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

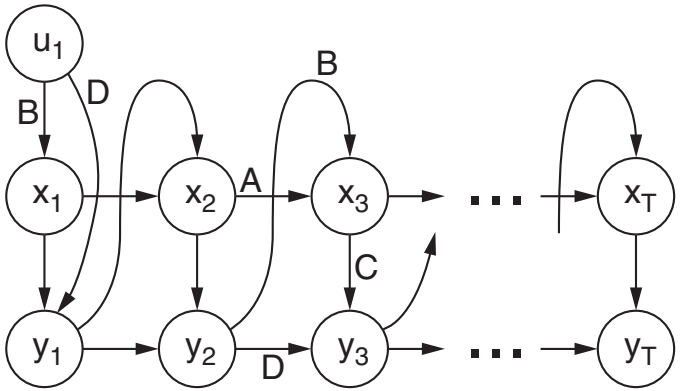
"Using gene expression data for reverse engineering of genetic networks is quite a hopeless exercise... Besides lacking power and being naive, the results from inference of these models can rarely be verified in any meaningful way".

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



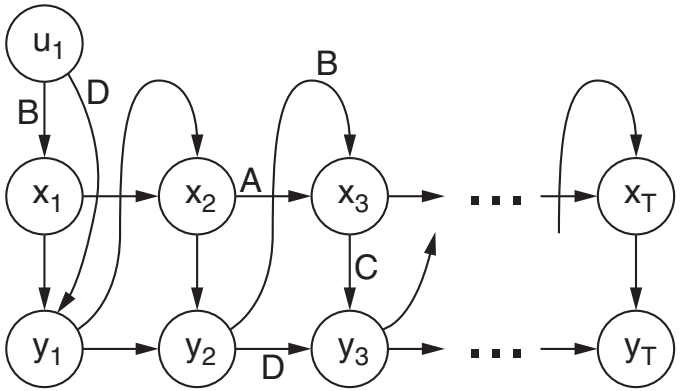
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

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



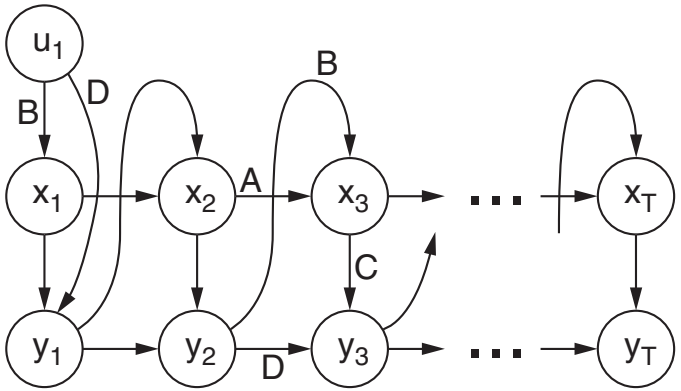
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

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



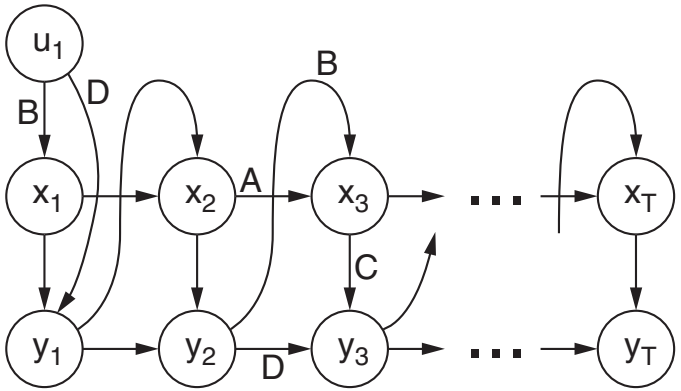
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

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., 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 $p(\theta|\mathcal{D})$, 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., 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 $p(\theta|\mathcal{D})$, 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., 2004).

Our Approach

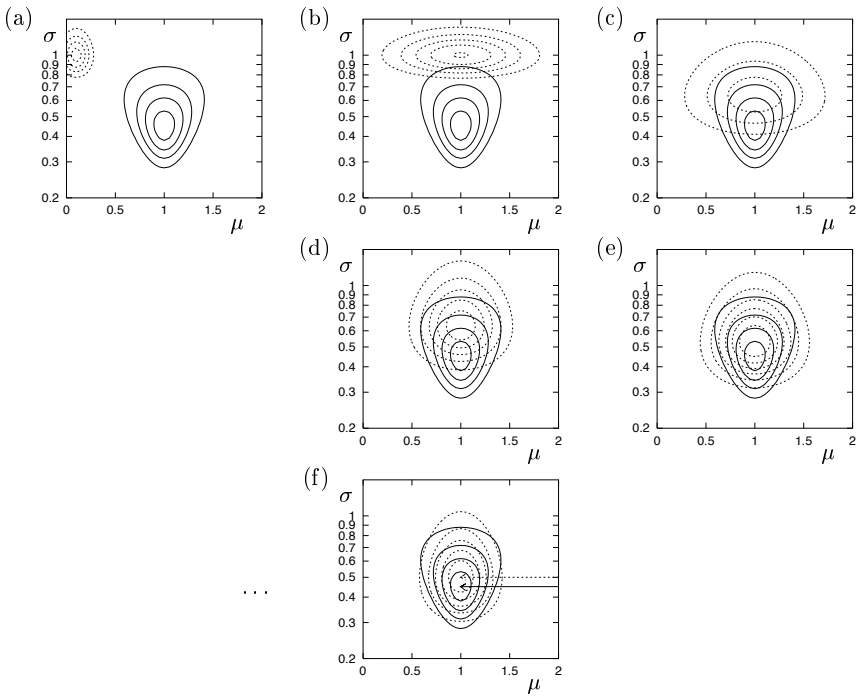
- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- We can use variational approximations to approximate Bayesian inference in state-space models (Beal et al., 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 $p(\theta|\mathcal{D})$, 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., 2004).

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.



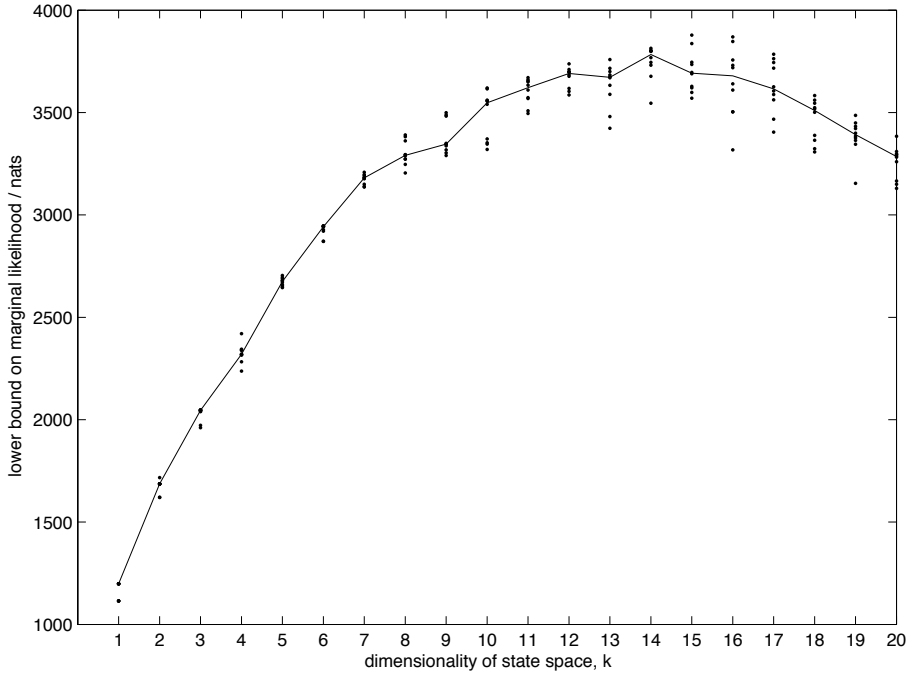
From David J.C. MacKay “Information Theory, Inference and Learning Algorithms”

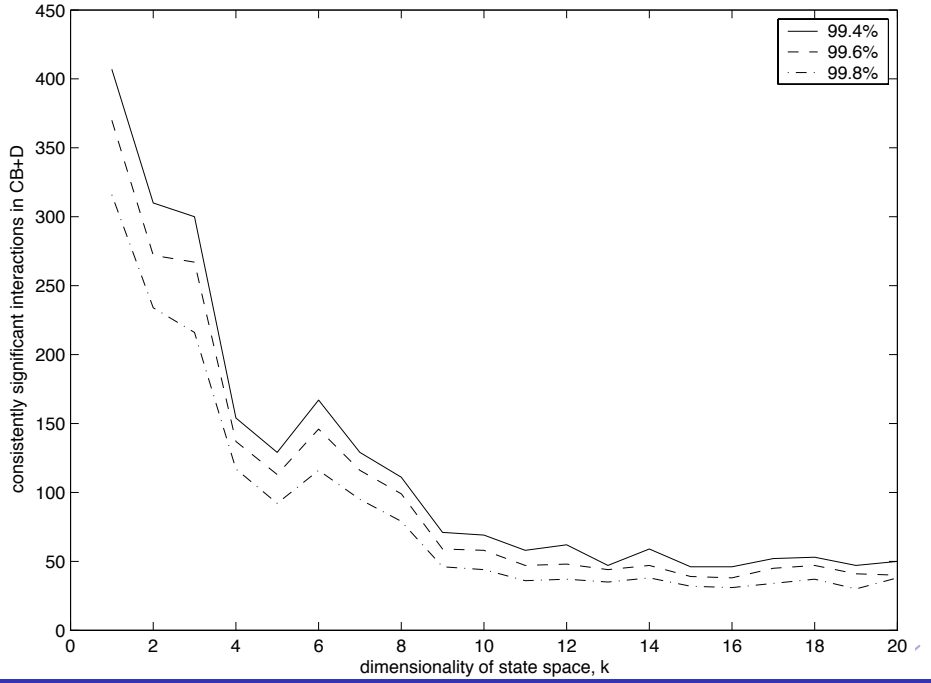
Lower Bounding the Marginal Likelihood

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

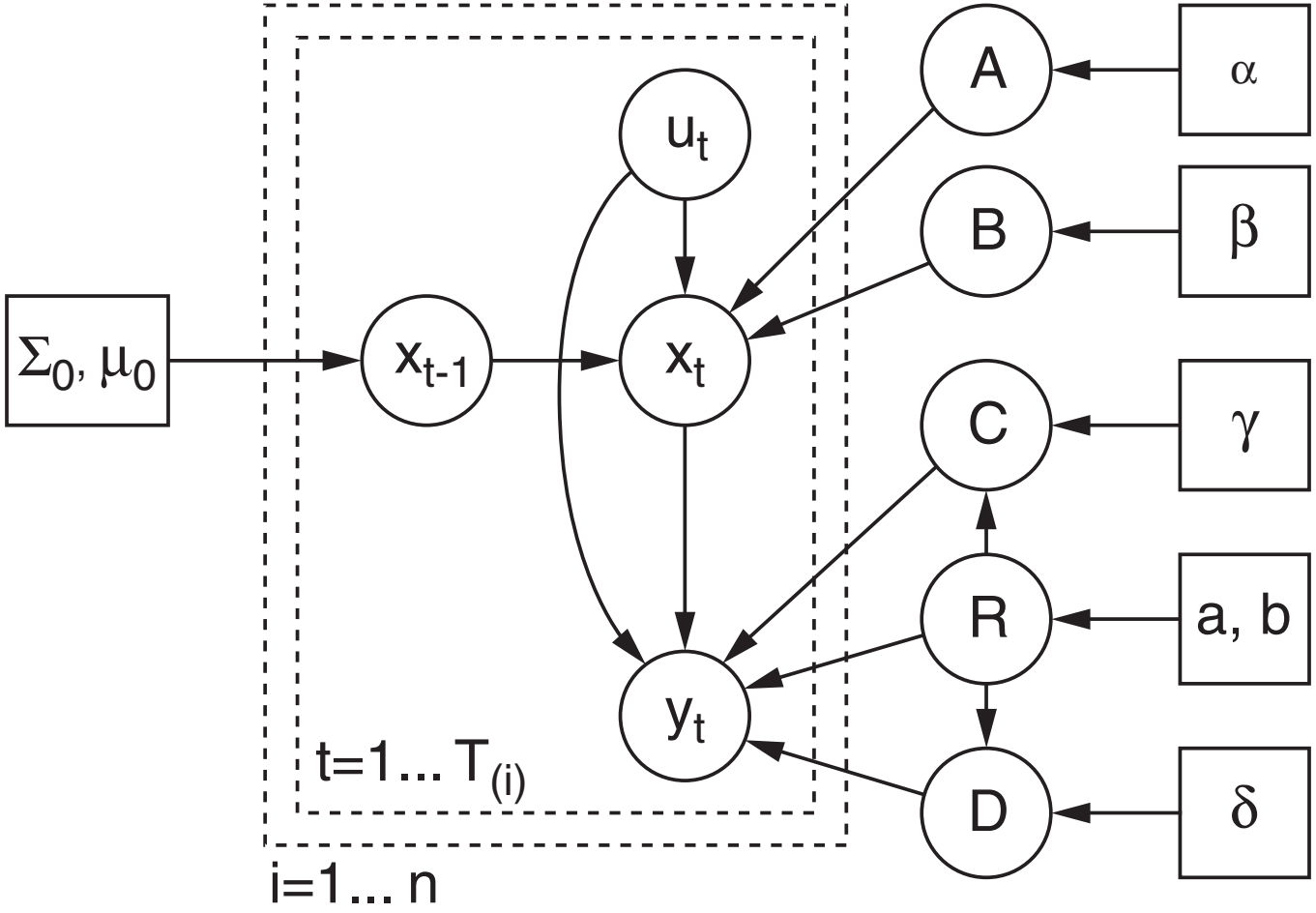
$$\ln p(\mathbf{y}|m) \quad = \quad \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\theta}(\theta), \mathbf{y}).$$

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

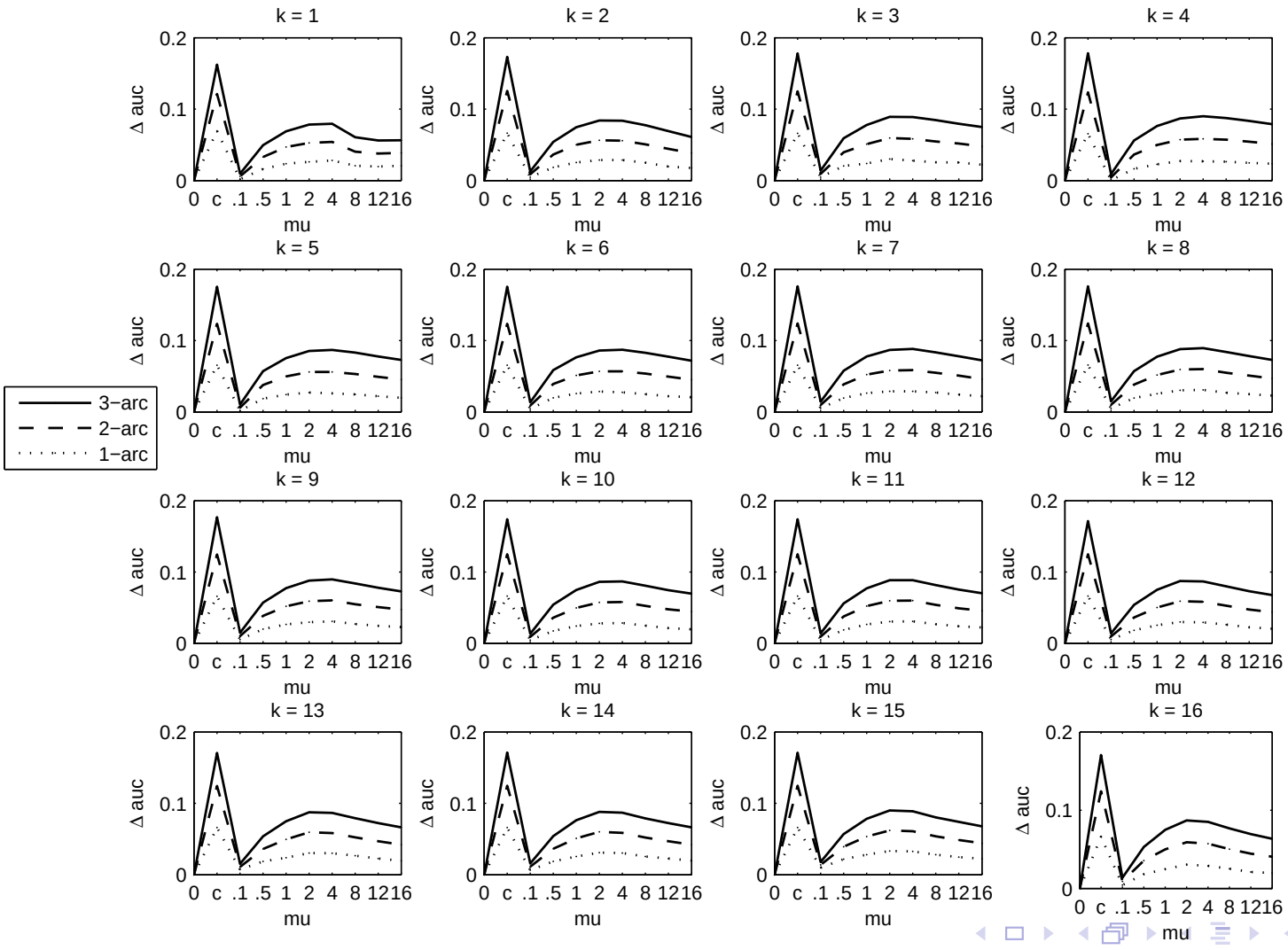




Prior Specification

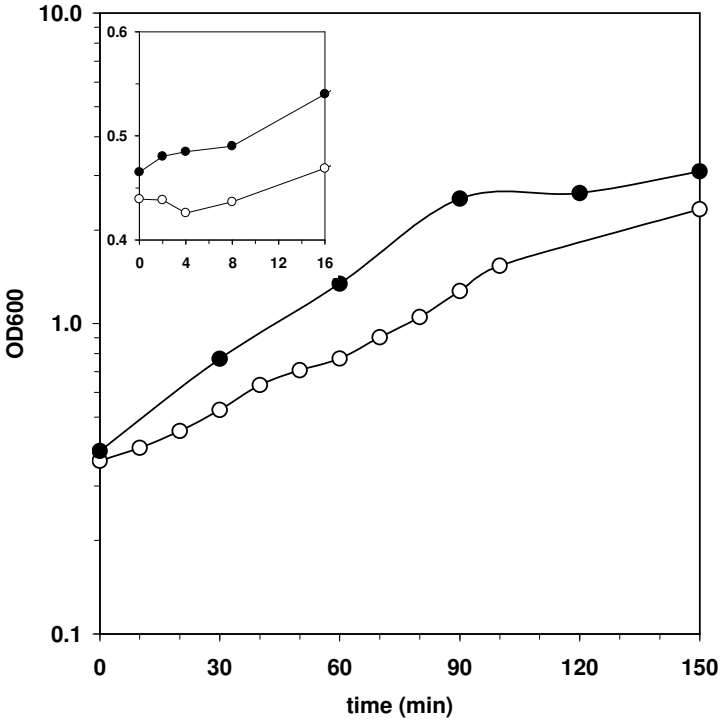


Incorporating Prior Information - 8 replicates



Experimental Setup

The experimental set up

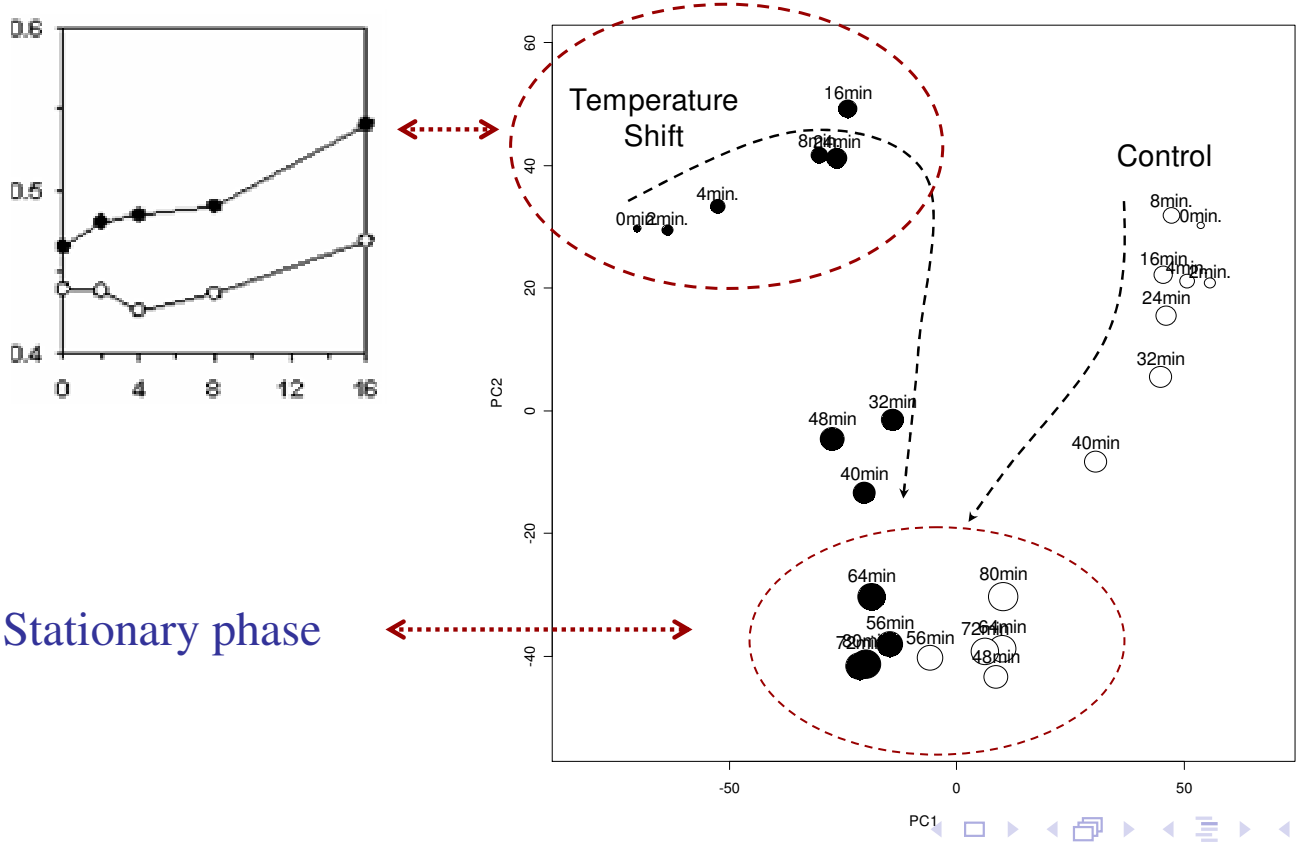


- A single channel microarray covering the entire transcriptional capacity of *E. coli* K12.
- A **control experiment** with cells grown at 37 C
- A **temperature shift experiment** where cells were grown at **10 C** and then **shifted at 37**.

•**13 time points** with **6 replicates** (a total of **78 measurements**).

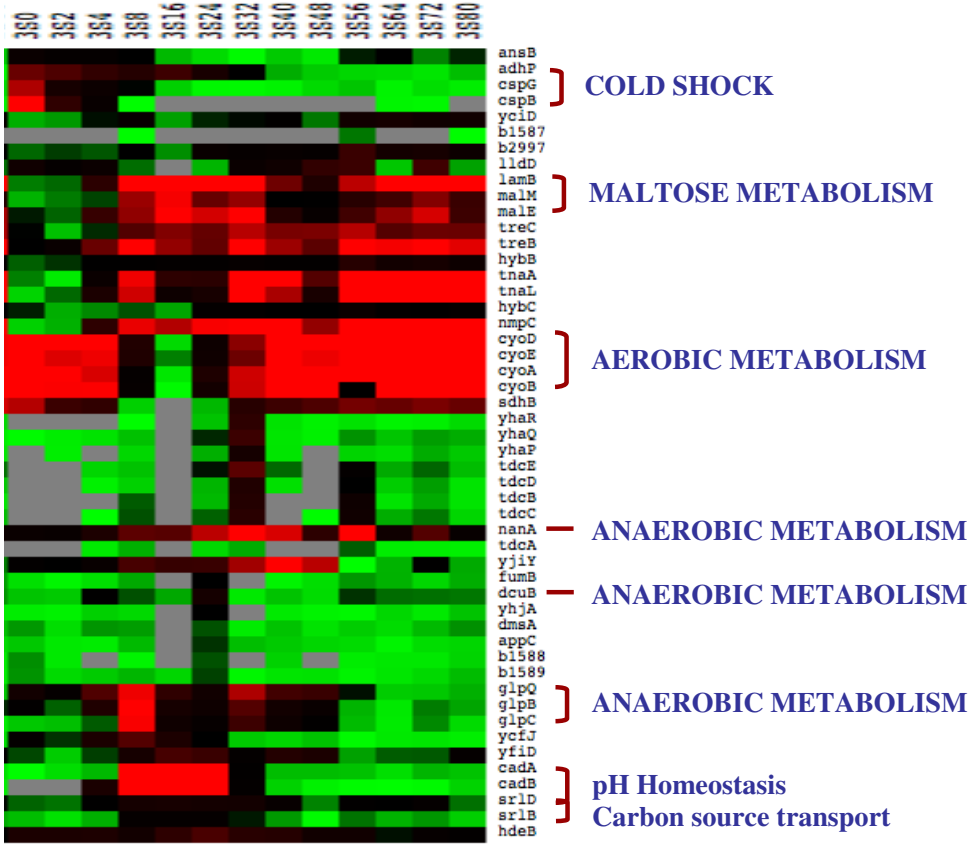
Principal Components Analysis

Dynamics of temperature adaptation at a genome level



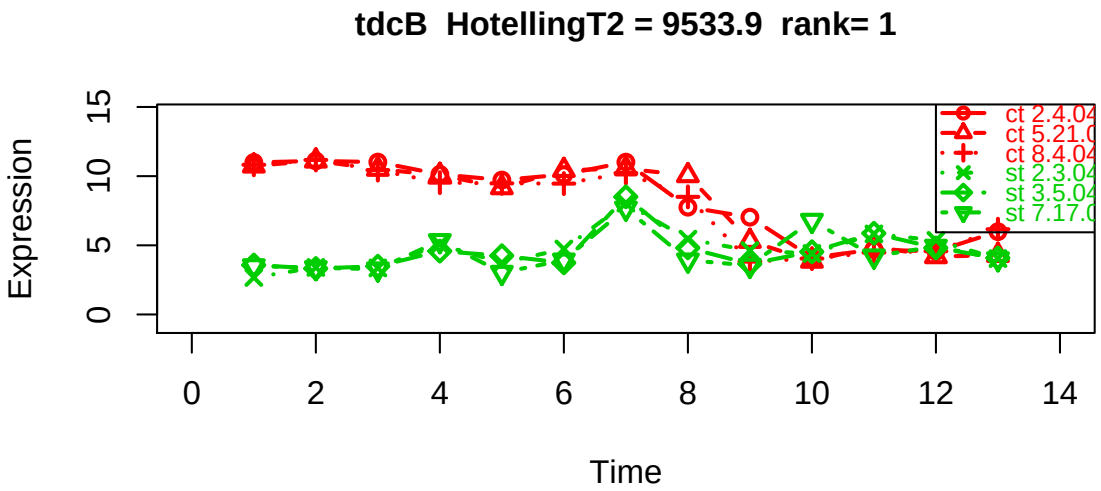
Stationary phase

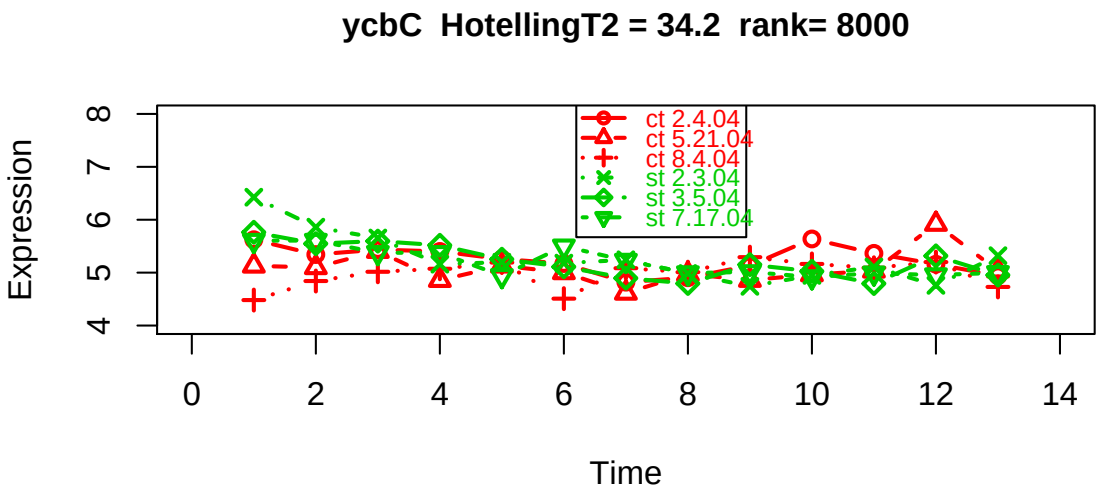
Functional Classification of Genes - Temperature Shifted Data



Pvalue	GO as name
2.87E-15	BP (electron transport)
1.26E-13	BP (generation of precursor metabolites and energy)
1.24E-06	BP (establishment of localization)
1.30E-06	BP (transport)
2.29E-06	BP (cellular respiration)
8.64E-05	BP (polyol metabolism)
0.000106	BP(tricarboxylic acid cycle)
0.000106	BP (coenzyme catabolism)
0.000106	BP (acetyl-CoA catabolism)
0.000106	BP (cofactor catabolism)
0.000173	BP (aerobic respiration)
0.000651	BP (acetyl-CoA metabolism)
0.000683	BP (energy derivation by oxidation of organic compounds)
0.00296	BP (main pathways of carbohydrate metabolism)
0.0064	BP (cellular physiological process)
0.00818	BP (galactitol metabolism)
0.00818	BP (alditol metabolism)
0.00818	BP (hexitol metabolism)
0.0192	BP (glycerol metabolism)
0.0193	BP (cellular process)
0.029	BP (alcohol metabolism)
0.0389	BP (carboxylic acid transport)
0.0411	BP (cellular metabolism)
0.0441	BP (anaerobic electron transport)
0.0441	BP (disaccharide transport)
0.0841	BP (carbohydrate transport)

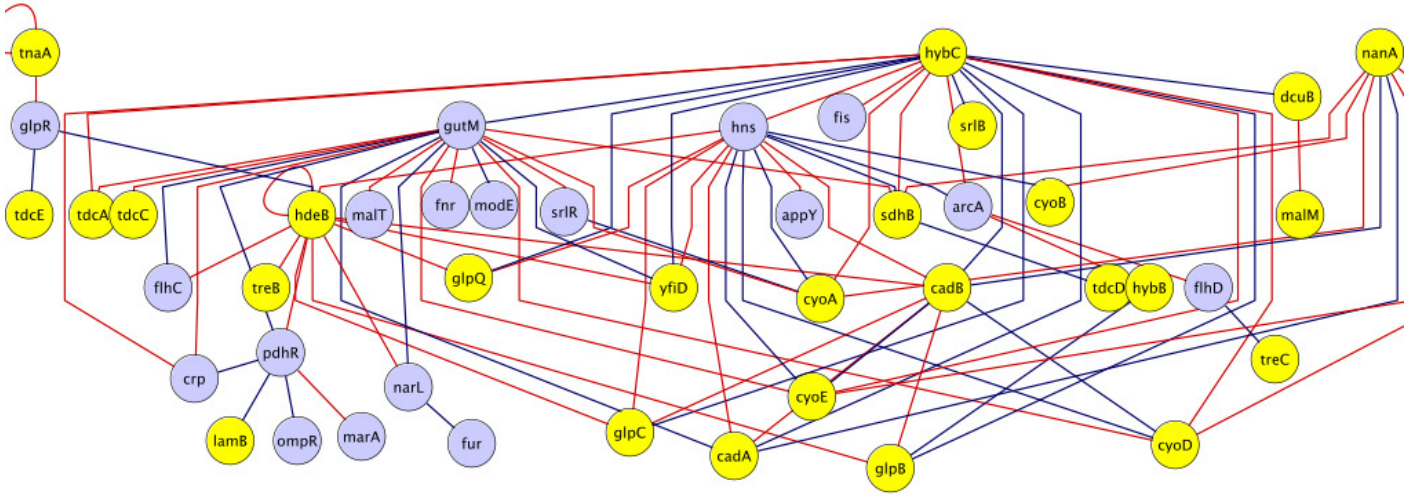
Gene Selection - Method of Tai and Speed





Inferred Network - Temperature Shifted Data

A network representing the adaptation of *E. coli* to 37 C

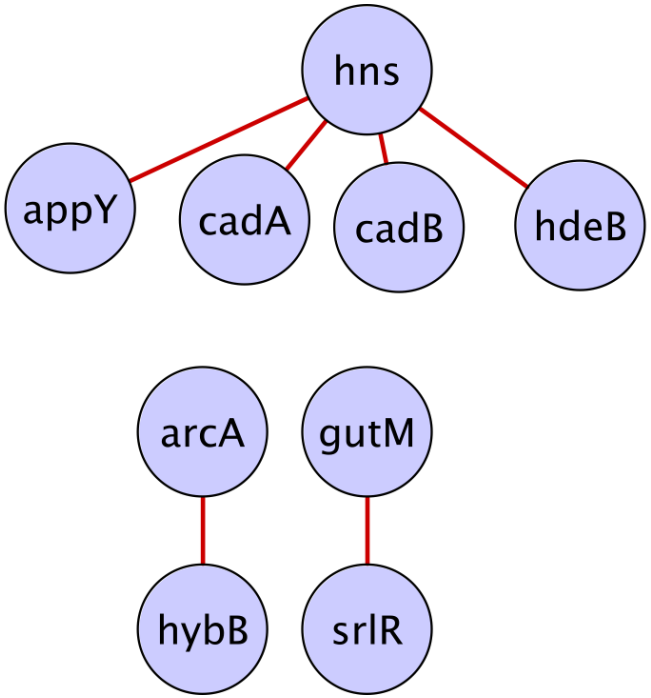


50 most differentially regulated genes

regulators

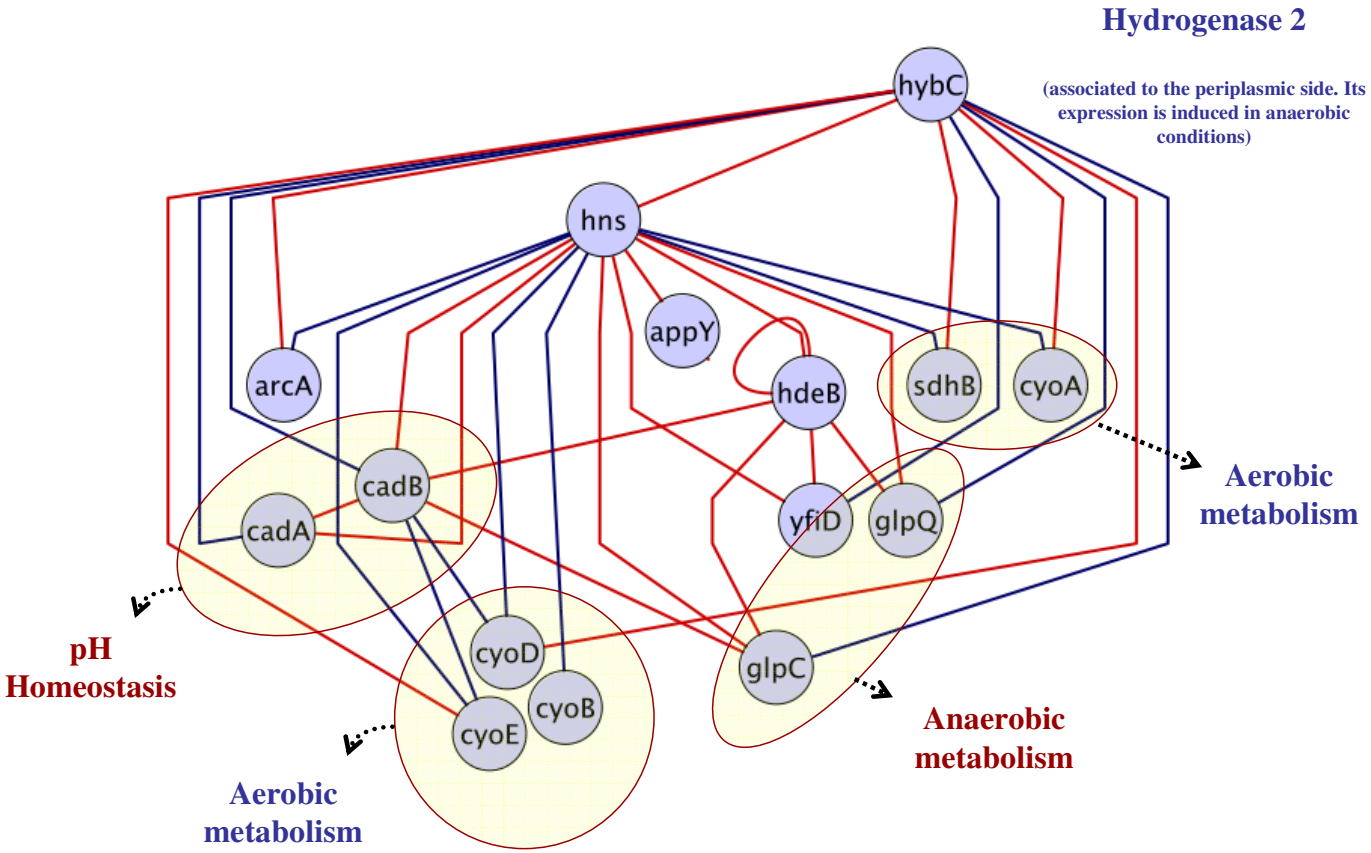
Regulatory Interactions in RegulonDB

Verified interactions in Regulon DB

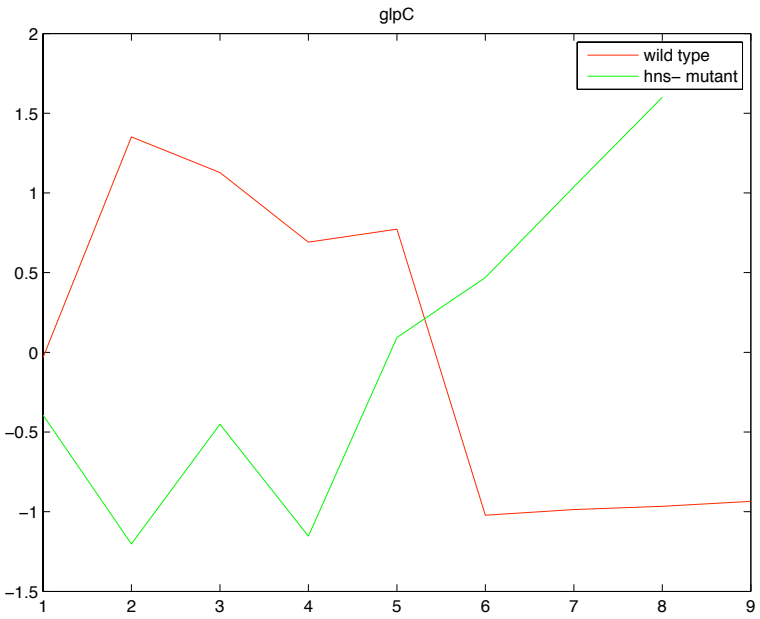


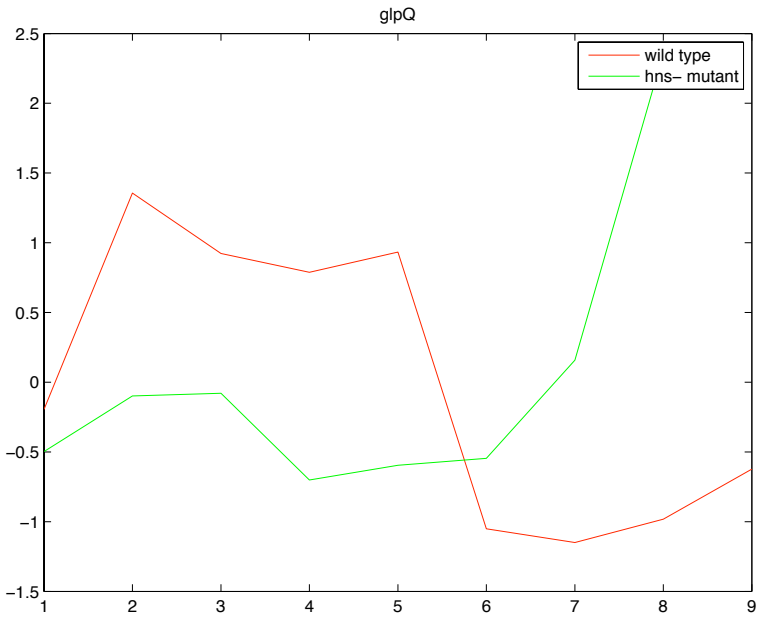
Subset of Interactions - Temperature Shifted Data

A role for hns in controlling Aerobic versus anaerobic switch?



Profiles of glpC and glpQ





Ongoing Work

- 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

Conclusions

- 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 *hns* as **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)