Probabilistic modelling of metabolic regulation in prokaryotes

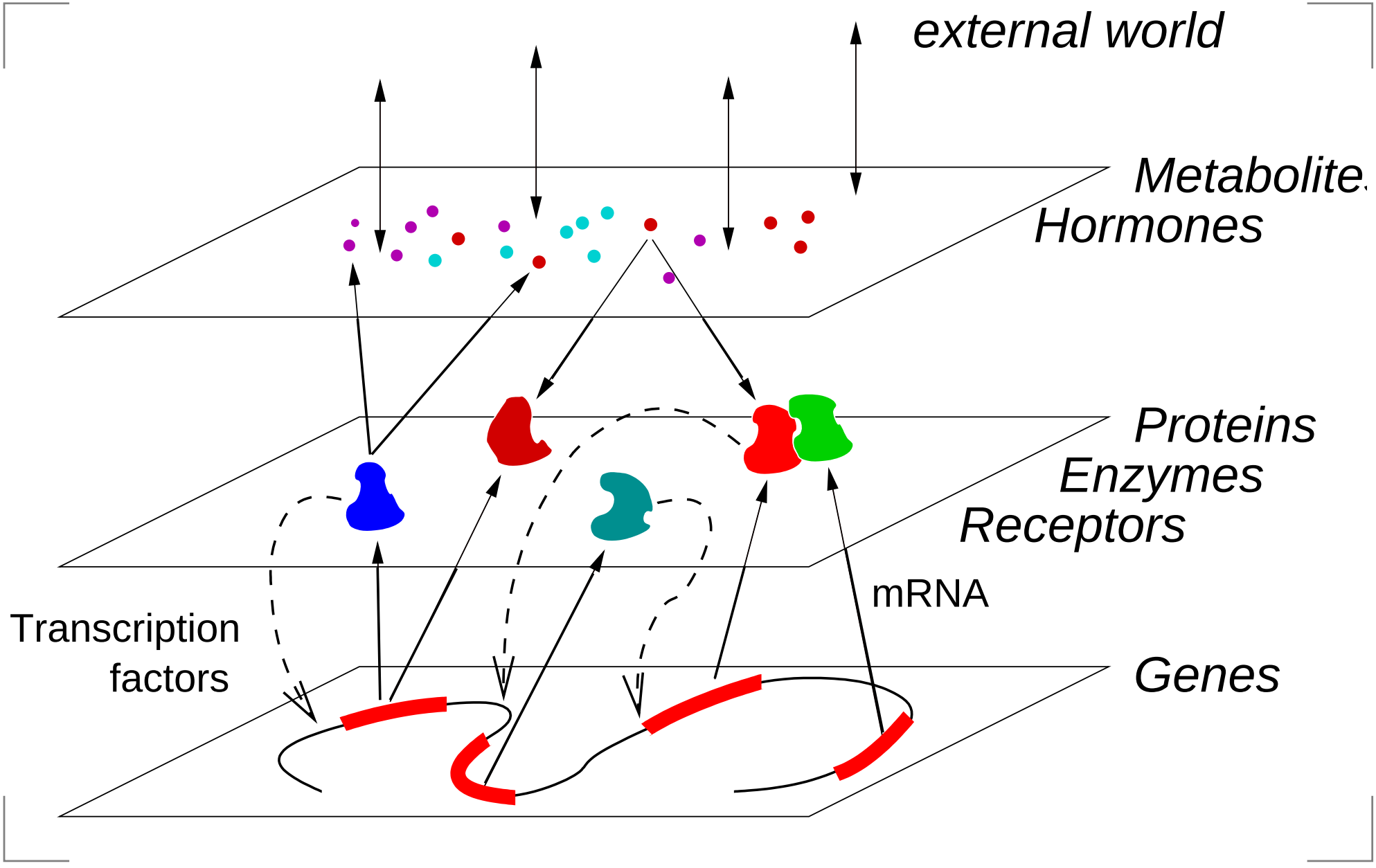
Lorenz Wernisch

School of Crystallography

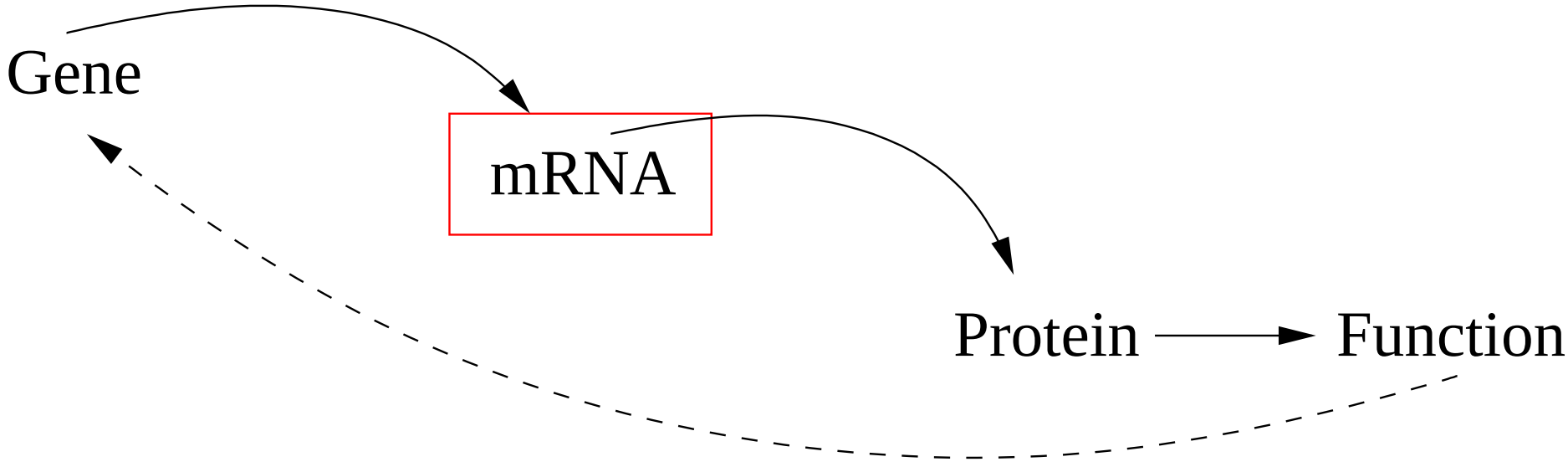
Birkbeck College

London

Regulatory networks

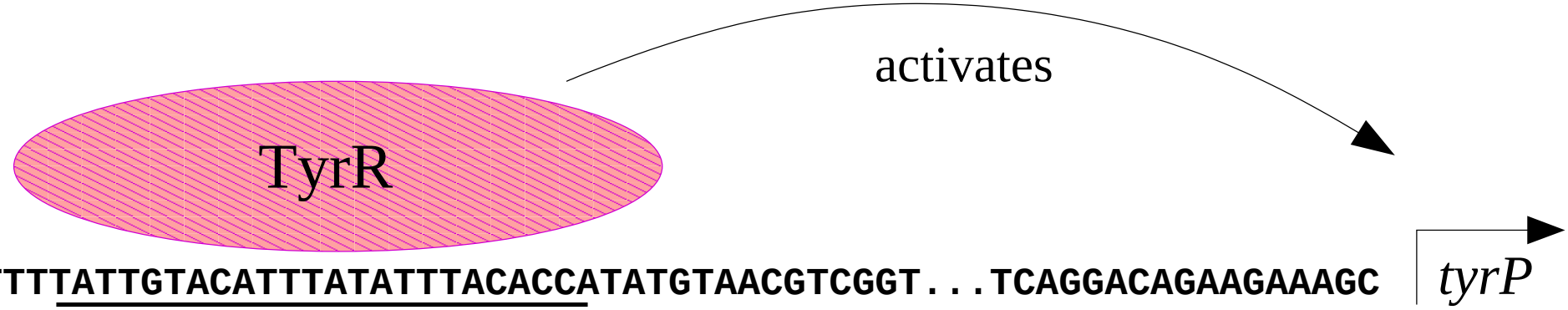


Gene expression



- Major cellular events are driven or accompanied by activated or supressed **gene expression**
- Gene activity is reflected in **mRNA concentration**
- Measure **mRNA concentration** in **microarrays** against a control

Link by transcription factors




```
TGTAAATTTATCTATACAGA aroF
TGTAAATAAAAATGTACGAA aroF
TGTAATTTATTATTTACACT aroL
TGTACATTTATATTTACACC tyrP
TTTAATTCAATTAAAACGAA aroP
TGTCAATGATTGTTGACAGA tyrR
TGTAAAATAATATATACAGC mtr
TGAAATTAATTTTAAAAAGG rbfA
TGAAATAATTAACAAACAAA aroG
```

Common regulatory motif in upstream regions of controled genes

link by known motifs, **Transfac** database

motif search (MCMC)

Microarray experiment

mRNA normal cell culture

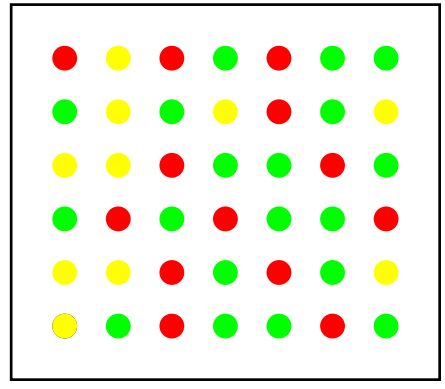
mRNA mutant/conditions

reverse transcriptase

cDNA marked with fluorescent Cy5

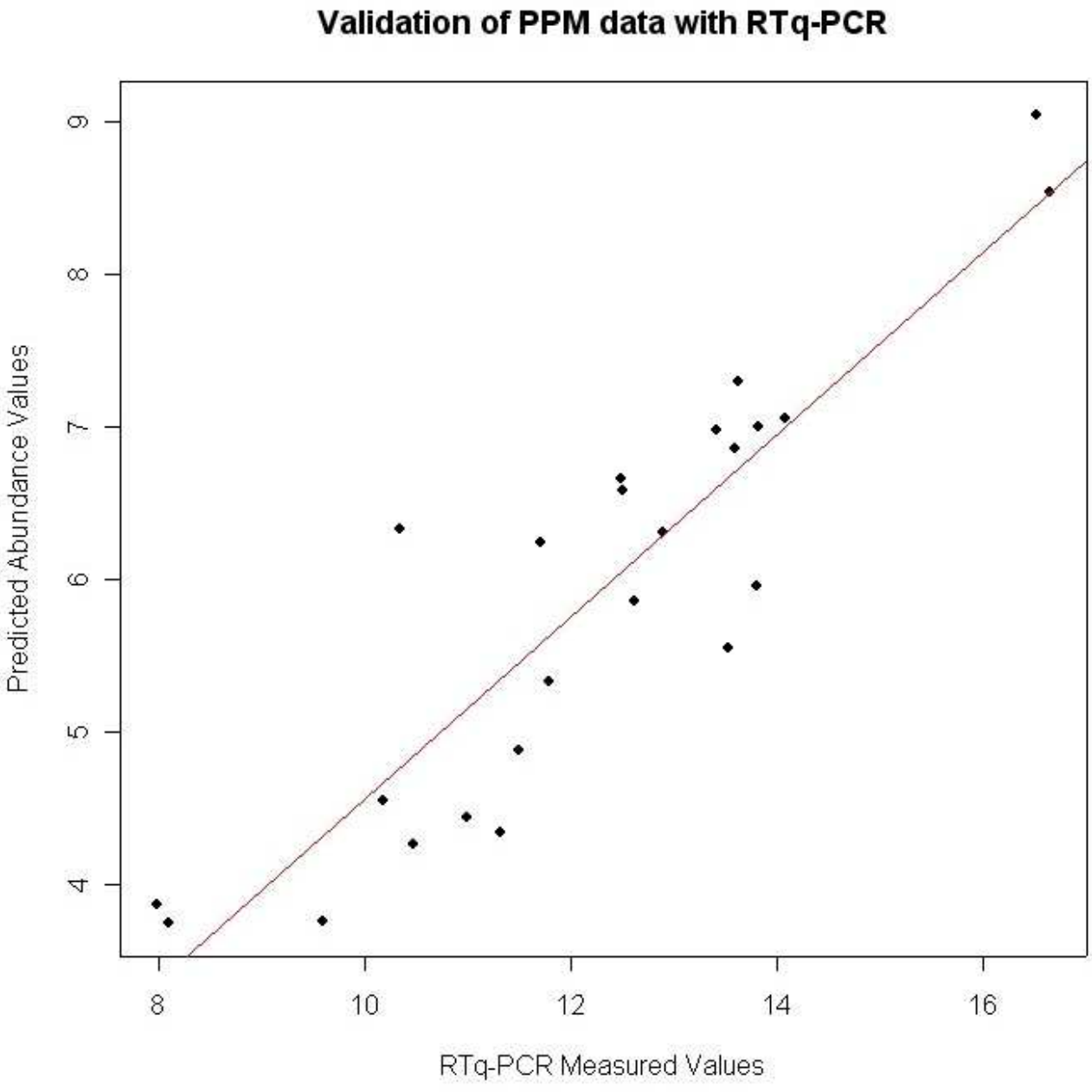
cDNA marked with fluorescent Cy3

genes are represented by DNA fragments on array



hybridization: cDNA sticks to gene on microarray

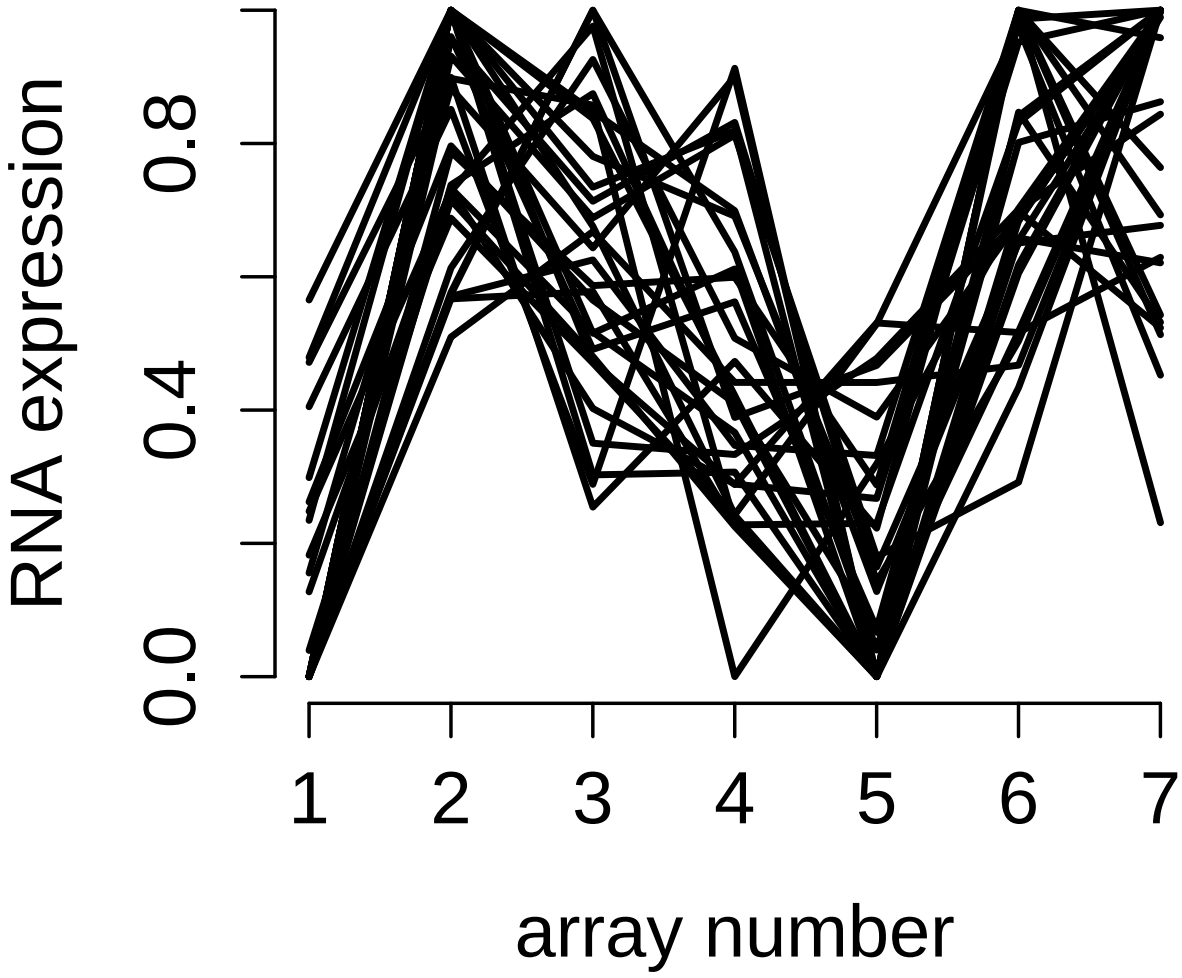
Genomic DNA as a control



Precise concentration of genomic DNA digested by restriction enzymes, pieces hybridise in equal amounts to probes

MAs agree with actual RNA concentration (Sidders and Stoker, RVC London)

RNA expression profiles

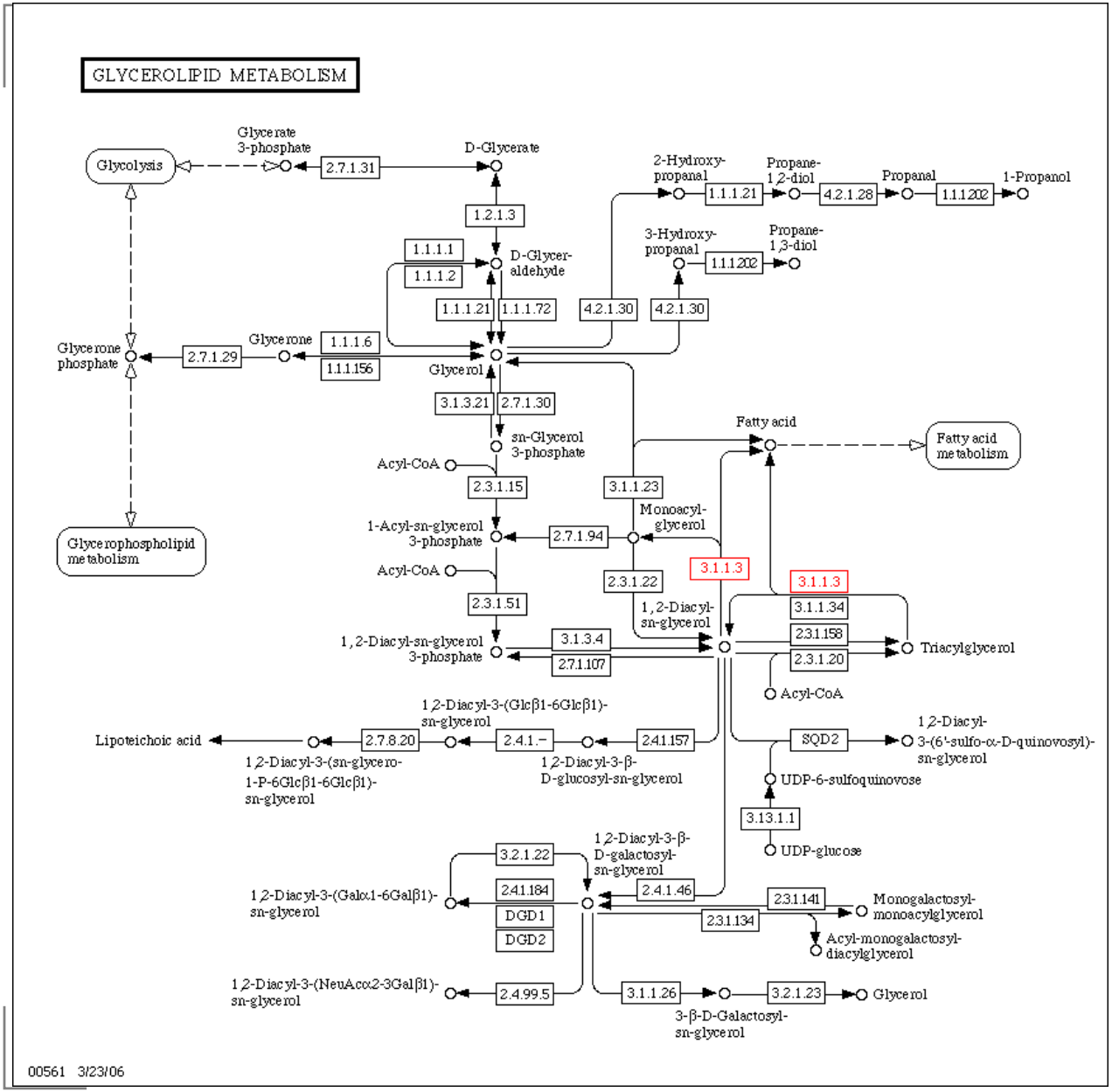


Cluster of gene profiles from time series on *Mycobacterium tuberculosis*

(J. Bacon, HPA, Porton Down)

Genes with similar **profiles** might have similar **function**

Metabolic network



KEGG database of 6810 reactions and 14238 compounds (Dec 2006)

Reactions collected in **pathways**

Pathway definition somewhat subjective and following textbook tradition

Enzymatic reactions

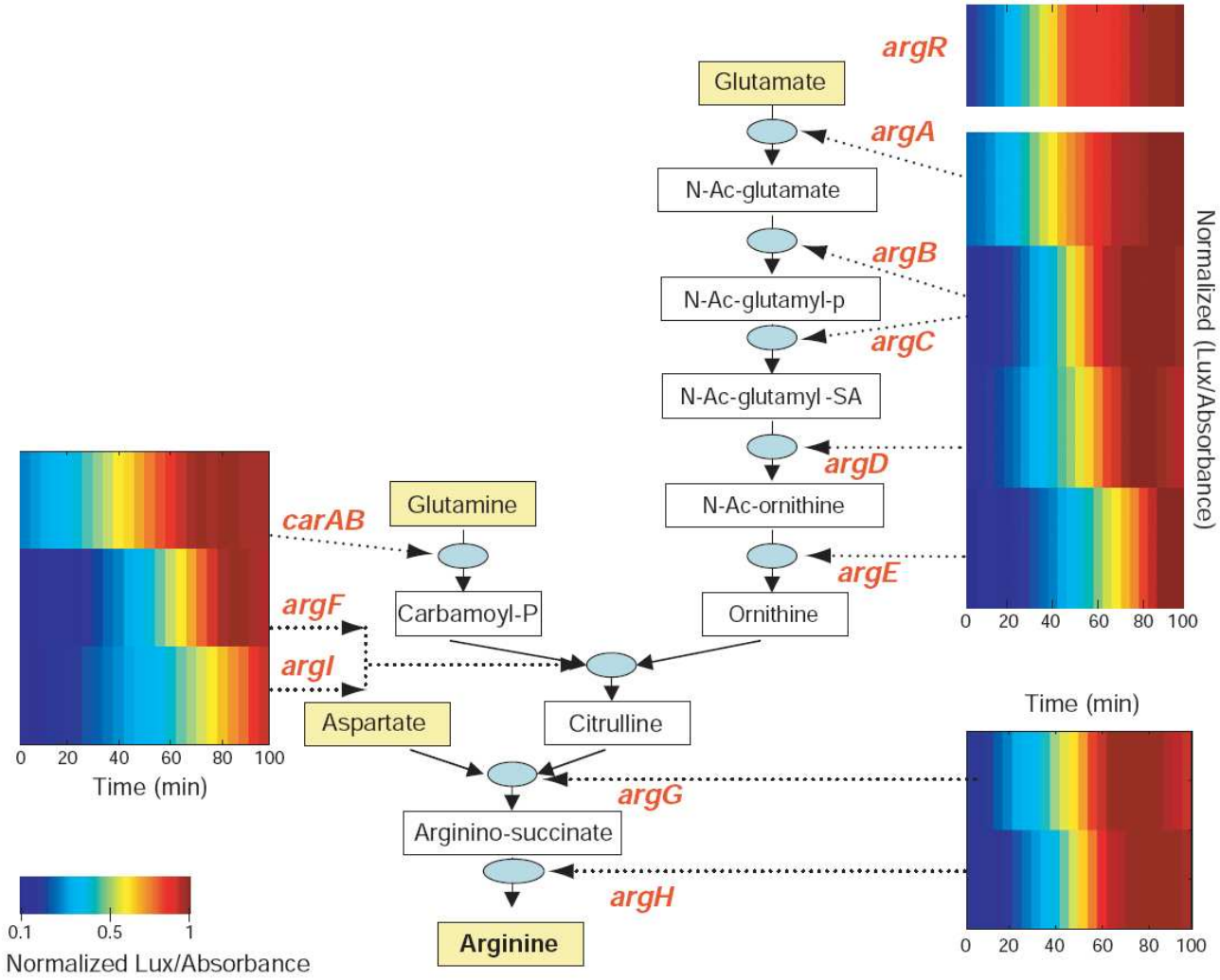
EC number provides hierarchical classification

1. Oxidoreductases
1.1 -CH-OH donor
1.1.1 NADP(P)+ acceptor
1.1.1.1 Alcohol dehydrogenase
. . .
2. Transferases
. . .

On the other hand, there are open reading frames (coding for genes) in a genome of interest: Rv0001c, . . . , Rv3916

Mapping largely unknown for *Mycobacterium tuberculosis*

Gene expression and metabolic pathways

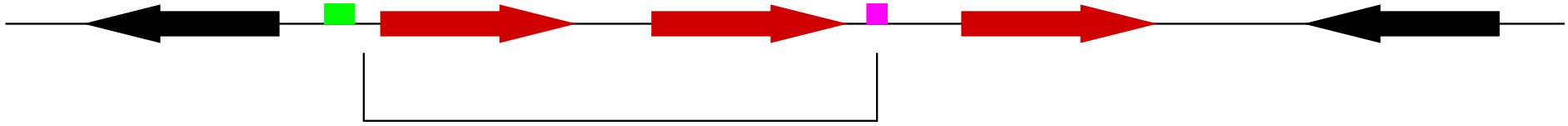


Removal of arginine activates Arg biosynthesis pathway

Measured are GFP activities linked to promoters of genes coding for enzymes

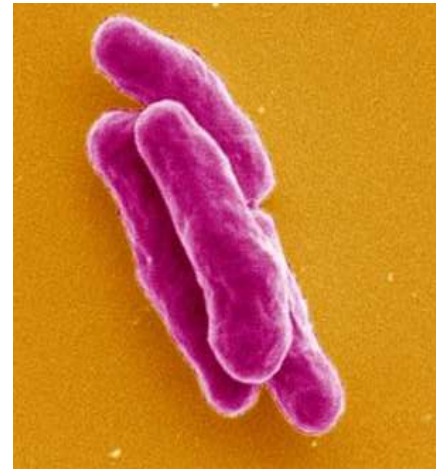
(U. Alon lab, Zaslaver et al. 04)

Operon structure



- Due to a common transcription factor binding site (green), the polymerase begins to transcribe both genes
- Due to a termination signal (magenta), transcription stops after the second gene, the two genes form an **operon**
- Longer operons might be subdivided in smaller **transcription units TU** which are always cotranscribed, while operons show some flexibility in transcription

Mycobacterium tuberculosis



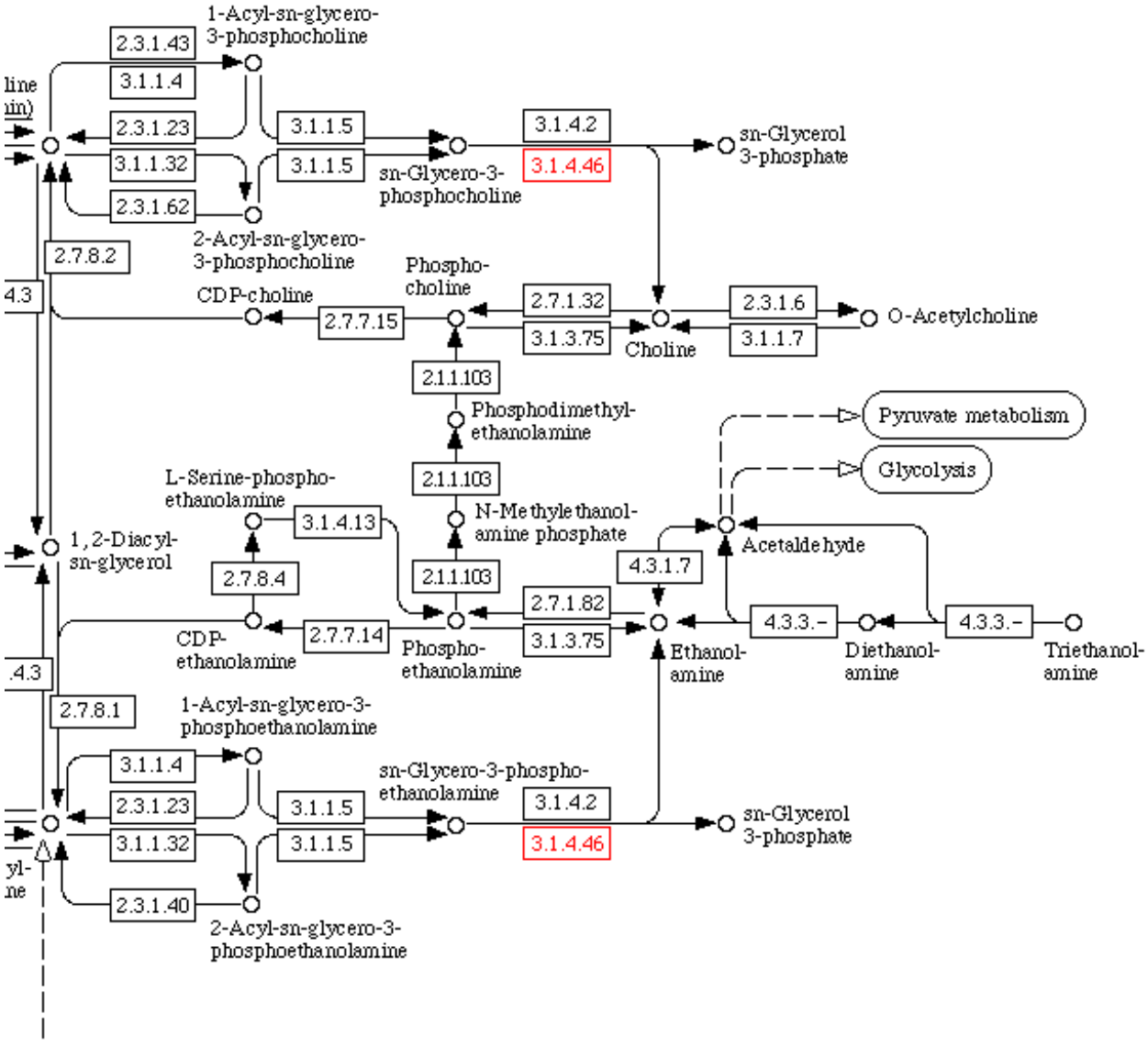
- Grows within macrophages (immune cells ingesting foreign bacteria)
- Primary infection often contained for decades: latent state (1/3 of human population)
- Highly infectious: inhaling a few cells is enough
- Slowly growing: doubling time 20 hrs (E coli 30 mins)
- Interesting lipid metabolism (complex cell wall, ability to live from lipids)
- About 3900 genes, 1200 with enzyme function (KEGG), few dozens experimentally verified

Assigning enzyme function to genes

Genome analysis strategies for assigning enzyme function (EC number) to genes in a genome:

- ENZYME database with representative proteins for each EC number (profiles in PRIAM)
- SHARKhunt (Pinney et al 05) realigns profiles (MUSCLE), creates HMMs (HMMER), search genome for matching sequences (Wise2) after prefiltering (PSI-TBLASTN)
- Results in about 1900 tentative EC number assignments for *M. tuberculosis* partly deviating from 1200 KEGG assignments

RV3842c to EC 3.1.4.46

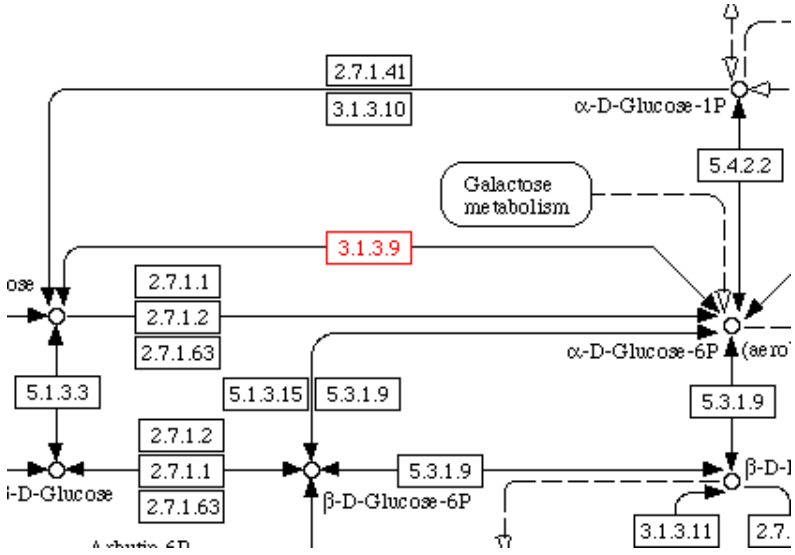


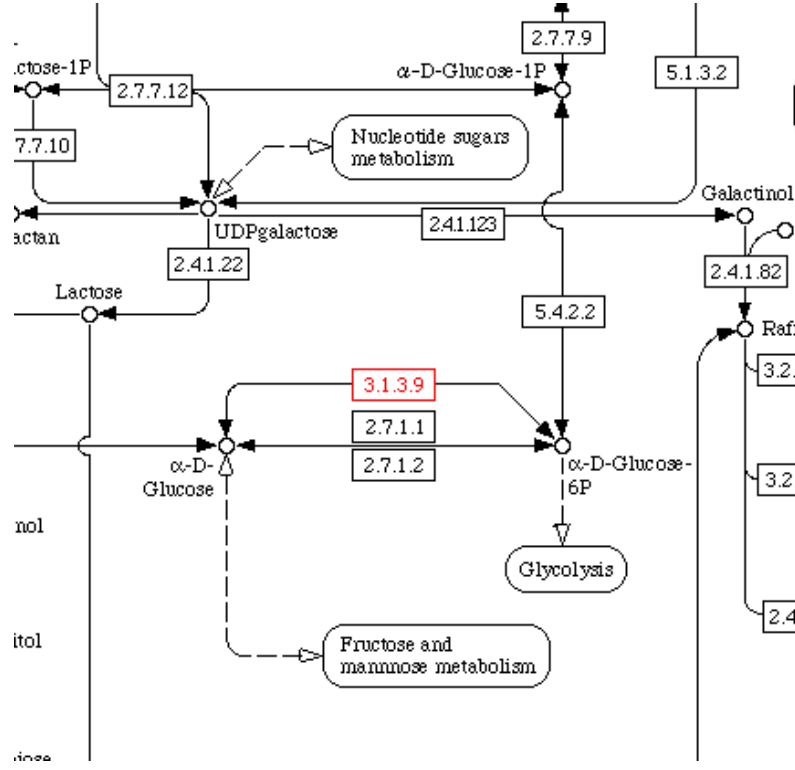
Gene Rv3842c induced under iron limitiation and mapped by SHARKhunt and KEGG to EC 3.1.4.46

Two roles in Glycerophospholipid metabolism

Rv0317c, Rv2277c, Rv3842c also mapped to same EC

RV3842c to EC 3.1.3.9

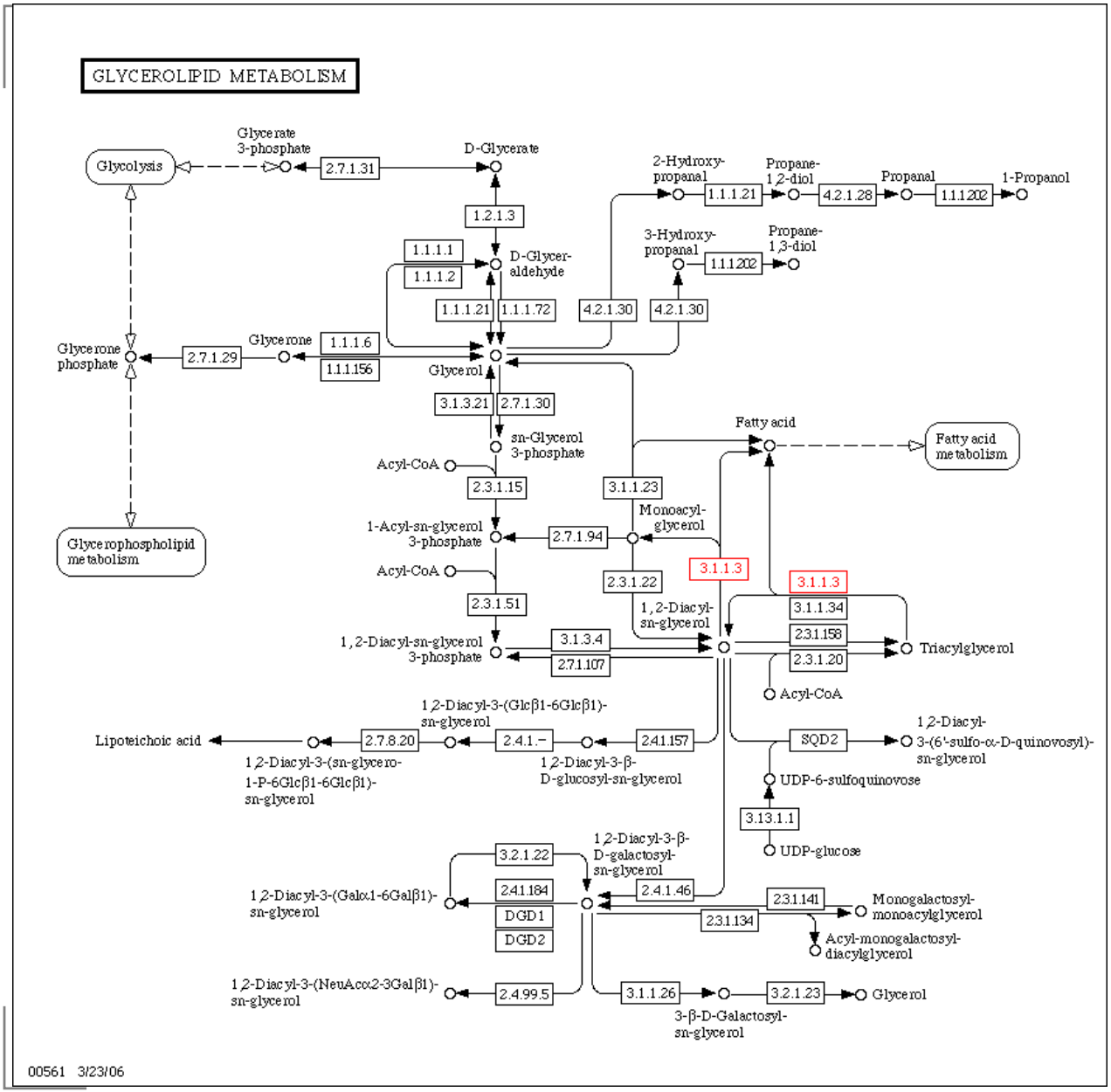




Gene Rv3842c also mapped to EC 3.1.4.46 in Glycolysis/Gluconeogenesis and Galactose metabolism (and more pathways not shown)

Rv2029c, Rv3842c also mapped to same EC number

EC 3.1.1.3



No less than 15(!) genes mapped to this enzyme with highly significant E-values by SHARKhunt

One of them Rv3097c induced under iron limitation

Questions

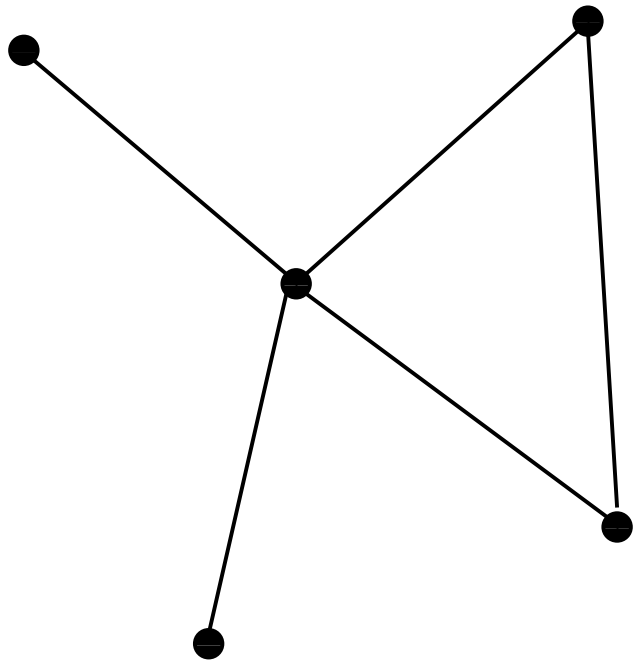
Have some information

- Gene expression (noisy)
- Metabolic pathways (precise)
- Organisation in operons (tentative)
- Transcription factor binding (very tentative)

Want to know

- Mapping of genes to enzymes
- Which connected components (pathways) of metabolic network are coregulated
- Which genes are coregulated by common transcription factor, in common operon

Markov networks



Nodes represent genes or enzymes, relation between nodes symmetric.
Edge for relations:

- Enzymes sharing metabolite
- Protein-protein interaction
- Genome neighborhood

Discrete networks: gene expression discretised (+1,0,-1), or functional classes

Continuous networks: gene expression

Probabilities in Markov nets

Potential functions $\exp(\alpha(x_i))$ for nodes $x_i \in \{0, 1\}$ and $\exp(\theta(x_i, x_j))$ for edges. Need to specify

- 2 values: $\alpha(0), \alpha(1)$
- 4 values: $\theta(0,0), \theta(1,1), \theta(0,1) = \theta(1,0)$

Probability of a configuration **very hard to compute**

$$x = (x_1, x_2, \ldots) = (1, 0, \ldots)$$

$$\mathrm{P}(x) = \frac{1}{Z} \exp\left(\sum_i \alpha(x_i) + \sum_{i,j} \theta(x_i, x_j)\right) = \frac{e^{E(x,\theta)}}{Z}$$

$Z = \sum_x \exp(E(x, \theta))$ **partition function**

Shifting $\theta(x_i, x_j)$

Shifting all $\theta(x, y)$ by the same θ_0 amounts to an extra factor $\exp(m\theta_0)$ (m number of edges in graph)

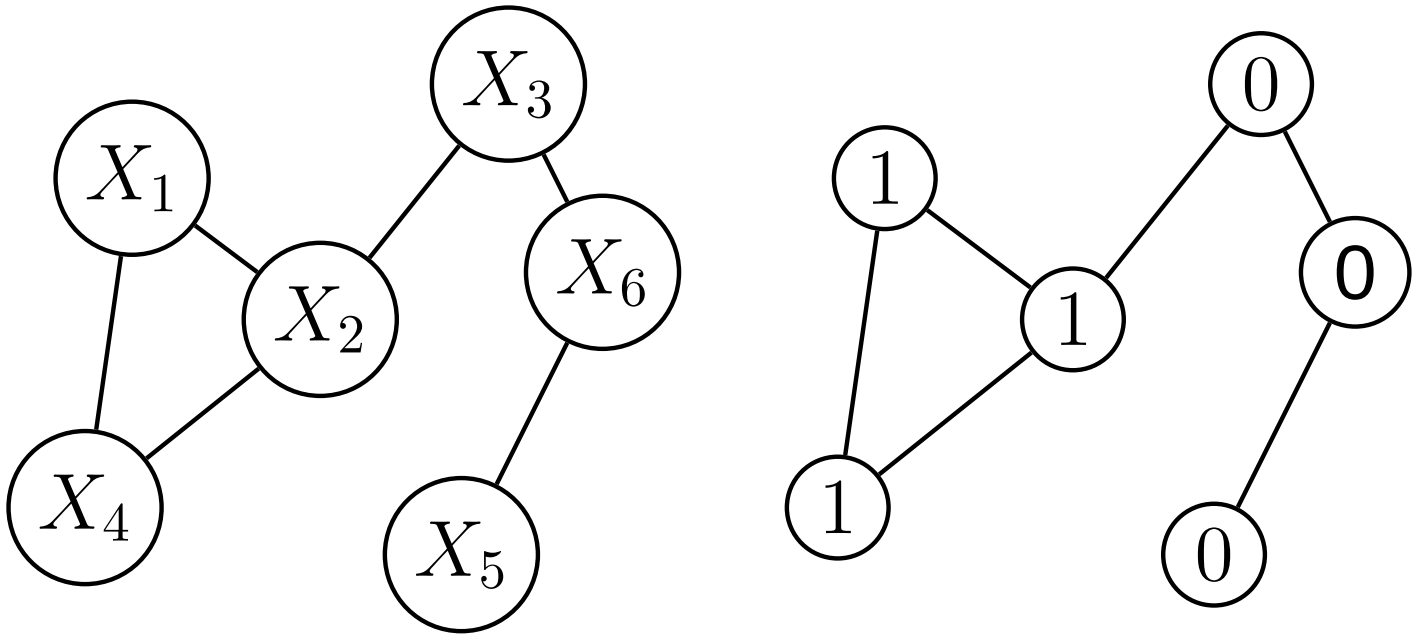
$$\frac{e^{E(x,\theta)}e^{m\theta_0}}{Ze^{m\theta_0}} = \frac{e^{E(x,\theta)}}{Z} = \mathrm{P}(x)$$

For example, shifting $\theta(x, y)$ so that

$$e^{\theta(0,0)} + e^{\theta(1,1)} + e^{\theta(0,1)} = 1$$

makes $\exp(\theta(x, y))$ comparable to frequencies of (x, y)-edges

Interpretation of parameters

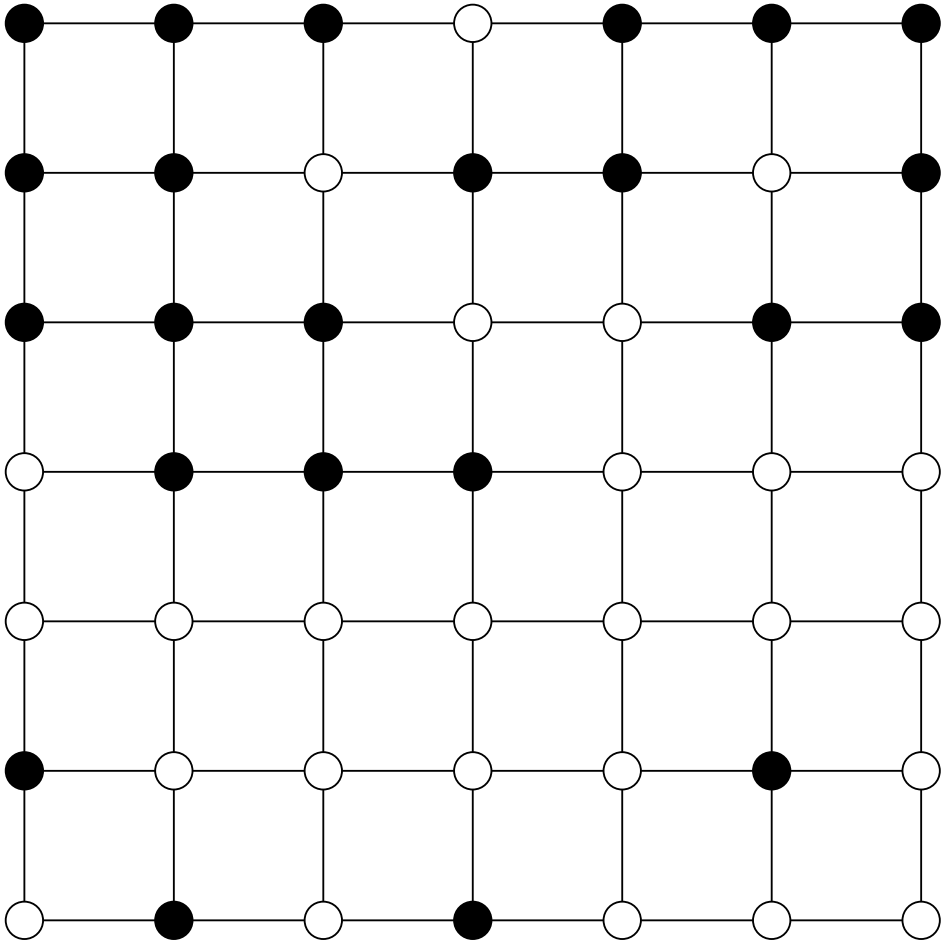


Edge frequencies: (0,0) 0.33, (1,1) 0.5, (0,1) 0.17

Markov net MLE: $e^{\theta_{00}} \approx 0.42$, $e^{\theta_{11}} \approx 0.46$, $e^{\theta_{01}} \approx 0.12$

Explanation(?): Most X configurations contain many (0,1), decreasing $e^{\theta_{01}}$ reduces Z and increases ML

Grid



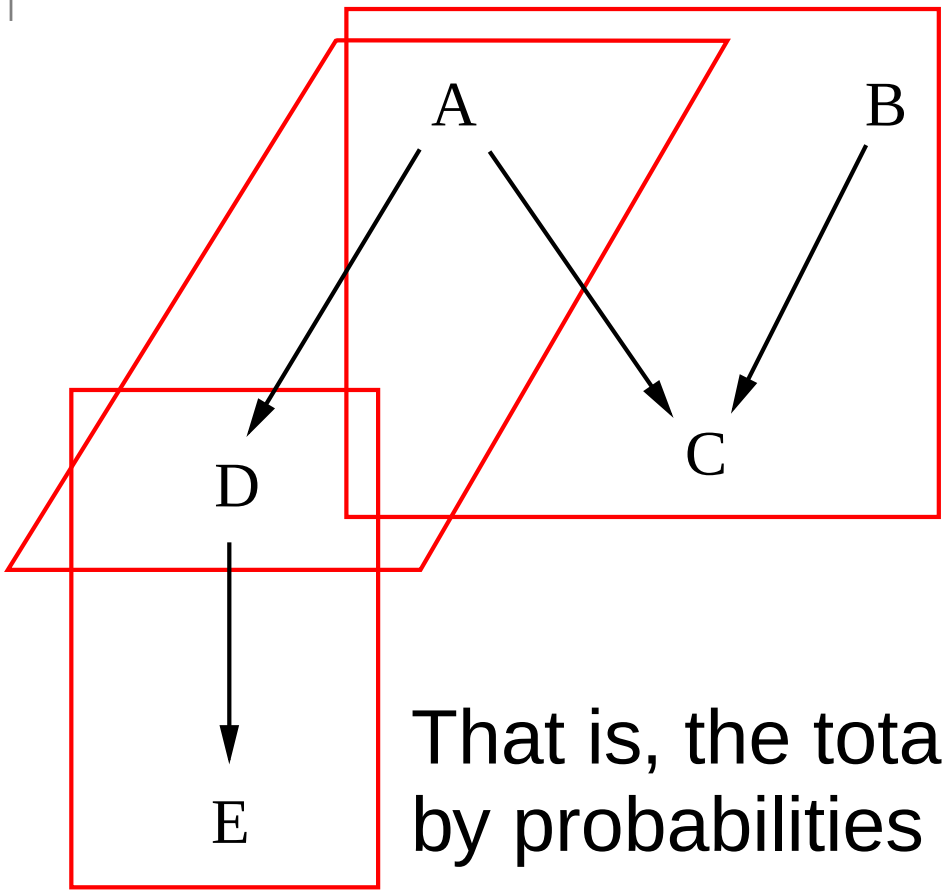
0–1: 37, 44%
0–0: 27, 32%
1–1: 20, 24%

Exact $Z(\theta)$ by dynamic programming (regular structure) with 2^7-state space gives MLE

$e^{\theta(0,1)} \approx 28\%$
$e^{\theta(0,0)} \approx 37\%$
$e^{\theta(1,1)} \approx 35\%$

Partition function takes dependencies between edge types into account!

Alternative: Bayesian networks



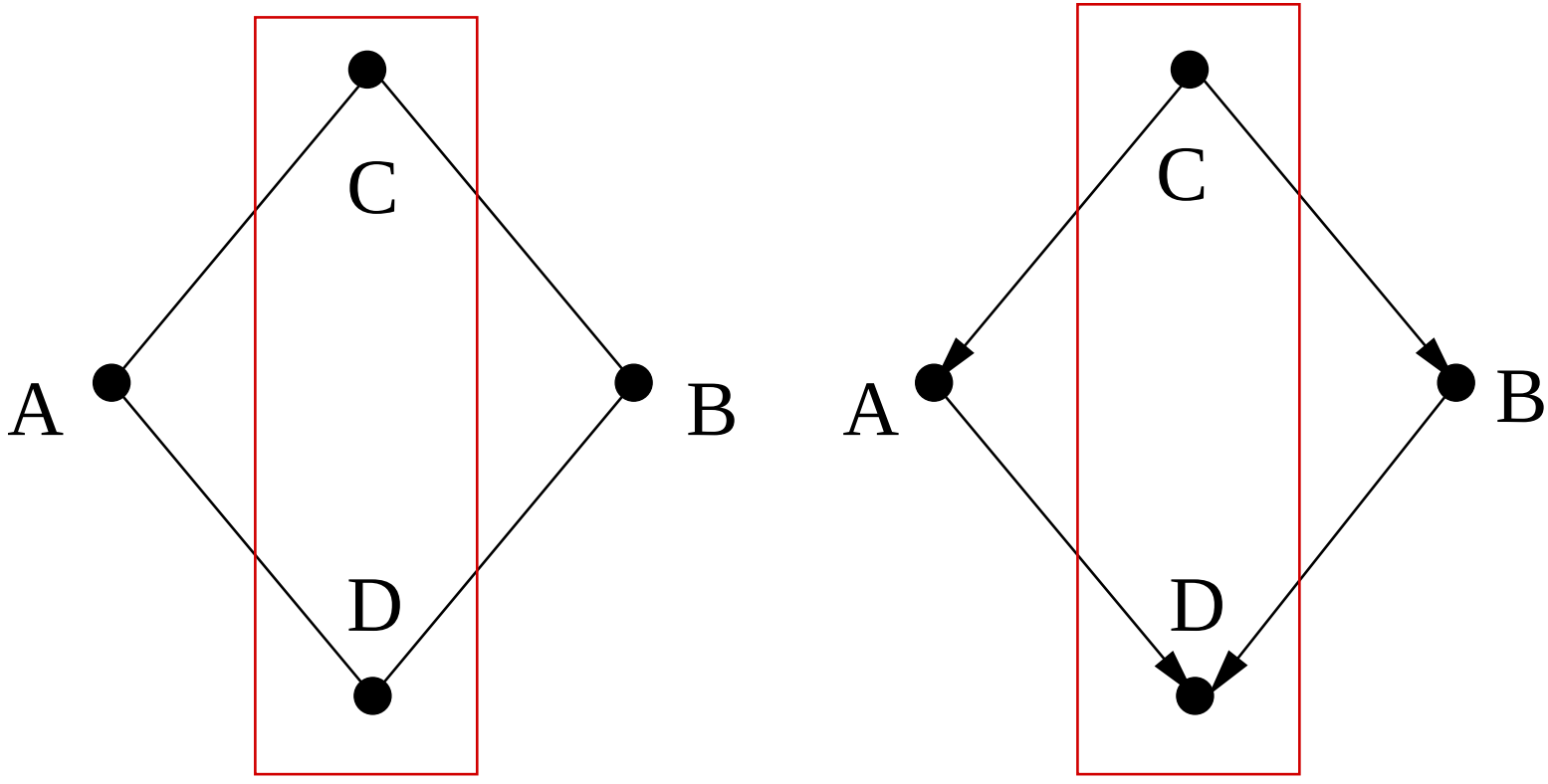
Probability of configuration **much easier to compute**

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_i \mathrm{P}(X_i | \mathrm{parents}(X_i))$$

That is, the total probability is given by probabilities on “core families”

$$\mathrm{P}(A, \ldots, E) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C|A,B)\mathrm{P}(D|A)\mathrm{P}(E|D)$$

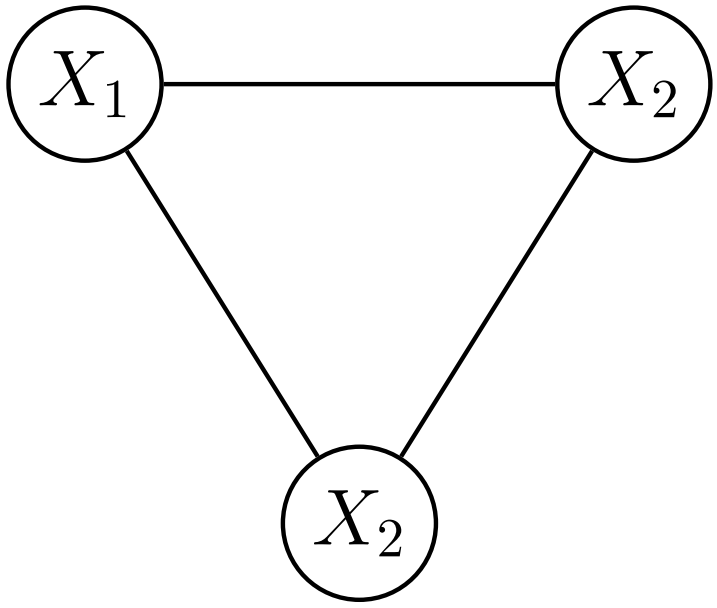
Differences Bayesian - Markov

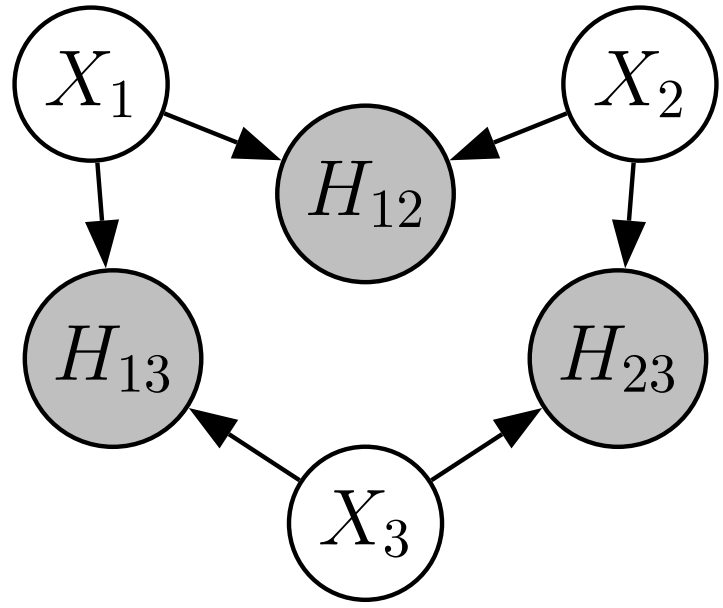


Markov net: A, B **independent** conditioned on C, D
(independence conditioned on separating node set)

Bayes net: A, B **dependent** conditioned on C, D

Markov to Bayes net





Conditioning on $H = (H_{12}, H_{13}, H_{23}) = (1, 1, 1)$:

$$P(x \mid H = 1) = \frac{\mathrm{P}(H = 1 \mid x)\,\mathrm{P}(x)}{\sum_X \mathrm{P}(H = 1 \mid X)\,\mathrm{P}(X)}$$

Sum over all configurations X is partition function Z!

Conditioning on H

The probability $P(H = 1 \mid x, y)$ gives us an interpretation of θ

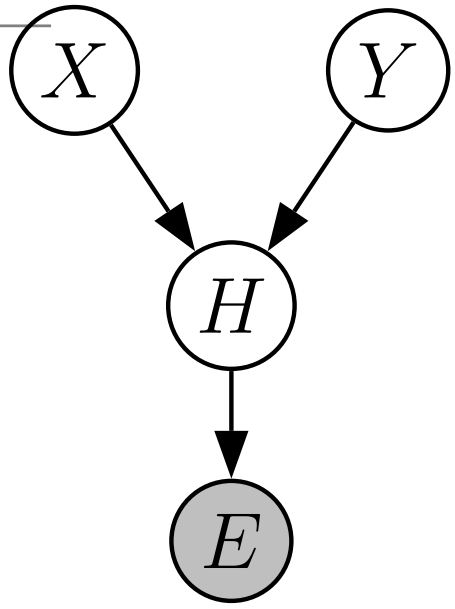
$$\theta(x, y) = \log P(H = 1 \mid x, y)$$

Conditioning on x makes H_{ij} independent, split into product

$$P(x \mid h, \theta) = \frac{1}{Z(\theta)} \mathrm{P}(x) \prod_{h_{ij}=1} e^{\theta(x_i, x_j)}$$

$\theta(x, y)$ either from independent information source or obtained by parameter estimation (up to a common factor for all x, y)

Indirect evidence



Evidence E for hidden H:

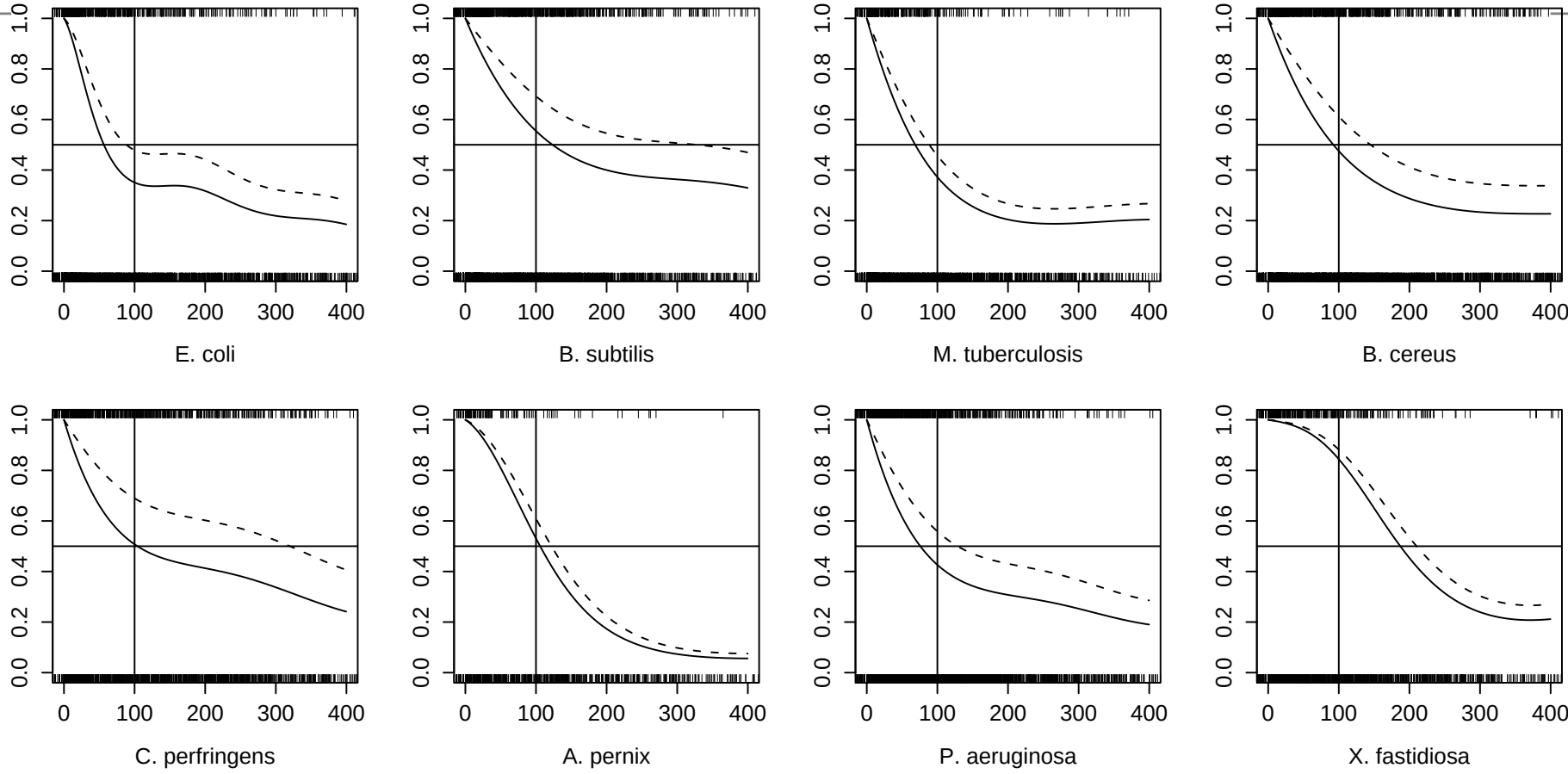
True positive rate
$\lambda_1 = \mathrm{P}(E = 1 \mid H = 1)$

False positive rate
$\lambda_0 = \mathrm{P}(E = 1 \mid H = 0)$
(from independent source)

$$\begin{aligned}\mathrm{P}(E = 1 \mid X, Y) &= \mathrm{P}(E = 1 \mid H = 1)\,\mathrm{P}(H = 1 \mid X, Y) \\ &\quad + \mathrm{P}(E = 1 \mid H = 0)\,\mathrm{P}(H = 0 \mid X, Y) \\ &= \lambda_0 + e^{\theta(X,Y)}(\lambda_1 - \lambda_0)\end{aligned}$$

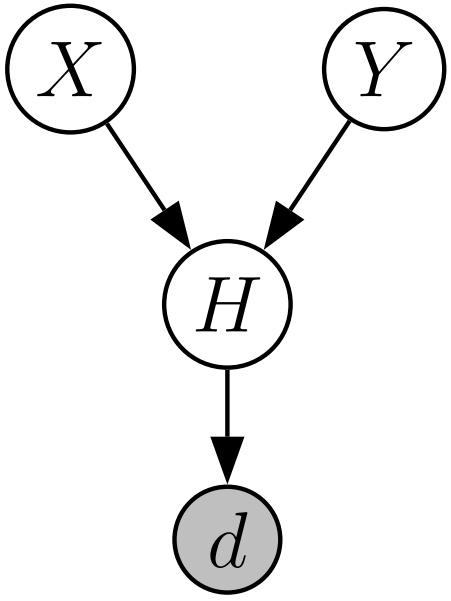
$\theta(x, y)$ **has to lie within** $[\log \lambda_0, \log \lambda_1]$ **for all** x, y**!**
If $\lambda_0 = \lambda_1$, no information from edge at all!

Evidence for operon pair



Posterior odds of gene pair being in operon given distance d in bp from synteny data (Edwards and LW, 2005): $\lambda_1(d) = P(d \mid O = 1)$ and $\lambda_0(d) = P(d \mid O = 0)$

Indirect continuous evidence



Value d as indicator of hidden H:

$H = 1$ distribution
$\lambda_1(d) = \mathrm{P}(d \mid H = 1)$

$H = 0$ distribution
$\lambda_0(d) = \mathrm{P}(d \mid H = 0)$
(from independent source)

$$P(x \mid d, \theta) = \frac{\mathrm{P}(d \mid x, \theta)\,\mathrm{P}(x)}{\sum_X \mathrm{P}(d \mid X, \theta)\,\mathrm{P}(X)}$$

$$= \frac{1}{Z(\theta, d)}\,\mathrm{P}(x) \prod_{d_{ij}} (\lambda_0(d_{ij}) + e^{\theta_{x_i x_j}} (\lambda_1(d_{ij}) - \lambda_0(d_{ij})))$$

Estimation of θ trickier, but methods below still work

Parameter estimation for Markov nets

Likelihood of θ (in MLE, MAP, or MCMC)

$$P(x \mid \theta) = \frac{1}{Z(\theta)} \exp(\sum_{edges} \theta_{ij}(x_i, x_j)) = \frac{1}{Z(\theta)} \exp(f_\theta(x))$$

Need to tackle partition function
$Z(\theta) = \sum_X \exp(\sum_{edges} \theta_{ij}(X_i, X_j))$

- Variational mean field
- Pseudolikelihood
- Loopy belief propagation

Many more suggested and analysed by Murray and Ghahramani 2005 "Doubly intractable distributions"

Variational bound

Variational lower bound on $\log Z(\theta)$

$$\log Z(\theta) = \log \sum_X \exp(f_\theta(X)) = \log \sum_X q(X) \frac{\exp(f_\theta(X))}{q(x)}$$

$$\geq \sum_X q(X) \log \exp(f_\theta(X)) - \sum_X q(X) \log q(X)$$

$$= E_q(f_\theta(X)) + H(q)$$

(using Jensens inequality with $q(X)$ any distribution)

Now find good $q(x)$ that maximises this lower bound, eg., from an arbitrary but "easy" family

Mean field approximation

For example, q **factorising over variables**:
$q(x) = \prod_i q(x_i)$ where $q\colon\ x_i = 1$ with probability q_i

$$E_q(\sum_{edges} \theta_{ij}(x_i, x_j))$$

$$= \sum_{edges} \theta_{ij}(1,1)q_i q_j + \theta_{ij}(0,0)(1-q_i)(1-q_j) + \cdots$$

$$H(q(x)) = \sum_i H(q(x_i))$$

$$= -\sum_i q_i \log q_i + (1-q_i)\log(1-q_i)$$

Easy to (numerically) optimise q_i

Pseudolikelihood, belief propagation

Pseudolikelihood circumvents calculating $Z(\theta)$

$$\mathrm{P}(x \mid \theta) \approx \prod_i \mathrm{P}(x_i \mid \text{neighbors of } x_i, \theta)$$

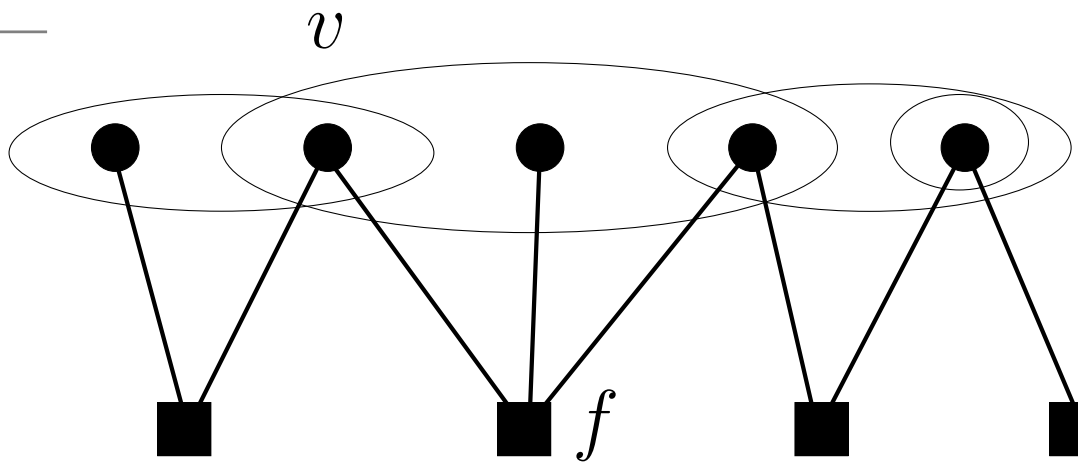
but results in low quality approximation

Bethe approximation of log partition function

$$\log Z(\theta) \approx E_{\mathrm{Bethe}} + H_{\mathrm{Bethe}}$$

approximating average E and entropy H obtained from **loopy belief propagation**

Belief propagation



variables

factors (edges)

$$m(v \to f)(x_v) = \prod_{\text{neigh } f' \neq f} n(f' \to v)(x_v)$$

$$n(f \to v)(x_v) = \sum_{x_f \setminus x_v} e^{g_f(x_f)} \prod_{\text{neigh } v' \neq v} m(v' \to f)(x_{v'})$$

With $g_f(x_f) = \theta_{ij}(x_i, x_j)$ for factor (edge) (i, j).
Exact on tree structures.

Bethe approximation

Run belief propagation until convergence, then:

$$b(x_v) \propto \prod_{\text{neigh } f} n(f \to v)(x_v)$$

$$b(x_f) \propto e^{f(x_f)} \prod_{\text{neigh } v} m(v \to f)(x_v)$$

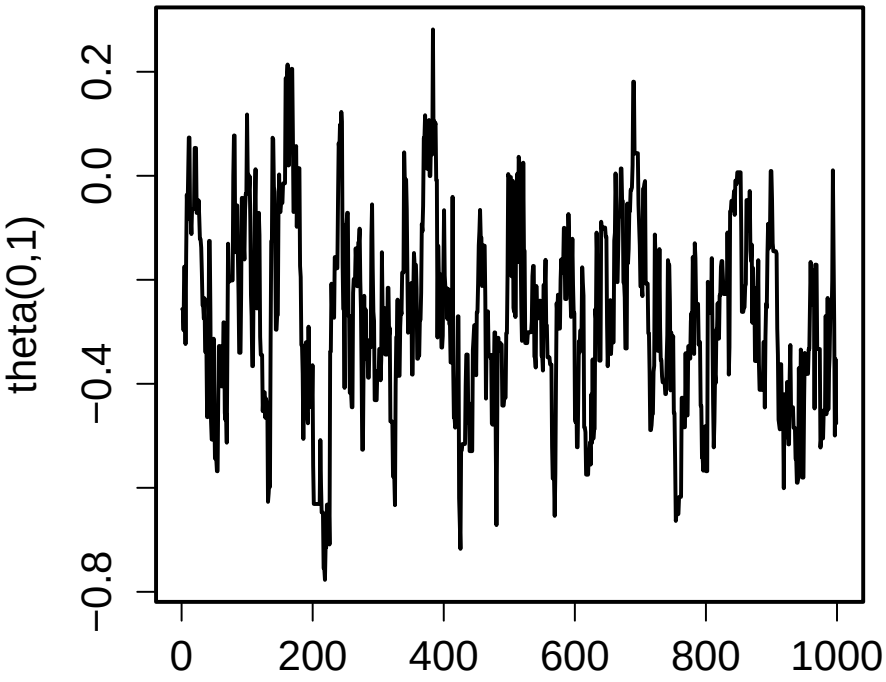
$$U_{\text{Bethe}} = -\sum_f \sum_{x_f} b(x_f) g_f(x_f)$$

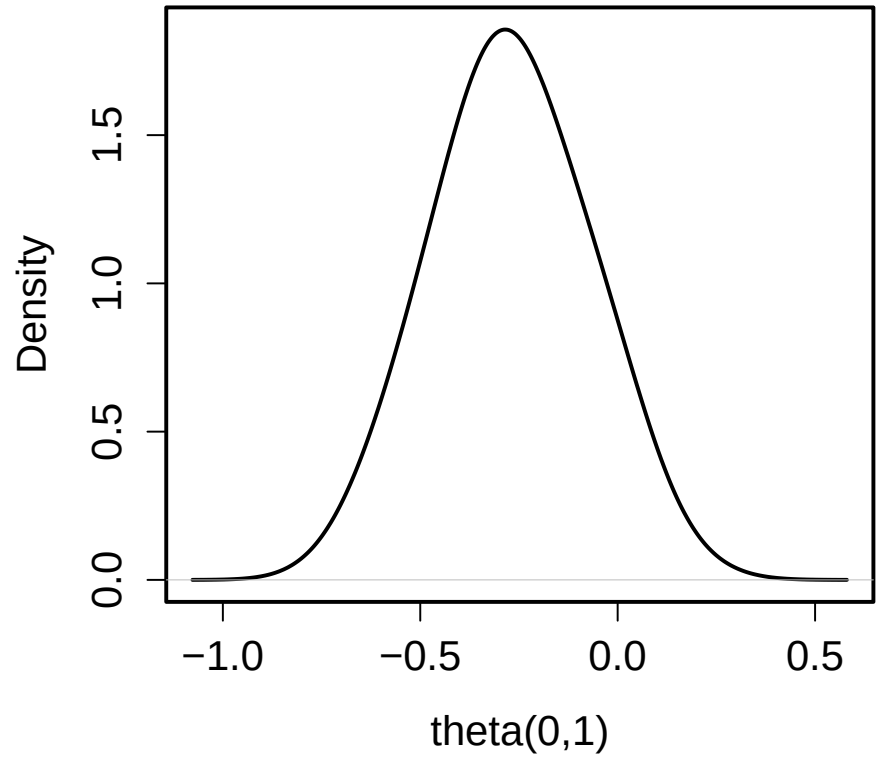
$$H_{\text{Bethe}} = -\sum_f \sum_{x_f} b(x_f) \log b(x_f)$$

$$+ \sum_v (d_v - 1) \sum_{x_v} b(x_v) \log b(x_v)$$

Parameter estimation: 7×7-grid

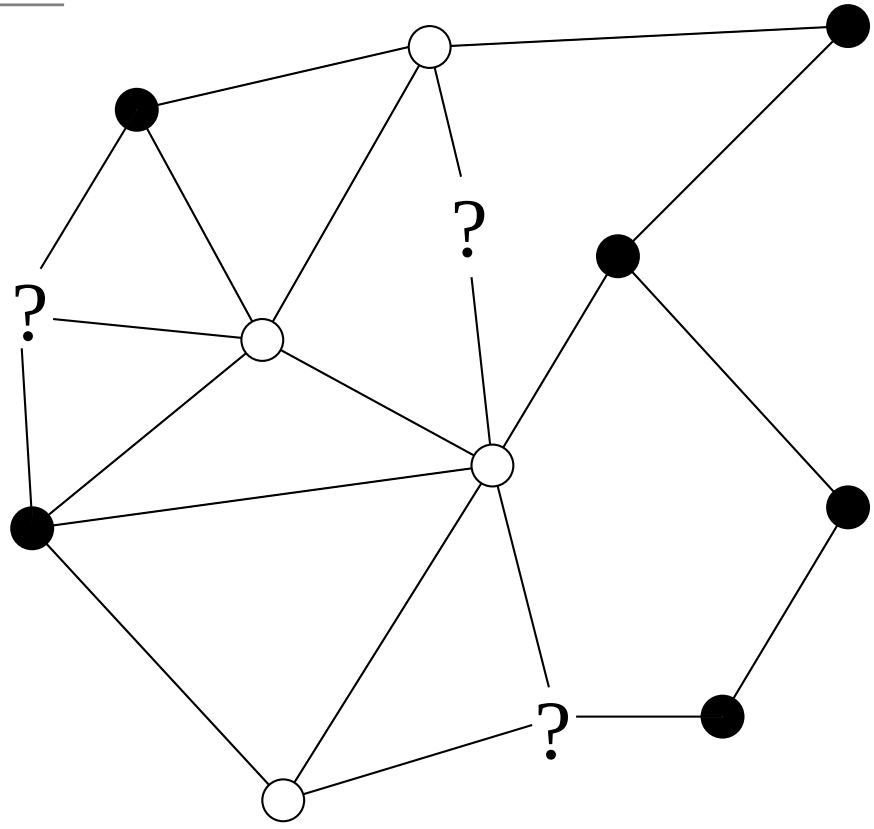
MCMC with belief propagation:





Exact: -0.2532, -0.0604
Bethe: -0.2576, -0.0597
Mean field: -0.2493, -0.0611
Pseudolikelihood: -0.191 -0.063

Hidden labels



Markov independence allows to assign labels to new nodes based on neighbors:

$$\mathrm{P}(x_i \mid \text{neighbors of } x_i)$$

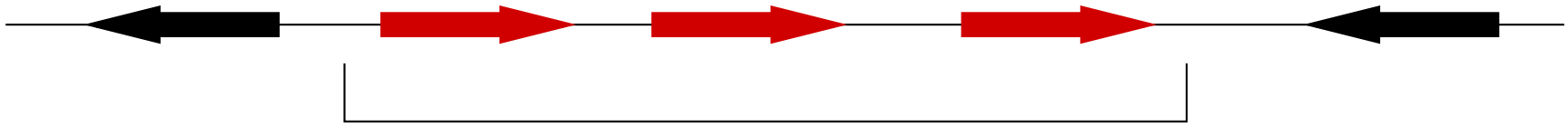
Training of parameters and hidden labels together MLE through Expectation Maximisation or MCMC

Functional assignment in E coli

4 sources of information on Escherichia coli

Two proteins related if

- part of the same **protein complex** (Genequiz)
- in the same **metabolic pathway** (EcoCyc)
- in the same **directon** in genome
- **expression profile correlation** high (7 time points, UV radiation, SMD)



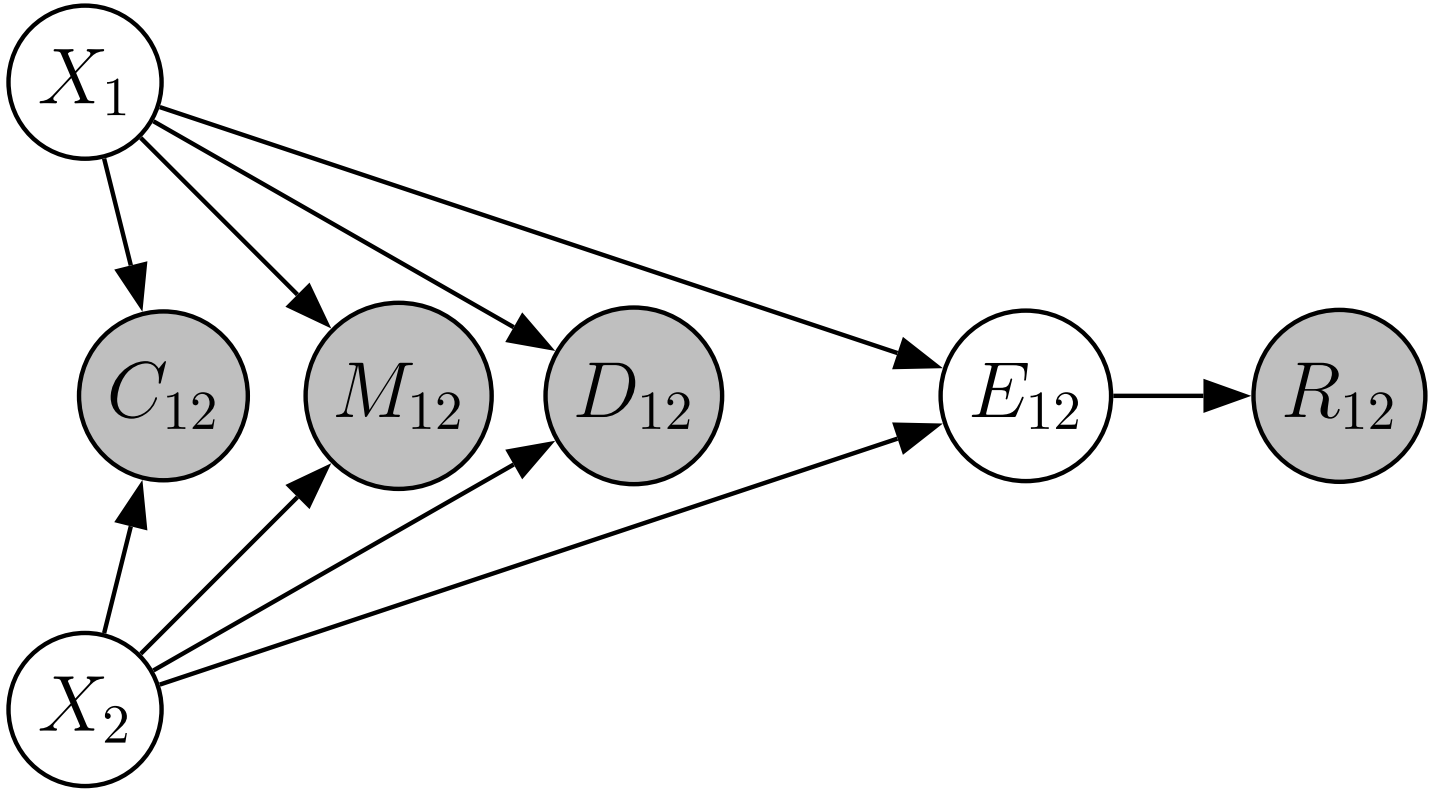
Functional classes of E coli

509 genes in 13 functional classes:

(1) Amino acid biosynthesis/23, (2) Biosynthesis of cofactors/15, (3) Cell envelope/55, (4) Cellular processes/42, (5) Central intermediary metabolism/31, (6) Energy metabolism/50, (7) Fatty acid and phospholipid metabolism/18, (8) Purines pyrimidines nucleosides and nucleotides/29, (9) Regulatory functions/101, (10) Replication/34, (11) Transcription/13, (12) Translation/7, (13) Transport and binding proteins/91

(Blattner)

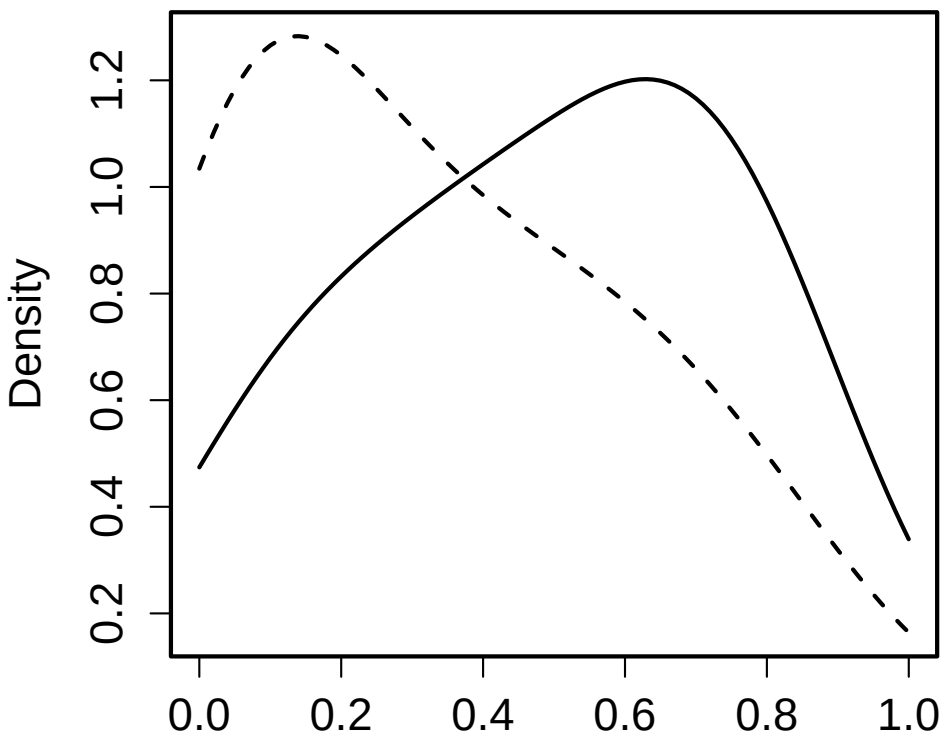
Combining information



Assumption: information sources independent conditioned on functional class

Parameter estimation by MLE and mean field approach

Soft evidence: expression correlation



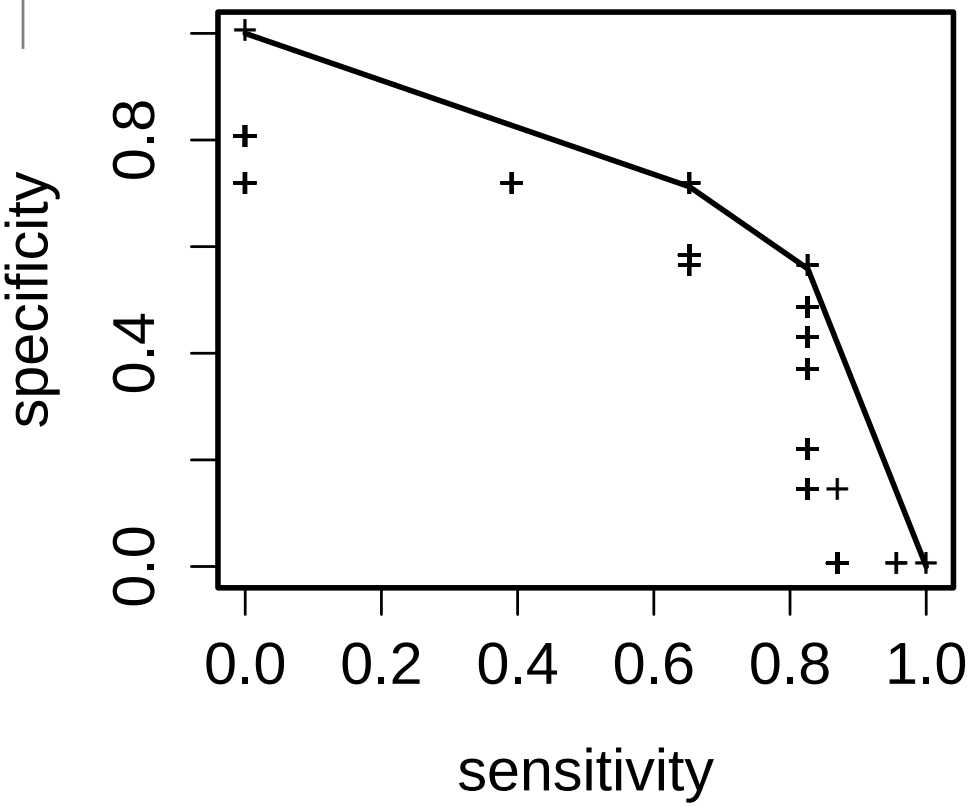
$\lambda_1(c^2)$ correlation density of gene expression profiles within functional classes

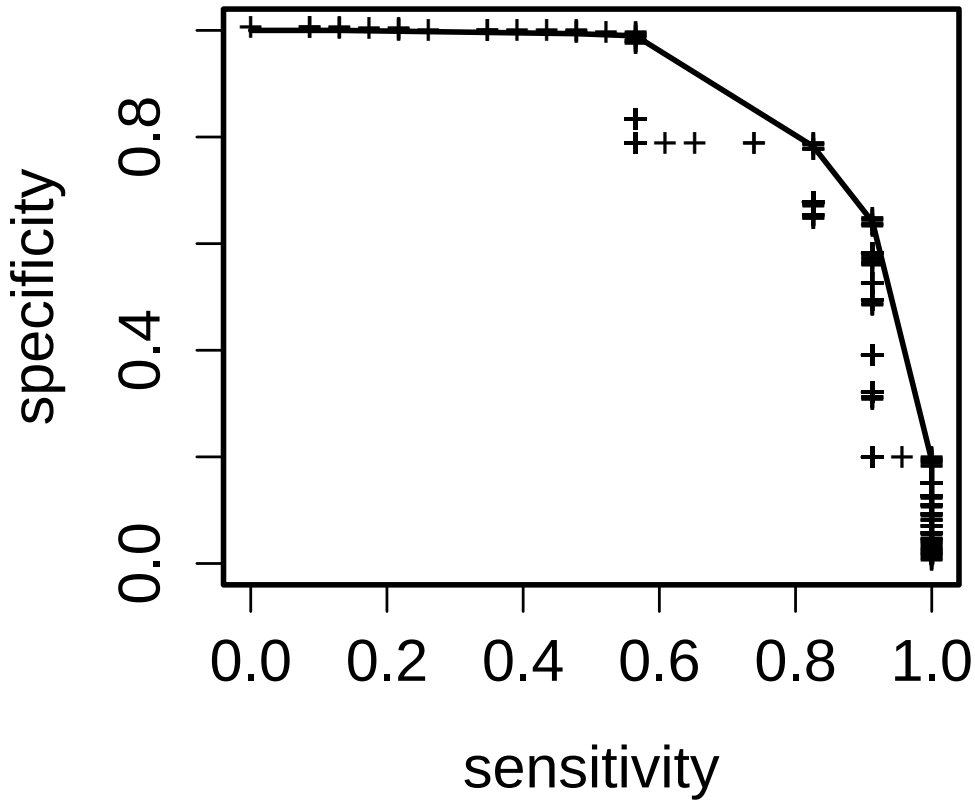
$\lambda_0(c^2)$ between functional classes

Information gain for two genes depends on $|\log \lambda_1(c^2) - \log \lambda_0(c^2)|$

Edge is empty if correlation lies on intersection

Leave-one-out cross validation





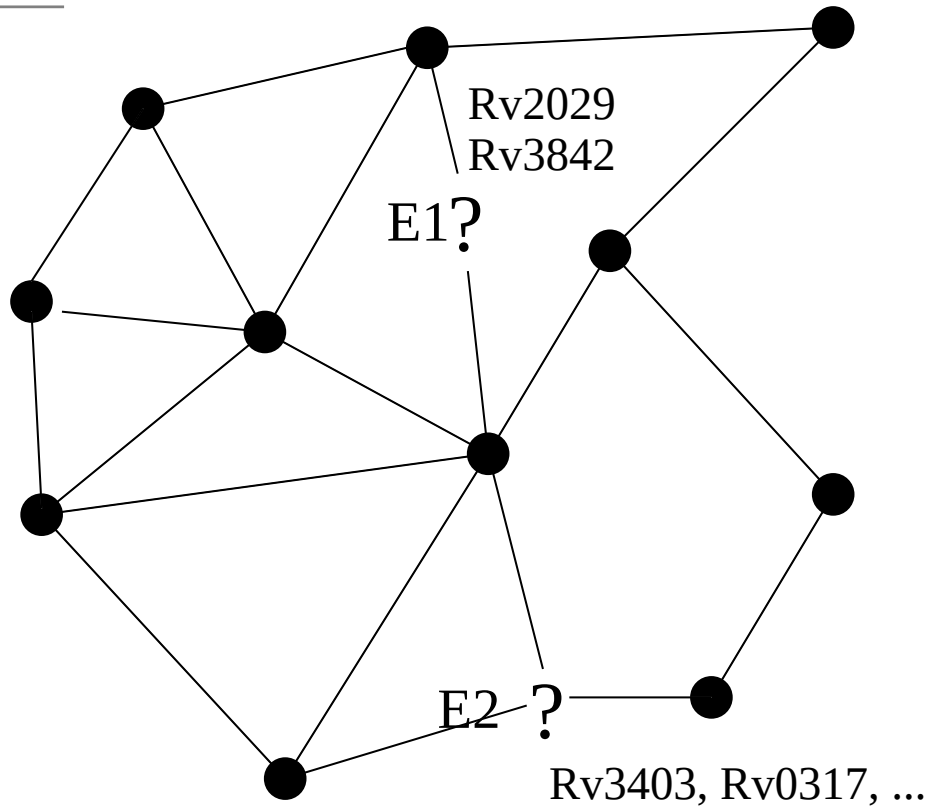
"Amino acid biosynthesis" MA net vs all nets

Area under ROC curve improves from 0.75 to 0.92

Combined area ROC

Class	MA	Compl	Path	Directon	Combined
1	**0.75**	0.51	**0.79**	0.661	0.92
2	**0.82**	0.51	0.53	0.56	0.82
3	0.60	0.51	0.62	0.53	0.71
4	0.67	0.57	0.59	0.68	0.84
5	0.61	0.51	0.61	0.56	0.72
6	0.62	0.52	0.62	0.61	0.76
7	0.60	0.51	**0.79**	0.66	0.84
8	0.67	0.53	**0.71**	0.52	0.78
9	0.57	0.52	0.61	0.59	0.72
10	0.68	0.53	0.57	0.65	0.78
11	0.65	0.55	0.59	**0.71**	0.87
12	**0.75**	0.51	0.58	0.59	0.85
13	0.59	0.52	0.58	0.62	0.72

Assignment of enzyme function



Metabolic network: assignment of some genes uncertain

Evaluate probability of expression pattern conditioned on graph structure

Inversion of usual graph search to explain data!

$$P(\mathsf{E1} = \mathsf{Rv2029}, \mathsf{E2} = \mathsf{Rv3403} \mid G)$$
$$P(\mathsf{E1} = \mathsf{Rv3842}, \mathsf{E2} = \mathsf{Rv3403} \mid G)$$
$$\ldots$$

Gaussian graphical models

- Allows continuous values instead of discrete labels
- Characterised by precision matrix $K = (k_{ij}) = \Sigma^{-1}$ (inverse of covariance matrix) with k_{ij} constrained to 0 for each missing edge in the graph
- Sampling from posterior of K is tricky but possible (Atay-Kayis and Massam, 05)
- Maximum likelihood estimate (Lauritzen, 96) gives excellent results on simulated networks
- Bayesian evaluation based on Gibbs sampling gives rather disappointing results