Transcription factor target identification with limited data using Gaussian process models

Antti Honkela
Neil D. Lawrence, Magnus Rattray and Michalis Titsias

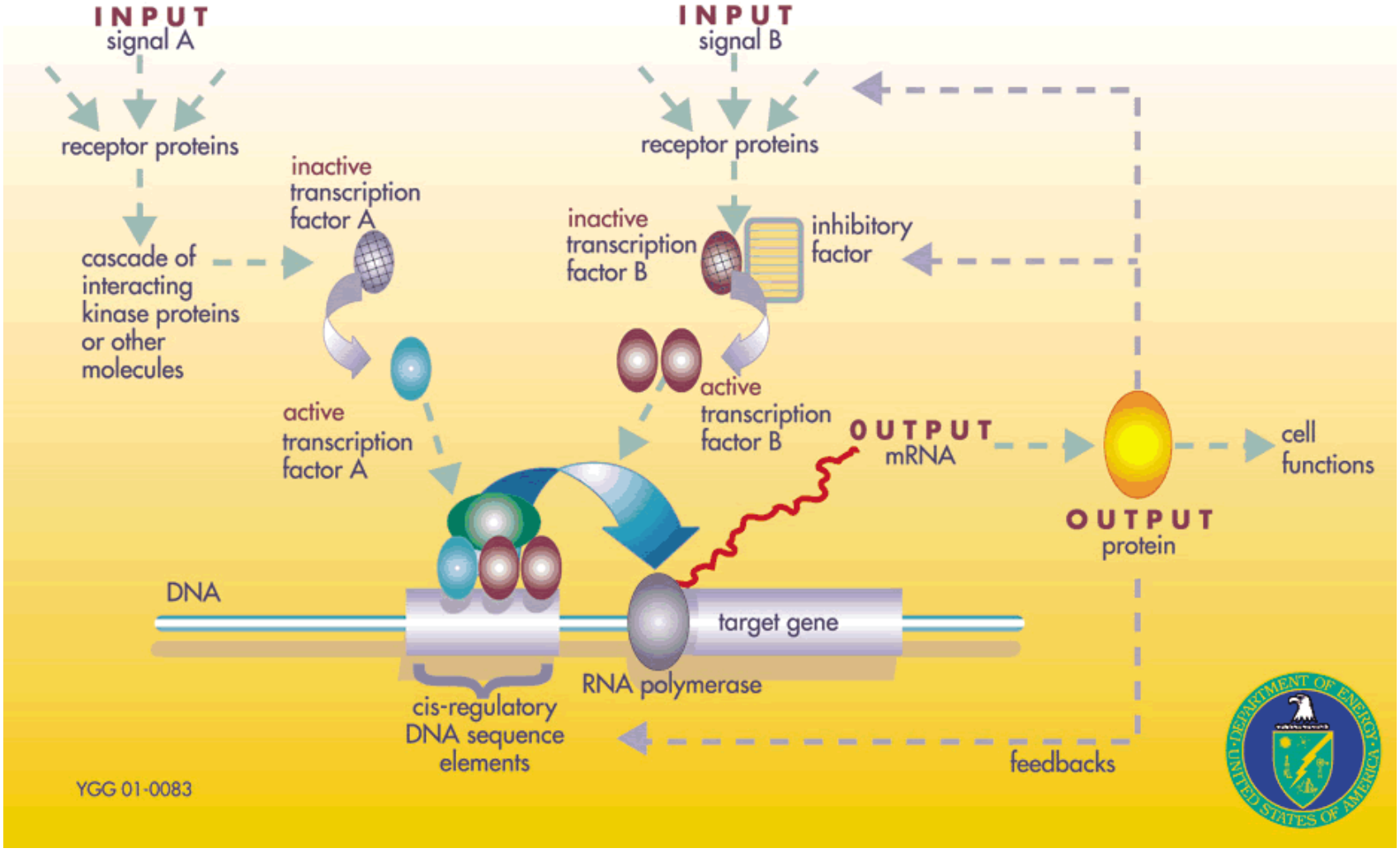
Aalto University School of Science and Technology
Department of Information and Computer Science

11 May 2010

Aalto University
School of Science
and Technology

The University of Manchester

A GENE REGULATORY NETWORK



The data

- High-throughput measurements of nucleic acids (esp. mRNA)
- Detecting proteins require more targeted techniques

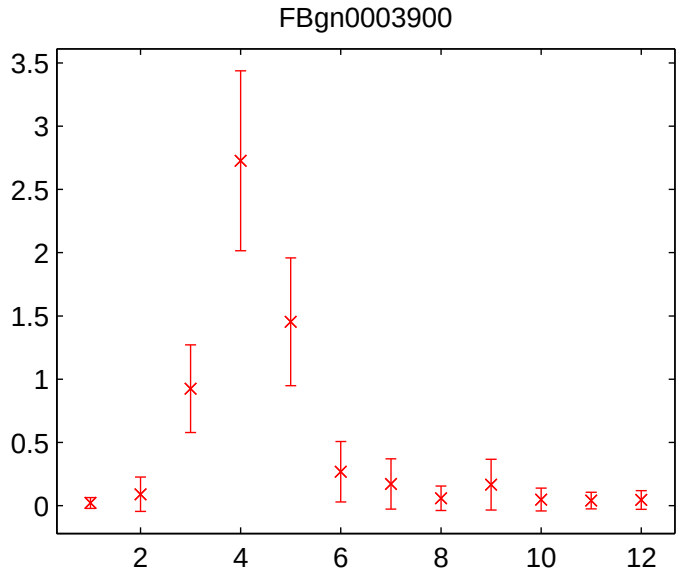


Figure: Sample expression time series

Our goal

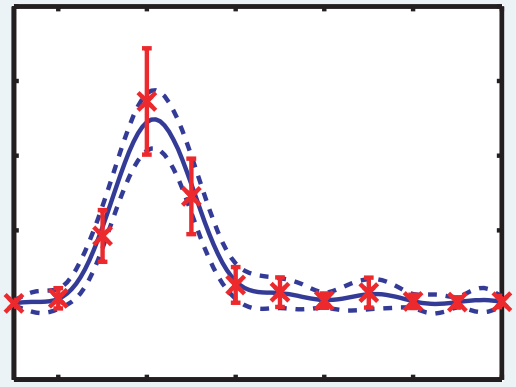
- To infer activities of unobserved chemical species through the effects they have on others
- Application: Infer local regulatory relationships, i.e. direct targets of transcription factors (TFs)
- Data: time series mRNA expression data
  (DNA (genes) $\xrightarrow{\text{transcription}}$ mRNA $\xrightarrow{\text{translation}}$ Protein)

Outline

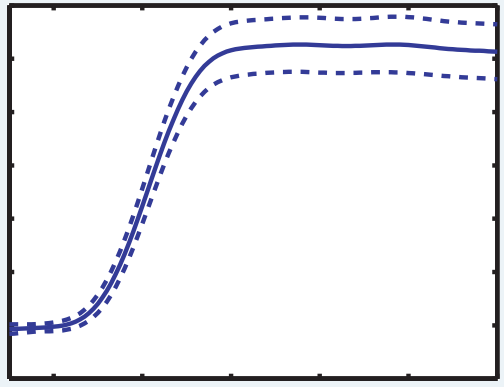
Outline

The ODE model

y(t) = TF mRNA profile



f(t) = protein activity profile

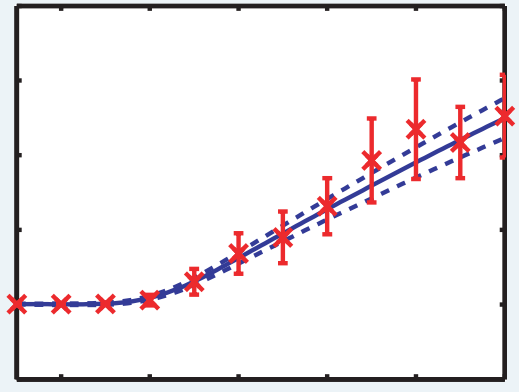


$$y(t) \sim \mathcal{GP}(m(t), k(t,t'))$$

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \sigma y(t) - \delta f(t)$$

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t)$$

x(t) = target expression profile



(Observations are denoted in red, inferred functions in blue.)

Gaussian process ODE modelling

- Use Gaussian process priors on activity time courses
  - Functional prior, specified by mean and covariance functions
  - No need for time discretisation
- Two alternatives for "bootstrapping" the TF protein activity
  - Training set of known targets (cf. **?**, Genome Biology)
  - Hierarchical model with TF translation from measured mRNA

Gaussian Processes

- Gaussian Process

$$f(t) \backsim \mathcal{GP}\left(m(t), k\left(t, t'\right)\right)$$

where

$$\begin{aligned} m(t) &= \mathbb{E}[f(t)] = \langle f(t) \rangle \\ k\left(t, t'\right) &= \mathbb{E}\left[\left(f(t) - m(t)\right)\left(f\left(t'\right) - m\left(t'\right)\right)\right] \end{aligned}$$

The joint covariance

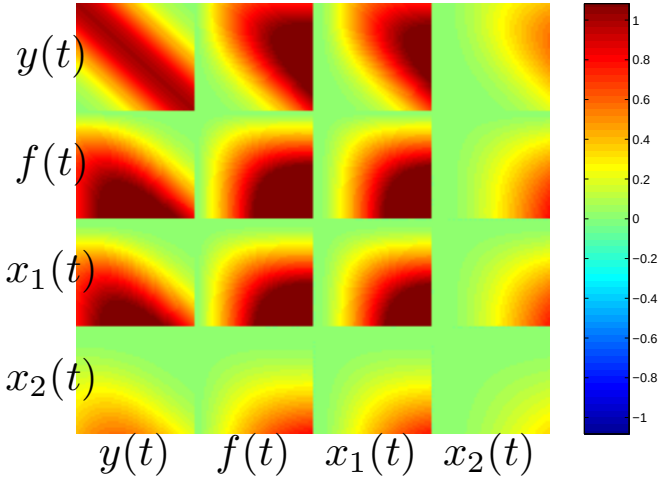
RBF covariance function for $y(t)$

$$f(t) = \sigma \exp(-\delta t) \int_0^t y(u) \exp(\delta u) \, \mathrm{d}u$$

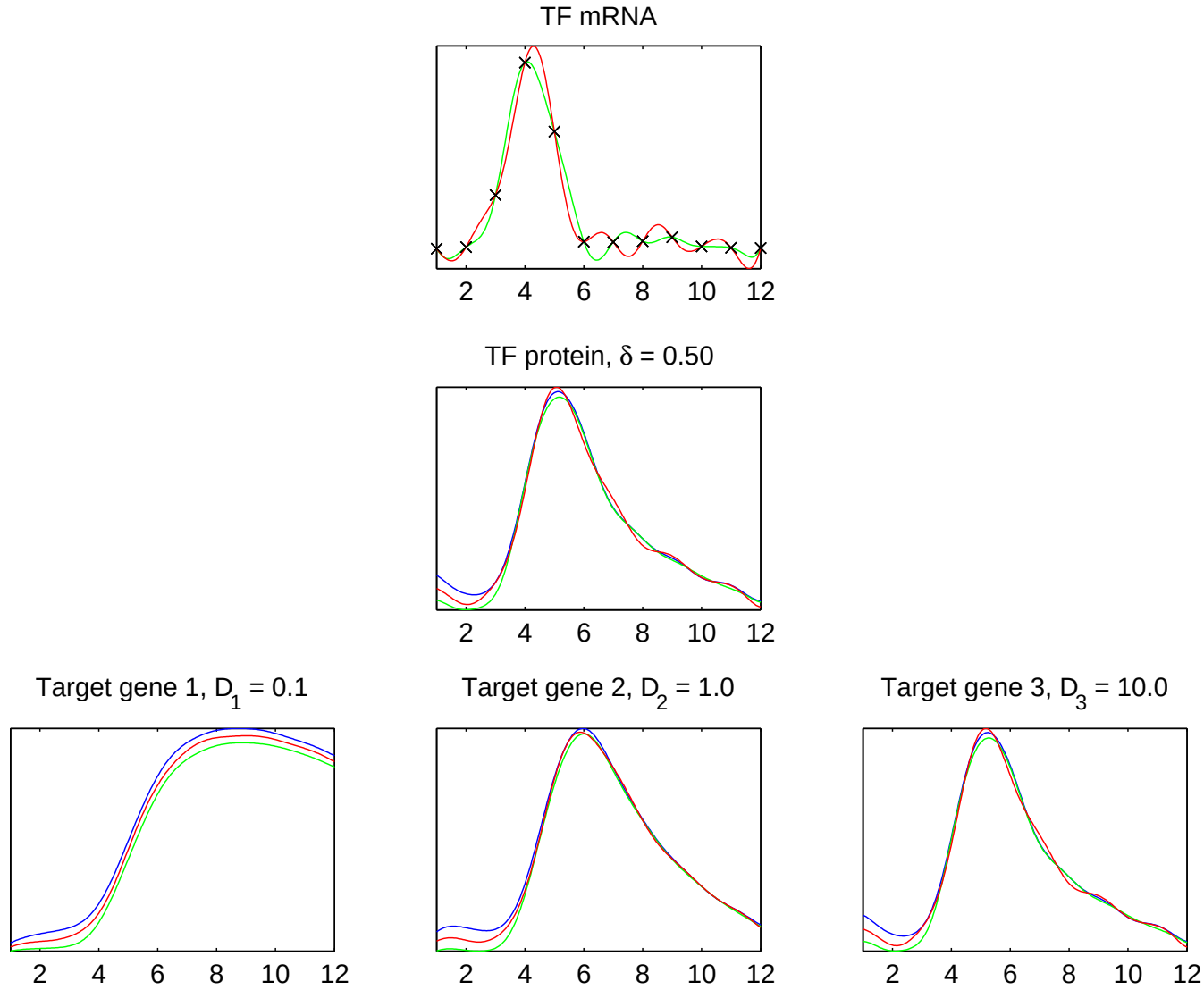
$$x_i(t) = \frac{B_i}{D_i} + S_i \exp(-D_i t) \int_0^t f(u) \exp(D_i u) \, \mathrm{d}u.$$

- Joint distribution for $x_1(t)$, $x_2(t)$, $f(t)$ and $y(t)$.
- Here:

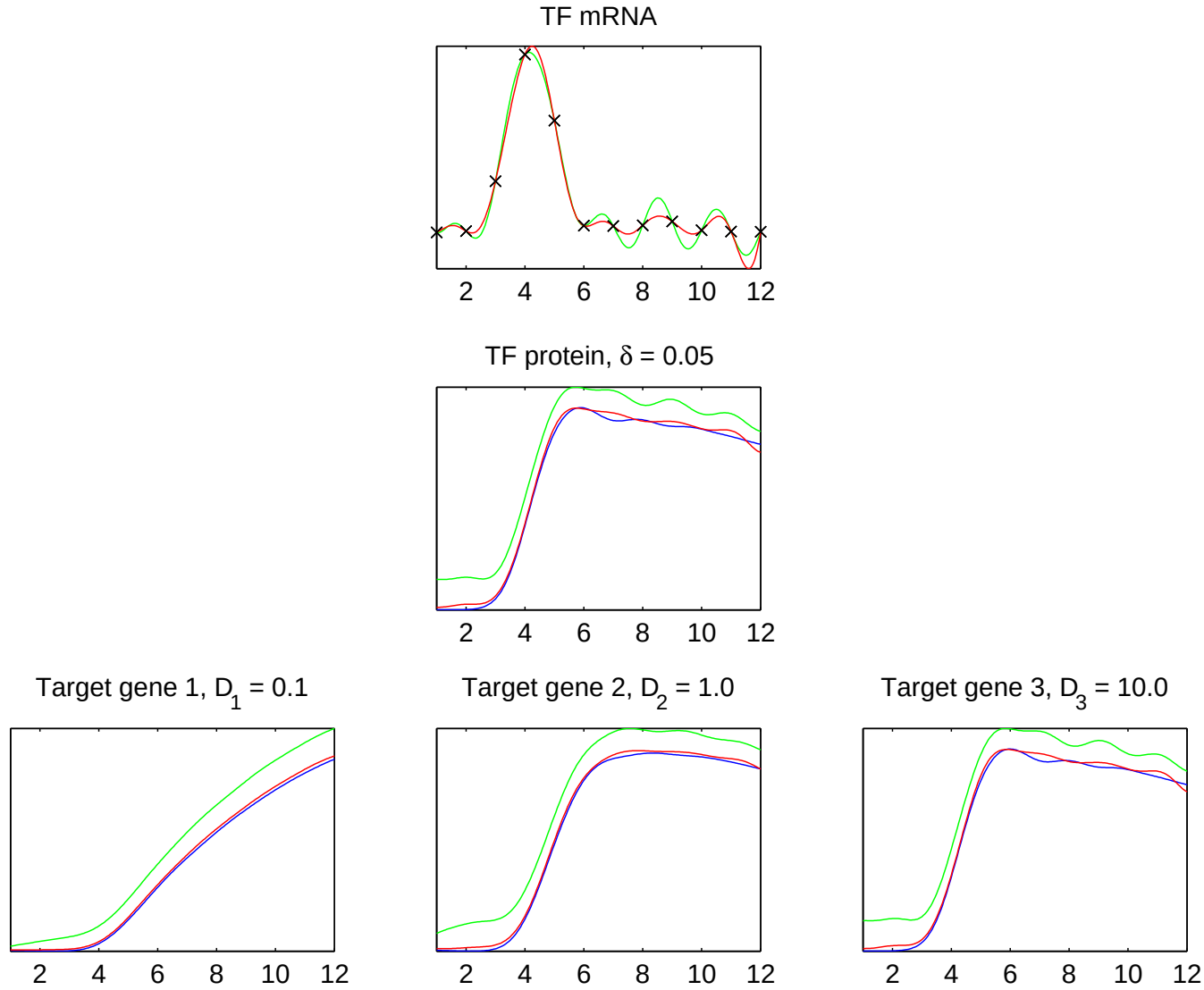
δ	D_1	S_1	D_2	S_2
0.1	5	5	0.5	0.5



Covariance samples



Covariance samples



Outline

Target ranking: motivation

- Finding target genes of TFs is an obvious first step in reverse engineering gene regulatory networks
- Typical techniques
  - Measure TF binding locations in the genome using ChIP-chip/-seq
  - Observe gene expression in mutants with the TF disabled (knockouts) or overexpressed
- Endless potential applications in understanding disease and other biological phenomena

Case study: Mesoderm and muscle development in Drosophila

- Focus on two TFs regulating mesoderm and muscle development in fruit fly Drosophila: Twist and Mef2
- Assume no post-translational regulation of these TFs
- Expression data: 12 time points at 1 h intervals, 3 repeats
- Data averaged over the whole embryo

Application of the models to target ranking

- First apply the two-layer model for each target gene independently
- Ranking by model likelihood

Application of the models to target ranking

- First apply the two-layer model for each target gene independently
- Ranking by model likelihood

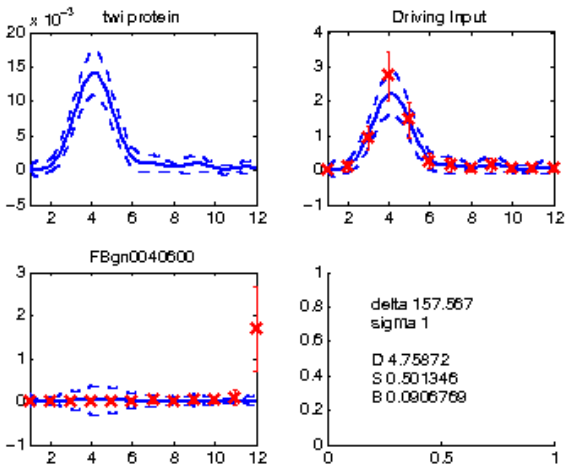


Figure: Fitted two-layer (GPDISIM) models

Application of the models to target ranking

- First apply the two-layer model for each target gene independently
- Ranking by model likelihood

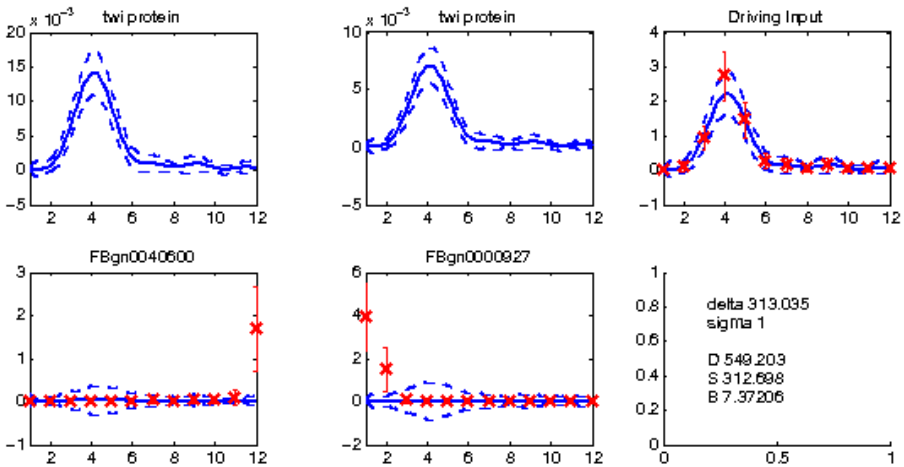


Figure: Fitted two-layer (GPDISIM) models

Application of the models to target ranking

- First apply the two-layer model for each target gene independently
- Ranking by model likelihood

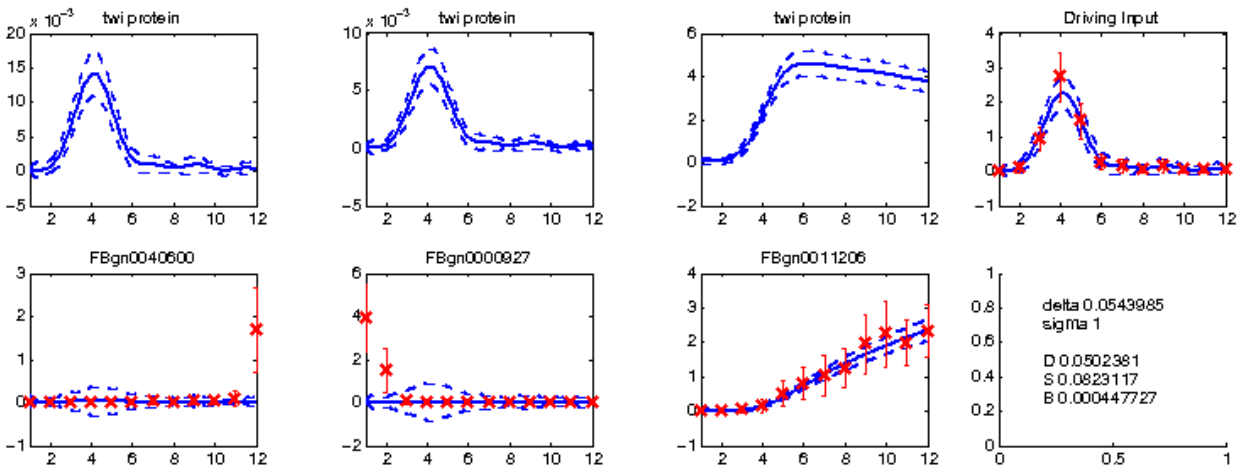


Figure: Fitted two-layer (GPDISIM) models

Application of the models to target ranking

- First apply the two-layer model for each target gene independently
- Ranking by model likelihood

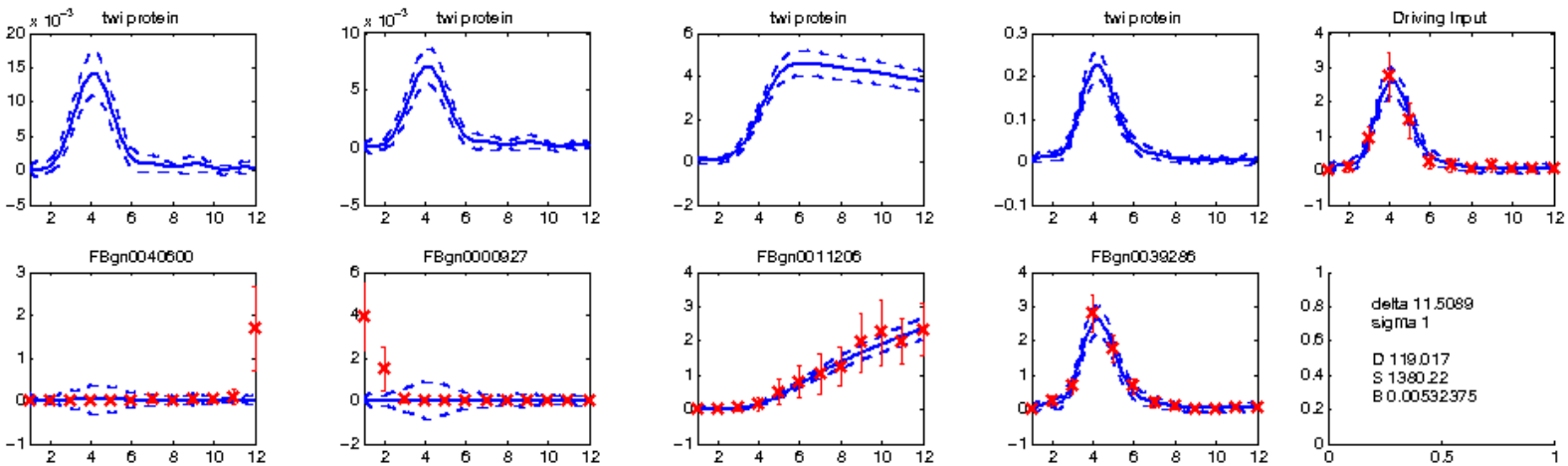


Figure: Fitted two-layer (GPDISIM) models

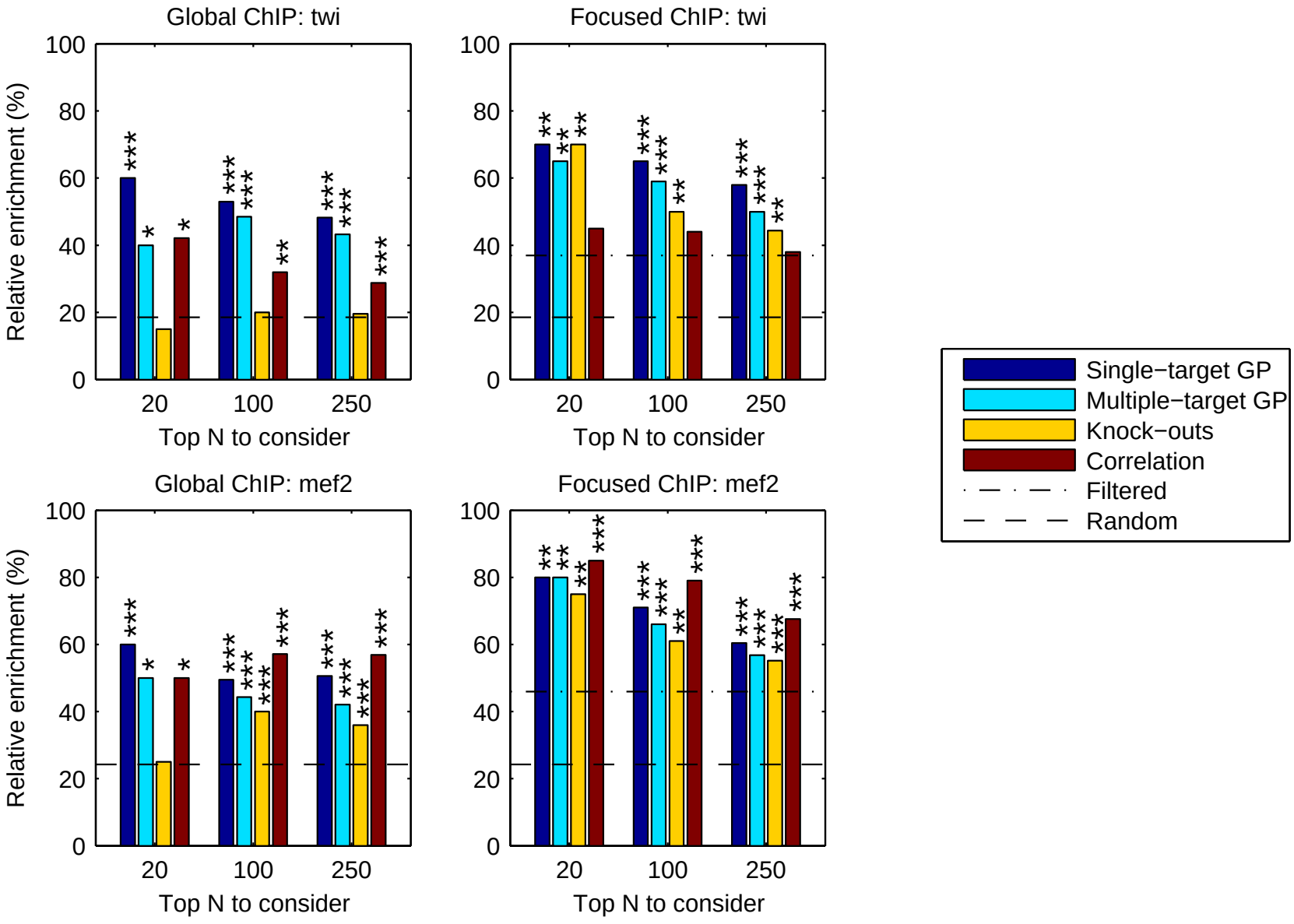
Application of the models to target ranking

- Need to exclude inactive genes
  - Threshold the average z-score of gene activity
  - Threshold the likelihood ratio against a model with $S = 0$
- Using a set of identified likely targets as a training set, learn multiple-target models for training set + each target individually

Evaluation methods

- Evaluate the ranking methods by taking a number of top-ranked targets and record the number of "positives" (**?**):
  - targets with ChIP-chip binding sites within 2 kb of gene
  - (targets differentially expressed in TF knock-outs)
- Compare against
  - Ranking by correlation of expression profiles
  - Ranking by q-value of differential expression in knock-outs
- Optionally focus on genes with annotated expression in tissues of interest

Results



'***': $p < 0.001$, '**': $p < 0.01$, '*': $p < 0.05$

Single-TF Target Ranking: Summary

- The two-layer translation/transcription model provides impressive target ranking results
- Works with very short time series, even 6-7 time points
- More details in the paper:

A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. M. Furlong, N. D. Lawrence, and M. Rattray.
Model-based method for transcription factor target identification with limited data.
Proc Natl Acad Sci U S A, 107(17):7793–7798, Apr 2010.
doi:10.1073/pnas.0914285107

Outline

Extending the model

- The linear single-TF model is about as far as exact inference takes us
- More complicated models require approximate techniques (e.g. MCMC)
- With MCMC there are no restrictions on the functional forms of models used

The full model

- Consider the ODE transcription regulation model for multiple TFs

$$\frac{dx_j(t)}{dt} = B_j + S_j g(f_1(t), \ldots, f_l(t); \mathbf{w}_j) - D_j x_j(t)$$

- $g(\cdot)$ a positive sigmoidal activation function
- $\mathbf{w}_j$ interaction weights between the jth gene and the set of l TFs

Gene regulation with multiple TFs

$$\frac{dx_j(t)}{dt} = B_j + S_j g(f_1(t), \ldots, f_I(t); \mathbf{w}_j) - D_j x_j(t),$$

- $g(\cdot)$ is assumed to be a multiple-TF hill function:

$$g(f_1(t), \ldots, f_I(t); \mathbf{w}_j) = \frac{\prod_{i=1}^{I} f_i(t)^{w_{ji}}}{\gamma_j^{\sum_{i=1}^{I} w_{ji}} + \prod_{i=1}^{I} f_i(t)^{w_{ji}}}$$

  where w_{ji} can be both positive and negative

- The above can also be written as the sigmoid function:

$$g(f_1(t), \ldots, f_I(t); \mathbf{w}_j) = \frac{1}{1 + e^{-w_{j0} - \sum_{i=1}^{I} w_{ji} \log f_i(t)}}$$

  where we defined $w_{j0} = -\sum_{i=1}^{I} w_{ji} \log \gamma_j$ as new parameter

Bayesian inference from mRNA data: priors

$$x_j(t) = \frac{B_j}{D_j} + \left(A_j - \frac{B_j}{D_j} \right) e^{-D_j t} + S_j \int_0^t g(f_1(u), \ldots, f_l(u); \mathbf{w}_j) e^{-D_j(t-u)} du$$

$$f_i(t) = \int_0^t y_i(t) e^{-d_i(t-u)} du$$

- We place priors on:
  - Kinetics: $\Theta = \{A_j, B_j, D_j, S_j\}_{j=1}^N$ (uniform or log normal)
  - Decays of TF mRNA: $\{d_i\}$, (uniform or log normal)
  - Interaction weights: $\{\mathbf{w}_j\}$, (Gaussian priors with optionally positivity constraints and/or spike and slab sparse priors)
  - mRNA functions $y_i(t)$: Gaussian processes (through a transformation that ensures positivity of $y_i(t)$)
  - Lengthscales of Gaussian processes (uniform or gamma) and noise variances in the likelihoods (gamma)

Bayesian inference from mRNA data: MCMC (Michalis)

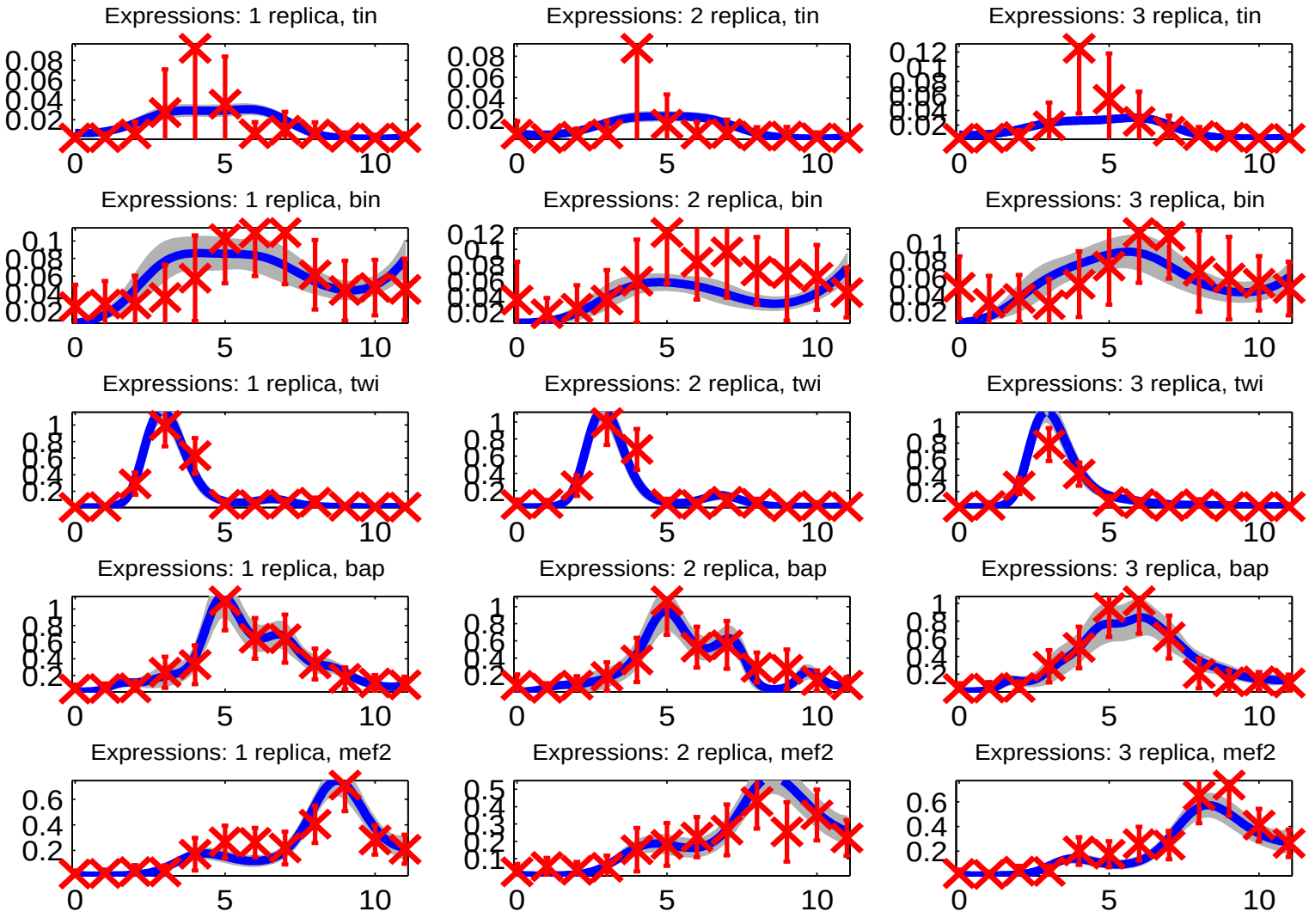
$$\text{joint} = p(\widetilde{X}|X)p(\widetilde{Y}|Y)p(\bar{Y}_i)p(\Theta)p(W)p(\{d_i\}_{i=1}^{I})p(\{\sigma_j^2\})p(\{\ell^2\}_{i=1}^{I})$$

- Many Metropolis-Hastings steps involving sampling Gaussian process functions, kinetic parameters, interaction weights, etc.
- We can afford training the model using MCMC in moderate-sized networks, e.g. with 100 genes and 5 TFs and 3 replicas, but not genome-wide (too slow for that)
- But once the model in trained, we can do genome-wide prediction (e.g. gene target identification) and this is fast

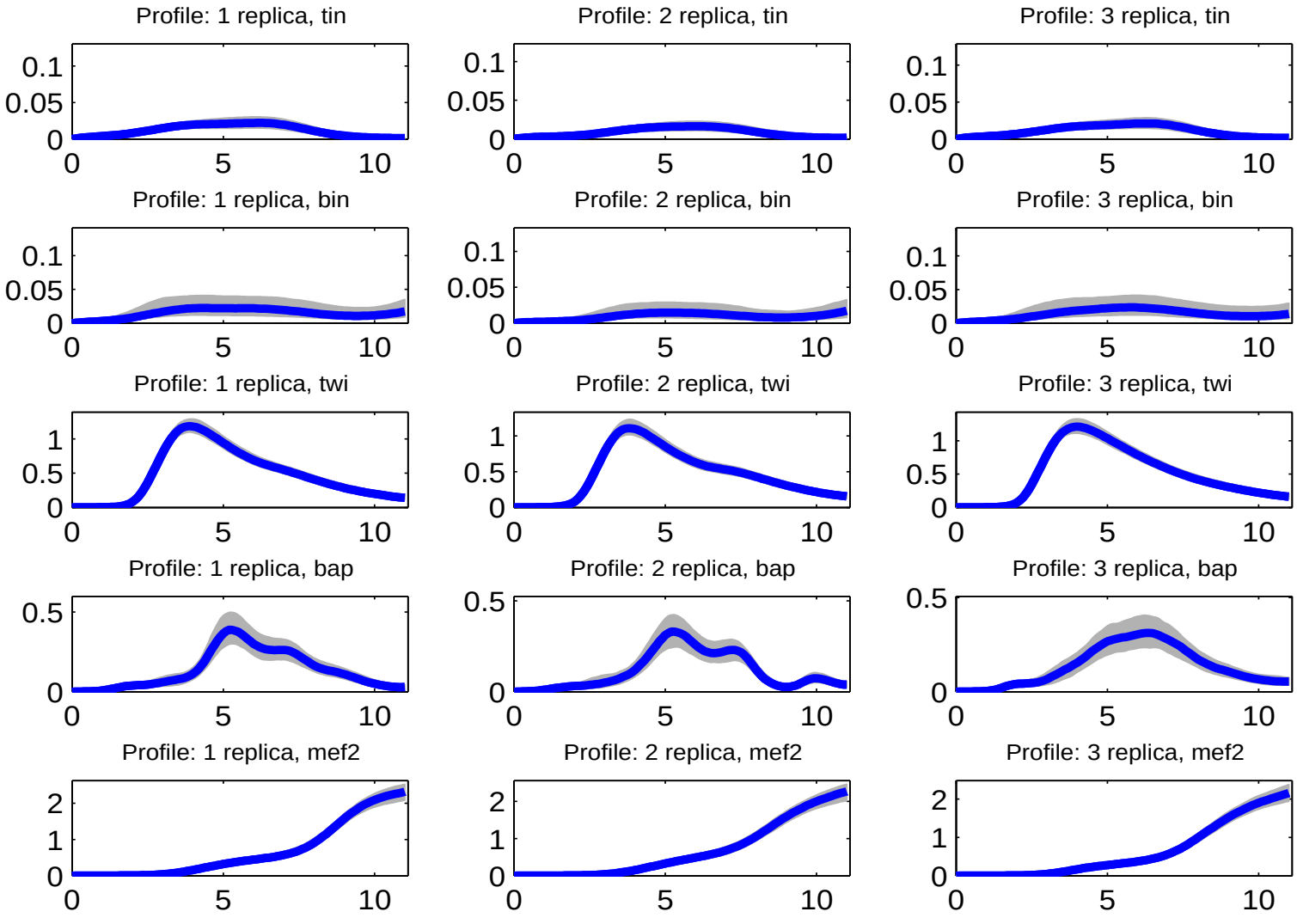
Experiments in Drosophila data

- Microarray dataset containing three replicas of 12 time points collected hourly throughout Drosophila embryogenesis in wild-type embryos
- 92 target genes
- 5 TFs (including Twist and Mef2 which regulate mesoderm and muscle development)
- ChIP information is used to define the (deterministic) sparse prior one interaction between TFs and target genes (**?**)

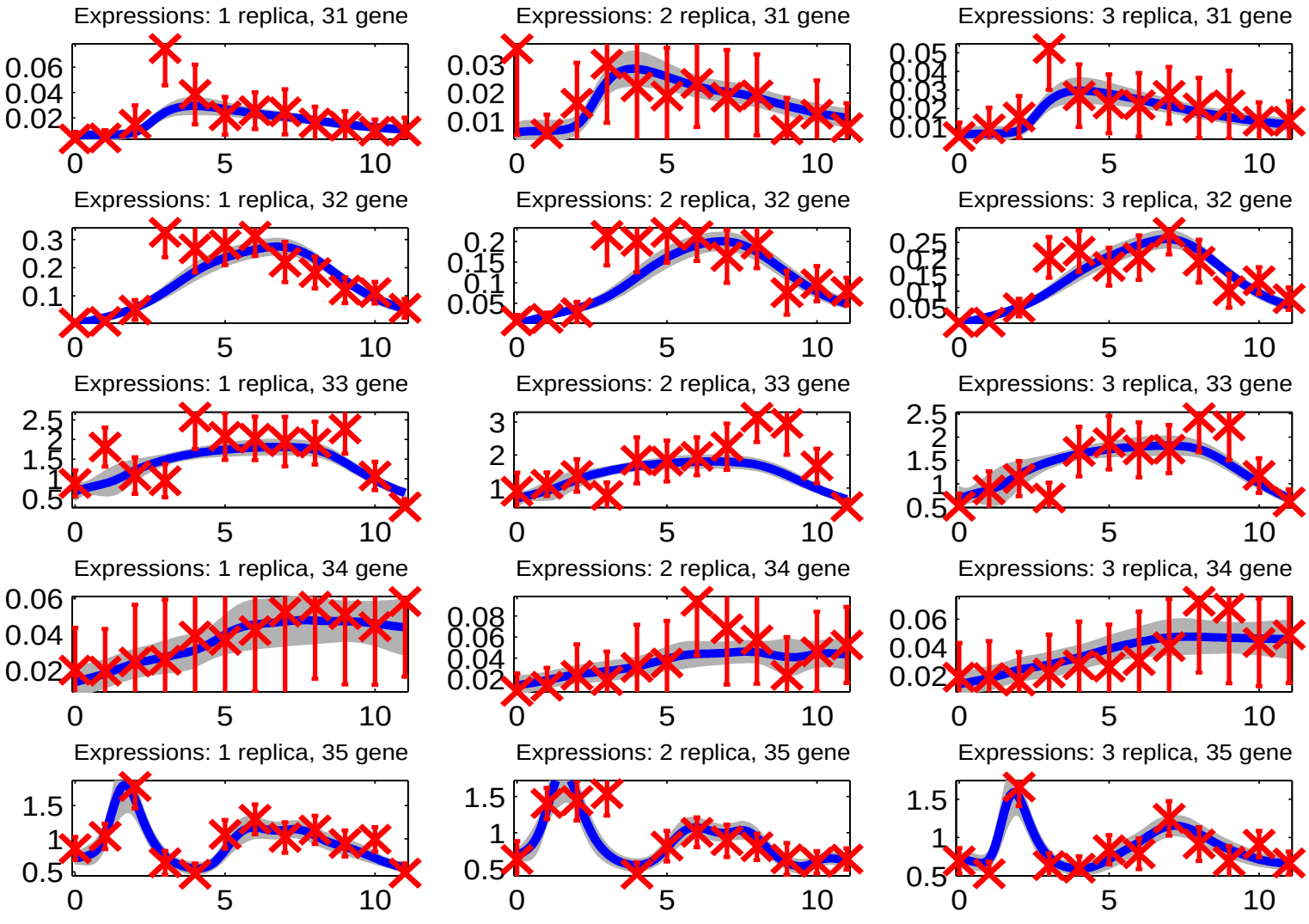
Experiments in Drosophila data



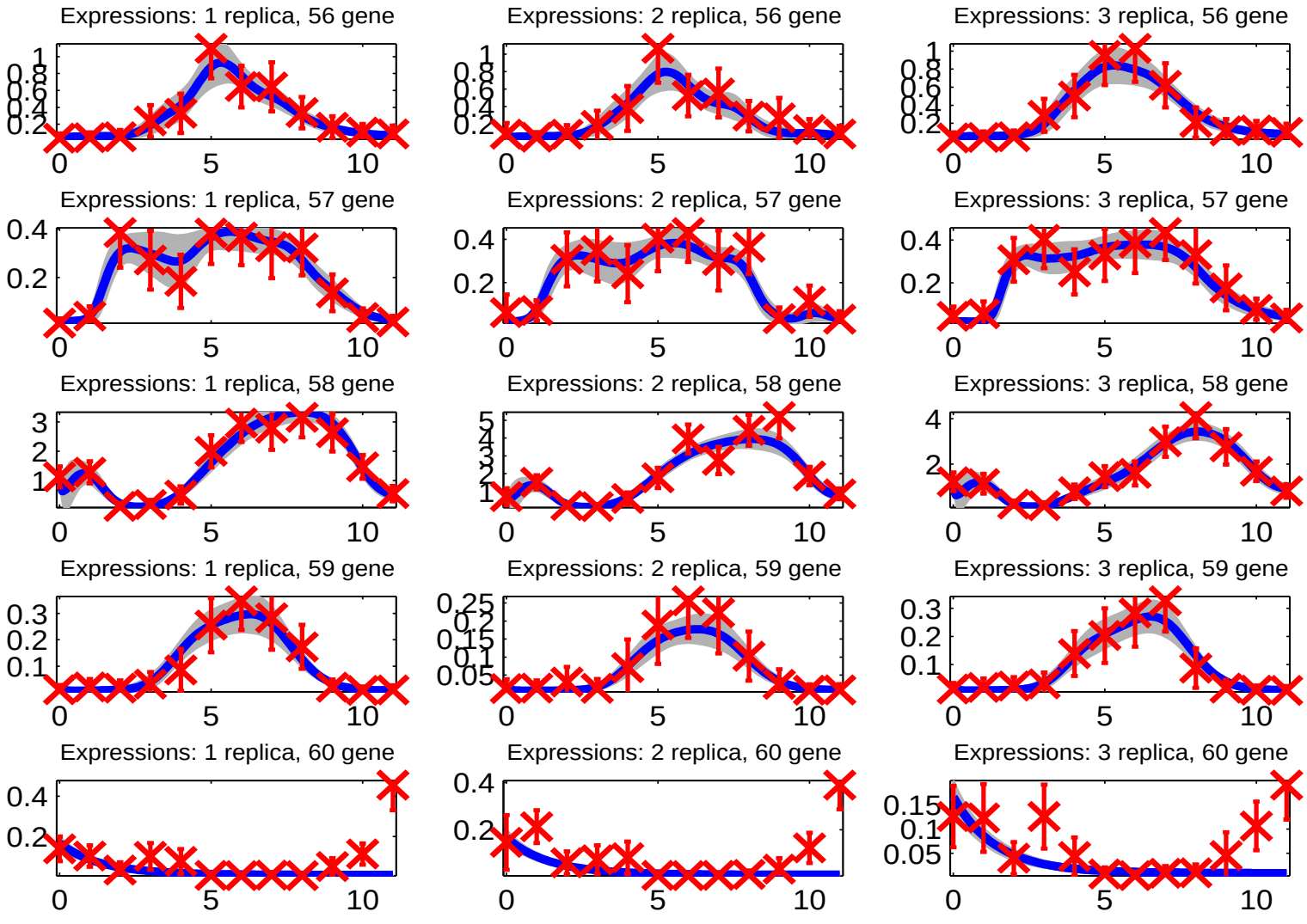
Experiments in Drosophila data



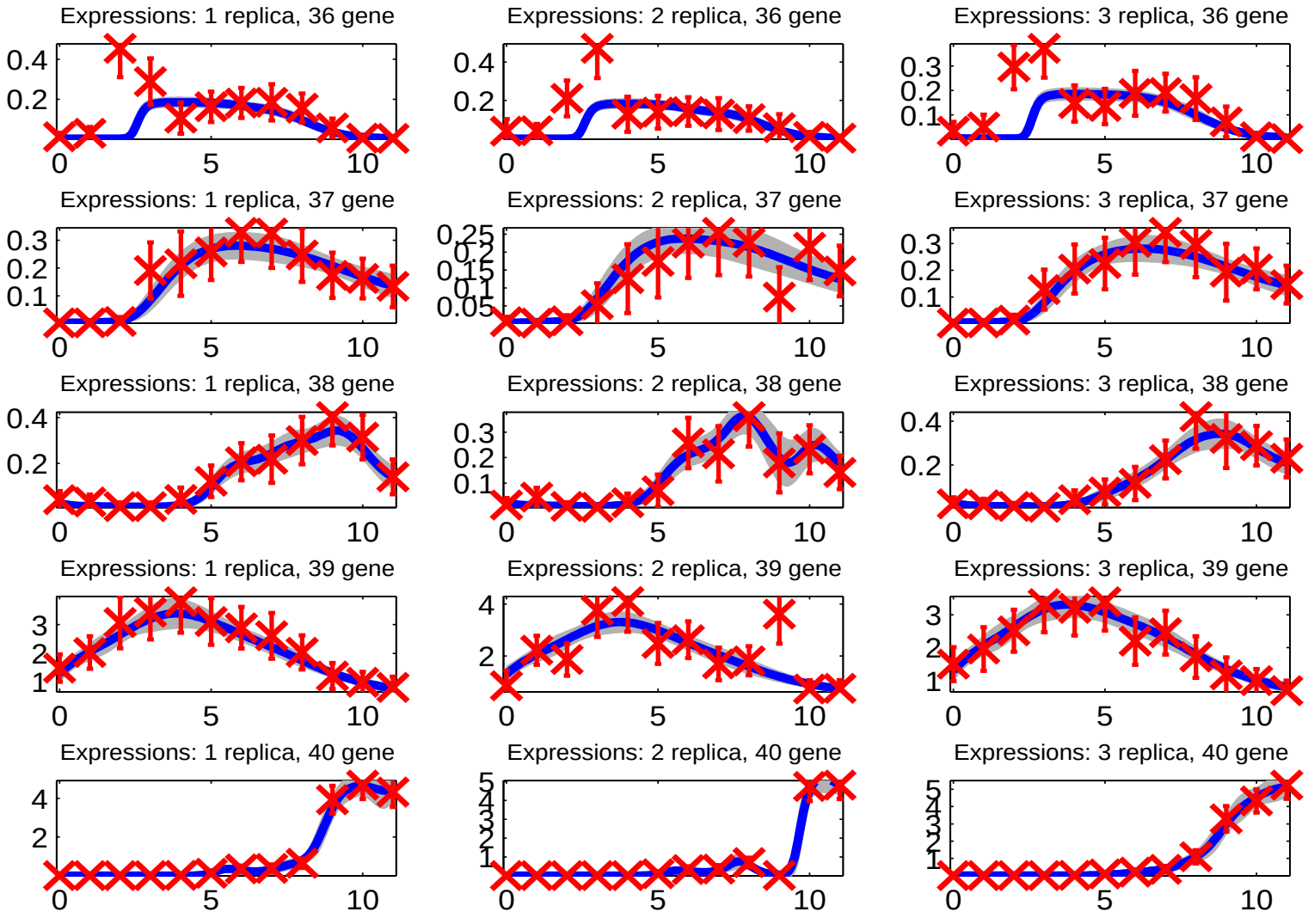
Experiments in Drosophila data



Experiments in Drosophila data



Experiments in Drosophila data



Genome-wide gene ranking/identification

- The trained model gives the posterior distribution of TF profiles
- It can be used to make (probabilistic) statements about if a certain TF combination regulates a test gene $*$?
  - Infer its interaction weights with a suitable prior
  - Compare models restricting a set of interaction weights to zero

Genome-wide gene ranking/identification

- Model comparison is based on **?** estimate of marginal likelihood
- Fast sampling, < 1 min per gene per model
  - All functions are drawn from posterior samples
  - Relatively few parameters to sample

Multiple-TF Models: Summary

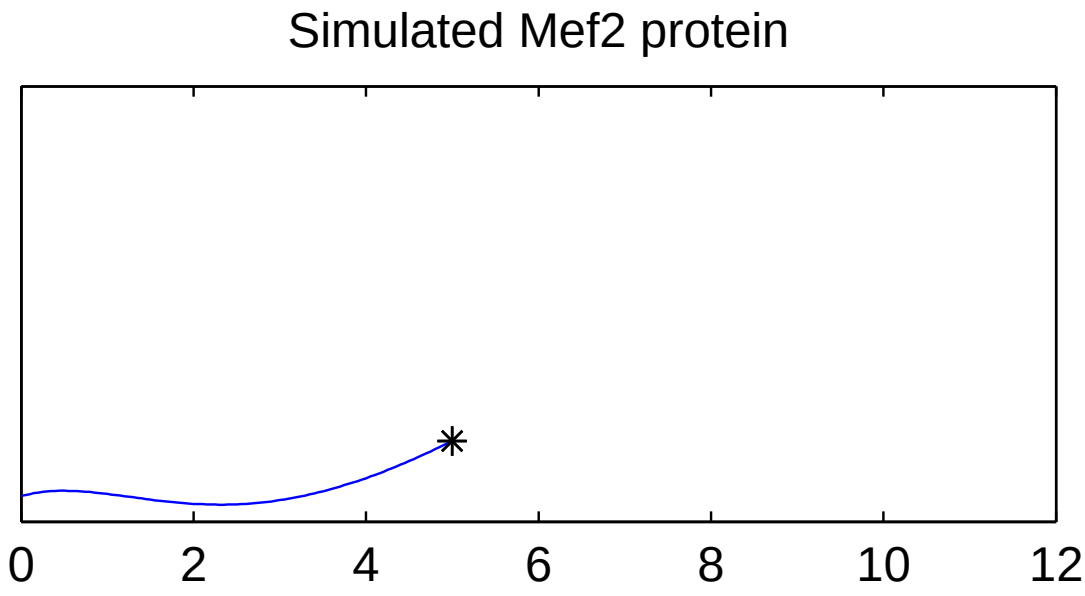
- Realistic models of combinatorial regulation
- MCMC techniques applicable to genome-wide screenings
- Amount of available data clearly a limiting factor in identifying the models

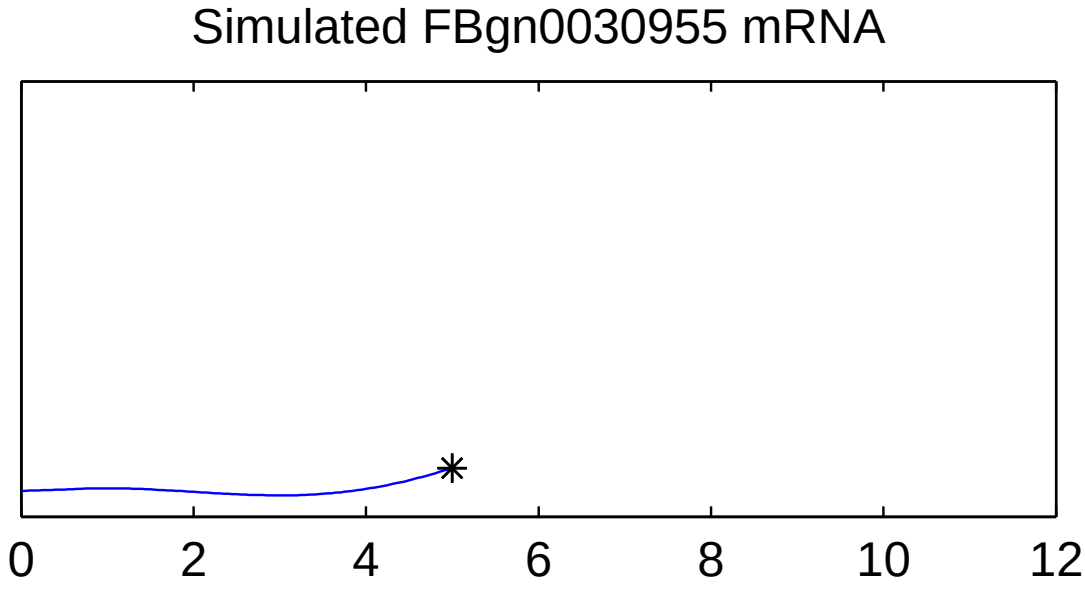
Outline

Molecular biology time series

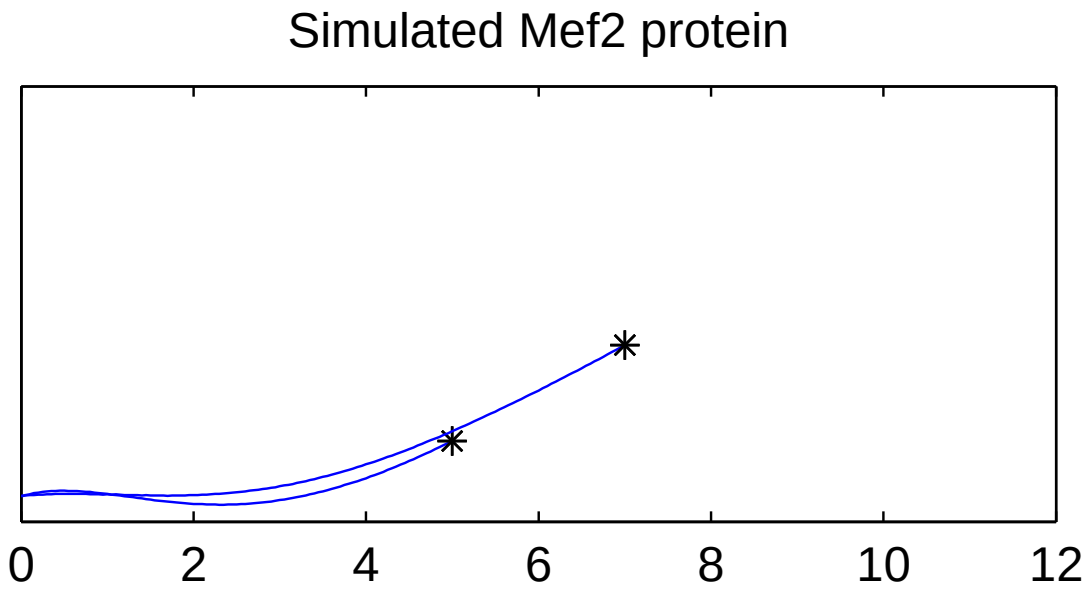
- Biological systems are dynamic, observing their time evolution very helpful
- Time series measurements of gene expression, protein activity, protein binding, ...
- Problem: most of these assays are highly disruptive to the sample
- Therefore: time series = series of independent experiments run for different lengths of time
- This has implications for modelling...

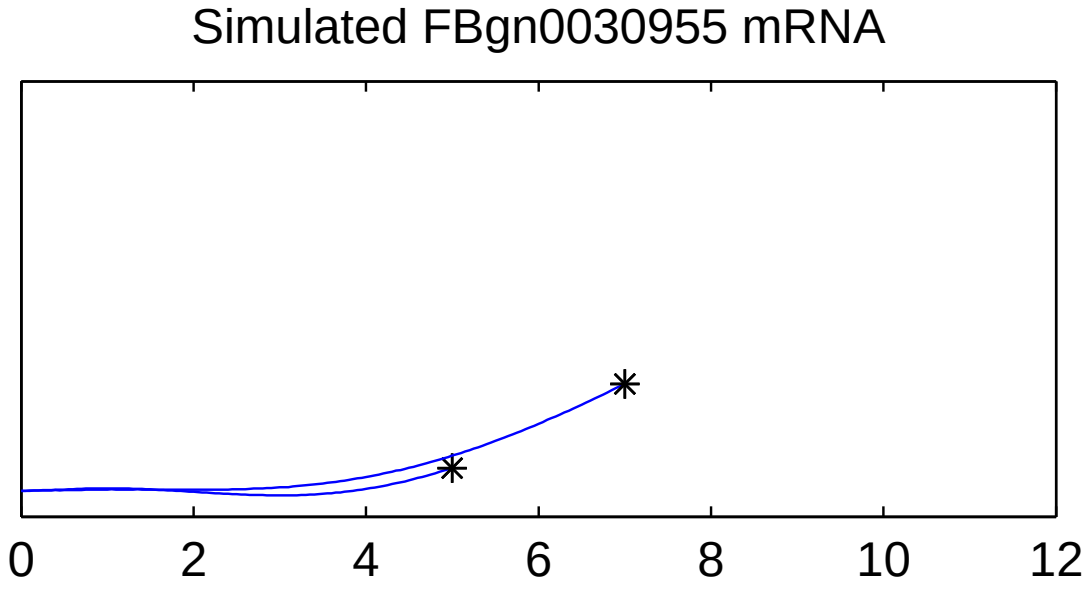
Simulated molecular biology time series



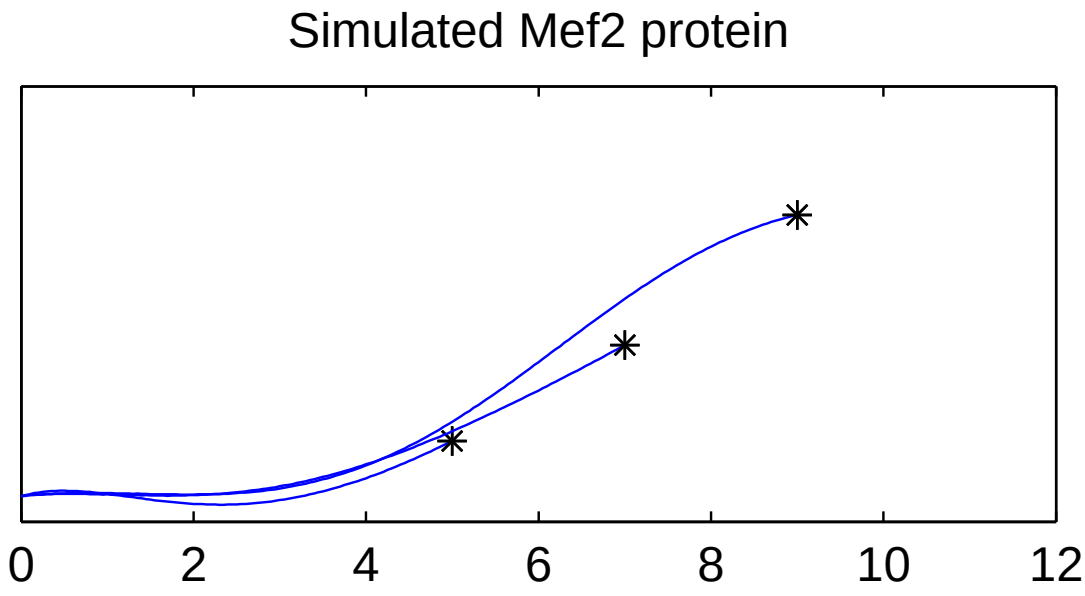


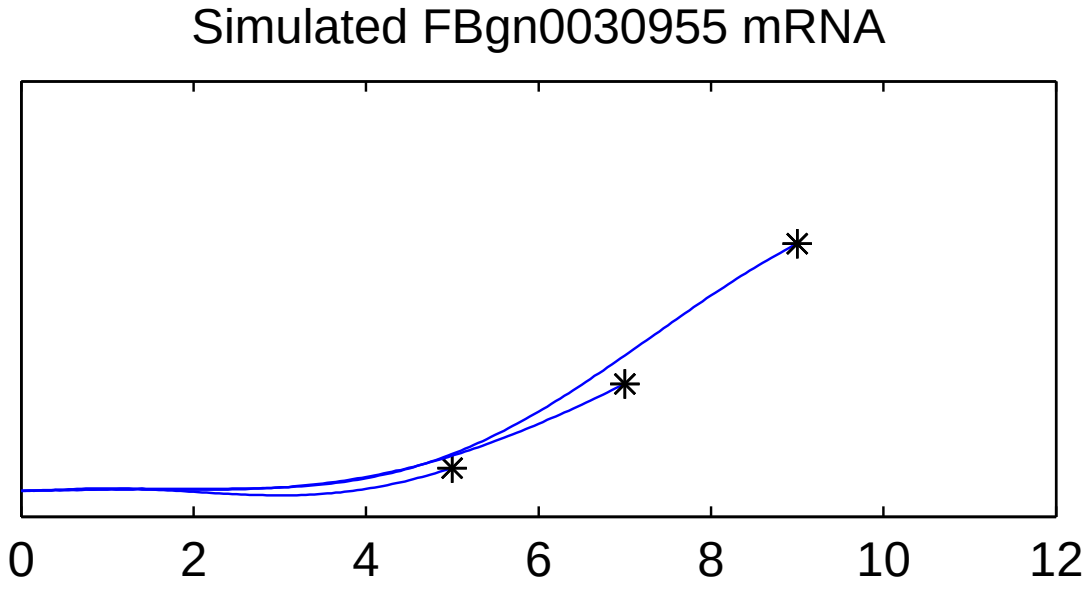
Simulated molecular biology time series



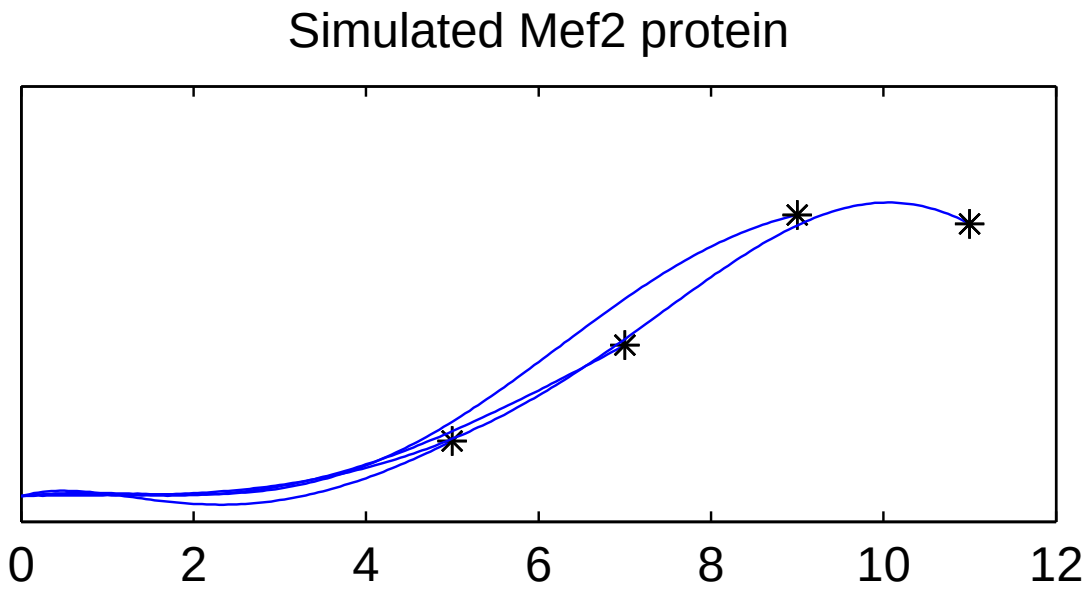


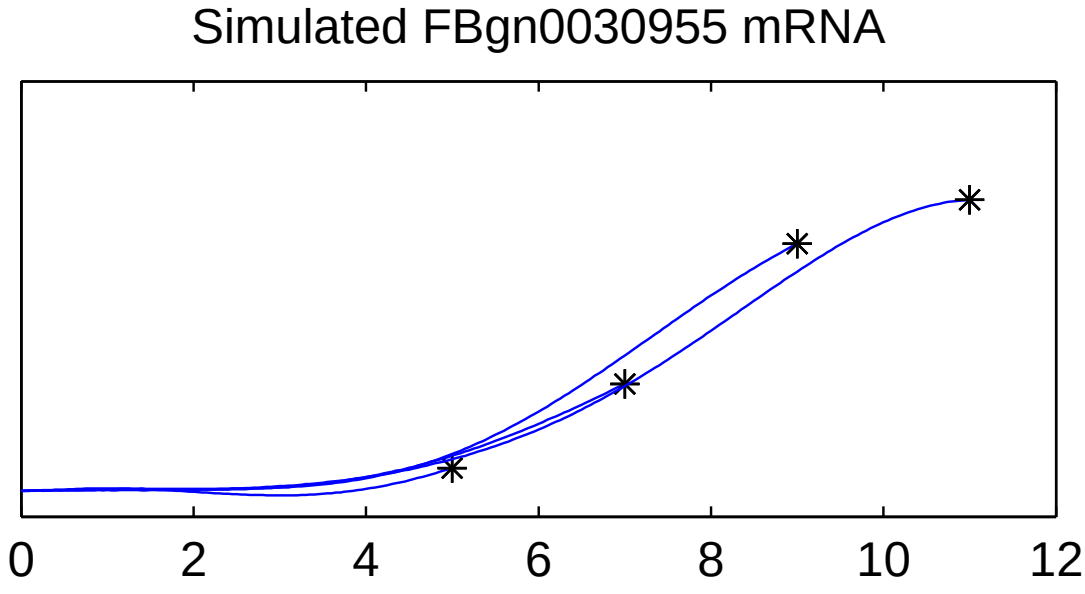
Simulated molecular biology time series



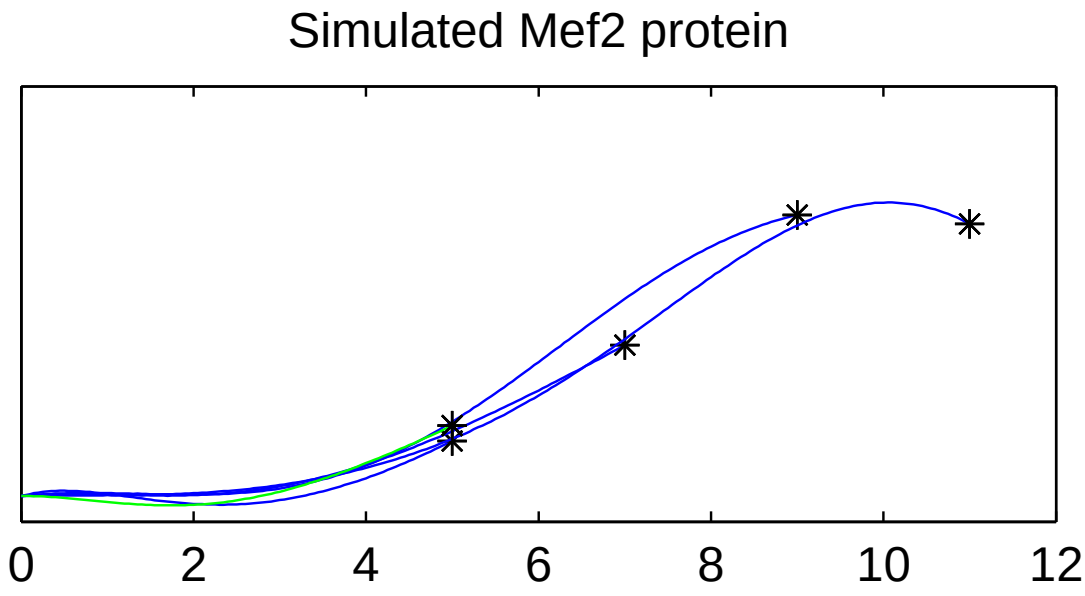


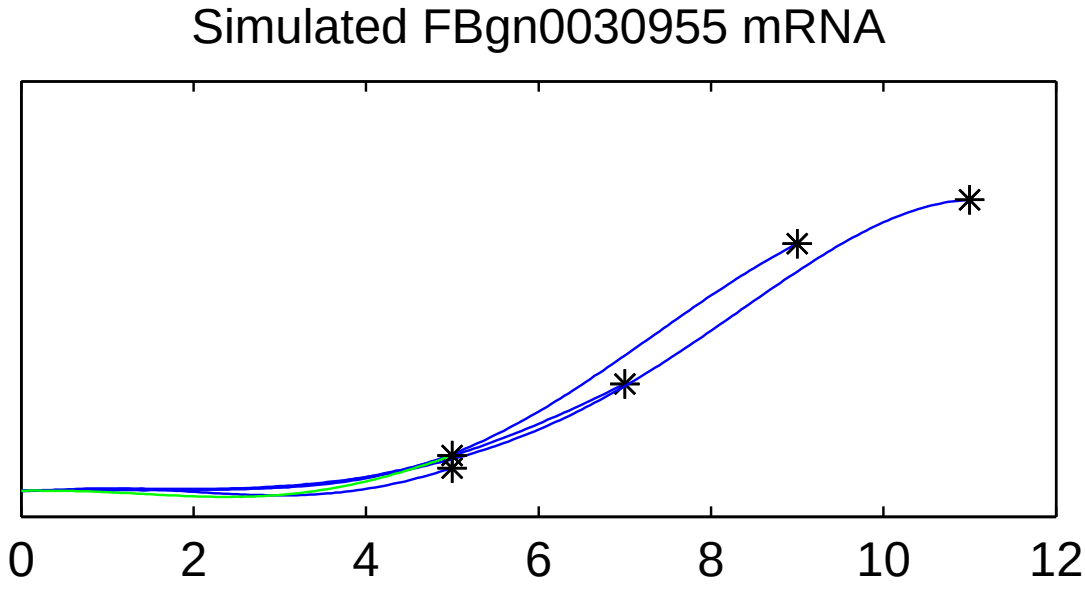
Simulated molecular biology time series



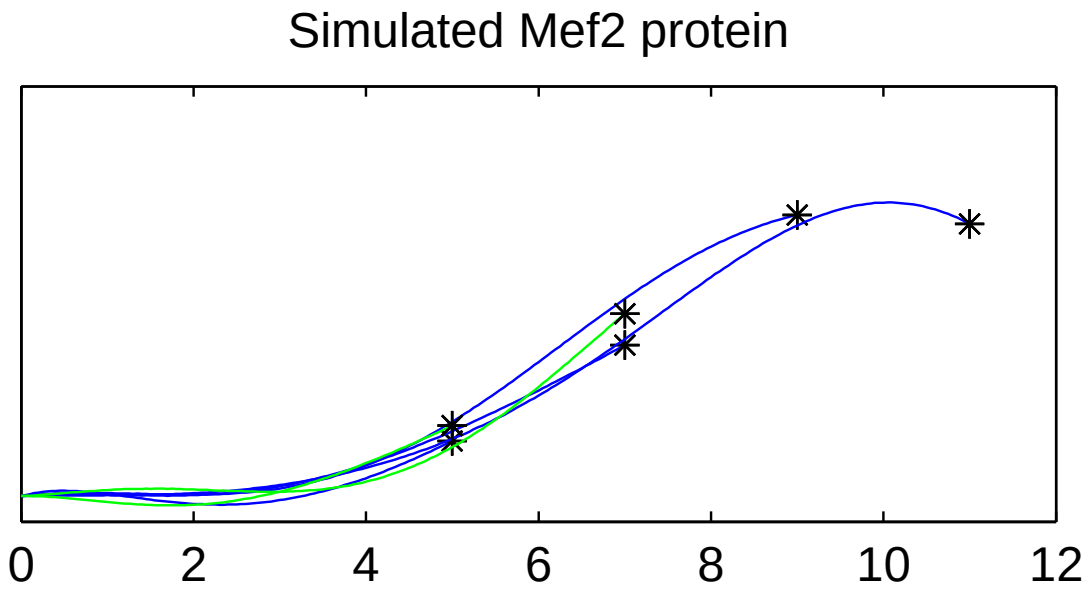


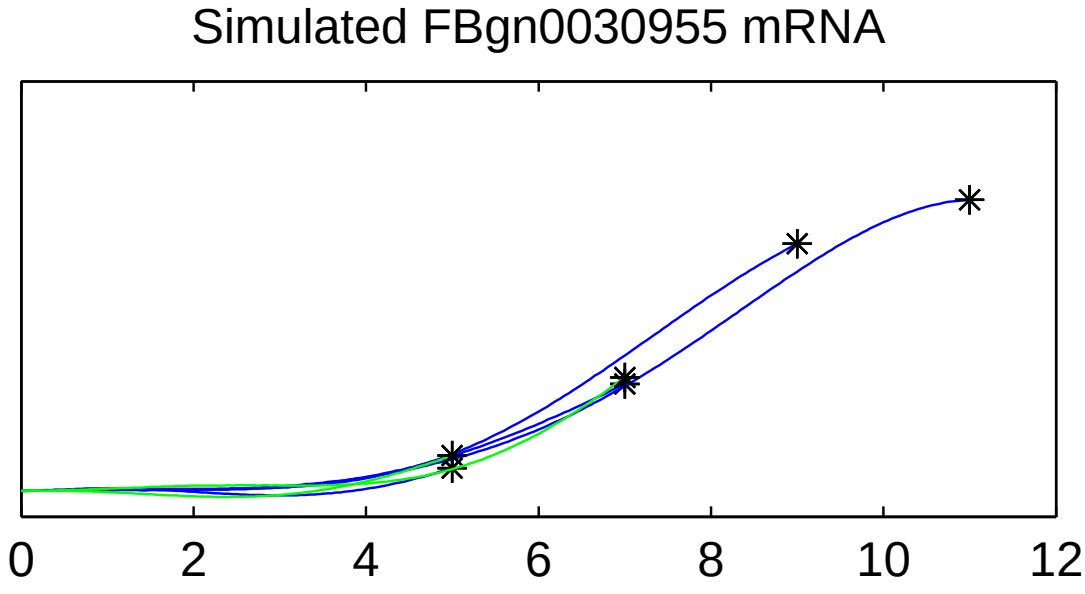
Simulated molecular biology time series



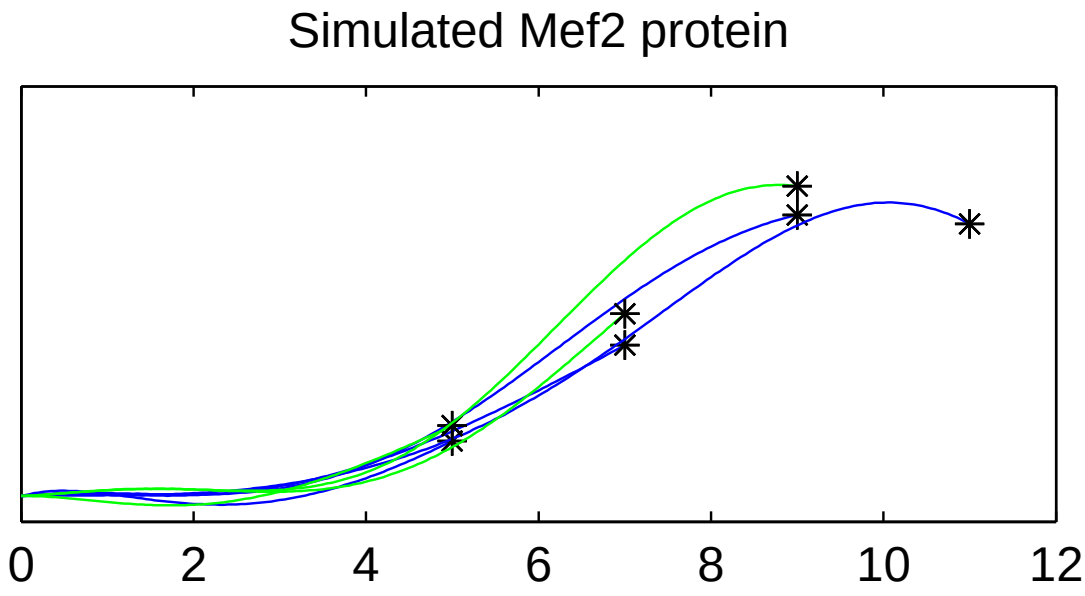


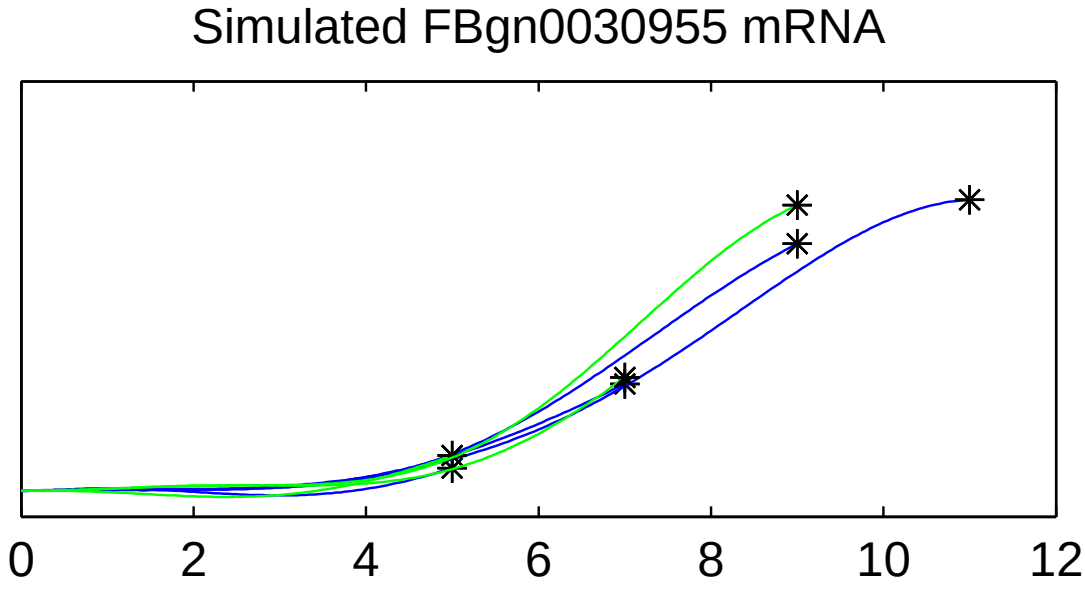
Simulated molecular biology time series



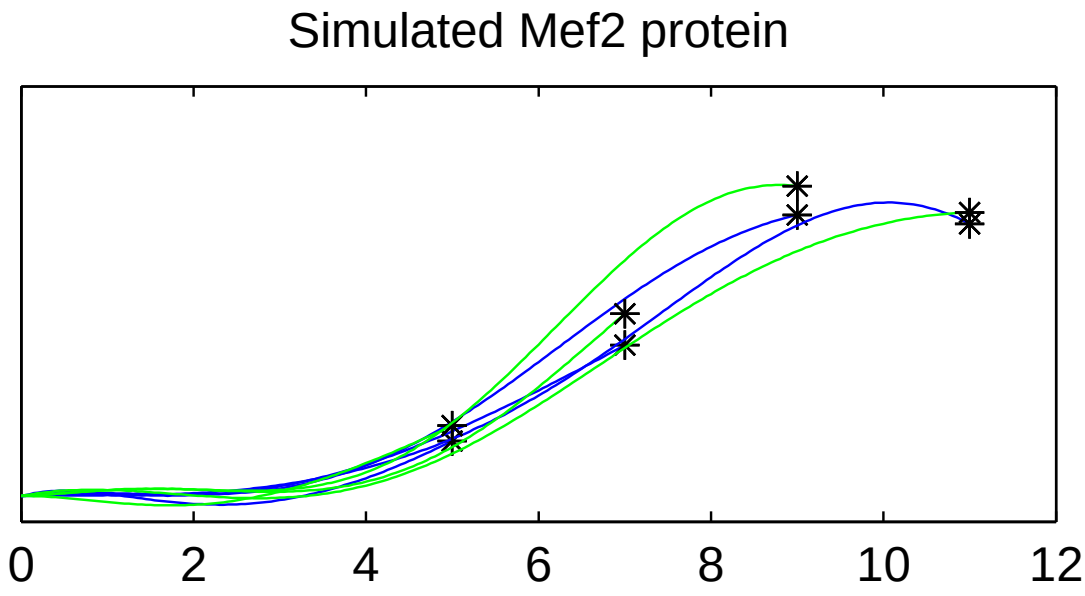


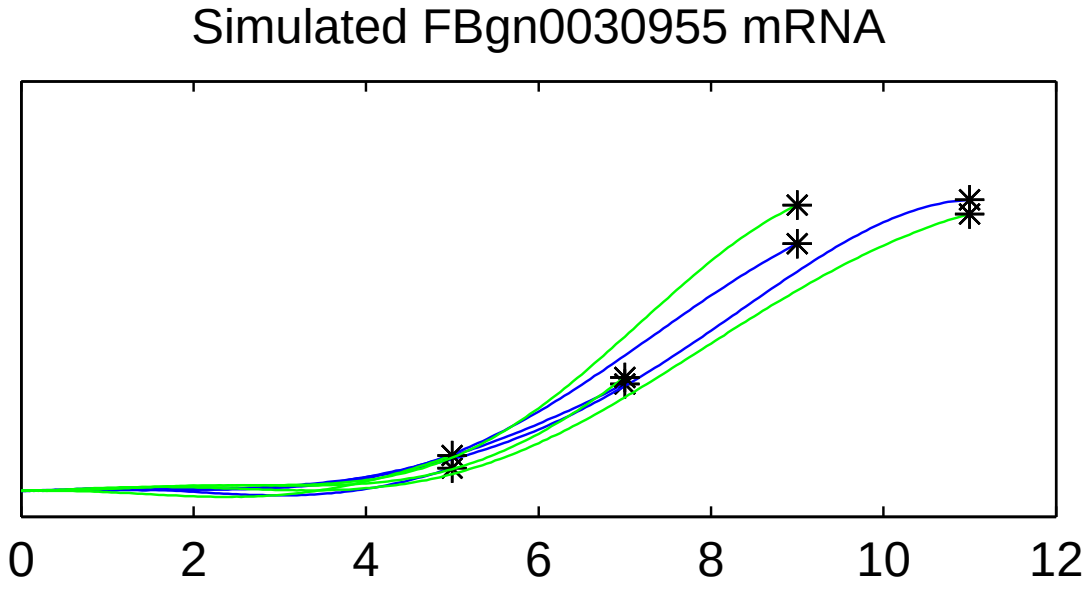
Simulated molecular biology time series



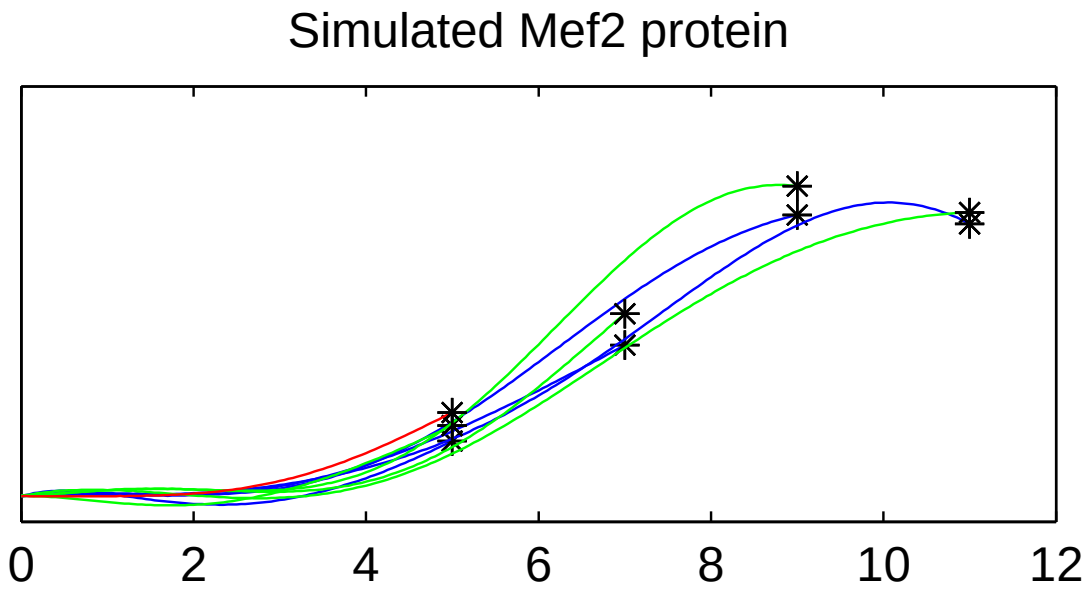


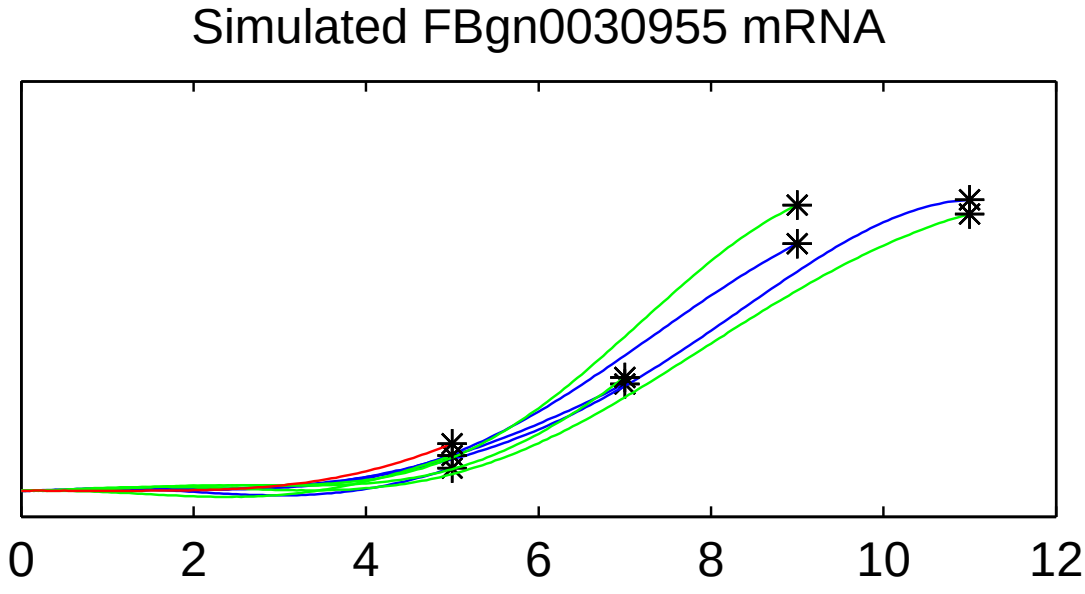
Simulated molecular biology time series



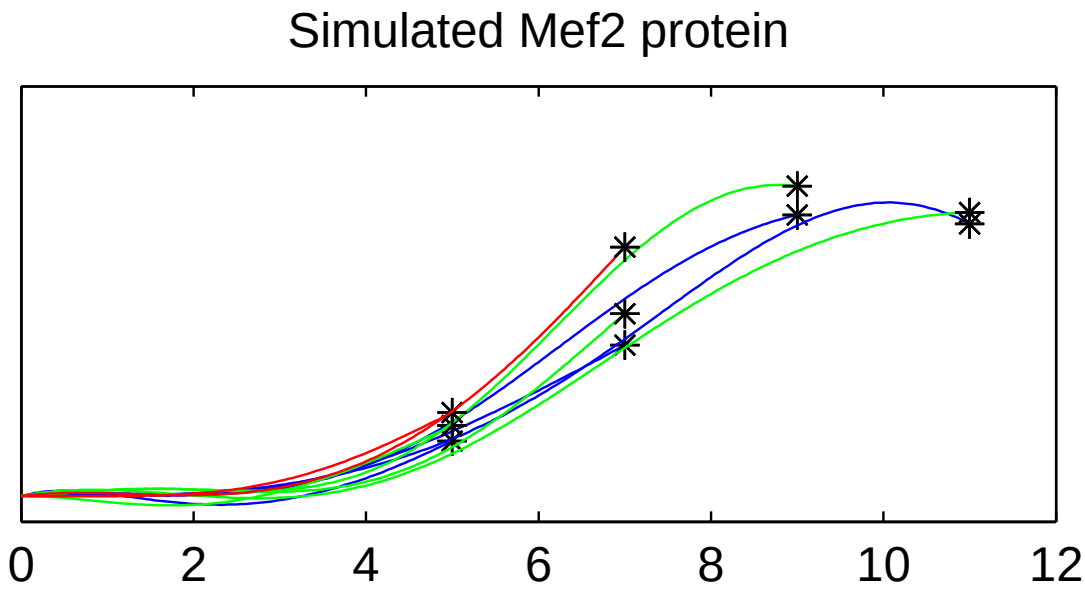


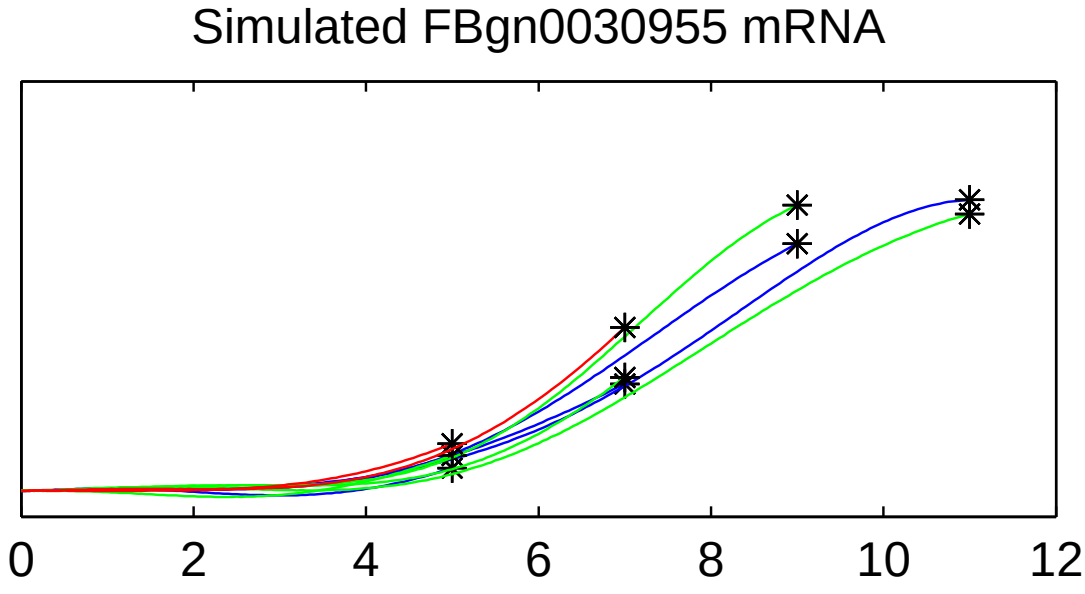
Simulated molecular biology time series



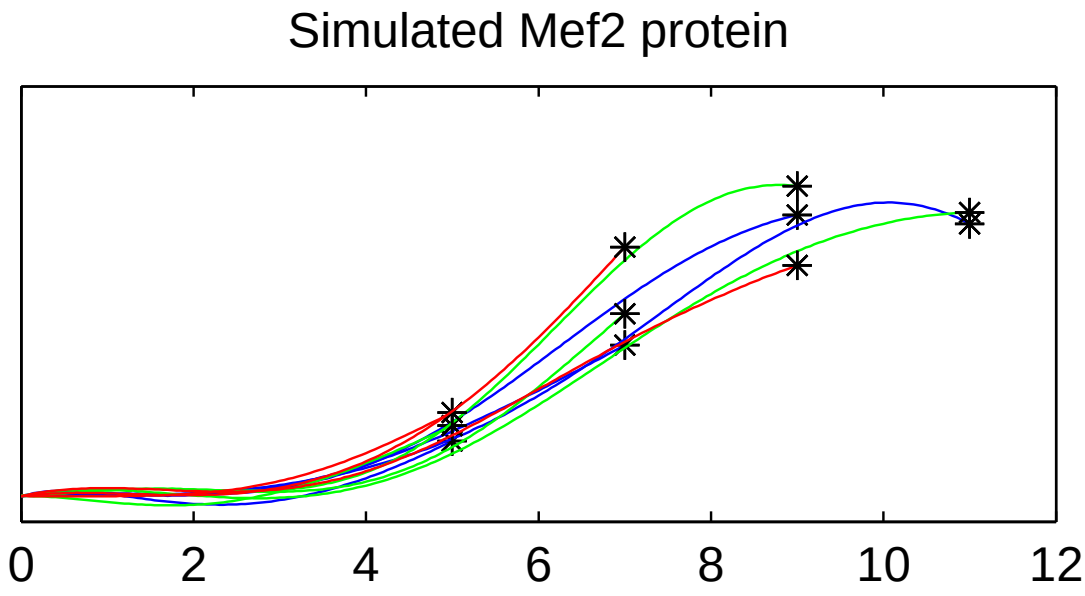


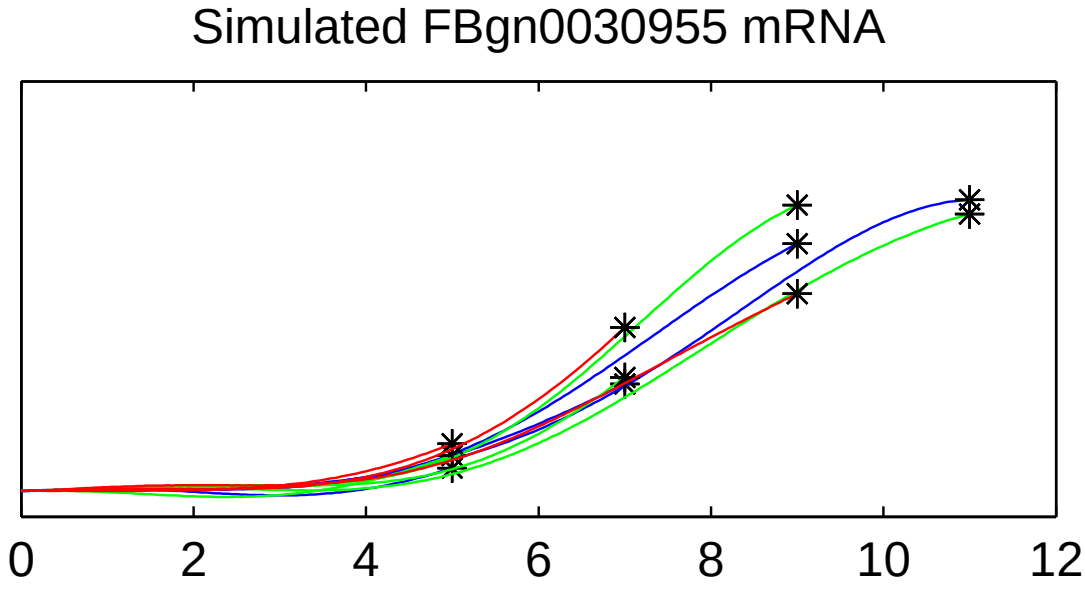
Simulated molecular biology time series



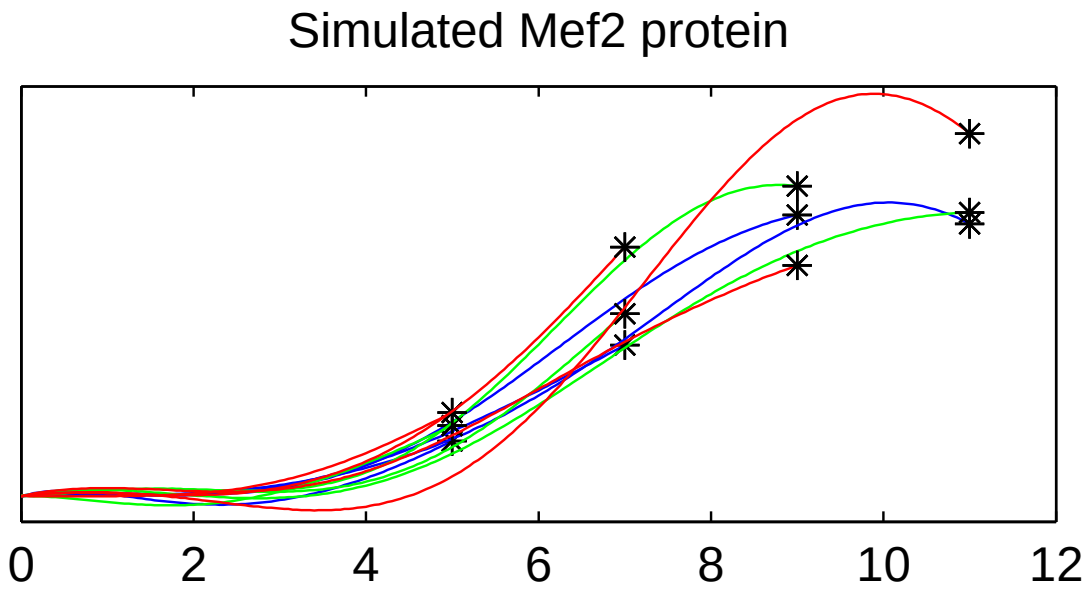


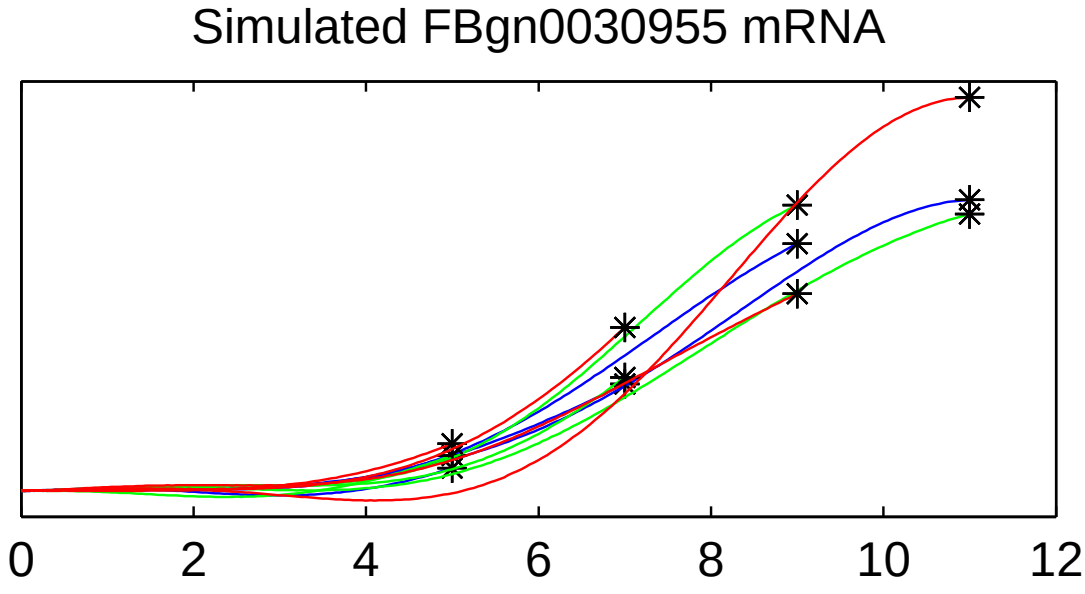
Simulated molecular biology time series



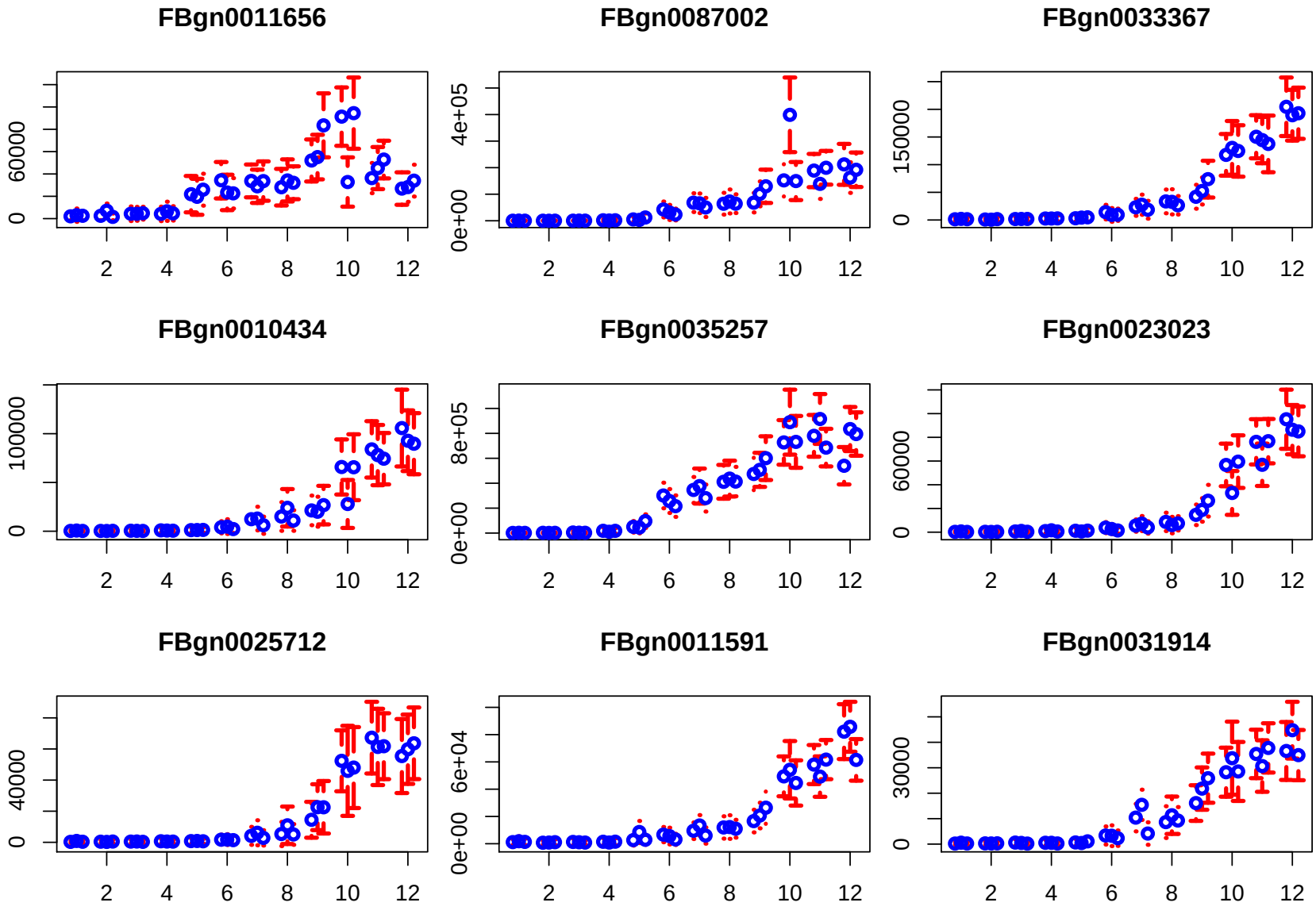


Simulated molecular biology time series





Real gene expression time series



Example model: Linear ODE model of transcription

- Linear Activation Model (**?**, Genome Biology)

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t)$$

- $x_j(t)$ – concentration of gene j's mRNA
- $f(t)$ – concentration of active transcription factor
- Model parameters: baseline B_j, sensitivity S_j and decay D_j
- Placing a Gaussian process (GP) prior on $f(t)$ leads to a joint GP over all concentration profiles (**?**, Bioinformatics)

How to connect the model to data?

1. Assume independent profiles for each complete (biological) repeat
   - Loses statistical power for extra independence assumptions
   - Is it meaningful to order the repeats?
2. Assume one shared underlying profile with independent observations
   - Potentially sensitive to outliers

Exchangeability analysis

Assume $x_j^k(t_i)$ observation of kth repeat of jth gene at ith time

	$x_\cdot^k(t_i) \leftrightarrow x_\cdot^{k'}(t_i)$ "swap arrays"	$x_j^k(t_i) \leftrightarrow x_j^{k'}(t_i)$ "swap single gene"
"Reality"	Yes	No
1. Independent profiles	No	No
2. Shared profile	Yes	Yes

Solution: hierarchical GP model

- Assume the underlying $f(t)$ is composed of a shared and an experiment-specific part $f_{ik}(t)$

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j[f_{\text{shared}}(t) + f_{ik}(t)] - D_j x_j(t)$$

- Covariance is of the same form as usual
- Introduces additional covariance terms for measurements from the same experiment
- Alternative parametrisations of variance of $f_{ik}(t)$
  - Shared across all experiments
  - Sampled independently for each experiment

Exchangeability analysis revisited

Assume $x_j^k(t_i)$ observation of kth repeat of jth gene at ith time

	$x_\cdot^k(t_i) \leftrightarrow x_\cdot^{k'}(t_i)$ "swap arrays"	$x_j^k(t_i) \leftrightarrow x_j^{k'}(t_i)$ "swap single gene"
"Reality"	Yes	No
1. Independent profiles	No	No
2. Shared profile	Yes	Yes
3. Hierarchical model	Yes	No

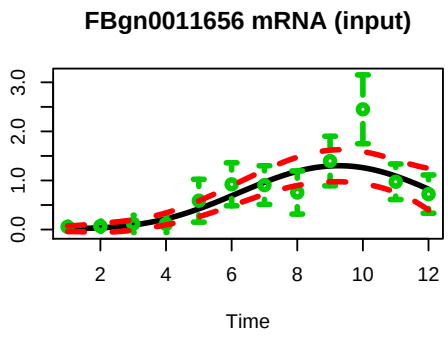
ODE model of translation and transcription

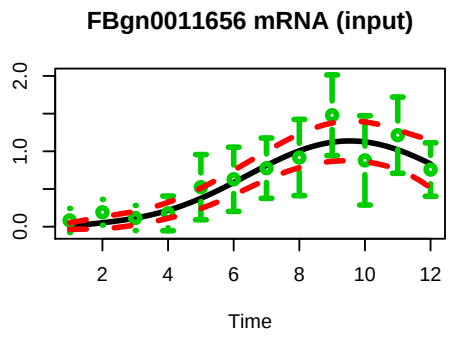
- Assume TF is transcriptionally regulated with related mRNA $y(t)$
- This yields a system of ODEs (**?**)

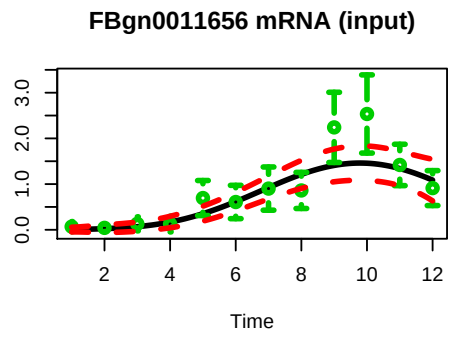
$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \sigma y(t) - \delta f(t)$$
$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t)$$

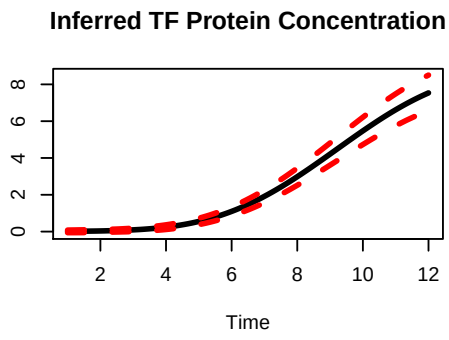
- The corresponding GP model can be derived analogously to the previous case

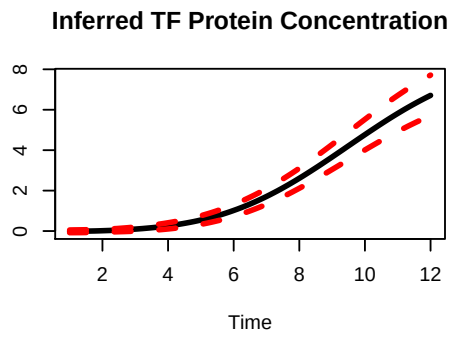
Independent profiles

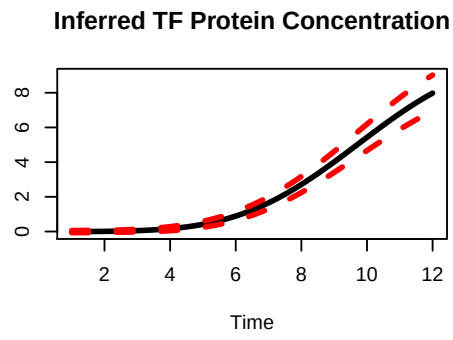


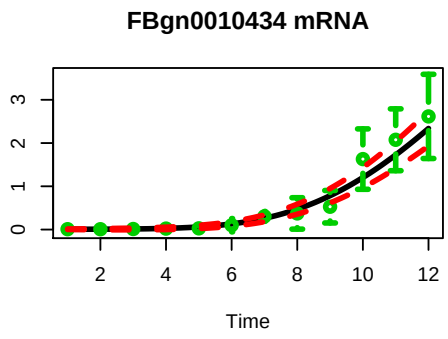


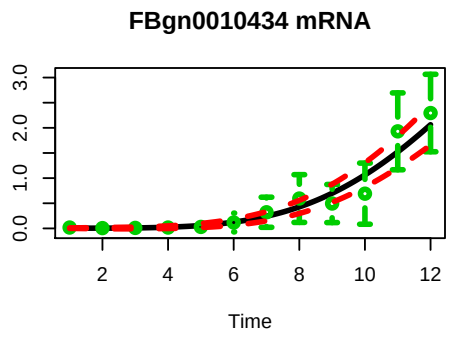


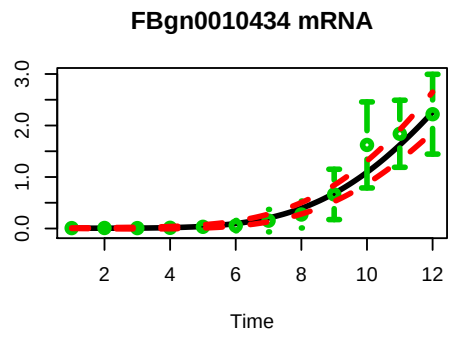




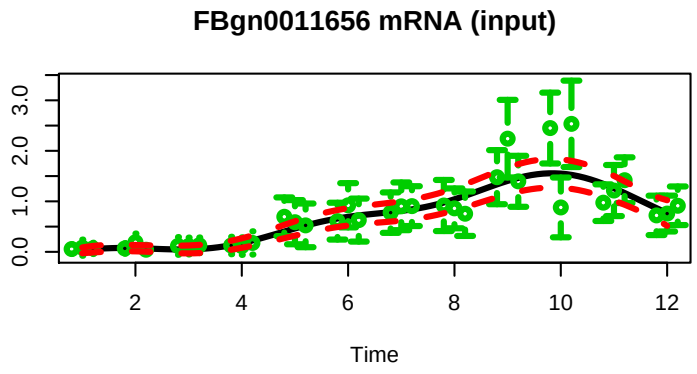


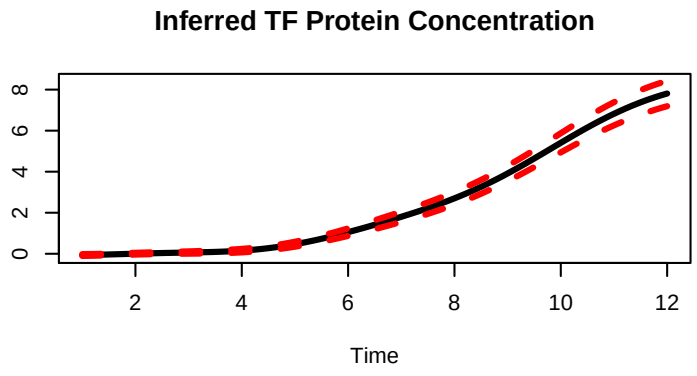


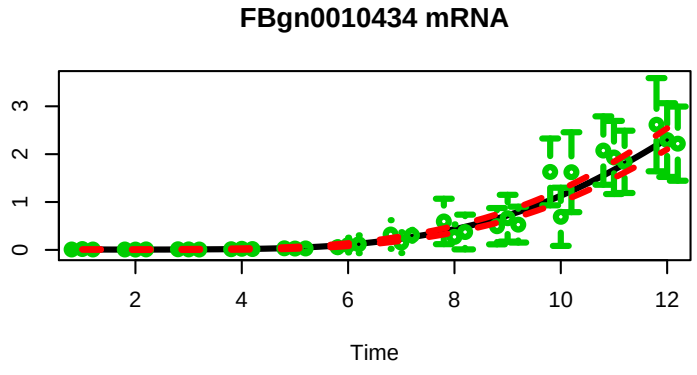




Hierarchical model







Conclusion

- Transcription factor target identification with ODE models
  - Very good performance with linear single-TF models
- Non-linear multiple-TF models also feasible
- Linear model can be extended to account for the experimental structure of time series assays
  - Previous approaches have invalid exchangeability assumptions
- Future work
  - Stochastic differential equation models
  - Incorporation of new data modalities

Acknowledgements

- Pei Gao (University of Cambridge)
- Charles Girardot and Eileen Furlong (EMBL Heidelberg)

References

Now available in Bioconductor:

tigre — Transcription factor Inference through Gaussian process Reconstruction of Expression