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

Outline

1. Motivation
   - Regulatory Networks
   - Regulation

2. Model
   - Kinetic Model of transcription
   - Noise Model
   - Inference

3. Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
Outline

1. Motivation
   - Regulatory Networks
     - Regulation

2. Model
   - Kinetic Model of transcription
   - Noise Model
   - Inference

3. Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
Modeling Regulatory Networks

Considerable effort has gone into the Reverse Engineering of regulatory networks from microarray data:

- Bayesian Networks (Friedman et al, 2000, Husmeier 2004)
- Conditional Correlation Analysis (Rice and Tu, 2004)
- State-space Models (Rangel et al, 2004)
- Linear Regression (Rogers and Girolami, 2005)

Each method imposes its own assumptions

- Bayesian Nets - data must be discretised
- Correlation, state-space, linear regression - all linear
Considerable effort has gone into the *Reverse Engineering* of regulatory networks from microarray data:

- Bayesian Networks (Friedman et al, 2000, Husmeier 2004)
- Conditional Correlation Analysis (Rice and Tu, 2004)
- State-space Models (Rangel et al, 2004)
- Linear Regression (Rogers and Girolami, 2005)

Each method imposes its own 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:
  - Bayesian Networks (Friedman et al, 2000, Husmeier 2004)
  - Conditional Correlation Analysis (Rice and Tu, 2004)
  - State-space Models (Rangel et al, 2004)
  - Linear Regression (Rogers and Girolami, 2005)

- Each method imposes its own 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
- e.g. \( e_g = \sum_{t \in \text{TF}} w_t e_t \)

- Modifications that are not observable on the microarray mean that this is often a poor assumption.
Outline

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Noise Model
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
Regulation

Gene Coding for TF

Transcription → mRNA → Translation

Gene Product

Translation → mRNA → Transcription

Target Gene

TF
The Microarray View

Gene Coding for TF

Transcription → mRNA → Translation

Gene Product

TF

Target Gene

Microarray
In many cases gene expression is **not an accurate replacement for transcription factor activity**.
**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
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 and is difficult to measure *in-vitro*.
- 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 - directly relating mRNA *level* to TFA level
  - Boulesteix and Strimmer (2005)
  - Sanguinetti *et al.* (2006)

- Fewer approaches using more realistic transcription models - relating mRNA *production* to TFA level
  - Nachman *et al.* (2004) Used non-linear model of transcription within the framework of Bayesian networks

- 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 - directly relating mRNA level to TFA level
  - Boulesteix and Strimmer (2005)
  - Kao et al. (2004)
  - Sanguinetti et al. (2006)

- Fewer approaches using more realistic transcription models - relating mRNA production to TFA level
  - 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 - directly relating mRNA level to TFA level
  - Boulesteix and Strimmer (2005)
  - Kao et al. (2004)
  - Sanguinetti et al. (2006)

- Fewer approaches using more realistic transcription models - relating mRNA production to TFA level
  - 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
   - Noise Model
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
The Model

- In order to infer the TFA from the target gene expression profiles, a model of transcription is required.
  - Most basic linear model?
    - Cannot capture non-linear effects - e.g. saturation
    - Models level of mRNA and not production
  - Linear ODE?
    - More Biologically intuitive but still incapable of modeling non-linear effects
  - Michaelis-Menten kinetic model
    - Can capture some non-linear effects
    - Still a reasonably simple model
  - Stochastic model?
    - Inappropriate data
  - Use Michaelis-Menten kinetic model
In order to infer the TFA from the target gene expression profiles, a model of transcription is required.

Most basic linear model?
- Cannot capture non-linear effects - e.g. saturation
- Models level of mRNA and not production

Linear ODE?
- More Biologically intuitive but still incapable of modeling non-linear effects

Michaelis-Menten kinetic model
- Can capture some non-linear effects
- Still a reasonably simple model

Stochastic model?
- Inappropriate data

Use Michaelis-Menten kinetic model
The Model

In order to infer the TFA from the target gene expression profiles, a model of transcription is required.

Most basic linear model?
- Cannot capture non-linear effects - e.g. saturation
- Models level of mRNA and not production

Linear ODE?
- More Biologically intuitive but still incapable of modeling non-linear effects

Michaelis-Menten kinetic model
- Can capture some non-linear effects
- Still a reasonably simple model

Stochastic model?
- Inappropriate data

Use Michaelis-Menten kinetic model
The Model

- In order to infer the TFA from the target gene expression profiles, a model of transcription is required.
- Most basic linear model?
  - Cannot capture non-linear effects - e.g. saturation
  - Models level of mRNA and not production

- Linear ODE?
  - More Biologically intuitive but still incapable of modeling non-linear effects

- Michaelis-Menten kinetic model
  - Can capture some non-linear effects
  - Still a reasonably simple model

- Stochastic model?
  - Inappropriate data

- Use Michaelis-Menten kinetic model
The Model

- In order to infer the TFA from the target gene expression profiles, a model of transcription is required.
- Most basic linear model?
  - Cannot capture non-linear effects - e.g. saturation
  - Models level of mRNA and not production
- Linear ODE?
  - More Biologically intuitive but still incapable of modeling non-linear effects
- Michaelis-Menten kinetic model
  - Can capture some non-linear effects
  - Still a reasonably simple model
- Stochastic model?
  - Inappropriate data
- Use Michaelis-Menten kinetic model
The Model

- In order to infer the TFA from the target gene expression profiles, a model of transcription is required.

- Most basic linear model?
  - Cannot capture non-linear effects - e.g. saturation
  - Models level of mRNA and not production

- Linear ODE?
  - More Biologically intuitive but still incapable of modeling non-linear effects

- Michaelis-Menten kinetic model
  - Can capture some non-linear effects
  - Still a reasonably simple model

- Stochastic model?
  - Inappropriate data

- Use **Michaelis-Menten** kinetic model
Some Definitions

- Observations at $T + 1$ time-points, $t_0, \ldots, t_i, \ldots t_T$
- Observed expression level of $r$th replicate of gene $g$, at time $t_i$ - $e_{gi}^r$
- Transcription factor activity (to be inferred) - $\eta_0, \ldots, \eta_i, \ldots, \eta_T$
- Expression level of gene $g$ at time $t_i$ predicted by model - $\mu_{gi}$
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

**Activation:**

\[ \dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_{gi} \]

- \( \alpha_g \) is the basal level of production
- \( \beta_g \) is the rate of production
- \( K_g \) is the half-saturation constant
- \( \delta_g \) is the level of (linear) decay
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

**Activation:**

\[
\dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_{gi}
\]

where...

- $\alpha_g$ is the basal level of production
- $\beta_g$ is the rate of production
- $K_g$ is the half-saturation constant
- $\delta_g$ is the level of (linear) decay
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

**Activation:**

\[
\dot{\mu}_g = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_g_i
\]

where...

- \(\alpha_g\) is the basal level of production
- \(\beta_g\) is the rate of production
- \(K_g\) is the half-saturation constant
- \(\delta_g\) is the level of (linear) decay
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

Activation:

\[ \dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_{gi} \]

where...

- \( \alpha_g \) is the basal level of production
- \( \beta_g \) is the rate of production
- \( K_g \) is the half-saturation constant
- \( \delta_g \) is the level of (linear) decay
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

**Activation:**

\[
\dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_{gi}
\]

where...

- \(\alpha_g\) is the basal level of production
- \(\beta_g\) is the rate of production
- \(K_g\) is the half-saturation constant
- \(\delta_g\) is the level of (linear) decay
Michaelis-Menten (MM)

The MM model defines the rate of change of gene expression - separate forms for a TF acting as an activator and a repressor

**Activation:**

\[
\dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + K_g} - \delta_g \mu_{gi}
\]

where...

- \(\alpha_g\) is the basal level of production
- \(\beta_g\) is the rate of production
- \(K_g\) is the half-saturation constant
- \(\delta_g\) is the level of (linear) decay

**Repression:**

\[
\dot{\mu}_{gi} = \alpha_g + \beta_g \frac{1}{\eta_i + K_g} - \delta_g \mu_{gi}
\]
Outline

1 Motivation
   • Regulatory Networks
   • Regulation

2 Model
   • Kinetic Model of transcription
   • Noise Model
   • Inference

3 Results
   • Synthetic Data
   • Fission Yeast Cell Cycle
   • Incorporating delays and shifts
Noise Model

- For MM kinetic model to make sense, we need to work in the original (not log) space. Hence lognormal noise model is appropriate.

- For a dataset consisting of $T$ time-points and $R$ replicates

$$p(e_g | \mu_{gt}, \sigma^2) = \prod_{t=0}^{T} \prod_{r=1}^{R} \frac{1}{\sqrt{2\pi} \sigma e_{gi}^r} \exp \left\{ -\frac{1}{2\sigma^2} \left( \log(e_{gi}^r) - m_{gi} \right)^2 \right\}$$

- $\sigma^2$ is the noise variance.
- $e_{gi}^r$ is the expression of gene $g$ at time $t_i$ in replicate $r$.
- $m_{gi}$ is the location of the log normal density given by
  $$m_{gi} = \log \mu_{gi} - \frac{1}{2} \sigma^2$$
Noise Model

For MM kinetic model to make sense, we need to work in the original (not log) space. Hence lognormal noise model is appropriate.

For a dataset consisting of $T$ time-points and $R$ replicates

$$\rho(e_g|\mu_{gt}, \sigma^2) = \prod_{t=0}^{T} \prod_{r=1}^{R} \frac{1}{\sqrt{2\pi}\sigma^2} e^{rgi} \exp \left\{-\frac{1}{2\sigma^2}(\log(e^{rgi}) - m_{gi})^2\right\}$$

where...

- $\sigma^2$ is the noise variance
- $e^{rgi}$ is the expression of gene $g$ at time $t_i$ in replicate $r$
- $m_{gi}$ is the location of the log normal density given by $m_{gi} = \log \mu_{gi} - \frac{1}{2}\sigma^2$
Noise Model

- For MM kinetic model to make sense, we need to work in the original (not log) space. Hence lognormal noise model is appropriate.

- For a dataset consisting of $T$ time-points and $R$ replicates

$$p(e_g|\mu_{gt}, \sigma^2) = \prod_{t=0}^{T} \prod_{r=1}^{R} \frac{1}{\sqrt{2\pi} \sigma e_{gi}^r} \exp \left\{ -\frac{1}{2\sigma^2} \left( \log(e_{gi}^r) - m_{gi} \right)^2 \right\}$$

where...

- $\sigma^2$ is the noise variance.
- $e_{gi}^r$ is the expression of gene $g$ at time $t_i$ in replicate $r$.
- $m_{gi}$ is the location of the log normal density given by $m_{gi} = \log \mu_{gi} - \frac{1}{2} \sigma^2$. 
For MM kinetic model to make sense, we need to work in the original (not log) space. Hence lognormal noise model is appropriate.

For a dataset consisting of $T$ time-points and $R$ replicates

$$p(e_g|\mu_{gt}, \sigma^2) = \prod_{t=0}^{T} \prod_{r=1}^{R} \frac{1}{\sqrt{2\pi}\sigma e_{gi}^r} \exp \left\{ -\frac{1}{2\sigma^2} \left( \log(e_{gi}^r) - m_{gi} \right)^2 \right\}$$

where...

- $\sigma^2$ is the noise variance
- $e_{gi}^r$ is the expression of gene $g$ at time $t_i$ in replicate $r$
- $m_{gi}$ is the location of the log normal density given by $m_{gi} = \log \mu_{gi} - \frac{1}{2} \sigma^2$
Model Overview

- TFA to be inferred
- Gene specific kinetic parameters to be inferred
- Observed Expression Data

\[ \eta \]

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

1 Motivation
   - Regulatory Networks
   - Regulation

2 Model
   - Kinetic Model of transcription
   - Noise Model
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
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
**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
   - Noise Model
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
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

Inferred

True

Rogers, Khanin, Girolami (Glasgow)

MCMC for TFA

PMSB 2006
Synthetic - Inferred $\eta$ profiles

![Graph showing inferred and true profiles]
Synthetic - Inferred $\eta$ profiles
Synthetic - Inferred Expression profiles

\[ \sigma^2 = 0.01 \]

\[ \sigma^2 = 0.05 \]

\[ \sigma^2 = 0.1 \]
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
   - Noise Model
   - Inference

3 Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
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
Fission Yeast - inferred $\eta$
Fission Yeast - inferred expression profiles
Fission Yeast - inferred expression profiles

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

\[ \sigma^2 \]
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
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
   - Noise Model
   - Inference

3. Results
   - Synthetic Data
   - Fission Yeast Cell Cycle
   - Incorporating delays and shifts
Gene-specific delays

Some genes do not fit the model very well due to gene-specific translation delays
Gene-specific delays

- Incorporate an extra \textbf{delay} parameter ($\tau_g$) for each gene
- Effective TFA at time $t$ for gene $g$ is now $\eta(t - \tau)$
- Place $\Gamma$ priors on $\tau_g$
- Fix one $\tau_g$ for identifiability
τ - fission yeast cell-cycle

Gene 8

Gene 9

Gene 12

MCMC for TFA

Rogers, Khanin, Girolami (Glasgow)

PMSB 2006 38 / 43
Example, gene 12

Blue - without delays, Red - with delays
Replicate-specific shifts

- Sample synchronisation is not perfect
- Results in replicates that are out of phase
- Introduce replicate specific shift parameter $\rho_r$
- Fix one $\rho_g$ for identifiability
Estimates could be used to align data prior to further analysis (c.f. Gilks et al 2005)
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).
- Discriminating between competing Biological hypothesis - e.g. possible post-translation modifications of SEP.
- Functional prior on $\eta$ - for example, Sanguinetti et al 2006. (PASCAL workshop on Gaussian Processes in Practice)
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).
- Discriminating between competing Biological hypothesis - e.g. possible post-translation modifications of SEP.
- Functional prior on $\eta$ - for example, Sanguinetti et al 2006. (PASCAL workshop on Gaussian Processes in Practice).
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