Using models of transcriptional regulation to uncover gene regulatory networks

Magnus Rattray
School of Computer Science, University of Manchester

joint work with Neil Lawrence and Antti Honkela

Imperial College, 15th February 2010

Talk outline

- Quick introduction to transcriptional regulation
- Our overall strategy for regulatory network inference
- Using simple activation models for target identification
- Closing the system with Gaussian process inference
- Empirical results on Drosophila mesoderm development

Transcriptional regulation of gene expression

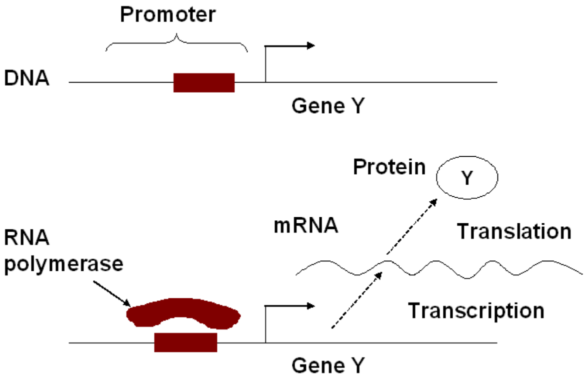


Figure from “An Introduction to Systems Biology” by U. Alon, 2006

Transcriptional regulation of gene expression

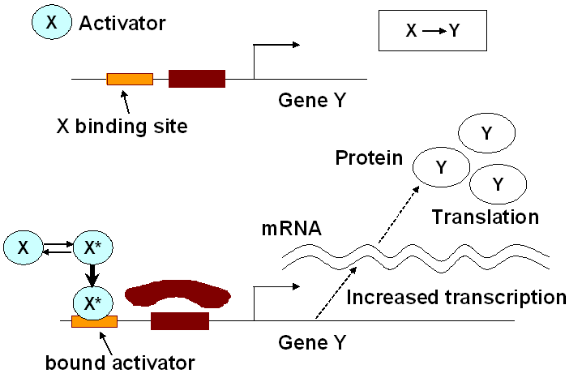


Figure from "An Introduction to Systems Biology" by U. Alon, 2006

Transcriptional regulation of gene expression

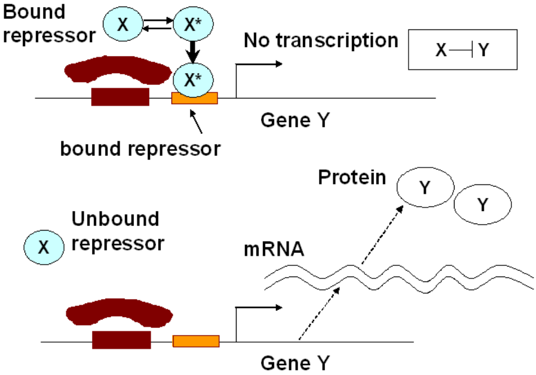
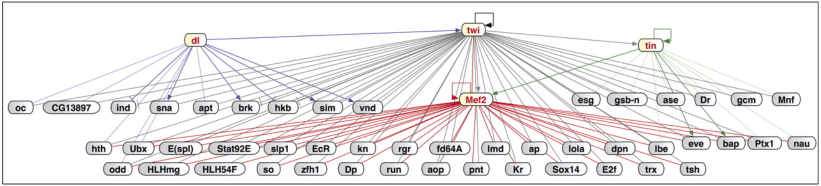


Figure from "An Introduction to Systems Biology" by U. Alon, 2006

Gene regulatory networks

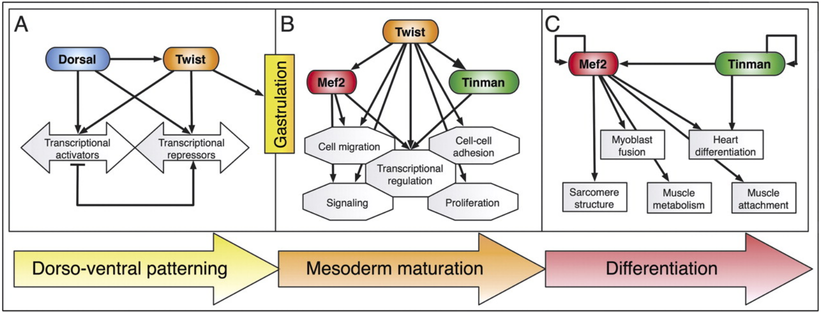
The core gene regulatory network controlling mesoderm development in Drosophila



Sandmann *et al.* Genes and Development 2007

Gene regulatory networks

The inferred network is used to help model biological processes



Sandmann *et al.* Genes and Development 2007

Inferring networks from data

- We have access to genome-wide data about. . .
  - Physical binding of transcription factors to DNA (ChIP)
  - Wild-type mRNA expression (microarrays/RNA-seq)
  - Mutant mRNA expression (microarrays/RNA-seq)
  - Spatial mRNA and protein expression (in situs)
  - DNA sequence

Inferring networks from data

- We have access to genome-wide data about. . .
  - Physical binding of transcription factors to DNA (ChIP)
  - Wild-type mRNA expression (microarrays/RNA-seq)
  - Mutant mRNA expression (microarrays/RNA-seq)
  - Spatial mRNA and protein expression (in situs)
  - DNA sequence
- None of the above data types provides a complete picture

Inferring networks from data

- We have access to genome-wide data about...
  - Physical binding of transcription factors to DNA (ChIP)
  - Wild-type mRNA expression (microarrays/RNA-seq)
  - Mutant mRNA expression (microarrays/RNA-seq)
  - Spatial mRNA and protein expression (in situs)
  - DNA sequence
- None of the above data types provides a complete picture
- We need to integrate them with our **model** of how gene regulation works

Inferring networks from data

- We have access to genome-wide data about. . .
  - Physical binding of transcription factors to DNA (ChIP)
  - Wild-type mRNA expression (microarrays/RNA-seq)
  - Mutant mRNA expression (microarrays/RNA-seq)
  - Spatial mRNA and protein expression (in situs)
  - DNA sequence
- None of the above data types provides a complete picture
- We need to integrate them with our **model** of how gene regulation works – Systems Biology provides a framework for modelling

Inferring networks from data

Our strategy:

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
    - Advantage: genome-wide coverage and *in vivo* method
    - Disadvantage: binding does not necessarily indicate regulation

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
  - Advantage: genome-wide coverage and *in vivo* method
  - Disadvantage: binding does not necessarily indicate regulation
- Fit regulation models to expression time-series data (microarray or RNA-seq) to identify functional enhancers

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
  - Advantage: genome-wide coverage and *in vivo* method
  - Disadvantage: binding does not necessarily indicate regulation
- Fit regulation models to expression time-series data (microarray or RNA-seq) to identify functional enhancers
  - Advantage: genome-wide coverage, quantitative data
  - Disadvantage: average over inhomogeneous cell population

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
    - Advantage: genome-wide coverage and *in vivo* method
    - Disadvantage: binding does not necessarily indicate regulation
- Fit regulation models to expression time-series data (microarray or RNA-seq) to identify functional enhancers
    - Advantage: genome-wide coverage, quantitative data
    - Disadvantage: average over inhomogeneous cell population
- Filter according to spatial expression patterns

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
  - Advantage: genome-wide coverage and *in vivo* method
  - Disadvantage: binding does not necessarily indicate regulation
- Fit regulation models to expression time-series data (microarray or RNA-seq) to identify functional enhancers
  - Advantage: genome-wide coverage, quantitative data
  - Disadvantage: average over inhomogeneous cell population
- Filter according to spatial expression patterns
  - Advantage: allows for spatial inhomogeneity
  - Disadvantage: less quantitative, less high-throughput

Inferring networks from data

Our strategy:

- Identify binding of transcription factors to DNA using chromatin immunoprecipitation (ChIP) experiments
  - Advantage: genome-wide coverage and in vivo method
  - Disadvantage: binding does not necessarily indicate regulation
- Fit regulation models to expression time-series data (microarray or RNA-seq) to identify functional enhancers
  - Advantage: genome-wide coverage, quantitative data
  - Disadvantage: average over inhomogeneous cell population
- Filter according to spatial expression patterns
  - Advantage: allows for spatial inhomogeneity
  - Disadvantage: less quantitative, less high-throughput

Modelling transcriptional regulation

Recall our simple picture of activation:

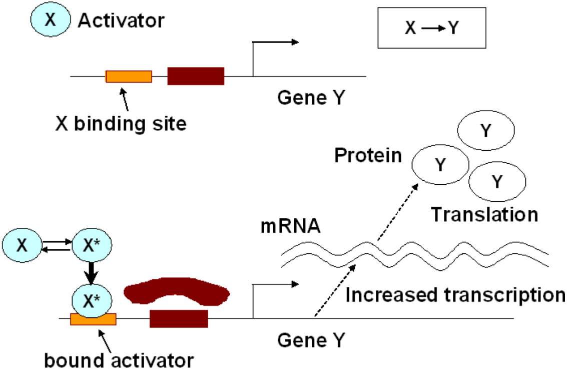


Figure from "An Introduction to Systems Biology" by U. Alon, 2006

Modelling transcriptional regulation

We model transcription factor translation and target activation:

$$\begin{aligned}\frac{\mathrm{d}f(t)}{\mathrm{d}t} &= m(t) - \delta f(t) \\ \frac{\mathrm{d}y_i(t)}{\mathrm{d}t} &= B_i + S_i f(t) - D_i y_i(t)\end{aligned}$$

- $m(t)$ – concentration of transcription factor mRNA
- $f(t)$ – concentration of transcription factor protein
- $y_i(t)$ – concentration of target gene i's mRNA

Modelling transcriptional regulation

We model transcription factor translation and target activation:

$$\begin{aligned}\frac{\mathrm{d}f(t)}{\mathrm{d}t} &= m(t) - \delta f(t) \\ \frac{\mathrm{d}y_i(t)}{\mathrm{d}t} &= B_i + S_i f(t) - D_i y_i(t)\end{aligned}$$

- $m(t)$ – concentration of transcription factor mRNA
- $f(t)$ – concentration of transcription factor protein
- $y_i(t)$ – concentration of target gene i's mRNA
- Application - identifying likely targets by their fit to the model

Modelling transcriptional regulation

We model transcription factor translation and target activation:

$$\begin{aligned}\frac{\mathrm{d}f(t)}{\mathrm{d}t} &= m(t) - \delta f(t)\\ \frac{\mathrm{d}y_i(t)}{\mathrm{d}t} &= B_i + S_i f(t) - D_i y_i(t)\end{aligned}$$

- $m(t)$ – concentration of transcription factor mRNA
- $f(t)$ – concentration of transcription factor protein
- $y_i(t)$ – concentration of target gene i's mRNA
- Application - identifying likely targets by their fit to the model
- Technical challenges - parameters $\theta = \{B_i, D_i, S_i, \delta\}$ unknown, few time points, noisy data, "open" system because of $m(t)$

Model-based inference

- Target expression data $Y = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_T)\}$

Model-based inference

- Target expression data $Y = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_T)\}$
- Trancription factor expression $M = \{m(t_1), m(t_2), \ldots, m(t_T)\}$

Model-based inference

- Target expression data $Y = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_T)\}$
- Trancription factor expression $M = \{m(t_1), m(t_2), \ldots, m(t_T)\}$ (NB. this is optional – we ignore translation layer if the transcription factor protein is regulated post-transcription)

Model-based inference

- Target expression data $Y = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_T)\}$
- Trancription factor expression $M = \{m(t_1), m(t_2), \ldots, m(t_T)\}$ (NB. this is optional – we ignore translation layer if the transcription factor protein is regulated post-transcription)
- Treat $m(t)$ [or $f(t)$] as functional parameters that can be “marginalised out” using non-parametric Bayesian methods

Model-based inference

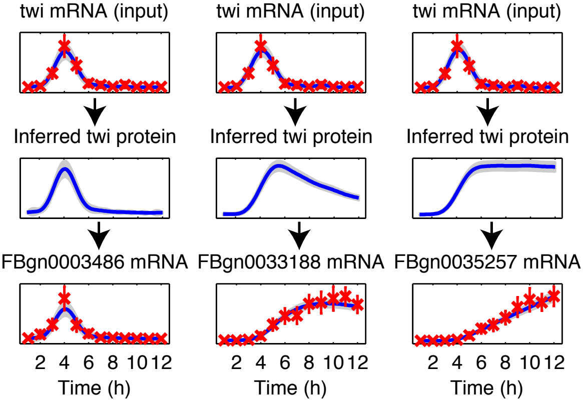
- Target expression data $Y = \{y_i(t_1), y_i(t_2), \ldots, y_i(t_T)\}$
- Trancription factor expression $M = \{m(t_1), m(t_2), \ldots, m(t_T)\}$ (NB. this is optional – we ignore translation layer if the transcription factor protein is regulated post-transcription)
- Treat $m(t)$ [or $f(t)$] as functional parameters that can be "marginalised out" using non-parametric Bayesian methods
- Fit model parameters $\theta = \{B_i, D_i, S_i, \delta\}$ by maximising the likelihood $p(Y, M|\theta)$ obtained by using a noise model

Model-based inference

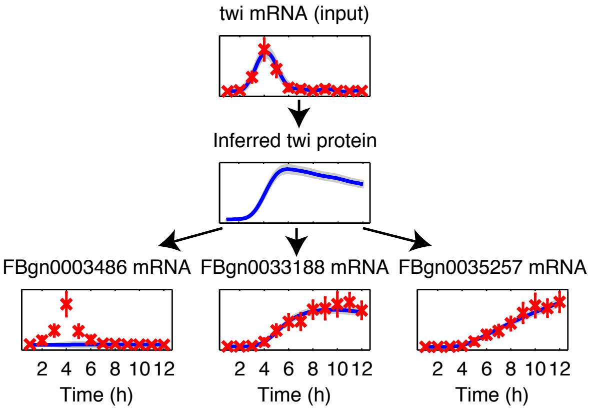
- Target expression data $Y = \{y_i(t_1), y_i(t_2), \dots, y_i(t_T)\}$
- Trancription factor expression $M = \{m(t_1), m(t_2), \dots, m(t_T)\}$ (NB. this is optional – we ignore translation layer if the transcription factor protein is regulated post-transcription)
- Treat $m(t)$ [or $f(t)$] as functional parameters that can be "marginalised out" using non-parametric Bayesian methods
- Fit model parameters $\theta = \{B_i, D_i, S_i, \delta\}$ by maximising the likelihood $p(Y, M|\theta)$ obtained by using a noise model
- Use likelihood score for genome-wide ranking of all genes as putative targets

Gao *et al.* Bioinformatics 24(16), i70-i75 (2008)

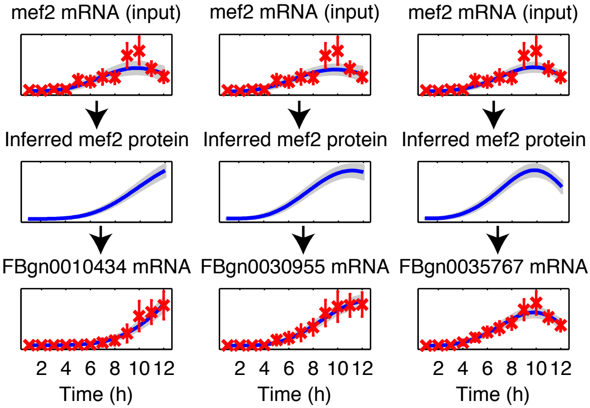
Fitting the model to data - Twist and Mef2



Fitting the model to data - Twist and Mef2



Fitting the model to data - Twist and Mef2



Gaussian process: definition

We model the transcription factor mRNA $m(t)$ as a sample drawn from a Gaussian process prior distribution

Gaussian process: definition

We model the transcription factor mRNA $m(t)$ as a sample drawn from a Gaussian process prior distribution

- A Gaussian Process (GP) is a distribution over functions $m(t)$

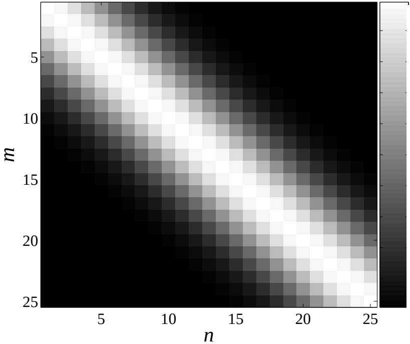
$$m(t) \backsim \mathcal{GP}\left(\mu_m(t), k_m(t, t')\right)$$

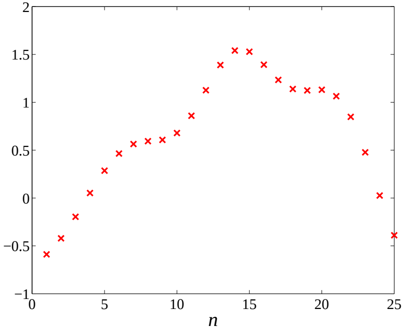
- It is characterised by a mean and covariance function

$$\begin{aligned} \mu_m(t) &= \mathbb{E}\left[m(t)\right] \\ k_m(t, t') &= \mathbb{E}\left[\left(m(t) - \mu(t)\right)\left(m(t') - \mu(t')\right)\right] \end{aligned}$$

- Any finite set of points sampled from the function are Gaussian distributed with covariance matrix elements $C_{ij} = k(t_i, t_j)$

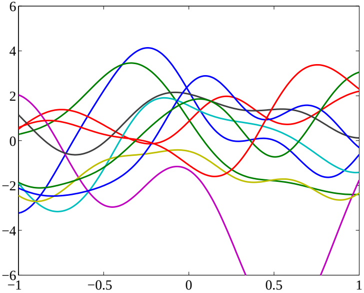
From a Gaussian distribution to a Gaussian process





Covariance function for $m(t)$

We assume a squared exponential covariance function for $m(t)$



$$\mu_m(t) = 0 \qquad k_m(t, t') = h \exp\left(-\frac{(t-t')^2}{l^2}\right)$$

Covariance function for the linear activation model

Recall the linear activation model

$$\begin{aligned}\frac{\mathrm{d}f(t)}{\mathrm{d}t} &= m(t) - \delta f(t)\\ \frac{\mathrm{d}y_i(t)}{\mathrm{d}t} &= B_i + S_i f(t) - D_i y_i(t)\end{aligned}$$

This differential equation can be solved for $f(t)$ and $y_i(t)$ as

$$\begin{aligned}f(t) &= \int_0^t e^{-\delta(t-u)} m(u)\mathrm{d}u\\ y_i(t) &= \frac{B_i}{D_i} + S_i \int_0^t e^{-D_i(t-u)} f(u)\mathrm{d}u\end{aligned}$$

Covariance function for the linear activation model

Recall the linear activation model

$$\begin{aligned}\frac{\mathrm{d}f(t)}{\mathrm{d}t} &= m(t) - \delta f(t)\\ \frac{\mathrm{d}y_i(t)}{\mathrm{d}t} &= B_i + S_i f(t) - D_i y_i(t)\end{aligned}$$

This differential equation can be solved for $f(t)$ and $y_i(t)$ as

$$\begin{aligned}f(t) &= \int_0^t e^{-\delta(t-u)} m(u)\mathrm{d}u\\ y_i(t) &= \frac{B_i}{D_i} + S_i \int_0^t e^{-D_i(t-u)} f(u)\mathrm{d}u\end{aligned}$$

Note: Both $f(t)$ and $y_i(t)$ are linear functions of $m(t)$

Covariance function for the linear activation model

Any linear operation on a GP $\Longrightarrow$ Related GP

$$m(t) \backsim \mathcal{GP}\left(0, k_m\left(t, t'\right)\right) \Longrightarrow y_i(t) \backsim \mathcal{GP}\left(\frac{B_i}{D_i}, k_{y_i}\left(t, t'\right)\right) \ .$$

The covariance of target gene mRNA $y_i(t)$ is defined:

$$k_{y_i}\left(t, t'\right) = S_i^2 \int_0^t \int_0^{t'} e^{-D_i(t-u)-D_i(t'-u')} k_f\left(u, u'\right) \mathrm{d}u \mathrm{d}u'$$

in terms of covariance of the TF protein $f(t)$ which is defined:

$$k_f\left(t, t'\right) = \int_0^t \int_0^{t'} e^{-\delta(t-u)-\delta(t'-u')} k_m\left(u, u'\right) \mathrm{d}u \mathrm{d}u' \ .$$

Computing the likelihood

We have a 2D process for the target and transcription factor mRNA

$$p(y, m|\theta) = \mathcal{GP}\left(\begin{bmatrix} 0 \\ \frac{B}{D} \end{bmatrix}, \begin{bmatrix} k_m & k_{my} \\ k_{ym} & k_y \end{bmatrix}\right)$$

with parameters $\theta = \{\delta, h, l, B, S, D\}$.

Computing the likelihood

We have a 2D process for the target and transcription factor mRNA

$$p(y, m|\theta) = \mathcal{GP}\left(\begin{bmatrix} 0 \\ \frac{B}{D} \end{bmatrix}, \begin{bmatrix} k_m & k_{my} \\ k_{ym} & k_y \end{bmatrix}\right)$$

with parameters $\theta = \{\delta, h, l, B, S, D\}$. Given noise-corrupted data $\boldsymbol{x} = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_T, \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$ then the data likelihood is:

$$\begin{aligned} L(\theta) = \log p(\boldsymbol{x}|\theta) &= \log \int p(\boldsymbol{x}|y, m) p(y, m|\theta) \, \mathrm{d}y \mathrm{d}m \\ &= \left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} C^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2} \log |C|\right] - T \log 2\pi \end{aligned}$$

where $C_{ij} = k(x_i, x_j) + \delta_{ij}\sigma_i^2$ is the data covariance.

Pros and cons of the Gaussian Process approach

Pros and cons of the Gaussian Process approach

Pros:

- The function $m(t)$ is integrated (marginalized) out of the likelihood so only two new parameters introduced
- All parameters can be efficiently estimated by maximum likelihood, allowing for genome-wide coverage
- No requirement for equal spacing of times
- Unobserved functions can be inferred very naturally

Pros and cons of the Gaussian Process approach

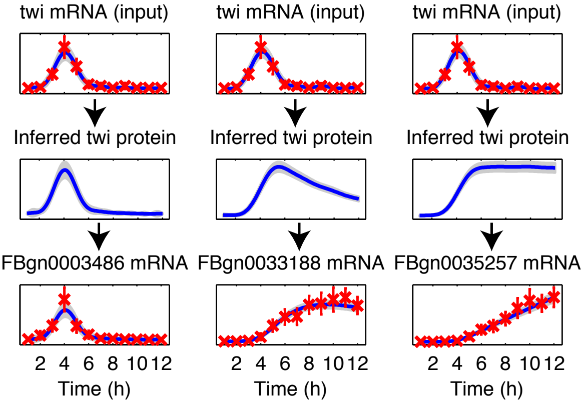
Pros:

- The function $m(t)$ is integrated (marginalized) out of the likelihood so only two new parameters introduced
- All parameters can be efficiently estimated by maximum likelihood, allowing for genome-wide coverage
- No requirement for equal spacing of times
- Unobserved functions can be inferred very naturally

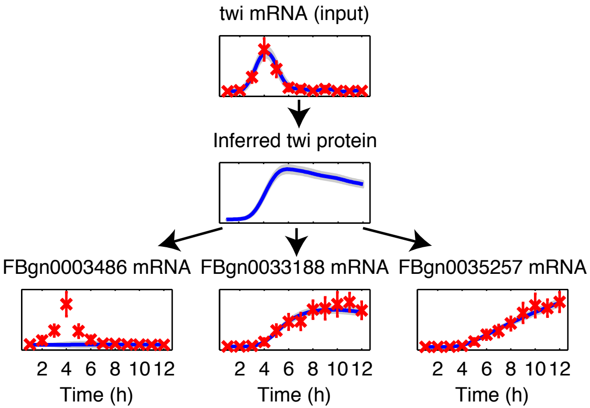
Cons:

- Concentrations should really be constrained positive
- Non-linear models require approximate inference methods

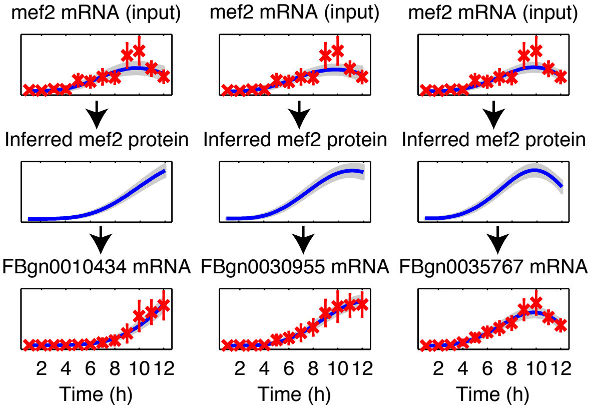
Fitting the model to data - Twist and Mef2



Fitting the model to data - Twist and Mef2

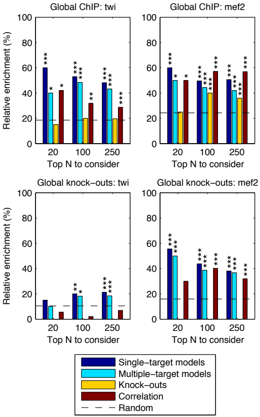


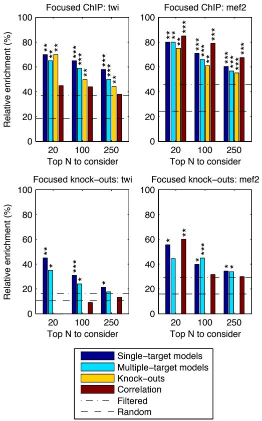
Fitting the model to data - Twist and Mef2



Ranking assessment

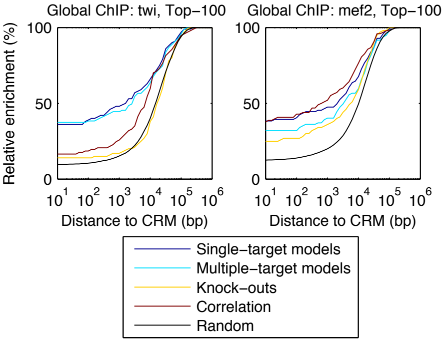
Evaluation of model-based ranking using ChIP and knock-out data





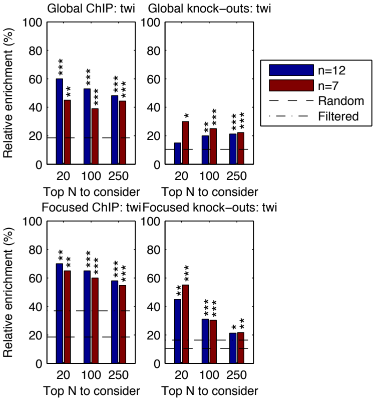
Ranking assessment

Changing the distance threshold between CRM and positive targets



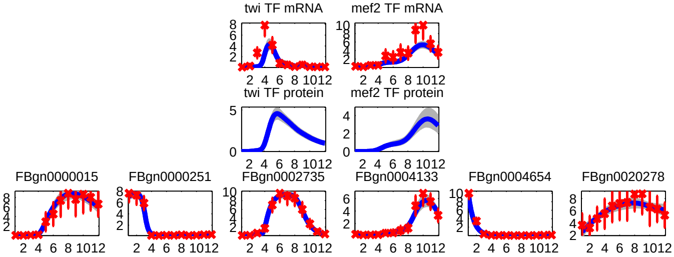
Ranking assessment

Changing the size of the dataset



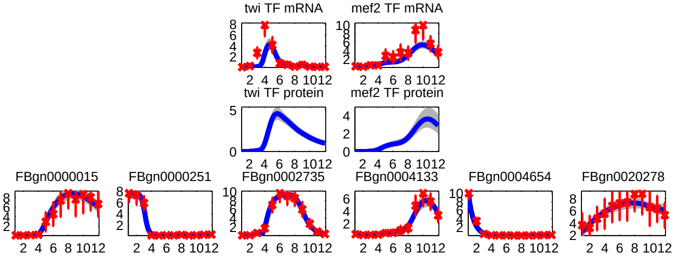
Modelling combinatorial regulation

Many target genes are regulated by multiple transcription factors



Modelling combinatorial regulation

Many target genes are regulated by multiple transcription factors



Network inference becomes much more challenging because the models become more complex and the space of models is huge.

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources - these may be noisy and incomplete

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources - these may be noisy and incomplete
- Model-based inference can make effective use of these data, even when models are highly simplified

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources - these may be noisy and incomplete
- Model-based inference can make effective use of these data, even when models are highly simplified
- Gaussian processes provide a useful way to close an open system without introducing too many additional parameters

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources - these may be noisy and incomplete
- Model-based inference can make effective use of these data, even when models are highly simplified
- Gaussian processes provide a useful way to close an open system without introducing too many additional parameters
- When models don't fit the data then we learn something

Conclusion

- Gene regulatory networks are fundamental to understanding many cellular processes
- Inference of gene regulatory networks requires input from multiple data sources - these may be noisy and incomplete
- Model-based inference can make effective use of these data, even when models are highly simplified
- Gaussian processes provide a useful way to close an open system without introducing too many additional parameters
- When models don't fit the data then we learn something – and then we need new models and new experiments

Acknowledgements

- Co-PIs: Neil Lawrence, Antti Honkela (HUT Finland)
- Experimental data: Eileen Furlong group (EMBL Heidelberg)

Funded by EPSRC award "Gaussian processes for systems identification with applications in Systems Biology", PASCAL2 Network of Excellence and the Academy of Finland

Advertisement