Supplementary information

Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks:

Hybrid Stochastic/Deterministic Simulation

Daniel E. Zak[α,β]

Gregory E. Gonye[β]

James S. Schwaber[β,⋆]

Francis J. Doyle III[γ]

[α] Department of Chemical Engineering, University of Delaware, Newark, DE 19716

[β] Department of Pathology, Cell Biology and Anatomy, Thomas Jefferson University Philadelphia, PA 19107

[γ] Department of Chemical Engineering, University of California, Santa Barbara, CA 93106

⋆ *Please address questions or comments about this material to: James Schwaber,* james.schwaber@mail.tju.edu*, (215) 503–7823, FAX: (215) 923–3808*

1 Introduction

In a stochastic framework, key components (transcripts, promoters, proteins) are expressed in integer numbers and reactions are treated as probabilistic events. This is appropriate for modeling genetic regulatory networks given the very low concentrations of some components (promoters and transcripts). Stochastic simulations also make possible the calculation of variances in transcript levels over time, thereby accounting for one source of noise in gene expression studies. These variances play an essential role in the the present study in the practical identifiability analysis.

The stochastic formalism has the disadvantage of requiring significant computational resources, however. For this reason we developed the efficient hybrid stochastic/ deterministic approach described presently. In our approach, a stochastic integrator (Gillespie's Direct Method [2]) was coupled with a deterministic integrator (Implicit Euler [3]). The stochastic integrator was used

for the components present in small numbers (promoters and transcripts) and the deterministic integrator was used for components present in large numbers (proteins, transcription factor dimers), thereby breaking the transcriptional module into stochastic and deterministic subsystems (Figure 1). We observed a speed up greater than 500–fold over the full stochastic simulation when this approach was implemented on the genetic regulatory network model. In the present document, specific details about the hybrid stochastic/ deterministic approach are presented, followed by a more detailed discussion of the errors and a discussion of the validity in the Gaussian assumption for the transcript levels used in the practical identifiability analysis.

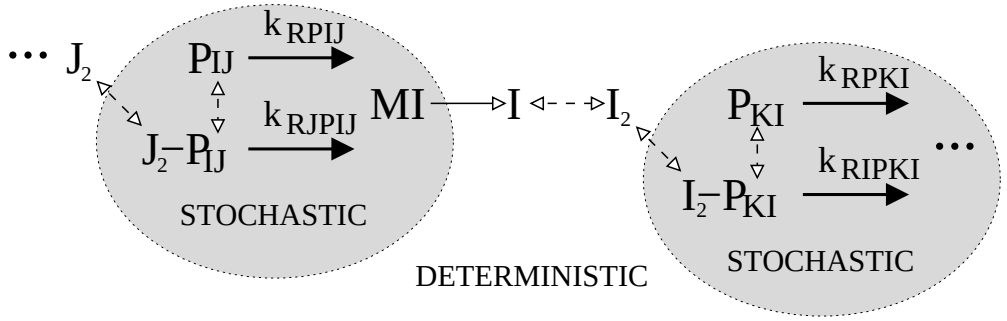

Figure 1: **Stochastic and deterministic subsystems in the transcriptional module:** The stochastic subsystem consists of those states typically present in small numbers (promoters and transcripts) while the deterministic subsystem consists of those components typically present in larger numbers (proteins, dimers).

2 Algorithm

The most common method for simulating systems in a stochastic framework (Gillespie's Direct Method) has a feature that the integration step size (and hence computational speed) depends inversely on the total sum of reaction rates in the system. Mathematically:

$$\Delta t = -log(r)/V_T \tag{1}$$

where Δt is the time step, r is a uniform pseudo–random number (0,1), and V_T is the sum of all of the reaction rates at the current simulation time.

In the hybrid approach described presently, the transcriptional regulatory module was first simplified by assuming that promoter binding and unbinding reactions do not influence the concentration of transcription factor dimers (J_2). The impact of this assumption on the numerical results was small because the transcription factor dimers were generally in great excess compared to the promoters. It is an important assumption, however, because it results in a simplified unidirectional coupling between the deterministic states and the stochastic states, simplifying the calculations. An algorithm for performing hybrid stochastic/ deterministic simulations may then be defined:

1. Calculate probabilities for stochastic reactions (promoter binding/unbinding, transcription, mRNA degradation) as per Gillespie's Direct Method.

2. Determine step size for stochastic reactions (Gillespie's Direct Method).

3. Update deterministic states with a deterministic ODE integrator (Implicit Euler) for the step size determined in (2), keeping the stochastic states constant.

4. Re–calculate reaction probabilities for stochastic reactions, determine the next one to occur (Gillespie's Direct Method), and execute it. Note that execution of the stochastic reactions will not affect the concentrations of the deterministic states.

5. Repeat (1)–(4) until end time.

The Implicit Euler integrator was chosen to integrate the deterministic system because it is computationally simple and robust to the varying integration step sizes that are characteristic of Gillespie's Direct Method.

3 Closed–form expressions for deterministic subsystems

The general form for the Implicit Euler integrator is:

$$\frac{dX}{dt} = f(X) \approx \frac{X^+ - X^-}{\Delta t} = f(X^+) \tag{2}$$

The integration is carried out by solving for the future values of the states (X^+) in terms of the parameters and the present values of the states (X^-).

The deterministic subsystems in the present genetic regulatory network, shown in Figure 2 are relatively simple, allowing derivation of closed–form expressions for their integration. Ordinary differential equations describing the subsystems, followed by closed–form expressions for integrating them with the Implicit Euler method, are given below. In all cases, Δt refers to the time–dependent integration time step obtained from the stochastic subsystem.

3.1 Mechanism A: Homodimerization

3.1.1 ODE

$$\begin{aligned} \frac{dj}{dt} &= k_{Tj} M_j - 2k_{j_2} j^2 + 2k_{Uj_2} j_2 - k_{Dj} j \\ \frac{dj_2}{dt} &= k_{j_2} j^2 - k_{Uj_2} j_2 - k_{Dj_2} j_2 \end{aligned} \tag{3}$$

where j is transcription factor monomer, j_2 is transcription factor dimer, M_j is an mRNA transcript, and the k's represent rate constants, with subscripts Tj indicating translation, j_2 indicating dimerization, Uj_2 indicating undimerization, Dj indicating monomer degradation, and Dj_2 indicating dimer degradation.

3.1.2 Closed–form expression for integration with the Implicit Euler method

Defining:

$$\begin{aligned} a_{Tj} &\equiv k_{Tj} \times \Delta t \\ a_{Dj} &\equiv k_{Dj} \times \Delta t \\ a_{j_2} &\equiv k_{j_2} \times \Delta t \\ a_{Uj_2} &\equiv k_{Uj_2} \times \Delta t \\ a_{Dj_2} &\equiv k_{Dj_2} \times \Delta t \end{aligned} \tag{4}$$

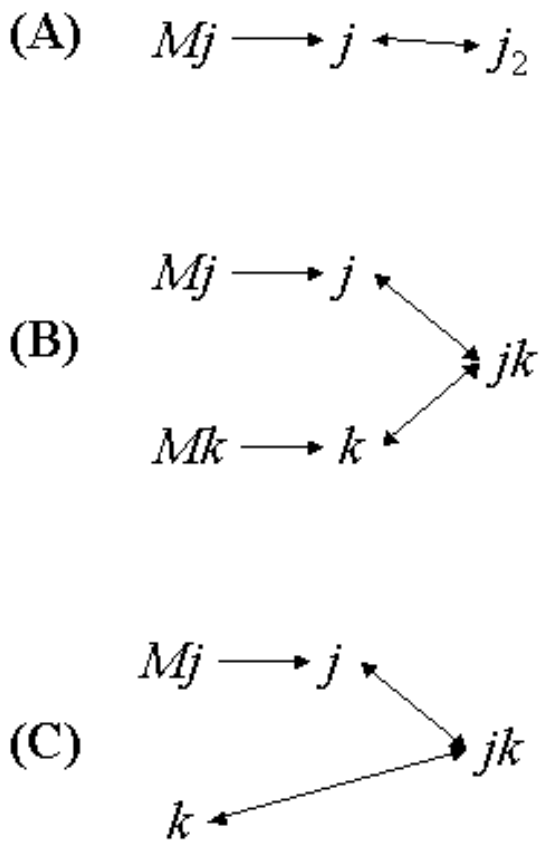

Figure 2: Deterministic subsystems in the genetic regulatory network model with closed-form expressions for Implicit Euler integration. M_j and M_k indicate transcripts from the stochastic subsystem. (A) Homodimerization (genes C, D, F, G, and K); (B) Heterodimerization (genes A and B); (C) Ligand binding (gene E and ligand Q).

Further defining:

$$\begin{aligned} a_{1j_2} \equiv & a_{Uj_2} + a_{Dj_2} \\ a \equiv & 2 \times a_{j_2} \frac{1 + a_{Dj_2}}{1 + a_{1j_2}} \\ b \equiv & 1 + a_{Dj} \\ c \equiv & j^- + a_{Tj} \times M_j + 2a_{Uj_2} \frac{j_2^-}{1 + a_{1j_2}} \end{aligned} \tag{5}$$

Gives for the future values of j and j_2:

$$\begin{aligned} j^+ = & \frac{-b + \sqrt{b^2 + 4ac}}{2a} \\ j_2^+ = & \frac{a_{j_2}(j^+)^2 + j_2^-}{1 + a_{1j_2}} \end{aligned} \tag{6}$$

3.2 Mechanism B: Heterodimerization

3.2.1 ODE

$$\begin{aligned}
\frac{dj}{dt} &= k_{Tj}M_j - k_{jk}j \times k + k_{Ujk}jk - k_{Dj}j \\
\frac{dk}{dt} &= k_{Tk}M_k - k_{jk}j \times k + k_{Ujk}jk - k_{Dk}k \\
\frac{djk}{dt} &= k_{jk}j \times k - k_{Ujk}jk - k_{Djk}jk
\end{aligned} \tag{7}$$

where j and k are transcription factor monomers, jk is transcription factor heterodimer, M_j and M_k are mRNA transcripts, and the k's represent rate constants, with subscripts Tj and Tk indicating translation, jk indicating heterodimerization, Ujk indicating undimerization, Dj and Dk indicating monomer degradation, and Djk indicating heterodimer degradation.

3.2.2 Closed–form expression for integration with the Implicit Euler method

Defining:

$$\begin{aligned}
a_{Tj} &\equiv k_{Tj} \times \Delta t \\
a_{Dj} &\equiv k_{Dj} \times \Delta t \\
a_{Tk} &\equiv k_{Tk} \times \Delta t \\
a_{Dk} &\equiv k_{Dk} \times \Delta t \\
a_{jk} &\equiv k_{jk} \times \Delta t \\
a_{Ujk} &\equiv k_{Ujk} \times \Delta t \\
a_{Djk} &\equiv k_{Djk} \times \Delta t
\end{aligned} \tag{8}$$

Further defining:

$$\begin{aligned}
a_{1jk} &\equiv a_{Dk} - a_{Dj} \\
a_{2jk} &\equiv a_{Ujk} + a_{Djk} \\
b_{jk} &\equiv \frac{j^- - k^- + a_{Tj}M_j - a_{Tk}M_k}{1 + a_{Dj}} \\
c_{jk} &\equiv \frac{jk^-}{1 + a_{2jk}} \\
e_{jk} &\equiv a_{jk} - a_{Ujk}\frac{a_{jk}}{1 + a_{2jk}} \\
g_{pjk} &\equiv a_{Tk}M_k + a_{Ujk}c_{jk} \\
f_{jk} &\equiv 1 + \frac{a_{1jk}}{1 + a_{Dj}} \\
a &\equiv e_{jk}f_{jk} \\
b &\equiv 1 + a_{Dk} + e_{jk}b_{jk} \\
c &\equiv k^- + g_{pjk}
\end{aligned} \tag{9}$$

Gives for the future values of j, k, and jk:

$$\begin{aligned} k^+ =& \frac{-b + \sqrt{b^2 + 4ac}}{2a} \\ j^+ =& b_{jk} + k^+ \frac{1 + a_{1jk}}{1 + a_{Dj}} \\ jk^+ =& c_{jk} + \frac{a_{jk}}{1 + a_{2jk}} j^+ k^+ \end{aligned} \tag{10}$$

3.3 Mechanism C: Ligand binding

3.3.1 ODE

$$\begin{aligned} \frac{dj}{dt} =& k_{Tj} M_j - k_{jk} j \times k + k_{Ujk} jk - k_{Dj} j \\ \frac{dk}{dt} =& S_k(t) - k_{jk} j \times k + k_{Ujk} jk - k_{Dk} k \\ \frac{djk}{dt} =& k_{jk} j \times k - k_{Ujk} jk - k_{Djk} jk \end{aligned} \tag{11}$$

where j is unbound receptor, k is free ligand, jk is the receptor–ligand complex, M_j is an mRNA transcripts for the receptor, $S_k(t)$ is the time–dependent rate at which ligand is injected into the system, and the k's represent rate constants, with subscripts Tj indicating translation, jk indicating ligand binding, Ujk indicating ligand unbinding, Dj indicating degradation of unbound receptor, Dk indicating degradation of free ligand, and Djk indicating degradation of the receptor–ligand complex.

3.3.2 Closed–form expression for integration with the Implicit Euler method

The closed form expression for integration of the ligand binding mechanism is the same as that for heterodimerization (mechanism B), except that $k_{Tk} M_k$ should be replaced by S_k.

4 Error

It is difficult to quantify precisely the error introduced by the hybrid stochastic/deterministic approximation, but there are indications that it is small for this system. The fastest reaction rate in the evolution of the mRNA levels for the single pulse study was 15 minutes^{-1} (for formation of EQ from E and Q), corresponding to a desired maximum step size of 0.0667 minutes. Only 27% of the step–sizes in the hybrid simulation for the pulse response were greater than 0.0667 minutes, with the average being 0.0521 minutes and the maximum being 0.8065 minutes.

The impact of approximating the deterministic subsystems with the Implicit Euler integrator was assessed by slightly modifying the hybrid formalism. The first three steps (calculation of reaction probabilities, determination of time step, and execution of deterministic integration) were carried out as described. Instead of executing the reactions in the fourth step according to Gillespie's Direct Method, however, the stochastic states were integrated for the length of the time step using a stiff integrator in MATLAB, while the deterministic states were held fixed. The results are shown in Figure 3, where the results of integrating the genetic regulatory network with this method are compared with those from the full deterministic simulation. Clearly the difference is very small, with the lines overlapping almost perfectly.

To demonstrate that the hybrid simulation adequately captured the stochastic nature of the system would require many runs using full stochastic simulations. Due to the size of the genetic regulatory network model, this was not feasible. The approach in the present work was to use a smaller system with similar structure. The Barkai and Leibler (2000) circadian oscillator [1] is an appropriate model for this task. Simulations for the oscillator using deterministic, full stochastic, and hybrid stochastic/ deterministic formalisms were performed. The fluctuations in the results of each were quantified by calculating the power spectral density (PSD). The width of the main peak of the PSD can be used as a measure of the noise present in a periodic system [4]. The resulting PSD plot is shown in Figure 7.

From Figure 7 it is clear that the hybrid stochastic/ deterministic integrator gave nearly identical results to the full stochastic integrator as their PSDs nearly overlap. This indicates that the hybrid stochastic/ deterministic formalism described presently may accurately capture the stochastic nature of the system.

5 Adaptation to constant input

As described in the main body of the text, the hybrid stochastic simulations yielded unexpected results for the case of an extended step in ligand concentration (Figure 3 of main text). Even though deterministic simulations predicted that prolonged ligand exposure would lead to prolonged down-regulation of genes A and C and prolonged up-regulation of genes B and D, this was true for only some of the hybrid stochastic simulations. A fraction of the hybrid stochastic simulations (*cells*) seemed to transiently adapt to the input. We have confirmed that this behavior is not an artifact of the hybrid approach by performing simulations with the full Gillespie algorithm. It arises from the fact that the transcript for the receptor, ME, at steady state in the presence of ligand will have a concentration of 0.4 molecules/cell in the deterministic simulation, which becomes 0 molecules/cell for the majority of the time in the stochastic simulations. When ME=0, there is a decrease in receptors/cell as compared to the deterministic simulation. In some cases, the receptor level becomes so low that transcription of F is reduced, transcription of MB and MD is reduced, and de-repression of A and C may occur. This general behavior is shown in Figure 5, for both hybrid stochastic/deterministic and full Gillespie stochastic simulations. It must be noted that, for both classes of simulations, there appear to be spontaneous bursts of transcription of MA, even when the level of MB has not been reduced. This result probably arises from the positive feedback regulation of the transcription of MA, which, when modified slightly becomes the circadian oscillator of Barkai and Leibler (2000) [1].

6 Distributions

The present section addresses the assumption of a Gaussian distribution of the measurements which was made in the practical identifiability analysis. When transcript levels are far from zero, the Gaussian assumption is a good one, as shown by the straight lines in the quantile–quantile (QQ) plots in Figure 5 (b, d, g) and Figure 6 (b, c, d, f). When the transcript levels are low, however, the distributions are distinctly non–Gaussian, as indicated by the curved QQ plots (Figure 5 (c, f, h) and Figure 6 (g)). The non–Gaussian nature arises largely because the transcript levels cannot take negative values.

For the case of low transcript levels, however, the variances are also small. This is important because under conditions where the transcript levels are small, the small variances most likely have a negligible effect on the parameter estimation accuracies, whether the distribution is Gaussian or

not. Thus the Fisher information matrix – based approach is still appropriate, and the calculated parameter estimation accuracies for these cases should still be expected to be faithful estimates.

In the case of the prolonged ligand step perturbation, the distributions in the transcript levels were observed to be non–Gaussian at some time points (Figure 6 (g, h)). At these times some cells are in the "HIGH" ligand state, while other cells escape to the "LOW" ligand state, leading to a very high variance about the mean behavior. Given the highly non–Gaussian nature of the distributions under these conditions, the parameter estimation accuracies obtained using the Fisher information matrix will not be rigorous lower bounds. Due to the high variance for these time points, however, their information content will be appropriately penalized and the calculated estimation accuracies should nevertheless be representative.

References

[1] N. Barkai and S. Leibler. Circadian clocks limited by noise. *Nature*, 403:267–268, 2000.

[2] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[3] M. T. Heath. *Scientific Computing.* McGraw Hill, New York, NY, 1997.

[4] D. E. Zak, F. J. Doyle, D. G. Vlachos, and J. S. Schwaber. Stochastic kinetic analysis of transcriptional feedback models for circadian rhythms. In *Proc. 40th IEEE Conf. Decision & Control*, pages 849–854, 2001b.

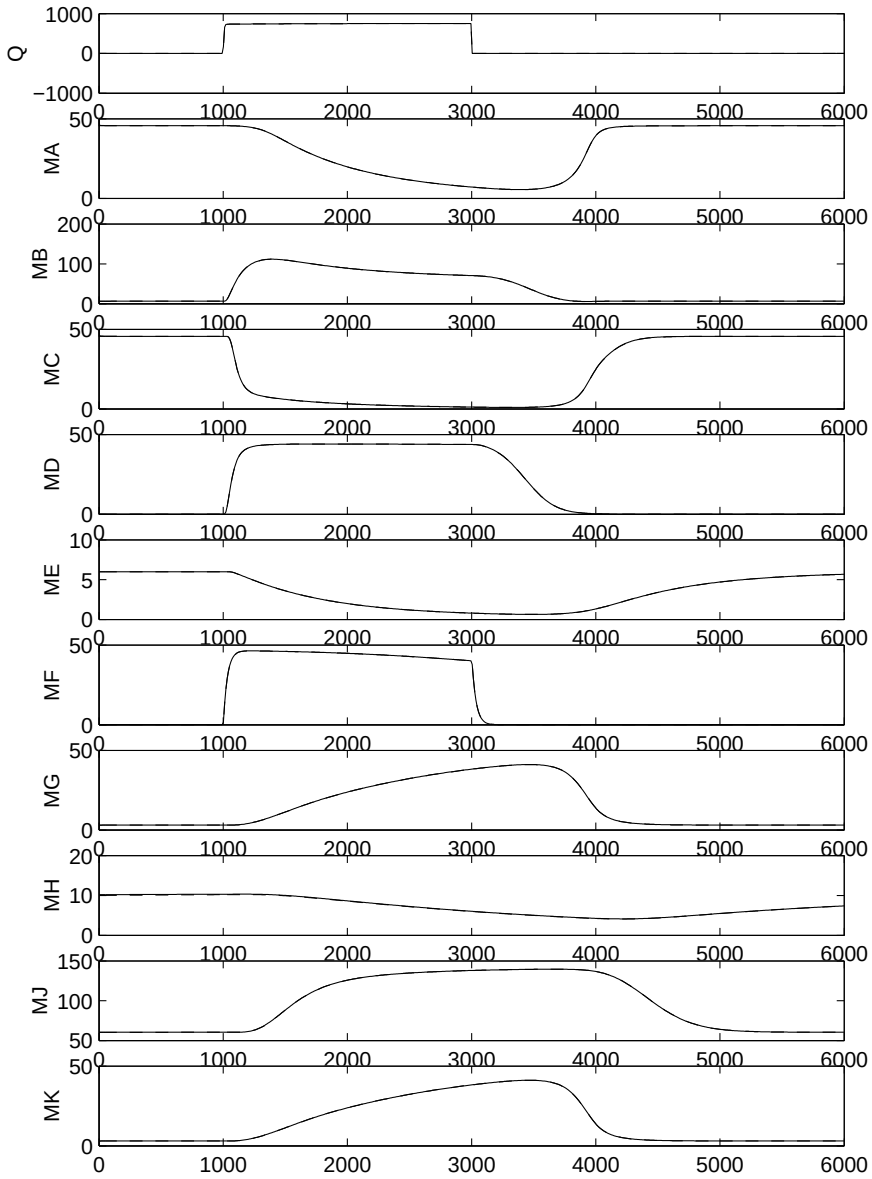

Figure 3: mRNA levels versus time for all genes for full solution and Implicit Euler approximated deterministic subsystem. Results from both simulations overlap almost perfectly, indicating that very little error in the integration of the deterministic subsystem was introduced by using the Implicit Euler method.

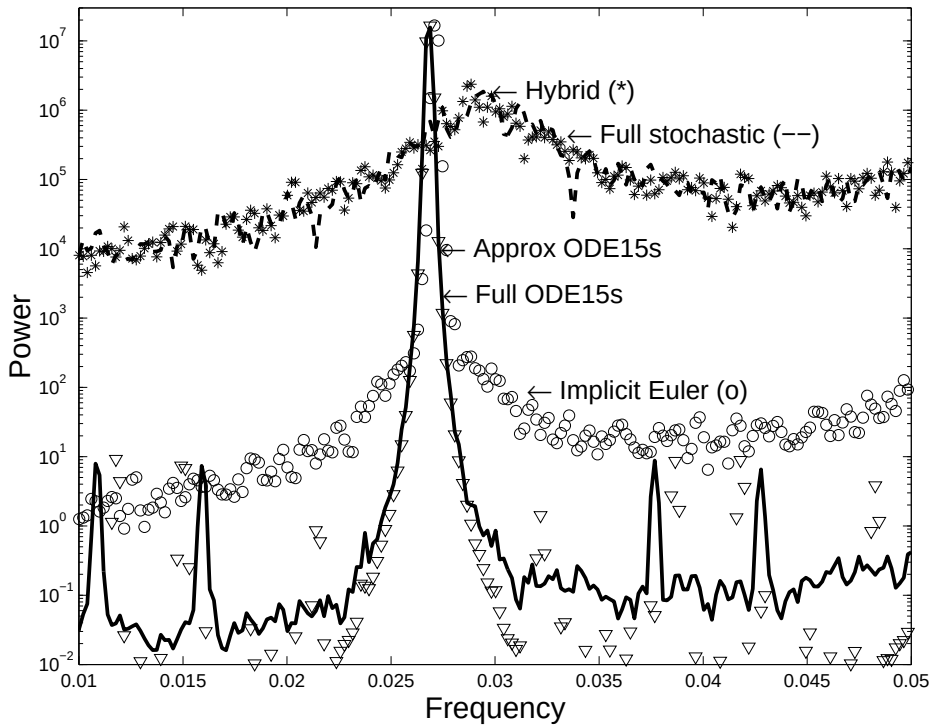

Figure 4: PSD plot (power versus frequency) for various methods of simulating the Barkain and Leibler (2000) circadian oscillator. Broader peaks indicate greater noise. The simulation methods include the full deterministic system, integrated using MATLAB ODE15s (solid line), the approximated deterministic system (assuming promoter binding by TFs does not affect TF concentration) using MATLAB ODE15s (triangles), a system where the Gillespie's Direct Method was used to determine the integration time step but integration was carried out for both subsystems using Implicit Euler integrators (to evaluate the error introduced by using Implicit Euler for integration)(circles), the full stochastic system using Gillespie's Direct Method (dashed line), and the hybrid stochastic/deterministic method (stars). Clearly the approximate deterministic approaches introduce little error as compared to the full deterministic simulation, while the hybrid stochastic/deterministic approach retains the noise character of the full stochastic simulation.

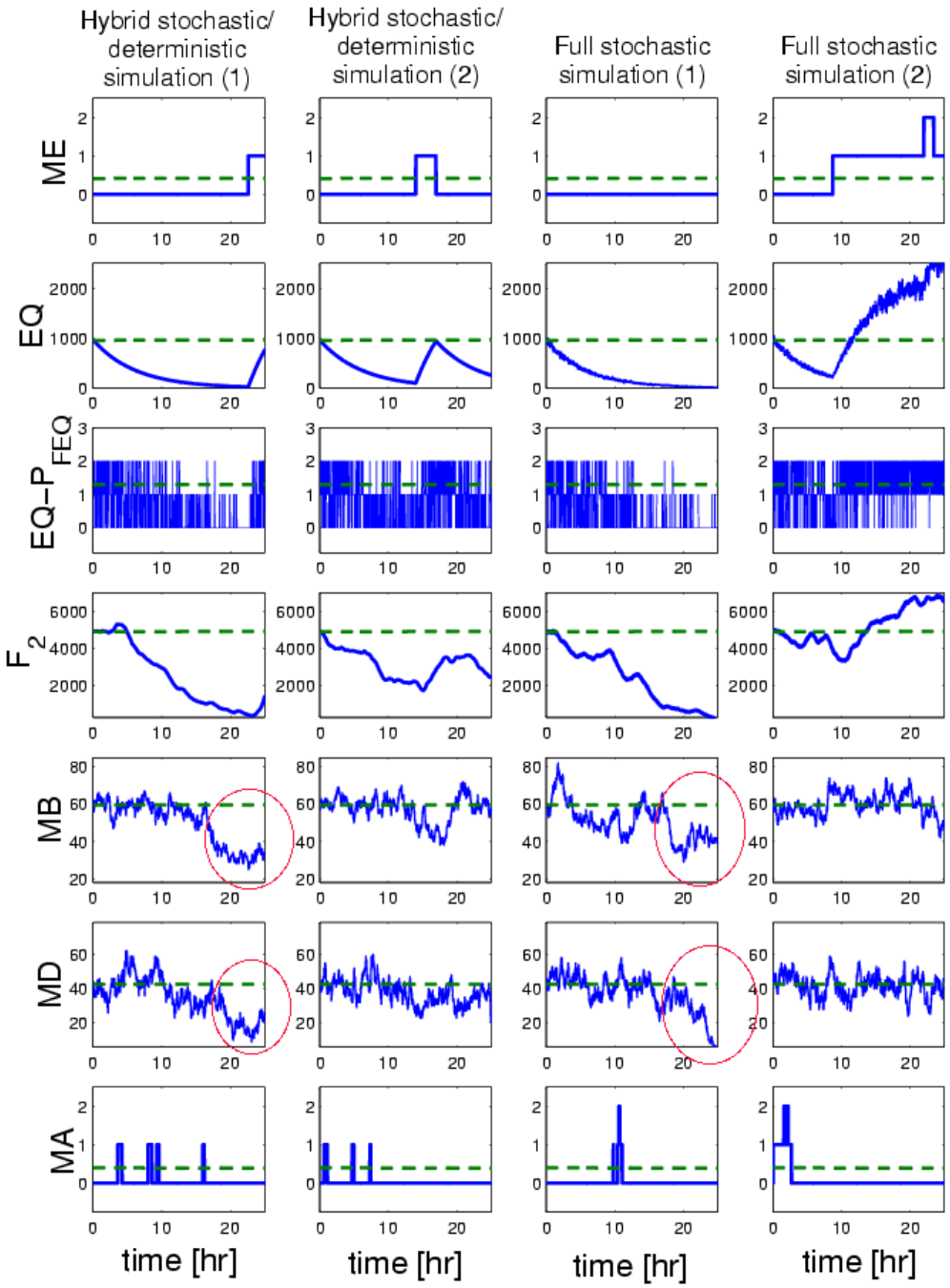

Figure 5: Demonstration of how fluctuations in receptor level can lead to transient adaptation of target genes in response to constant ligand input, for both hybrid stochastic/deterministic and full Gillespie stochastic simulations. **Columns 1 and 3**: Hybrid and full stochastic simulations that show transient adaptation, respectively. When the receptor transcript level (ME) becomes zero, the amount of available receptor E drops, leading to a decrease in EQ, the receptor-ligand complex that activates transcription of MF, as seen by a decrease in the amount of complexes between EQ and the F promoter ($EQ - P_{FEQ}$). Decreased transcription of MF leads to decreased F_2 transcription factor, which leads to decreased transcription of MD and MB (circles). If prolonged, decreased transcription of MD and MB can lead to de–repression of MA and MC (not shown). **Columns 2 and 4**: examples of hybrid and full stochastic simulations where sufficient ME is transcribed to prevent transient adaptation to the ligand input. Note for all cases that transient bursts of transcription of MA appear to occur spontaneously.

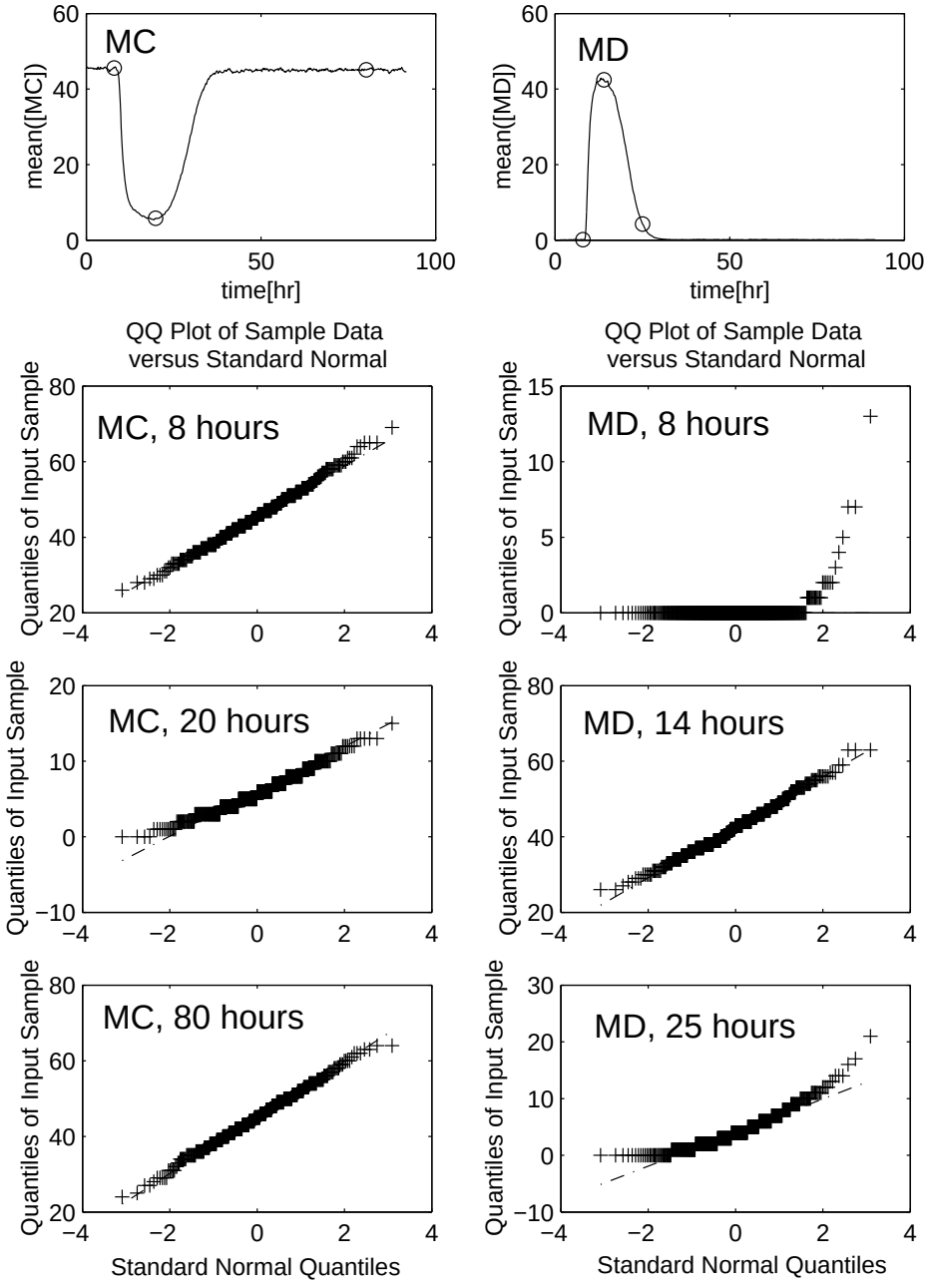

Figure 6: Quantile–quantile plots for MC and MD at different times in the single pulse perturbation time course. Straight lines indicate an approximately Gaussian distribution.

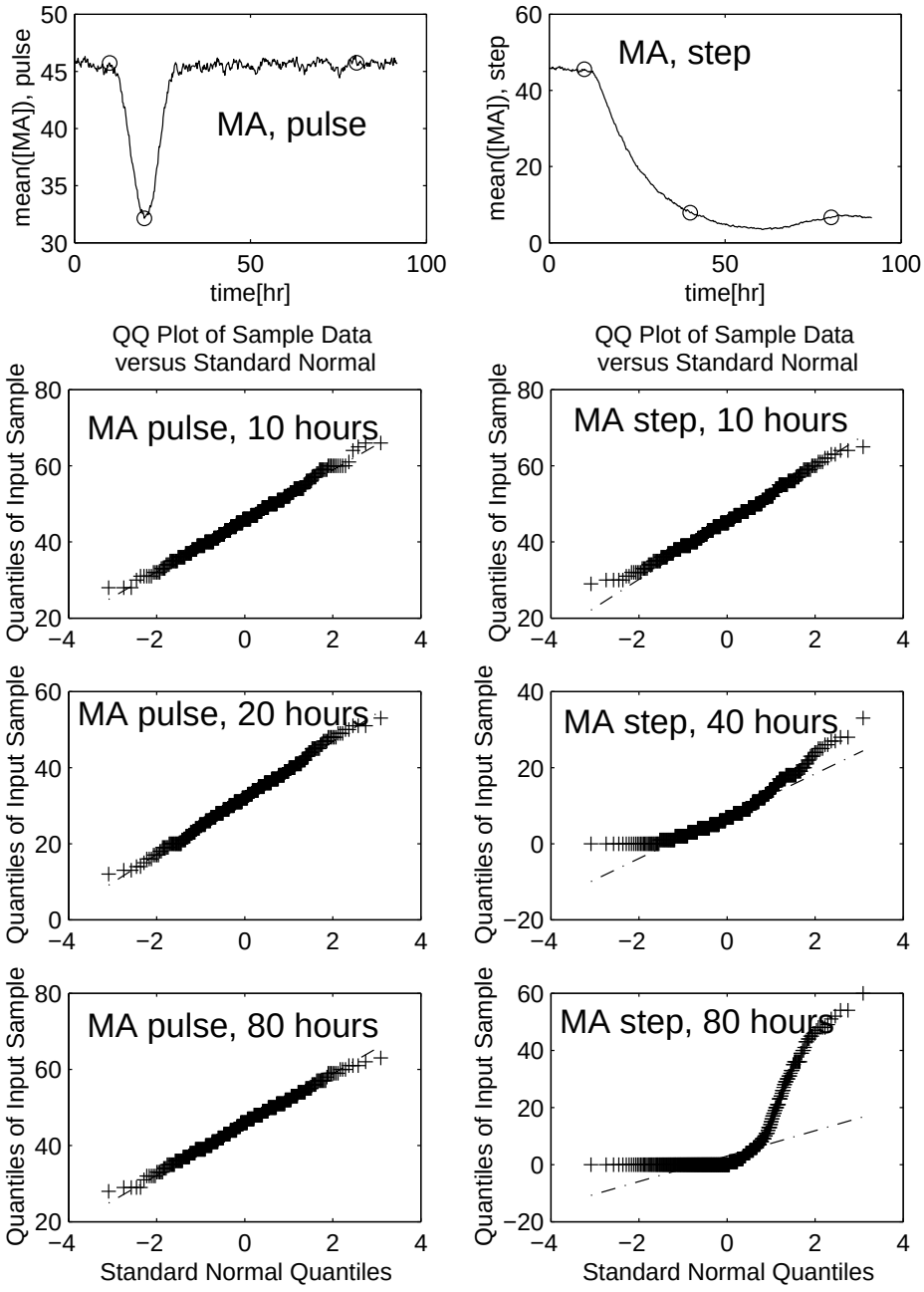

Figure 7: Quantile–quantile plots for MA at different times in the single pulse and single step perturbation time courses. Straight lines indicate an approximately Gaussian distribution.