Hierarchical Bayesian Modelling Identifies Shared Gene Function

Peter Sykacek[1 2] & Gos Micklem[1 3]

Departments of Genetics[1], Pathology[2] & CCBI[3]
University of Cambridge
ps408@cam.ac.uk

Data & Biology:	Richard Clarkson, Cris Print
Methodological Discussions:	David J. C. MacKay & Inference Group

jump 2 TOC

Problem Statement

- Assumption: Several microarray experiments are obtained such that slides can be mapped to a biological state of interest.
- Shared gene function: Genes are across experiments informative about that biological states.
- Task: find those genes! Actually two problems:
  - Cross annotation of genes (potentially different species)
  - Calculate a measure across experiments

This talk shows how we may obtain such a measure using a probabilistic approach.

Biological States of Experiments

Many active processes in a Mammary Gland tc. (lact. day & hrs involution)

biol. state	L_0	L_5	L_{10}	I_{12}	I_{24}	I_{48}	I_{72}	I_{96}
Type 1 Apoptosis	-	-	-	+	+	?	-	-
Type 2 Apoptosis	-	-	-	-	-	?	+	+
Apoptosis	-	-	-	+	+	+	+	+
Differentiation	+	+	+	?	-	-	-	-
Inflammation	?	-	-	+	+	?	-	-
Remodelling	-(?)	-	-	-	-	?	+	+

Biological States of Experiments

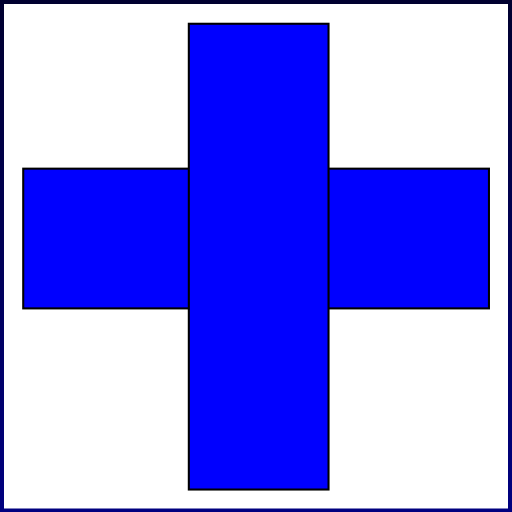
Many active processes in a Mammary Gland tc. (lact. day & hrs involution)

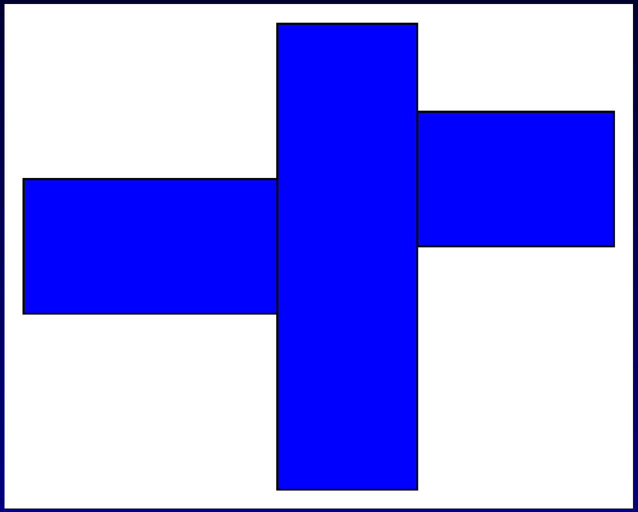
biol. state	L_0	L_5	L_{10}	I_{12}	I_{24}	I_{48}	I_{72}	I_{96}
Type 1 Apoptosis	-	-	-	+	+	?	-	-
Type 2 Apoptosis	-	-	-	-	-	?	+	+
Apoptosis	-	-	-	+	+	+	+	+
Differentiation	+	+	+	?	-	-	-	-
Inflammation	?	-	-	+	+	?	-	-
Remodelling	-(?)	-	-	-	-	?	+	+

Serum deprived Ednothelial cells provide a focus on Apoptosis and Differentiation (hours)

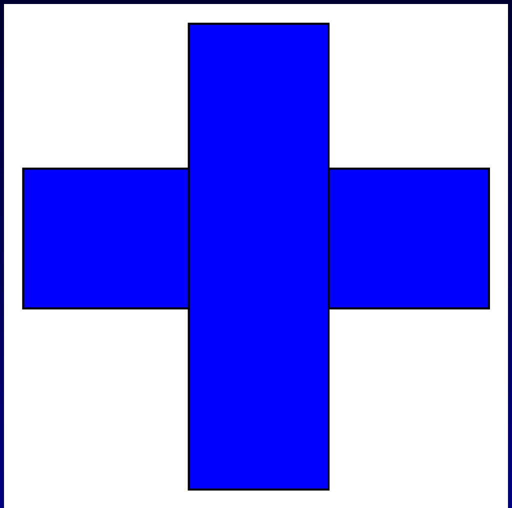
biol. state	t_0	t_{28}	t_{48}
Type 2 Apoptosis	-	+	+
Apoptosis	-	+	+
Differentiation	+	-	-

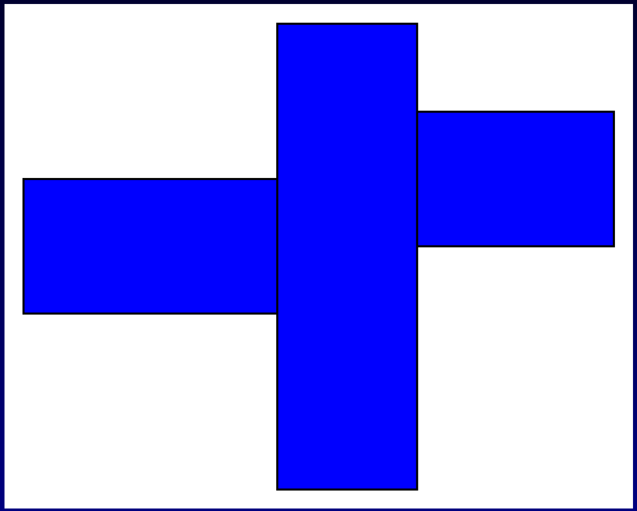
Guess the Correct “Model”

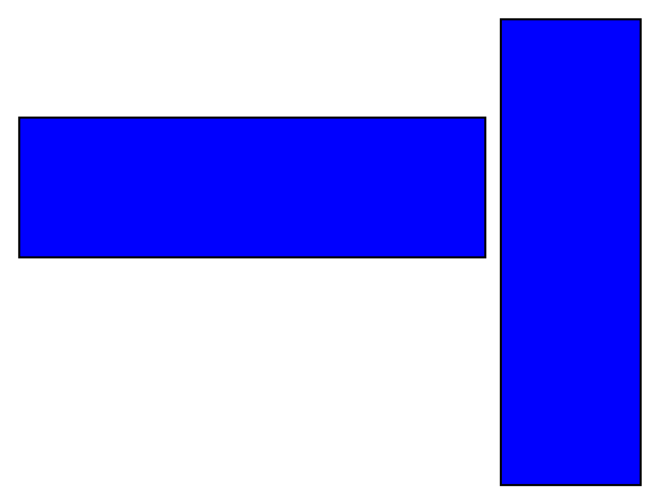


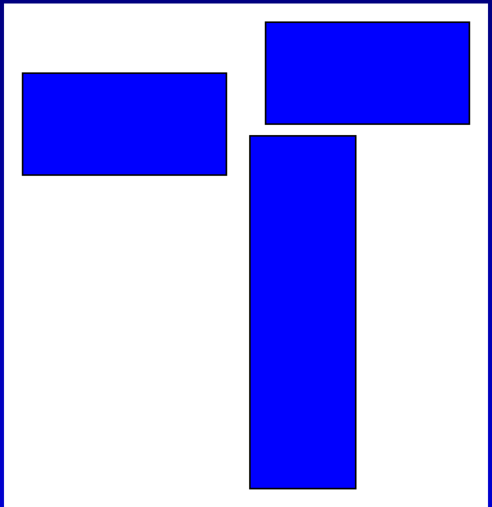


Guess the Correct “Model”



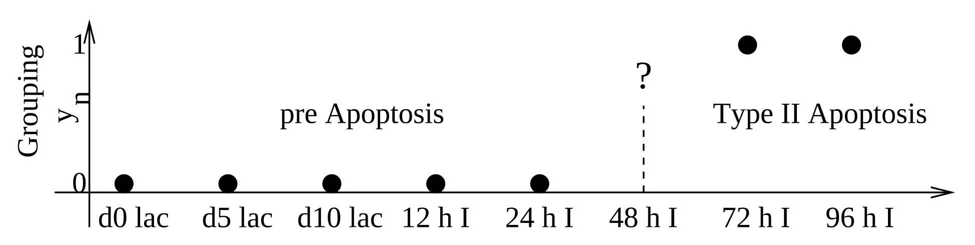




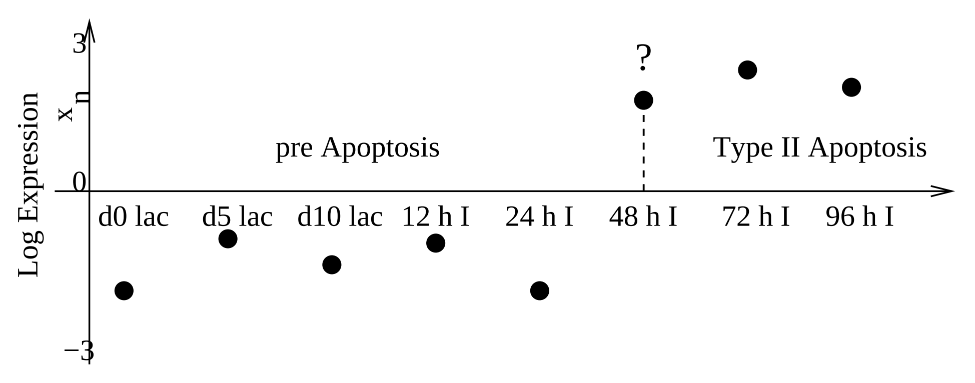


Now Guess Correct Label

Labels

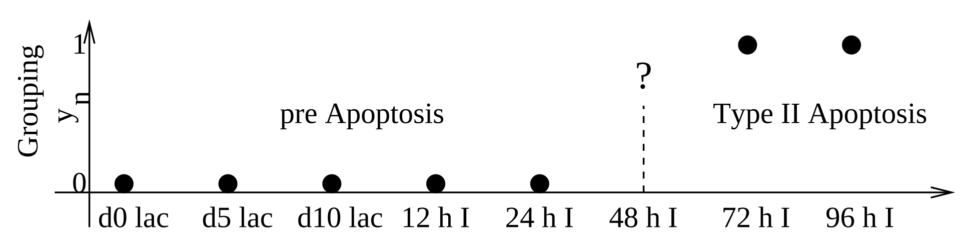


Expression Values Gene A

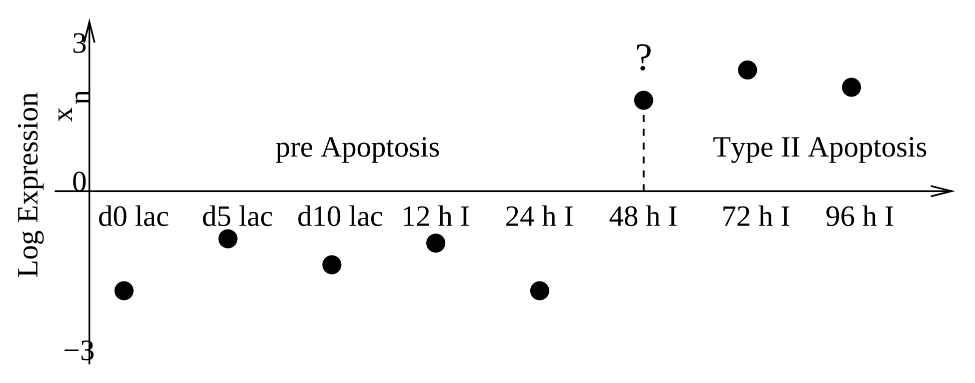


Now Guess Correct Label

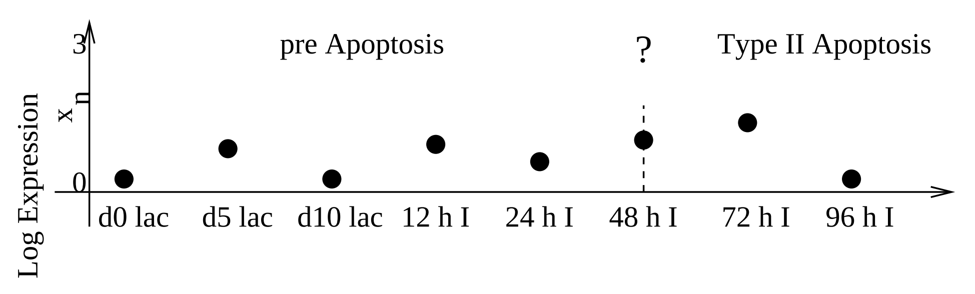
Labels



Expression Values Gene A

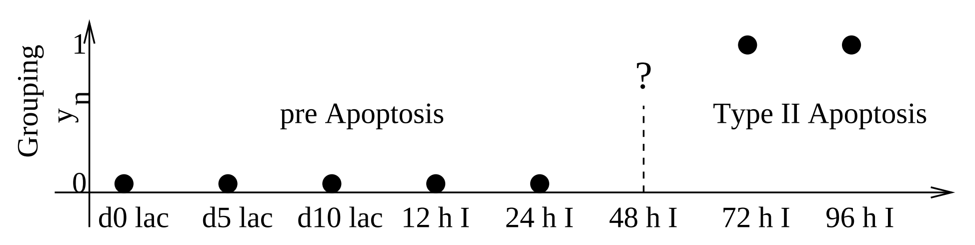


Expression Values Gene B

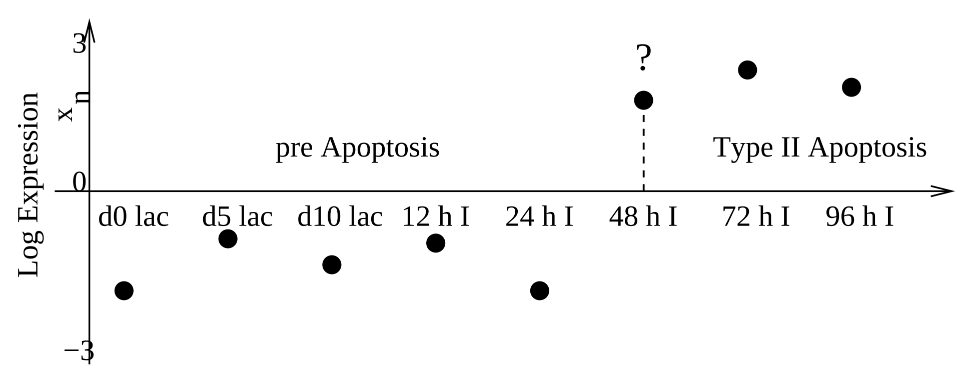


Now Guess Correct Label

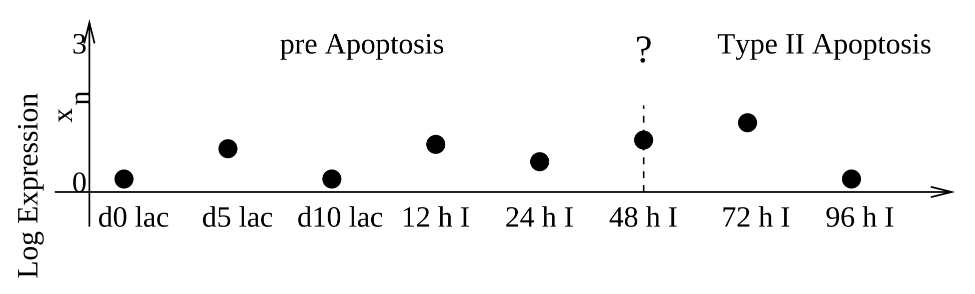
Labels



Expression Values Gene A

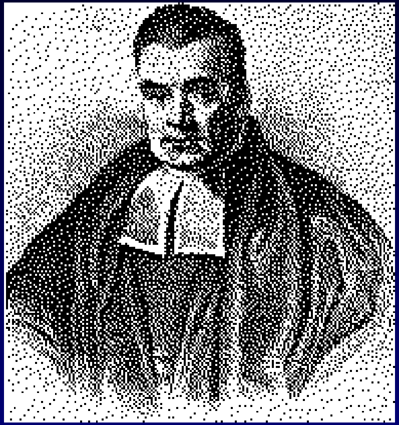


Expression Values Gene B



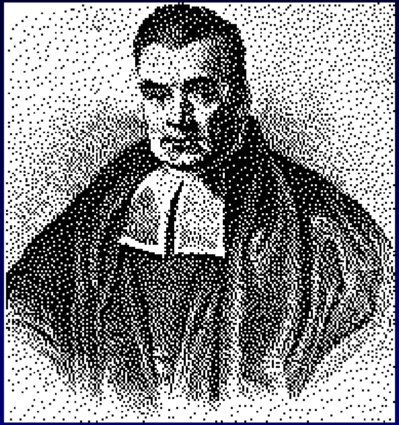
Guessing gene function is like guessing the model. We assess gene function only if predictions from expression values are sufficiently better than the default.

Probabilistic Approach



Thomas Bayes (1701 - 1763)
Learning from data based on a
decision theoretic framework

Probabilistic Approach

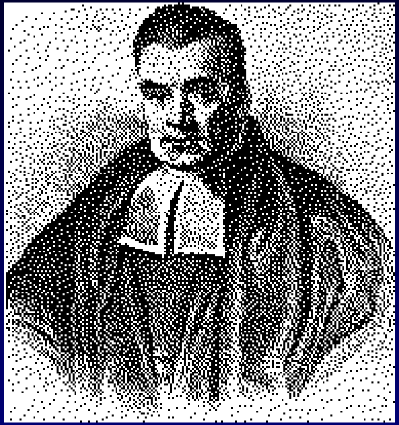


Thomas Bayes (1701 - 1763)
Learning from data based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

First consequence: we must revise beliefs according to Bayes theorem

Probabilistic Approach



Thomas Bayes (1701 - 1763)
Learning from data based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

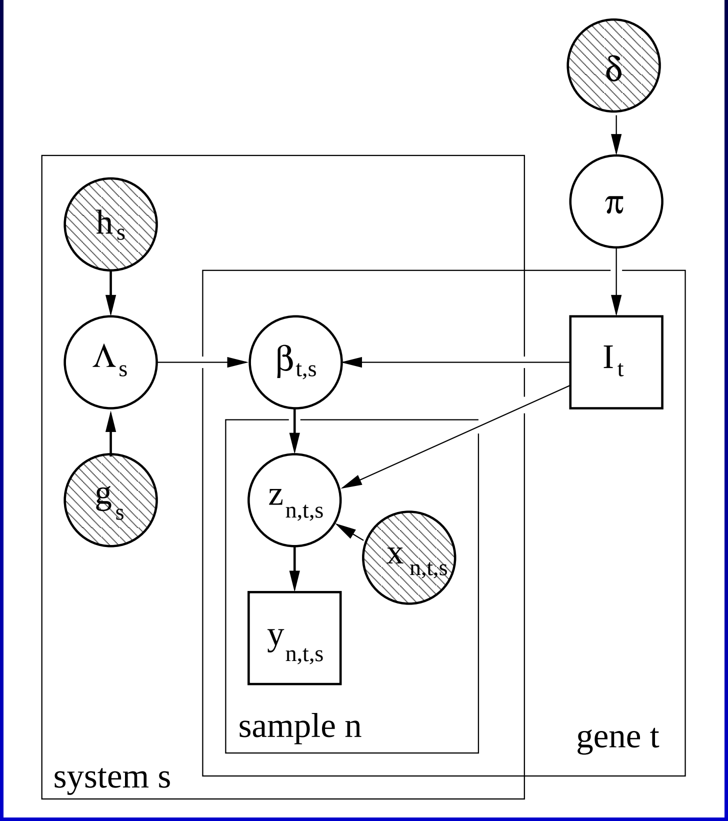
First consequence: we must revise beliefs according to Bayes theorem

$\alpha_{opt} = \text{argmax}_{\alpha} < u(\alpha) >$, where $< u(\alpha) >= \int_I u(\alpha, I)p(I|\mathcal{D})dI.$

Second consequence: Decisions by maximising expected utilities

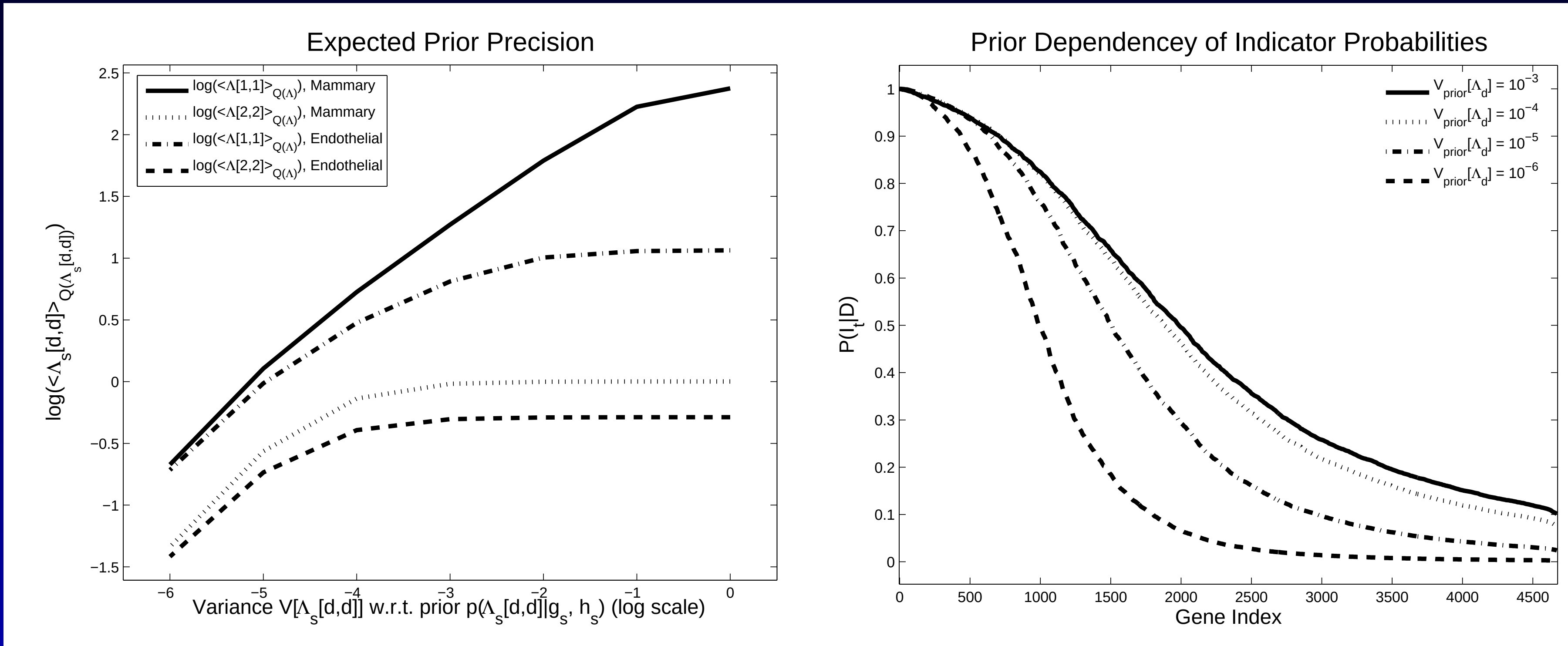
Hierarchical Models

moderate unnecessary side effects of prior choices and provide “data driven” results.



- all genes contribute to inference of $\mathbf{\Lambda}_s$
- hierarchical priors for sensitivity analysis
- $Q(I_t)$ approximates gene measure
- using *one* model gets all marginals right

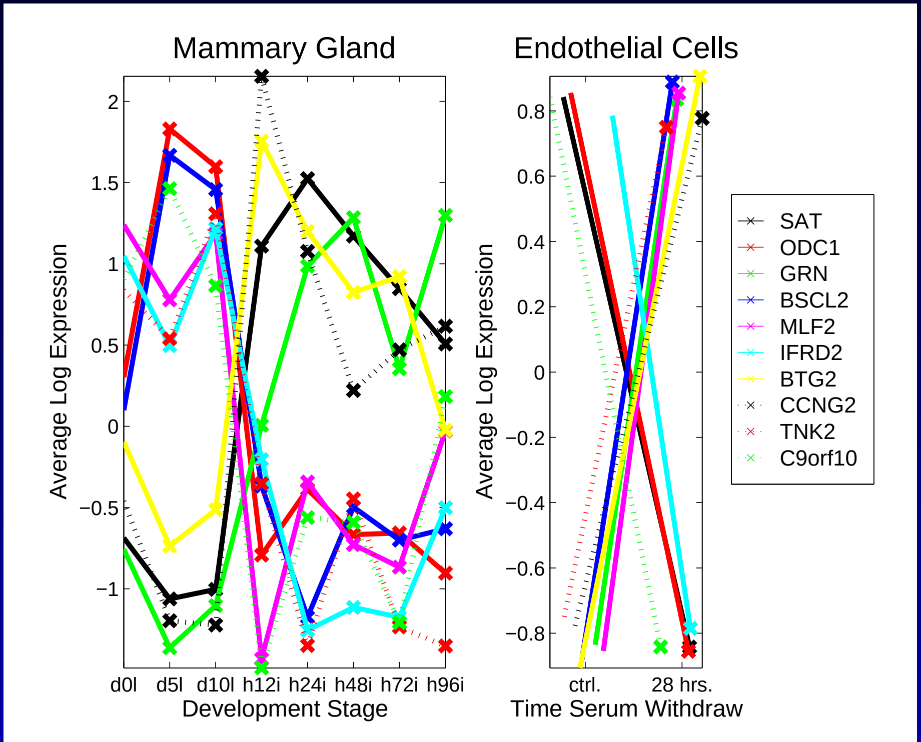
Sensitivity Check



For the hyper parameters this suggests $g \leq 0.01$ and $h \leq 1$.

We also conclude that equal cost results in many potential candiate genes.

Top Ten



Top 10 $P(I_t = 1|\mathcal{D}_1, \mathcal{D}_2)$ for Mammary lactation vs. involution *and* Endothelial cell death (result updated 01 2007).

Gene Symbol	$P(I_t\|\mathcal{D})$
SAT	0.99951
ODC1	0.99921
GRN	0.99921
BSCL2	0.99919
MLF2	0.99884
IFRD2	0.99867
BTG2	0.99843
CCNG2	0.99826
TNK2	0.99789
C9orf10	0.99783

Summary

- A realtively straight forward approach provides a principled measure of shared gene function.
- Beware of non hierarchical models - arbitrary gene measures can be adjusted for using the “right” prior.
- Don’t be afraid: a probabilistic (or Bayesian) approach is just common sense expressed by mathematical equations.

Table of Contents

- Problem Statement
- Biological States of Experiments
- Probabilistic Concepts
- Hierarchical Models
- Combined Results
- Summary

Variational Bayes

Mean field ansatz plus Jensens inequality. For all pdfs $Q(\theta)$:

$$\begin{aligned}
&\log\left(\int_\theta p(D|\theta)p(\theta)d\theta\right) \geq \\
&\qquad \int_\theta \Big(\log(p(D|\theta)) + \log(p(\theta)) - \log(Q(\theta))\Big)Q(\theta)d\theta \\
&\qquad = \log(p(D)) + \int_\theta \Big(\log(p(\theta|D)) - \log(Q(\theta))\Big)Q(\theta)d\theta
\end{aligned}$$

the last integral is a negative Kullback Leibler divergence and thus smaller or equal zero.

+ easy to compute; - systematic error as only an approximation.

back

jump 2 TOC

Variational Bayes II

Joint Distribution implied by the previous DAG

$$p(I_t, \boldsymbol{\beta}_{1,t}, S_{1,t}, D_{1,t} | \boldsymbol{\Lambda}_1, \pi_t, \gamma, X_{1,t}) = P(I_t|\pi_t) p(\boldsymbol{\beta}_{1,t}|\boldsymbol{\Lambda}_1, I_t) \times \prod_n \Big(p(s_{1,t,n}|\boldsymbol{\beta}_{1,t}, \boldsymbol{x}_{1,t,n}, I_t, \gamma) P(y_{1,t,n}|s_{1,t,n}, I_t) \Big)$$

where $S_{1,t} = \{s_{1,t,1}, ..., s_{1,t,N}\}$ and $D_{1,t} = \{y_{1,t,1}, ..., y_{1,t,N}\}$.

- Approximate posterior by a mean field expansion $Q(\boldsymbol{\beta}_{1,t}|I_t) \prod_n Q(s_{1,t,n}|I_t)$.
- Write down negative free energy and maximize the functional iteratively w.r.t. all Q-distributions.
- The negative free energy $F_{\max}(Q)$ approximates the log marginal likelihood and thus $P(I_t|D_{1,t}, \boldsymbol{\Lambda}_1, \pi_t, \gamma, X_{1,t})$.

back