Decoding Underlying Behaviour from Destructive Time Series Experiments through Gaussian Process Models

Antti Honkela[1], Neil D. Lawrence[2], and Magnus Rattray[2]

[1] Aalto University, Department of Information and Computer Science
Helsinki, Finland
[2] University of Manchester, School of Computer Science

March 30, 2010

Aalto University
School of Science
and Technology

The University of Manchester

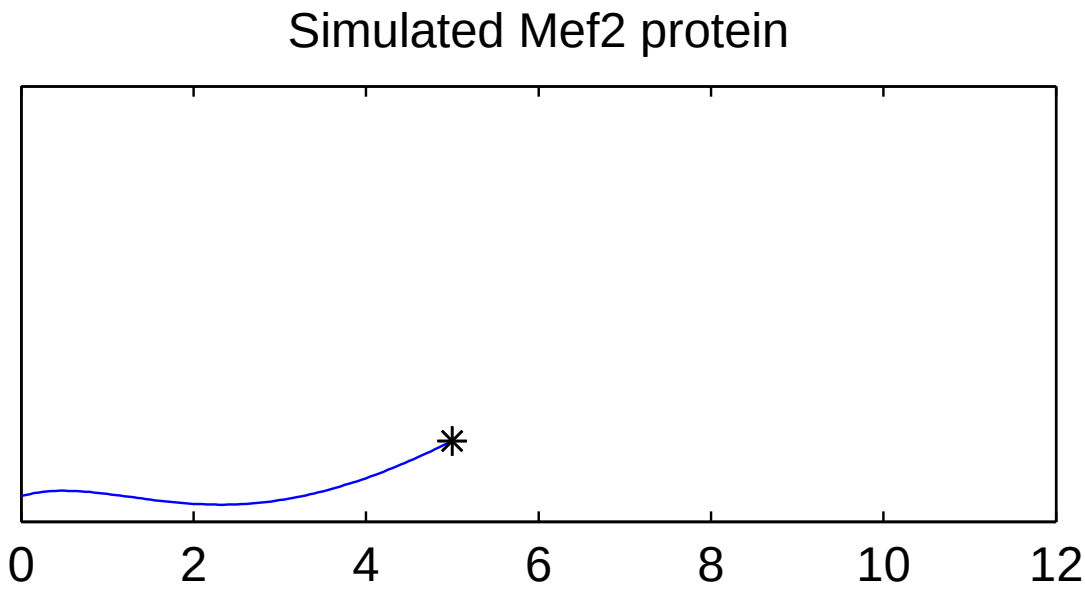
Molecular biology time series

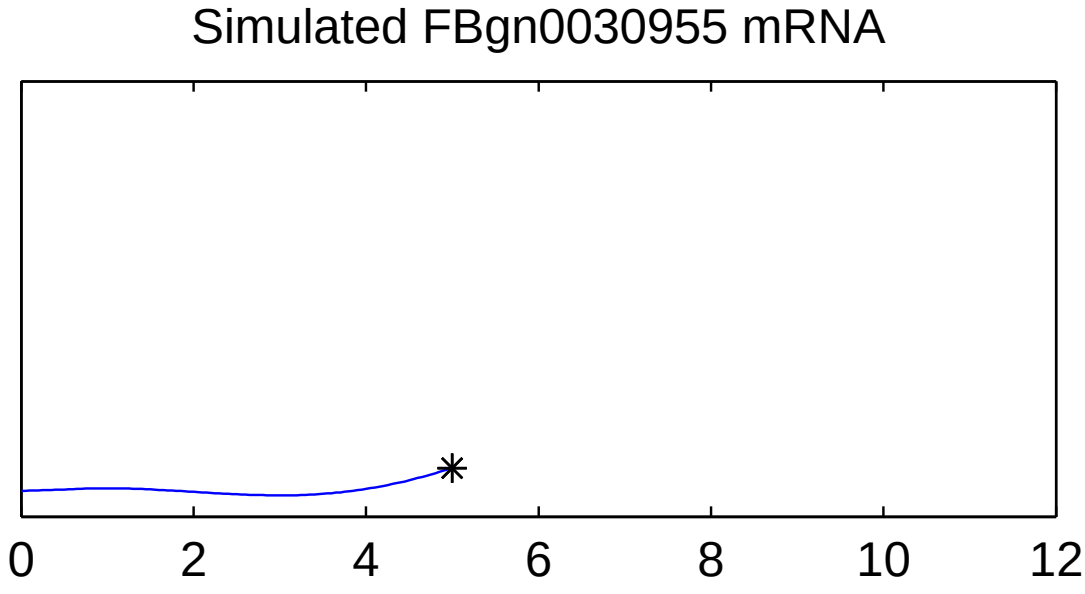
- Biological systems are dynamic, observing their time evolution very helpful
- Time series measurements of gene expression, protein activity, protein binding, ...
- Problem: most of these assays are highly disruptive to the sample
- Therefore: time series = series of independent experiments run for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

Outline

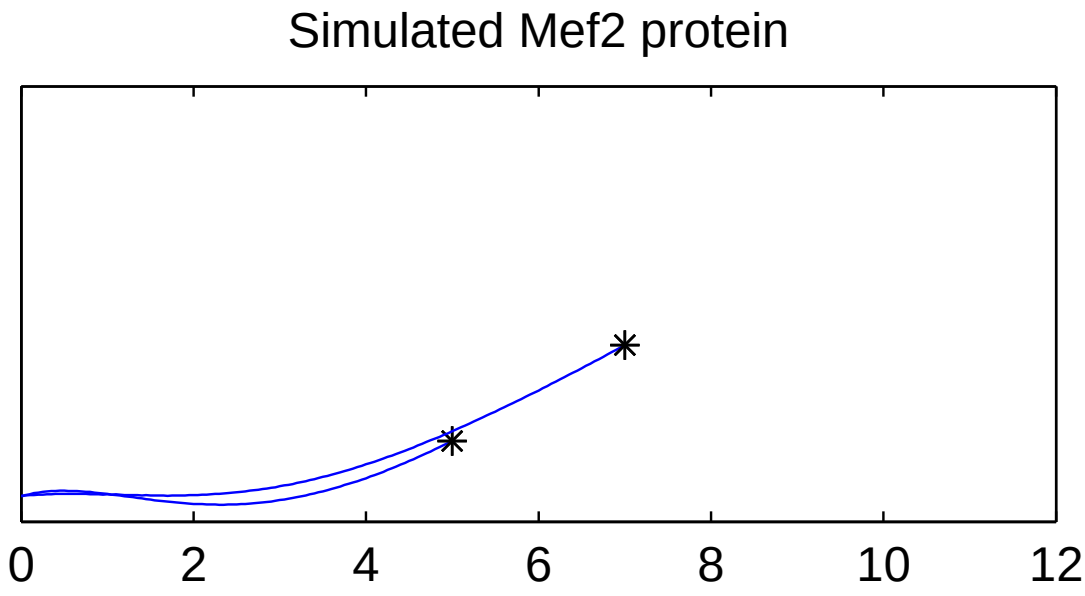
Outline

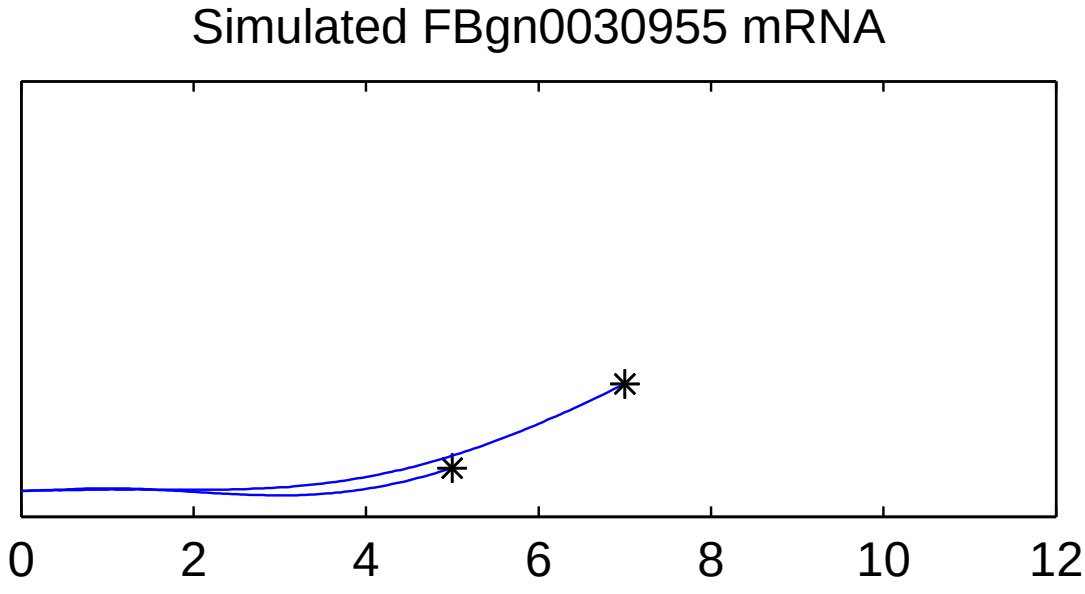
Simulated molecular biology time series



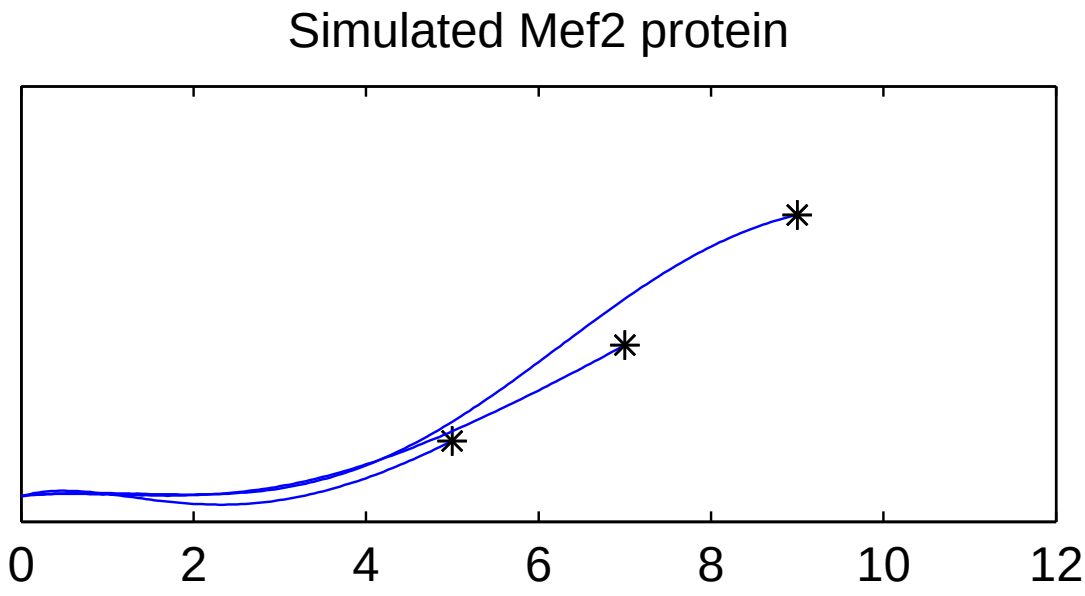


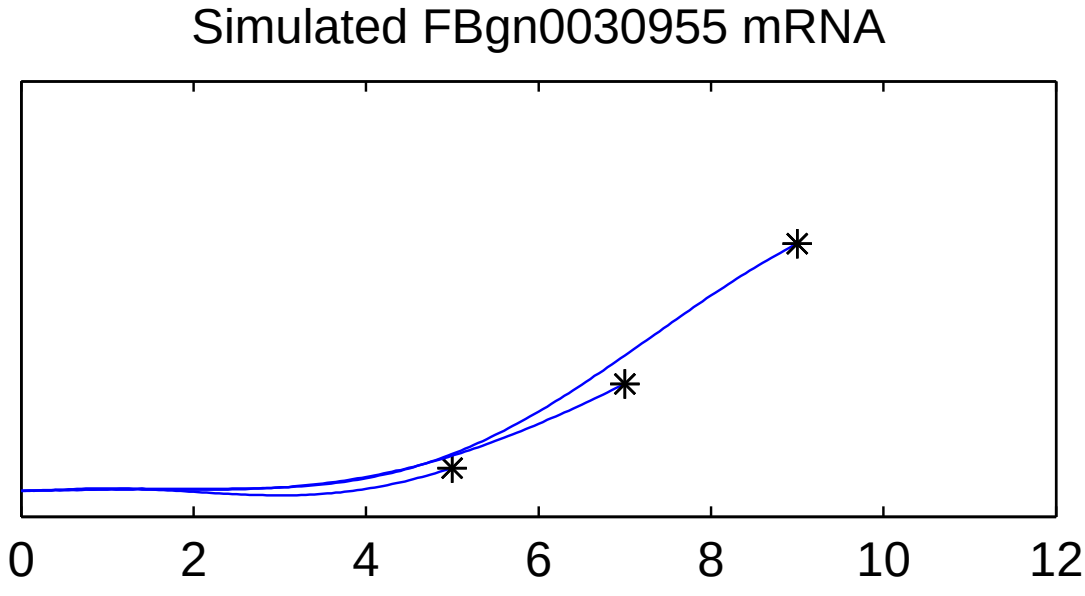
Simulated molecular biology time series



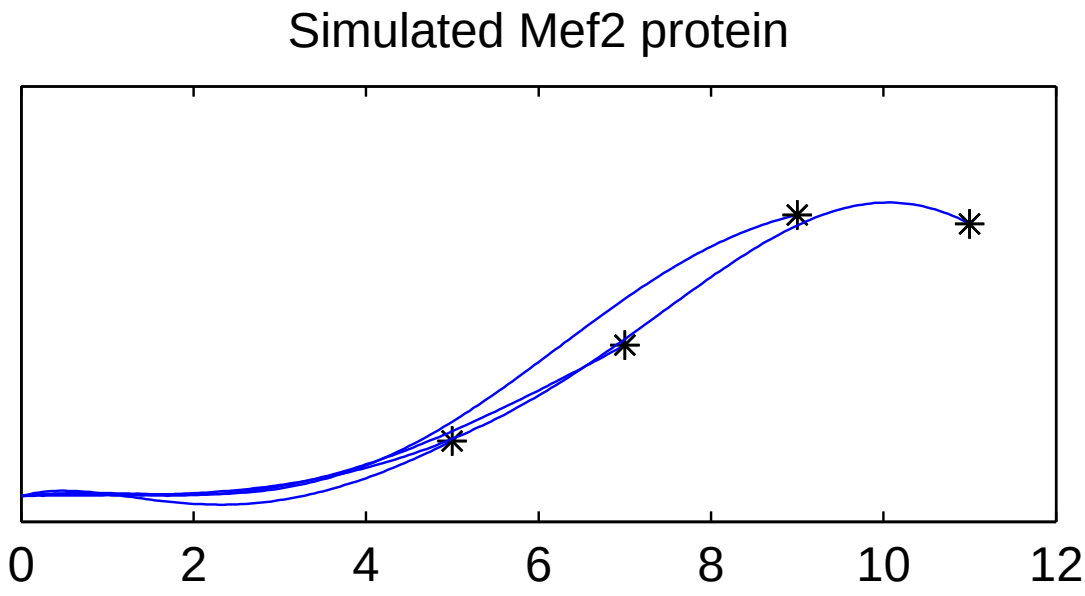


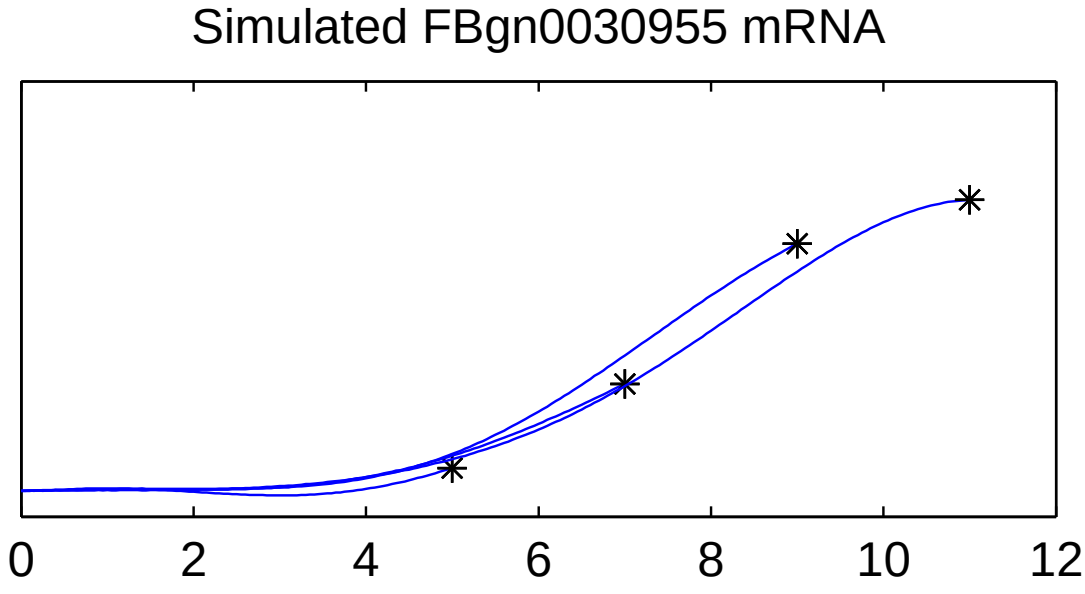
Simulated molecular biology time series



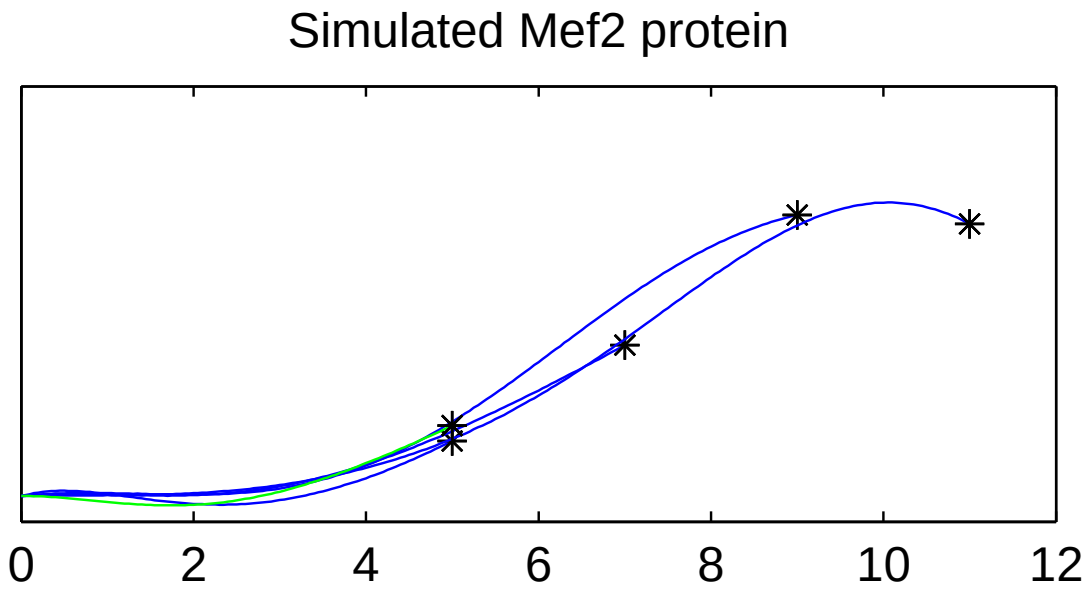


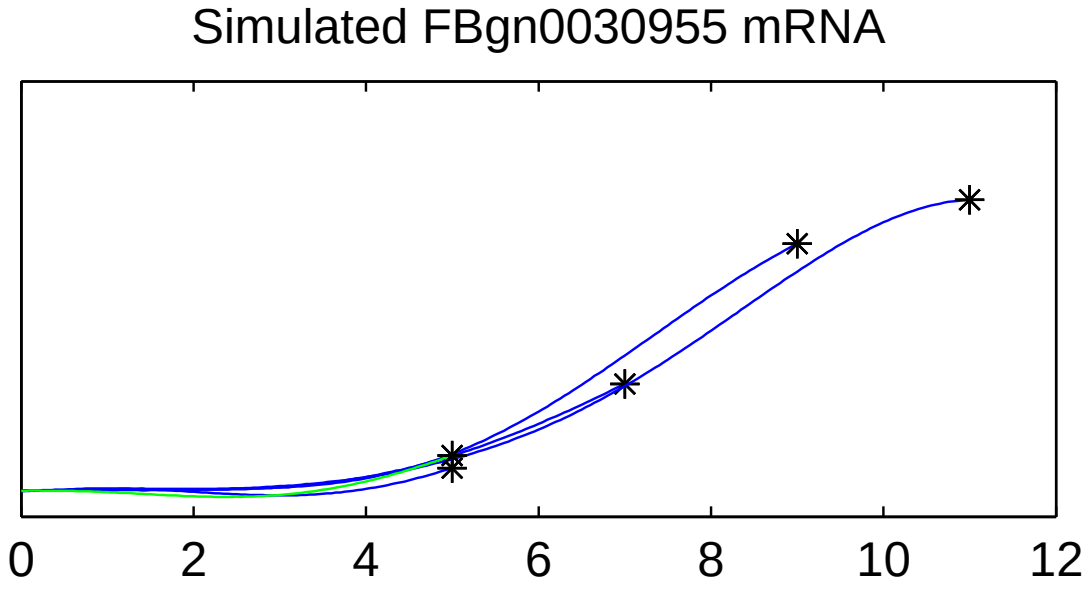
Simulated molecular biology time series



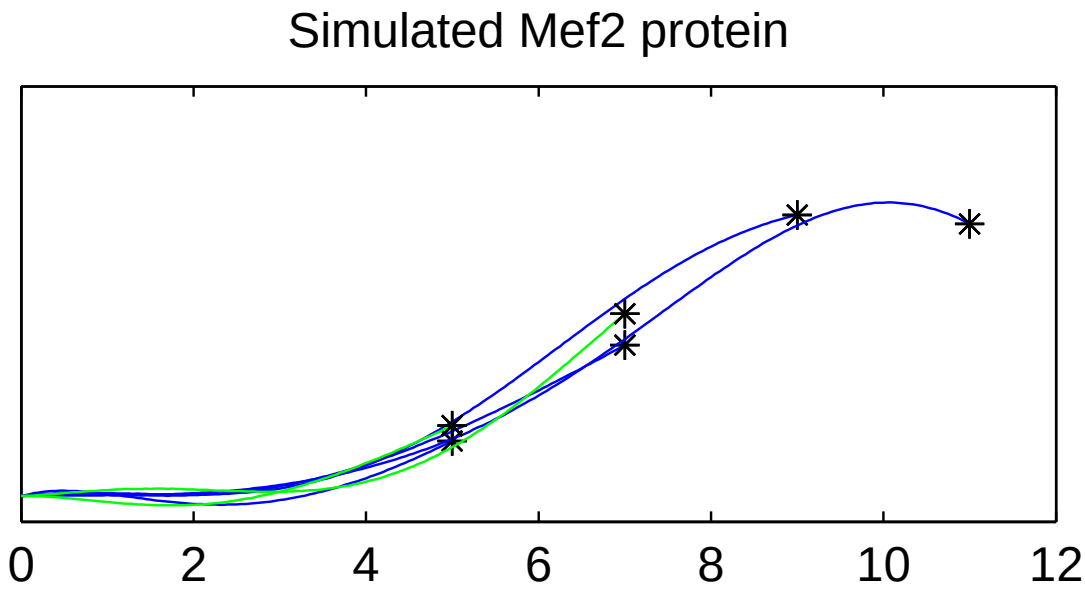


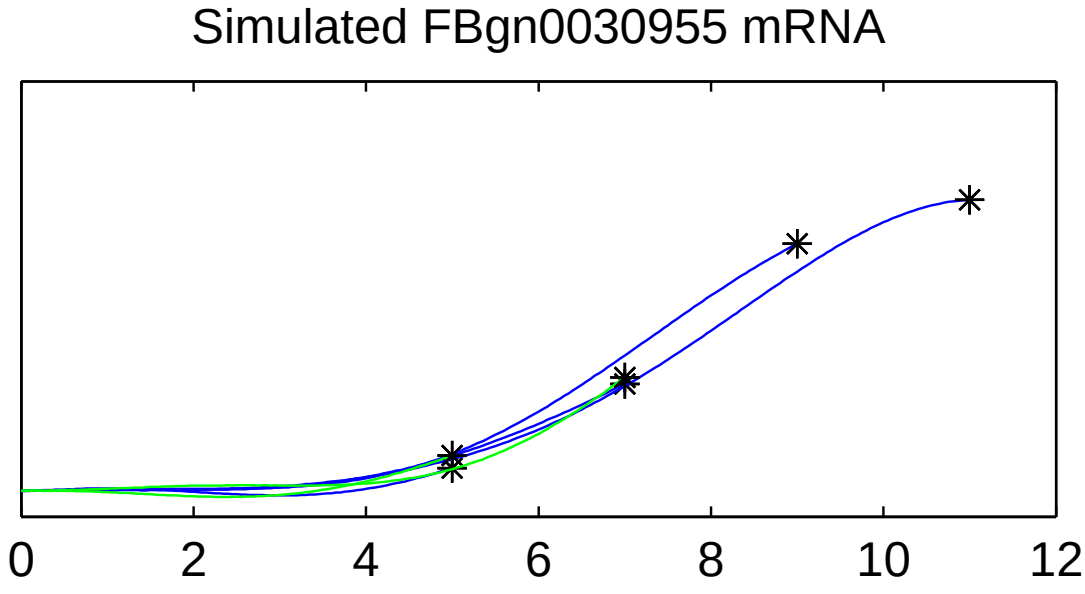
Simulated molecular biology time series



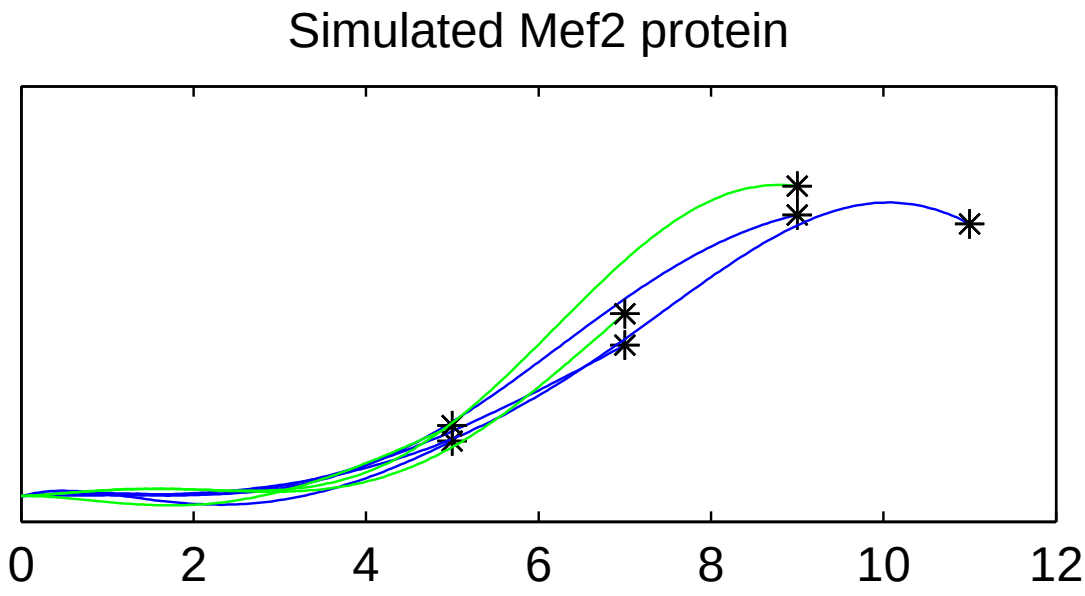


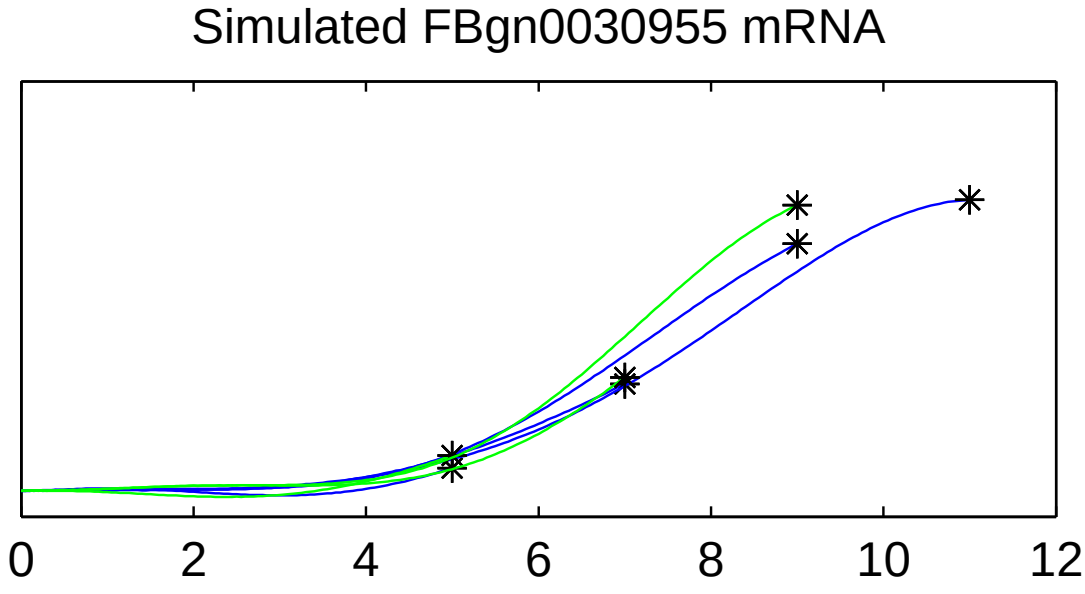
Simulated molecular biology time series



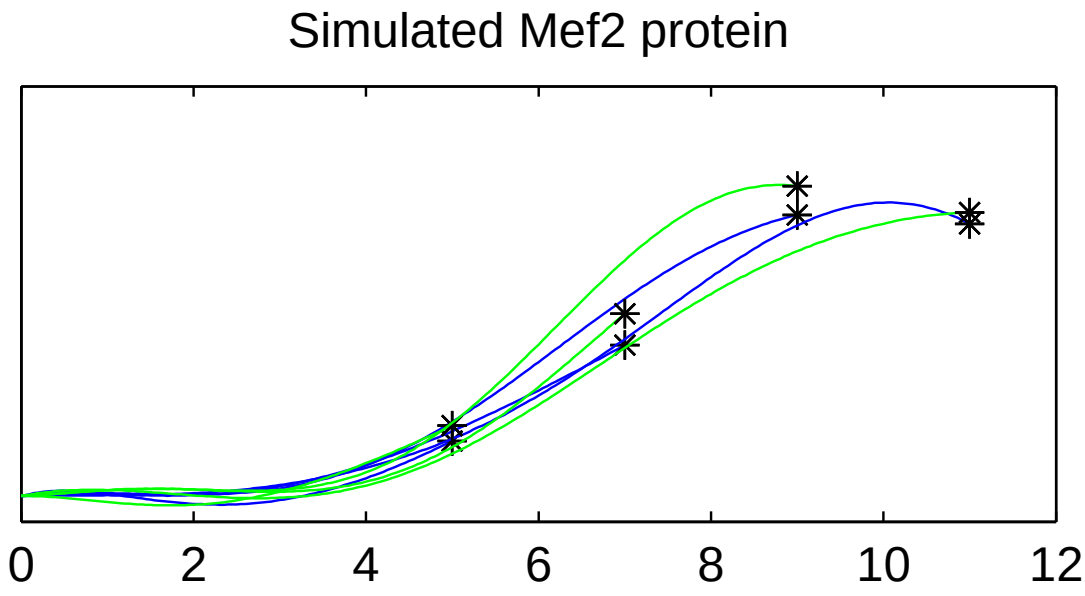


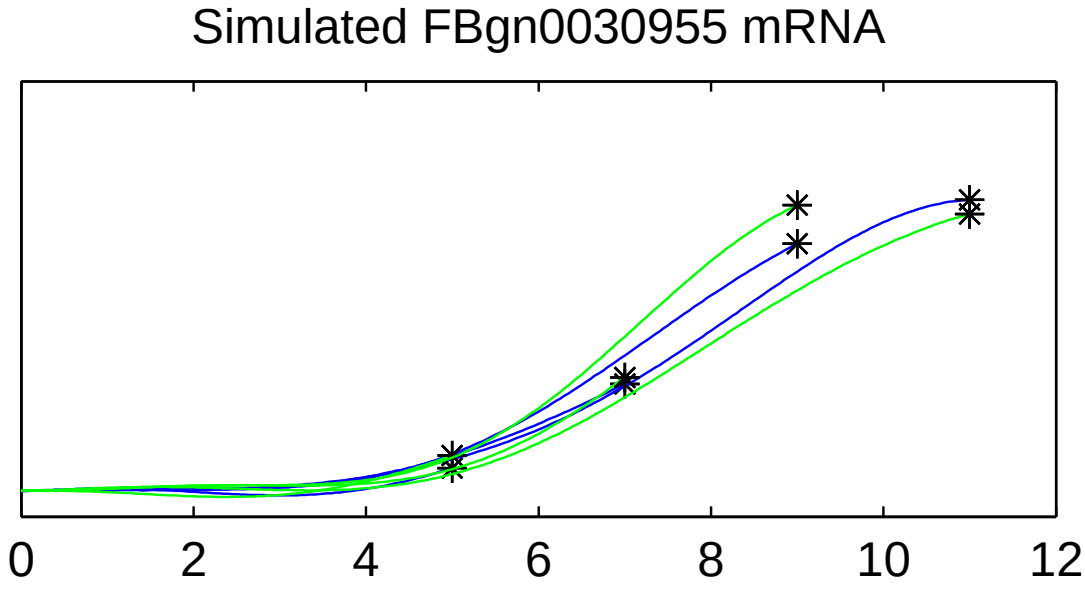
Simulated molecular biology time series



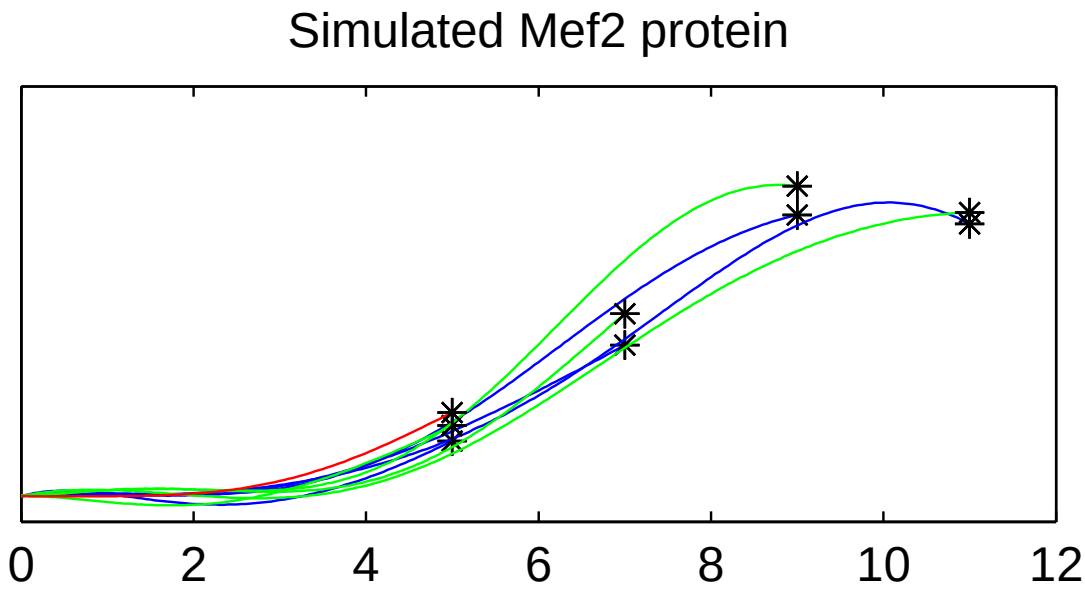


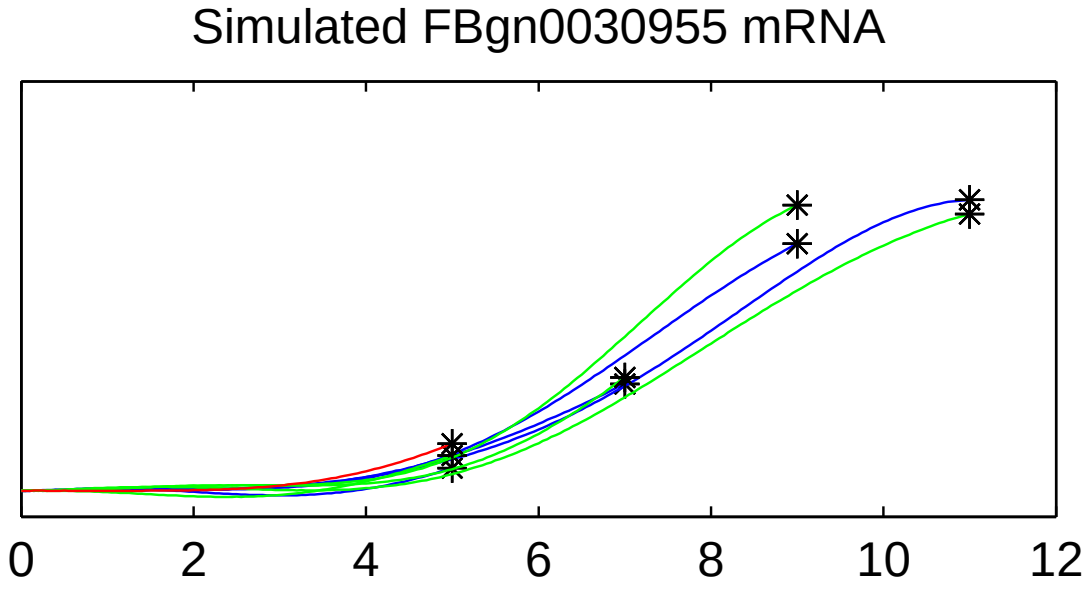
Simulated molecular biology time series



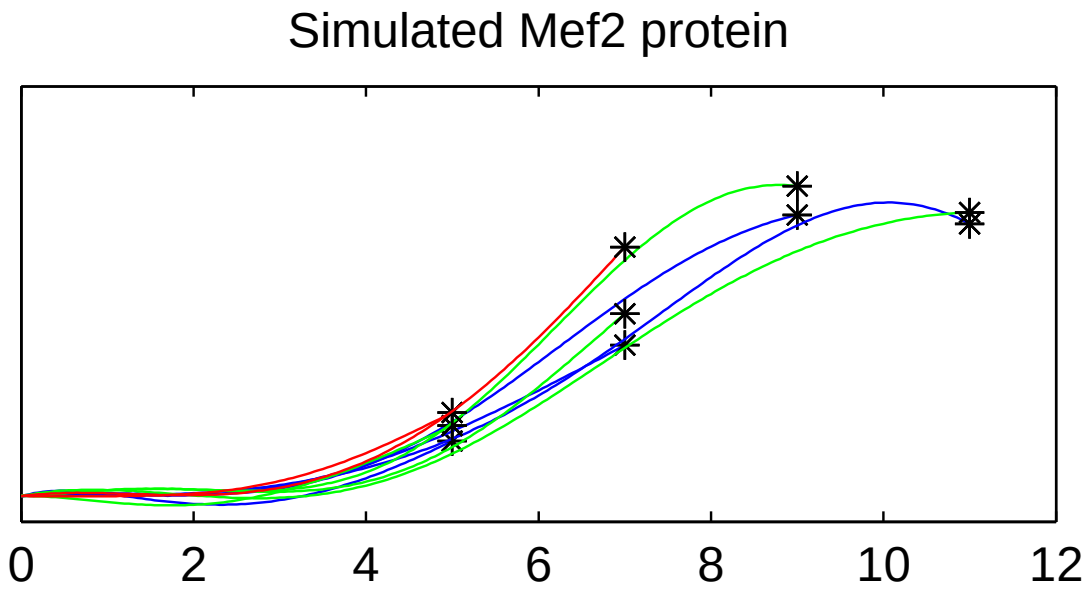


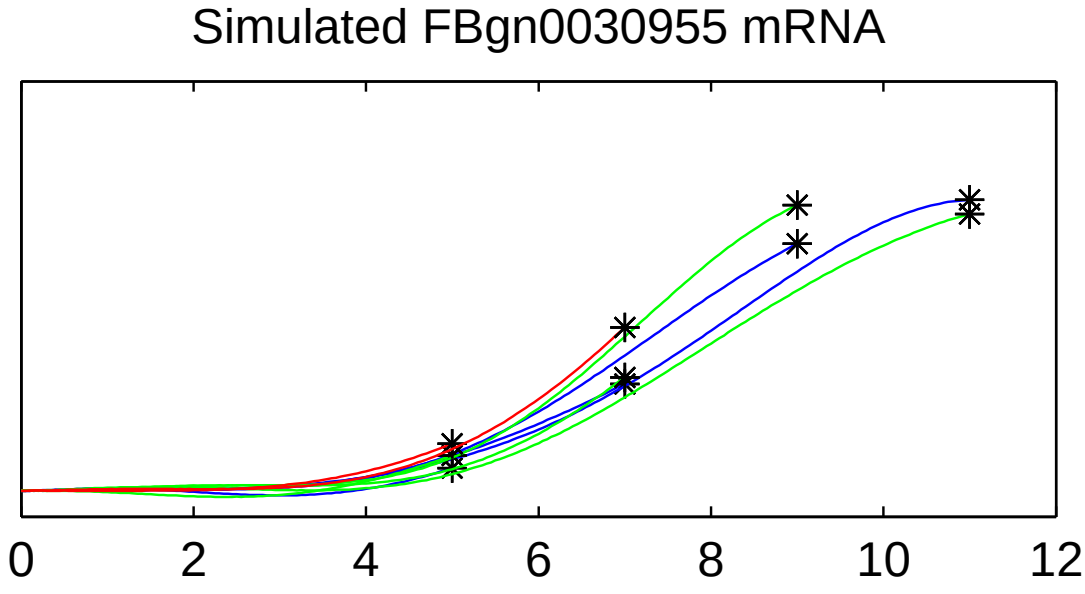
Simulated molecular biology time series



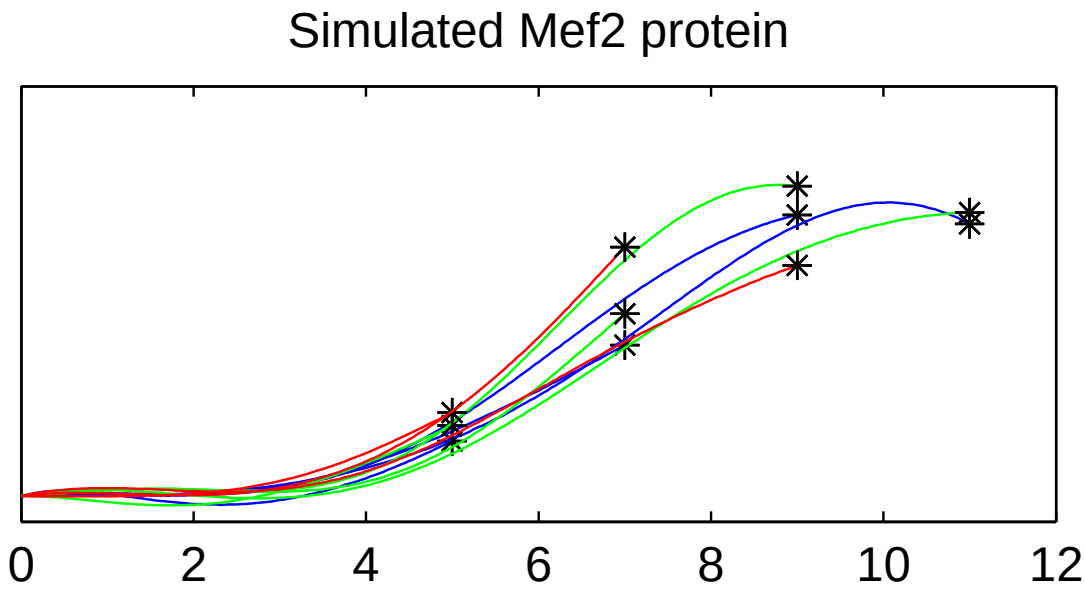


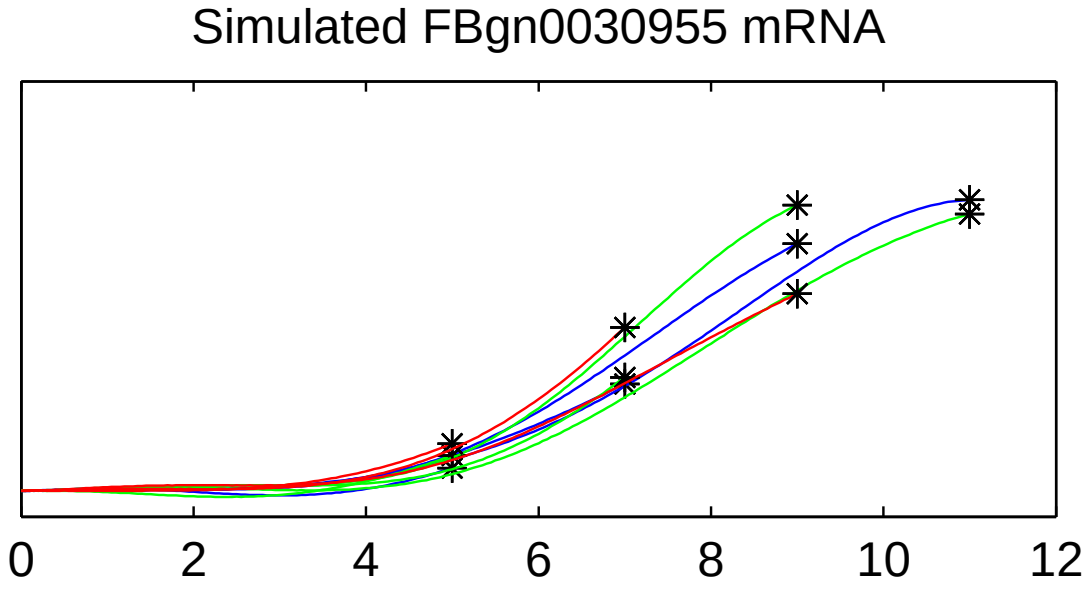
Simulated molecular biology time series



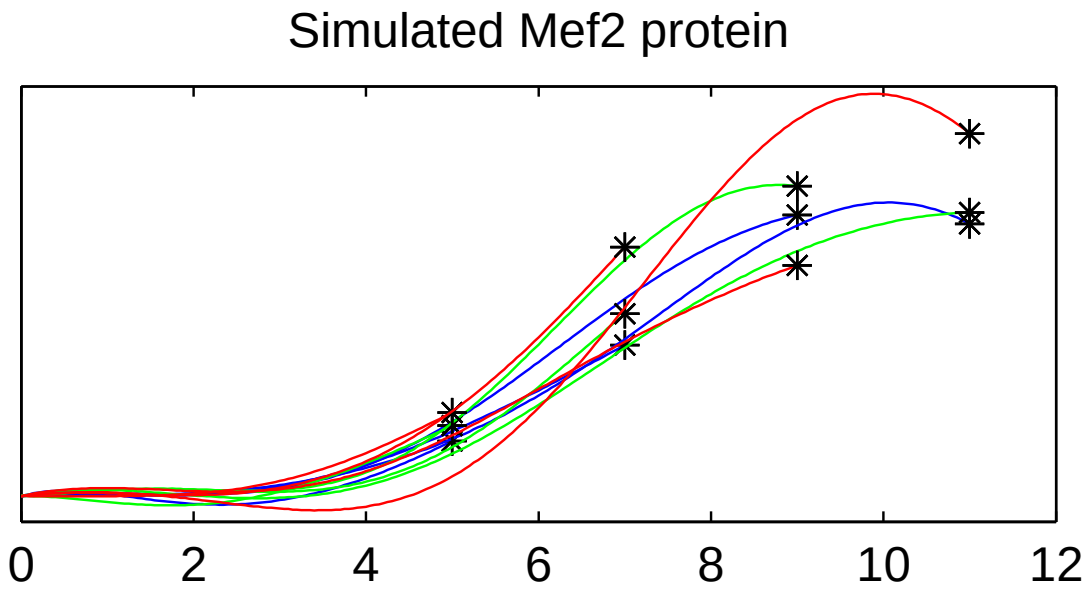


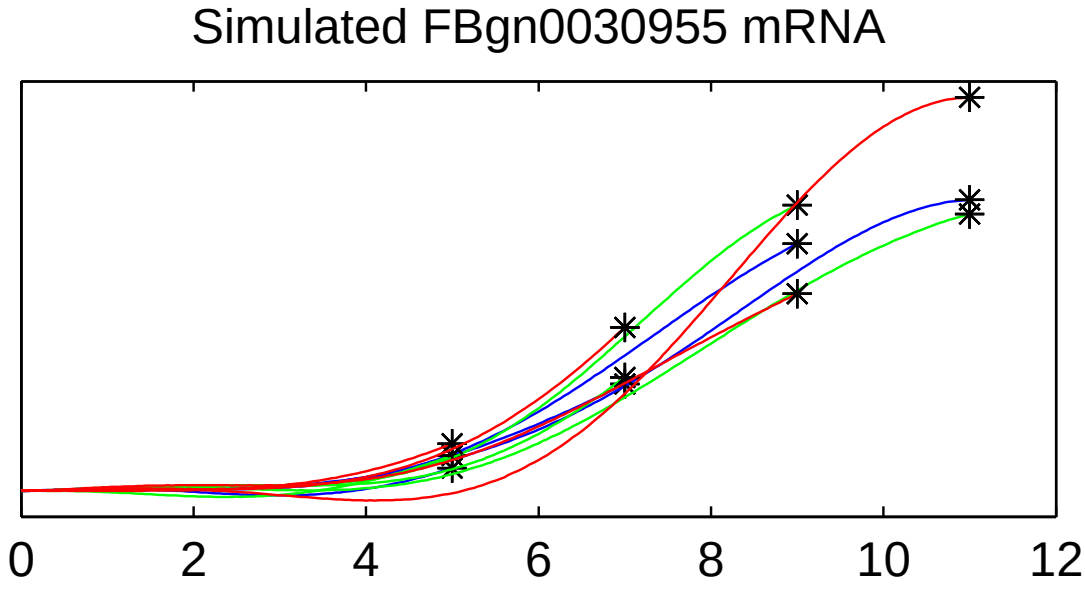
Simulated molecular biology time series



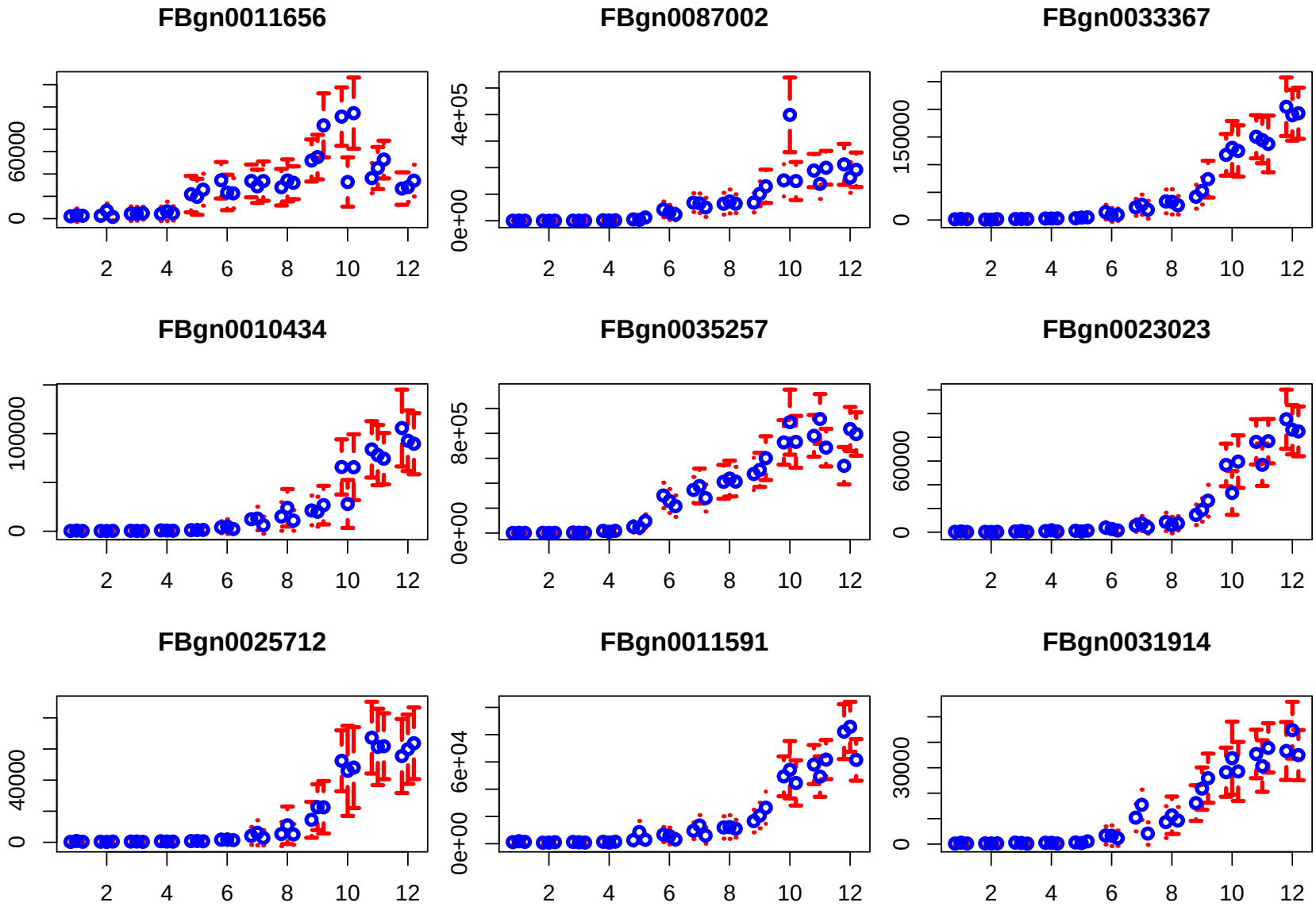


Simulated molecular biology time series





Real gene expression time series



Outline

Example model: Linear ODE model of transcription

- Linear Activation Model (Barenco et al., 2006, 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 (Gao et al., 2008, 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

Outline

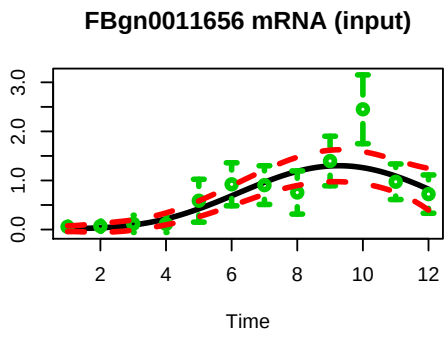
ODE model of translation and transcription

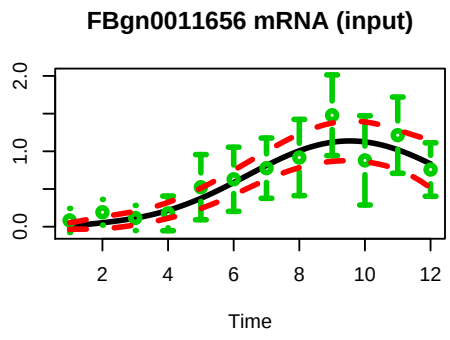
- Assume TF is transcriptionally regulated with related mRNA $y(t)$
- This yields a system of ODEs (Gao et al., 2008)

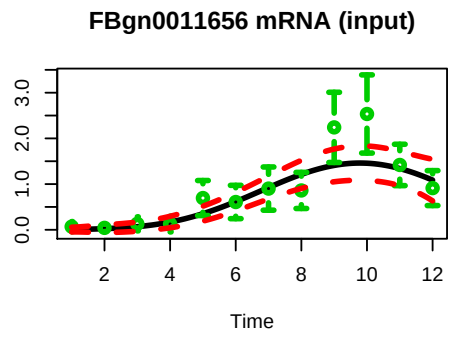
$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \sigma y(t) - \delta f(t)$$
$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t)$$

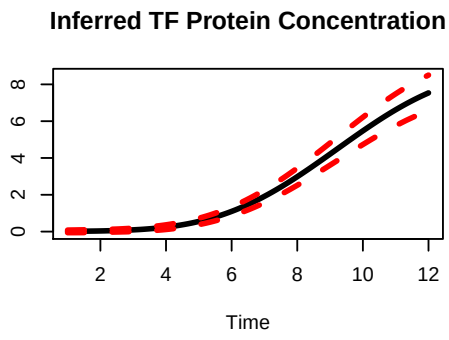
- The corresponding GP model can be derived analogously to the previous case

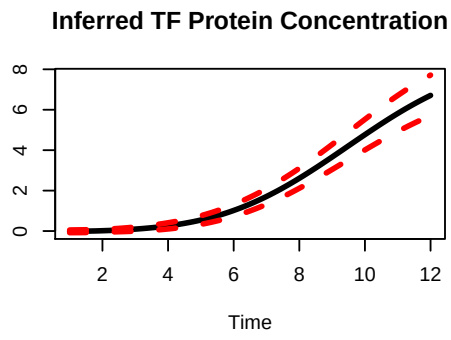
Independent profiles

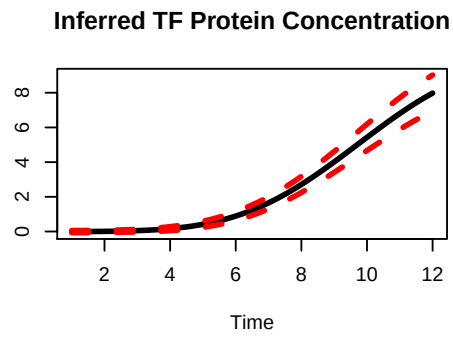


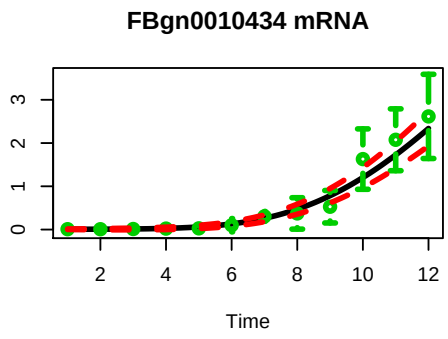


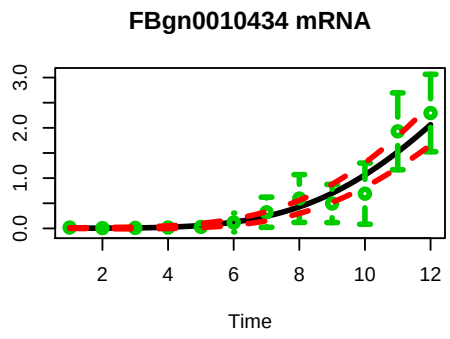


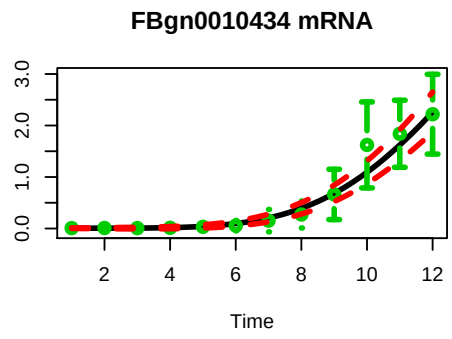




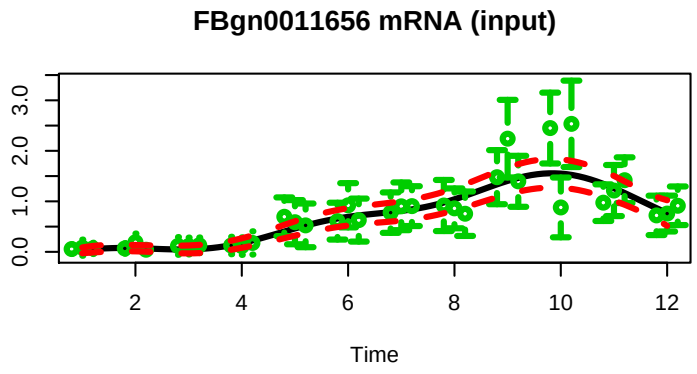


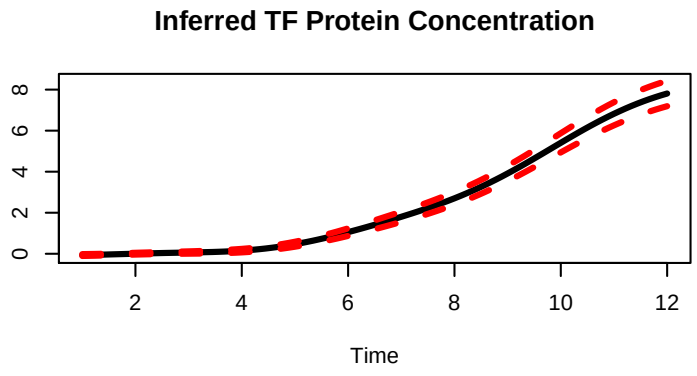


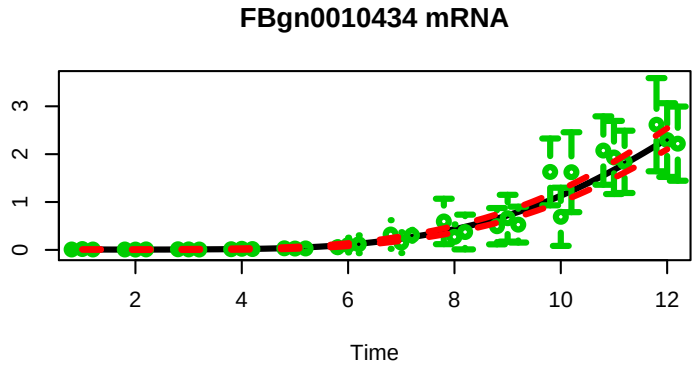




Hierarchical model







Conclusion

- Previous models of time series expression data are wrong
  - Invalid exchangeability assumptions
- Proposed hierarchical model rectifies this
- Open problems / work in progress
  - Need to move beyond Gaussian likelihoods?
  - How to do MCMC in these models?

Acknowledgements

Funding:
Academy of Finland
EU Network of Excellence PASCAL2

Coming soon to Bioconductor:

tiger — Transcription factor Inference through Gaussian process Expression Reconstruction

References

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006. [PDF]. [DOI].

P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75, 2008. [PDF]. [DOI].