Gene Regulatory Network Inference: In Silico Hypotheses and Experimental Validation

David L. Wild

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

September 3, 2007
Outline

1. Motivation and Background
2. In Silicon Hypotheses and Experimental Validation
3. Ongoing Work and Conclusions
"Using gene expression data for reverse engineering of genetic networks is quite a hopeless exercise... Besides lacking power and being naive, the results from inference of these models can rarely be verified in any meaningful way".

An anonymous BBSRC grant referee
Temperature adaptation in *E. coli*

- Temperature is an **important factor** in the adaptation of *E. coli* to the environment.

- **Heat shock** response in Bacteria is characterized by the expression of chaperones and proteases which role is to cope with heat induced alterations of protein conformation. regulated by rpoH and rpo E

- No master regulators have been identified in **cold shock**. It is associated with a temporary inhibition of transcription and translation. Some of the proteins are protective.

- Shift between **10 C to 37 C** is likely to be an important factor in controlling adaptation to the host.
A Gaussian State-Space Model with Feedback

Output equation: \[ y_t = Cx_t + Dy_{t-1} + v_t \]
State dynamics equation: \[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

Key Concept: \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) 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:
\[ y_t = Cx_t + Dy_{t-1} + v_t \]

State dynamics equation:
\[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

**Key Concept:** \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) models the many unmeasured (hidden) factors such as

- genes that have not been 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:
\[ y_t = Cx_t + Dy_{t-1} + v_t \]

State dynamics equation:
\[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

**Key Concept:** \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) 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: \[ y_t = C x_t + D y_{t-1} + v_t \]

State dynamics equation: \[ x_t = A x_{t-1} + B y_{t-1} + w_t \]

Key Concept: \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) 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 $p(\theta|D)$, 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., 2004).
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 $p(\theta|D)$, 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., 2004).
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 $p(\theta|D)$, 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., 2004).
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 $p(\theta|D)$, 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., 2004).
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 $p(\theta|D)$, 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., 2004).
Variational Bayesian Approach

Variational free energy minimization is a method of approximating a complex distribution $p(x)$ by a simpler distribution $q(x; \theta)$. We adjust the parameters $\theta$ 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 $q(x, \theta) \approx q_x(x)q_\theta(\theta)$:

$$\ln p(y|m) = F_m(q_x(x), q_\theta(\theta), y).$$

Maximizing this lower bound, $F_m$, leads to EM-like iterative updates. $-F_m$ is a variational free energy.
Prior Specification

\[ \Sigma_0, \mu_0 \rightarrow X_{t-1} \rightarrow X_t \rightarrow u_t \rightarrow A \rightarrow \alpha \]
\[ \rightarrow B \rightarrow \beta \]
\[ \rightarrow C \rightarrow \gamma \]
\[ \rightarrow R \rightarrow a, b \]
\[ \rightarrow D \rightarrow \delta \]

i = 1 \ldots n

t = 1 \ldots T_{(i)}
Incorporating Prior Information - 8 replicates
The experimental set up

- A single channel microarray covering the entire transcriptional capacity of *E. coli* K12.
- A control experiment with cells grown at 37°C.
- A temperature shift experiment where cells were grown at 10°C and then shifted at 37°C.
- 13 time points with 6 replicates (a total of 78 measurements).
**Motivation and Background**

**In Silicon Hypotheses and Experimental Validation**

**Ongoing Work and Conclusions**

**Functional Classification of Genes - Temperature Shifted Data**

- **COLD SHOCK**
- **MALTOSE METABOLISM**
- **AEROBIC METABOLISM**
- **ANAEROBIC METABOLISM**
- **pH Homeostasis**
- **Carbon source transport**

**Gene Expression Data**

- **Gene Expression Values**
  - Protein turnover
  - Specific activity
  - Growth rate
  - Cellular response

**Pathway Analysis**

- **GO Terms**
  - p-value
  - Biological Process
  - Molecular Function
  - Cellular Component

**Significant Pathways**

- **cardiac muscle contraction**
- **cellular metabolic process**
- **cellular physiological process**
- **angiogenesis**
- **carbohydrate transport**
Gene Selection - Method of Tai and Speed

Gene Expression Graphs:

- **tdcB**
  - Hotelling $T^2 = 9533.9$  rank= 1
  - Expression at different time points:
    - 2.4.04, 5.21.04, 8.4.04, 7.17.04

- **ycbC**
  - Hotelling $T^2 = 34.2$  rank= 8000
  - Expression at different time points:
    - 2.4.04, 5.21.04, 8.4.04, 3.5.04, 7.17.04
A network representing the adaptation of *E. coli* to 37°C

![Network Diagram]

- 50 most differentially regulated genes
- Regulators
Verified interactions in Regulon DB
A role for hns in controlling Aerobic versus anaerobic switch?

Hydrogenase 2
(associated to the periplasmic side. Its expression is induced in anaerobic conditions)

Aerobic metabolism

Anaerobic metabolism

pH Homeostasis

Aerobic metabolism
Profiles of glpC and glpQ
Chip-Chip experiments with hns and RNA polymerase to determine co-location

Incorporating verified interactions and non-interactions as priors
Ongoing Work

- Chip-Chip experiments with hns and RNA polymerase to determine co-location
- Incorporating verified interactions and non-interactions as priors
Experimental evidence of a novel regulatory role for *hns* as **activator** in cold shock

Models produce plausible biological hypotheses which can be experimentally validated
Conclusions

- Experimental evidence of a novel regulatory role for \textit{hns} as \textbf{activator} in cold shock
- Models produce plausible biological hypotheses which can be experimentally validated
Acknowledgements

- Juan Li and Matthew Beal (SUNY Buffalo)
- Francesco Falciani (Birmingham) and Gianni Deho (Milan)
- This work is supported by NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)
Acknowledgements

- Juan Li and Matthew Beal (SUNY Buffalo)
- Francesco Falciani (Birmingham) and Gianni Deho (Milan)

This work is supported by NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)
Acknowledgements

- Juan Li and Matthew Beal (SUNY Buffalo)
- Francesco Falciani (Birmingham) and Gianni Deho (Milan)
- This work is supported by NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)