Inference of Key Transcriptional Regulators in Endothelial Cell Apoptosis using Bayesian State Space Models

David L. Wild

Systems Biology Centre, University of Warwick
Keck Graduate Institute, Claremont, CA

joint work with

Claudia Rangel, Irma Aguilar-Delfin

December 4, 2008

Outline

Challenge

- Response of HUVEC to serum withdrawal, triggering apoptosis
- Timecourse with only a few measurements
- Challenge is to identify candidate regulators

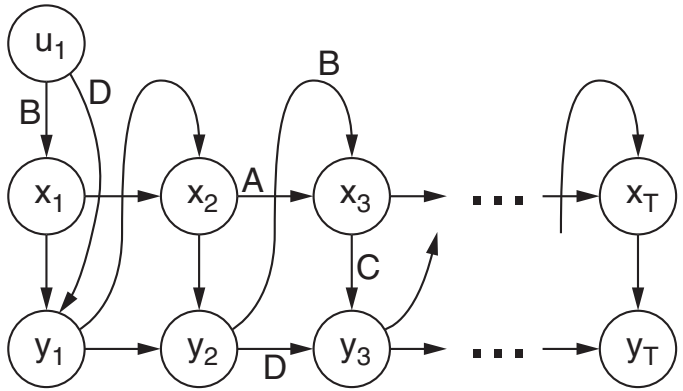
Challenge

- Response of HUVEC to serum withdrawal, triggering apoptosis
- Timecourse with only a few measurements
- Challenge is to identify candidate regulators

Challenge

- Response of HUVEC to serum withdrawal, triggering apoptosis
- Timecourse with only a few measurements
- Challenge is to identify candidate regulators

A Gaussian State-Space Model with Feedback



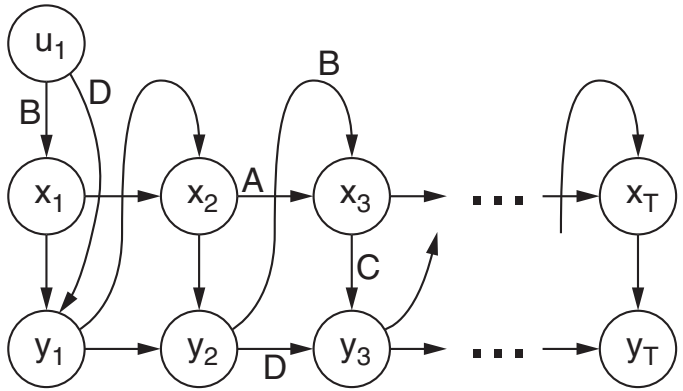
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



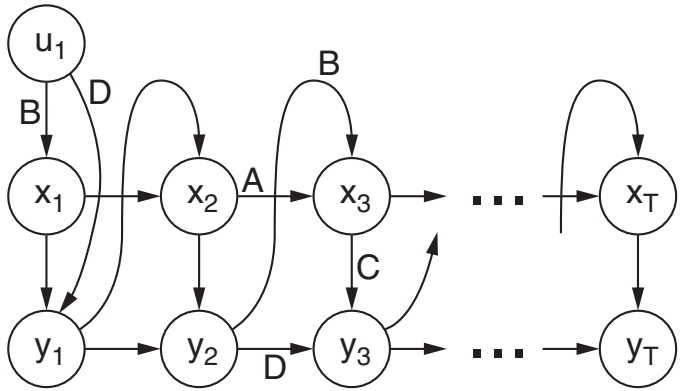
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



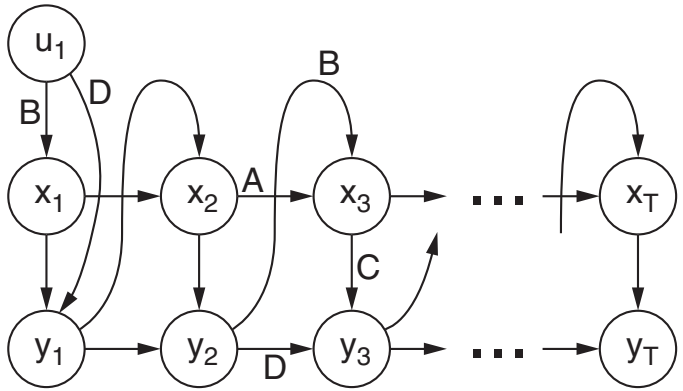
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 2005).

Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 2005).

Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 2005).

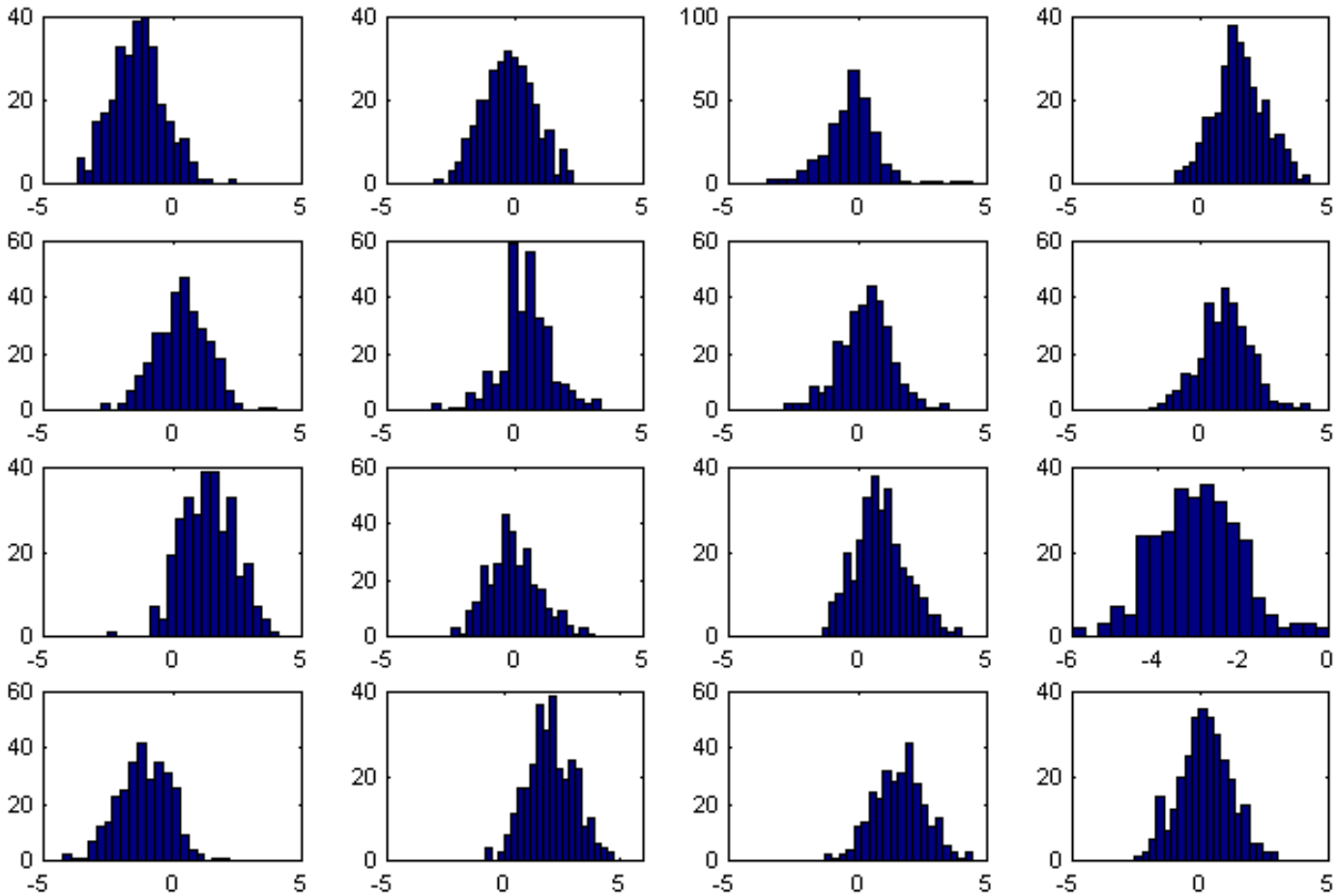
Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 2005).

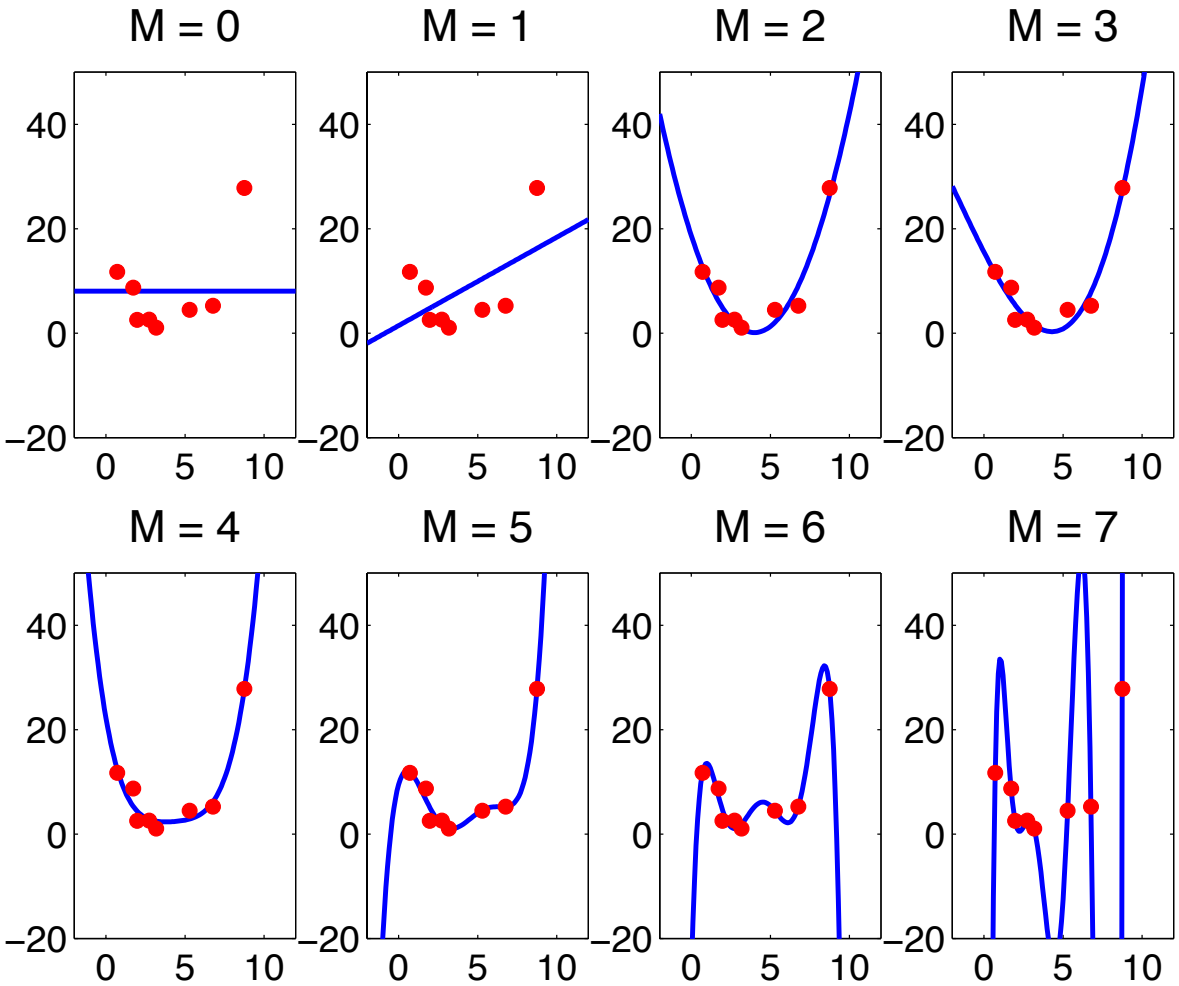
Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 2005).

Parameter Distributions



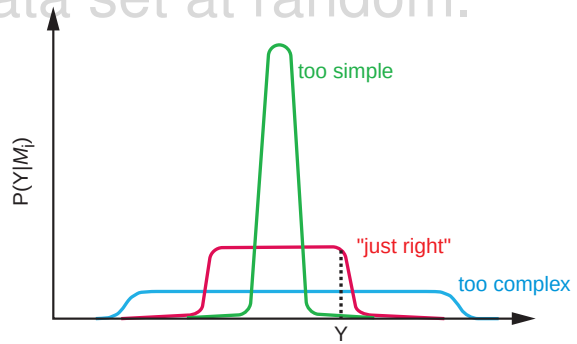
Model structure and overfitting: a simple example



Using Bayesian Occam's Razor to Learn Model Structure

Select the model class m_i with the highest probability given the data by computing the Marginal Likelihood ("evidence"):
Interpretation: The probability that *randomly selected* parameters from the prior would generate the data set.

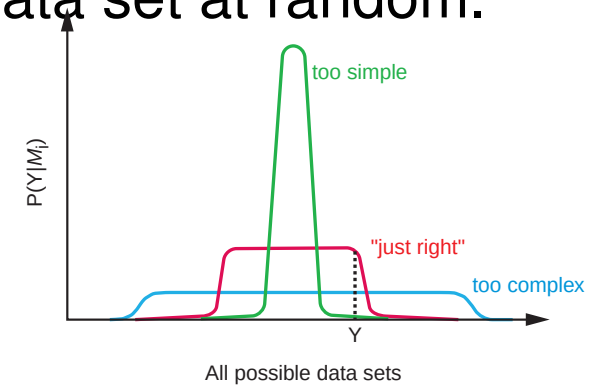
- Model classes that are too simple are unlikely to generate the data set.
- Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



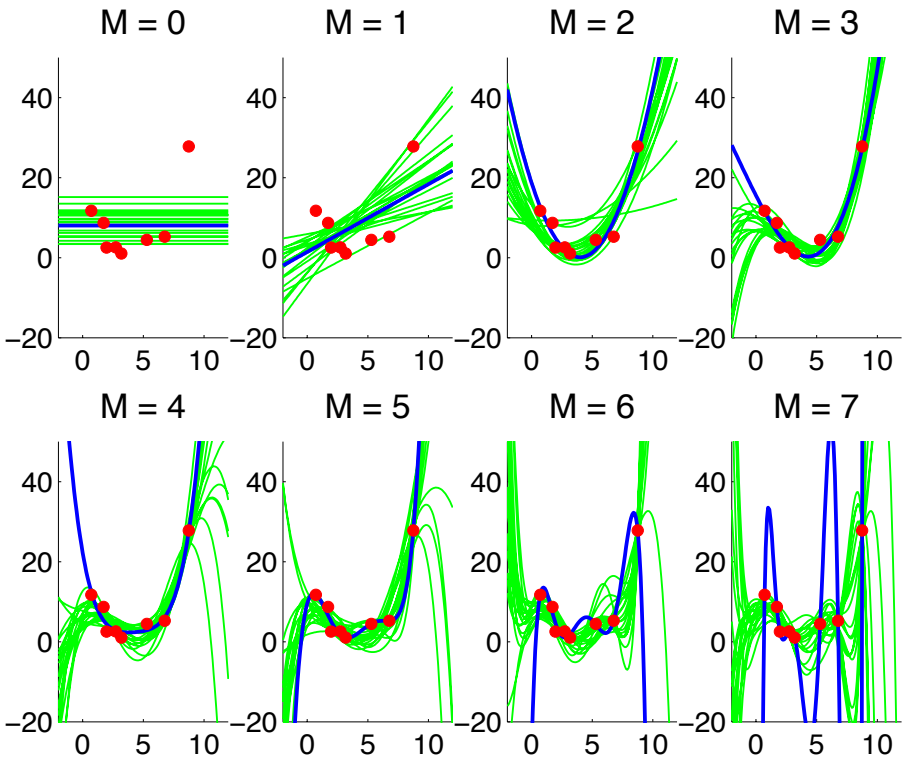
Using Bayesian Occam's Razor to Learn Model Structure

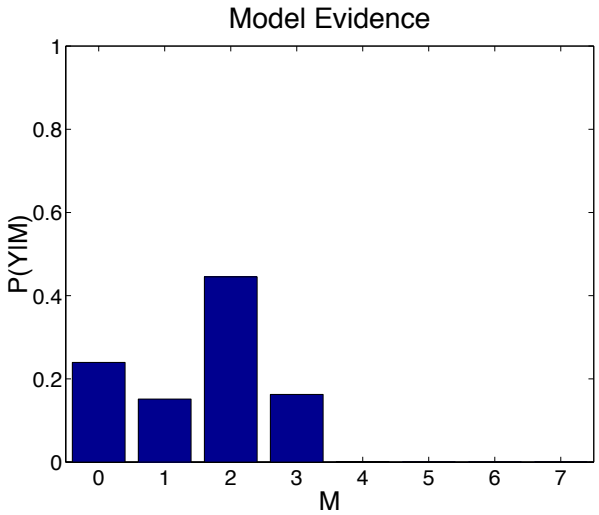
Select the model class m_i with the highest probability given the data by computing the Marginal Likelihood ("evidence"):
Interpretation: The probability that *randomly selected* parameters from the prior would generate the data set.

- Model classes that are too simple are unlikely to generate the data set.
- Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Bayesian Model Selection: Occam's Razor at Work

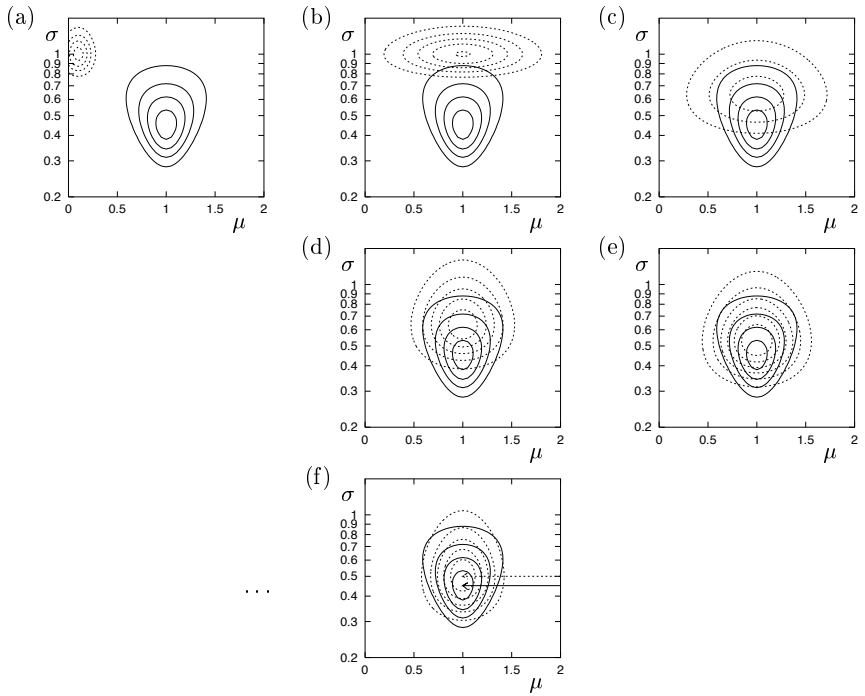




e.g. for quadratic (M=2): $y = a_0 + a_1 x + a_2 x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$ and $\theta_2 = [a_0\ a_1\ a_2\ \tau]$

Variational Bayesian Approach

Variational free energy minimization is a method of approximating a complex distribution $p(\mathbf{x})$ by a simpler distribution $q(\mathbf{x};\theta)$. We adust the parameters θ so as to get q to best approximate p in some sense.



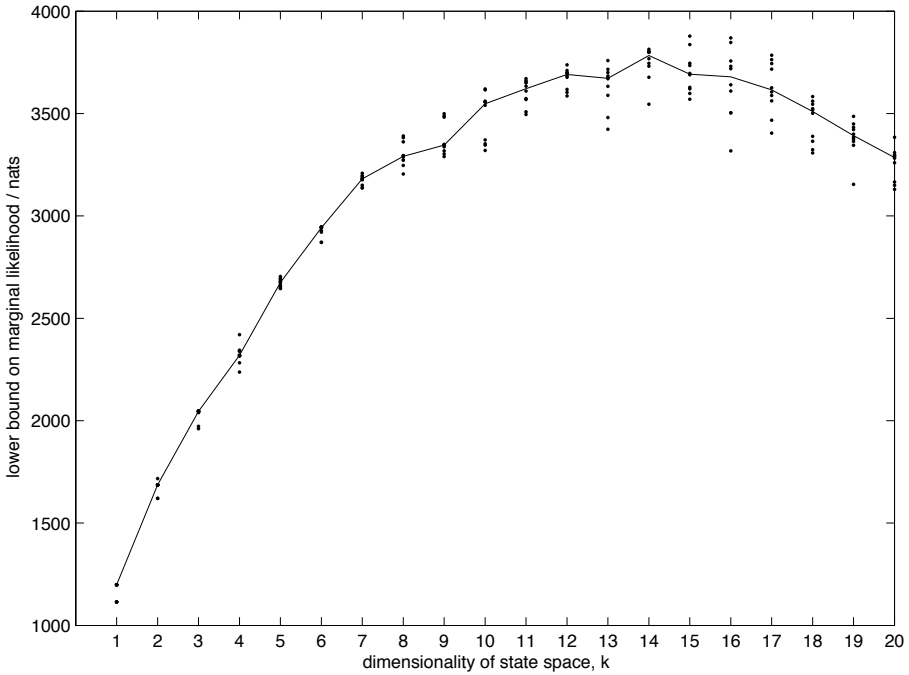
From David J.C. MacKay “Information Theory, Inference and Learning Algorithms”

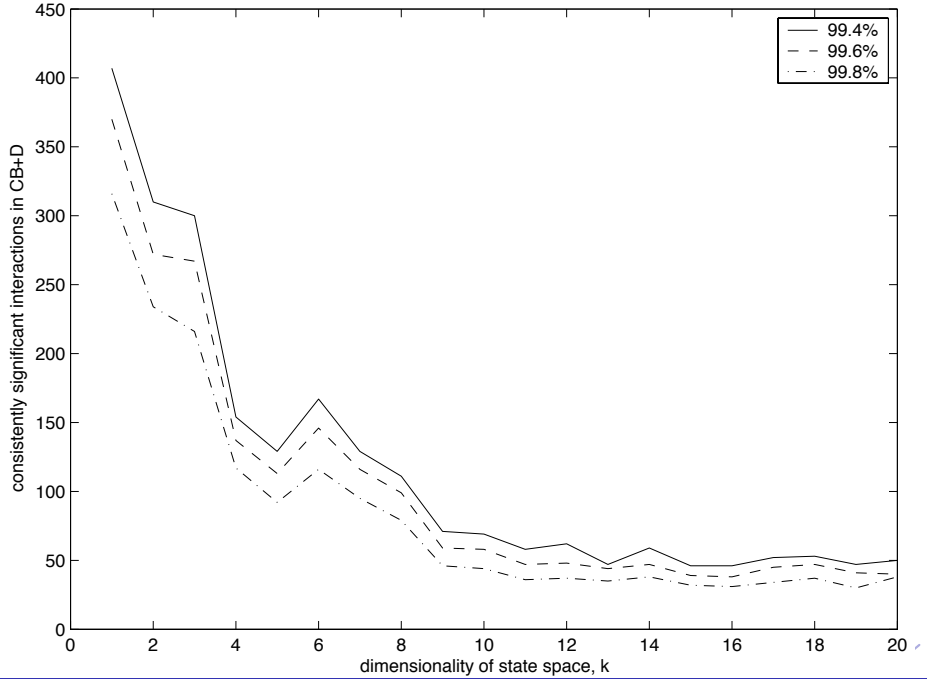
Lower Bounding the Marginal Likelihood

We can also lower bound the marginal likelihood:
Using a simpler, factorised approximation to
$q(\mathbf{x}, \theta) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\theta}(\theta)$:

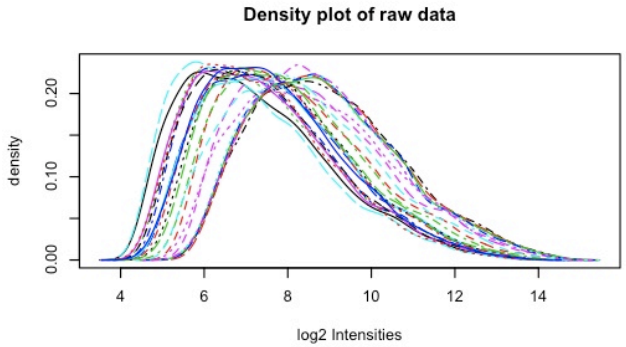
$$\ln p(\mathbf{y}|m) \quad = \quad \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\theta}(\theta), \mathbf{y}).$$

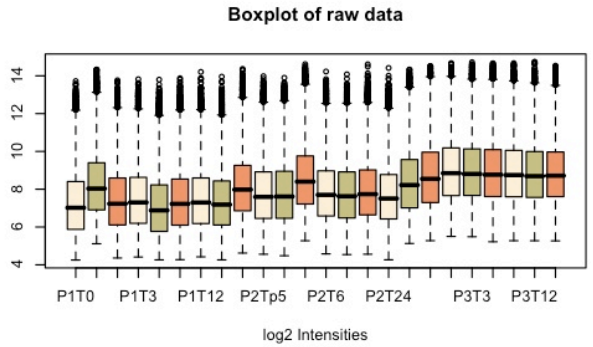
Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** iterative updates. $-\mathcal{F}_m$ is a variational free energy

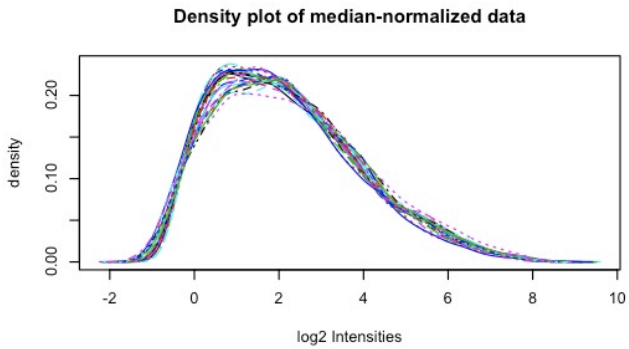


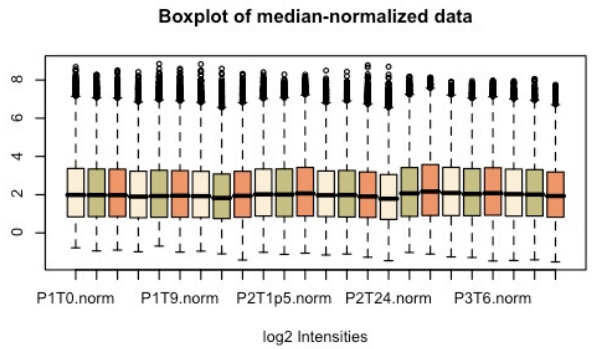


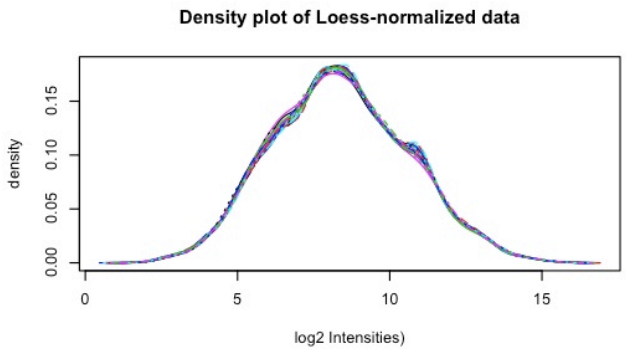
Data Normalization

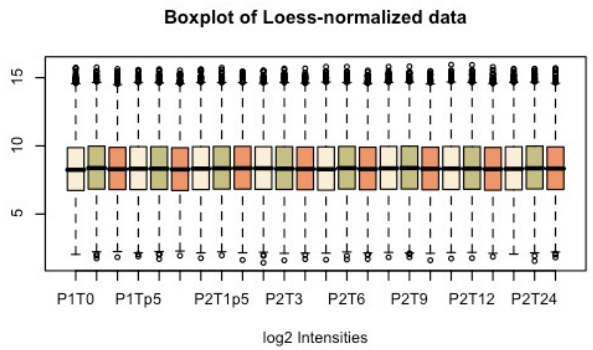




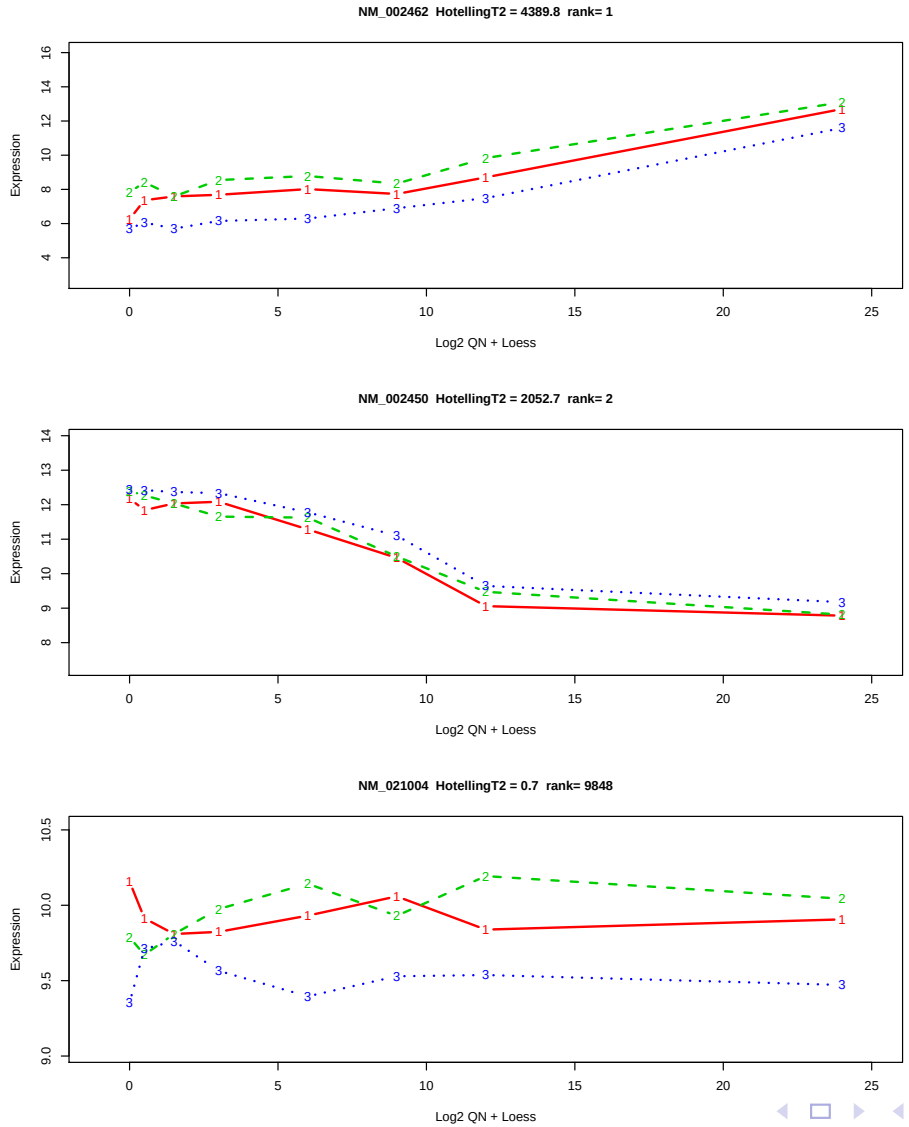




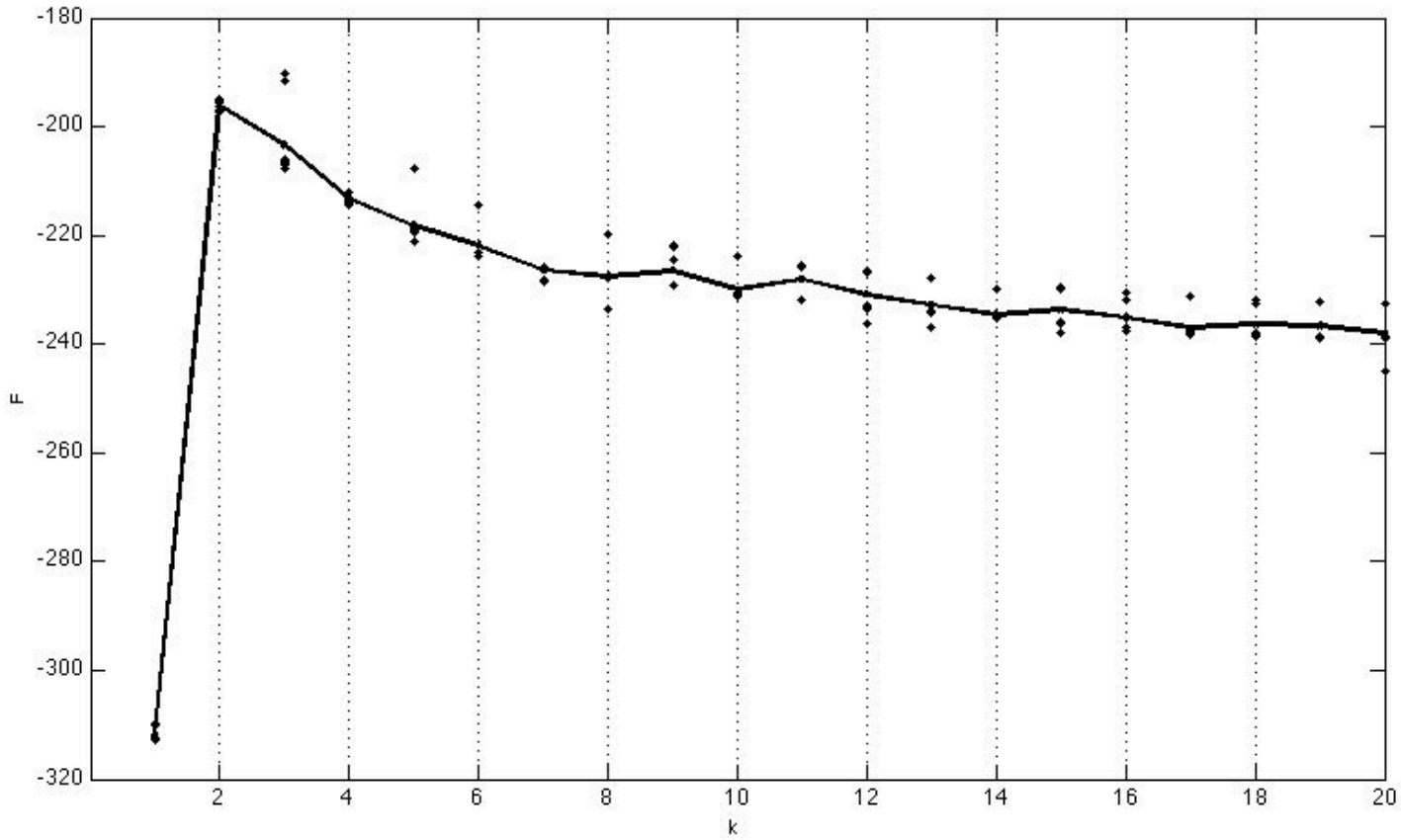




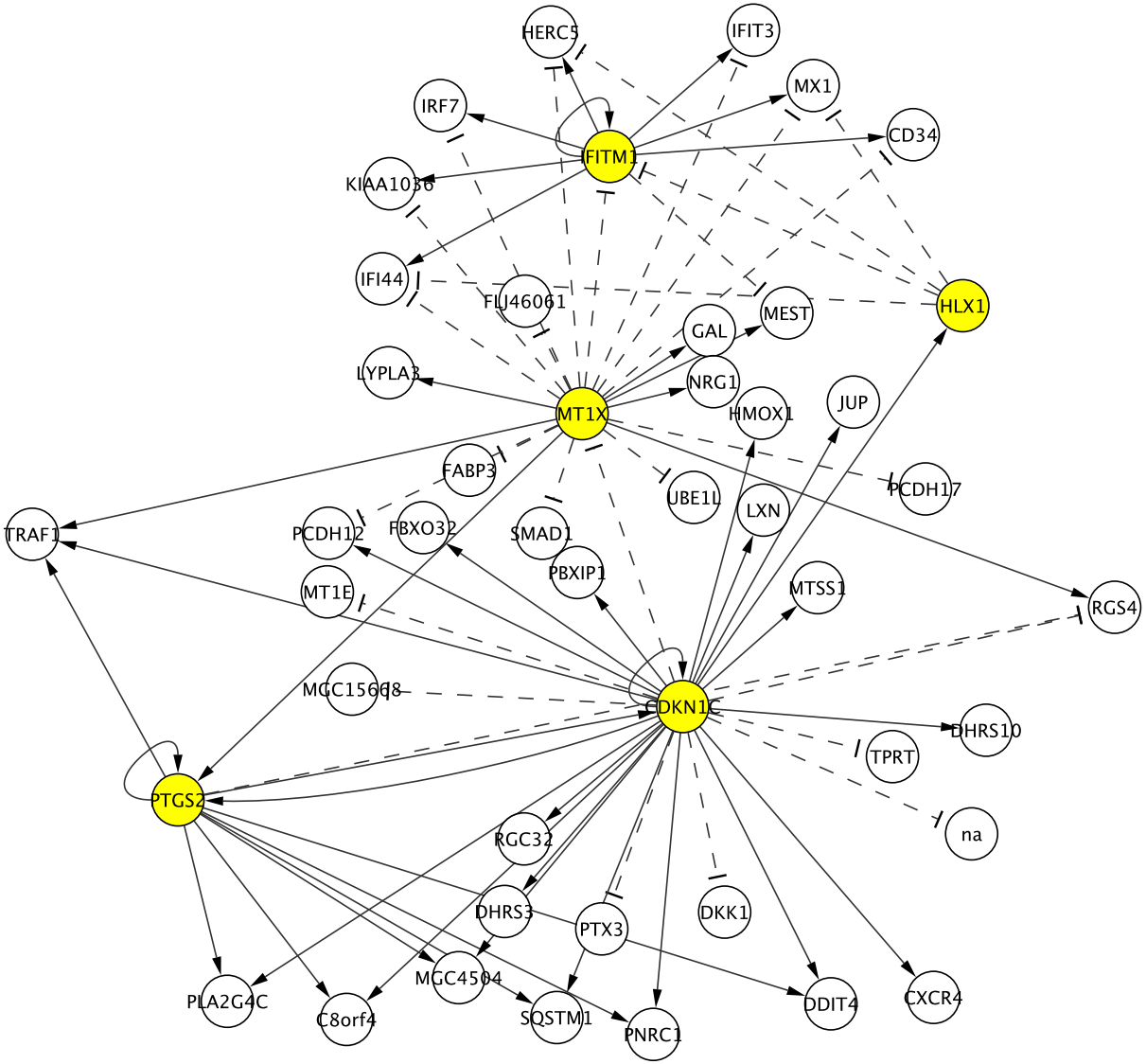
Gene Selection - Method of Tai and Speed



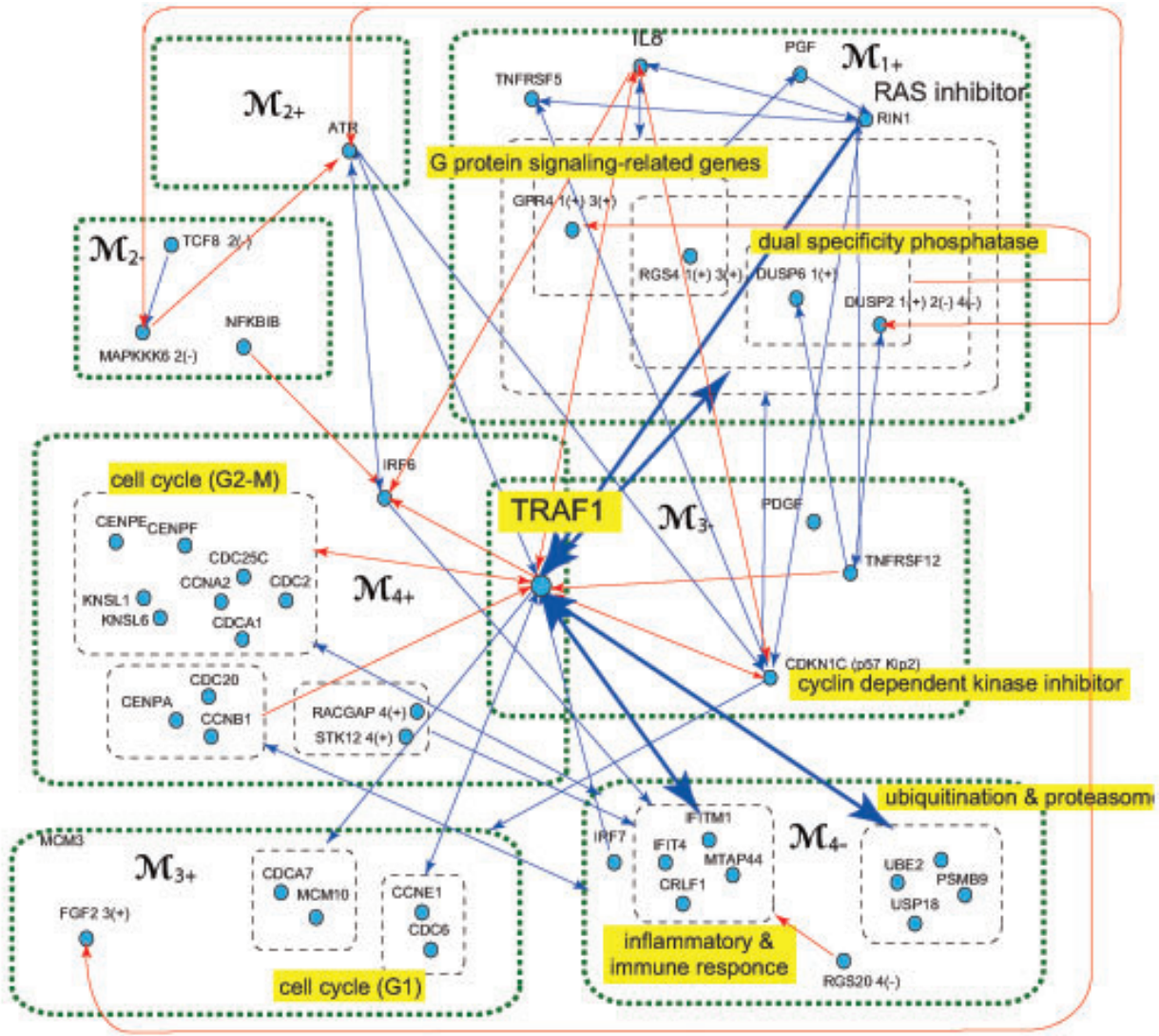
Model Selection



Inferred Network - Top 50 Ranked Genes



Inferred Network from Hirose et al. (2008)



Conclusions

- VBSSM model produces plausible biological hypotheses which can be experimentally validated
- *Candidate regulators* predicted as major hubs in inferred network
- Contradictory but *experimentally testable* hypothesis to Hirose et al. (2008)

Conclusions

- VBSSM model produces plausible biological hypotheses which can be experimentally validated
- *Candidate regulators* predicted as major hubs in inferred network
- Contradictory but *experimentally testable* hypothesis to Hirose et al. (2008)

Conclusions

- VBSSM model produces plausible biological hypotheses which can be experimentally validated
- *Candidate regulators* predicted as major hubs in inferred network
- Contradictory but *experimentally testable* hypothesis to Hirose et al. (2008)

Acknowledgements

- This work is supported by NSF Grant Number CCF-0524331 and and EU Marie-Curie IRG Fellowship (46444)