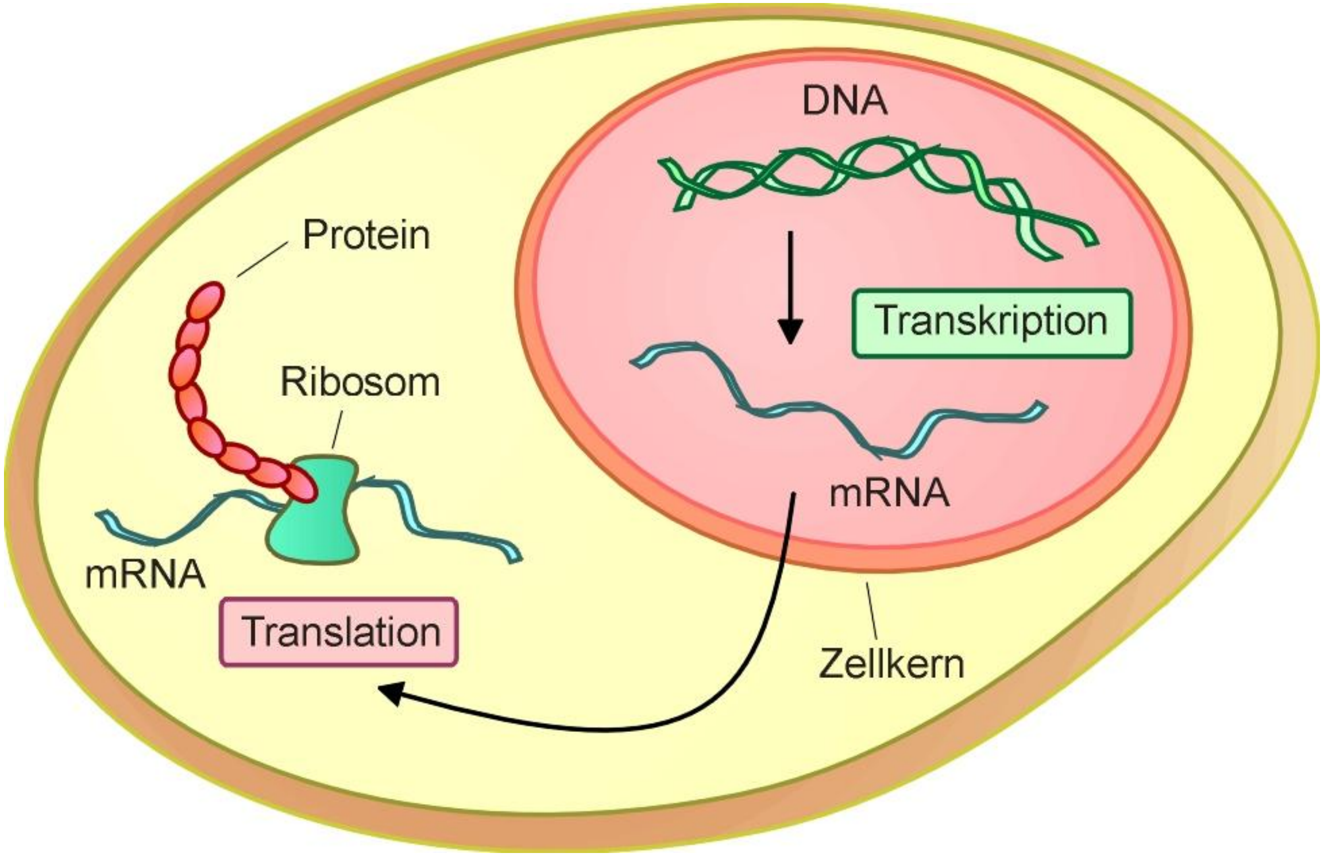
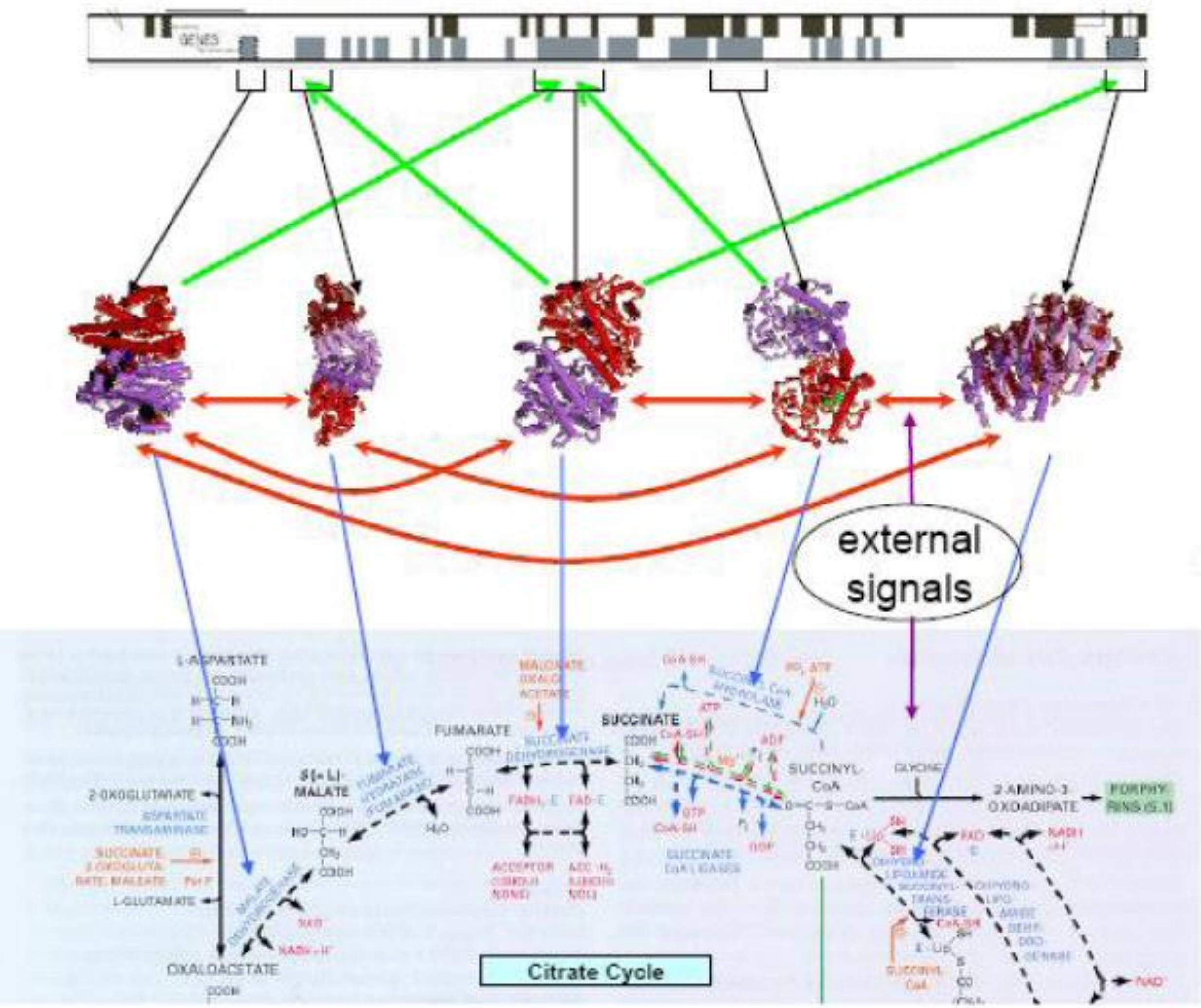
Systems Biology

Dirk Husmeier

BioSS

Systems Biology





GENOME

gene regulation

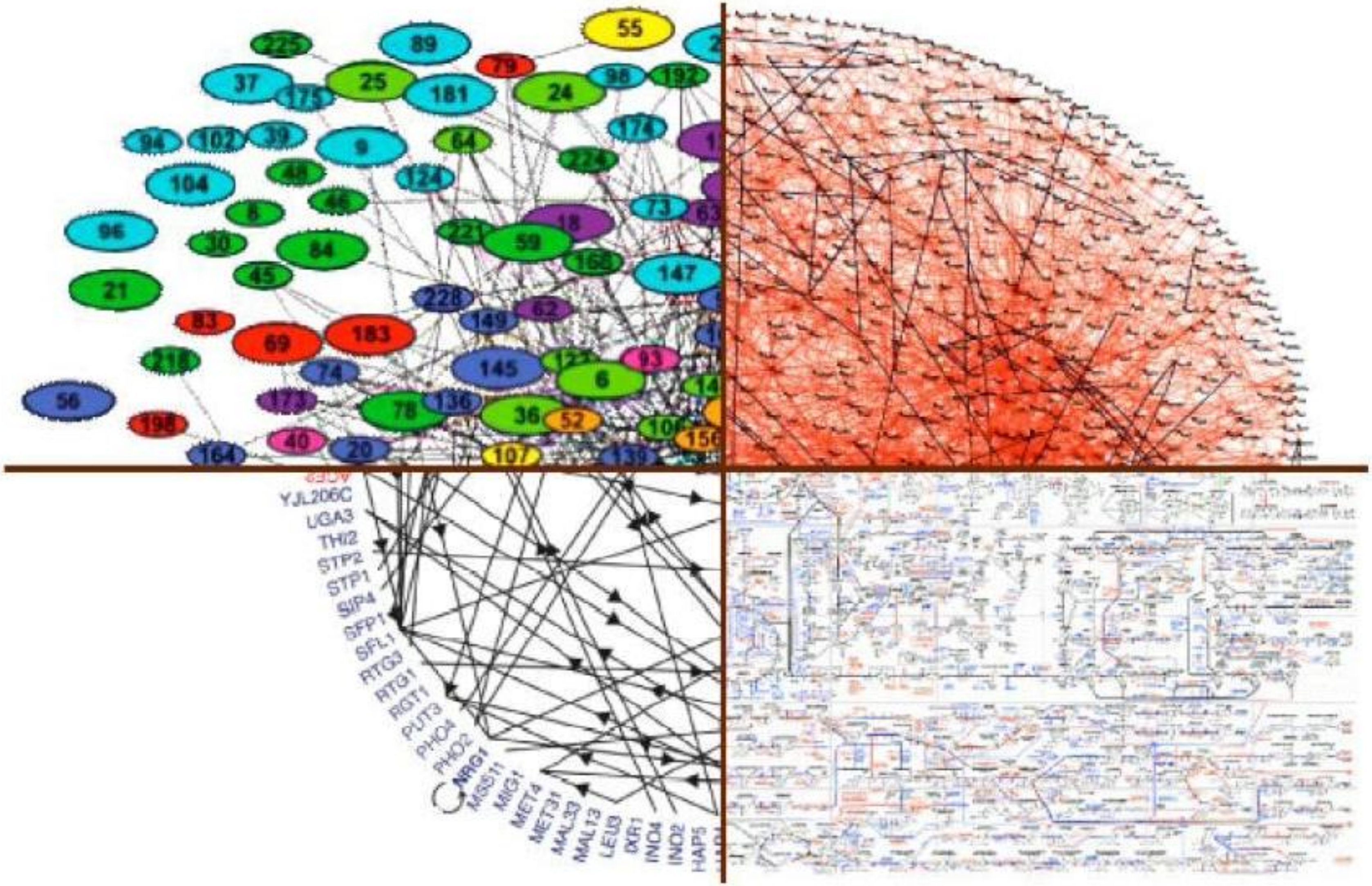
PROTEOME

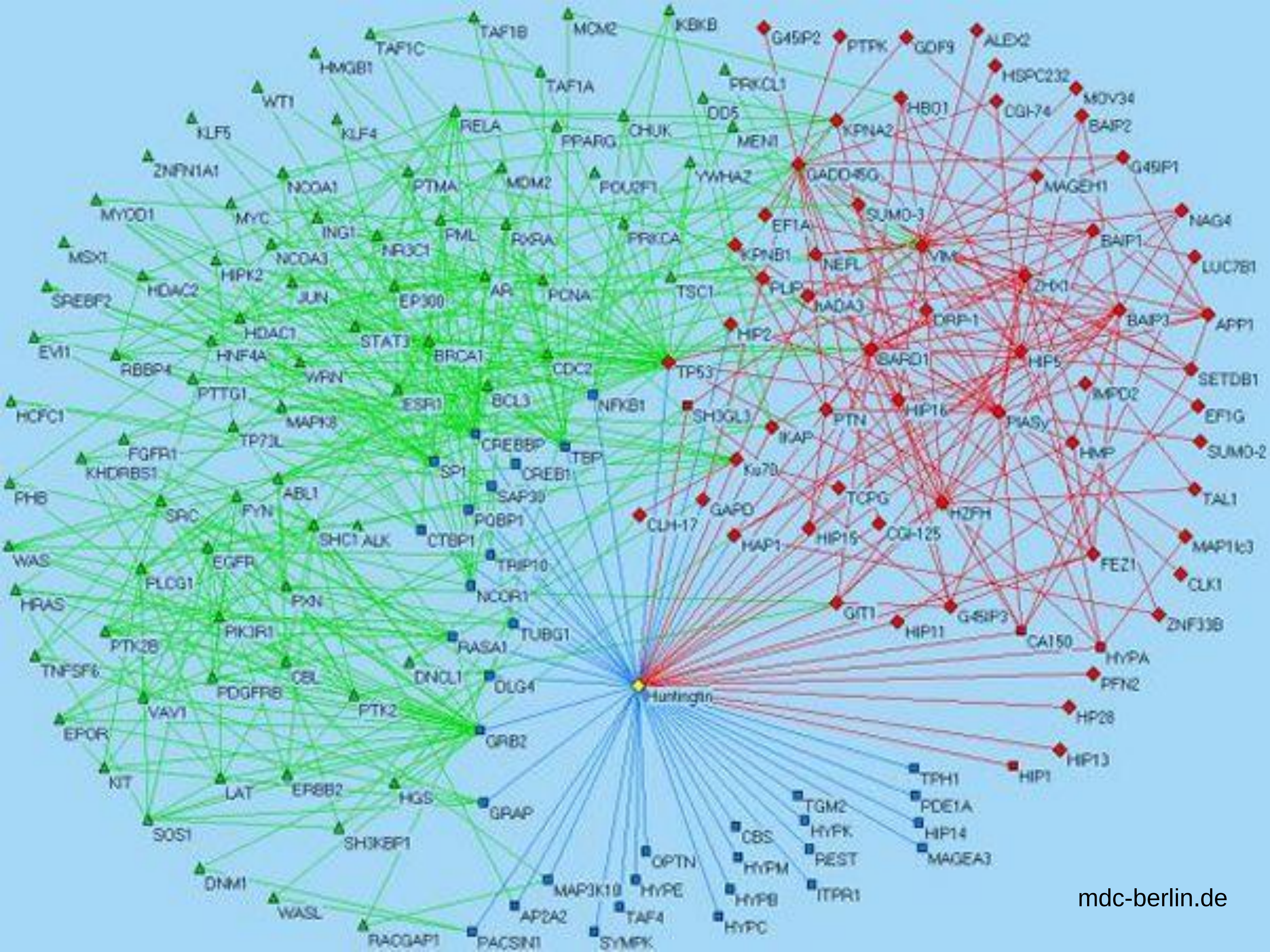
protein-protein interactions

signal transduction

METABOLISM

Bio-chemical reactions


55
225
89
79
37
25
181
24
98
192
175
94
102
39
9
64
174
104
48
124
224
8
46
73
96
30
84
221
59
18
168
45
147
21
228
62
83
149
69
183
218
74
145
93
6
56
173
78
136
36
52
198
40
106
156
164
20
107
139
YJL206C
UGA3
THI2
STP2
STP1
SIP4
SFP1
SFL1
RTG3
RTG1
RGT1
PUT3
PHO4
PHO2
NRG1
MSS11
MIG1
MET4
MET31
MAL33
MAL13
LEU3
IXR1
INO4
INO2
HAP5



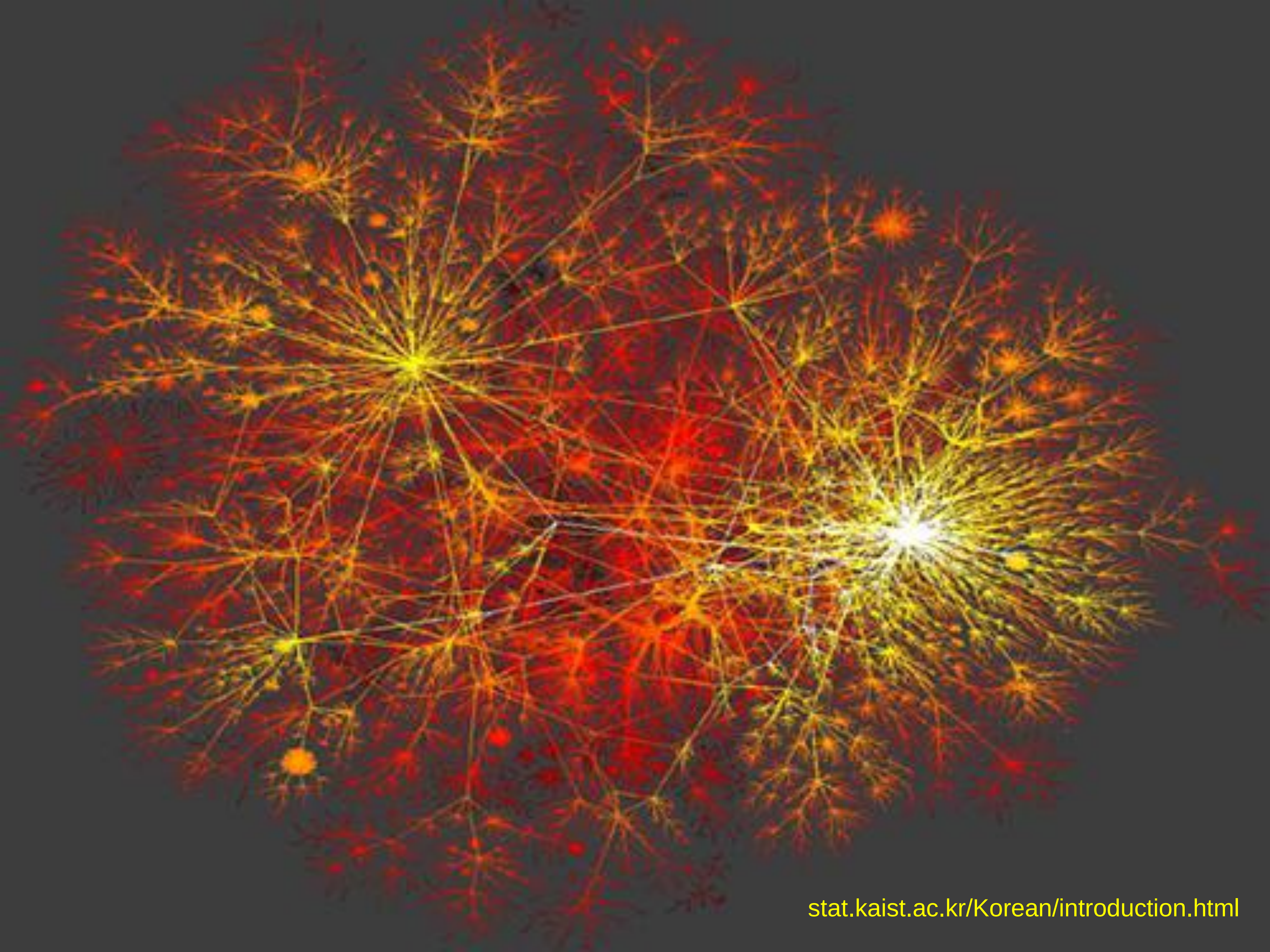
mdc-berlin.de

Topics in systems biology

- Network characterization
- Active pathways
- Network reconstruction

Topics in systems biology

- **Network characterization**
- Active pathways
- Network reconstruction

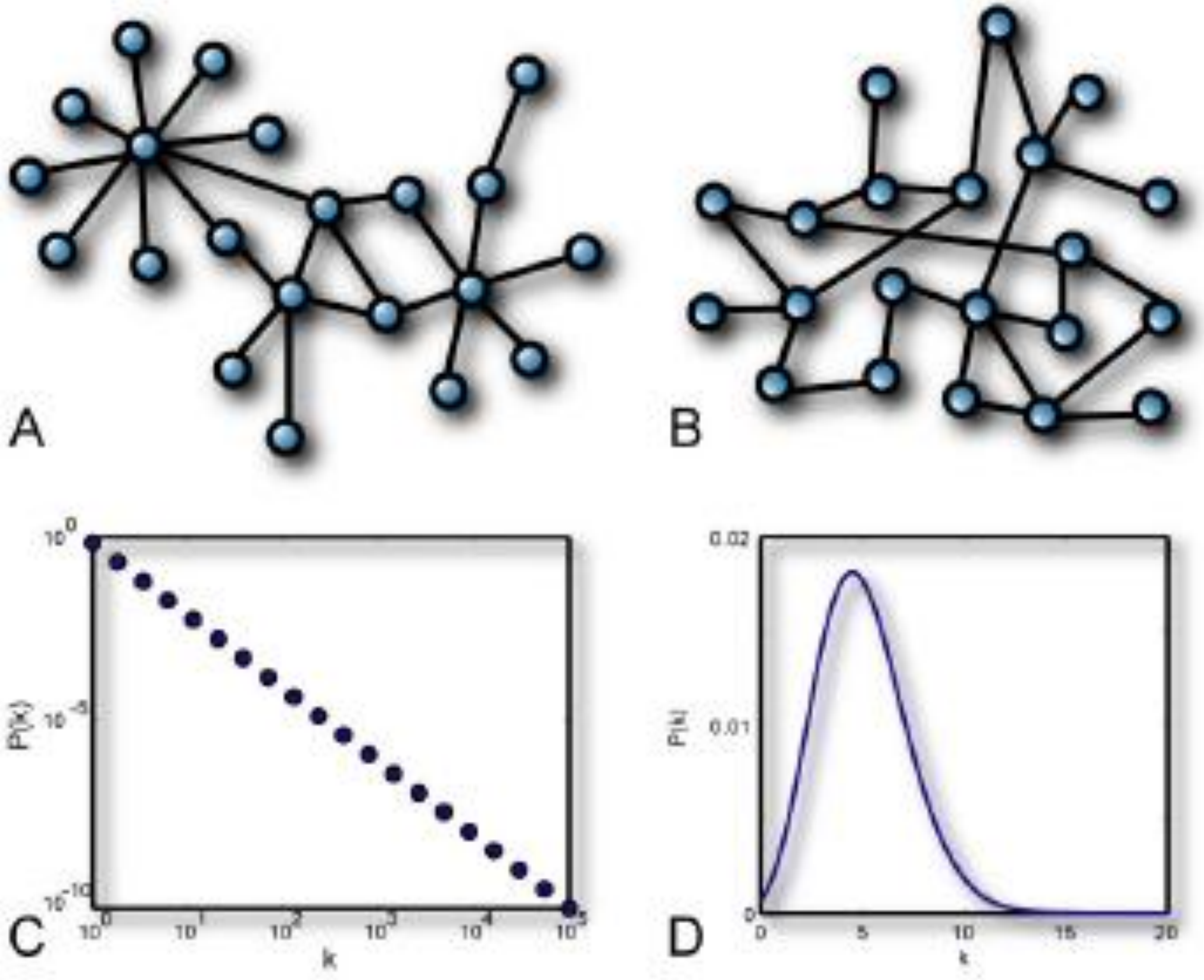


stat.kaist.ac.kr/Korean/introduction.html

Network statistics

- Degree of a node: The number of edges attached to it.
- Degree distribution: Distribution of the individual node degrees for the entire network.
- Power law degree distribution: $P(k) \sim k^{-\alpha}$
- Clustering coefficient: Measure of the average neighbourhood of a graph. Probability that two nodes that are connected to a third node are themselves connected.
- Network diameter: Mean shortest path between all nodes in the network.

Degree distribution and power law



Log P(k)

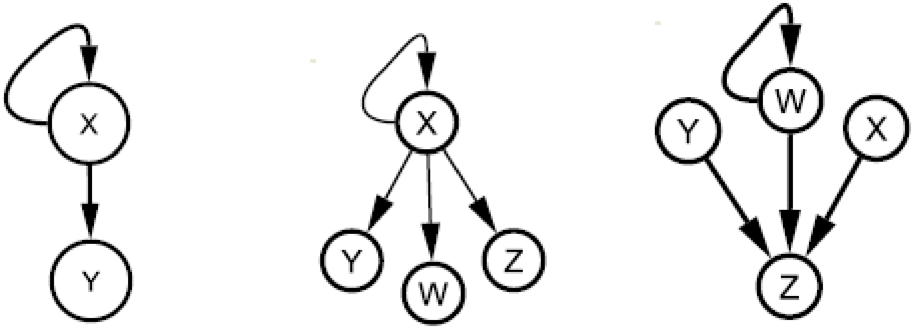
P(k)

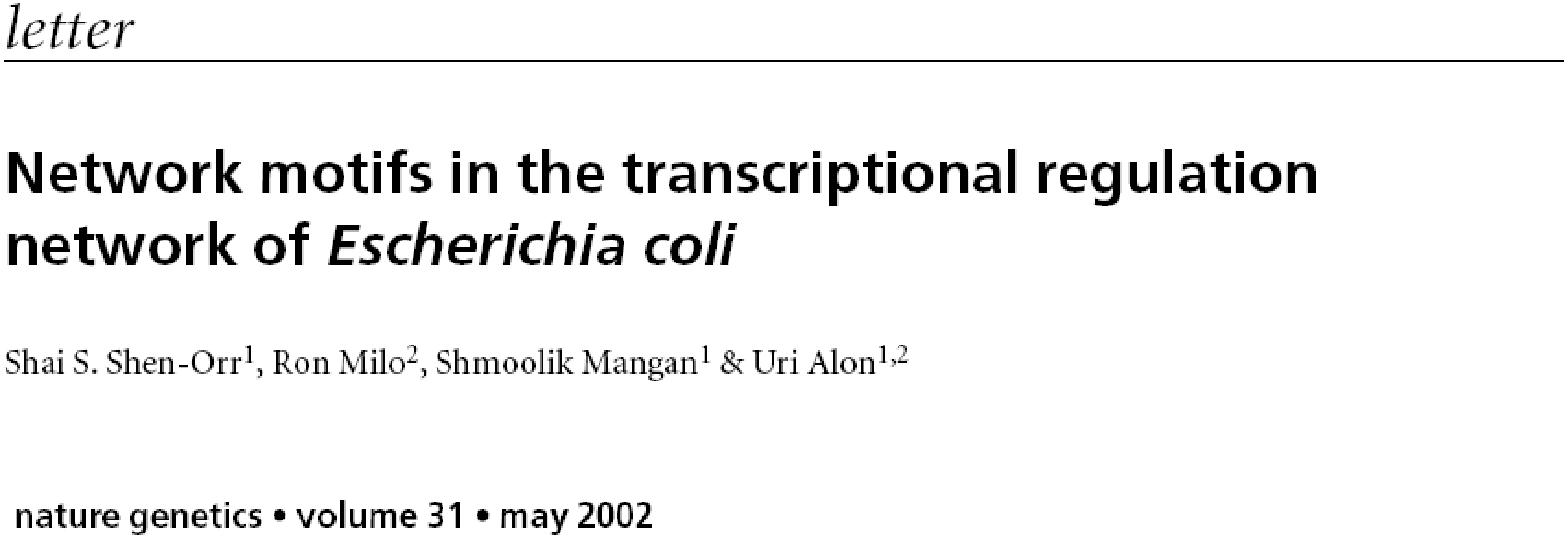
Log k

k

From https://nwb.slis.indiana.edu/community/uploads/CustomFillings/1.jpg

Network motifs





letter

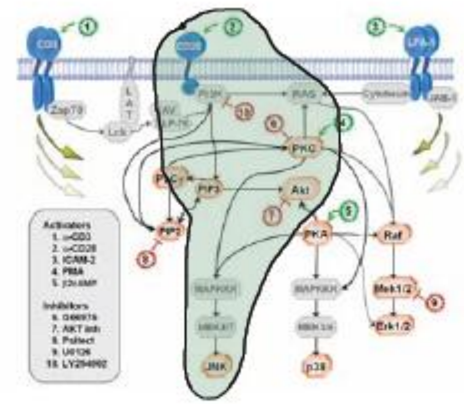
Network motifs in the transcriptional regulation network of *Escherichia coli*

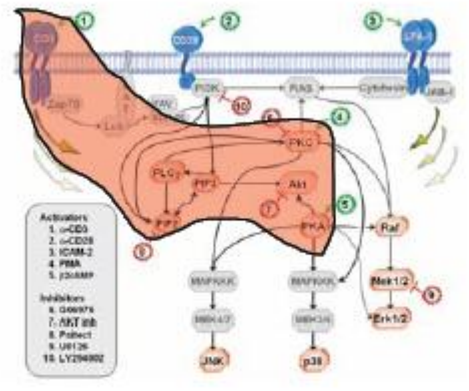
Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

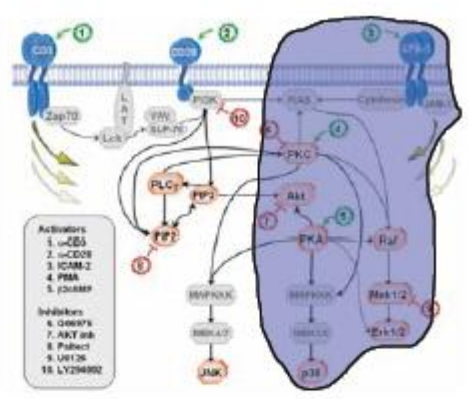
nature genetics • volume 31 • may 2002

Topics in systems biology

- Network characterization
- **Active pathways**
- Network reconstruction







BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 8 2008, pages 1078–1084
doi:10.1093/bioinformatics/btn066

Systems biology

MMG: a probabilistic tool to identify submodules of metabolic pathways

Guido Sanguinetti[1],*, Josselin Noirel[2] and Phillip C. Wright[2]

[1]Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Road, Sheffield, S1 4DP, UK and [2]Biological and Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

Received on November 16, 2007; revised on February 13, 2008; accepted on February 16, 2008

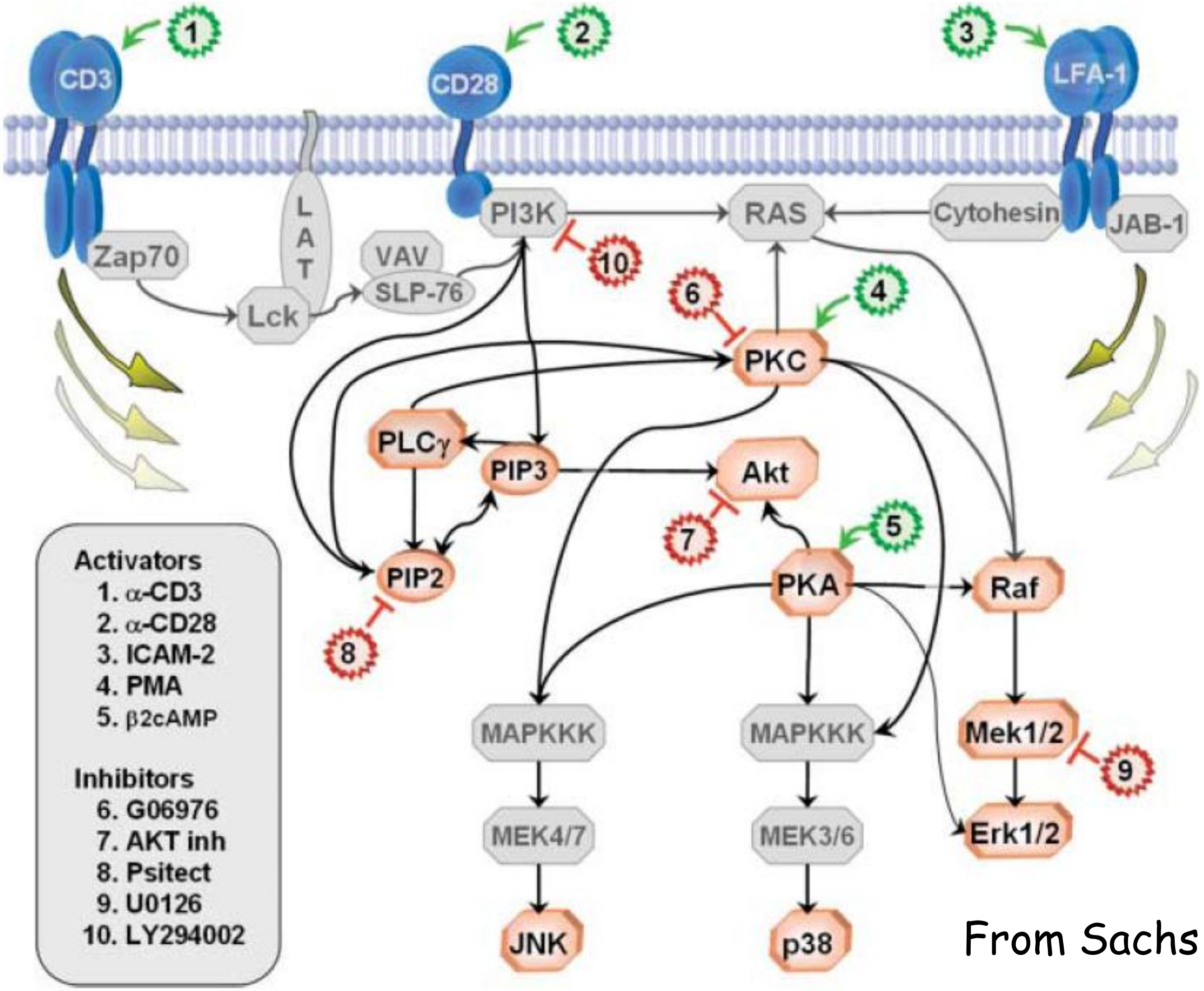
Advance Access publication February 21, 2008

Associate Editor: Jonathan Wren

Topics in systems biology

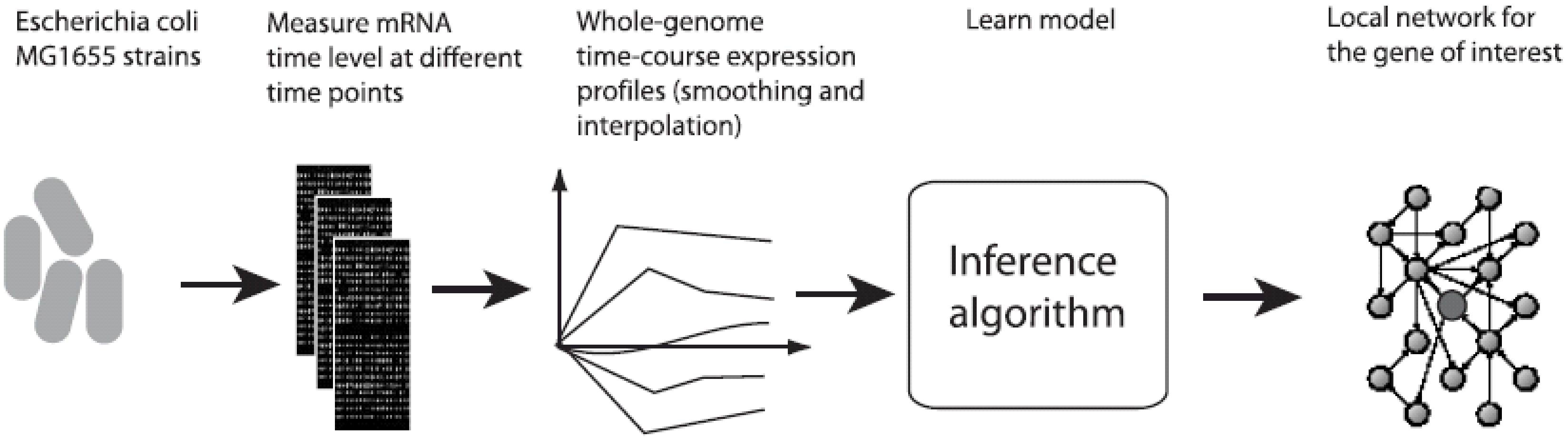
- Network characterization
- Active pathways
- **Network reconstruction**

Can we learn the signalling pathway from data?



From Sachs et al Science 2005

Network reconstruction from postgenomic data



Escherichia coli MG1655 strains

Measure mRNA time level at different time points

Whole-genome time-course expression profiles (smoothing and interpolation)

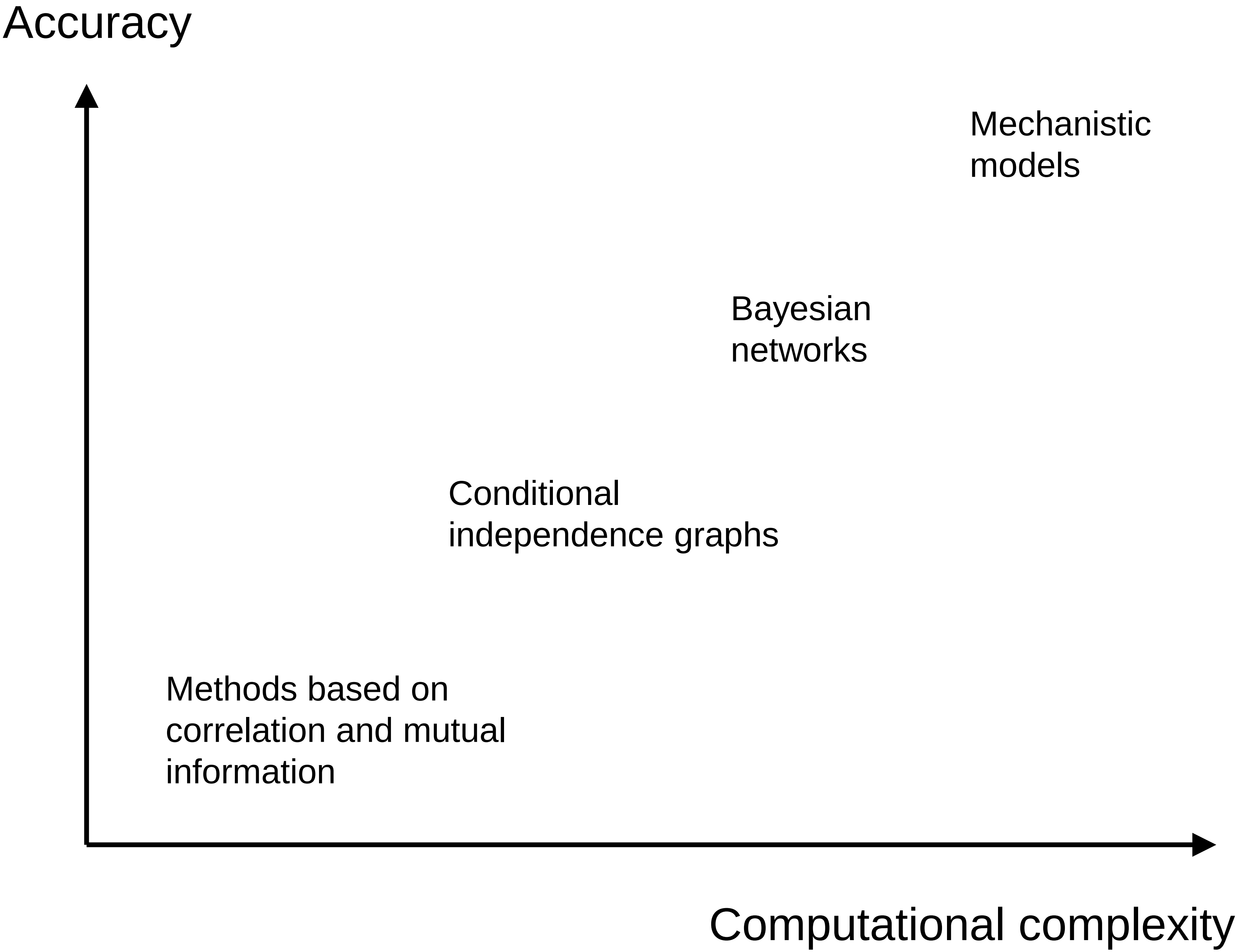
Learn model

Inference algorithm

Local network for the gene of interest

Mukesh Bansal[1,2], Giusy Della Gatta[1,3] and Diego di Bernardo[1,2,*]

Vol. 22 no. 7 2006, pages 815–822
doi:10.1093/bioinformatics/btl003



Accuracy

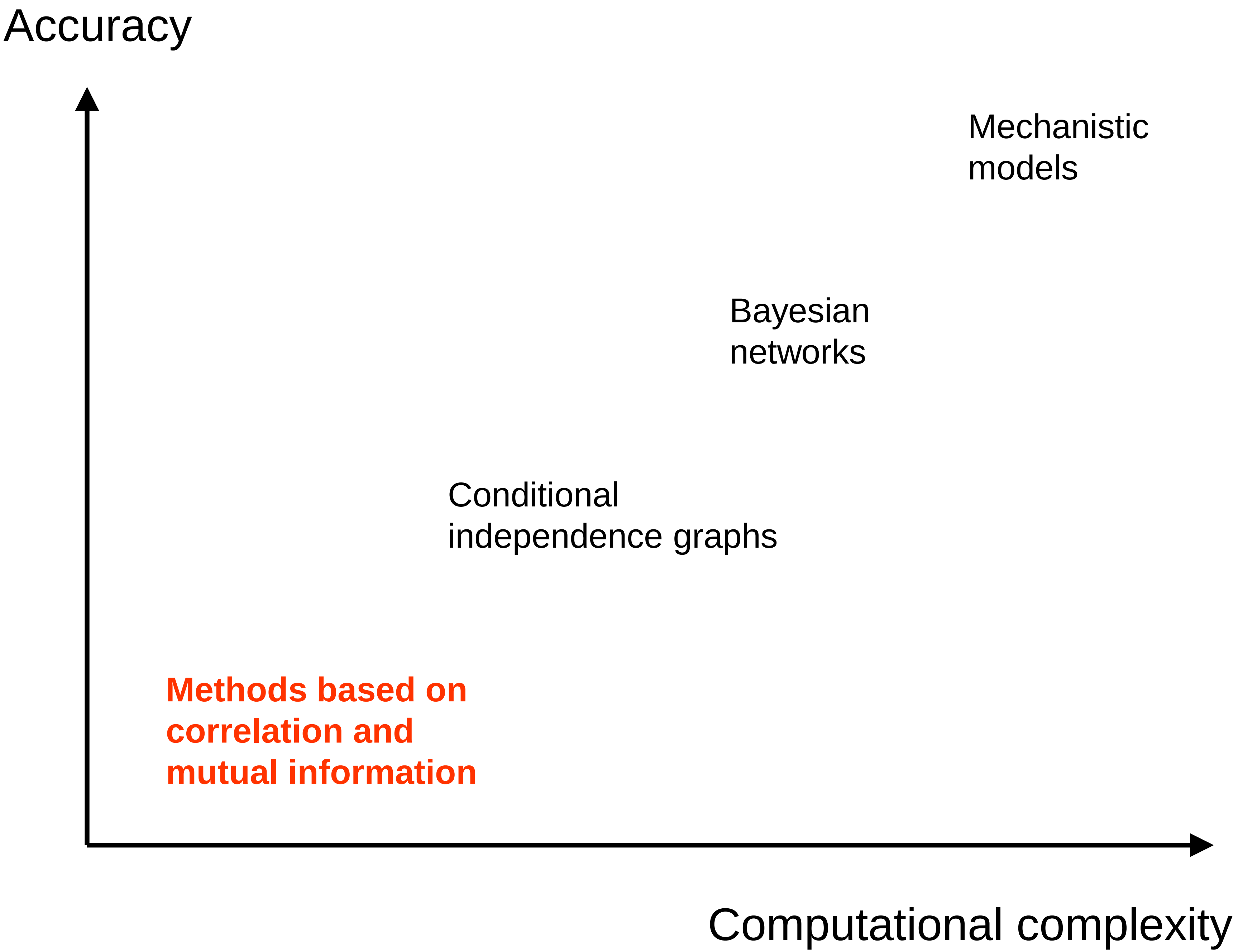
Mechanistic models

Bayesian networks

Conditional independence graphs

Methods based on correlation and mutual information

Computational complexity



Accuracy

Mechanistic models

Bayesian networks

Conditional independence graphs

Methods based on correlation and mutual information

Computational complexity

Pacific Symposium on Biocomputing 5:415-426 (2000)

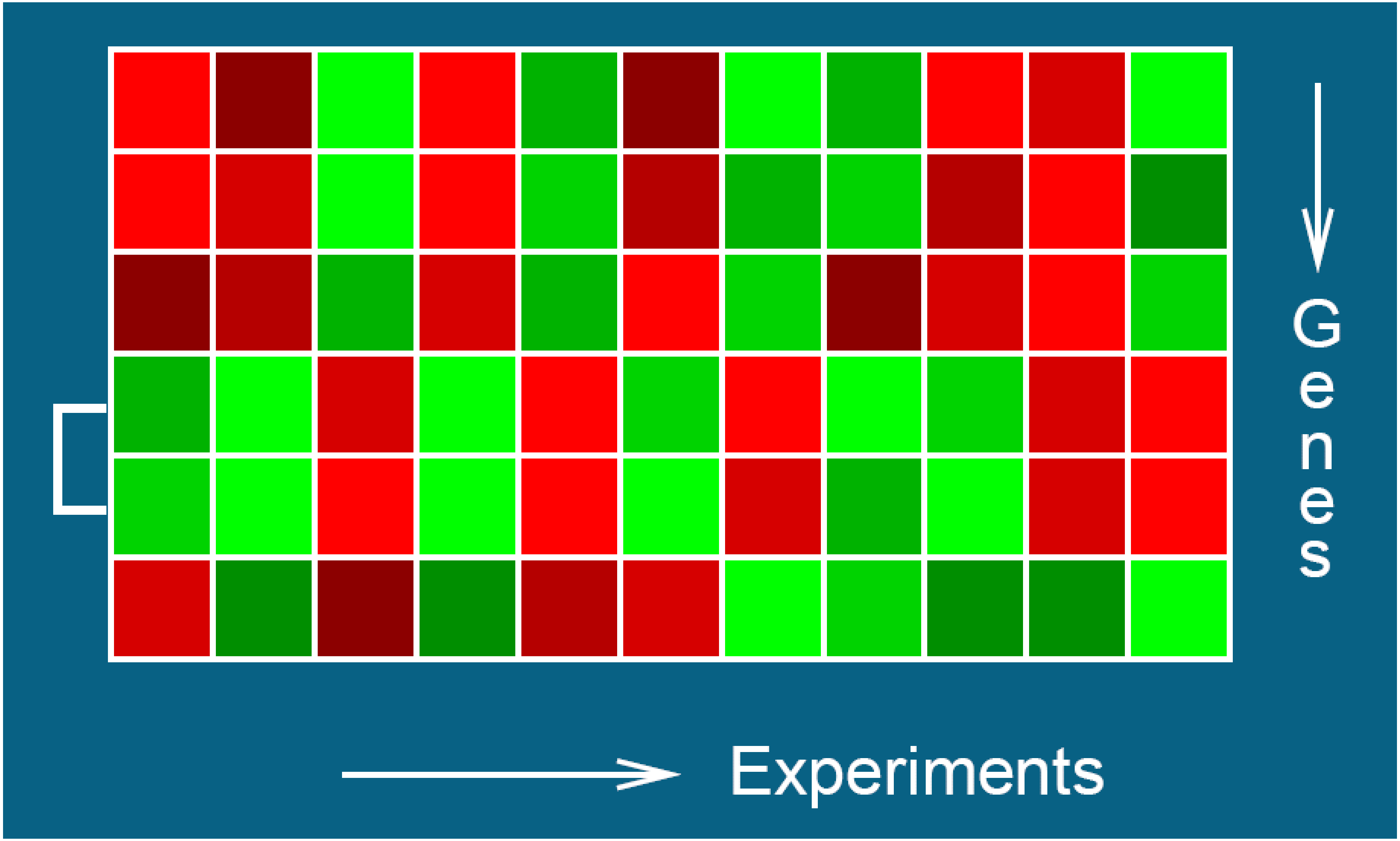
MUTUAL INFORMATION RELEVANCE NETWORKS: FUNCTIONAL GENOMIC CLUSTERING USING PAIRWISE ENTROPY MEASUREMENTS

A. J. BUTTE, I. S. KOHANE
Children's Hospital Informatics Program and
Division of Endocrinology,
300 Longwood Avenue,
Boston, MA 02115, USA

Relevance networks

(Butte and Kohane, 2000)

1. Choose a measure of association A(.,.)
2. Define a threshold value t_A
3. For all pairs of domain variables (X,Y) compute their association A(X,Y)
4. Connect those variables (X,Y) by an undirected edge whose association A(X,Y) exceeds the predefined threshold value t_A



Genes

Experiments

Association scores

$$\mathrm{corr}(x,y) = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2}\right)\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \overline{y})^2}\right)}$$

$$\mathrm{MI}(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{r} P(x=i, y=j)\log\frac{P(x=i, y=j)}{P(x=i)P(y=j)}$$

Association scores

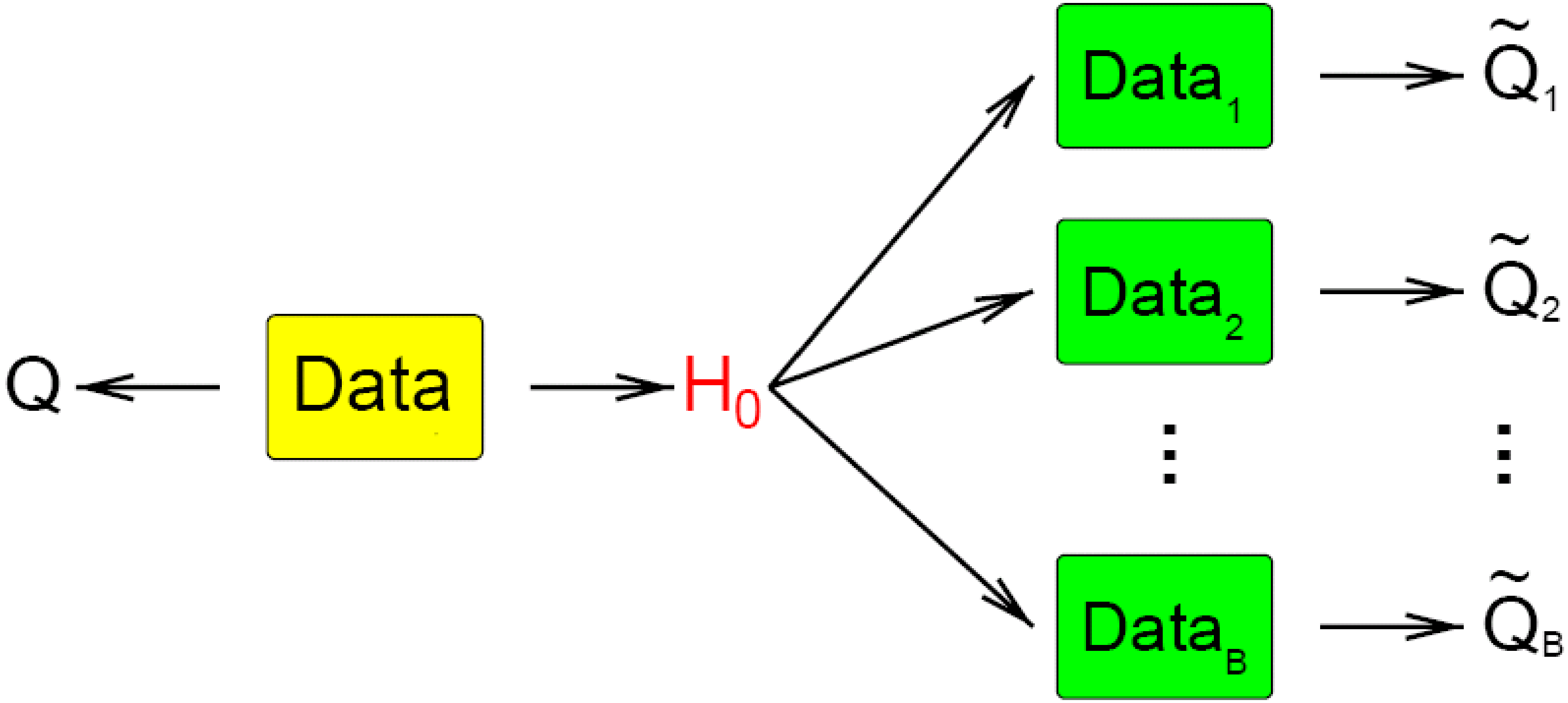
$$\mathrm{corr}(x,y) = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i-\overline{x})(y_i-\overline{y})}{\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i-\overline{x})^2}\right)\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i-\overline{y})^2}\right)}$$

$$\mathrm{MI}(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{r} P(x=i,y=j)\log\frac{P(x=i,y=j)}{P(x=i)P(y=j)}$$

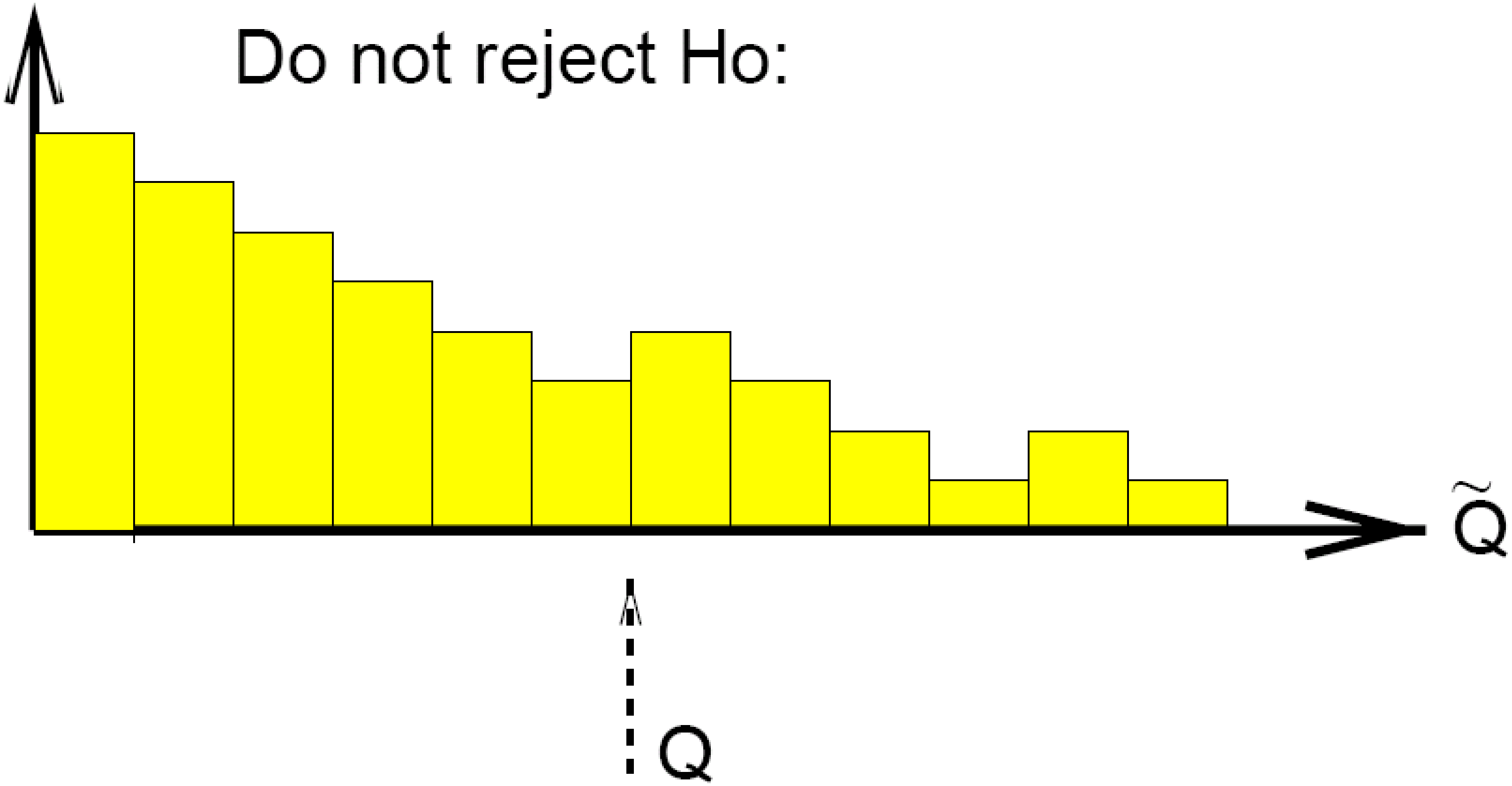
How to choose the threshold ?

→ Bootstrapping or randomization test

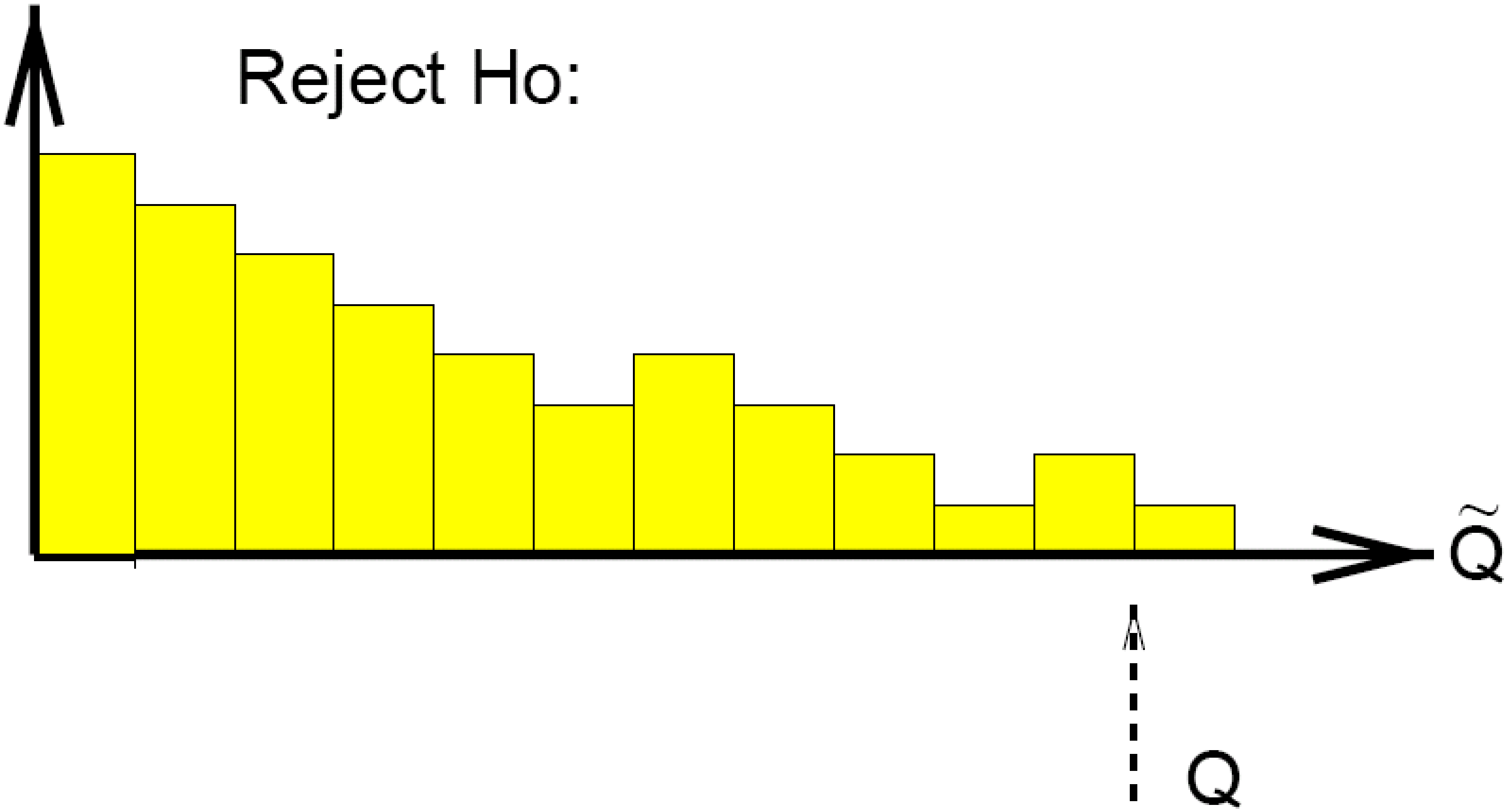
Frequentist statistics, hypothesis testing

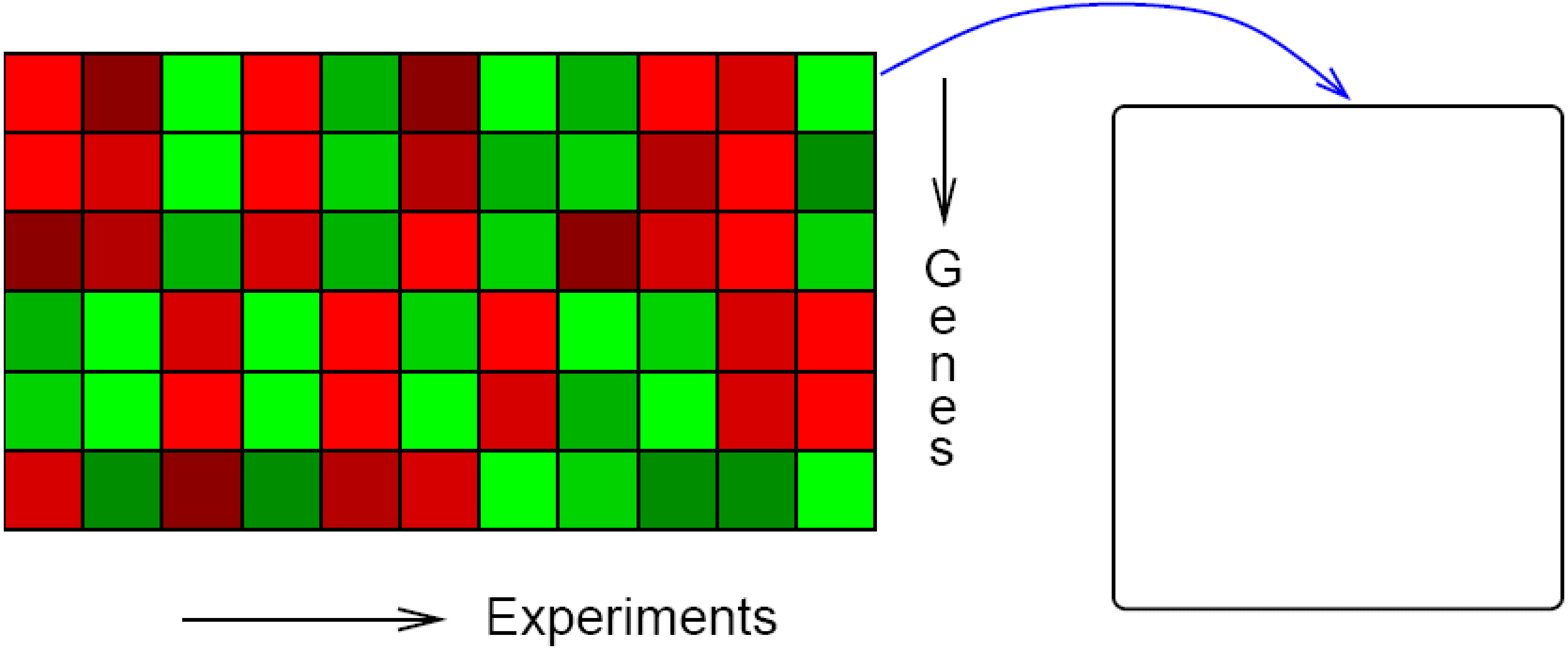


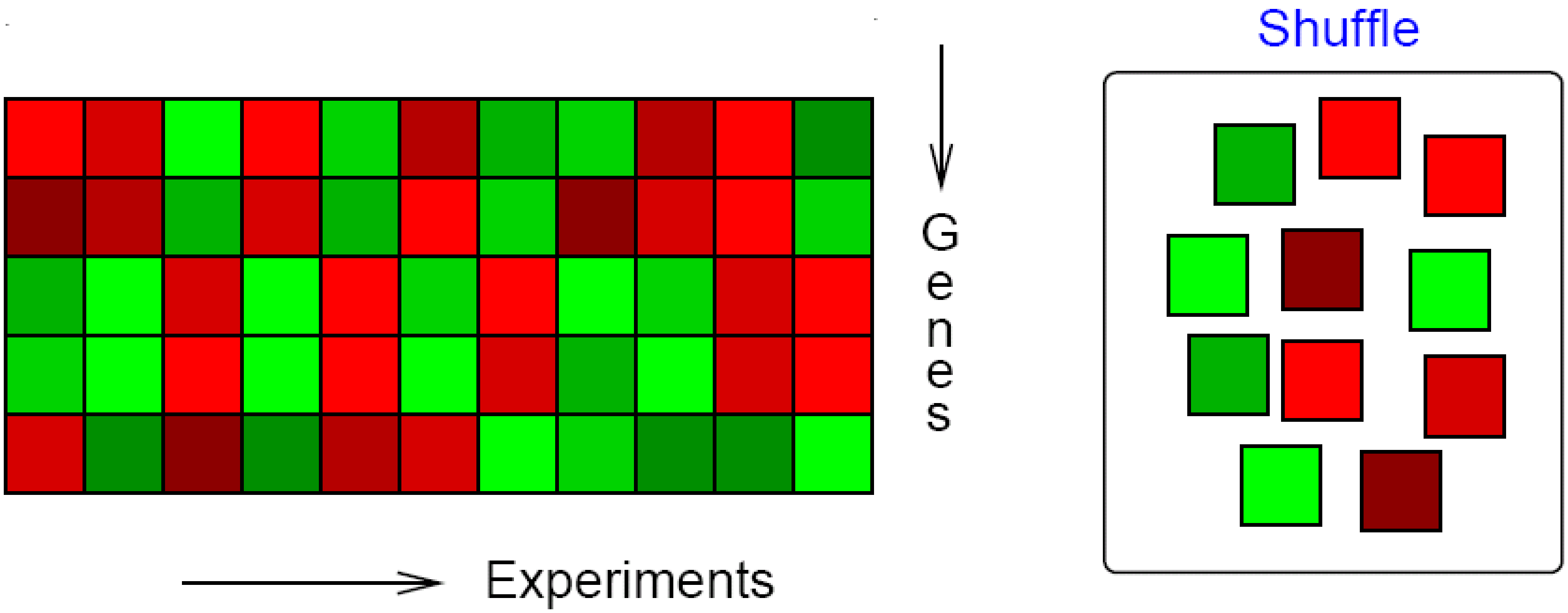
Result not significant: no interaction

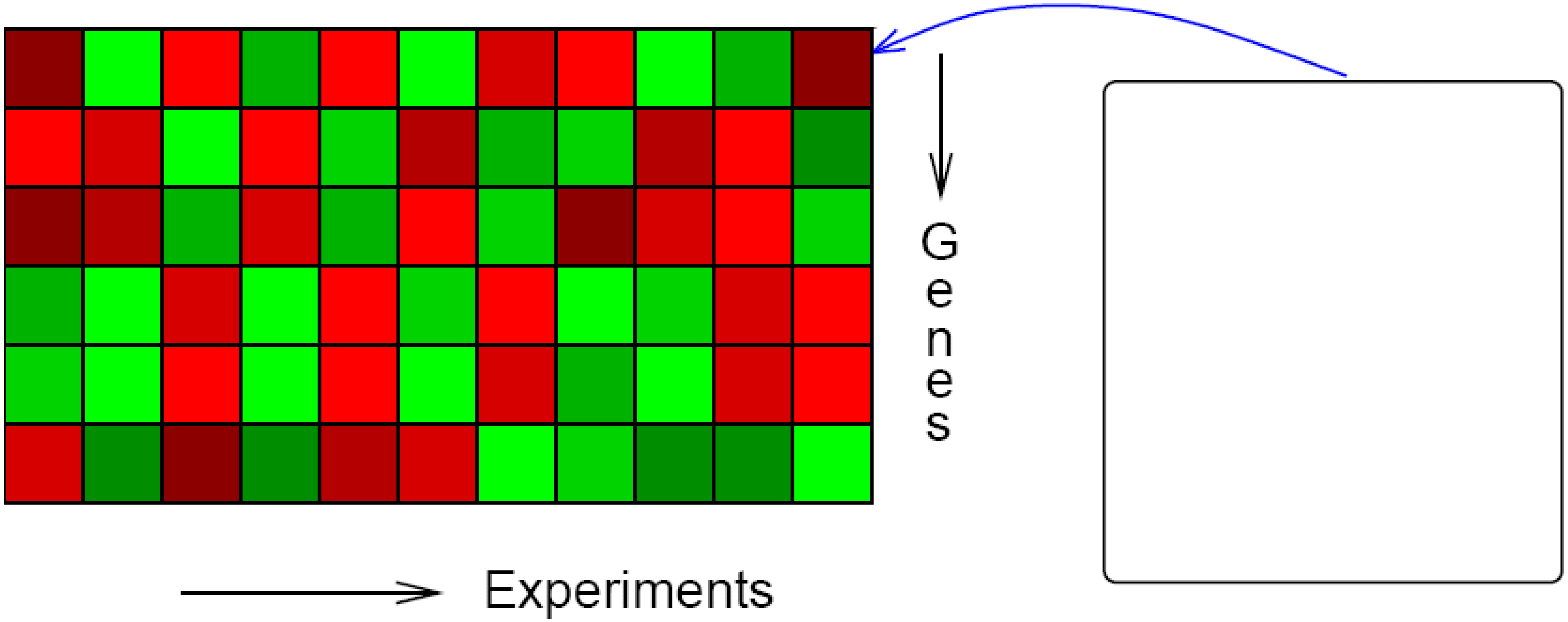


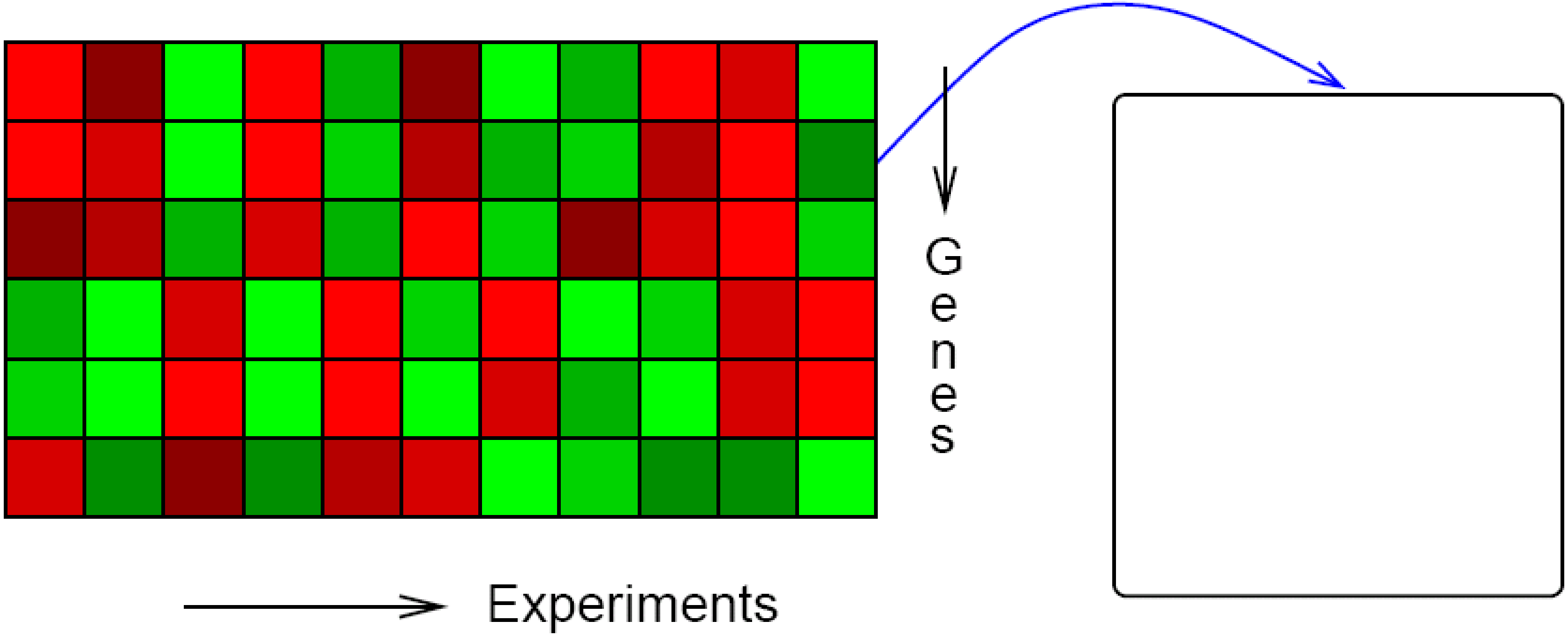
Significant interaction

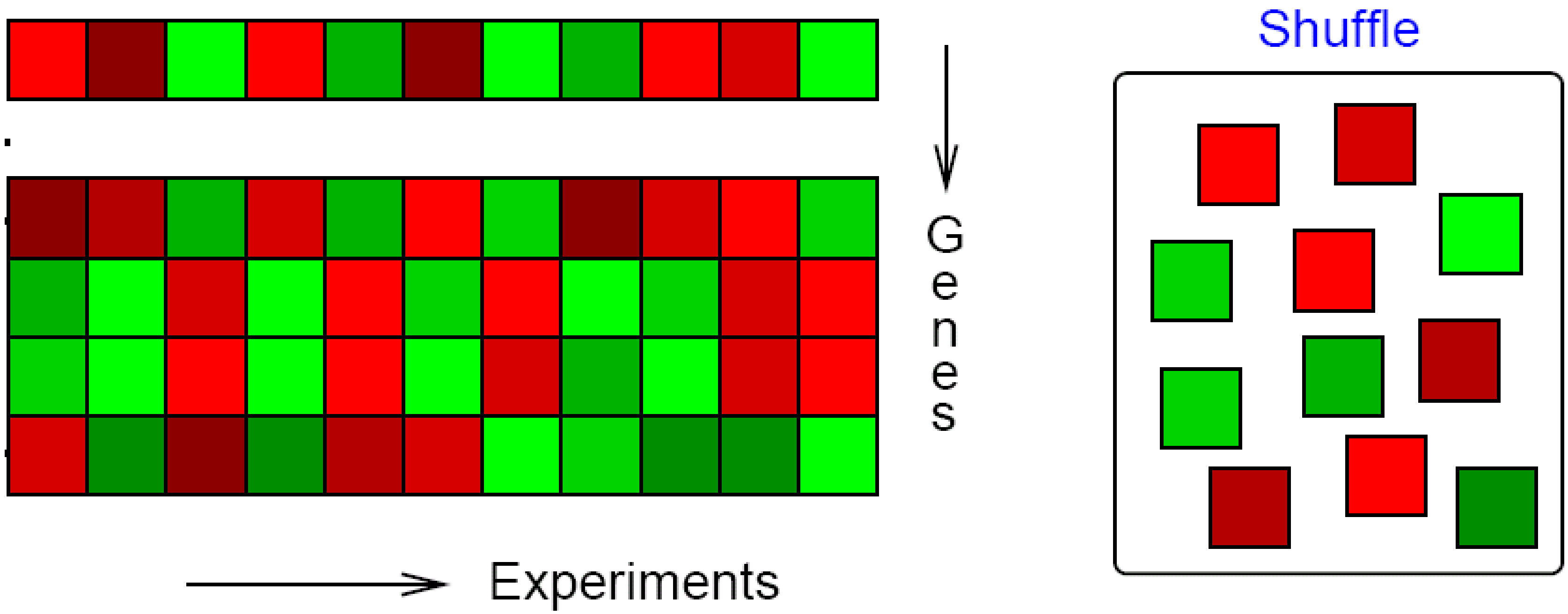



Genes
Experiments


Shuffle
Genes
Experiments


Genes
Experiments

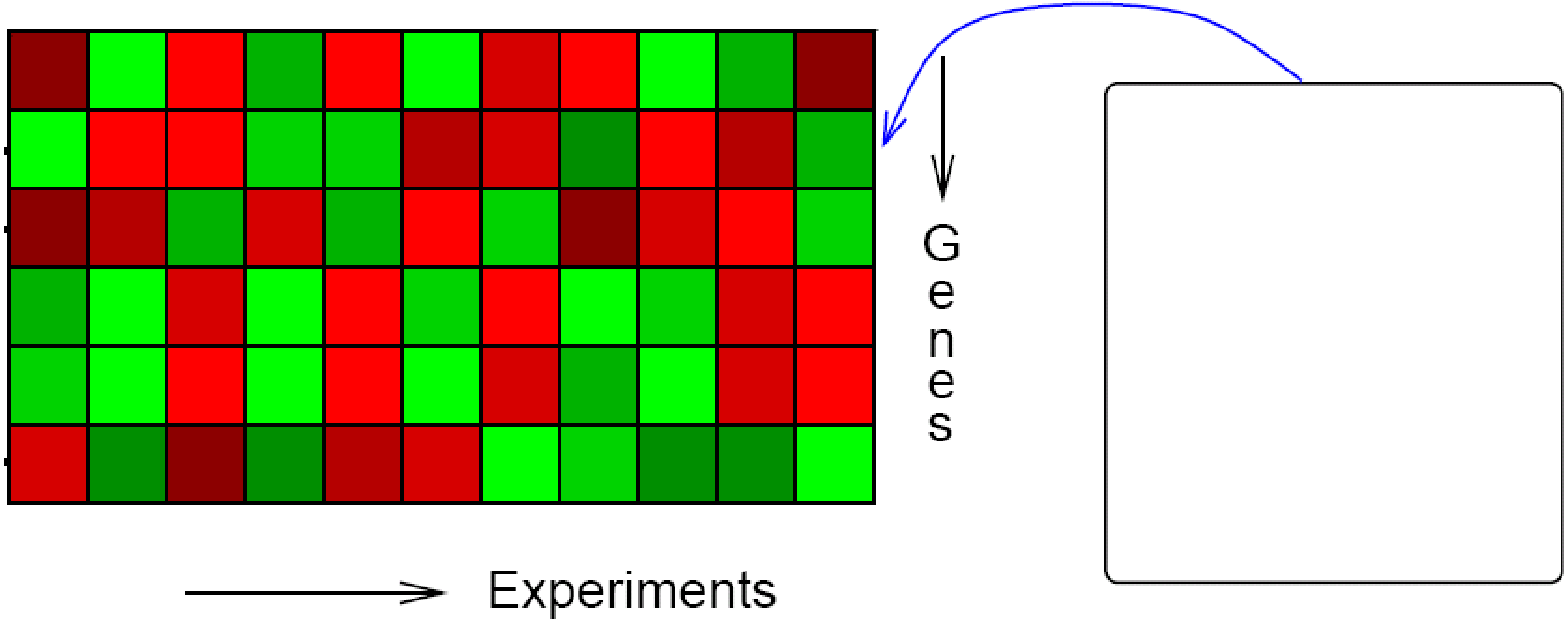

Genes
Experiments

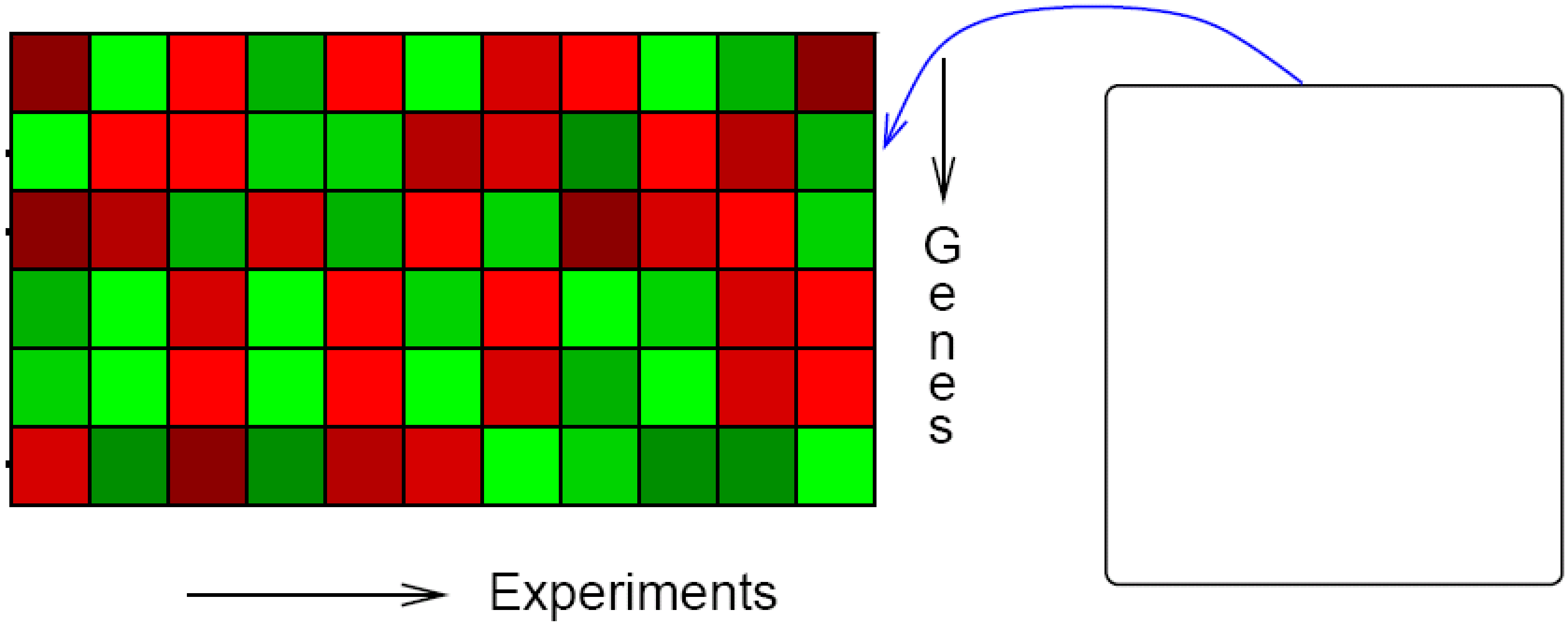


Shuffle

Genes

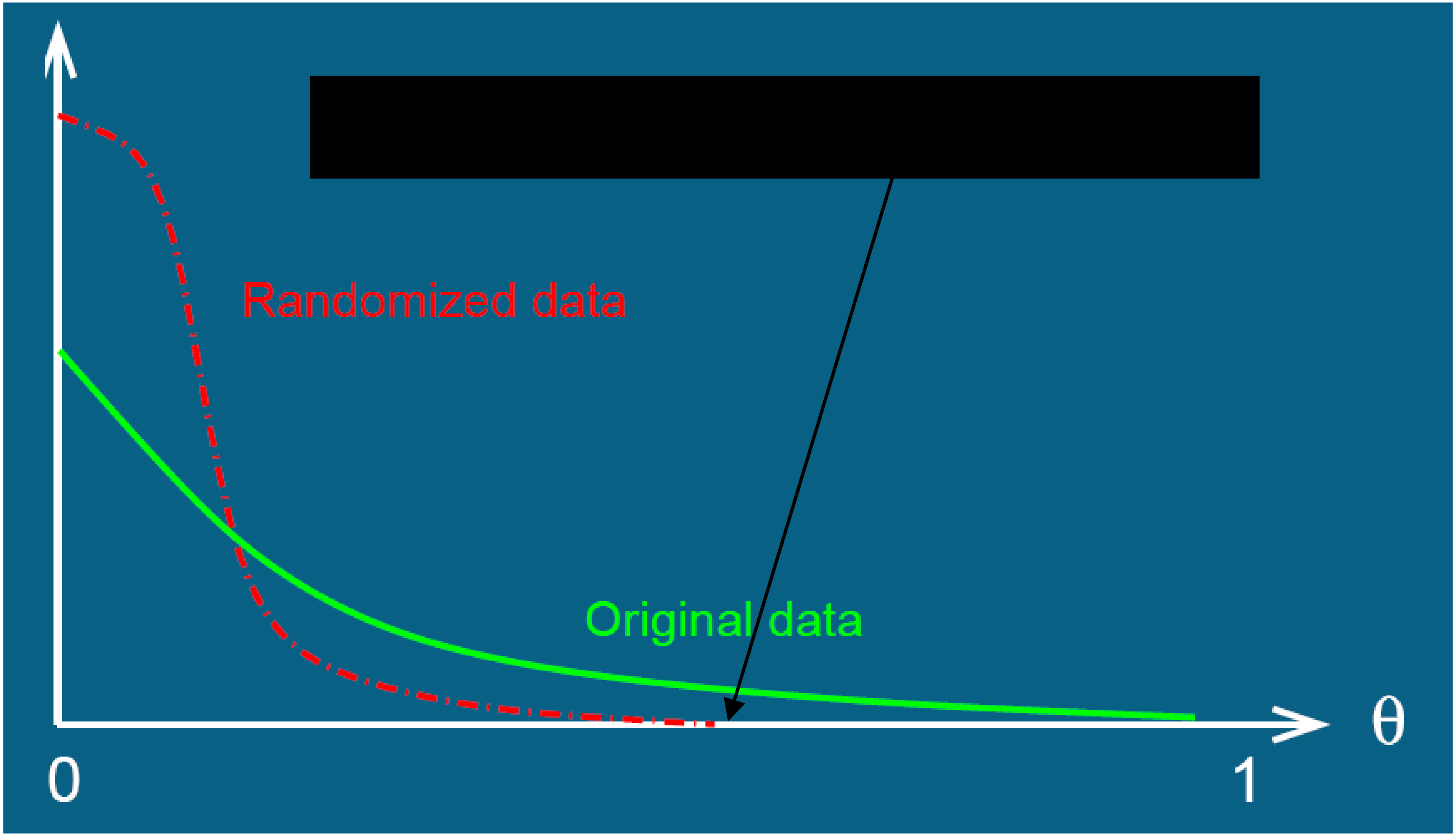
Experiments


Genes
Experiments



and so on …

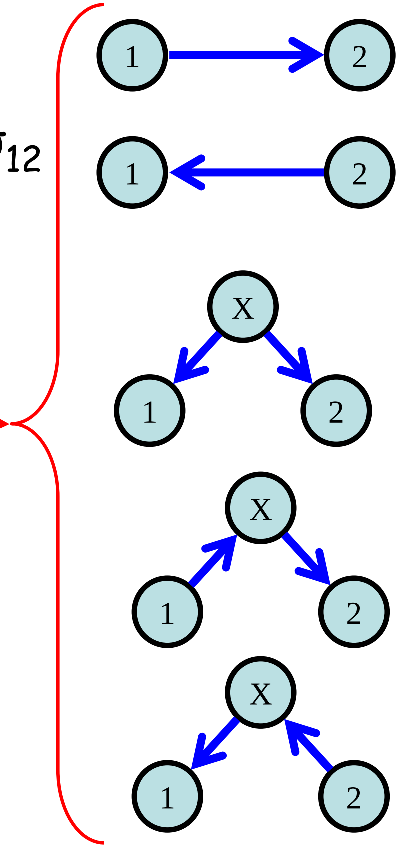
Number of edges with association score greater than θ



[redacted]

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_{nn} \end{pmatrix}$$

strong correlation σ_{12}



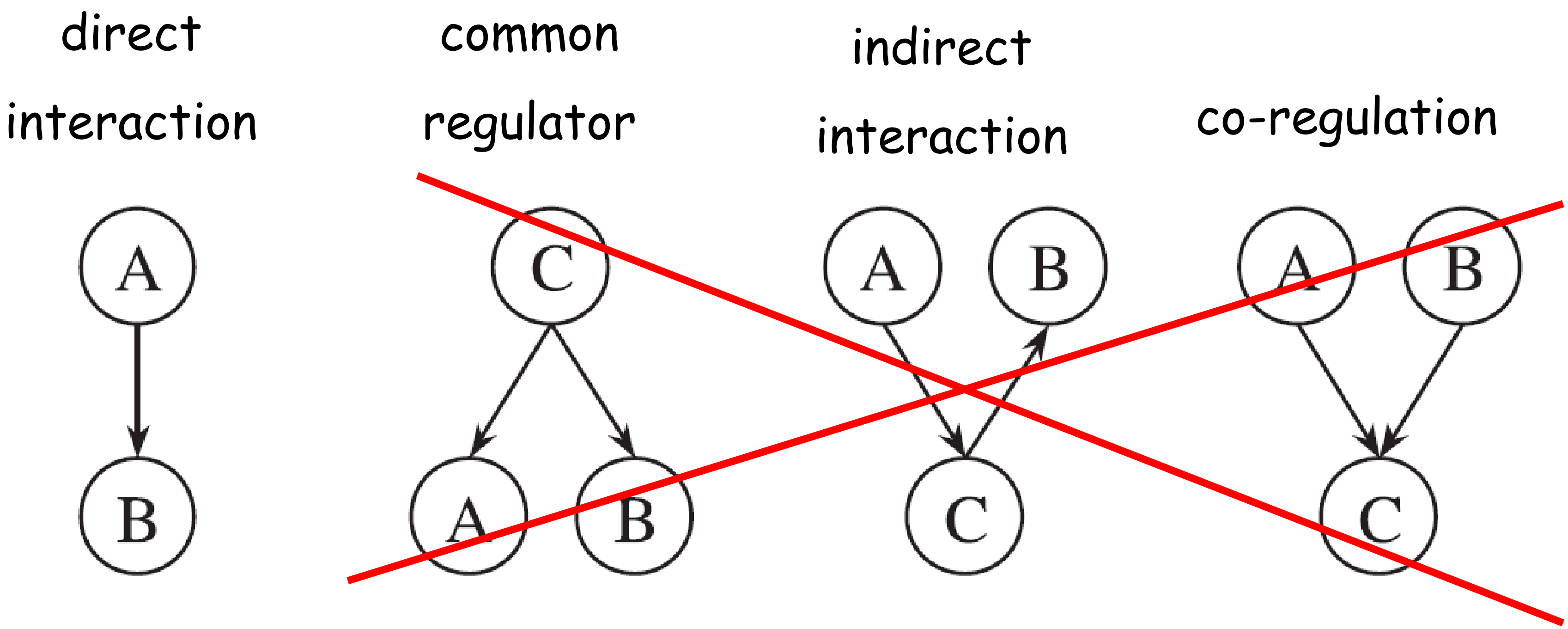
direct interaction

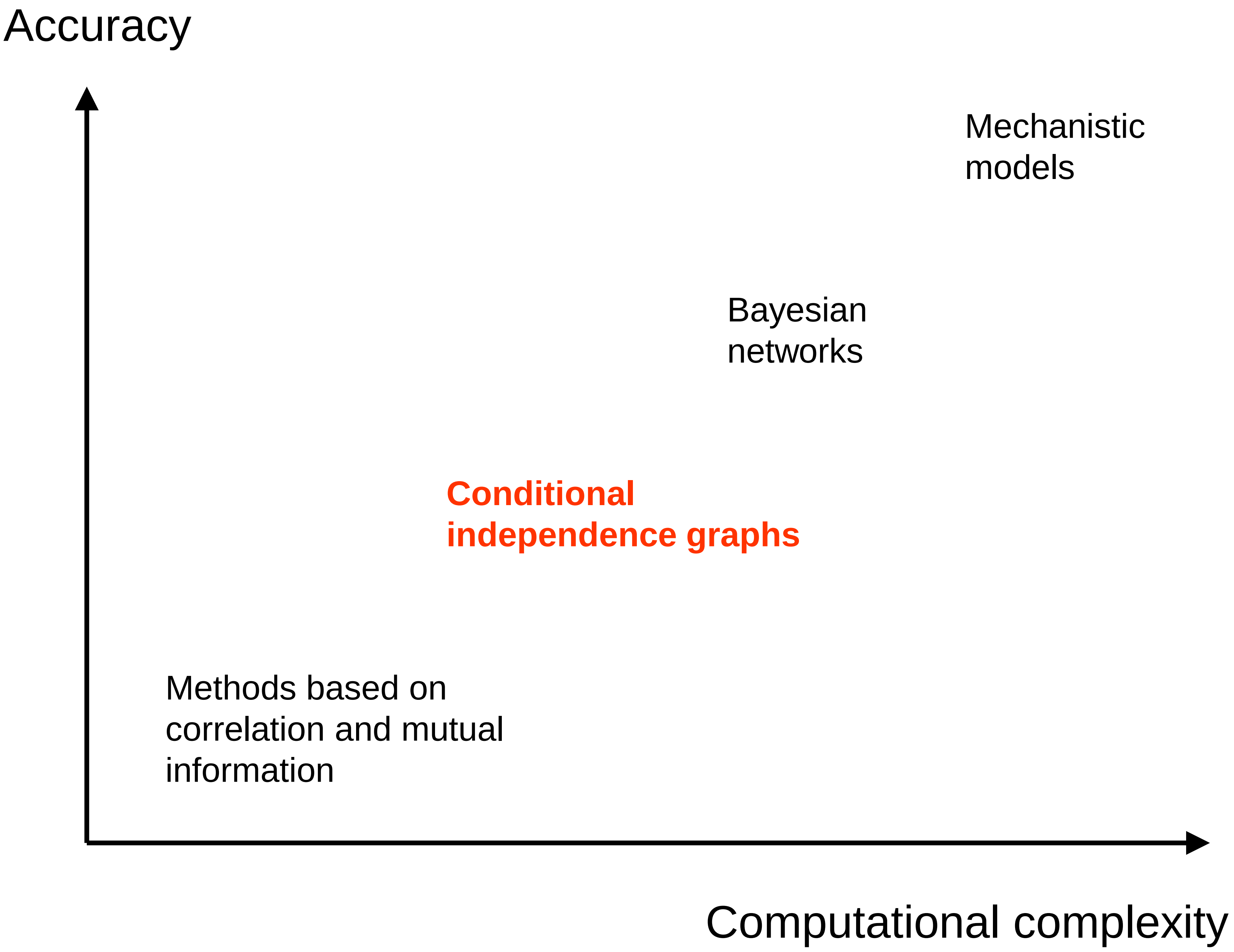
common regulator

indirect interaction

Shortcomings

Pairwise associations do not take the context of the system into consideration





Accuracy

Mechanistic models

Bayesian networks

Conditional independence graphs

Methods based on correlation and mutual information

Computational complexity

Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a \\ \mathbf{x}_b\end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}$$

Inverse of the co-variance matrix

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) =$$

$$-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).$$

the exponent in a general Gaussian distribution can be written

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

pick out all terms that are second order in $\mathbf{x}_a$

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a$$

from which we can immediately conclude that the covariance $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}.$$

Graphical Gaussian Models (GGMs)

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \cdots & \pi_{1n} \\ \pi_{21} & \pi_{22} & \pi_{23} & \cdots & \pi_{2n} \\ \pi_{31} & \pi_{32} & \pi_{33} & \cdots & \pi_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_{n1} & \pi_{n2} & \pi_{n3} & \cdots & \pi_{nn} \end{pmatrix}$$

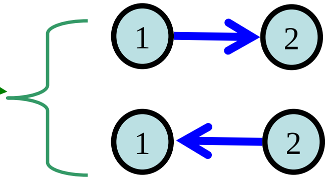
$$\pi_{ij} = \frac{-1 \cdot (\Sigma^{-1})_{ij}}{\sqrt{(\Sigma^{-1})_{ii} \cdot (\Sigma^{-1})_{jj}}}$$

Inverse of the covariance matrix

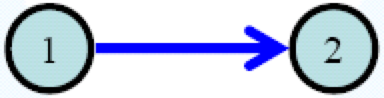
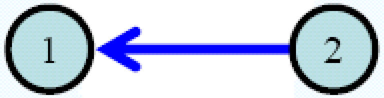
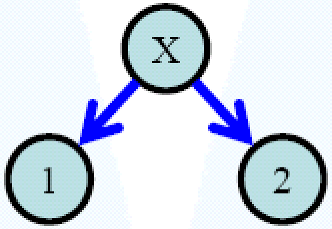
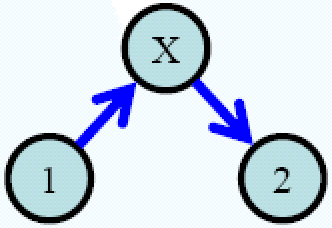
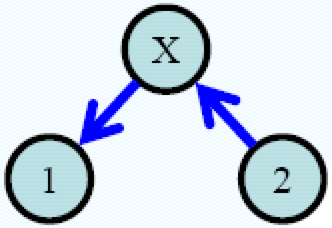
strong partial correlation π_{12}

Partial correlation, i.e. correlation conditional on all other domain variables

$Corr(X_1,X_2|X_3,\ldots,X_n)$



Direct interaction

	Correlation	Partial correlation
	high	high
	high	high
	high	low
	high	low
	high	low

Graphical Gaussian Models (GGMs)

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \cdots & \pi_{1n} \\ \pi_{21} & \pi_{22} & \pi_{23} & \cdots & \pi_{2n} \\ \pi_{31} & \pi_{32} & \pi_{33} & \cdots & \pi_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_{n1} & \pi_{n2} & \pi_{n3} & \cdots & \pi_{nn} \end{pmatrix}$$

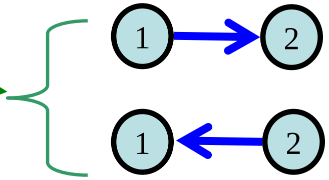
$$\pi_{ij} = \frac{-1 \cdot (\Sigma^{-1})_{ij}}{\sqrt{(\Sigma^{-1})_{ii} \cdot (\Sigma^{-1})_{jj}}}$$

Inverse of the covariance matrix

strong partial correlation π_{12}

Partial correlation, i.e. correlation conditional on all other domain variables

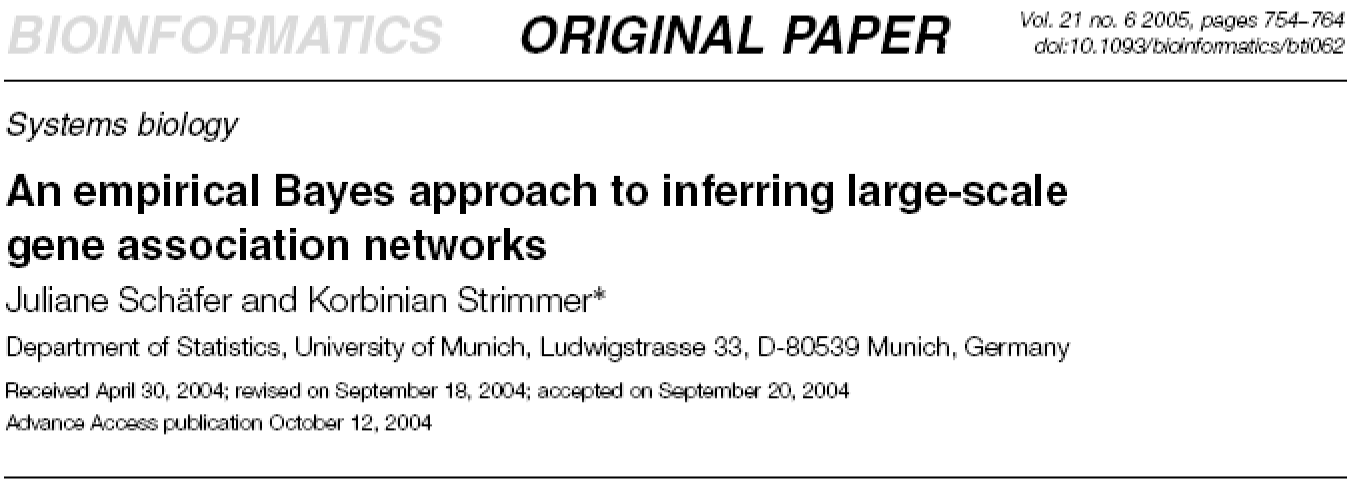
$\mathrm{Corr}(X_1, X_2 | X_3, \ldots, X_n)$



Direct interaction

Problem: #observations < #variables

➔ Covariance matrix is singular



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 21 no. 6 2005, pages 754–764
doi:10.1093/bioinformatics/bti062

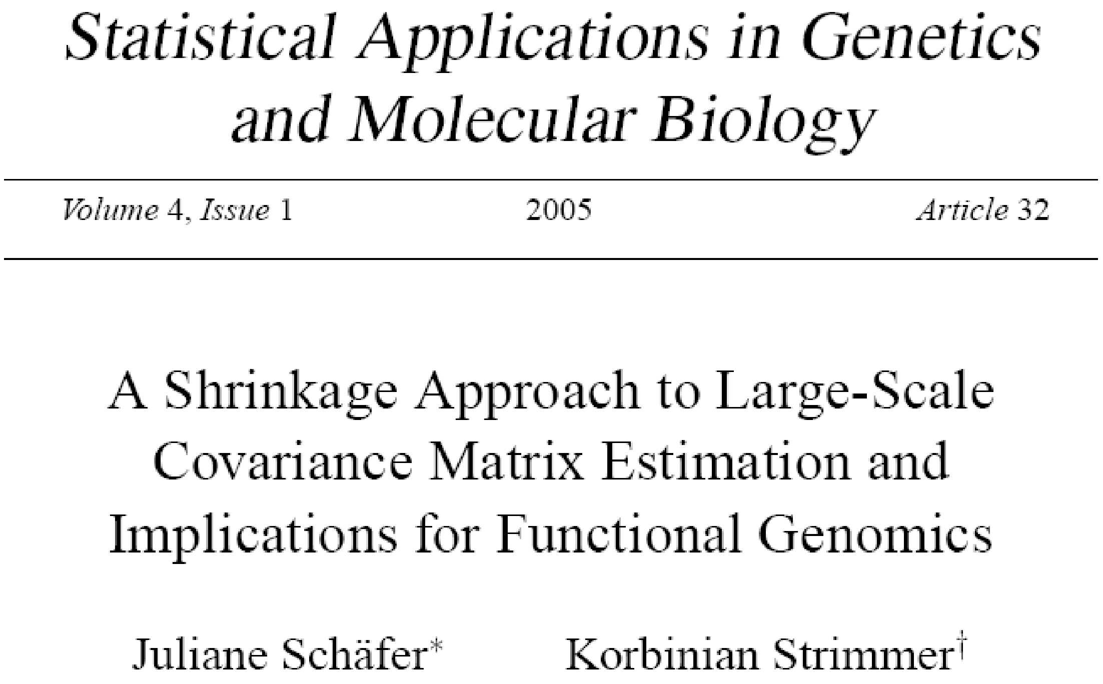
Systems biology

An empirical Bayes approach to inferring large-scale gene association networks

Juliane Schäfer and Korbinian Strimmer*

Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany

Received April 30, 2004; revised on September 18, 2004; accepted on September 20, 2004
Advance Access publication October 12, 2004



Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 32

A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics

Juliane Schäfer* Korbinian Strimmer†

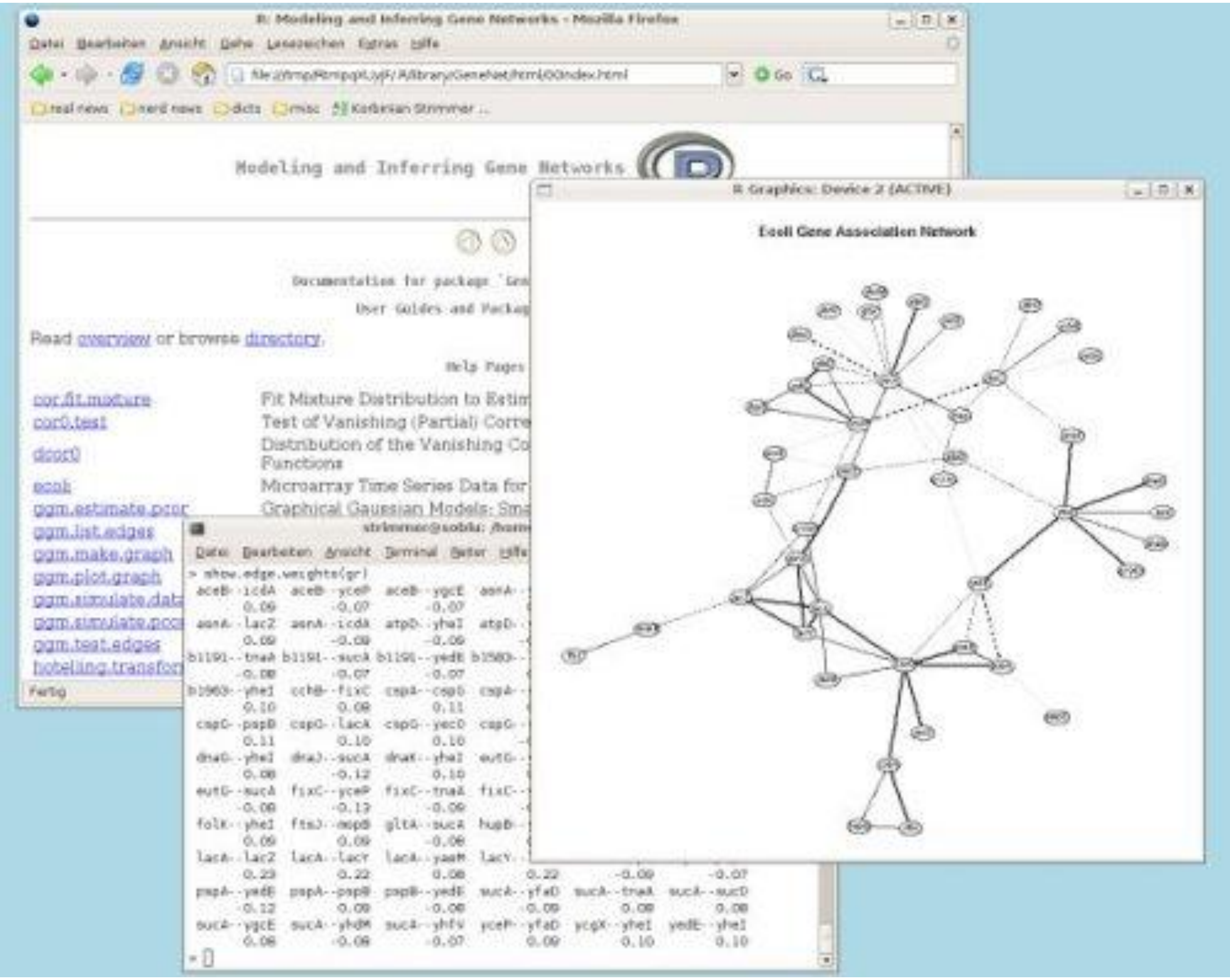
Summary of the GGM algorithm, part 1

- Partial correlations, as opposed to standard correlations, capture the influence of the whole system. Mathematically, this is the correlation between two nodes conditional on the rest of the system.
- The partial correlations can be computed from the inverse of the covariance matrix.
- The true covariance matrix is usually unknown → approximated by the empirical covariance matrix, estimated from the data.
- Empirical covariance matrix → over-fitting problem, can be ill-conditioned or rank-deficient (singular) → inversion impossible.
- Regularization: add the identity matrix, weighted by some constant, to the empirical covariance matrix → matrix non-singular. Possible problem: bias.

Summary of the GGM algorithm, part 2

- Set the trade-off parameter so as to minimize the expected difference between the (unknown) true covariance matrix and the estimated matrix.
- Statistical decision theory: closed-from expression for the optimal trade-off parameter (Ledoit-Wolf lemma).
- Catch: this expression depends on expectation values with respect to other data sets generated from the same processes. Cannot be computed in practice.
- Heuristics: replace expectation values by the actually observed values.

GeneNet (Strimmer et al.)



R: Modeling and Inferring Gene Networks - Mozilla Firefox

Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
file:///tmp/RtmppKJyjF/.R/library/GeneNet/html/00Index.html
Go
real news
nerd news
dicts
misc
Korbinian Strimmer ...

Modeling and Inferring Gene Networks

Documentation for package `Gen

User Guides and Packag

Read overview or browse directory.

Help Pages

cor.fit.mixture Fit Mixture Distribution to Estim

cor0.test Test of Vanishing (Partial) Corre

dcor0 Distribution of the Vanishing Co Functions

ecoli Microarray Time Series Data for

ggm.estimate.pcor Graphical Gaussian Models: Sma

ggm.list.edges

ggm.make.graph

ggm.plot.graph

ggm.simulate.data

ggm.simulate.pcor

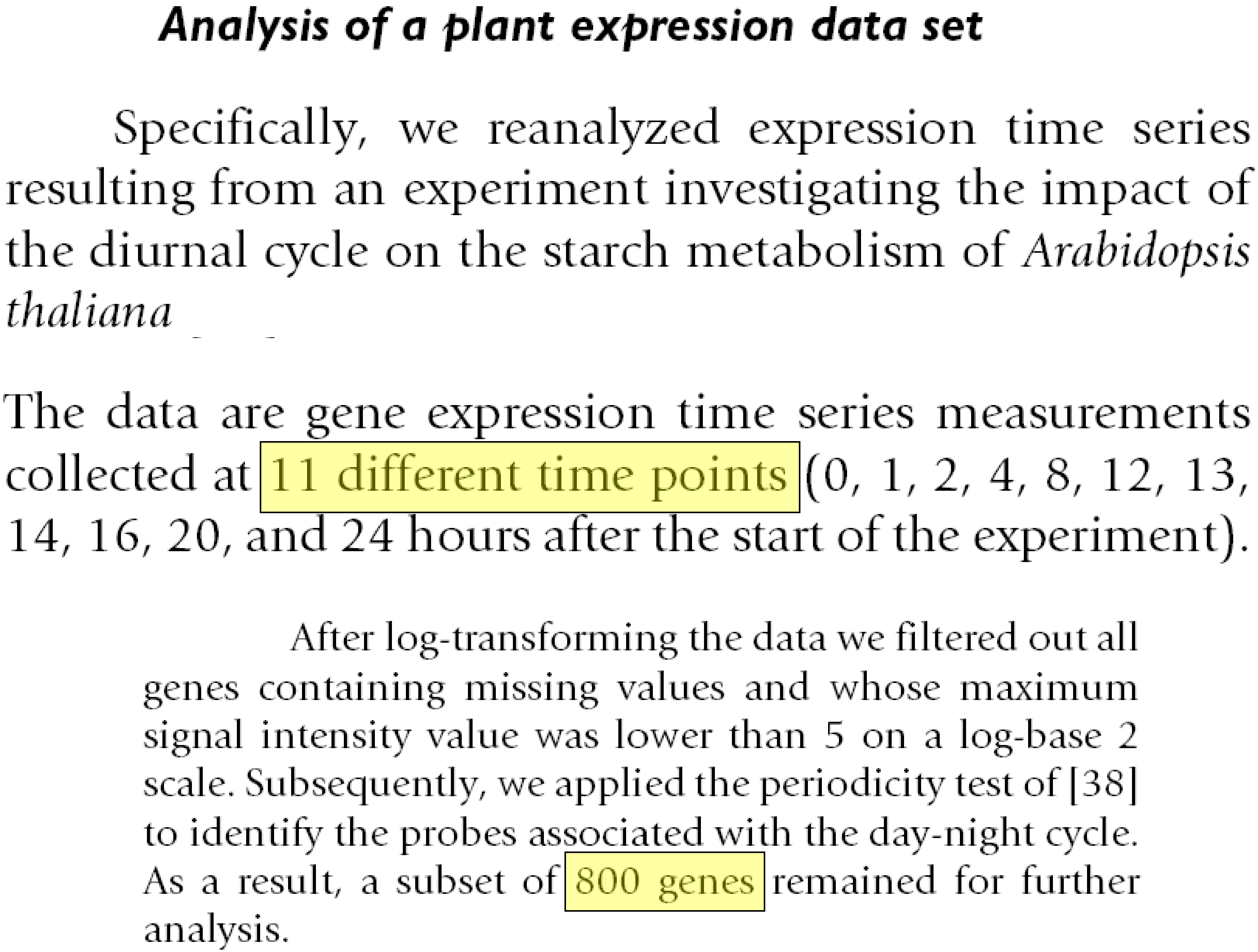
ggm.test.edges

hotelling.transform

```
> show.edge.weights(gr)
 aceB--icdA  aceB--yceP  aceB--ygcE  aenA--
       0.09       -0.07       -0.07
 aenA--lacZ  aenA--icdA  atpD--yheI  atpD--
       0.09       -0.08       -0.06
b1191--tnaA b1191--sucA b1191--yedE b1583--
      -0.08       -0.07       -0.07
b1963--yheI  cchB--fixC  cspA--cspG  cspA--
       0.10        0.08        0.11
 cspG--pspB  cspG--lacA  cspG--yecO  cspG--
       0.11        0.10        0.10
 dnaG--yheI  dnaJ--sucA  dnaK--yheI  eutG--
       0.08       -0.12        0.10
 eutG--sucA  fixC--yceP  fixC--tnaA  fixC--
      -0.08       -0.13       -0.09
 folK--yheI  ftsJ--mopB  gltA--sucA  hupB--
       0.09        0.09       -0.08
 lacA--lacZ  lacA--lacY  lacA--yaeM  lacY--
       0.23        0.22        0.08        0.22       -0.09       -0.07
 pspA--yedE  pspA--pspB  pspB--yedE  sucA--yfaD  sucA--tnaA  sucA--sucD
      -0.12        0.09       -0.08        0.09        0.08        0.08
 sucA--ygcE  sucA--yhdM  sucA--yhfV  yceP--yfaD  ycgX--yheI  yedE--yheI
       0.08       -0.08       -0.07        0.09        0.10        0.10
> 
```

Availble from http://strimmerlab.org/software/genenet/

Application in Schaefer & Strimmer (text copied form their paper)



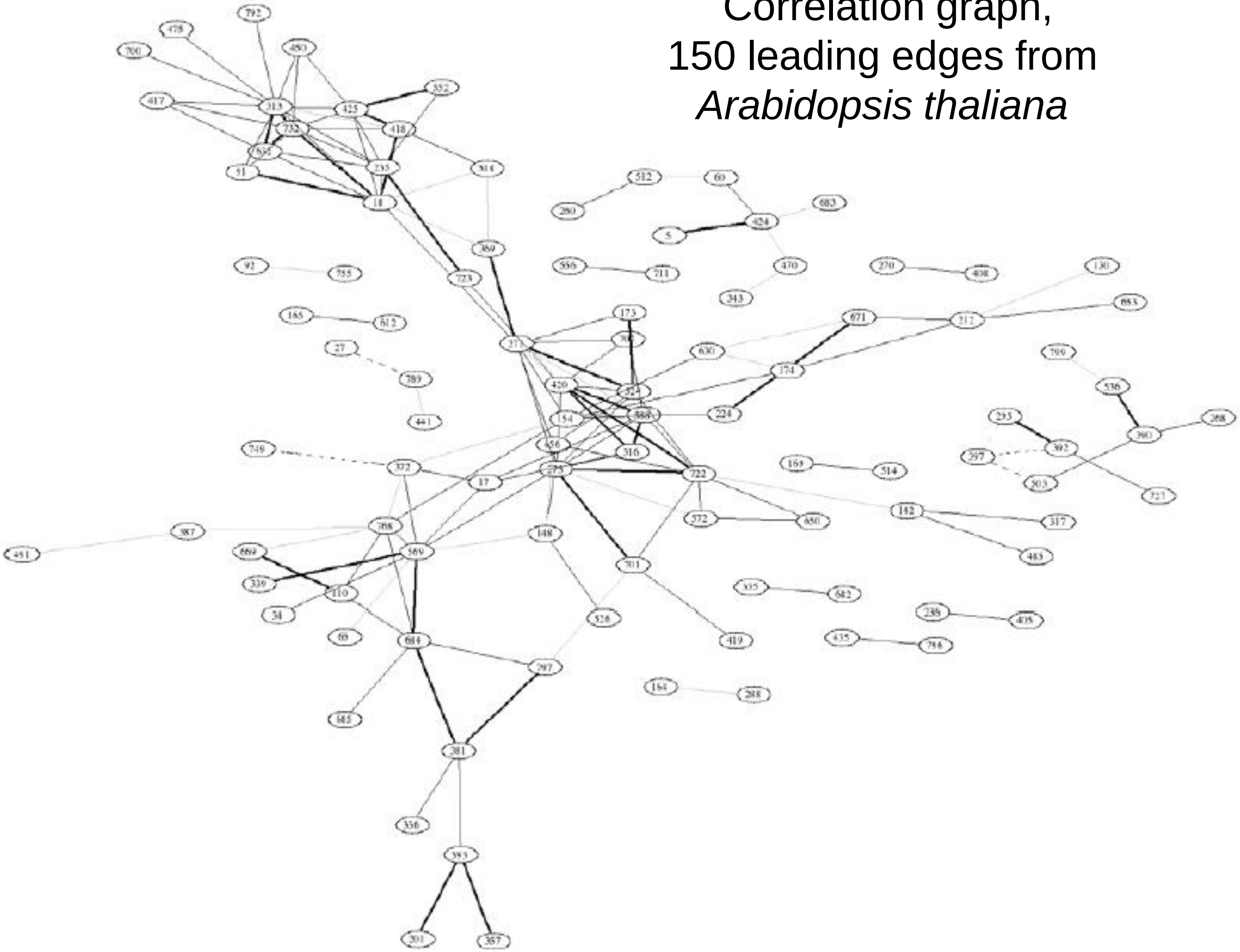
Analysis of a plant expression data set

Specifically, we reanalyzed expression time series resulting from an experiment investigating the impact of the diurnal cycle on the starch metabolism of *Arabidopsis thaliana*

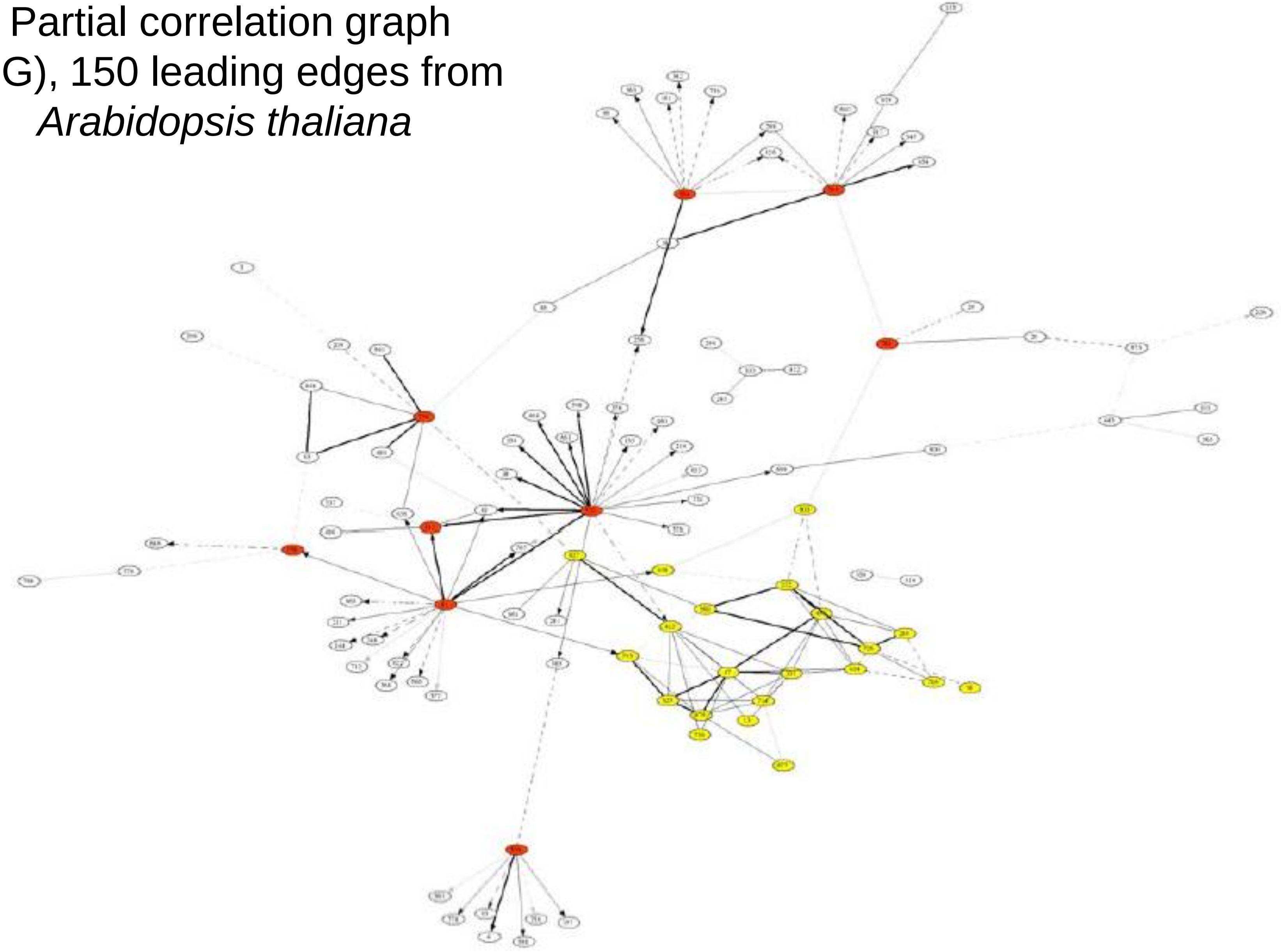
The data are gene expression time series measurements collected at 11 different time points (0, 1, 2, 4, 8, 12, 13, 14, 16, 20, and 24 hours after the start of the experiment).

After log-transforming the data we filtered out all genes containing missing values and whose maximum signal intensity value was lower than 5 on a log-base 2 scale. Subsequently, we applied the periodicity test of [38] to identify the probes associated with the day-night cycle. As a result, a subset of 800 genes remained for further analysis.

Correlation graph, 150 leading edges from *Arabidopsis thaliana*

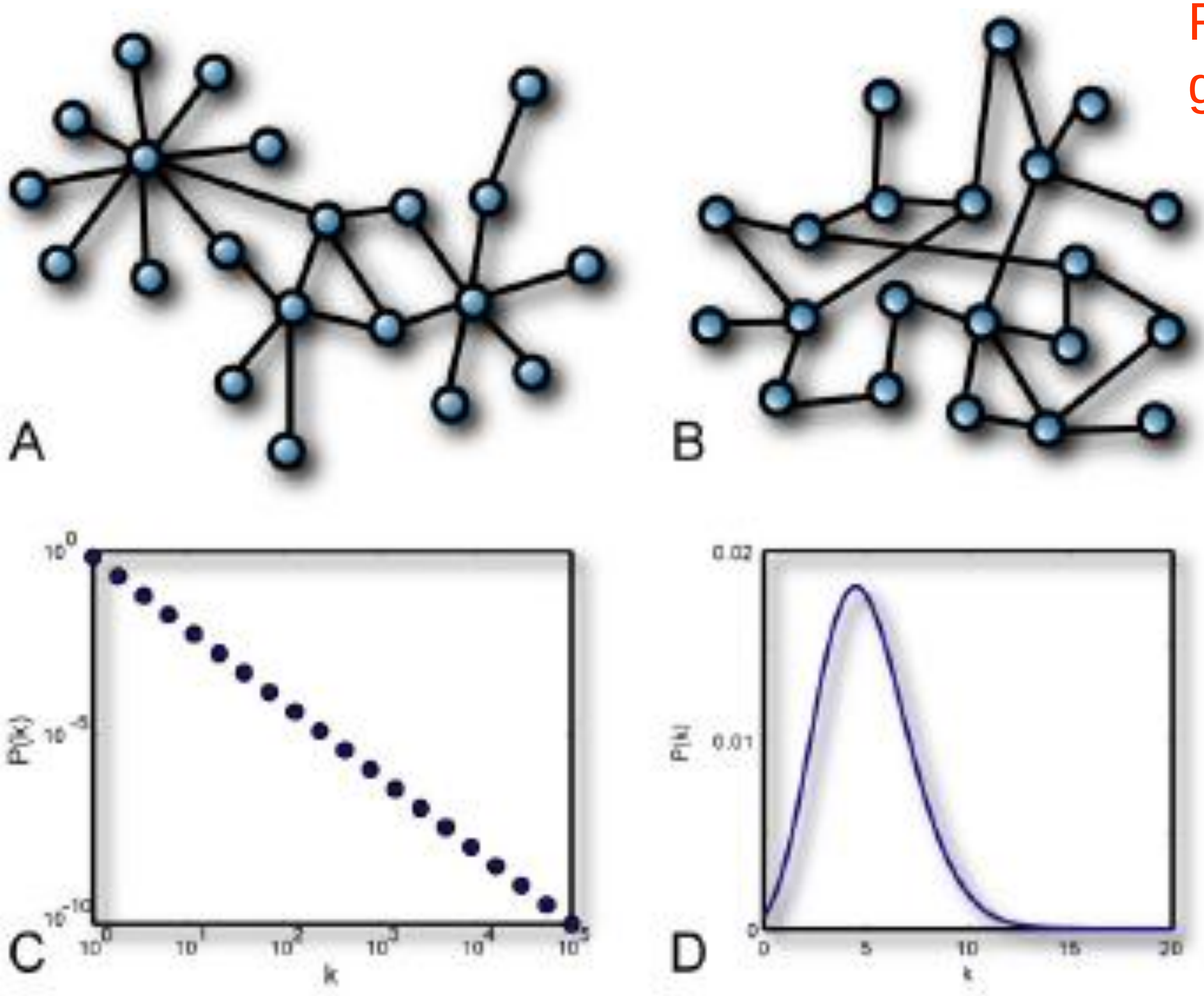


Partial correlation graph (CIG), 150 leading edges from *Arabidopsis thaliana*



Degree distribution and power law

Power law



Random graph

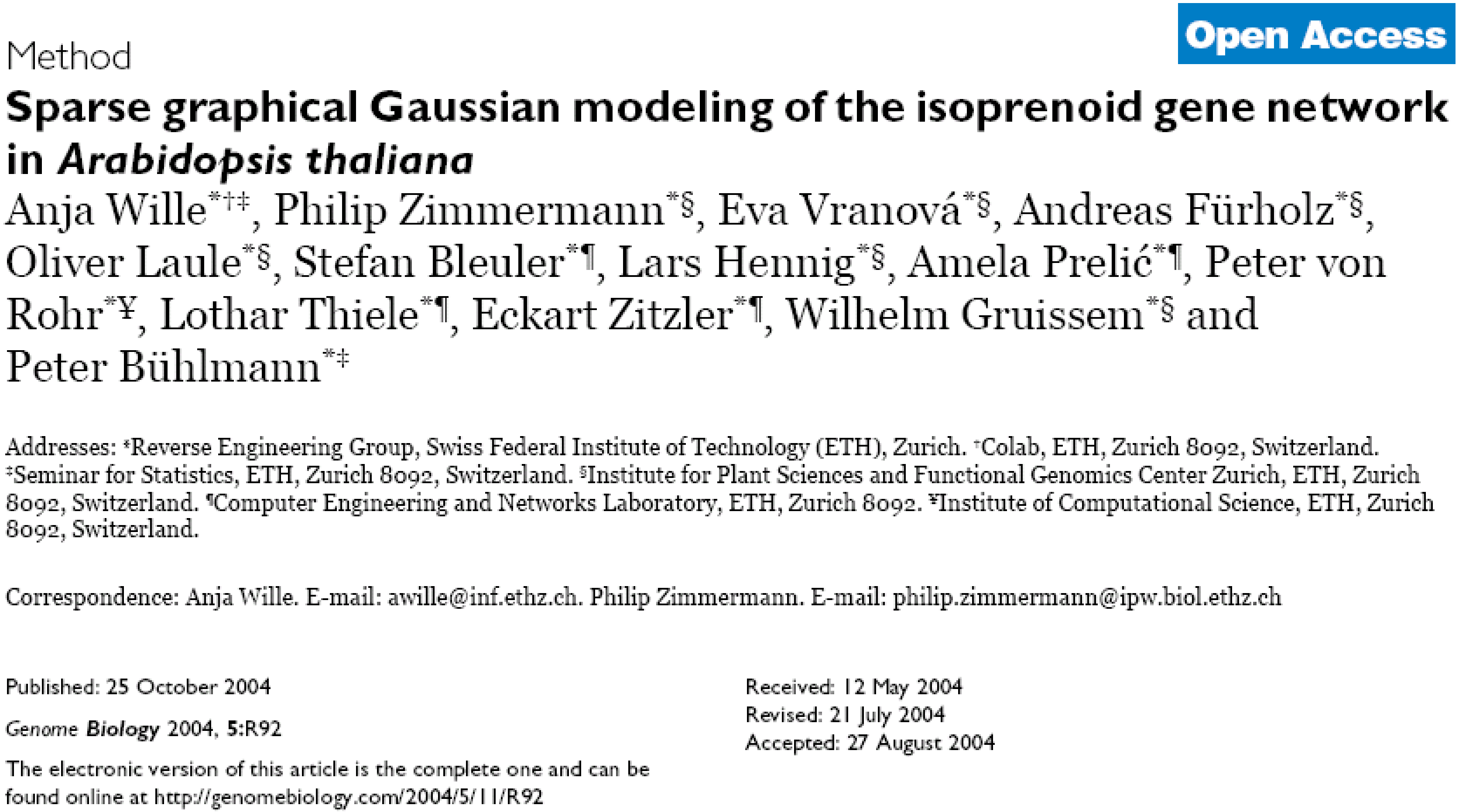
Log P(k)

P(k)

Log k

k

From https://nwb.slis.indiana.edu/community/uploads/CustomFillings/1.jpg



Method

Open Access

Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*

Anja Wille[*†‡], Philip Zimmermann[*§], Eva Vranová[*§], Andreas Fürholz[*§], Oliver Laule[*§], Stefan Bleuler[*¶], Lars Hennig[*§], Amela Prelić[*¶], Peter von Rohr[*¥], Lothar Thiele[*¶], Eckart Zitzler[*¶], Wilhelm Gruissem[*§] and Peter Bühlmann[*‡]

Addresses: [*]Reverse Engineering Group, Swiss Federal Institute of Technology (ETH), Zurich. [†]Colab, ETH, Zurich 8092, Switzerland. [‡]Seminar for Statistics, ETH, Zurich 8092, Switzerland. [§]Institute for Plant Sciences and Functional Genomics Center Zurich, ETH, Zurich 8092, Switzerland. [¶]Computer Engineering and Networks Laboratory, ETH, Zurich 8092. [¥]Institute of Computational Science, ETH, Zurich 8092, Switzerland.

Correspondence: Anja Wille. E-mail: awille@inf.ethz.ch. Philip Zimmermann. E-mail: philip.zimmermann@ipw.biol.ethz.ch

Published: 25 October 2004

Genome **Biology** 2004, **5**:R92

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/11/R92

Received: 12 May 2004
Revised: 21 July 2004
Accepted: 27 August 2004

Genome **Biology** 2004, **5**:R92

Crosstalk between two metabolic pathways, from microarray data

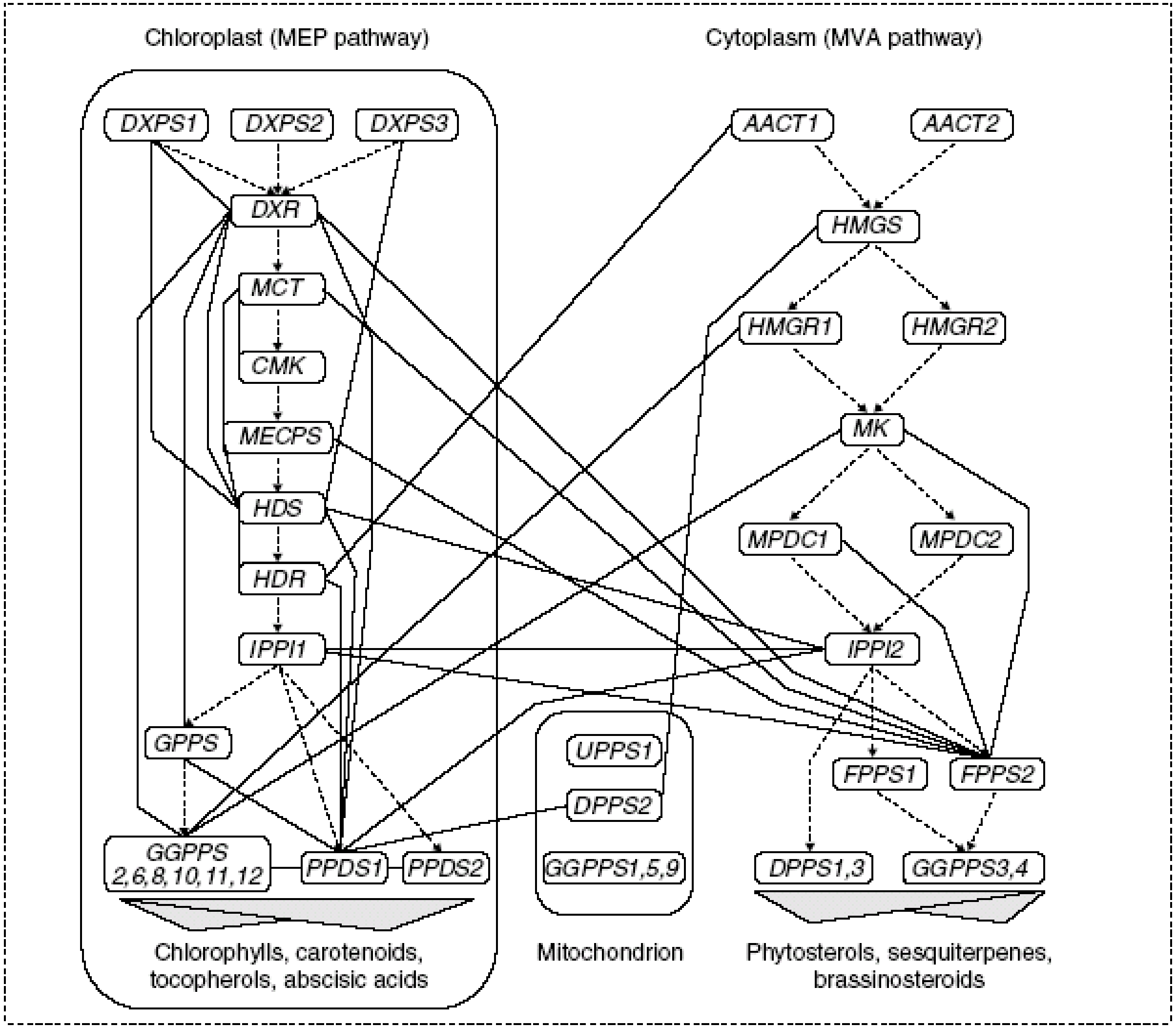


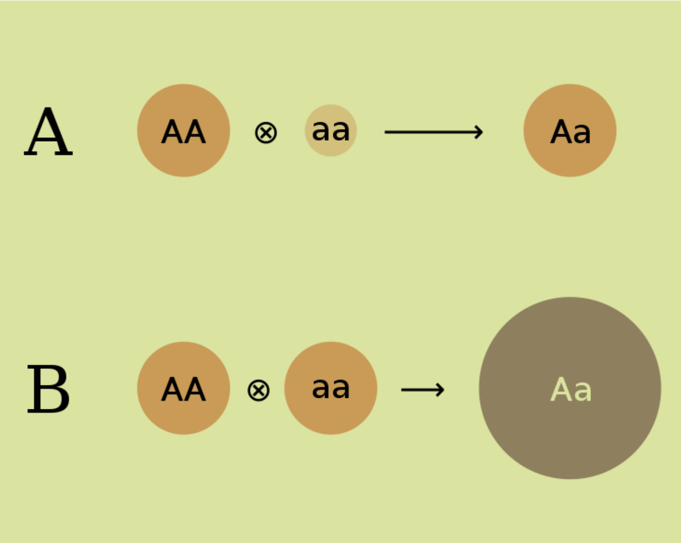
Figure 2
Bootstrapped GGM of the isoprenoid pathway with a cutoff at 0.8. The solid undirected edges connecting individual genes (in boxes) represent the GGM. Dotted directed edges mark the metabolic network, and are not part of the GGM. The grey shading indicates metabolic links to downstream pathways.

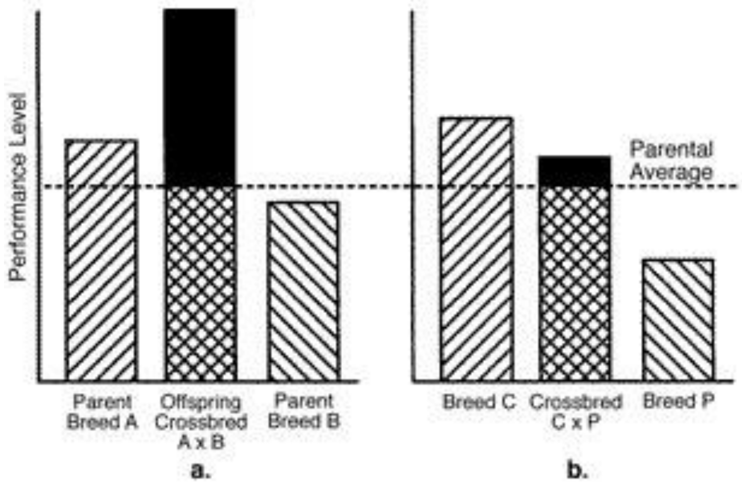
Theor Appl Genet (2010) 120:249–259
DOI 10.1007/s00122-009-1214-z

ORIGINAL PAPER

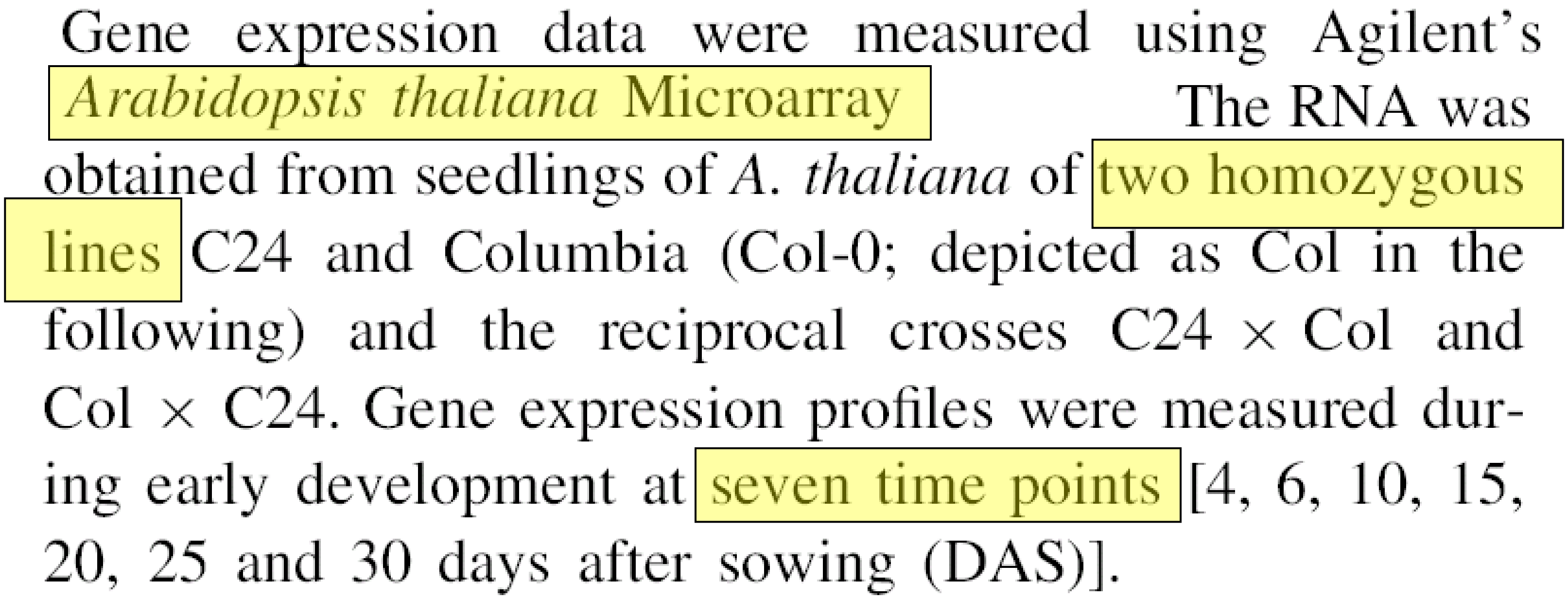
Enriched partial correlations in genome-wide gene expression profiles of hybrids (*A. thaliana*): a systems biological approach towards the molecular basis of heterosis

Sandra Andorf · Joachim Selbig · Thomas Altmann ·
Kathrin Poos · Hanna Witucka-Wall · Dirk Repsilber





Network hypothesis of heterosis: additional alleles → additional regulatory interactions in the molecular network



Gene expression data were measured using Agilent's *Arabidopsis thaliana* Microarray The RNA was obtained from seedlings of *A. thaliana* of two homozygous lines C24 and Columbia (Col-0; depicted as Col in the following) and the reciprocal crosses C24 × Col and Col × C24. Gene expression profiles were measured during early development at seven time points [4, 6, 10, 15, 20, 25 and 30 days after sowing (DAS)].

GGMs applied to 1000 genes

Problem: short time series
Modify the research question

Rather than asking:

“How does the network structure change as a consequence of additional alleles at the heterozygous loci?"

which could not be answered with the given amount of data – the authors asked the question:

“What is the impact of heterozygosity on the overall connectivity of the molecular regulatory network?"

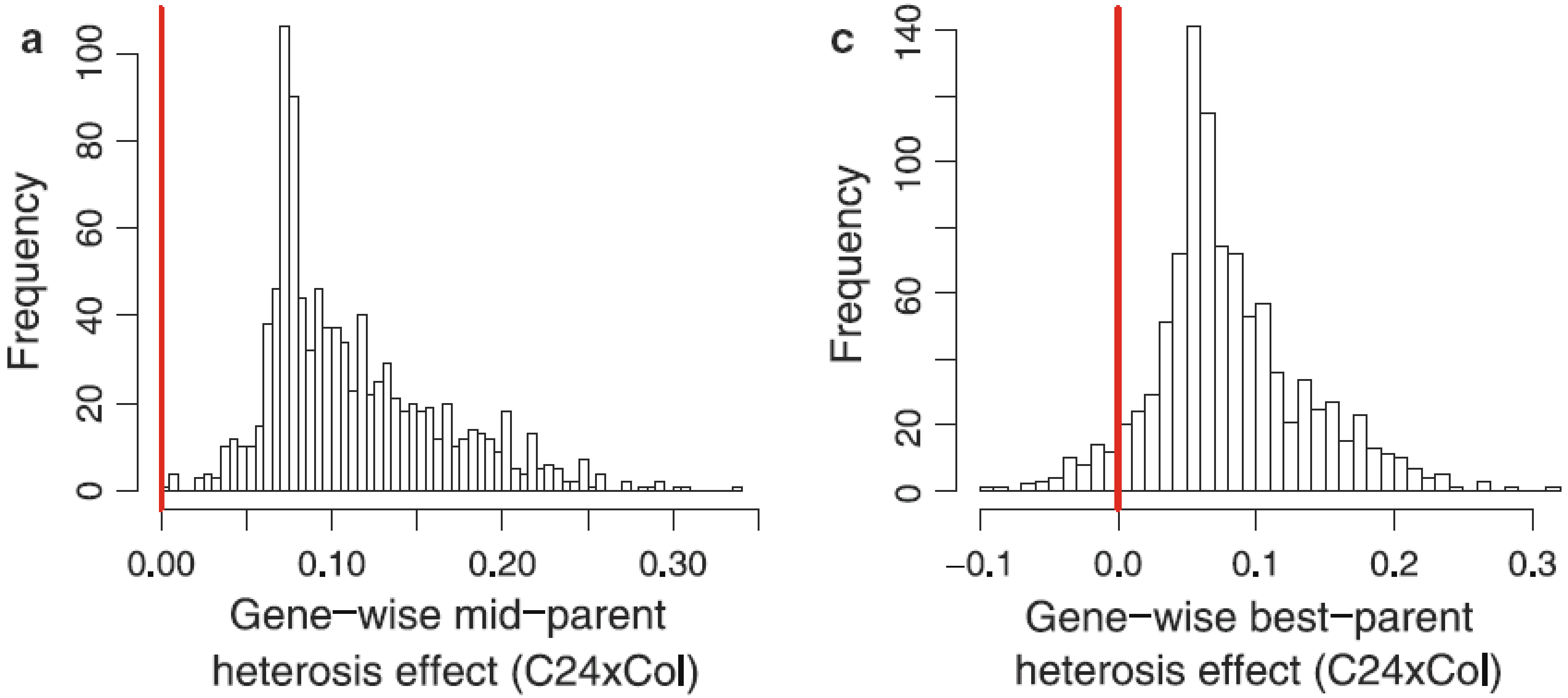
Spectrum of partial correlation coefficients

heterozygous

homozygous

$$h_{w,f}^{\text{mid-heterosis}} = h_{w,f} - h_f^{\text{mid-parent}}$$

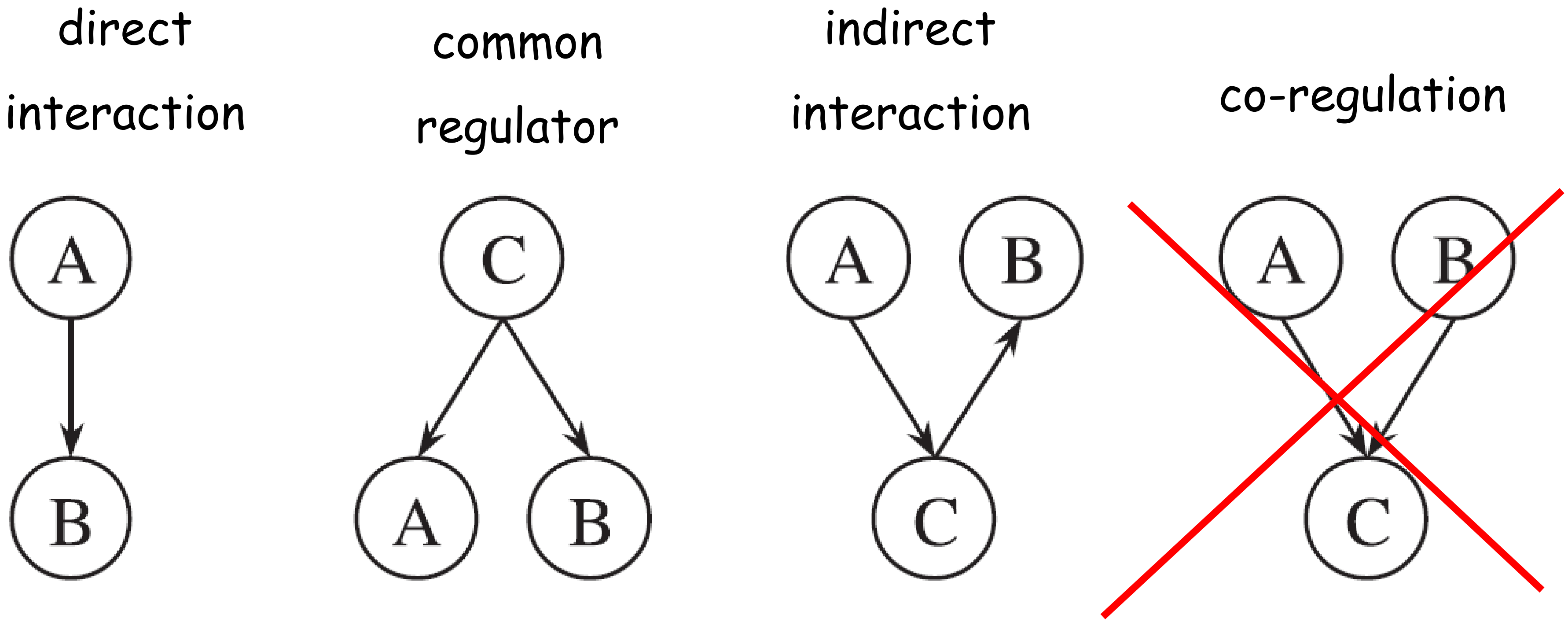
$$h_{w,f}^{\text{best-heterosis}} = h_{w,f} - h_f^{\text{best-parent}}$$



Shortcomings of GGMs

Pairwise interactions conditional on the whole systems, but:

no proper scoring of the whole network



direct interaction

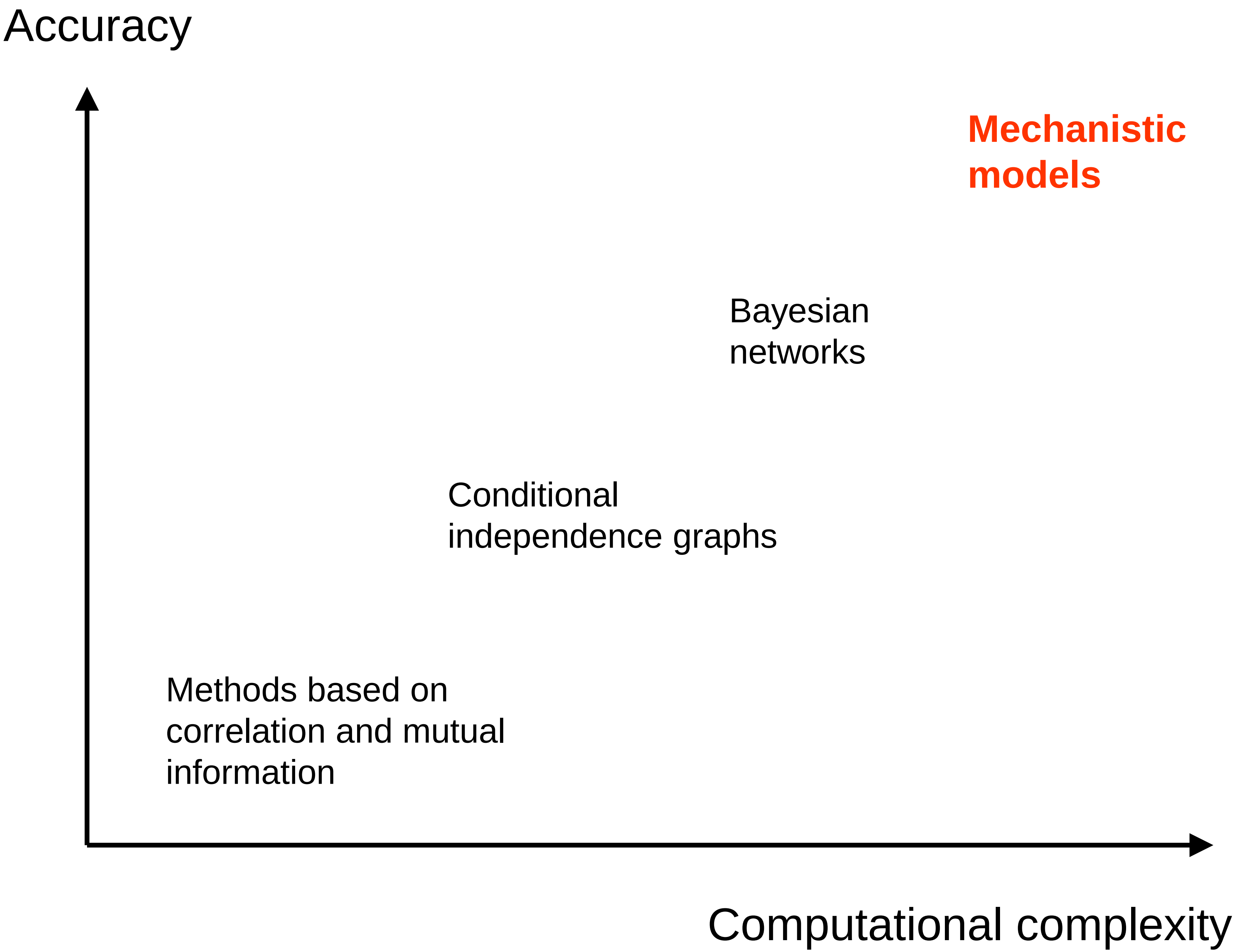
common regulator

indirect interaction

co-regulation

P(A,B)=P(A)·P(B)

But: P(A,B|C)≠P(A|C)·P(B|C)



Accuracy

Mechanistic models

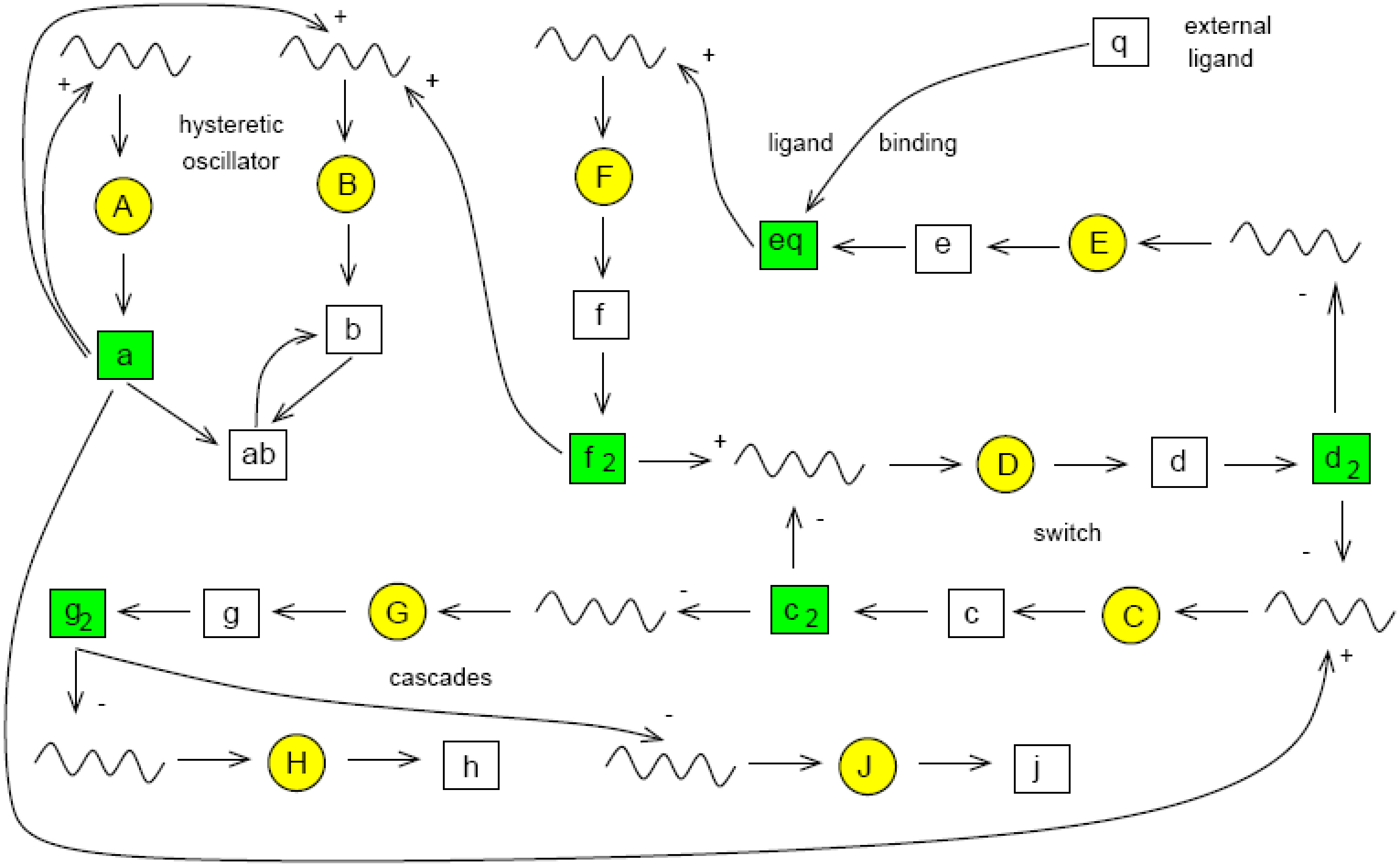
Bayesian networks

Conditional independence graphs

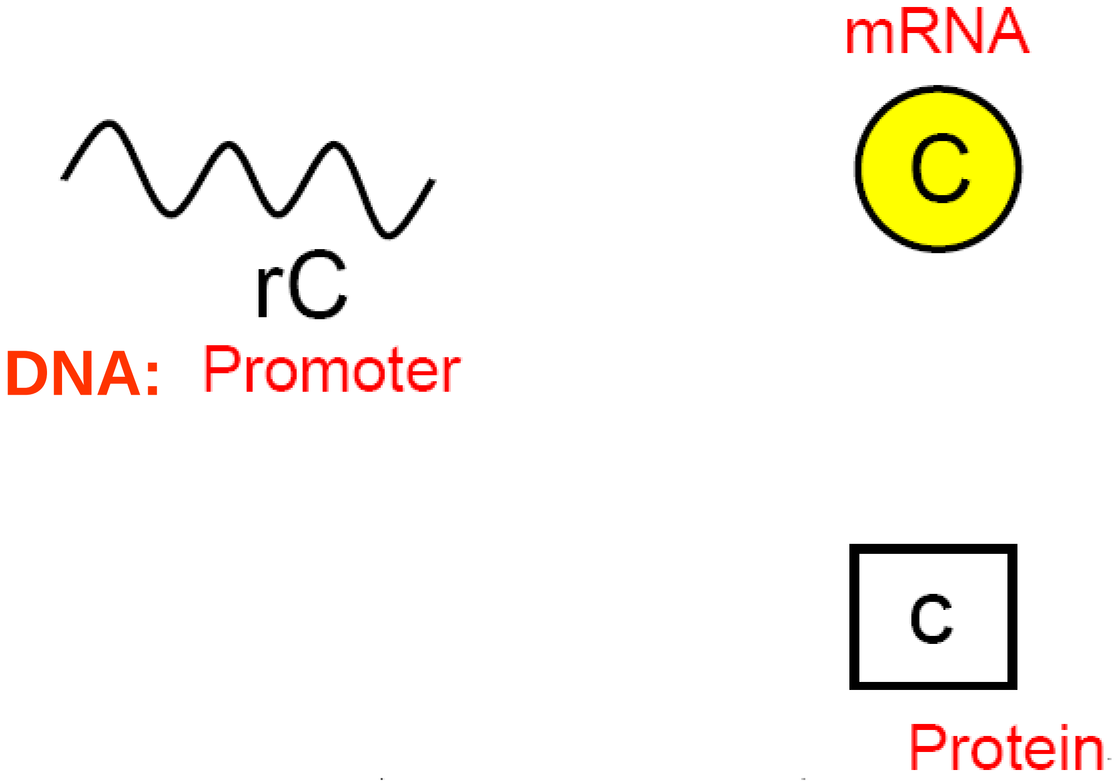
Methods based on correlation and mutual information

Computational complexity

Regulatory network



Elementary molecular components



mRNA

C

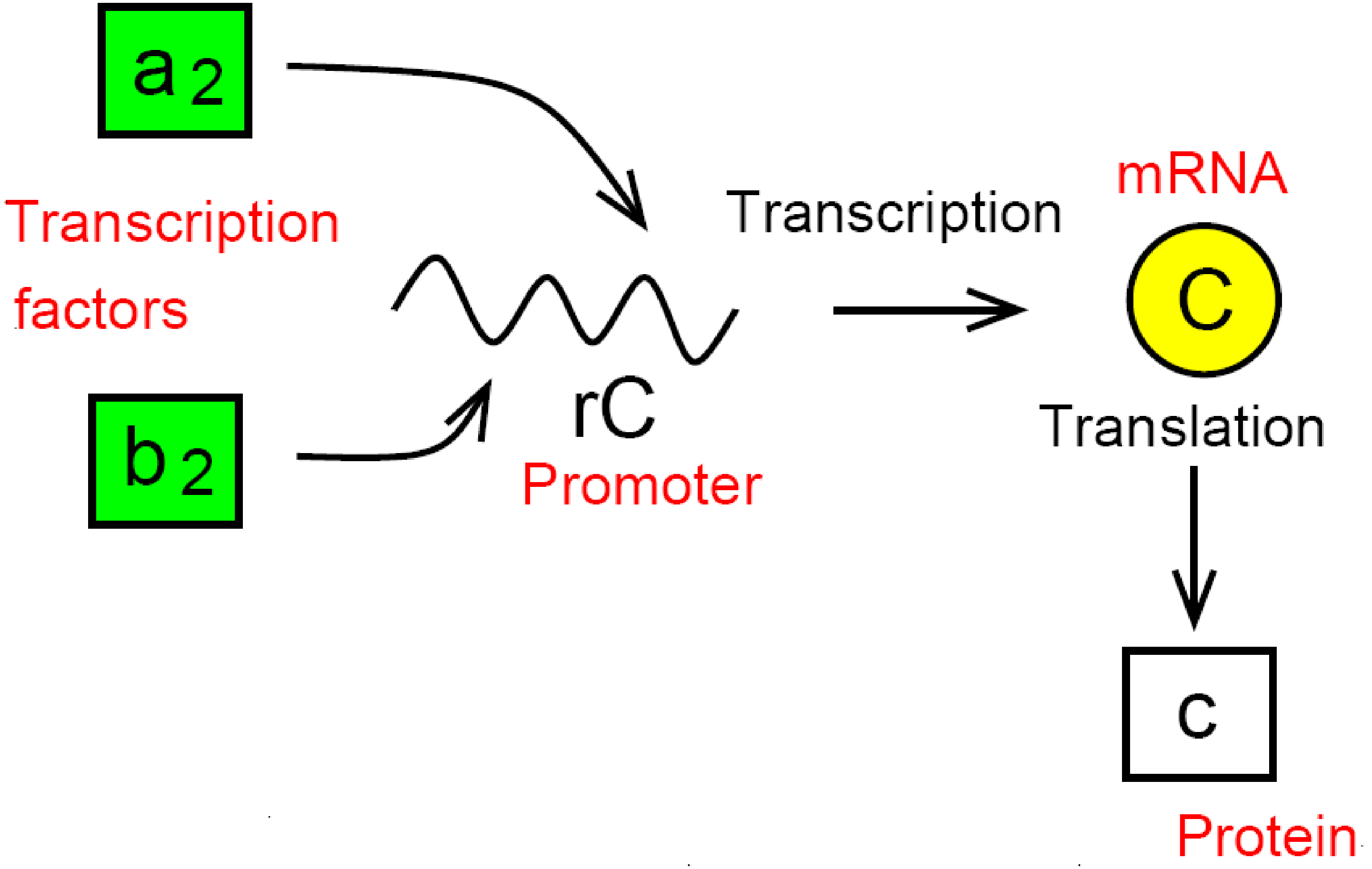
rC

DNA: Promoter

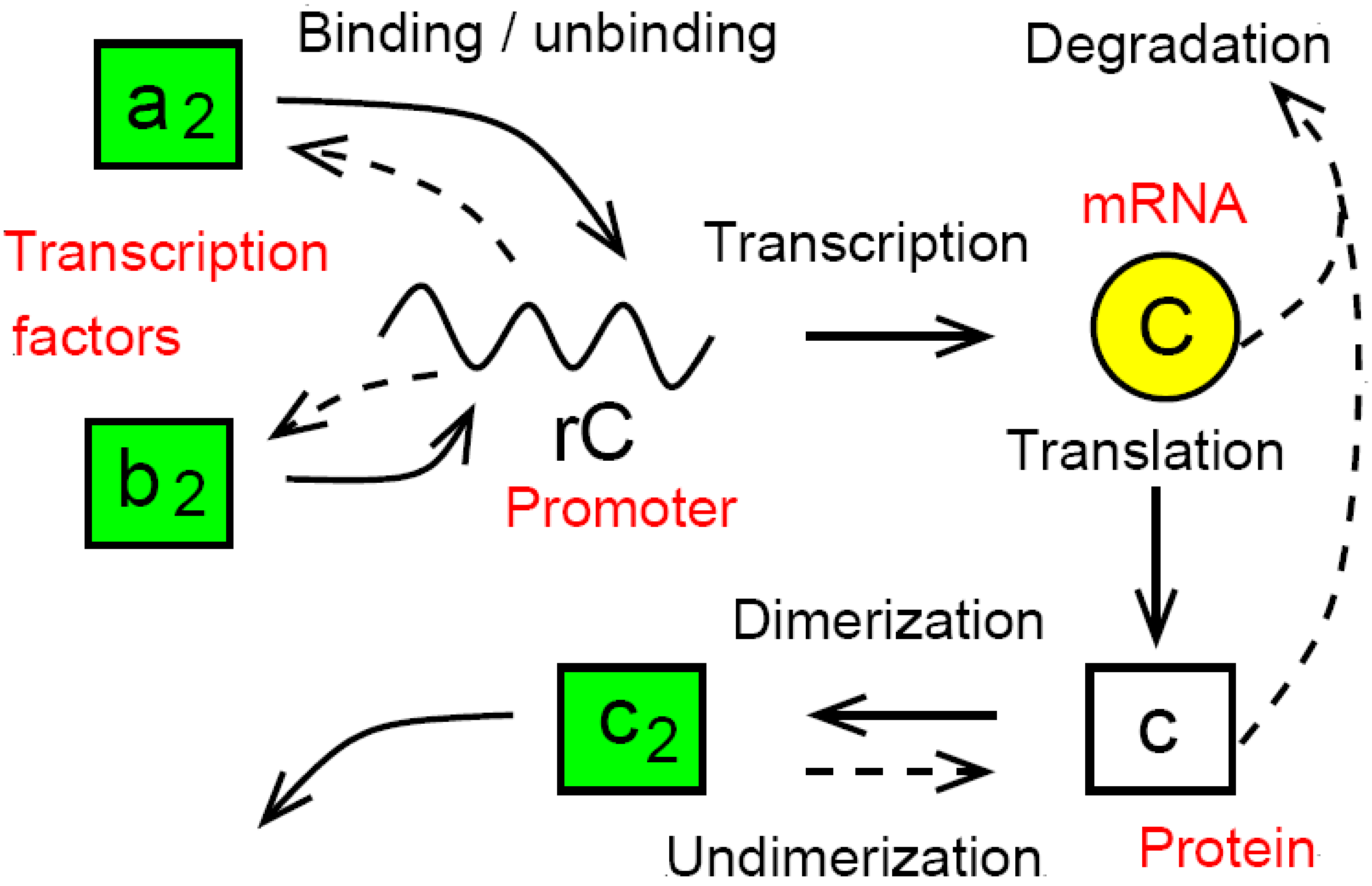
c

Protein

Elementary molecular biological processes



Elementary molecular biological processes



Description with differential equations

$$\frac{d}{dt}[a2.rC] = \lambda^{+}_{a_2.rC}[a_2][rC] - \lambda^{-}_{a_2.rC}[a_2.rC]$$

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

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

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

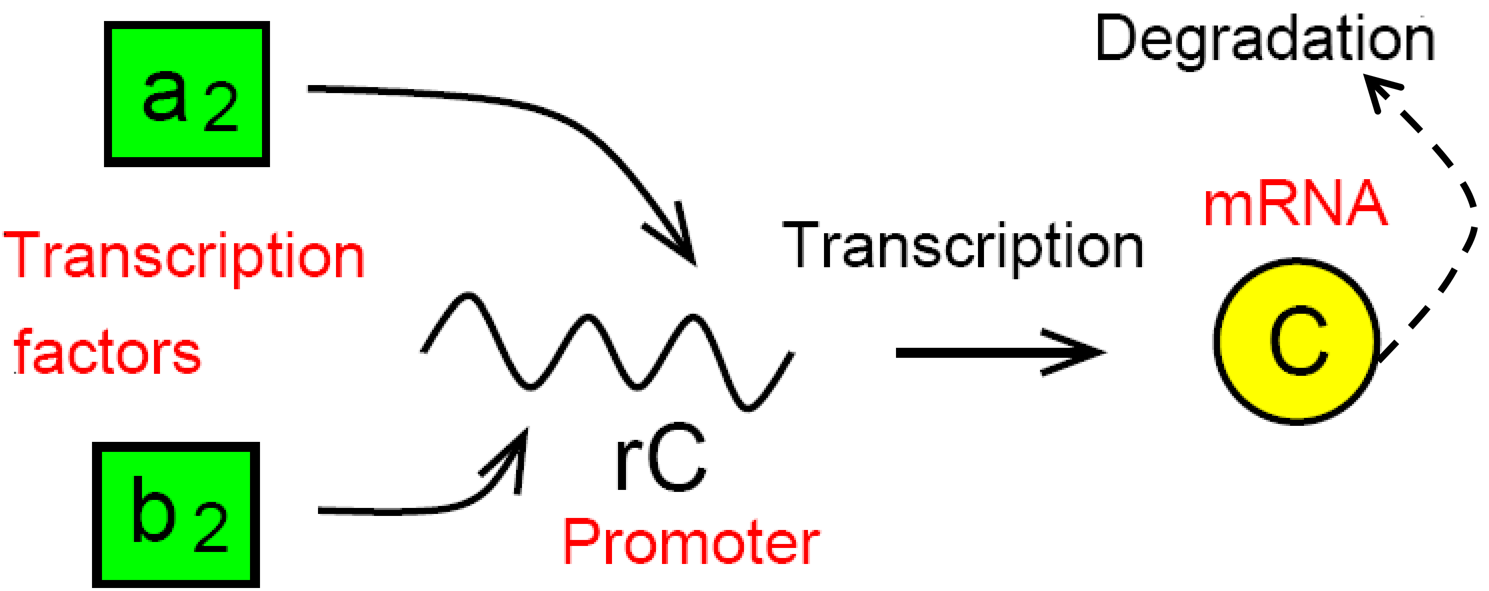
Description with differential equations

$$\frac{d}{dt}[a2.rC] = \lambda^{+}_{a_2.rC}[a_2][rC] - \lambda^{-}_{a_2.rC}[a_2.rC]$$

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

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

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



Concentrations

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

Rates

Kinetic parameters **q**

Description with differential equations

Concentrations

$$\frac{d}{dt}[a2.rC] = \lambda^{+}_{a_2.rC}[a_2][rC] - \lambda^{-}_{a_2.rC}[a_2.rC]$$

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

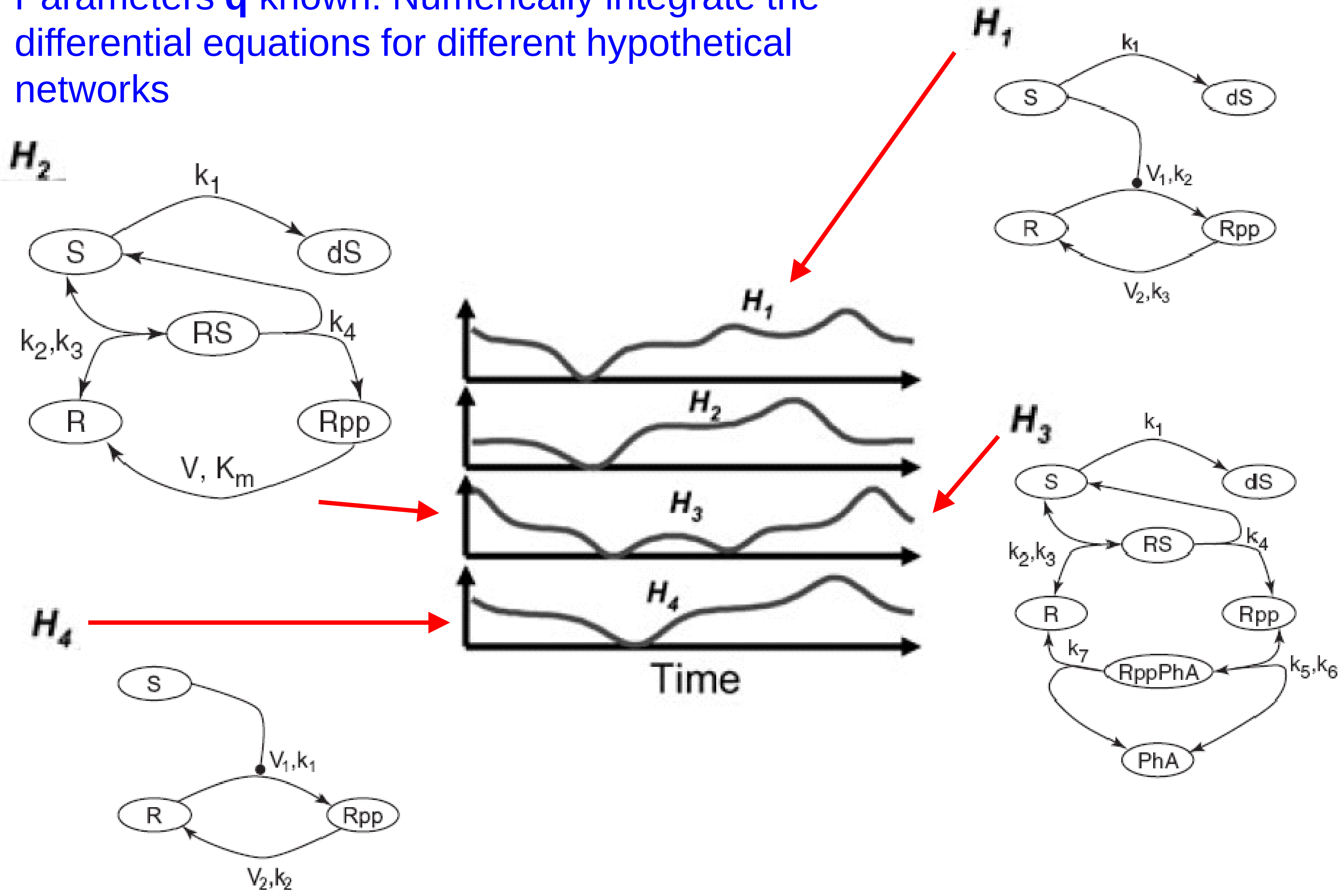
$$\frac{d}{dt}[c] = \lambda_{Cc}[C] - \lambda_{c}[c]$$

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

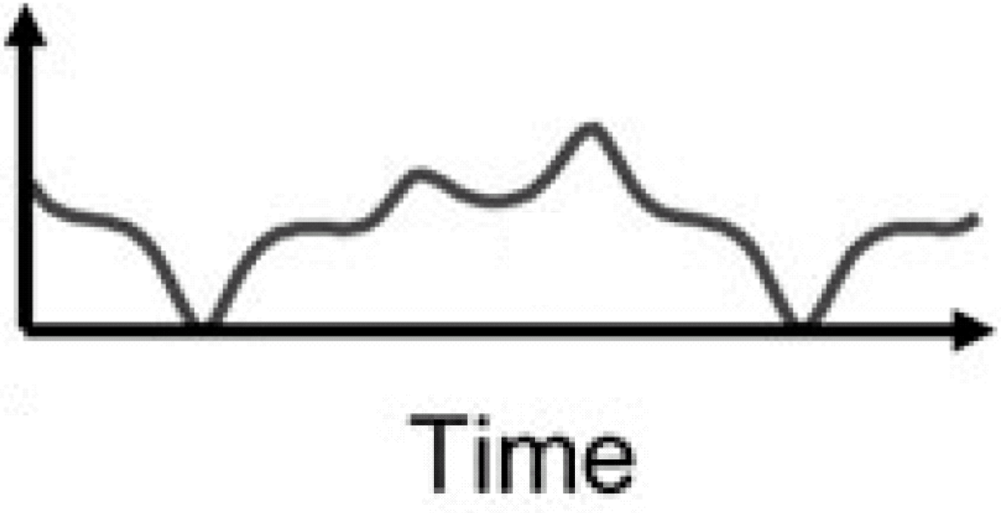
Kinetic parameters **q**

Rates

Parameters **q** known: Numerically integrate the differential equations for different hypothetical networks



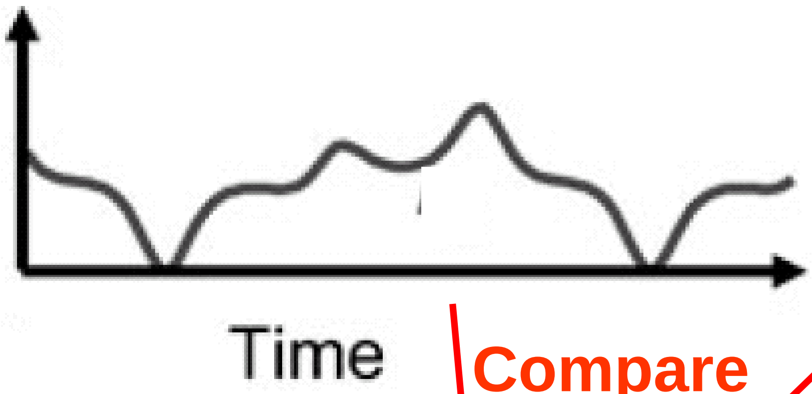
Experiment:
Gene expression time series



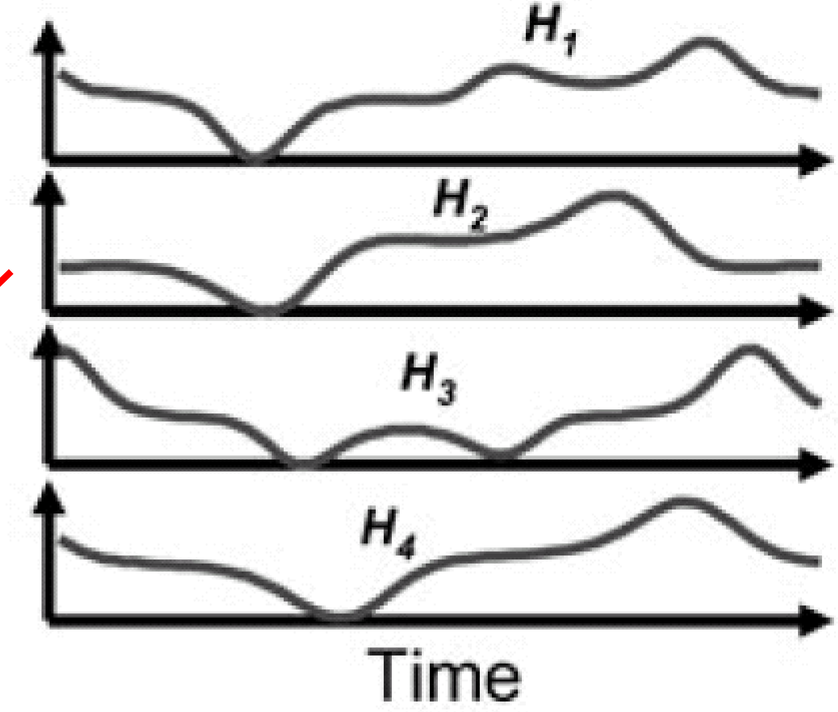
Can we infer the correct gene regulatory network?

Model selection for known parameters **q**

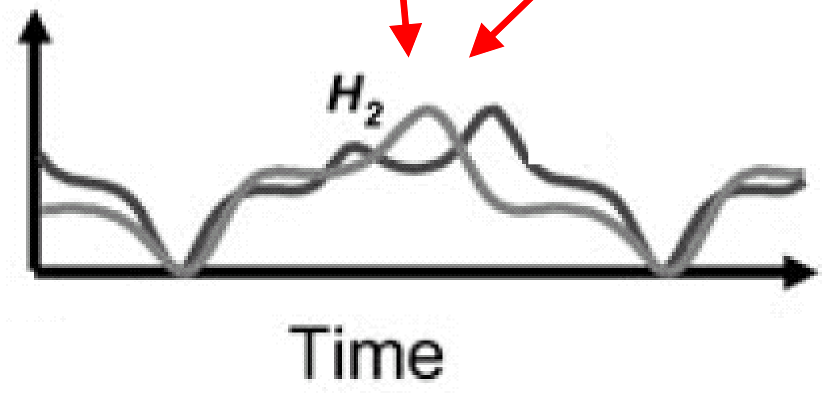
Measured gene expression time series



Gene expression time series **predicted** with different models



Compare

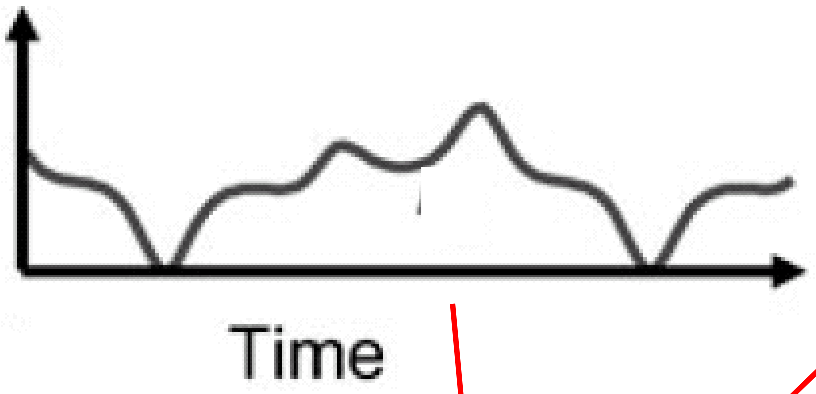


Highest likelihood: best model

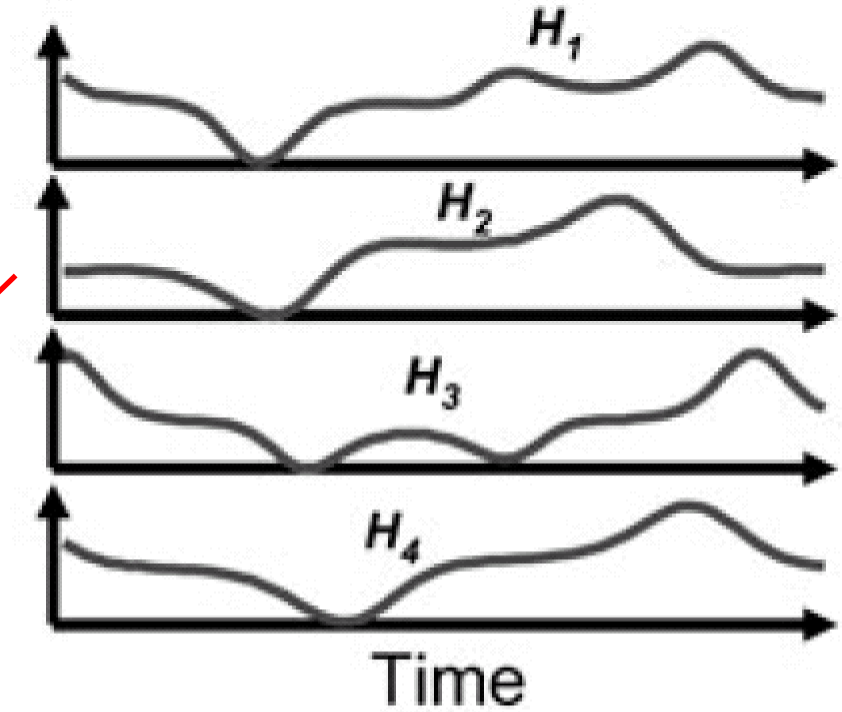
$$P(\mathcal{D}|\mathbf{q}, \mathcal{M})$$

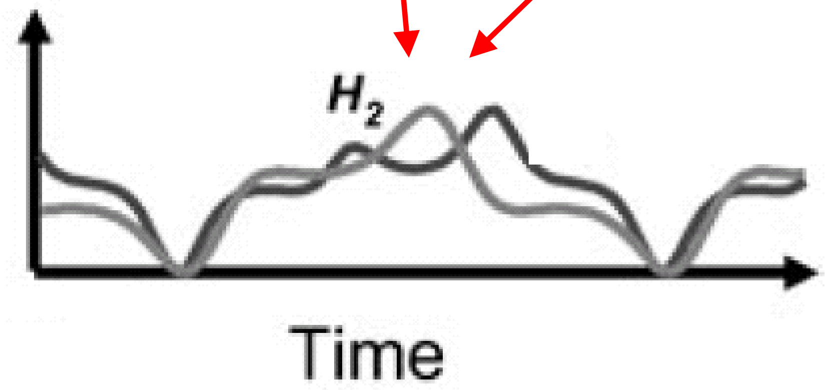
Model selection for **unknown** parameters **q**

Measured gene expression time series



Gene expression time series predicted with different models

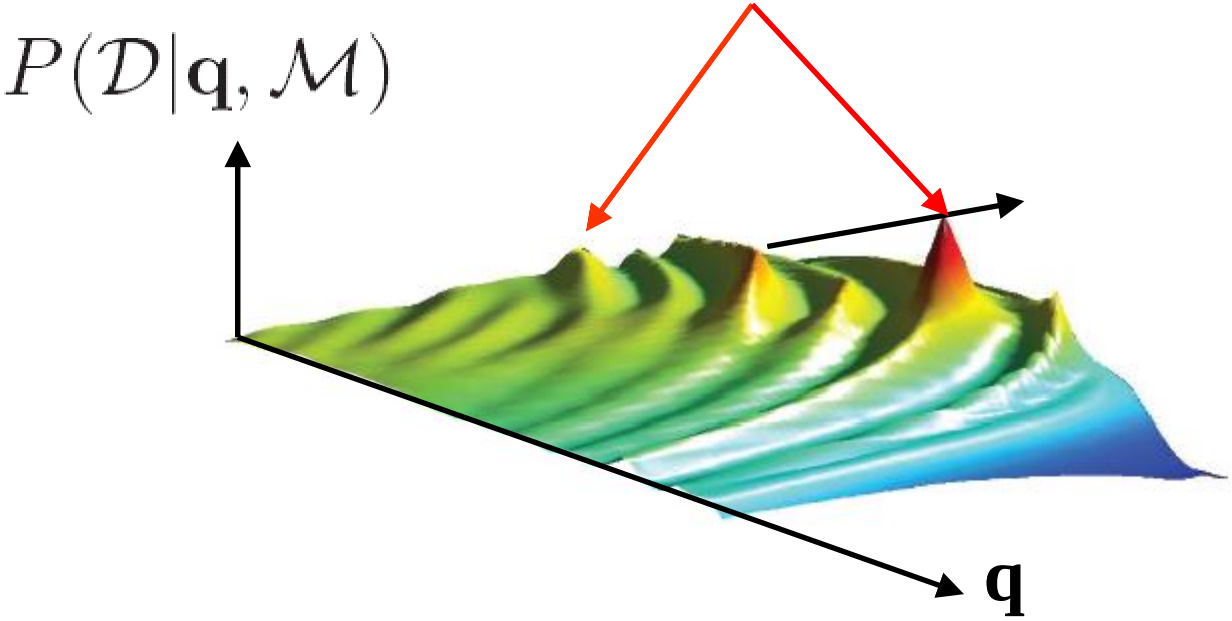




Joint maximum likelihood:

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

1) Practical problem: numerical optimization

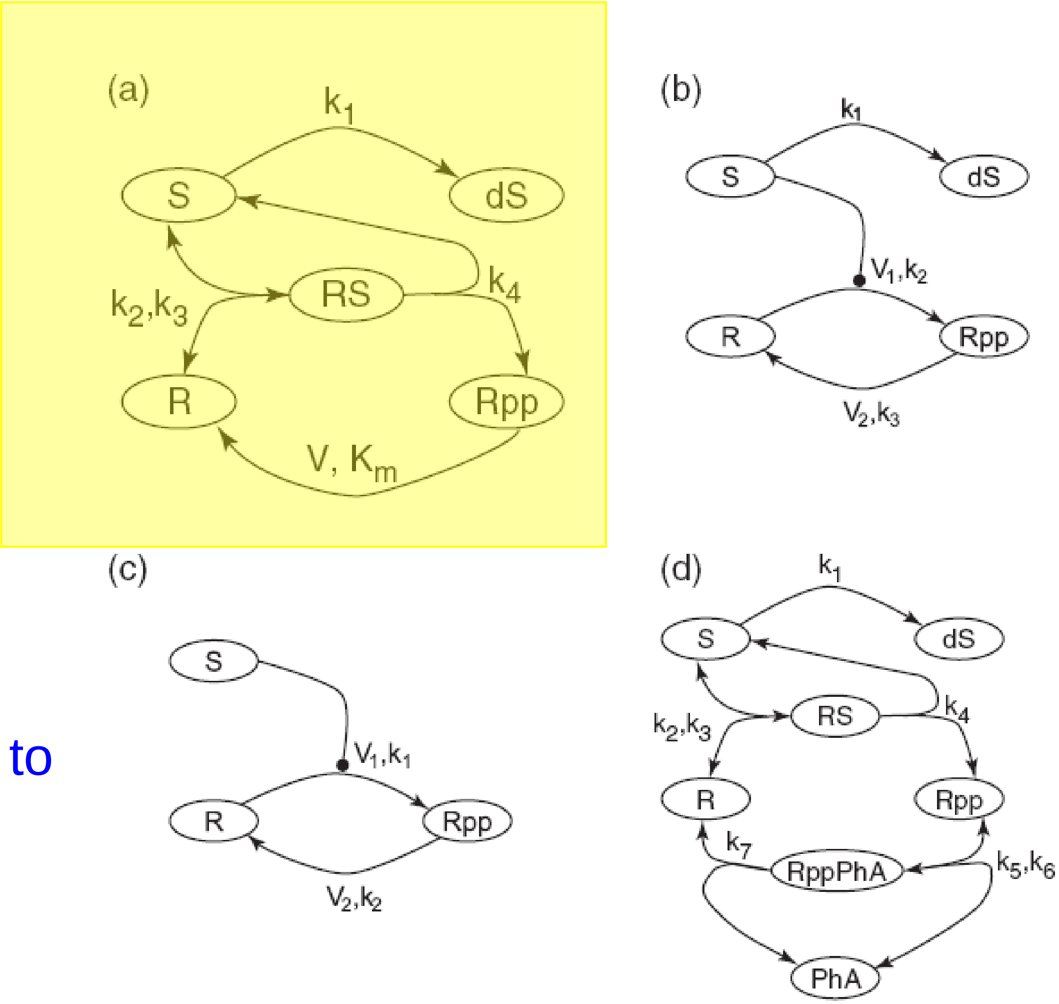


2) Conceptual problem: overfitting

ML estimate increases on increasing the network complexity

Overfitting problem

True pathway



Poorer fit to the data

Poorer fit to the data

Equal or better fit to the data

Regularization

E.g.: BIC

Data misfit term

Regularization term

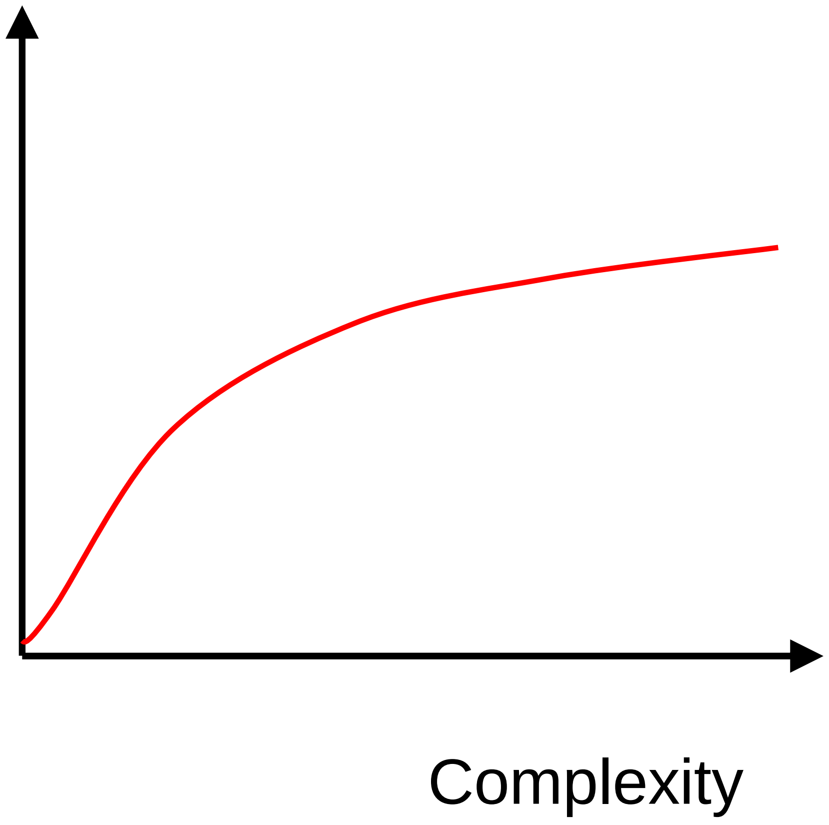
$$\log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{k}{2} \log N$$

Maximum likelihood parameters

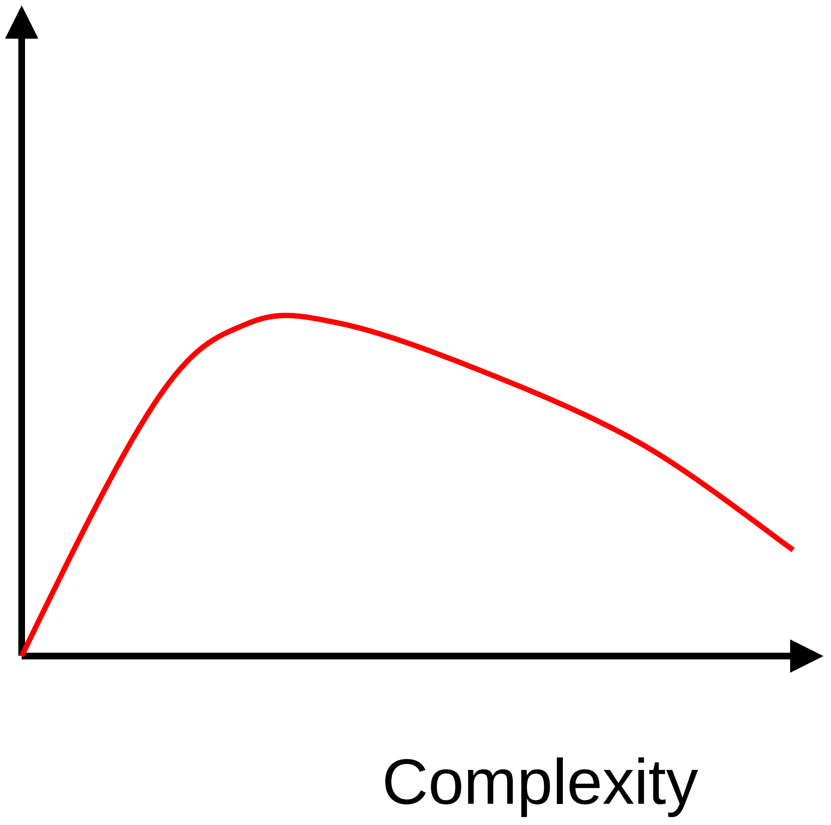
Number of parameters

Number of data points

Likelihood



BIC



Model selection: find the best pathway

Select the model $\mathcal{M}$ with the highest posterior probability:

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

This requires an integration over the whole parameter space:

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

Comparison with BIC

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

$$P(\mathcal{D}|\mathcal{M}) \;=\; \int P(\mathcal{D}|\mathbf{q},\mathcal{M})\xcancel{P(\mathbf{q}|\mathcal{M})}d\mathbf{q} \;=\; \int \exp\Big[-E(\mathbf{q})\Big]d\mathbf{q}$$

$$E(\mathbf{q}) \;=\; -\log P(\mathcal{D}|\mathbf{q},\mathcal{M})$$

Comparison with BIC

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

$$P(\mathcal{D}|\mathcal{M}) \;=\; \int P(\mathcal{D}|\mathbf{q},\mathcal{M})\xcancel{P(\mathbf{q}|\mathcal{M})}d\mathbf{q} \;=\; \int \exp\Big[-E(\mathbf{q})\Big]d\mathbf{q}$$

$$E(\mathbf{q}) \;=\; -\log P(\mathcal{D}|\mathbf{q},\mathcal{M})$$

$$E(\mathbf{q}) \;\approx\; E(\hat{\mathbf{q}}) + \frac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})^{\dagger}\mathbf{H}(\mathbf{q}-\hat{\mathbf{q}})$$

$$\begin{aligned} P(\mathcal{D}|\mathcal{M}) \;&\approx\; \exp\Big[-E(\hat{\mathbf{q}})\Big]\int \exp\Big[-\frac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})^{\dagger}\mathbf{H}(\mathbf{q}-\hat{\mathbf{q}})\Big]d\mathbf{q} \\ &=\; P(\mathcal{D}|\hat{\mathbf{q}},\mathcal{M})\sqrt{\frac{(2\pi)^k}{\det\mathbf{H}}} \end{aligned}$$

Comparison with BIC

$$P(\mathcal{D}|\mathcal{M}) \quad = \quad P(\mathcal{D}|\hat{\mathbf{q}},\mathcal{M})\sqrt{\frac{(2\pi)^k}{\det\mathbf{H}}}$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}},\mathcal{M}) - \frac{1}{2}\log\det\mathbf{H} + \frac{k}{2}\log(2\pi)$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}},\mathcal{M}) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{\varepsilon_i}{2\pi}\right)$$

Comparison with BIC

$$P(\mathcal{D}|\mathcal{M}) \quad = \quad P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M})\sqrt{\frac{(2\pi)^k}{\det \mathbf{H}}}$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{1}{2}\log\det\mathbf{H} + \frac{k}{2}\log(2\pi)$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{\varepsilon_i}{2\pi}\right)$$

$$\varepsilon_i = \alpha_i N$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{\alpha_i}{2\pi}\right) - \frac{k}{2}\log N$$

Comparison with BIC

$$P(\mathcal{D}|\mathcal{M}) \quad = \quad P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M})\sqrt{\frac{(2\pi)^k}{\det \mathbf{H}}}$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{1}{2}\log \det \mathbf{H} + \frac{k}{2}\log(2\pi)$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{\varepsilon_i}{2\pi}\right)$$

$$\varepsilon_i = \alpha_i N$$

$$\log P(\mathcal{D}|\mathcal{M}) \quad = \quad \log P(\mathcal{D}|\hat{\mathbf{q}}, \mathcal{M}) - \xcancel{\frac{1}{2}\sum_{i=1}^{k}\log\left(\frac{\alpha_i}{2\pi}\right)} - \frac{k}{2}\log N$$

BIC approximation

Model selection: find the best pathway

Select the model $\mathcal{M}$ with the highest posterior probability:

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

This requires an integration over the whole parameter space:

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

Model selection: find the best pathway

Select the model $\mathcal{M}$ with the highest posterior probability:

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

This requires an integration over the whole parameter space:

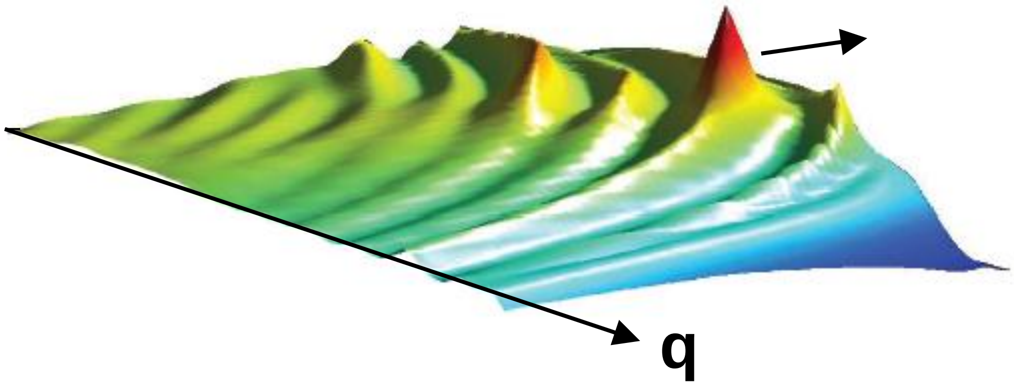
$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\mathbf{q},\mathcal{M})P(\mathbf{q}|\mathcal{M})d\mathbf{q}$$

This integral is usually analytically intractable

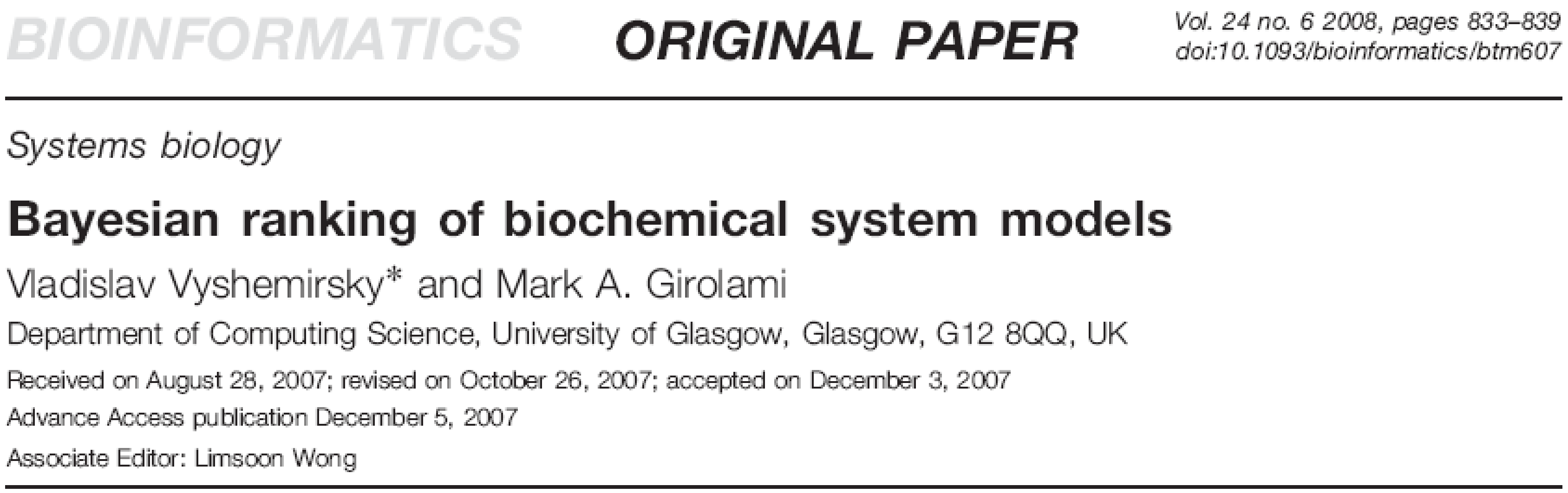
Complexity problem

This requires an integration over the whole parameter space:

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



The numerical approximation is highly non-trivial



BIOINFORMATICS **ORIGINAL PAPER**

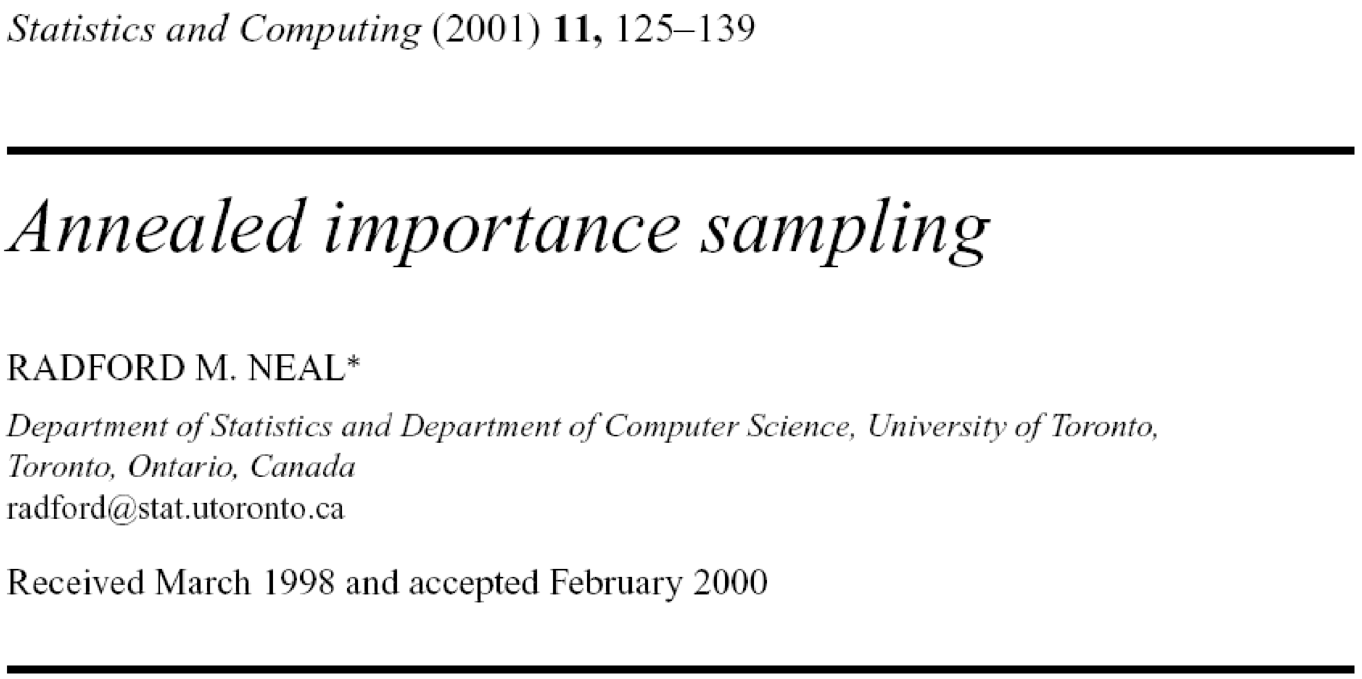
Vol. 24 no. 6 2008, pages 833–839
doi:10.1093/bioinformatics/btm607

Systems biology

Bayesian ranking of biochemical system models

Vladislav Vyshemirsky* and Mark A. Girolami
Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK

Received on August 28, 2007; revised on October 26, 2007; accepted on December 3, 2007
Advance Access publication December 5, 2007
Associate Editor: Limsoon Wong



Statistics and Computing (2001) **11,** 125–139

Annealed importance sampling

RADFORD M. NEAL*
Department of Statistics and Department of Computer Science, University of Toronto,
Toronto, Ontario, Canada
radford@stat.utoronto.ca

Received March 1998 and accepted February 2000

Numerical integration by sampling from the prior

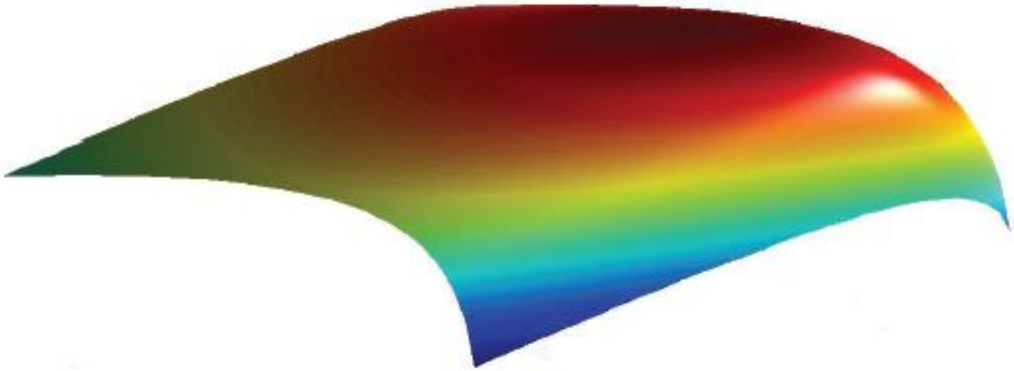
Model: S Parameters: ϕ

$$P(\boldsymbol{\mathcal{D}}|S) = \int P(\boldsymbol{\mathcal{D}}|\phi, S)P(\phi|S)d\phi$$

$$P(\boldsymbol{\mathcal{D}}|S) \approx \frac{1}{N}\sum_{t=1}^{N} P(\boldsymbol{\mathcal{D}}|\phi_t, S)$$

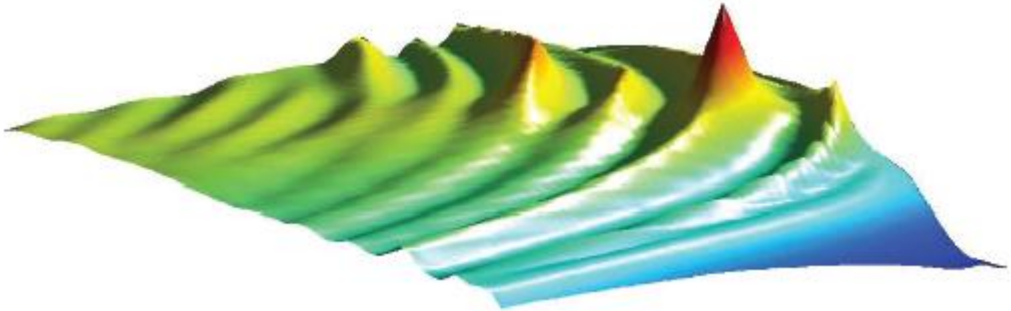
where $\{\phi_t\}$ is a sample from the prior distribution $P(\phi|S)$

Problem: Extremely poor convergence in high dimensions



Prior distribution

$P(\phi|S)$



Likelihood function

$P(\mathcal{D}|\phi, S)$.

Taken from the MSc thesis by Ben Calderhead

Numerical integration by sampling from the posterior

Model: S Parameters: ϕ

$$P(\boldsymbol{\mathcal{D}}|\phi, S)P(\phi|S) = P(\phi|\boldsymbol{\mathcal{D}}, S)P(\boldsymbol{\mathcal{D}}|S)$$

$$\int \frac{P(\phi|S)}{P(\boldsymbol{\mathcal{D}}|S)} d\phi = \int \frac{P(\phi|\boldsymbol{\mathcal{D}}, S)}{P(\boldsymbol{\mathcal{D}}|\phi, S)} d\phi$$

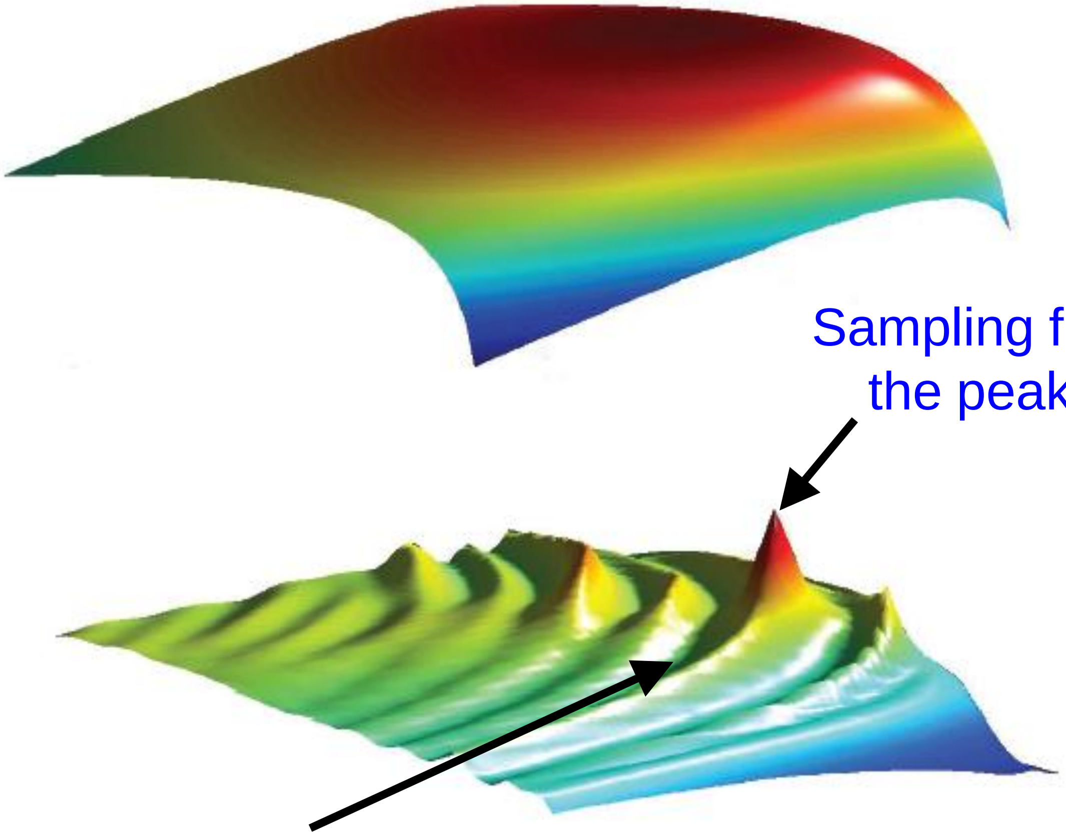
$$\frac{1}{P(\boldsymbol{\mathcal{D}}|S)} = \int \frac{P(\phi|\boldsymbol{\mathcal{D}}, S)}{P(\boldsymbol{\mathcal{D}}|\phi, S)} d\phi$$

$$\frac{1}{P(\boldsymbol{\mathcal{D}}|S)} \approx \frac{1}{N}\sum_{t=1}^{N} \frac{1}{P(\boldsymbol{\mathcal{D}}|\phi_t, S)}$$

where $\{\phi_t\}$ is a sample from the posterior distribution $P(\phi|\boldsymbol{\mathcal{D}}, S)$

Problem: Poor convergence in high dimensions and instability

Taken from the MSc thesis by Ben Calderhead



Sampling from the peaks

Main contributions to the integral from the valleys

Prior distribution

$$P(\phi|S)$$

Likelihood function

$$P(\mathcal{D}|\phi, S).$$

$$\approx$$

Posterior distribution

$$P(\phi|\mathcal{D}, S)$$

Importance sampling

$$P(\boldsymbol{\mathcal{D}}|S) = \int P(\boldsymbol{\mathcal{D}}|\phi, S)P(\phi|S)d\phi$$

Arbitrary (possibly unnormalized) distribution $Q(\phi)$

$$\frac{P(\boldsymbol{\mathcal{D}}|S)}{Z_Q} = \int \frac{P(\boldsymbol{\mathcal{D}}|\phi, S)P(\phi|S)}{Q(\phi)} \boxed{\frac{Q(\phi)}{Z_Q}} d\phi$$

$$\frac{P(\boldsymbol{\mathcal{D}}|S)}{Z_Q} \longleftarrow \frac{1}{N}\sum_{t=1}^{N} c_t$$

$$c_t = \frac{P(\boldsymbol{\mathcal{D}}|\phi_t, S)P(\phi_t|S)}{Q(\phi_t)}$$

sampled from

Statistics and Computing (2001) **11,** 125–139

Annealed importance sampling

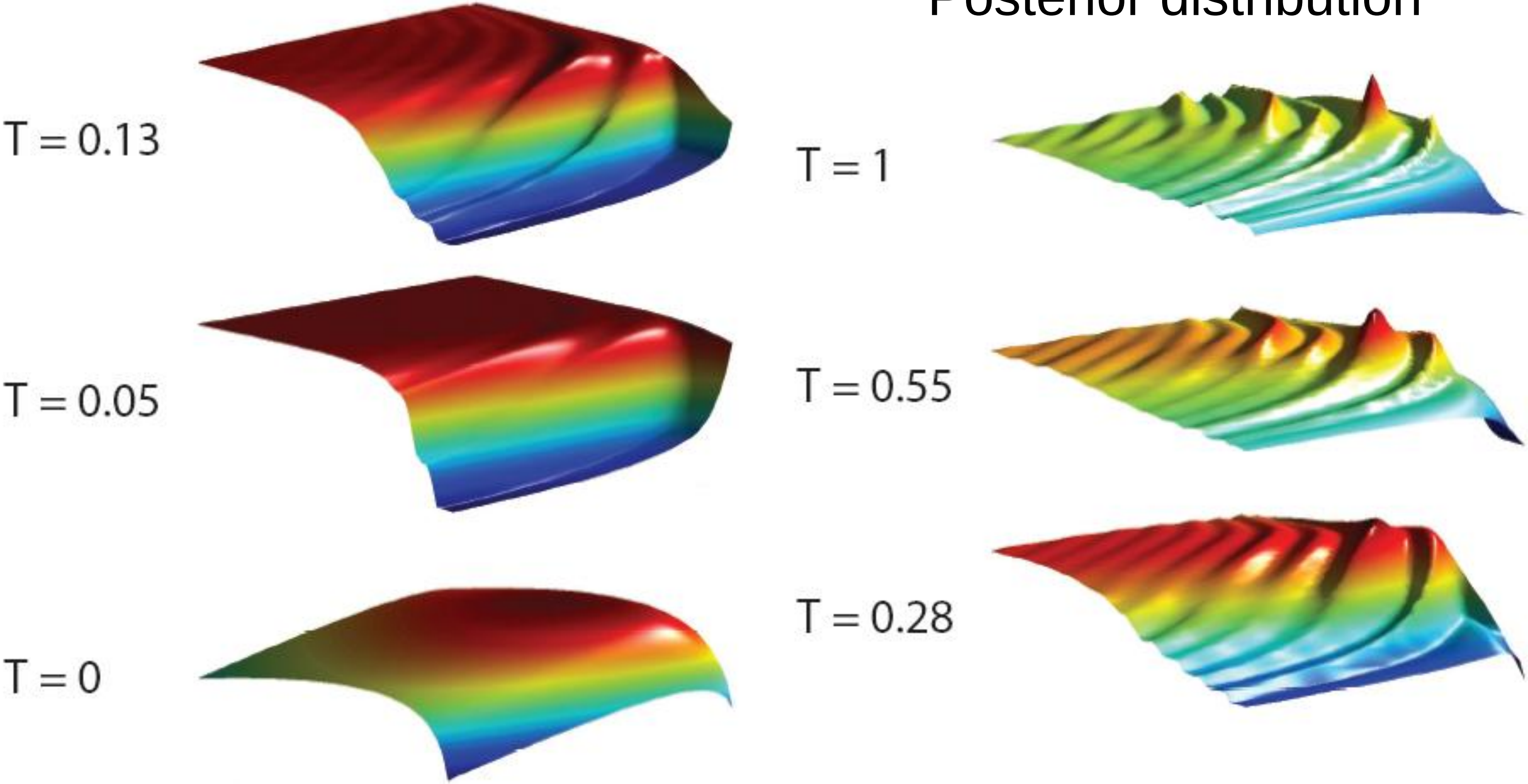
RADFORD M. NEAL*

Department of Statistics and Department of Computer Science, University of Toronto,
Toronto, Ontario, Canada
radford@stat.utoronto.ca

Received March 1998 and accepted February 2000

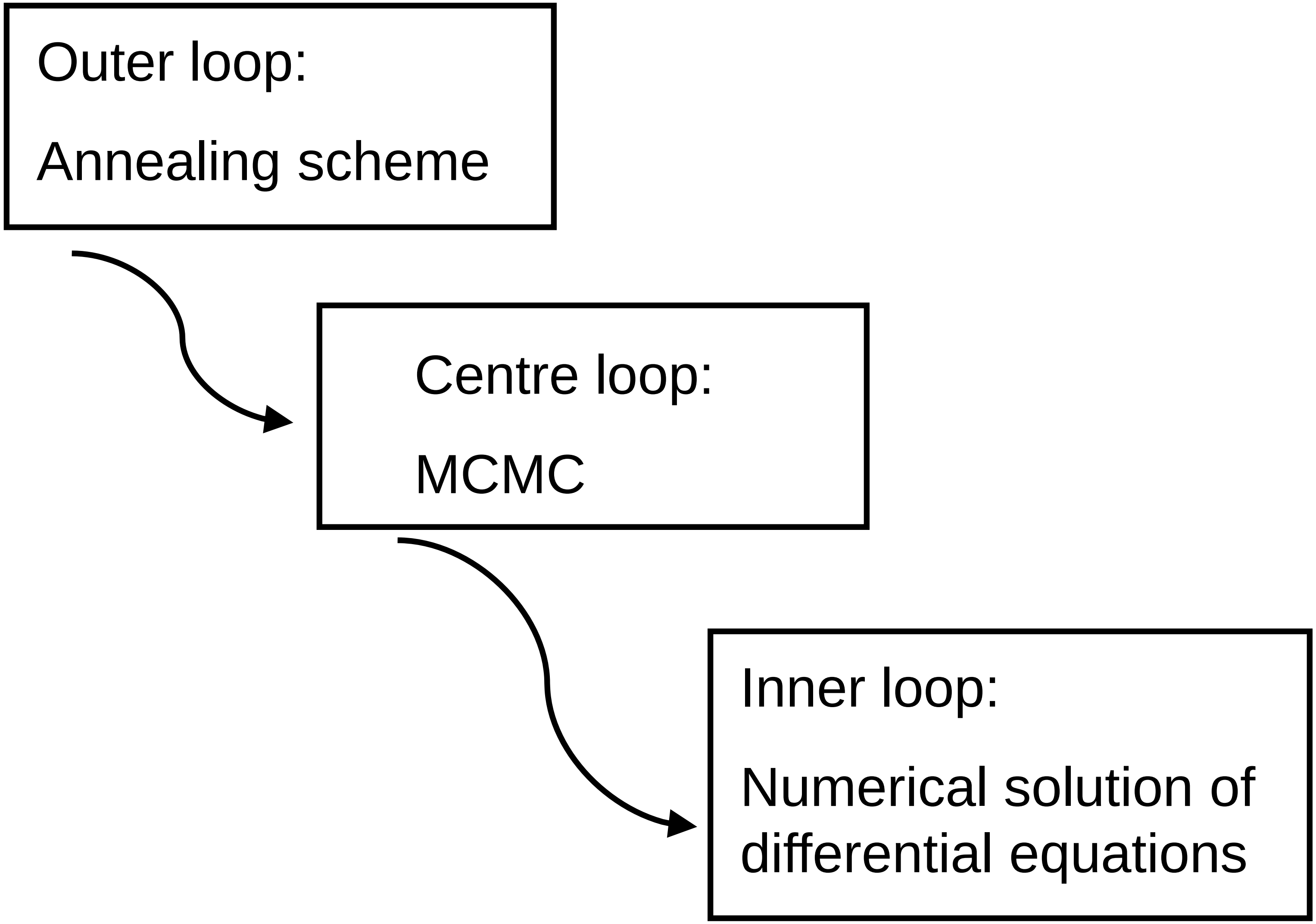
Illustration of annealed importance sampling

Posterior distribution



Prior distribution

Taken from the MSc thesis by Ben Calderhead,



Outer loop:

Annealing scheme

Centre loop:

MCMC

Inner loop:

Numerical solution of differential equations

BIOINFORMATICS ORIGINAL PAPER

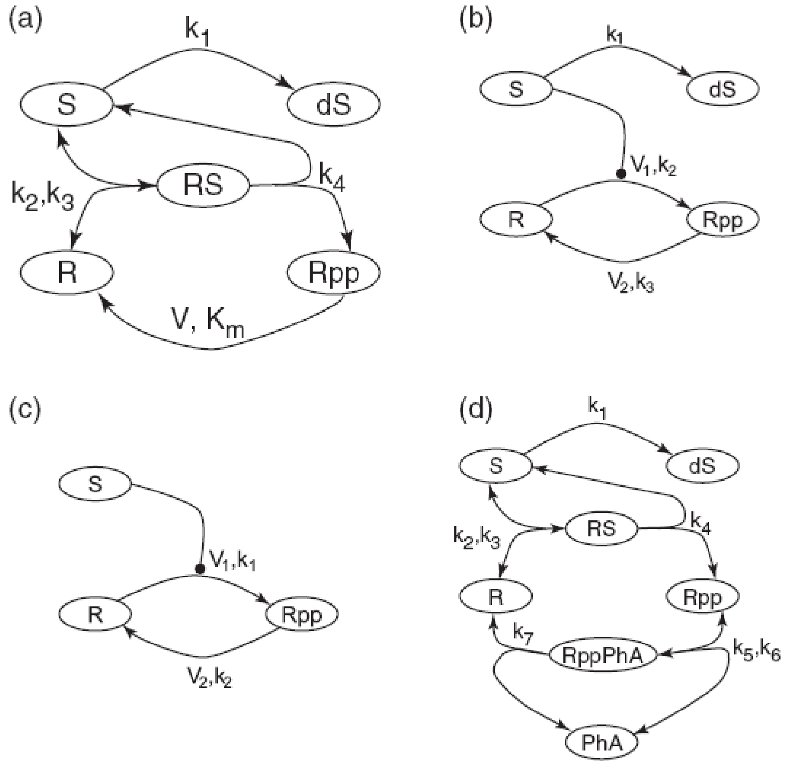
Vol. 24 no. 6 2008, pages 833–839
doi:10.1093/bioinformatics/btm607

Systems biology

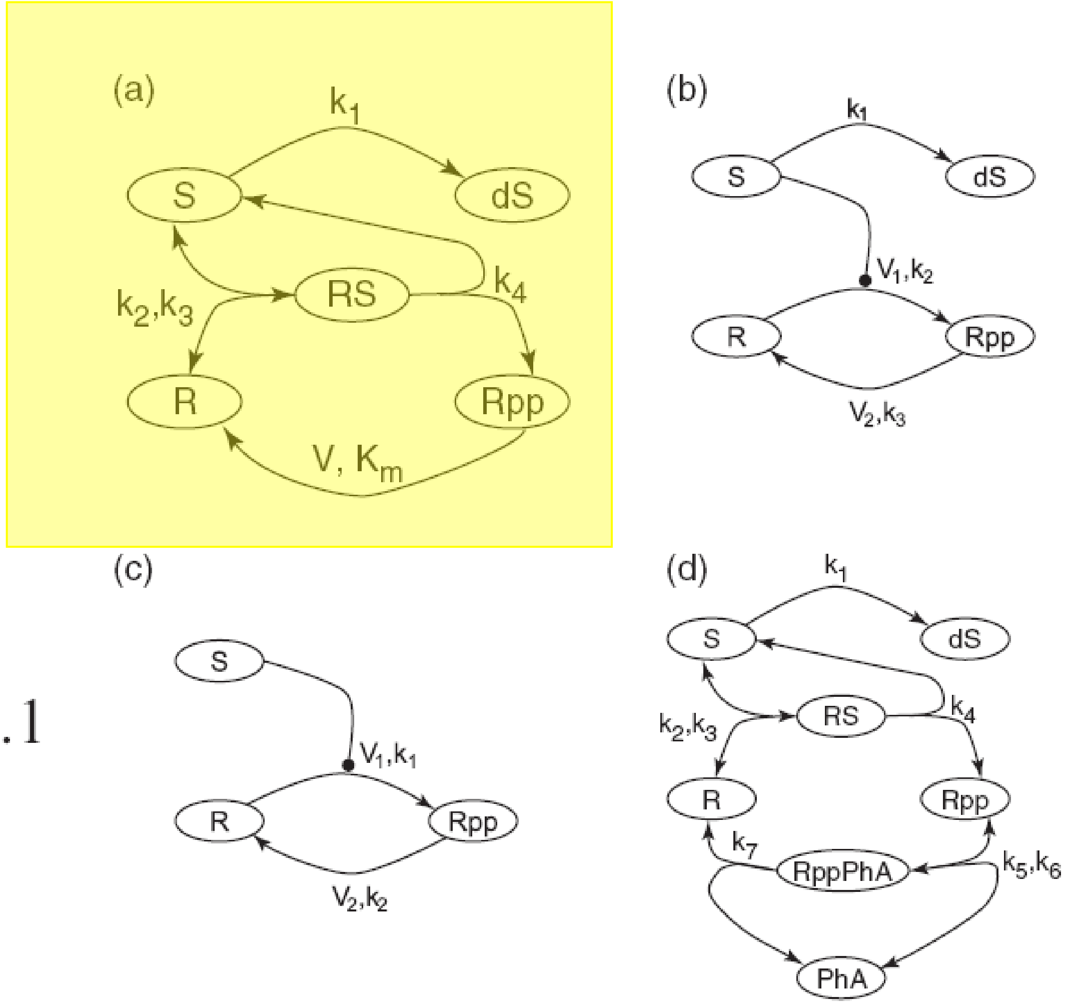
Bayesian ranking of biochemical system models

Vladislav Vyshemirsky* and Mark A. Girolami
Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK

Received on August 28, 2007; revised on October 26, 2007; accepted on December 3, 2007
Advance Access publication December 5, 2007
Associate Editor: Limsoon Wong



Marginal likelihoods for the alternative pathways



44.6 ± 0.8

28.9 ± 0.3

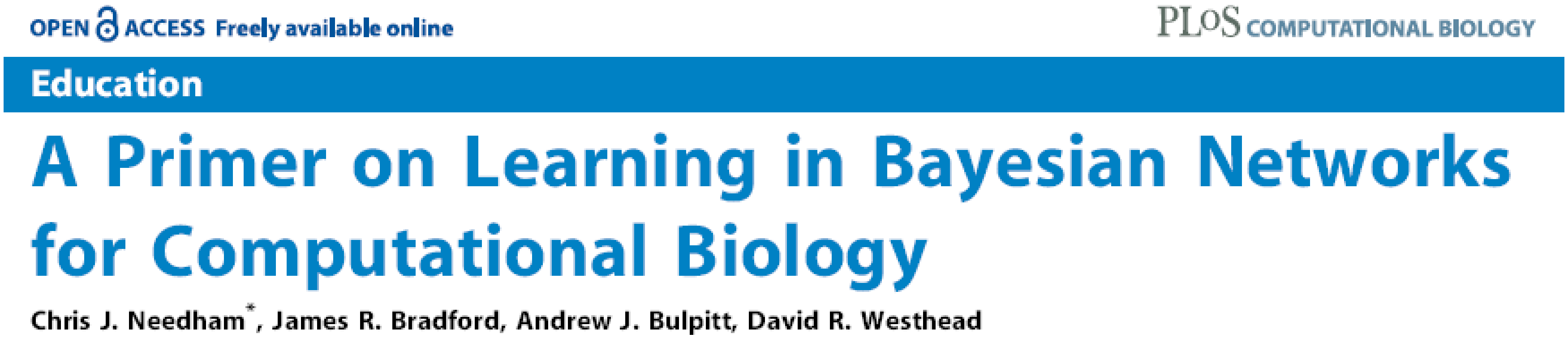
-1.1 ± 0.1

35.0 ± 0.7

Computational expensive, network reconstruction *ab initio* unfeasible

Objective: Reconstruction of regulatory networks *ab initio*

Higher level of abstraction:
Bayesian networks



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

A Primer on Learning in Bayesian Networks for Computational Biology

Chris J. Needham*, James R. Bradford, Andrew J. Bulpitt, David R. Westhead

August 2007

Volume 3

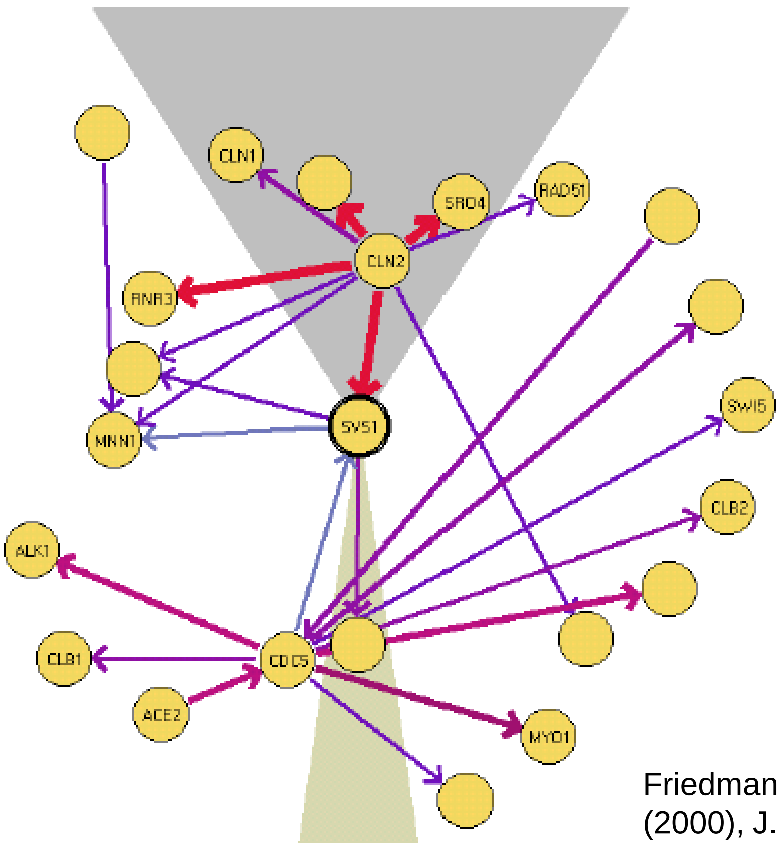
Issue 8

Marriage between

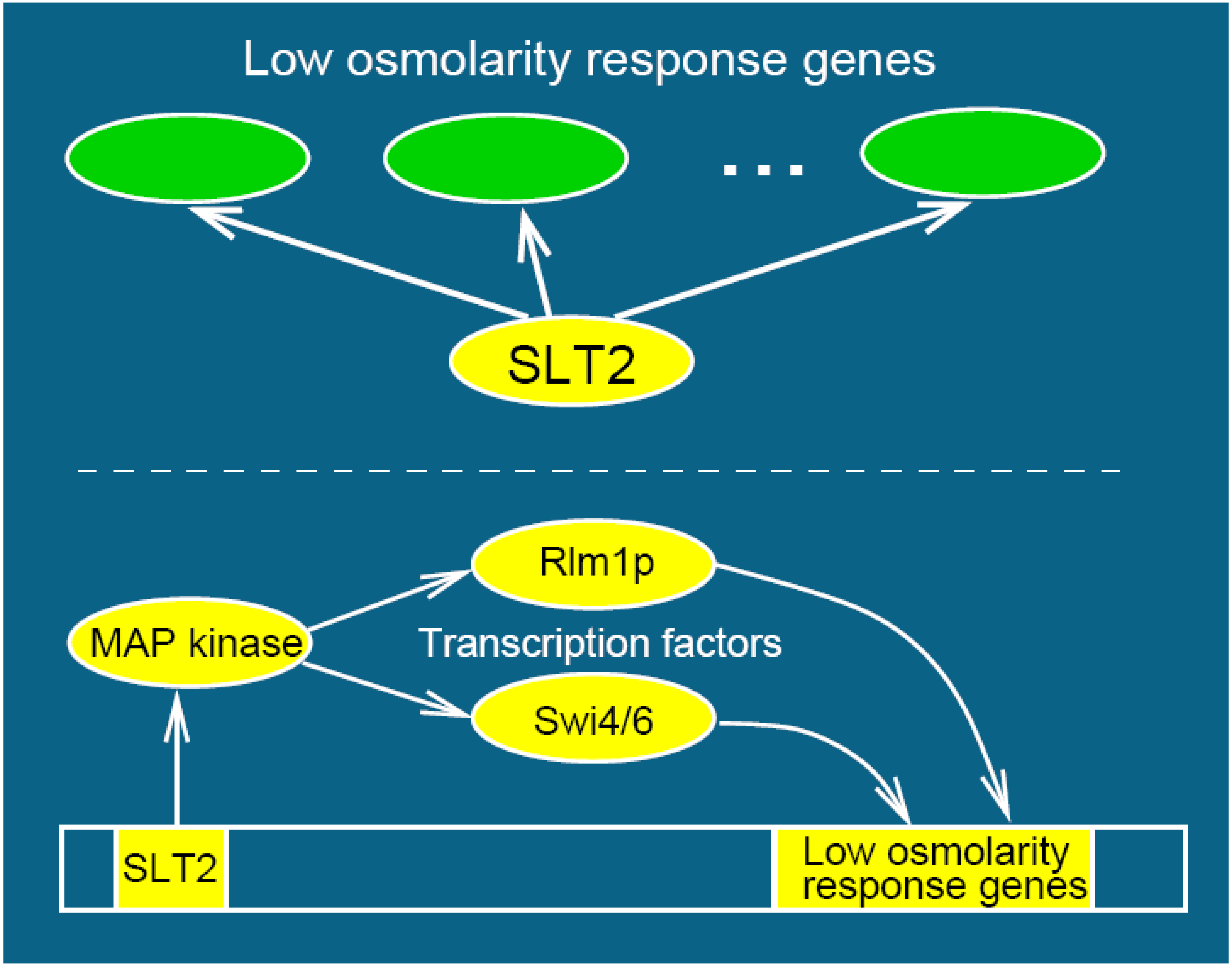
graph theory

and

probability theory



Friedman et al. (2000), J. Comp. Biol. 7, 601-620

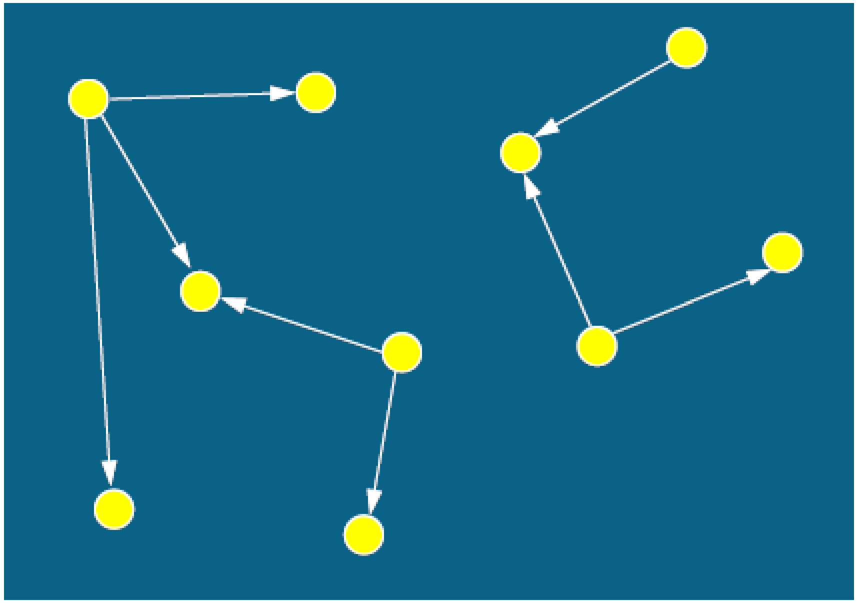


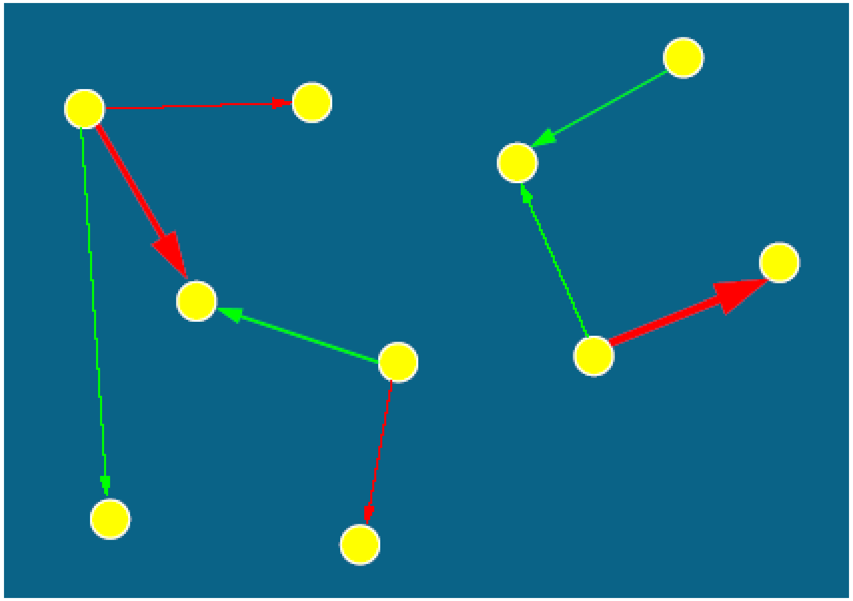
Bayes net

ODE model

Model $\mathcal{M}$

Parameters $\mathbf{q}$

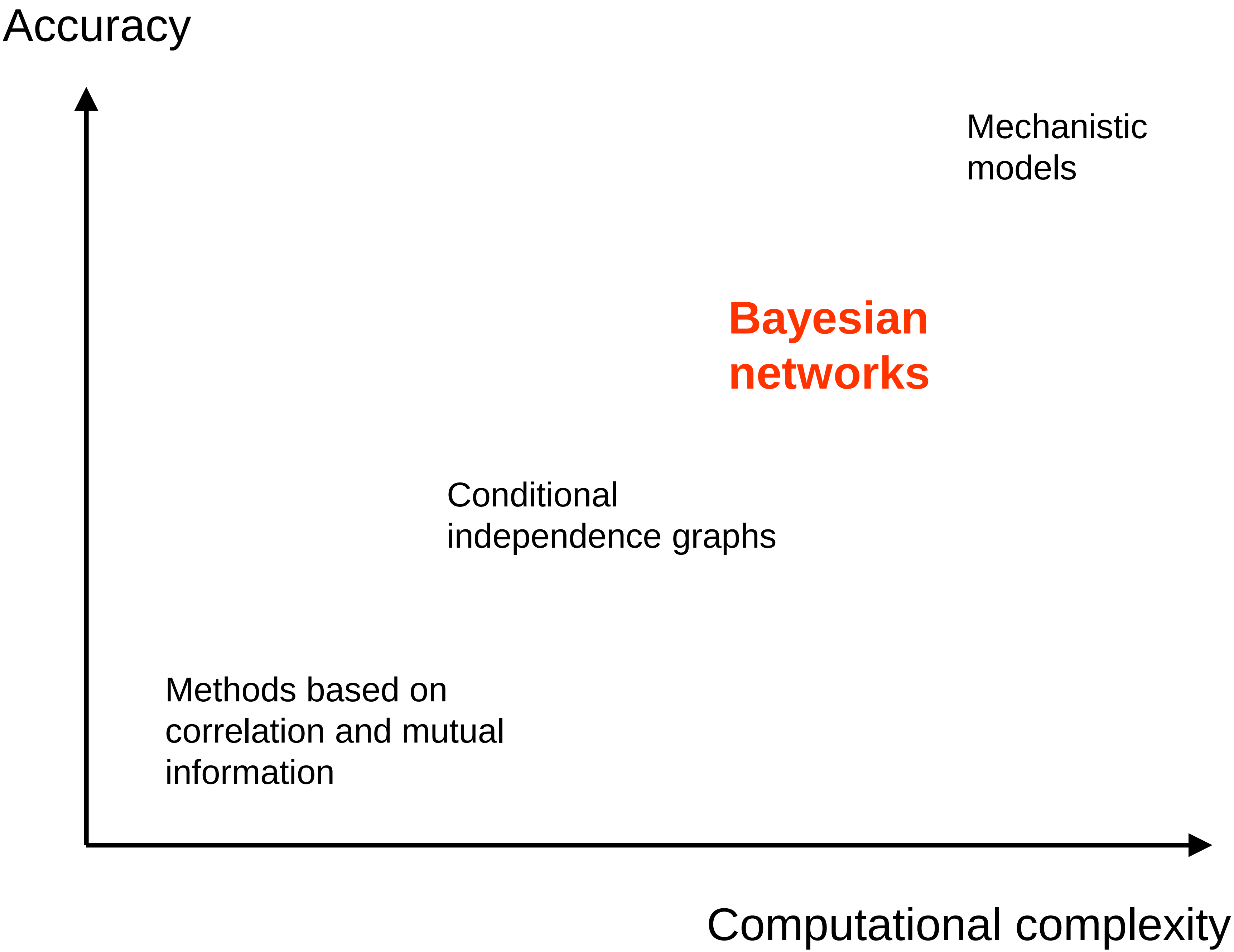




$$P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|\mathbf{q},\mathcal{M})P(\mathbf{q}|\mathcal{M})d\mathbf{q}$$

Under certain regularity conditions:

Integral analytically tractable!



Accuracy

Mechanistic models

Bayesian networks

Conditional independence graphs

Methods based on correlation and mutual information

Computational complexity