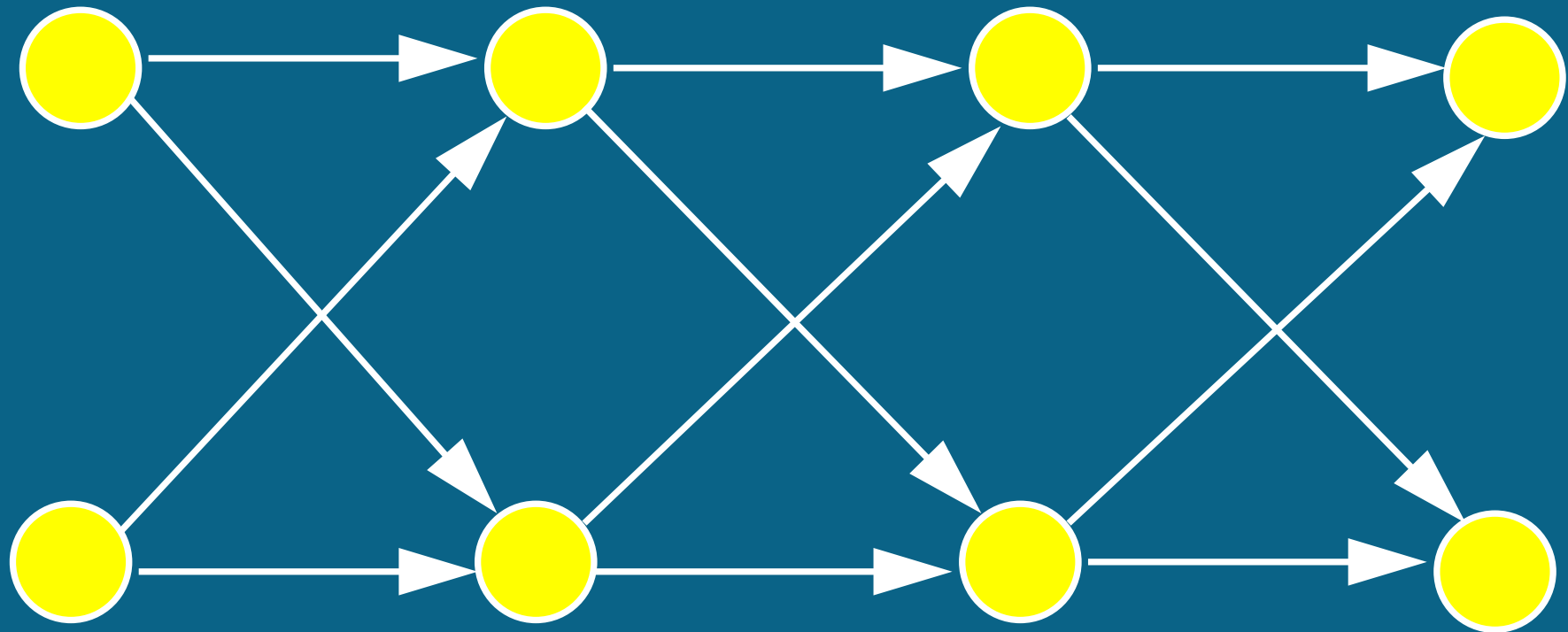
Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering:
Learning networks from data
- **Dynamic Bayesian networks**
- Estimating the reliability of inference









t=1

t=2

t=3

t=4

Outline of the talk

- Recapitulation: Bayesian networks
- Reverse engineering:
Learning networks from data
- **Estimating the reliability of inference**

Objective:

Reverse engineering → Learn networks from data

Problem:

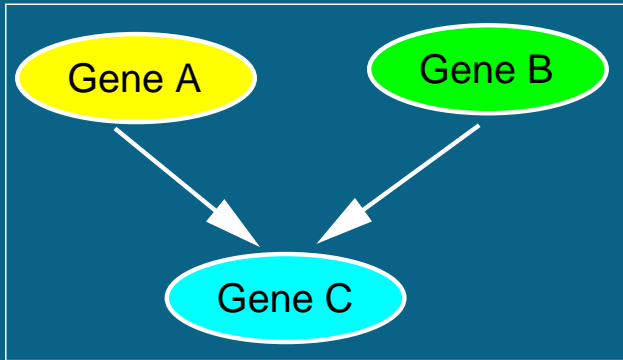
Data are sparse → Intrinsic uncertainty of inference

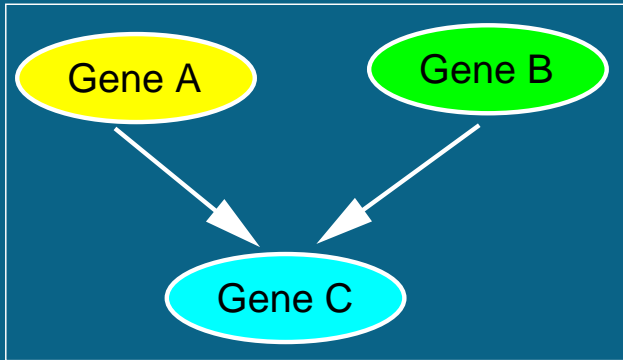
Estimate the reliability of inference

- Synthetic data
- Real data
- Realistic simulation

Estimate the reliability of inference

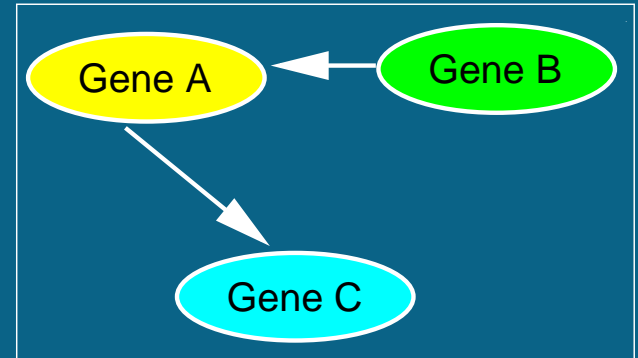
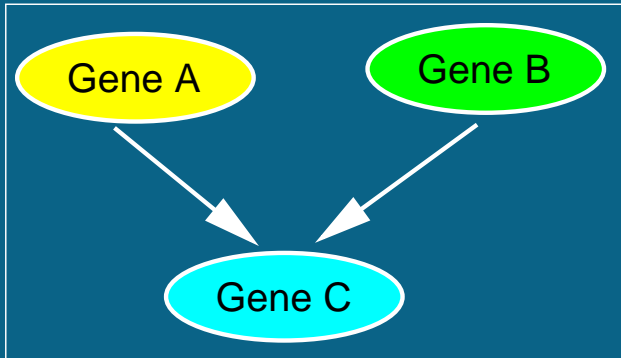
- **Synthetic data**
- Real data
- Realistic simulation





generate

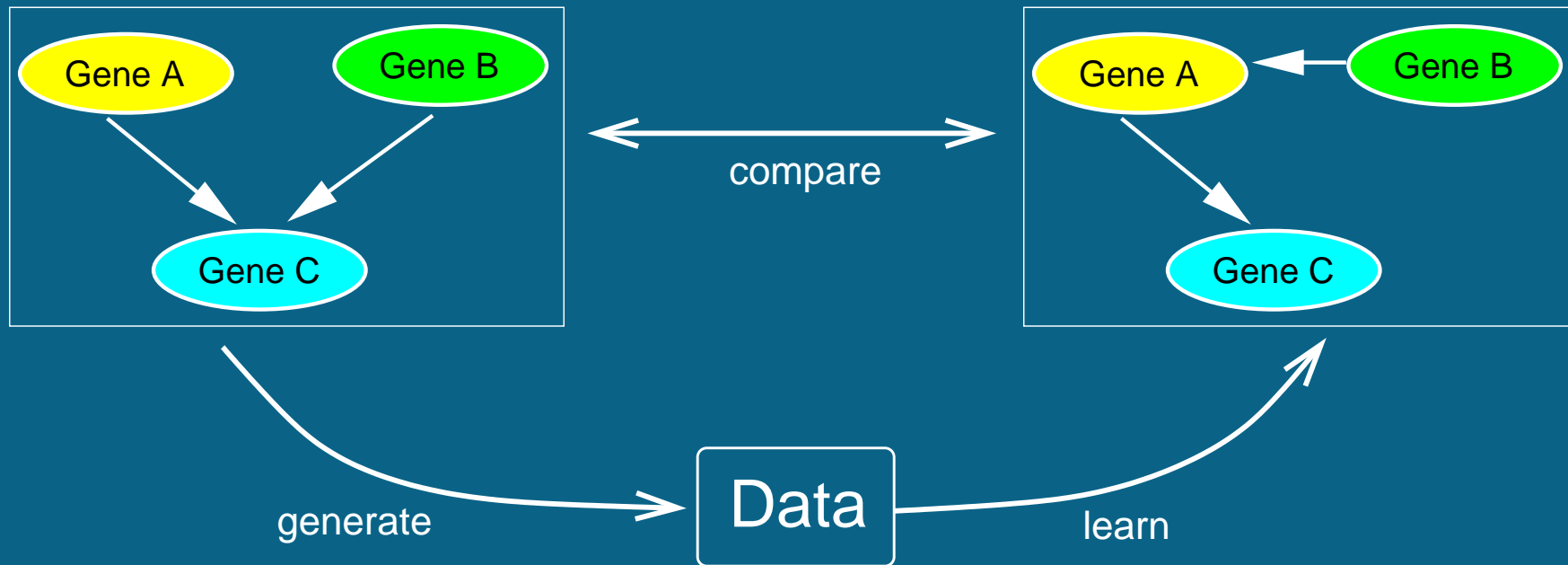


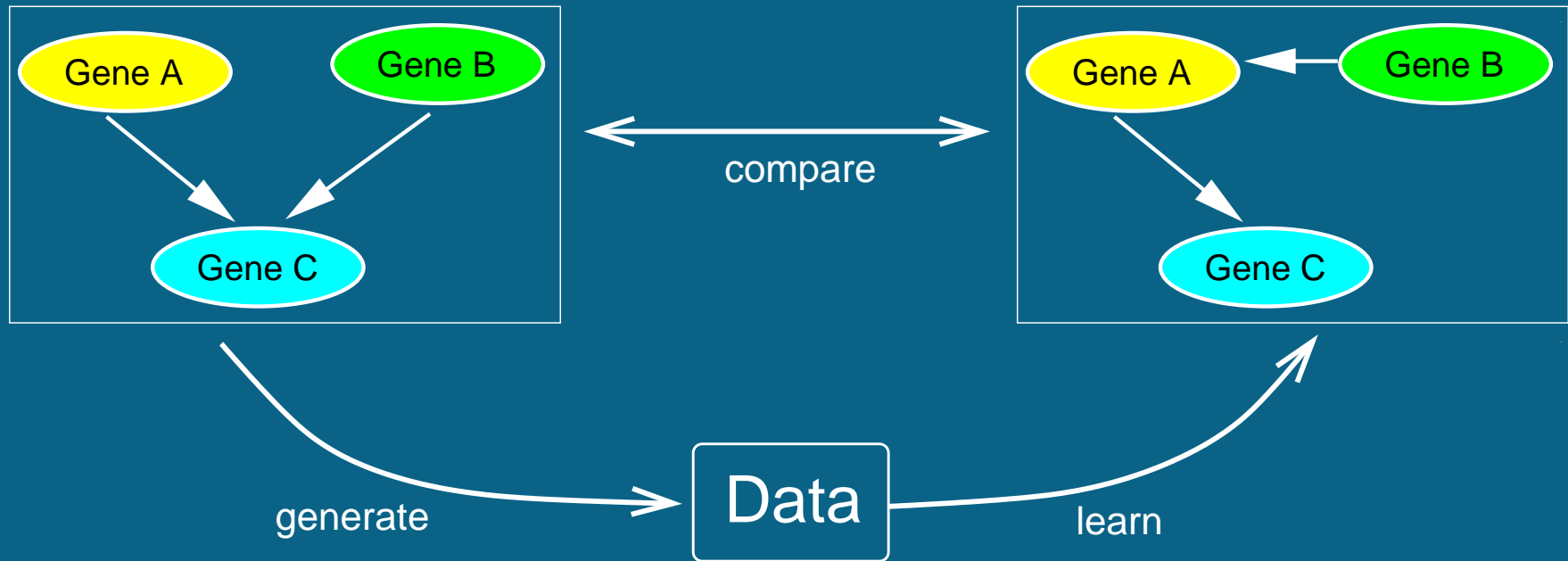


generate

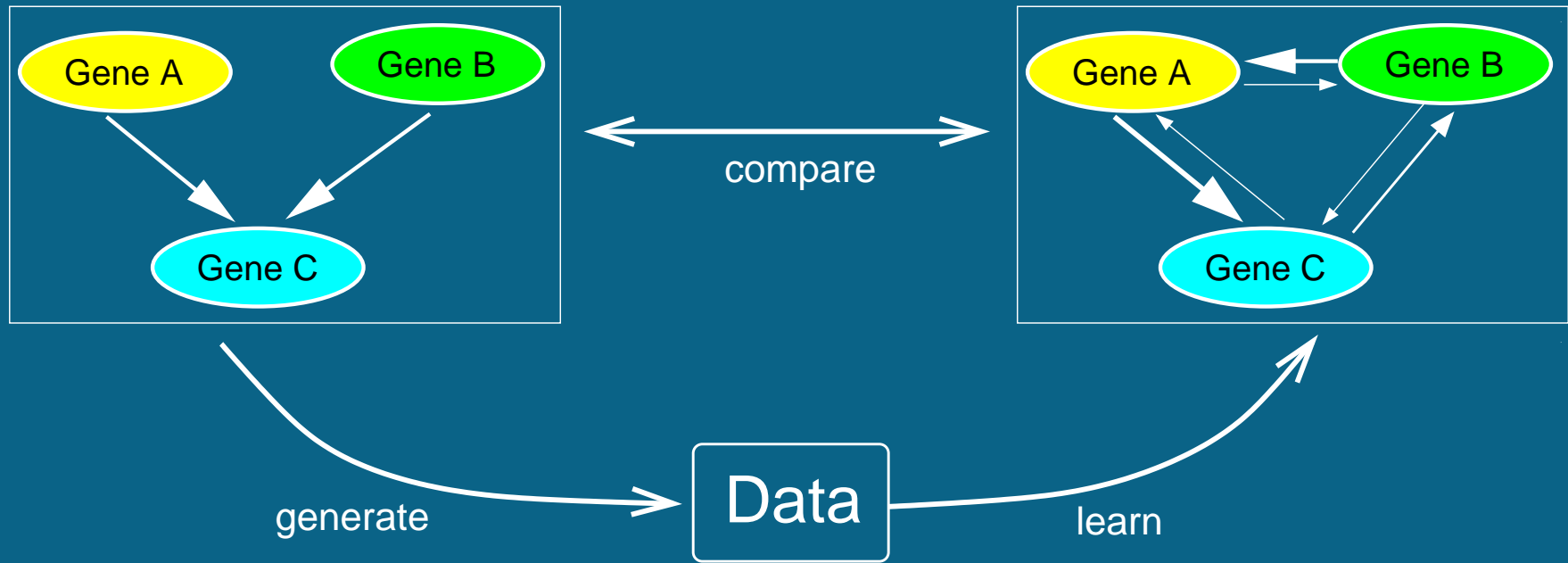
Data

learn

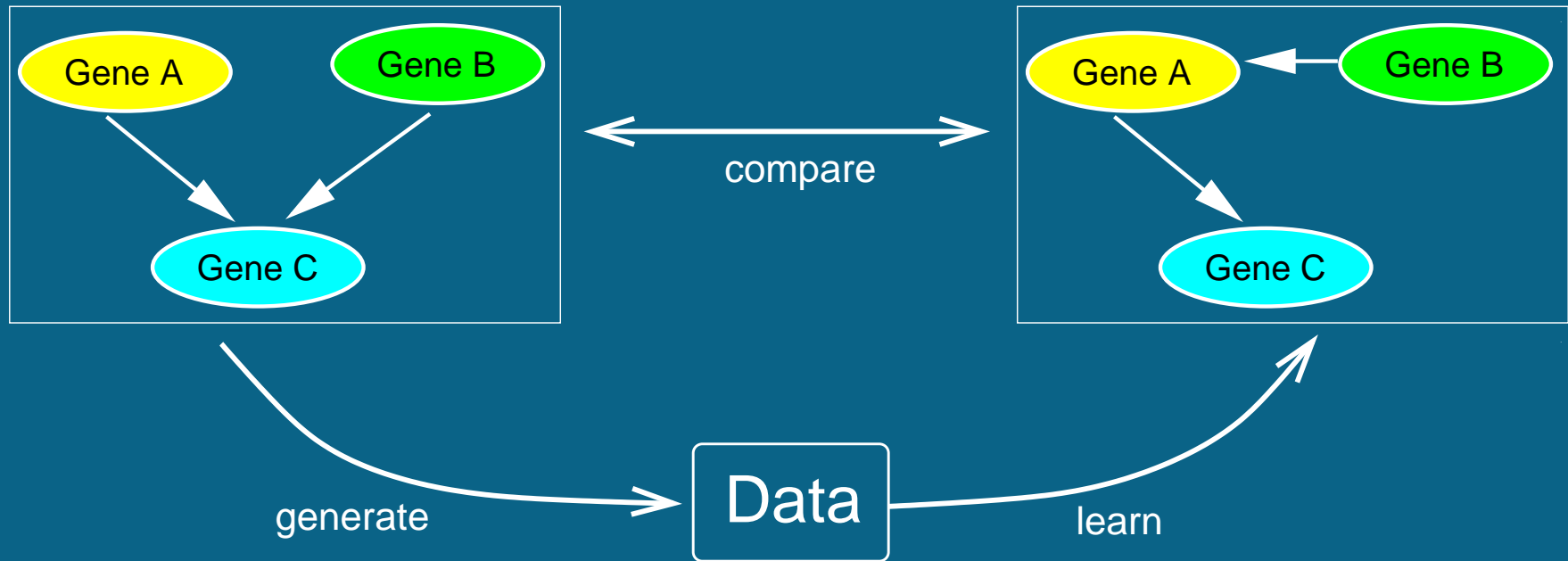




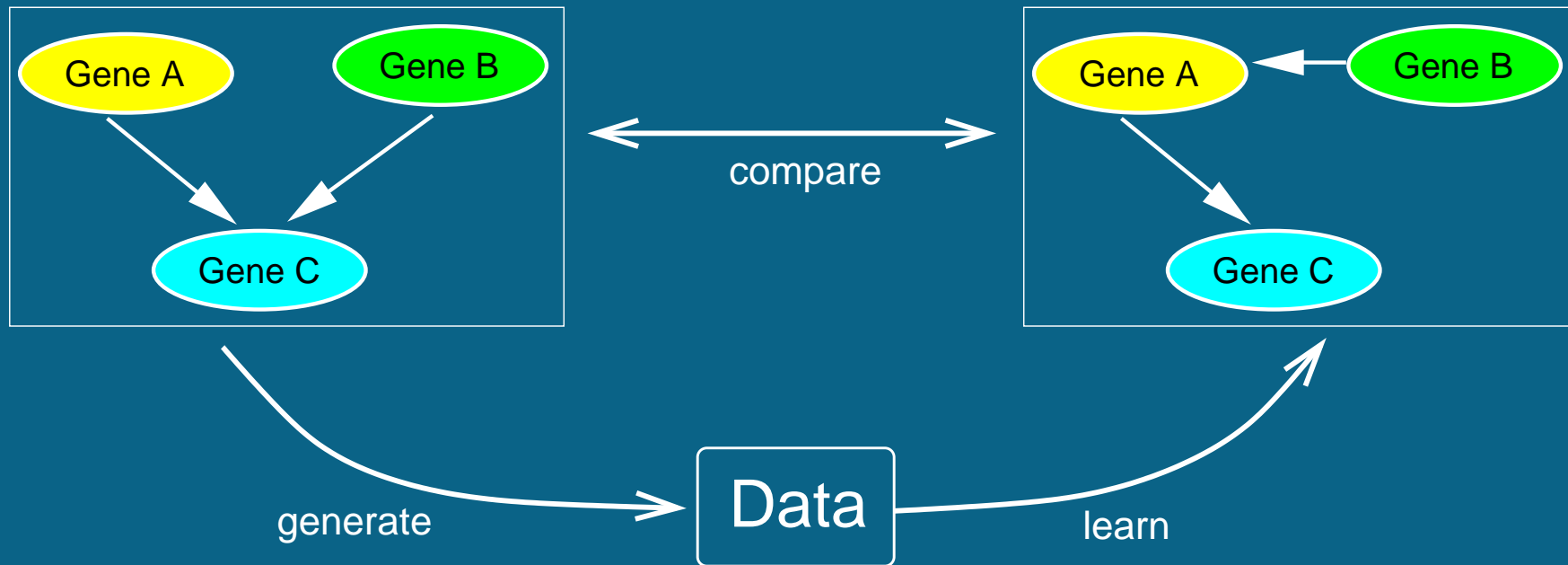
Deterministic inference



Probabilistic inference



Thresholding

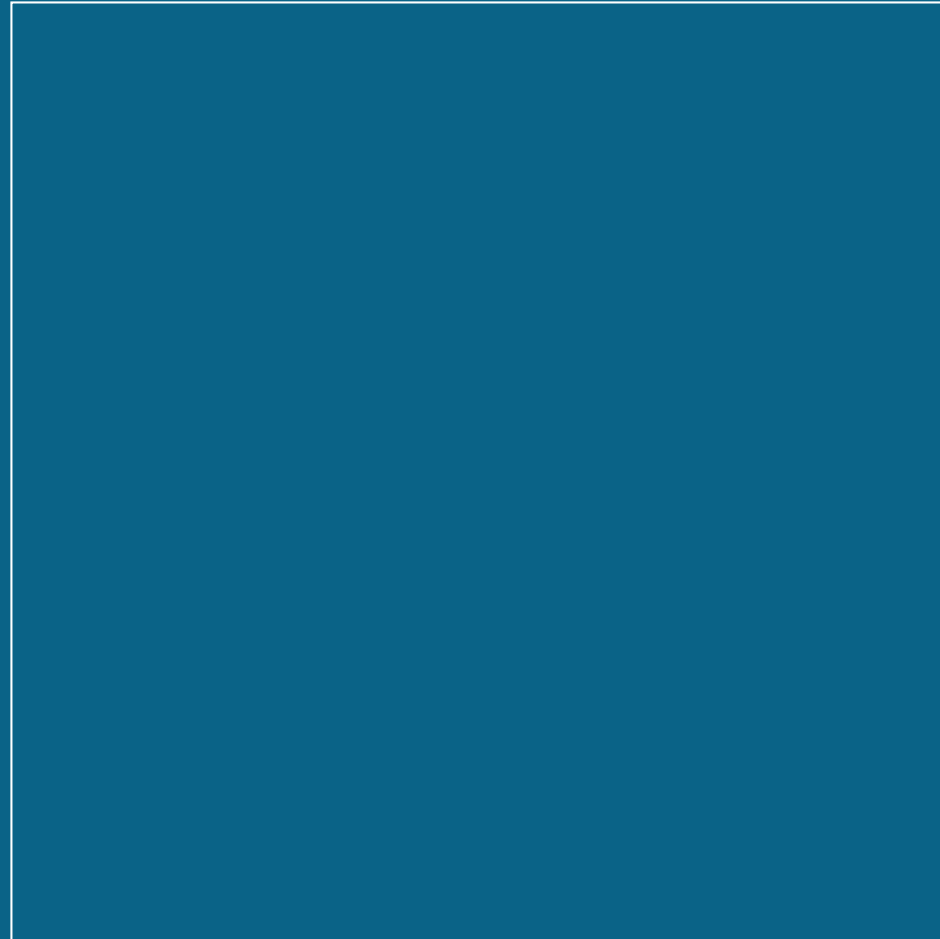


Thresholding

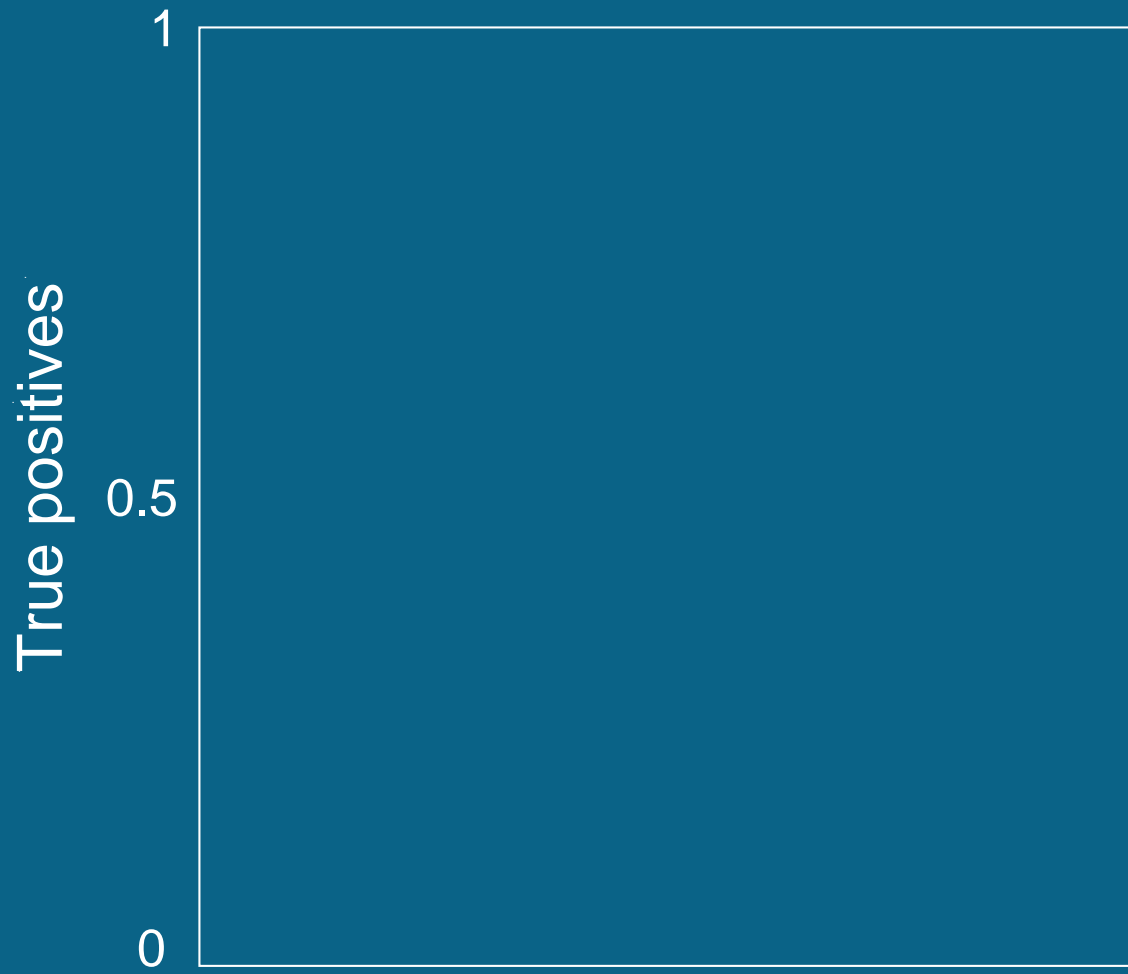
True positives

False positives

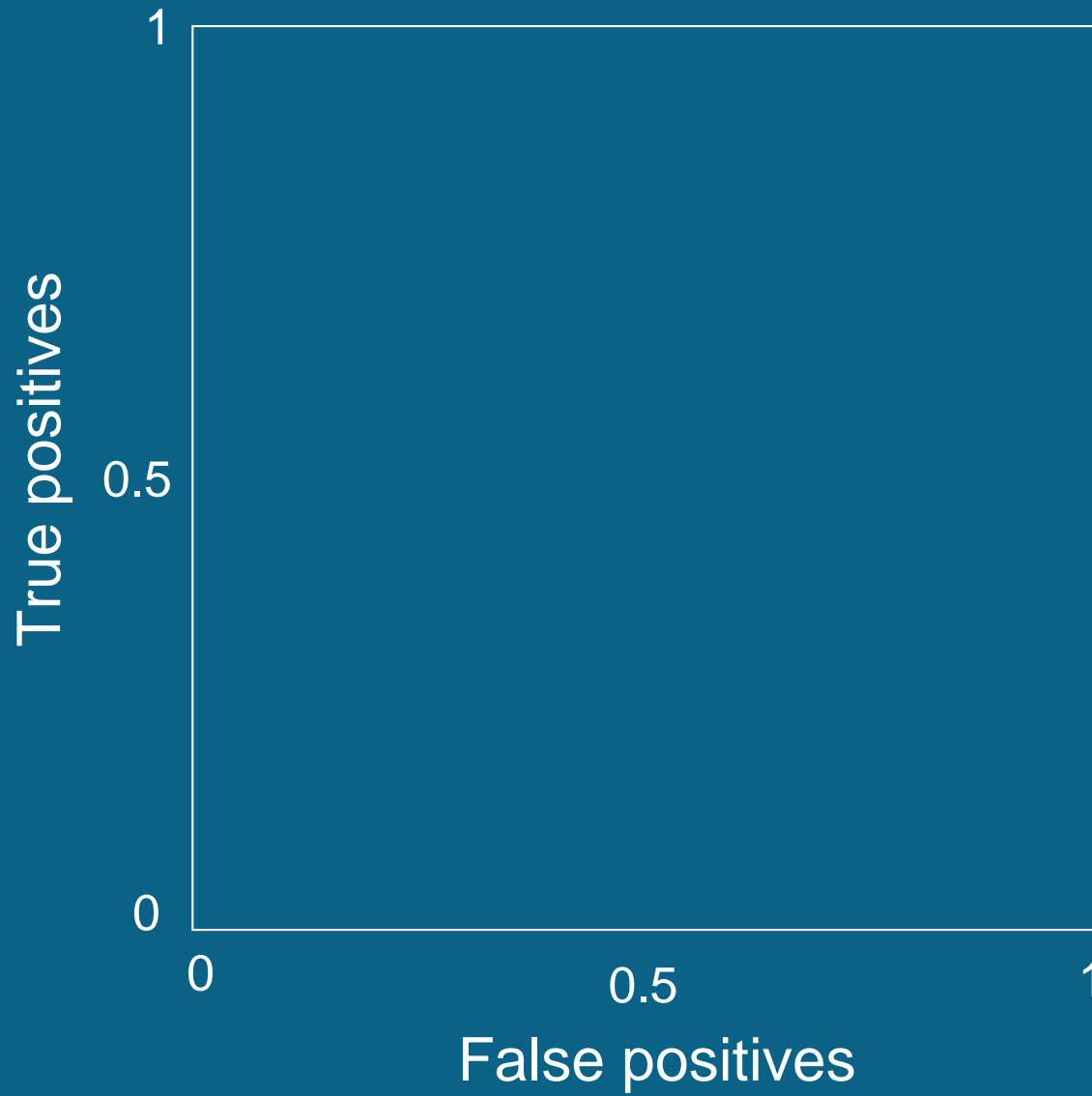
ROC curve



ROC curve



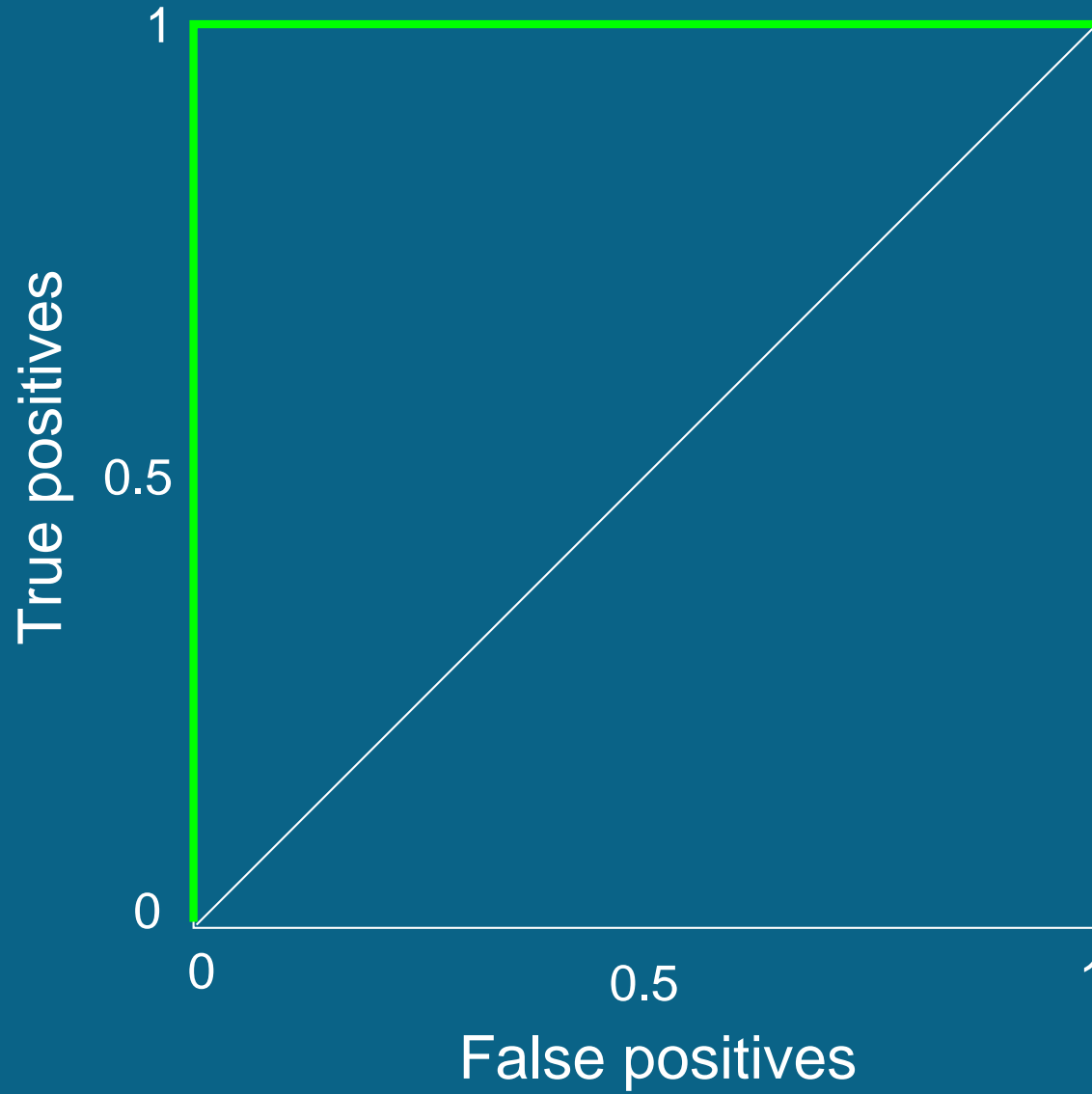
ROC curve



Random predictor



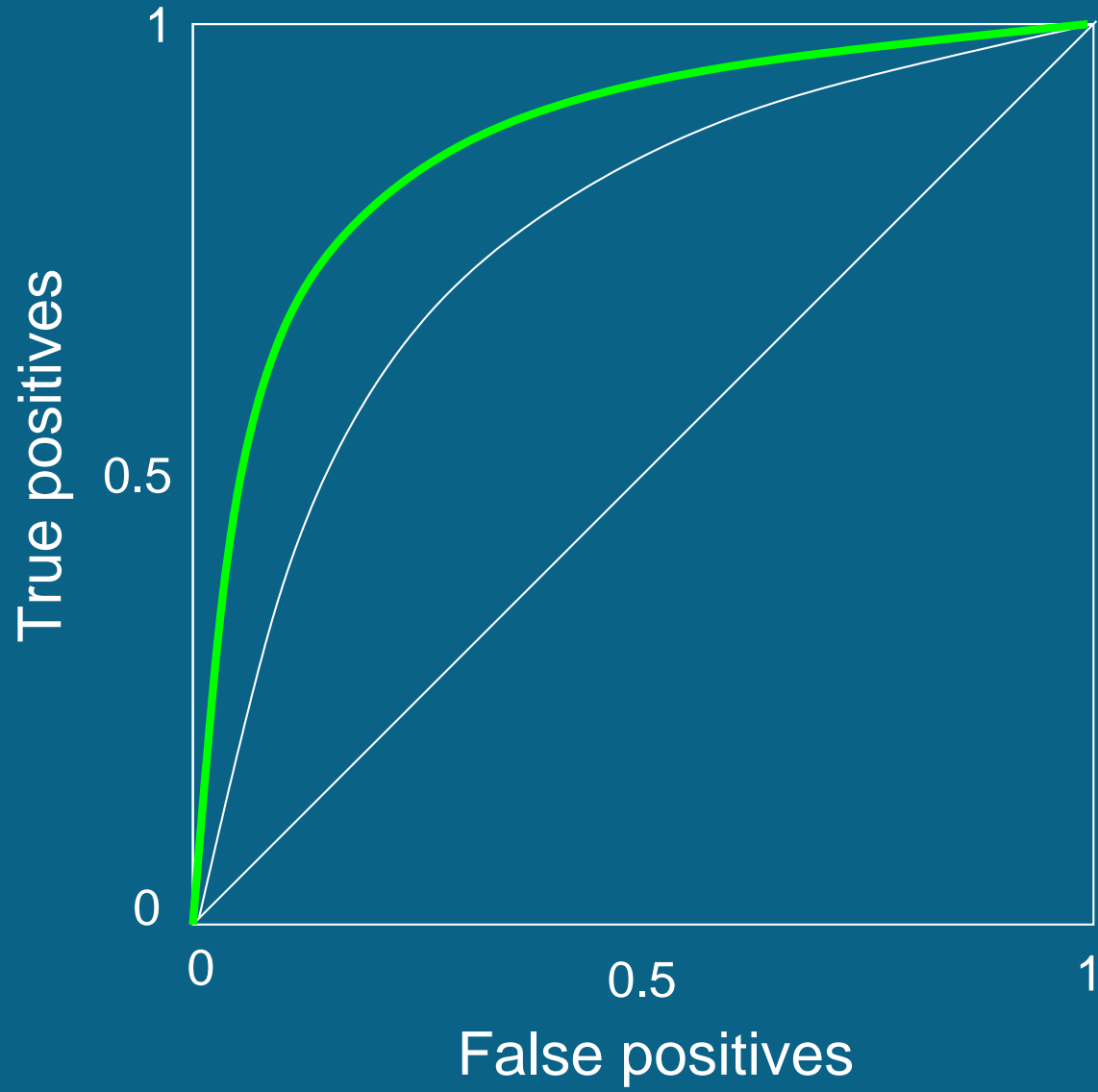
Perfect predictor



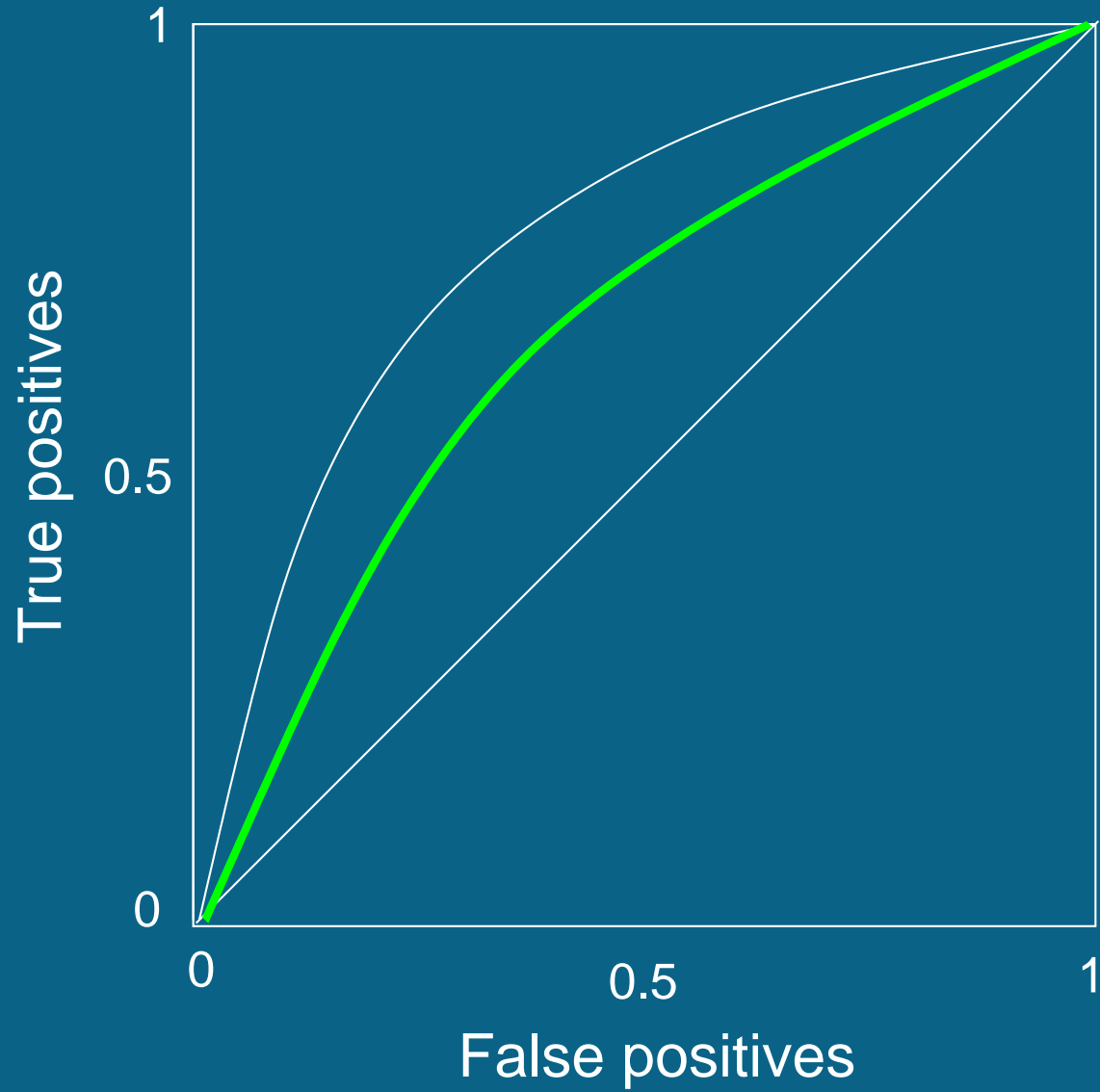
Realistic predictor



Better predictor

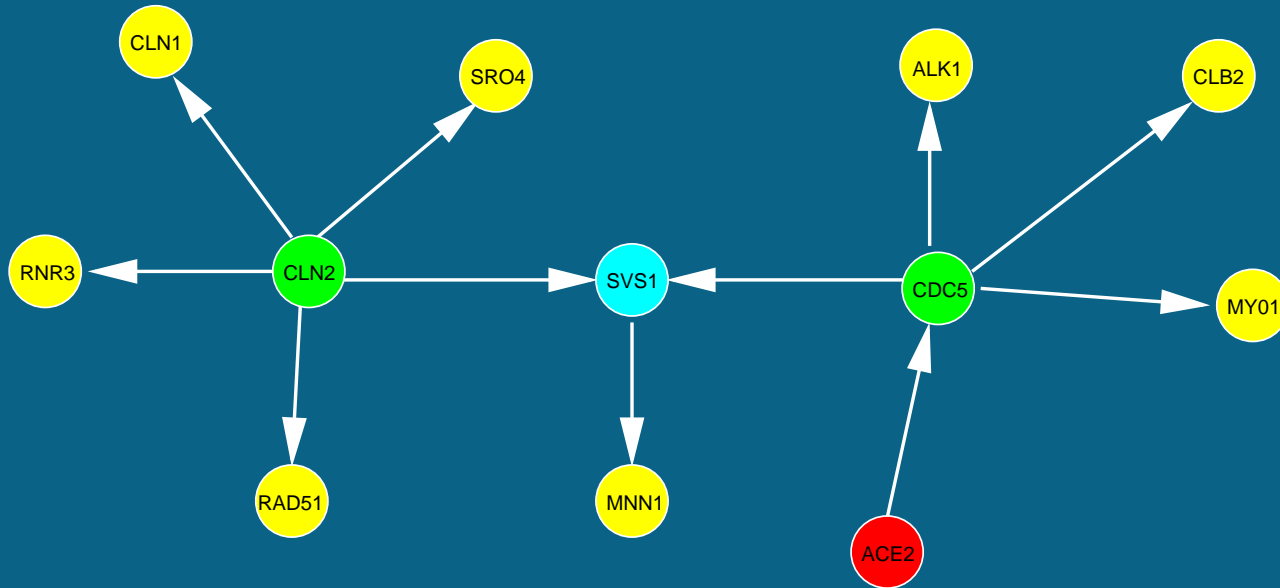


Poorer predictor



Data: binary

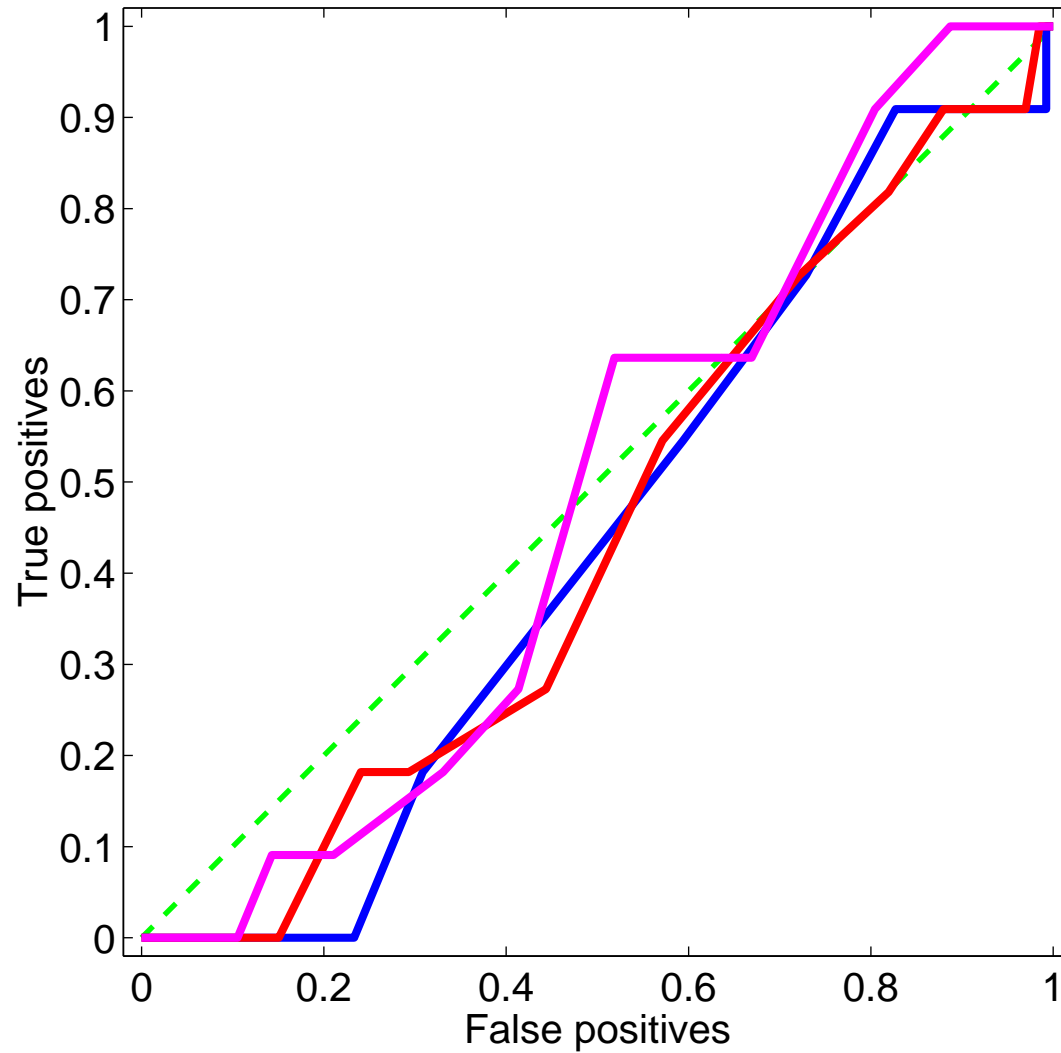
Model:



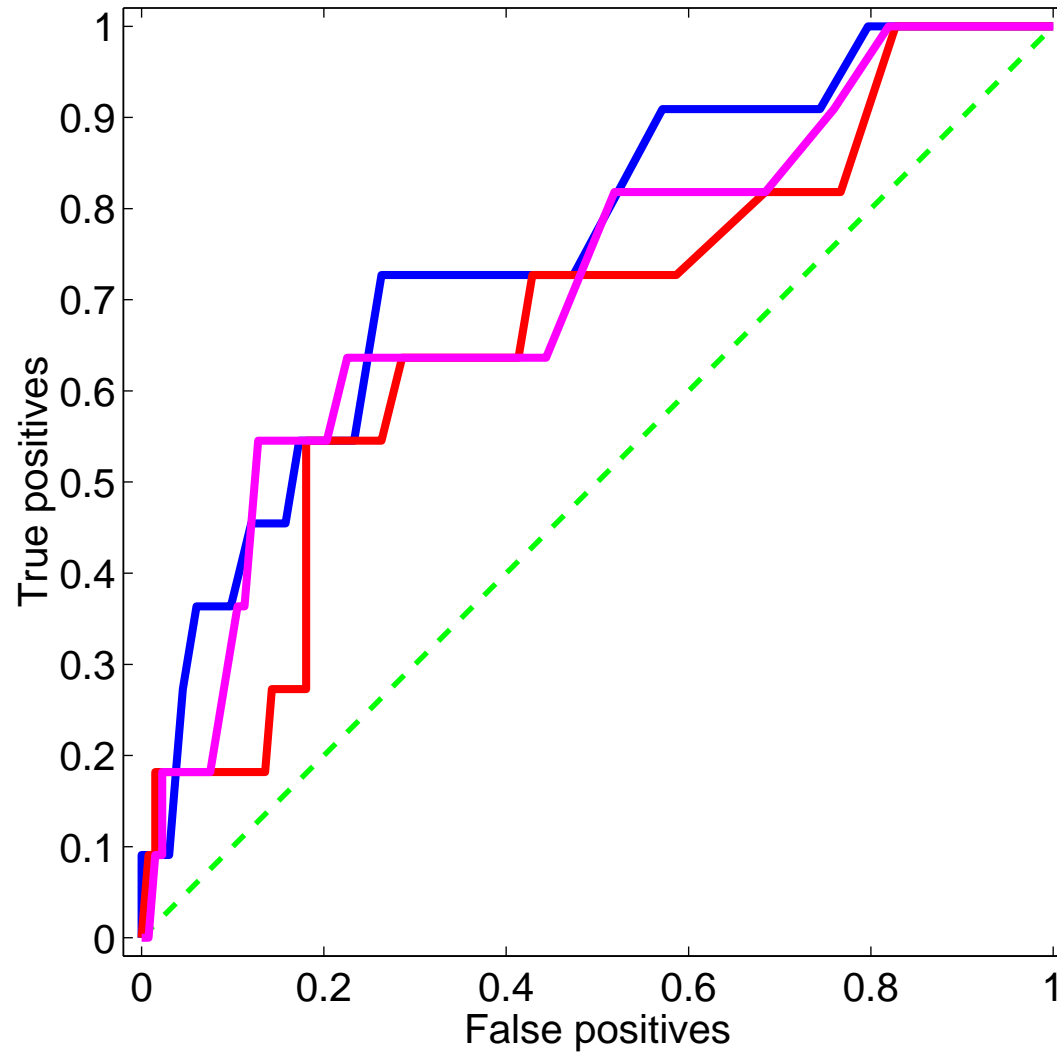
Parameters:

Noisy boolean: $P \in \{0.1, 0.9\}$

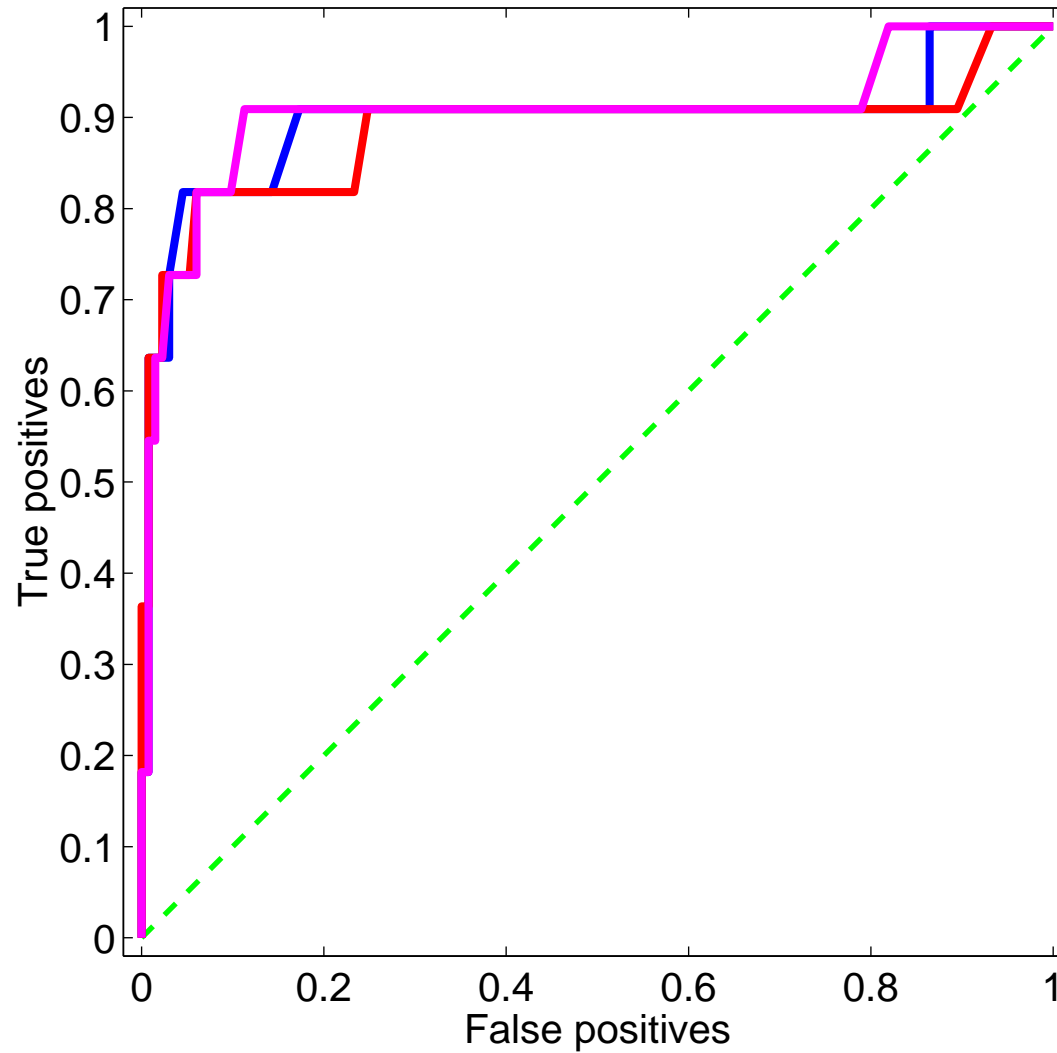
ROC curve: Sample size= 3



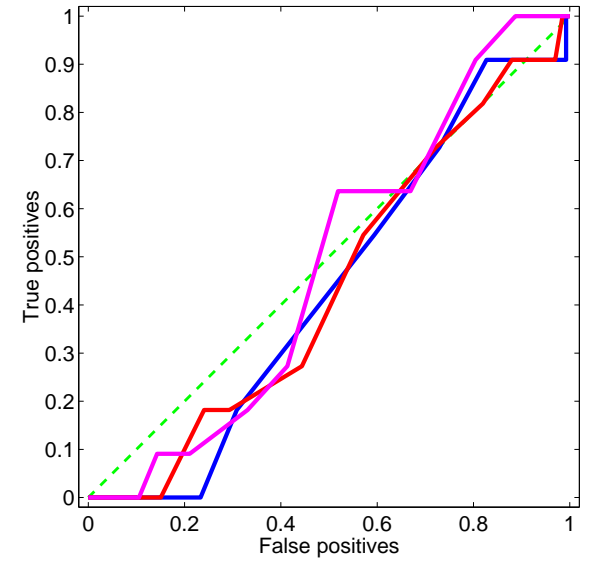
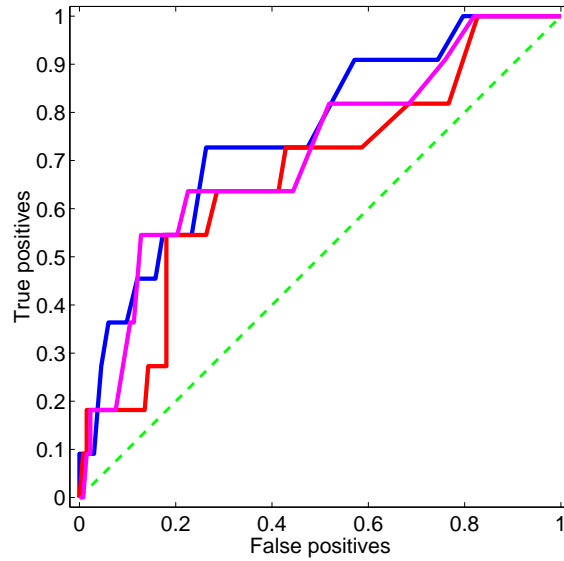
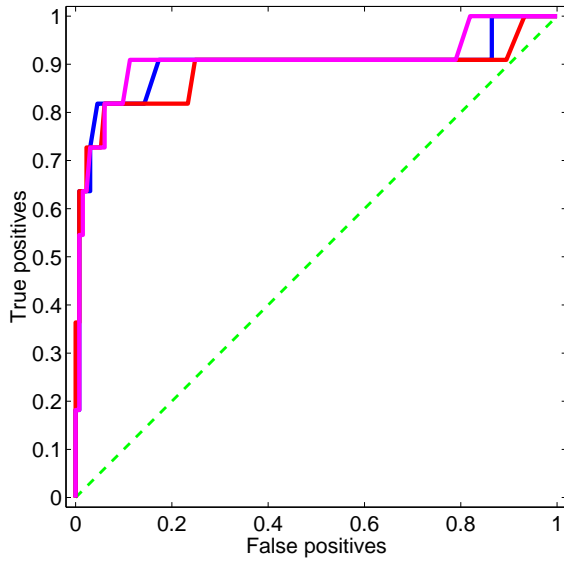
ROC curve: Sample size= 6



ROC curve: Sample size= 12

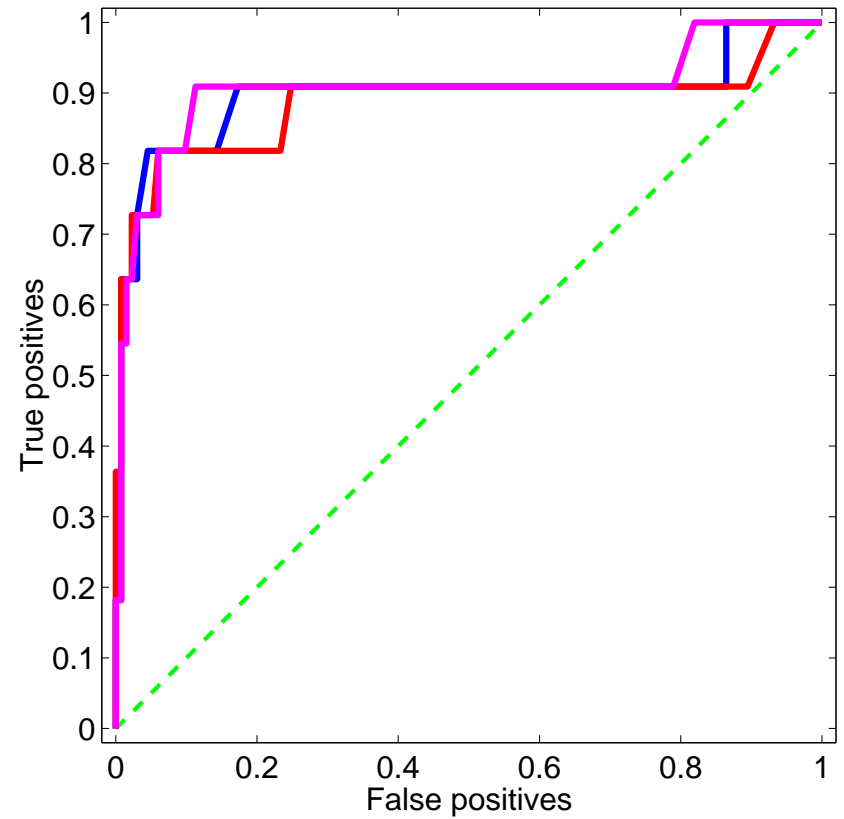
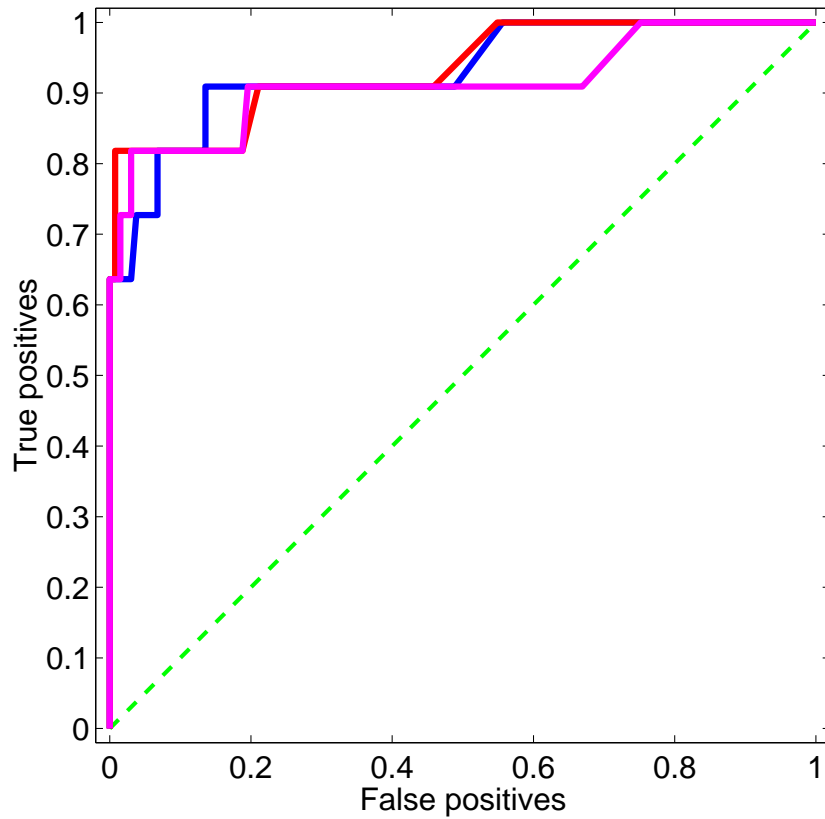


Training set size= **12** (left), **6** (middle), **3** (right)



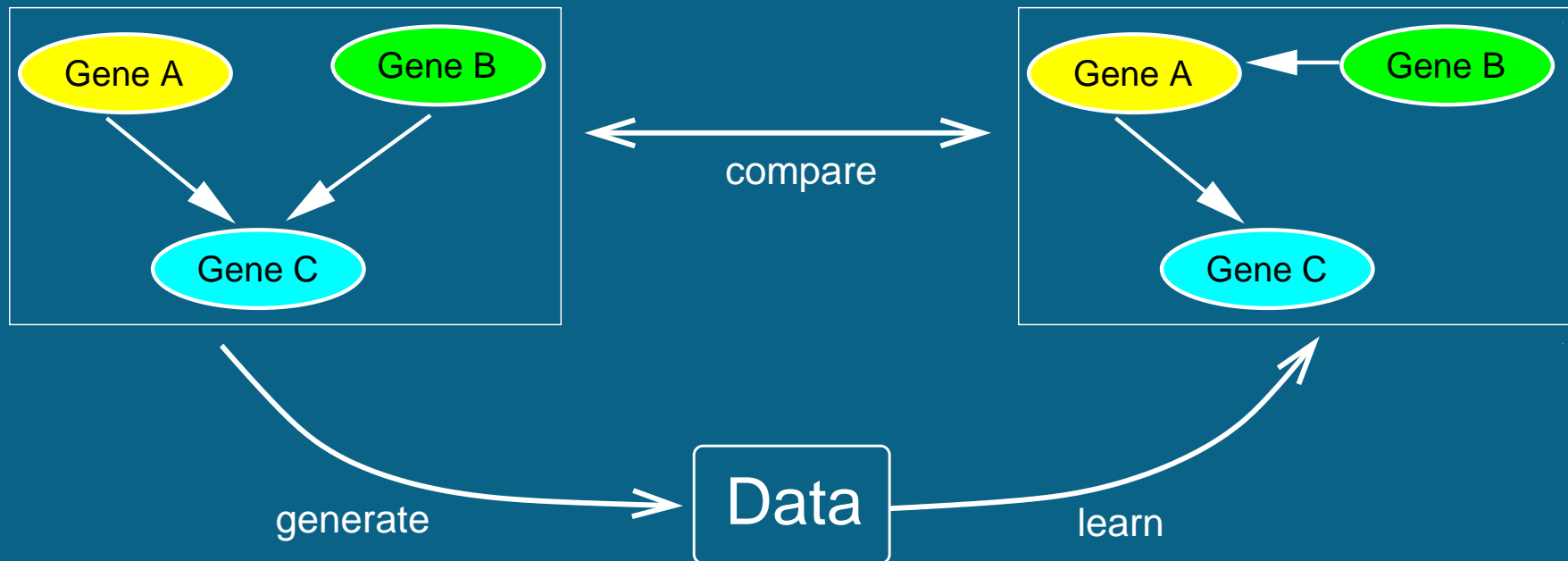
Max fan-in= 2 (left) versus 4 (right)

Training set size = 12



Disadvantage:

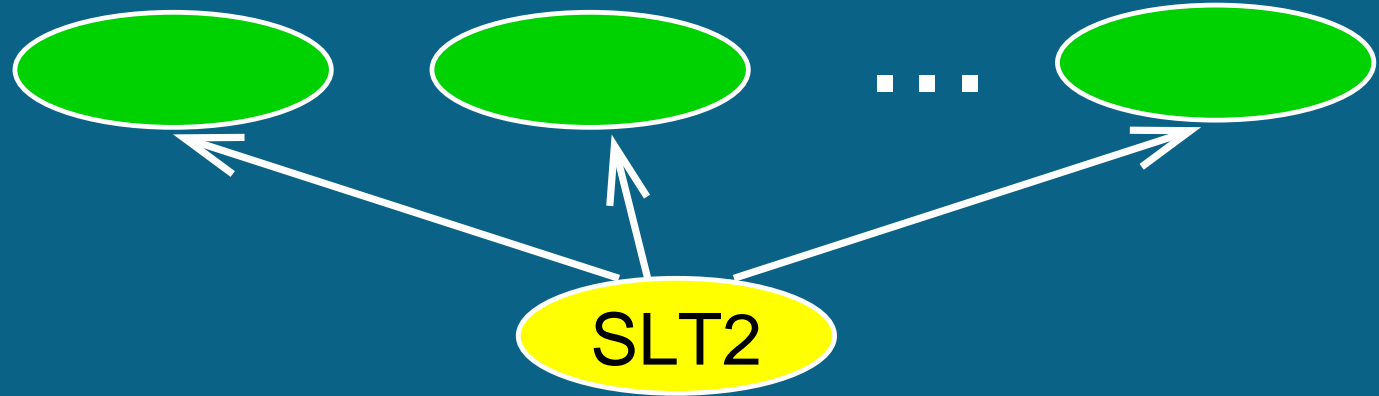
Unrealistic, **no mismatch** between the model used for **data generation** and the model used for **inference**.



Estimate the reliability of inference

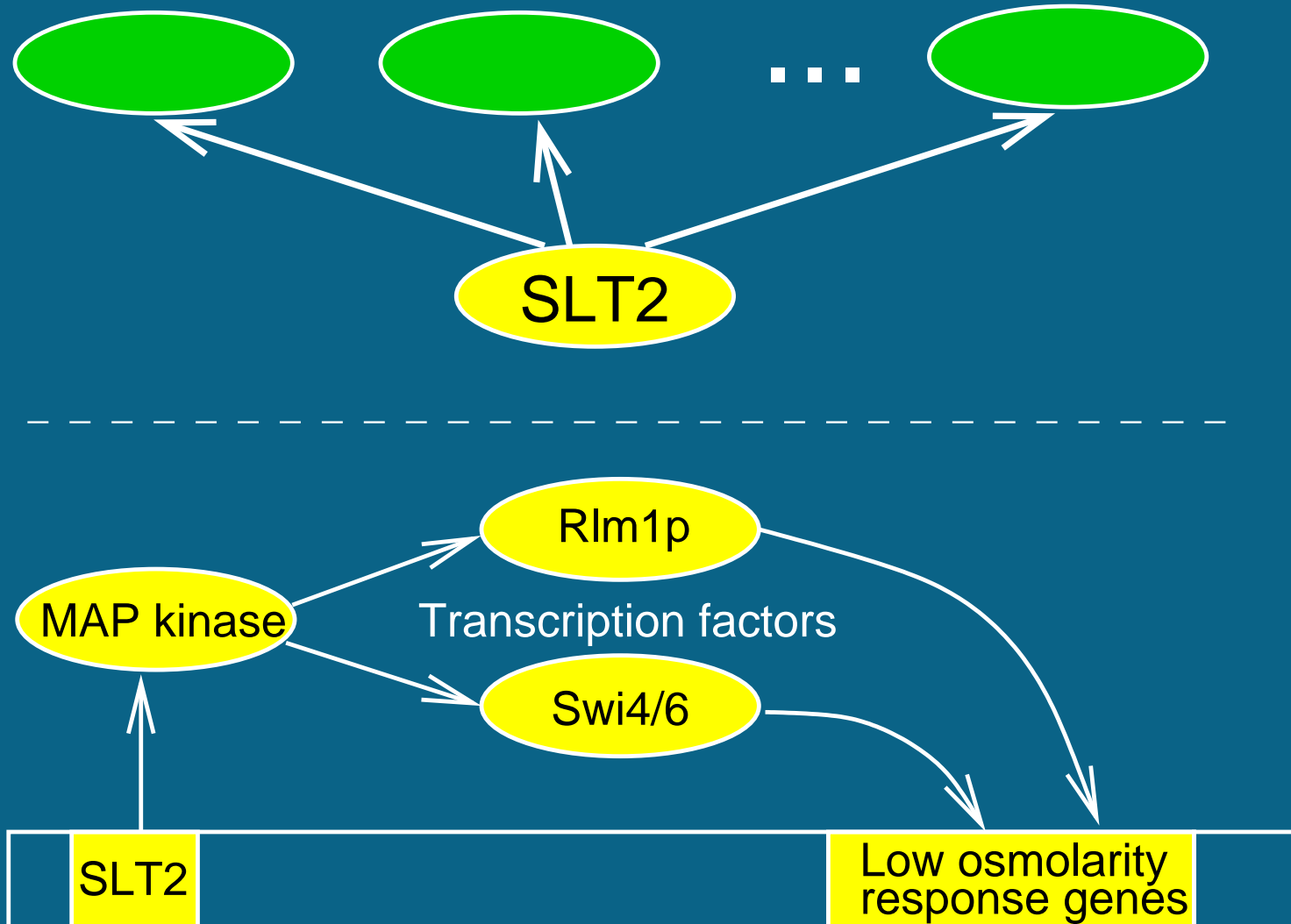
- Synthetic data
- **Real data**
- Realistic simulation

Low osmolarity response genes



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

Low osmolarity response genes



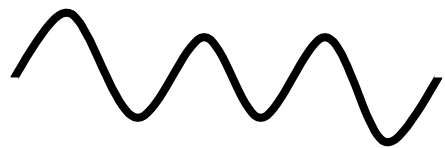
Disadvantage:

No **gold standards** !

It may be possible to estimate the **sensitivity**,
but not the **specificity**.

Estimate the reliability of inference

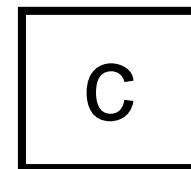
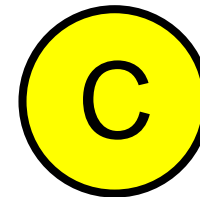
- Synthetic data
- Real data
- **Realistic simulation**



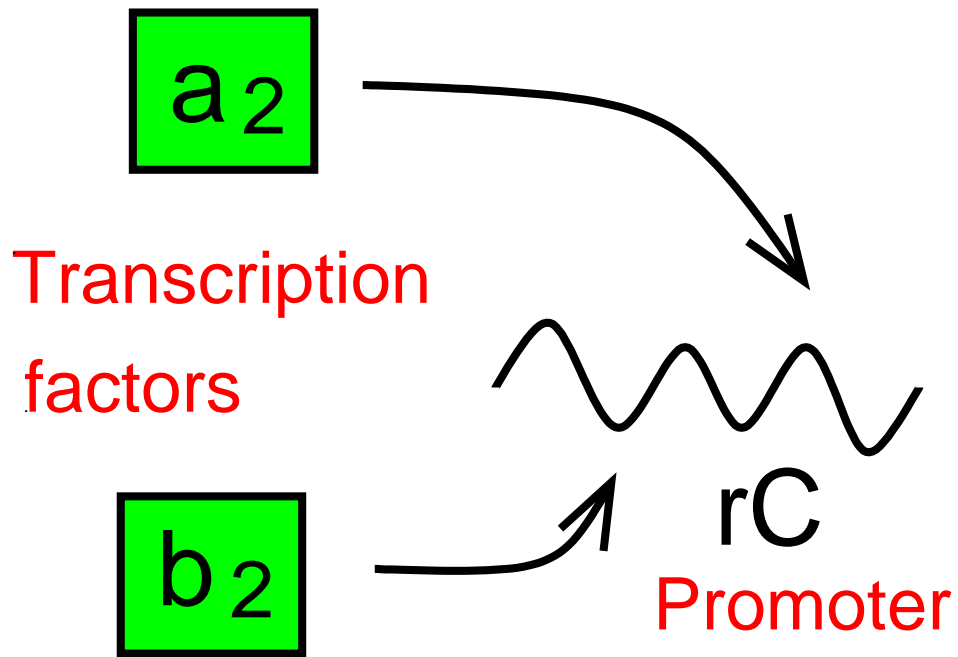
rC

Promoter

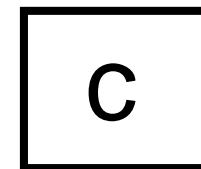
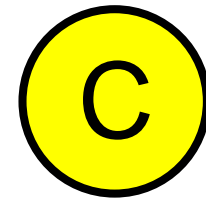
mRNA



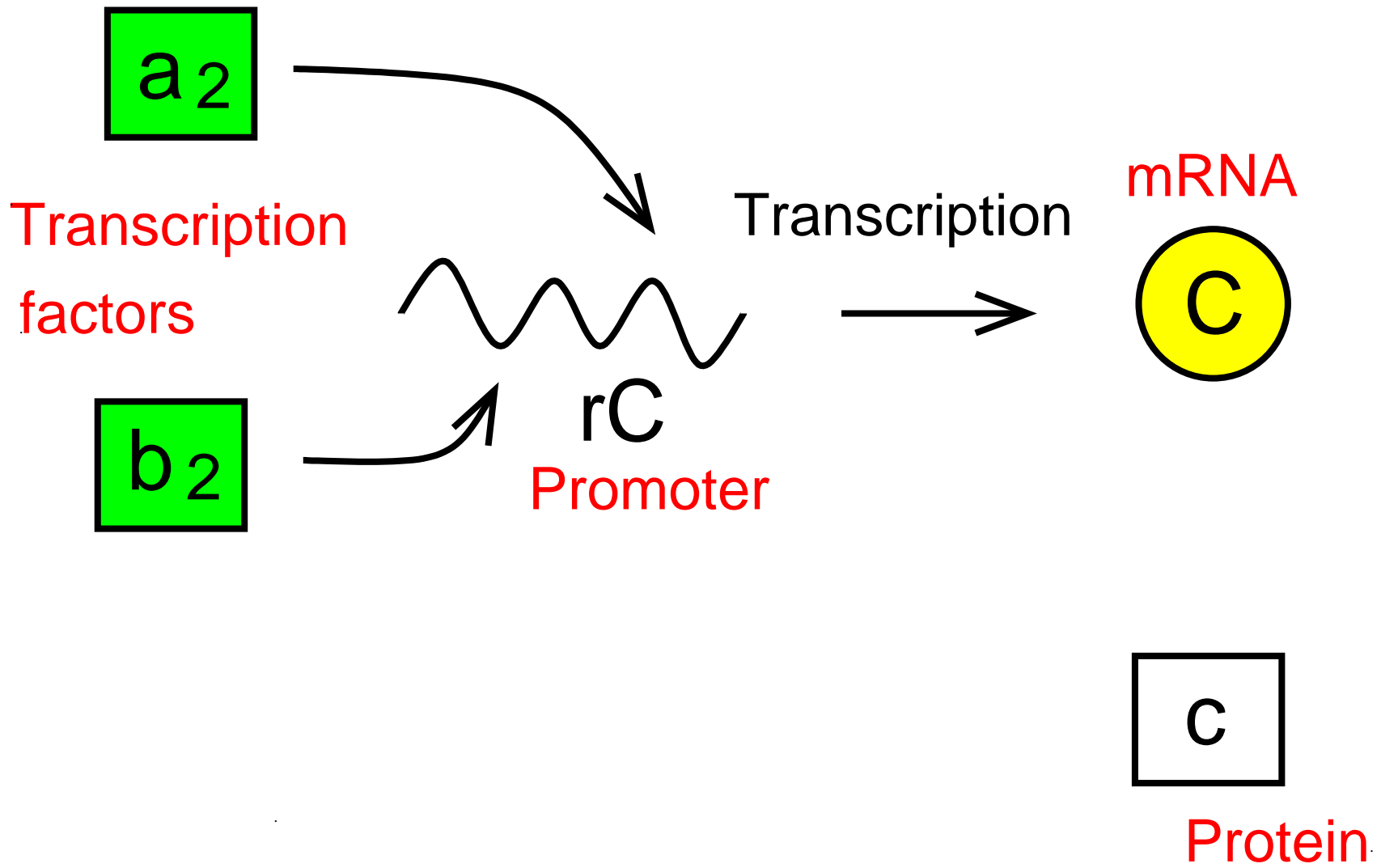
Protein

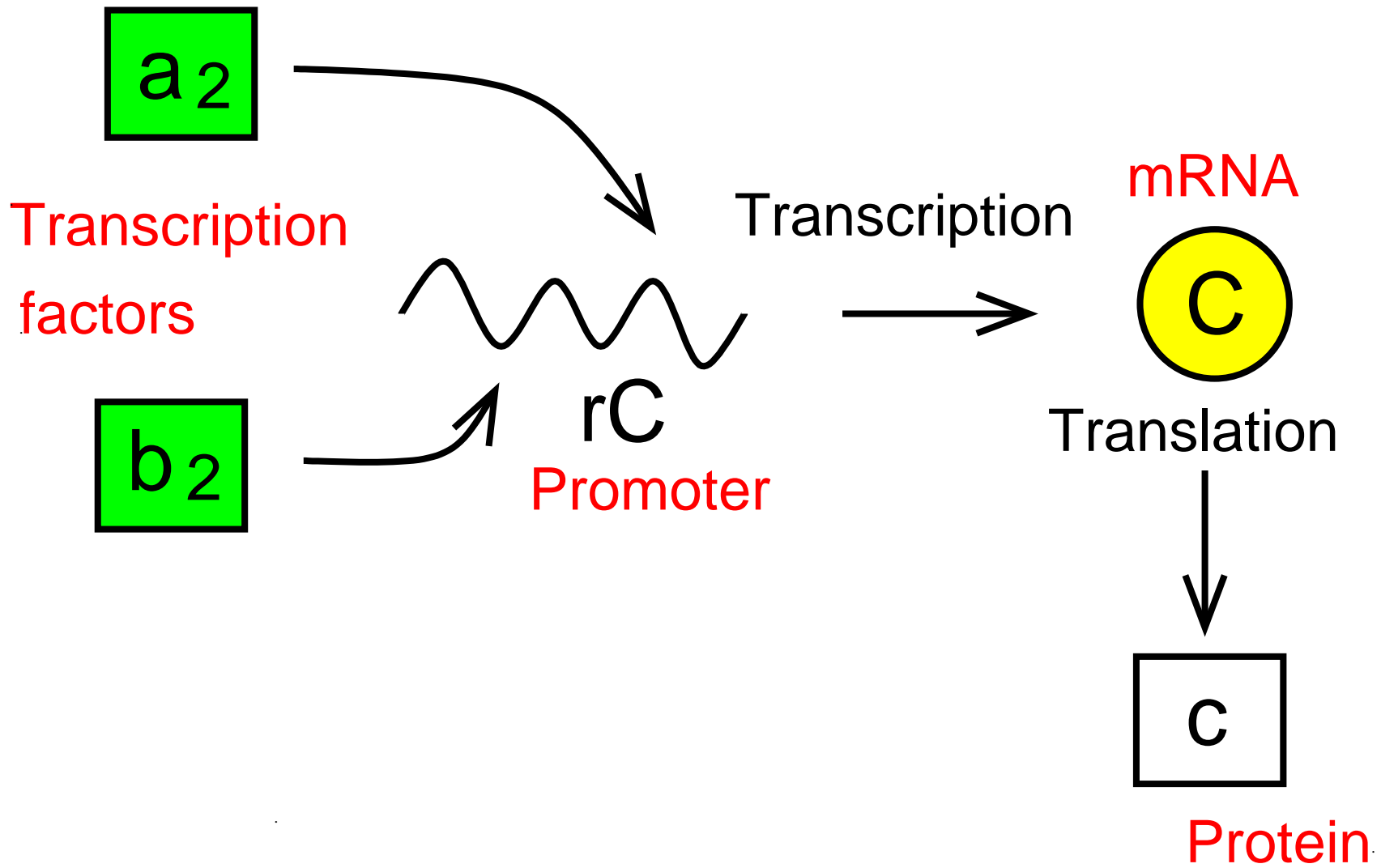


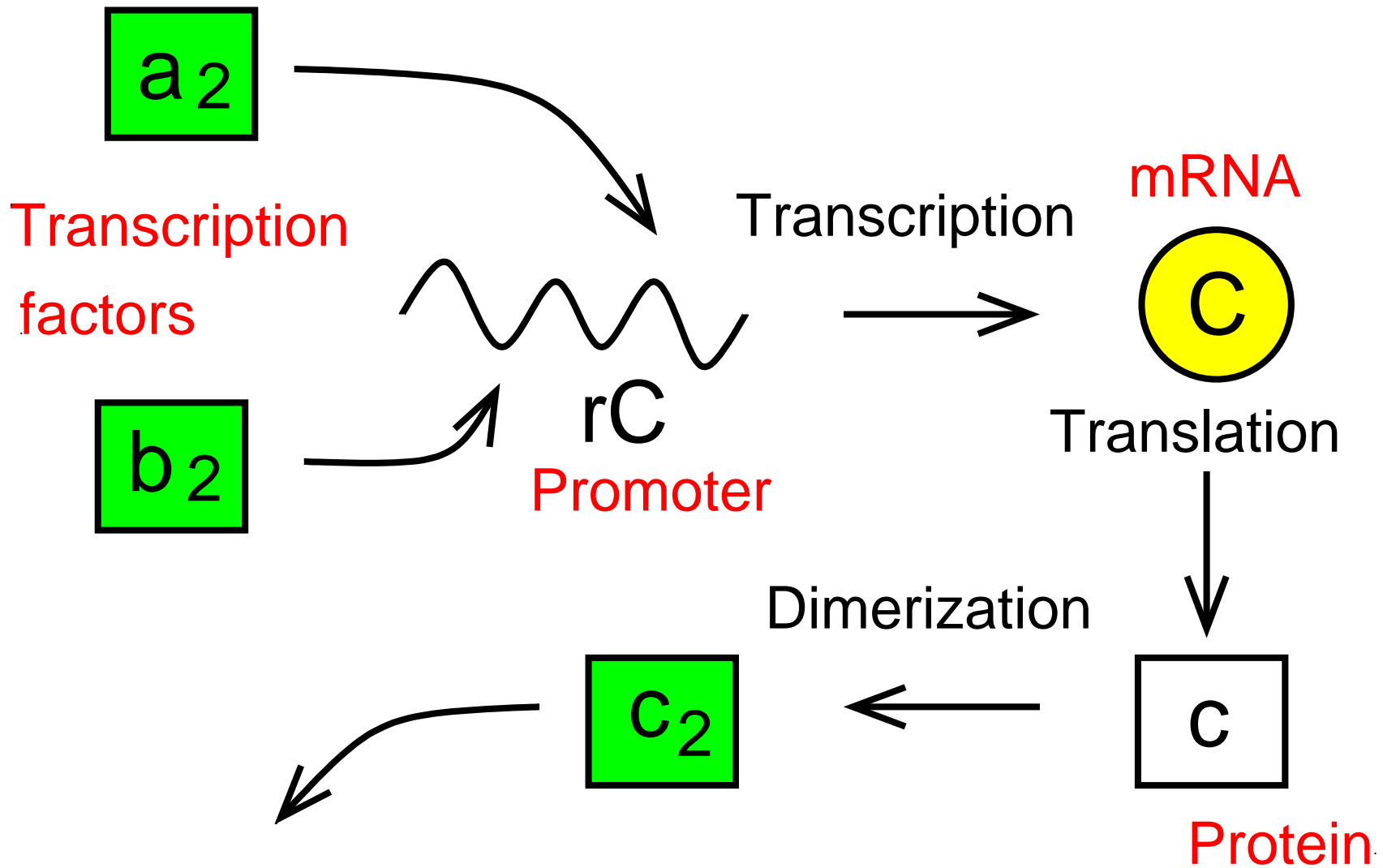
mRNA

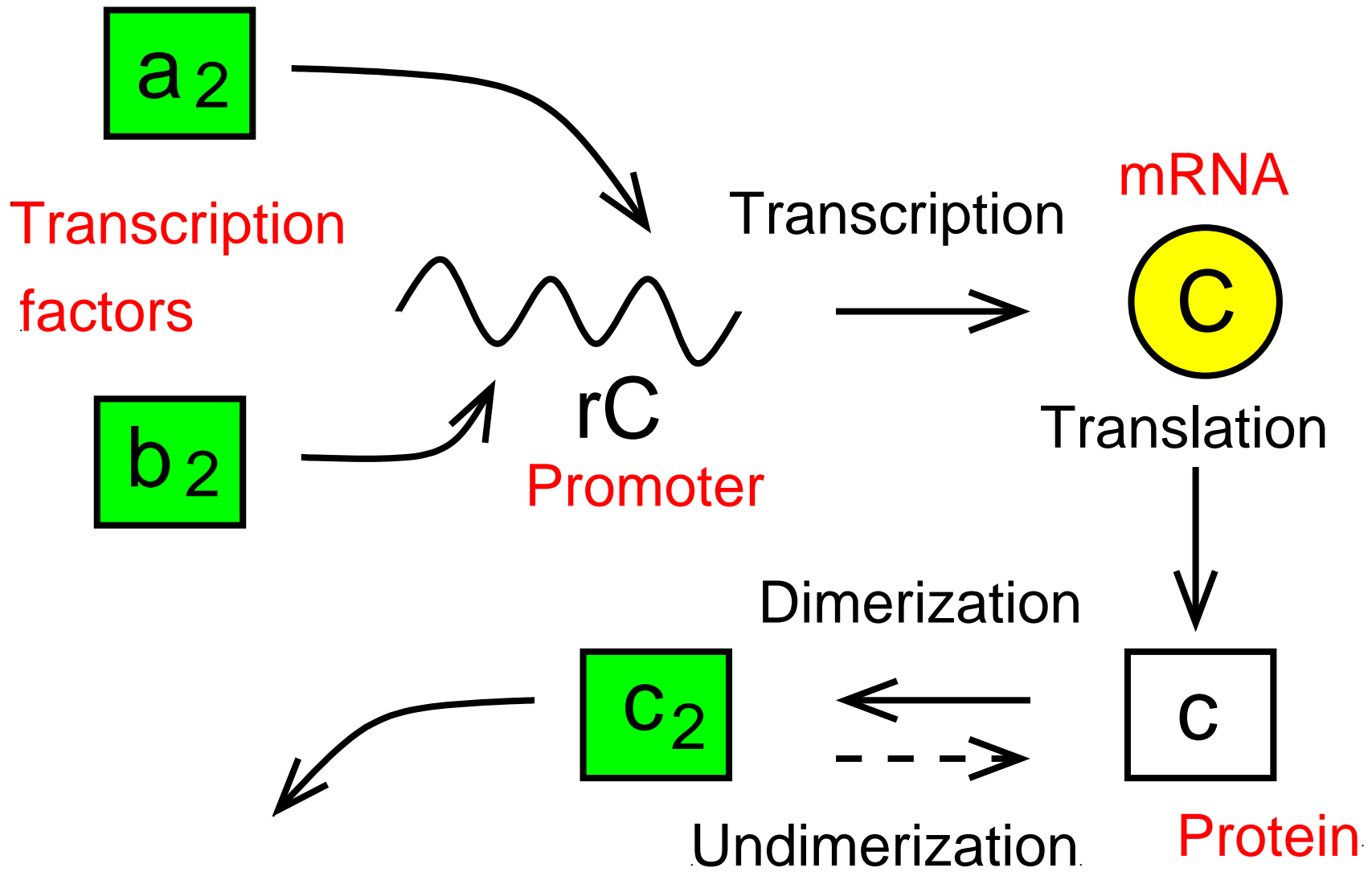


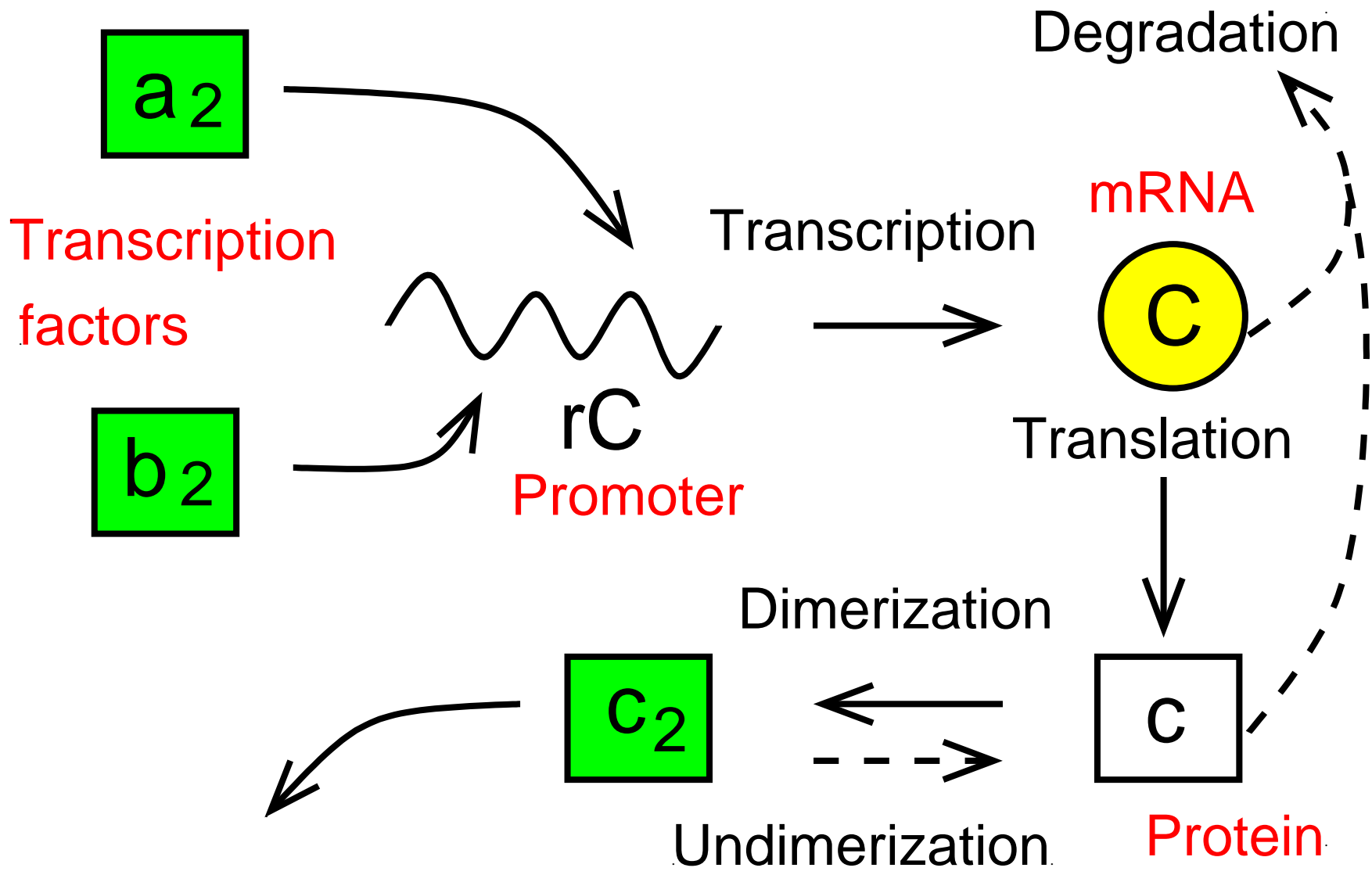
Protein

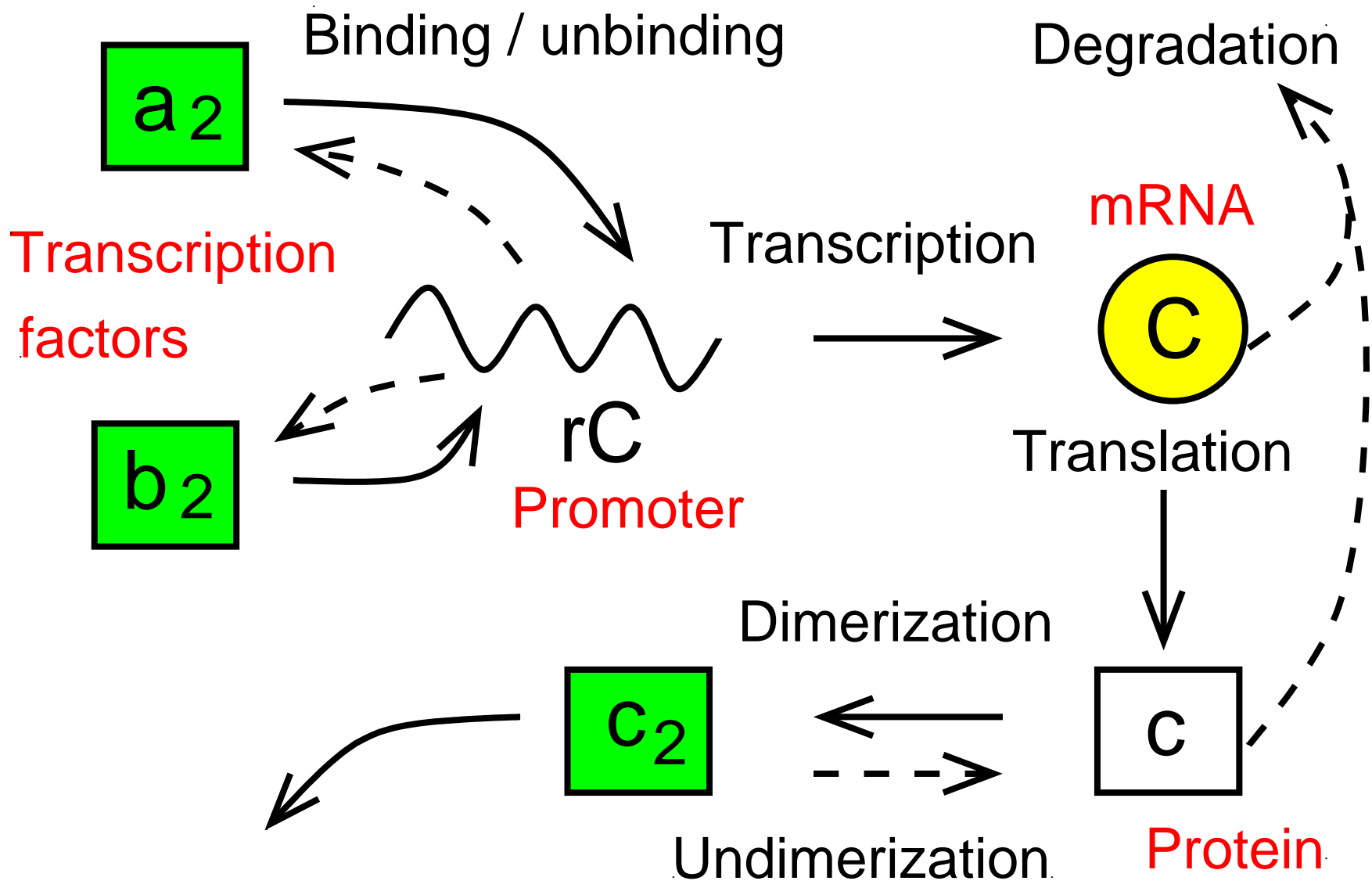










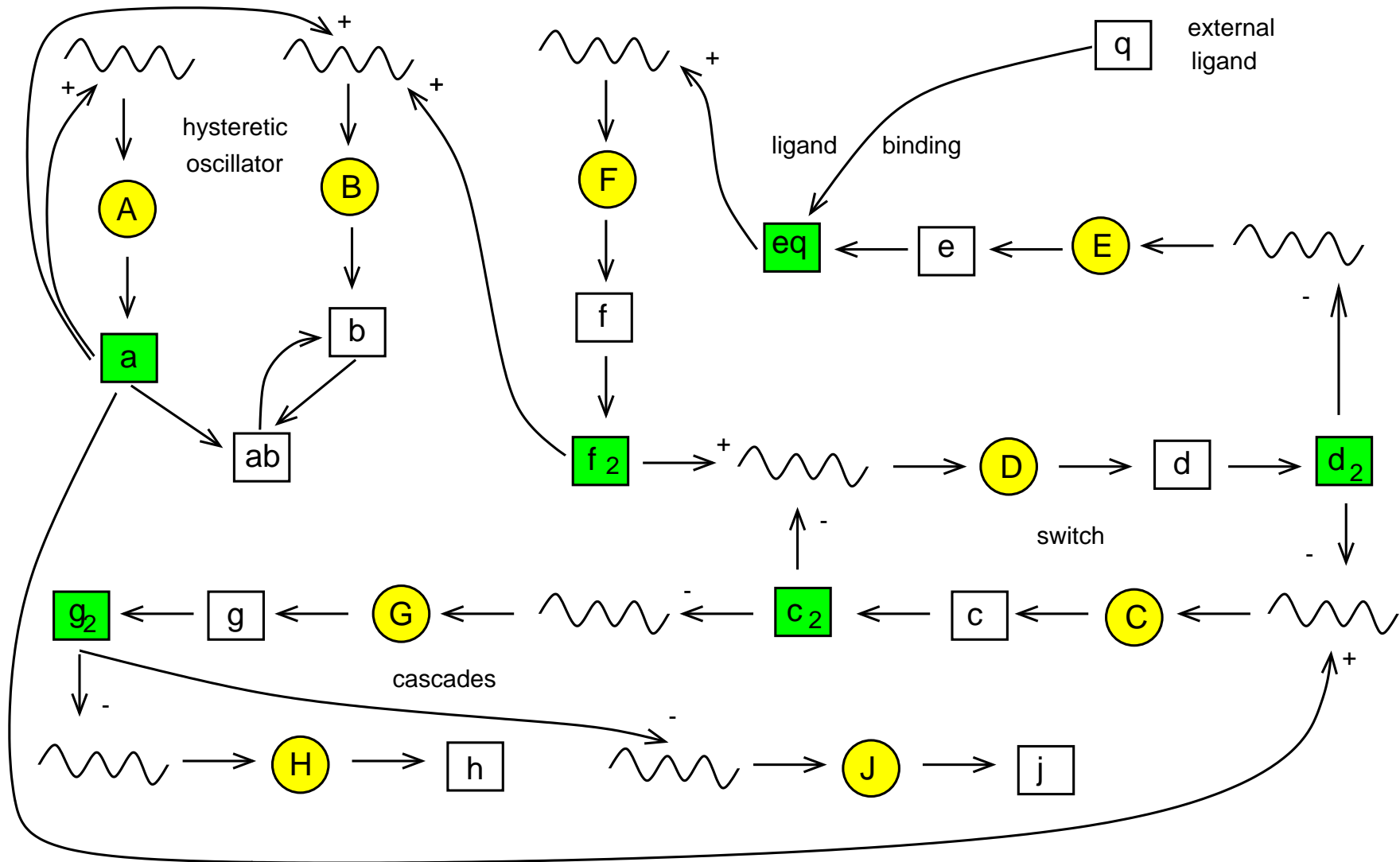


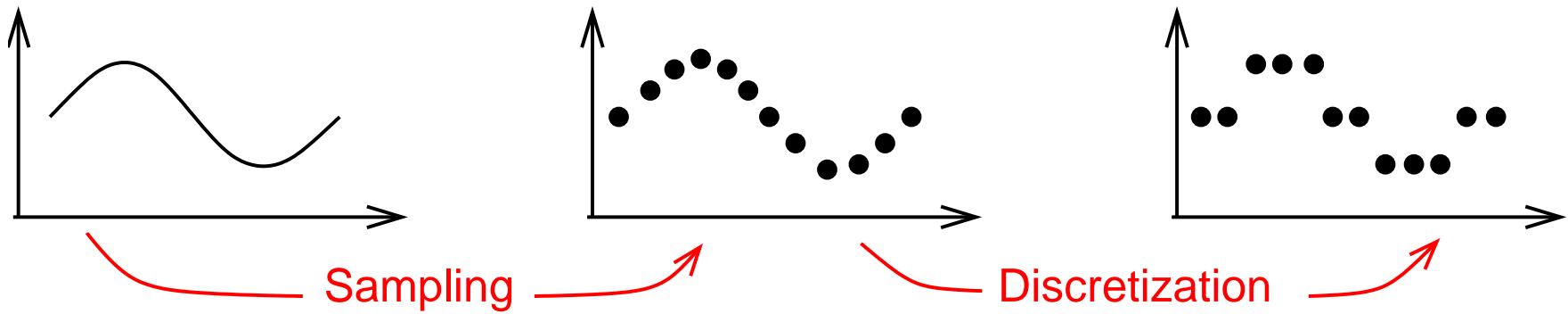
$$\frac{d}{dt}[a_2.rC] = \lambda_{a_2.rC}^+[a_2][rC] - \lambda_{a_2.rC}^-[a_2.rC]$$

$$\frac{d}{dt}[C] = \lambda_{rC}[rC] + \lambda_{a_2.rC}[a_2.rC] + \lambda_{b_2.rC}[b_2.rC] - \lambda_C[C]$$

$$\frac{d}{dt}[c] = \lambda_{Cc}[C] - \lambda_c[c]$$

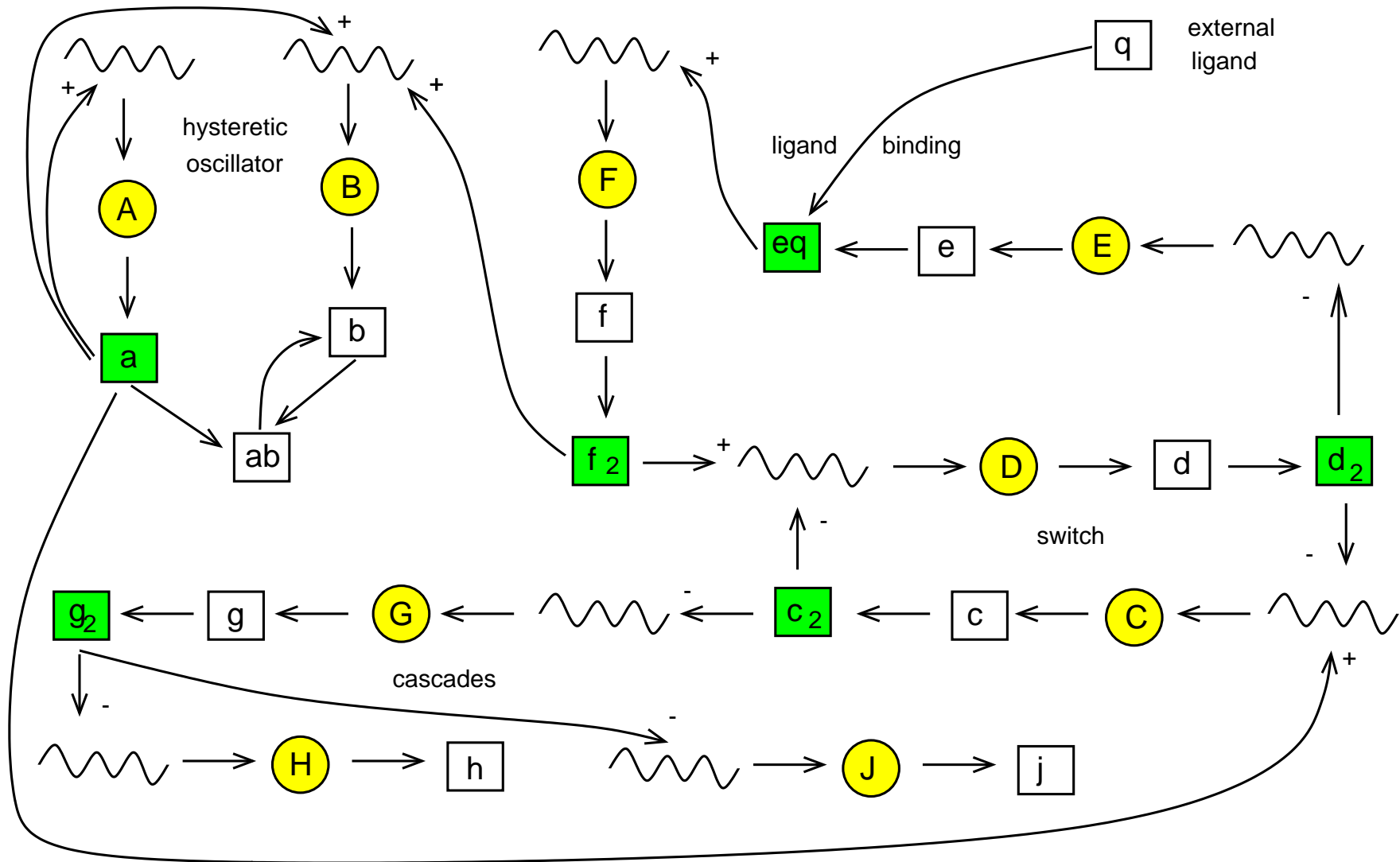
$$\frac{d}{dt}[c_2] = \lambda_{cc}^+[c]^2 - \lambda_{cc}^-[c_2]$$

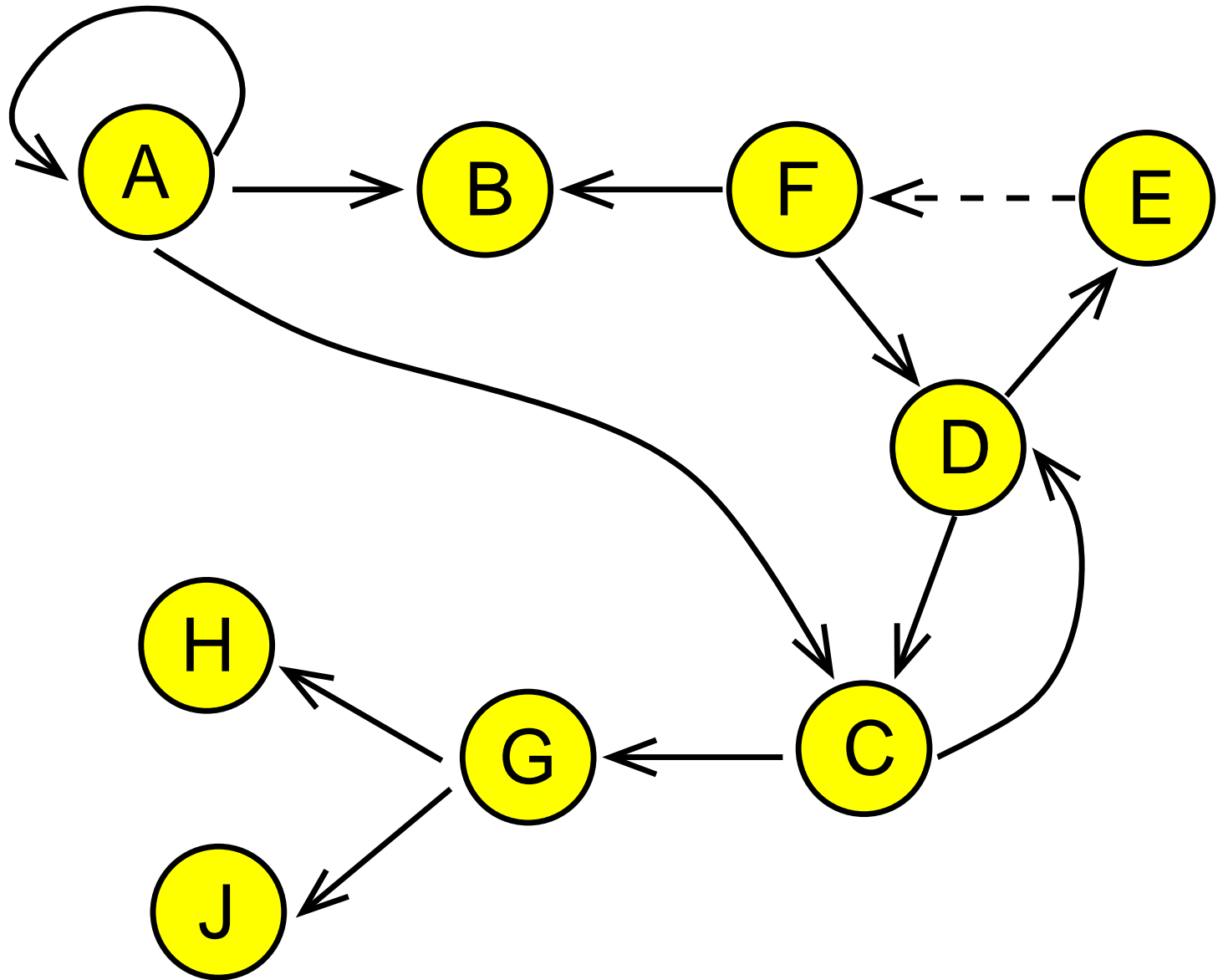


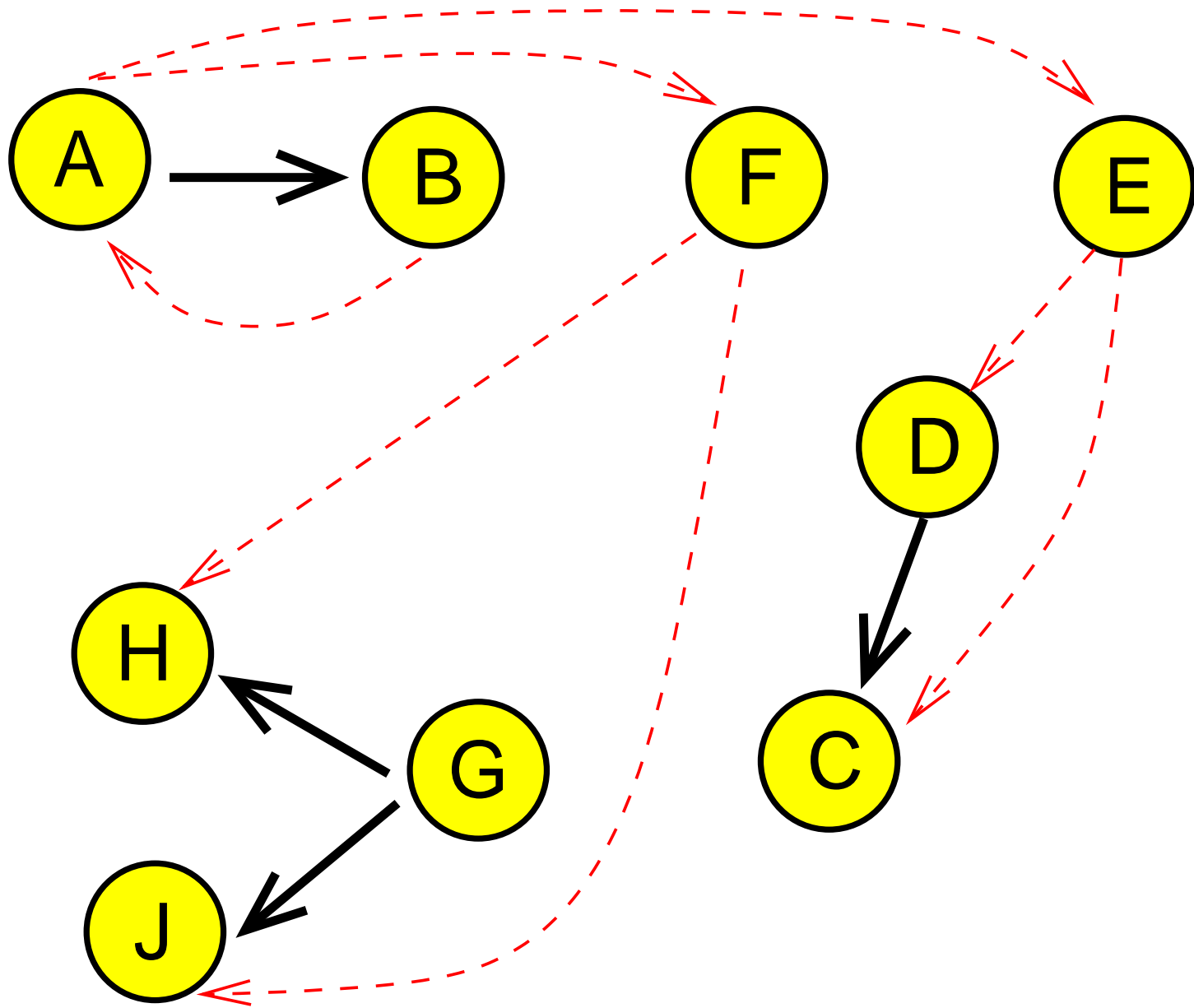


12 time points

Recover the **true genetic network**
with
reverse engineering.







Simulation Experiments

Ligand injection for 10 minutes.

Simulation Experiments

Ligand injection for 10 minutes.

Equilibrium

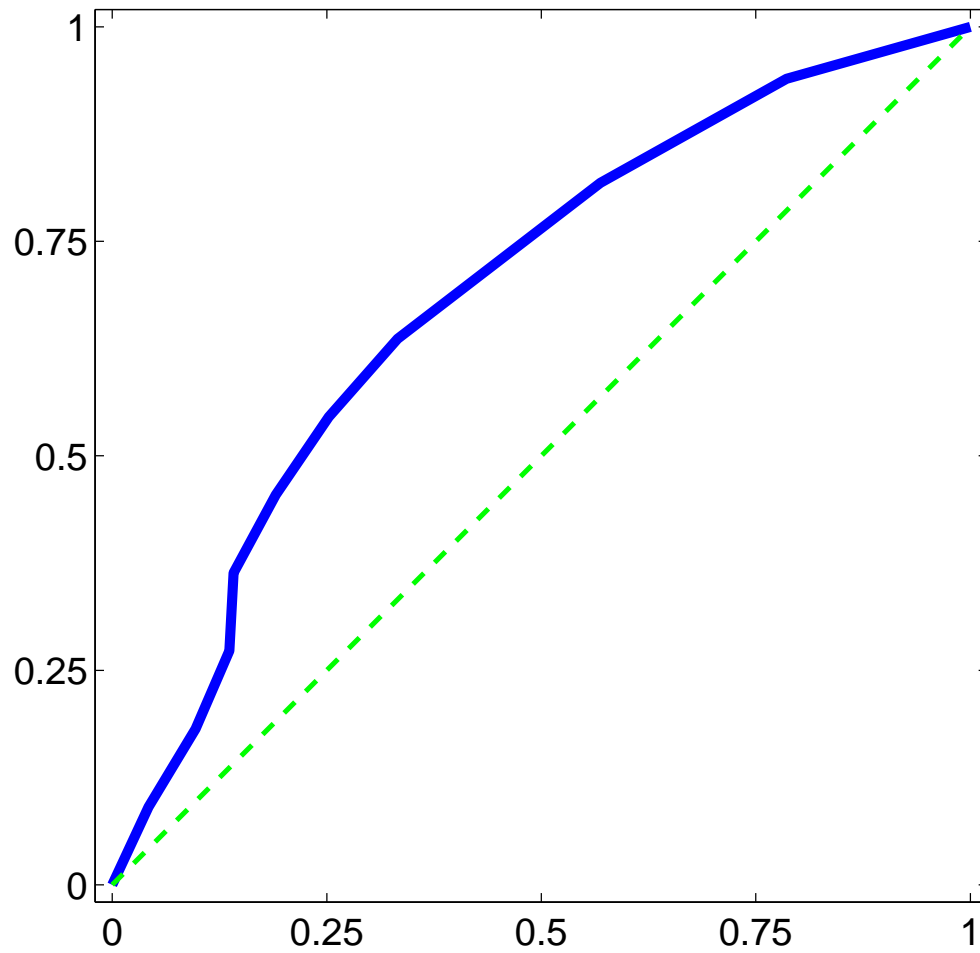
12 data points collected over 4000 min
in equi-distant intervals.

Disequilibrium

12 data points collected over 500 min
in equi-distant intervals.

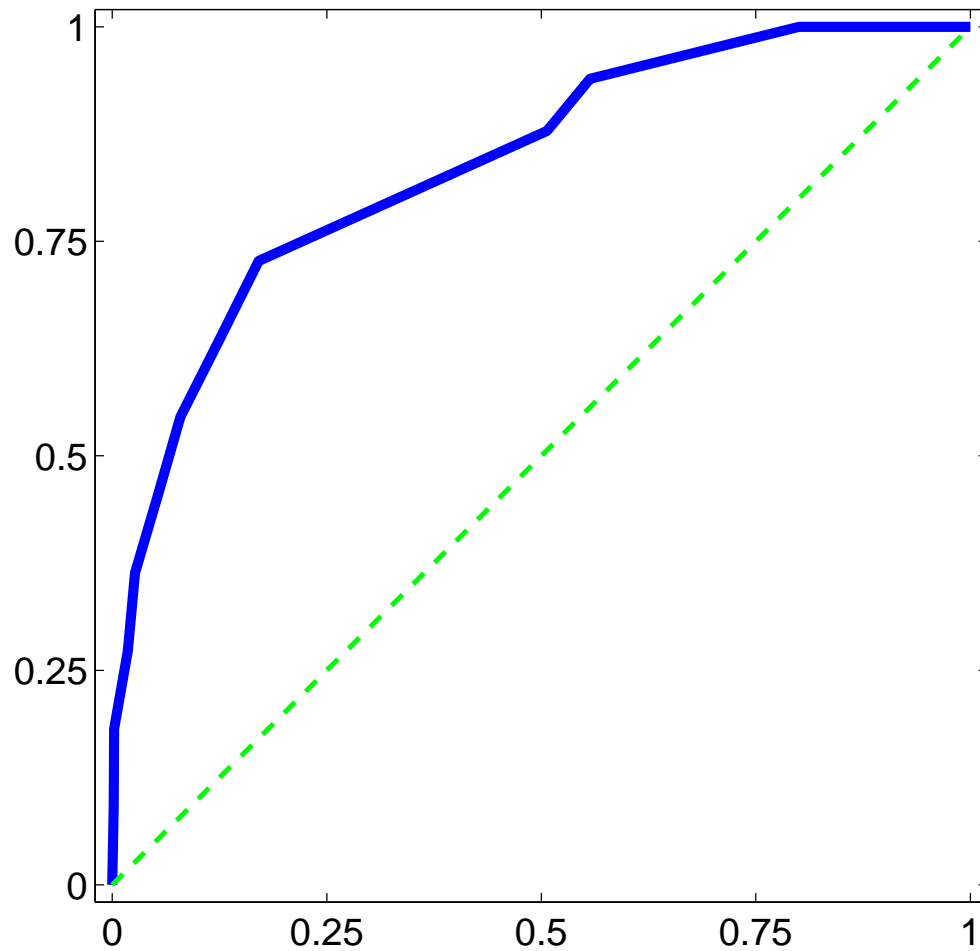
ROC curve: Equilibrium

True positives (vertical axis) \longleftrightarrow False positives (horizontal axis)

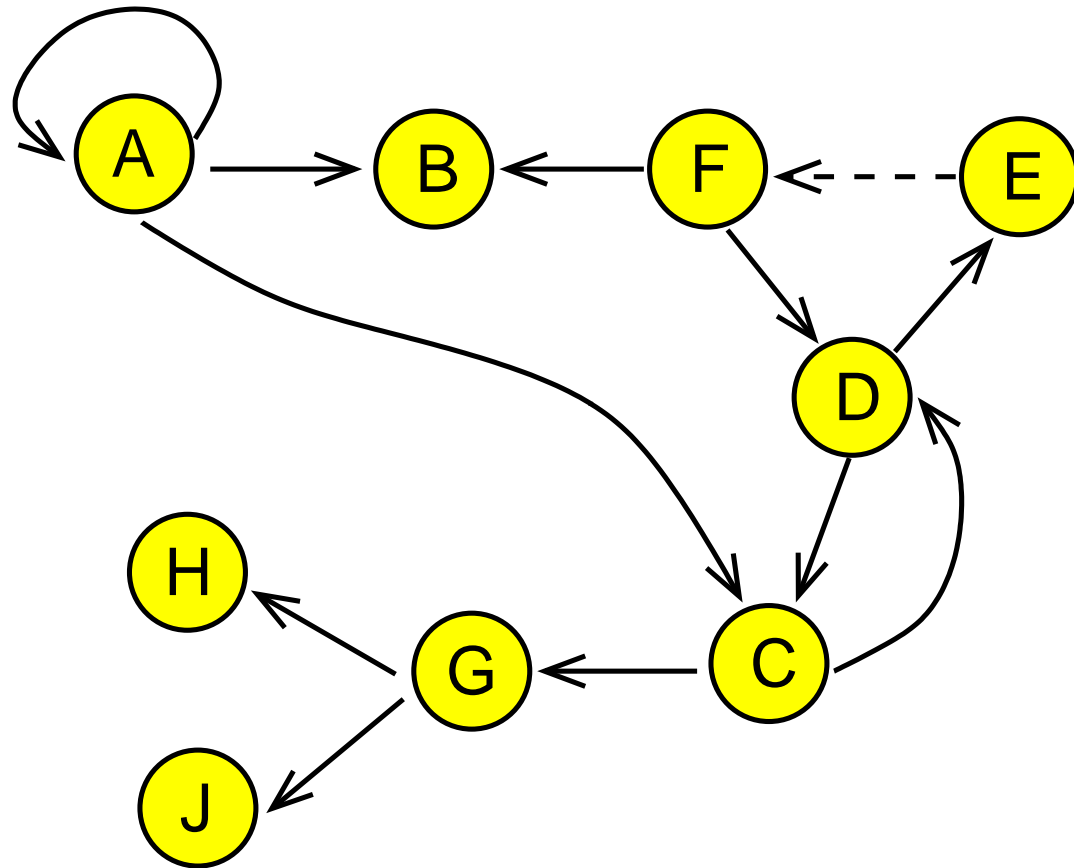


ROC curve: Disequilibrium

True positives (vertical axis) \longleftrightarrow False positives (horizontal axis)



Structure Prior



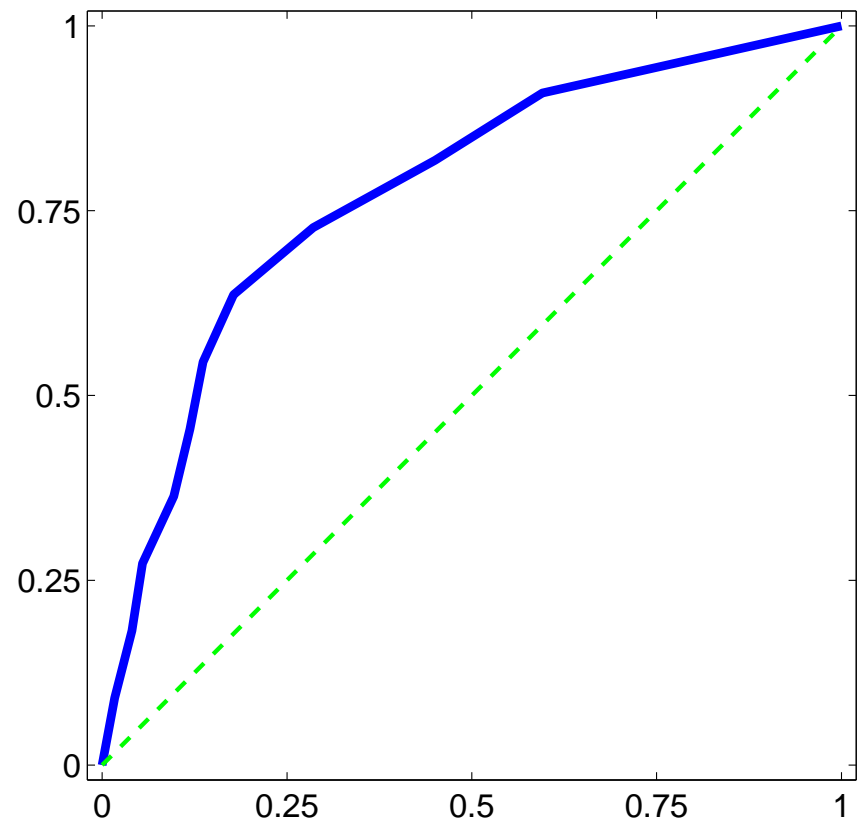
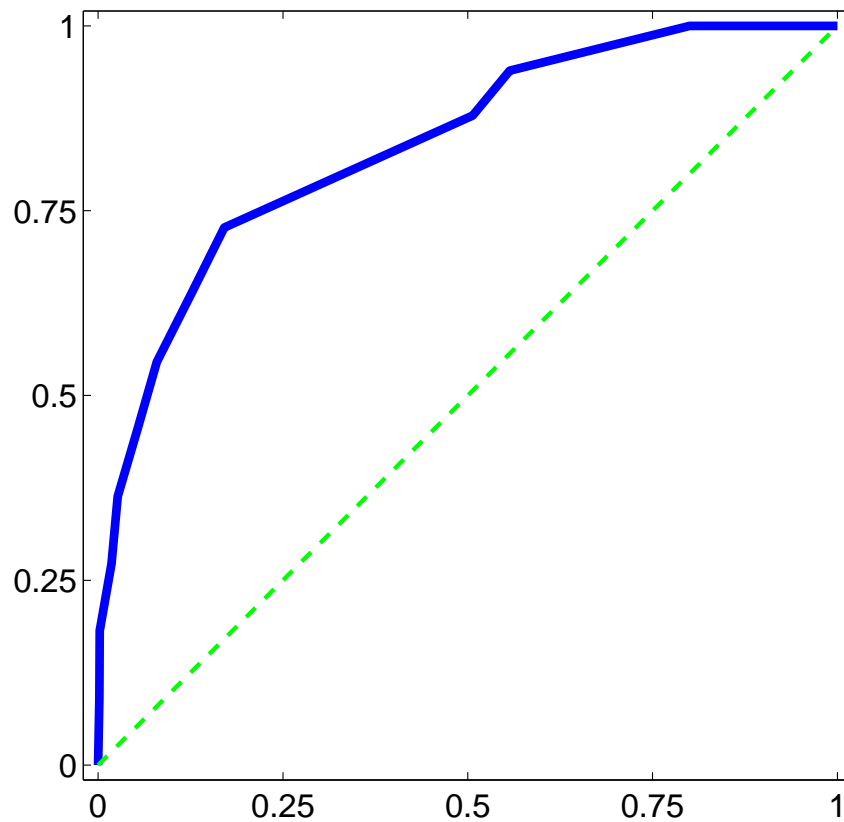
Max fan-in = 2, 3, 4

ROC curves

True positives (vertical axis) \longleftrightarrow False positives (horizontal axis)

Left: **max fan-in = 2**

Right: **max fan-in = 3**

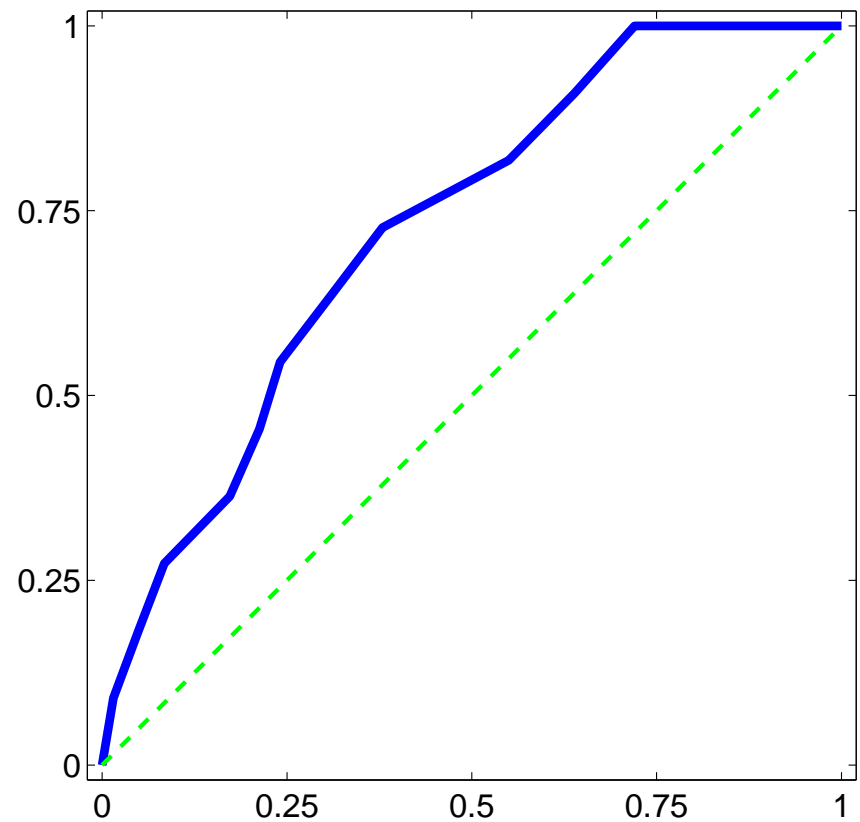
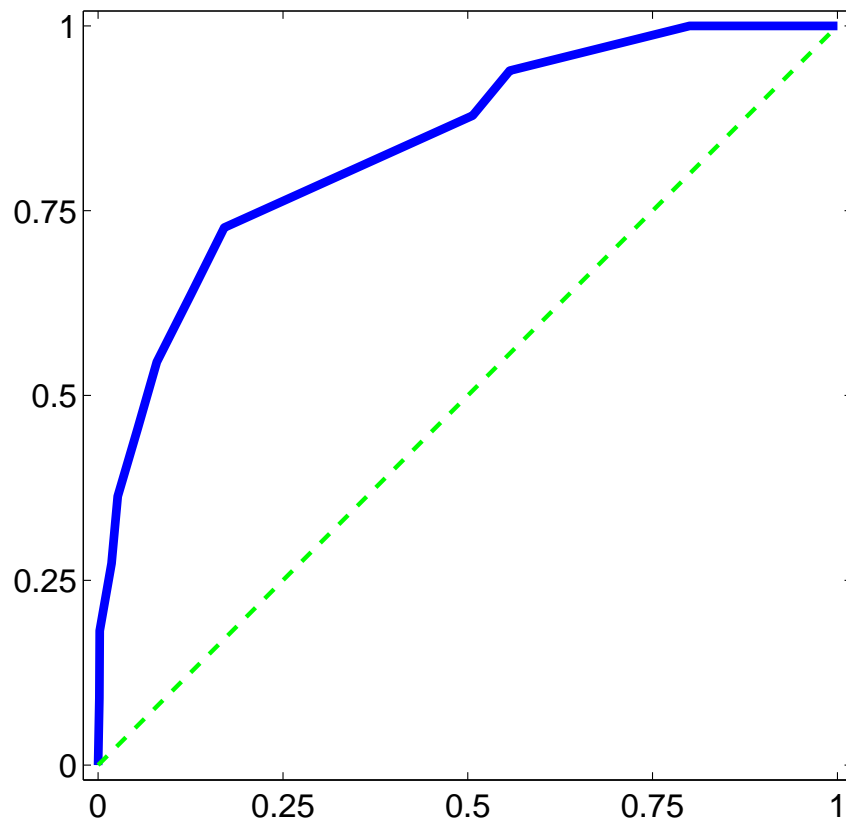


ROC curves

True positives (vertical axis) \longleftrightarrow False positives (horizontal axis)

Left: **max fan-in = 2**

Right: **max fan-in = 4**



Sequence information

$$\frac{P(y \rightarrow rX | r \in B[y])}{P(y \rightarrow rX | r \notin B[y])} = 2$$

$y \rightarrow rX$ denotes the event that transcription factor y binds to the promoter r upstream of gene X , and $B[y]$ is the set of (known) binding motifs for y .

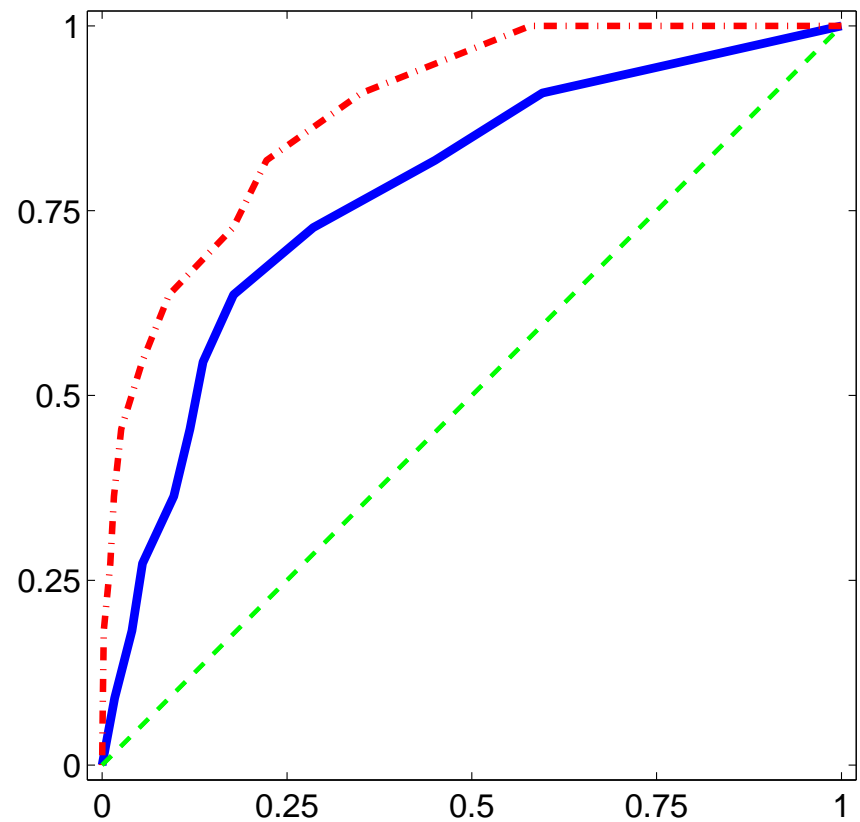
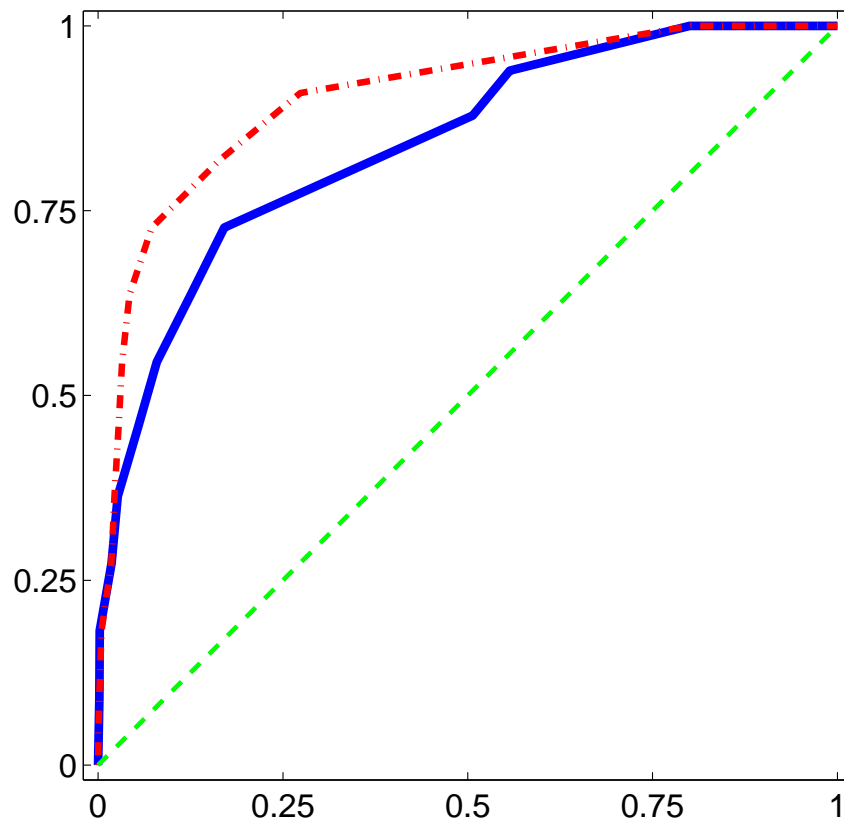
In words: The equation expresses that on identifying a binding motif for transcription factor y in the upstream region of gene X , this transcription factor is twice as likely to bind to X than in the absence of such a motif.

ROC curves

True positives (vertical axis) \longleftrightarrow False positives (horizontal axis)

Left: max fan-in = 2

Right: max fan-in = 3



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
- But: **Obscured by noise**

Conclusions

- Learning the **global network** → impossible
- Intrinsic **uncertainty** due to lack of data
- Inference of **local substructures** possible
- But: **Obscured by noise**
- **Simulation studies** and **ROC curves**:
Estimate the **sensitivity** and **specificity**