Model Based Identification of Transcription Factor Regulatory Activity via Markov Chain Monte Carlo

Simon Rogers\textsuperscript{1}, Raya Khanin\textsuperscript{2} and Mark Girolami\textsuperscript{1}

\textsuperscript{1}Bioinformatics Research Centre
Department of Computing Science
University of Glasgow
\textsuperscript{2}Department of Statistics
University of Glasgow

Glasgow Systems Biology Workshop, 2nd May 2006
Outline

1 Motivation
   • Regulatory Networks
   • Regulation

2 Model
   • Kinetic Model of transcription
   • Inference

3 Results
   • Synthetic Data
   • Fission Yeast Cell Cycle
   • Dataset Comparison
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Considerable effort has gone into the *Reverse Engineering* of regulatory networks from microarray data:
- To infer network topology
- To model the kinetics of specific regulatory interactions

All methods impose different assumptions
- Bayesian Nets - data must be *discretised*
- Correlation, state-space, linear regression - all *linear*
Modeling Regulatory Networks

- Considerable effort has gone into the *Reverse Engineering* of regulatory networks from microarray data:
  - To infer network topology
  - To model the kinetics of specific regulatory interactions

- All methods impose different assumptions
  - Bayesian Nets - data must be *discretised*
  - Correlation, state-space, linear regression - all *linear*
Modeling Regulatory Networks

- Considerable effort has gone into the Reverse Engineering of regulatory networks from microarray data:
  - To infer network topology
  - To model the kinetics of specific regulatory interactions

- All methods impose different assumptions
  - Bayesian Nets - data must be \textit{discretised}
  - Correlation, state-space, linear regression - all \textit{linear}
Additionally, all of these approaches make one other major assumption...

- Expression of the gene coding the TF is equivalent to the activity of the TF

In many examples, this is not the case
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Unobservable Modifications

In many cases gene expression is **not an accurate replacement for transcription factor activity**.
Examples 1, HIF

- **HIF**: HIF-1 is an important TF that stimulates tumour growth and metastases
- No over-expression of HIF-1 gene found in human breast cancer samples
- **But**: Over-expression of HIF-1 protein was found
- Other mechanisms must be responsible
- Vleugel *et al.*, 2004, Cell.Oncol.26
Examples 2, Fission Yeast

expression of SEP

expression of SEP's targets

Rogers, Khanin, Girolami (Glasgow)
Goal

- Translation and post-translational modifications result in a lack of correlation between gene expression and protein activity level.
- Therefore, the activity profile (TFA) cannot be approximated by transcription factor expression.
- We would like to be able to infer the levels of TF activity from the expression profiles of the target genes.
Previous Approaches

Several approaches based on linear (or log-linear) models of transcription
- Boulesteix and Strimmer (2005)
- Kao et al. (2004)

Fewer approaches using more realistic transcription models
- Nachman et al. (2004) Used non-linear model of transcription within the framework of Bayesian networks
- Khanin et al. (2005) Michaelis-Menten model of transcription, TFA inferred via Maximum Likelihood

The work here extends on the previous work of Khanin et al.
Previous Approaches

- Several approaches based on linear (or log-linear) models of transcription
  - Boulesteix and Strimmer (2005)
  - Kao et al. (2004)

- Fewer approaches using more realistic transcription models
  - Nachman et al. (2004) Used non-linear model of transcription within the framework of Bayesian networks
  - Khanin et al. (2005) Michaelis-Menten model of transcription, TFA inferred via Maximum Likelihood

- The work here extends on the previous work of Khanin et al.
Previous Approaches

- Several approaches based on linear (or log-linear) models of transcription
  - Boulesteix and Strimmer (2005)
  - Kao et al. (2004)

- Fewer approaches using more realistic transcription models
  - Nachman et al. (2004) Used non-linear model of transcription within the framework of Bayesian networks
  - Khanin et al. (2005) Michaelis-Menten model of transcription, TFA inferred via Maximum Likelihood

- The work here extends on the previous work of Khanin et al.
Aside - Single Input Motifs (SIMs)

Common network *motifs* - consisting of one TF regulating several target genes
Outline

1. Motivation
   - Regulatory Networks
   - Regulation

2. Model
   - Kinetic Model of transcription
   - Inference

3. Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Transcription Model

Most previous approaches have assumed a linear model of transcription

Fail to capture non-linearities e.g. saturation

The Michaelis-Menten model has been used previously for regulation

\[
\dot{\mu}_i = \alpha + p(\eta_i) - \delta \mu_i \tag{1}
\]

where \( p(\eta_i) = \beta \frac{\eta_i}{\eta_i + K} \) \tag{2}
Model Overview

TFA to be inferred

\( \eta \)

Gene specific kinetic parameters to be inferred

\( \beta, \alpha, K, \delta \)

Observed Expression Data

\( \mu(t) \)
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Khanin *et al.* used Maximum Likelihood to produce point estimates of $\eta$ and the kinetic parameters for each gene.

Point estimates of parameter values provide little information:
- Calculation of confidence intervals is non-trivial.

Full Bayesian inference would be more desirable:
- Full posteriors over parameters provide information regarding confidence and parameter sensitivity.
- Prior knowledge regarding parameter values and TFA profiles can be easily encoded through prior distributions.
- Straightforward to extend - discussed in future work.
- Implementation more straightforward.
Khanin et al. used Maximum Likelihood to produce point estimates of $\eta$ and the kinetic parameters for each gene.

Point estimates of parameter values provide little information:
- Calculation of confidence intervals is non-trivial.

Full Bayesian inference would be more desirable:
- Full posteriors over parameters provide information regarding confidence and parameter sensitivity.
- Prior knowledge regarding parameter values and TFA profiles can be easily encoded through prior distributions.
- Straightforward to extend - discussed in future work.
- Implementation more straightforward.
Inference

- **Sampling:** Metropolis Algorithm with Gaussian jumping distribution
- **Priors:** Uniform priors for parameters and $\eta$, Gamma prior for $\sigma^2$. 
Inference

- **Sampling**: Metropolis Algorithm with Gaussian jumping distribution
- **Priors**: Uniform priors for parameters and $\eta$, Gamma prior for $\sigma^2$. 
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Synthetic Dataset

- 10 Genes
- 10 time points
- 3 replicates
- Activation
- 3 Separate datasets with $\sigma^2 = 0.01, 0.05, 0.1$
Synthetic - Inferred $\eta$ profiles
Synthetic - Inferred Expression profiles

\[ \sigma^2 = 0.01 \]

\[ \sigma^2 = 0.05 \]

\[ \sigma^2 = 0.1 \]
Synthetic - Inferred Expression profiles

\[ p(\sigma^2 | \ldots) \]

\( \sigma^2 \)
Outline

1. Motivation
   - Regulatory Networks
   - Regulation

2. Model
   - Kinetic Model of transcription
   - Inference

3. Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Cell-cycle regulation in Fission Yeast
Fission Yeast Dataset

- 20 time points (samples taken every 15 minutes)
- 3 Replicates
- From Rustici et al, Nature Genetics 2004
- Lots of other data available

Rogers, Khanin, Girolami (Glasgow)
Fission Yeast - inferred $\eta$
Fission Yeast - inferred expression profiles
Fission Yeast - inferred expression profiles
Fission Yeast - inferred expression profiles

\[ p(\sigma^2 | \ldots) \]

\( \sigma^2 \)

\( \sigma^2 \)

\( \sigma^2 \)

\( \sigma^2 \)
Comparison

- Could we have created the same model without inferring $\eta$?
- Try fixing $\eta$ equal to the expression of SEP

Rogers,Khanin,Girolami (Glasgow) MCMC for TFA 28 / 37
Comparison

Could we have created the same model without inferring $\eta$?
Try fixing $\eta$ equal to the expression of SEP
Comparison

- Could we have created the same model without inferring $\eta$?
- Try fixing $\eta$ equal to the expression of SEP
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Dataset Comparison
Combining datasets

- So far we have considered expression with multiple replicates
- What about multiple datasets?
  - Kinetic parameters should be conserved
  - Only $\eta$ and noise $\sigma^2$ should change
  - Can combine hoping to improve inference

Datasets under different $\eta$ conditions
Different Sample Synchronisations
Combining datasets

- So far we have considered expression with multiple replicates
- What about multiple datasets?
  - Kinetic parameters should be conserved
  - Only $\eta$ and noise $\sigma^2$ should change
  - Can combine hoping to improve inference
    - Datasets under different $\eta$ conditions
    - Different Sample Synchronisations
Combining datasets

\[ \eta_1 \]

\[ \eta_2 \]

\[ \beta, \alpha, K, \delta \]

Observed Expression Data
Synthetic Combination Experiment

- 3 genes, 10 time-points
- Create 2 datasets from different true $\eta$ profiles
  - Dataset 1 - High Noise, 1 replicate
  - Dataset 2 - Low Noise, 3 replicates
- Infer $\eta$ profile for dataset 1 using just dataset 1
- Infer $\eta$ profile for dataset 1 using both datasets (with shared kinetic parameters)
- Will investigate whether or not inference is improved via addition of data under different $\eta$ conditions
Synthetic Combination Experiment

- 3 genes, 10 time-points
- Create 2 datasets from different true $\eta$ profiles
  - Dataset 1 - High Noise, 1 replicate
  - Dataset 2 - Low Noise, 3 replicates
- Infer $\eta$ profile for dataset 1 using just dataset 1
- Infer $\eta$ profile for dataset 1 using both datasets (with shared kinetic parameters)
- Will investigate whether or not inference is improved via addition of data under different $\eta$ conditions
Synthetic Combination Experiment

- 3 genes, 10 time-points
- Create 2 datasets from *different* true $\eta$ profiles
  - Dataset 1 - High Noise, 1 replicate
  - Dataset 2 - Low Noise, 3 replicates
- Infer $\eta$ profile for dataset 1 using *just* dataset 1
- Infer $\eta$ profile for dataset 1 using *both* datasets (with shared kinetic parameters)
- Will investigate whether or not inference is improved via addition of data under different $\eta$ conditions
Synthetic Combination Experiment

- 3 genes, 10 time-points
- Create 2 datasets from different true $\eta$ profiles
  - Dataset 1 - High Noise, 1 replicate
  - Dataset 2 - Low Noise, 3 replicates
- Infer $\eta$ profile for dataset 1 using just dataset 1
- Infer $\eta$ profile for dataset 1 using both datasets (with shared kinetic parameters)
- Will investigate whether or not inference is improved via addition of data under different $\eta$ conditions
Synthetic Combination Experiment

- 3 genes, 10 time-points
- Create 2 datasets from different true $\eta$ profiles
  - Dataset 1 - High Noise, 1 replicate
  - Dataset 2 - Low Noise, 3 replicates
- Infer $\eta$ profile for dataset 1 using just dataset 1
- Infer $\eta$ profile for dataset 1 using both datasets (with shared kinetic parameters)
- Will investigate whether or not inference is improved via addition of data under different $\eta$ conditions
Inferred $\eta$ profile on single dataset
Inferred $\eta$ profile on combined dataset

![Graph showing inferred and true profiles with MCMC for TFA by Rogers, Khanin, and Girolami (Glasgow)]
Typical gene expression profile
Summary

- The expression of the gene coding for a TF can not generally be used to approximate the TFA.
- Using Bayesian inference and a non-linear kinetic model, it is possible to infer the TFA from the expression of the target genes.
- In this setting, it is possible to combine datasets to improve inference when data is sparse.

Future work
- Extend to other network motifs (MIM, FFL, etc).
- Incorporating models of translation, delays etc.
- Discriminating between competing Biological hypothesis - e.g. possible post-translation modifications of SEP.
Summary

- The expression of the gene coding for a TF can not generally be used to approximate the TFA.
- Using Bayesian inference and a non-linear kinetic model, it is possible to infer the TFA from the expression of the target genes.
- In this setting, it is possible to combine datasets to improve inference when data is sparse.

Future work
- Extend to other network motifs (MIM, FFL, etc).
- Incorporating models of translation, delays etc.
- Discriminating between competing Biological hypothesis - e.g. possible post-translation modifications of SEP.
Acknowledgments

- Simon Rogers and Mark Girolami are supported by EPSRC grant EP/C010620/1 "Stochastic Modelling and Statistical Inference of Gene Regulatory Pathways: Integrating Multiple Sources of Data"
- Raya Khanin is supported by a RCUK fellowship in the Department of Statistics