Parameter estimation in stochastic biochemical reactions

S. Reinker, R.M. Altman and J. Timmer

Abstract: Gene regulatory, signal transduction and metabolic networks are major areas of interest in the newly emerging field of systems biology. In living cells, stochastic dynamics play an important role; however, the kinetic parameters of biochemical reactions necessary for modelling these processes are often not accessible directly through experiments. The problem of estimating stochastic reaction constants from molecule count data measured, with error, at discrete time points is considered. For modelling the system, a hidden Markov process is used, where the hidden states are the true molecule counts, and the transitions between those states correspond to reaction events following collisions of molecules. Two different algorithms are proposed for estimating the unknown model parameters. The first is an approximate maximum likelihood method that gives good estimates of the reaction parameters in systems with few possible reactions in each sampling interval. The second algorithm, treating the data as exact measurements, approximates the number of reactions in each sampling interval by solving a simple linear equation. Maximising the likelihood based on these approximations can provide good results, even in complex reaction systems.

1 Introduction

The goal of the recently emerging interest in modelling of biological processes is to understand and describe quantitively the dynamics of living cells. In order to attain this goal, new experimental procedures and modelling techniques are needed to generate and analyse relevant biological data. The latest experimental imaging techniques enable researchers to measure biochemical processes at the molecular level, including small molecule counts of reactants. In systems with only a few molecules, classic mass action kinetics are no longer valid for describing the reaction dynamics. Therefore the usual approach to estimating reaction rates by fitting solutions of law of mass action differential equations to reactant concentrations is not appropriate.

Instead, reaction rates are determined based on the probability of collision and reaction of individual molecules. In this article, we present two algorithms for estimating the kinetic reaction constants from time series of molecule counts measured with error. This work is a contribution to the solution of a core problem in systems biology: the identification of the structure and the dynamics of biochemical networks.

The importance of noise in cell dynamics has been pointed out in numerous recent articles. In many cases, when small numbers of molecules react in a stochastic manner, macroscopic concepts such as chemical concentration cannot be used to describe the stochastic dynamics. One of the earliest prominent stochastic models of cellular dynamics is the lambda phage switch [1] that determines the outcome of a viral infection. Another example is the observed phenomenon of individual differences among genetically identical bacteria, which is thought to be based on varying stochastic gene expression [2]. Molecular noise also can lead to behavioural variability in bacteria [3]. Through a stochastic resonance-like phenomenon, cells can use stochastic fluctuations to enhance their sensitivity to external signals [4]. A stochastic switch in enzyme kinetics was described by Samoilov *et al.* [5].

In gene regulatory networks, noise is believed to play an important role because only a small number of mRNA molecules are generated in transcription [6, 7]. Also, transcription factor binding and gene expression are highly stochastic and likely the original sources of cellular noise [8]. Paulsson [9] gave an exhaustive review of the different sources of noise that can be separated into intrinsic and extrinsic sources, as investigated theoretically by Swain *et al.* [10] and experimentally by Elowitz *et al.* [11]. Gene regulation networks can be subdivided into modules and motifs [12] that perform different functions. For example, feedback loops can dampen and control the amount of noise [13] and thus increase reliability of signal transduction. Thattai *et al.* [14], as well as Kierzek *et al.* [15], suggested in modelling studies that noise in gene expression arises mainly at the translational level, whereas Kepler and Elston [16] investigated the origin and characteristics of noise from transcriptional processes. A recent experimental study of rates of gene transcription by fluorescence methods in single cells demonstrated the stochastic nature of mRNA production, which occurs in bursts when polymerase is bound to DNA [17].

Even though stochastic effects are present in all aspects of cellular dynamics, they can be detrimental to processes that are vital for the cell's survival. For example, the circadian clock, which many living systems

IEE Proceedings online no. 20050105

doi:10.1049/ip-syb:20050105

Paper first received 15th December 2005 and in revised form 9th May 2006

S. Reinker and J. Timmer are with the Department of Mathematics and Physics, FDM, University of Freiburg, Hermann-Herder-Str. 3, 79104 Freiburg, Germany

R.M. Altman is with the Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada, V5A 1S6

E-mail: reinker@physik.uni-freiburg.de

possess, must be robust against random molecular events. Small numbers of molecules are involved in the clock mechanism, and Barkai and Leibler [18] showed that the accuracy of the generated rhythm is limited by the amount of noise.

There is evidence that cells have adjusted to the stochasticity inherent in biochemical reactions and have developed strategies to control it and even use it to their advantage [19]. Noise seems to be minimised in the expression of genes that are critically important to cell function [20]. One mechanism of noise attenuation in gene expression may be post-transcriptional control [21], resulting in reliable expression of important genes. Shibata and Fujimoto [22] discussed the limitations of faithful signal transduction by intrinsic and extrinsic noise.

Most previous studies of parameter identification in regulatory networks have dealt only with deterministic models. In signal transduction networks, the estimation of kinetic parameters can be performed by local and global optimisation methods [23, 24] and Bayesian networks [25]. In differential equation models of gene networks, Perkins *et al.* [26] estimated parameters from expression data. The connection between stochastic and deterministic effects in biochemical reaction networks was discussed by Puchalka and Kierzek [27].

For the forward problem of simulating stochastic chemical systems, Gillespie [28] introduced a now widely used algorithm. In most theoretical studies that employ this method, parameter values are set to biologically plausible levels or estimated from in vitro experiments on the macroscopic level [16, 27, 29]. For the inverse problem of obtaining parameters, kinetic rate constants are normally estimated by fitting law-of-mass action dynamics to experimentally measured concentration curves. The stochastic rate constants can be derived from these deterministic reaction rate constants [30]. However, this is not possible in many biological systems where the components cannot be isolated from cells. For example, genetic transcription and regulation mechanisms rely on the whole machinery of the cell and cannot be studied separately. Thus, these small-scale reactions are accessible only by in situ measurement techniques, for example, fluorescence methods [31].

Some recent studies have attempted to estimate stochastic rate constants from exact molecule count data, assuming that the counts follow a Markov chain (as per Gillespie [28]). Becskei *et al.* [8] fitted a log-normal error model to stochastic gene expression data, whereas Golightly and Wilkinson [32] used a stochastic differential equation approximation. Boys *et al.* [33] demonstrated that Bayesian inference methods can give estimates of stochastic reaction constants for both continuously and discretely sampled data. These authors also explored parameter estimation in data-poor scenarios when only some of the involved species can be measured. However, we are not aware of any literature on the problem of estimating the reaction parameters in stochastic systems with small molecule counts which are measured with error.

In this article, we propose two different algorithms for stochastic kinetic parameter estimation from data with observation errors, assuming that the (unobserved) molecule counts follow a Markov chain. The resulting model is a hidden Markov model (HMM), where the hidden states are the true molecule counts, and transitions between these states correspond to reaction events following the collision of molecules.

Our first algorithm follows the classical approach of maximum likelihood estimation of HMMs, and specifies an analytic expression for the likelihood. As discretely sampled data contain information only about the change in molecule counts – not the order or type of reactions that occurred during the sampling interval – this expression is a complicated sum of the probabilities of each possible sample path. Therefore evaluation and maximisation are computationally expensive. We demonstrate that this method can be used to estimate the parameters in simple systems where the state space is low dimensional, and the reaction rates are slow, relative to the sampling rate.

Our second algorithm treats the data as exact measurements and approximates the number of reactions in each sampling interval by solving a simple linear equation. Parameter estimates are then obtained by maximising the likelihood based on these values. As the true number of molecules at each time point is assumed to follow a Markov chain, this likelihood is much simpler, and hence much faster to evaluate and maximise than the HMM likelihood. Simulation studies show that this method can give good results even with measurement error.

2 Definition of the system

For a fixed time point, let X_n, $n = 1, \ldots, N$, denote the number of molecules of species n. A general system of (bio-)chemical reactions can be written in typical chemistry stoichiometric notation as

$$l_{\mu 1}X_1 + l_{\mu 2}X_2 + \cdots \longrightarrow r_{\mu 1}X_1 + r_{\mu 2}X_2 + \ldots \quad (1)$$

Here, the coefficients $l_{\mu n}$ and $r_{\mu n}$ denote the numbers of molecules of species n consumed and produced, respectively, in reaction μ, $\mu = 1, \ldots, M$. Let $l_\mu = (l_{\mu 1}, \ldots, l_{\mu N})$ and $r_\mu = (r_{\mu 1}, \ldots, r_{\mu N})$ so that a reaction of type μ changes the numbers of molecules by $\nu_\mu \equiv -l_\mu + r_\mu$.

In reaction systems, the number of molecules of each species evolves over time. Let $\boldsymbol{X}(t) = (X_1(t), \ldots, X_N(t))$, where $X_n(t)$, $n = 1, \ldots, N$, denotes the number of molecules of species n at time t. As argued by Gillespie [28], the process $\boldsymbol{X}(t)$ follows a continuous-time Markov chain, with changes in state determined by reactions among molecules. In particular, it is a jump process with state-dependent transition probabilities.

Chemical reactions can occur if molecules collide, and hence the rate of reaction of type μ at a time t depends on the state of the system at this time. In particular, the probability that reaction μ results from a collision is a function of a reaction constant c_μ and the number of possible collisions between molecules of the species involved. The values $\boldsymbol{c} = (c_1, \ldots, c_M)$ are similar to deterministic law-of-mass-action reaction constants. The propensity for reaction μ at time t is then

$$a_\mu = c_\mu \prod_n [X_n(t)]^{l_{\mu n}} \quad (2)$$

where $a_\mu \, \mathrm{d}\tau$ is the probability that reaction μ will occur in the time interval $[t, t + \mathrm{d}\tau]$, given that the system is in state $\boldsymbol{X}(t)$ at time t [28].

In this work, we assume a 'well-stirred' medium, that is there are no spatial effects such as local accumulation of molecules or metabolic channelling. Then, the joint distribution of the waiting time until the next reaction (t) and the type of reaction that occurs (μ) is

$$P(\mu, t) = a_\mu \exp(-a_0 t) \quad (3)$$

where $a_0 = \sum_\mu a_\mu$ is the overall reaction propensity.

In practice, it is not possible to measure $\boldsymbol{X}(t)$ exactly and continuously in time. However, with new experimental

techniques such as fluorescence methods, it is possible to measure small numbers of molecules (with some error) at discrete time points. Denote the observed measured count of species n at time t_i by $O_n(t_i)$, $t_0 = 0$, $t_1, \ldots, t_R$, and let $\boldsymbol{O}(t_i) = (O_1(t_i), \ldots, O_N(t_i))$.

We assume that $O_n(t_i) = X_n(t_i) + \epsilon_n(t_i)$, and that the measurement errors $\epsilon_n(t_i)$ are i.i.d. Gaussian distributed with mean 0 and variance σ. This assumption implies that, conditional on $X_n(t_i)$, $O_n(t_i)$ is independent of all other observations and all other molecule counts. In other words, the process $\boldsymbol{O}(t_i)$ follows a HMM.

Finally, in this article, we consider only the case where the sampled time points are equally spaced at intervals of length Δt, although this assumption could easily be relaxed. In this case, the unobserved process $\boldsymbol{X}(t_i)$ is a time-homogeneous Markov chain.

We present two simple examples of such reaction systems with the goal of estimating the reaction parameters $c_1, \ldots, c_M$, and σ from the observed data.

2.1 Example: a simple gene expression model

As a simple example, we use the following simple system of stochastic gene transcription, as investigated and modelled by Golding *et al.* [17]. In *E. coli*, a green fluorescent protein tag was introduced to create fluorescent mRNA from a specific gene. The resulting fluorescence can be measured in individual cells and converted into mRNA count data. Because there is only one copy of the tagged DNA sequence, as well as a small number of polymerase proteins that initiate transcription in each cell, the transcription process is highly stochastic with very different time courses of mRNA copy numbers in different cells [17]. Using high-resolution fluorescence microscopy, the experimenters were able to measure mRNA copy numbers with good accuracy (error <1 molecule) and at high sampling rates (2 measurements/min).

We use the model described by Golding *et al.* [17] to describe the gene transcription process. The DNA for the tagged mRNA is switched on and off by polymerase binding and unbinding, respectively. Only polymerase-bound DNA is transcribed into mRNA that consequently grows in count monotonically. The system takes the form

$$\begin{aligned} R_1&: \quad \mathrm{DNA_{OFF}} \rightarrow \mathrm{DNA_{ON}} \\ R_2&: \quad \mathrm{DNA_{ON}} \rightarrow \mathrm{DNA_{OFF}} \\ R_3&: \quad \mathrm{DNA_{ON}} \rightarrow \mathrm{DNA_{ON}} + \mathrm{mRNA} \end{aligned}$$

The reaction rates associated with R_1, R_2 and R_3 are c_1, c_2 and c_3, respectively. During a reaction of type μ, the state $\boldsymbol{X} = (X_1, X_2, X_3) = (\mathrm{DNA_{OFF}}, \mathrm{DNA_{ON}}, \mathrm{mRNA})$ changes by $\nu_1 = (-1, 1, 0)$ if R_1 occurs, $\nu_2 = (1, -1, 0)$ if R_2 occurs and $\nu_3 = (0, 0, 1)$ if R_3 occurs.

As initial conditions in this system, we set $\mathrm{DNA_{OFF}} = 1$, $\mathrm{DNA_{ON}} = 0$ and $\mathrm{mRNA} = 0$, and as parameter values $c_1 = 0.0270\ \mathrm{min}^{-1}$, $c_2 = 0.1667\ \mathrm{min}^{-1}$ and $c_3 = 0.40\ \mathrm{min}^{-1}$ [17]. Fig. 1 shows mRNA counts of different realisations of this process at a fixed sampling rate, $\Delta t = 0.5$ min. When DNA is switched on by polymerase binding, transcription of the gene into mRNA starts, and the mRNA count grows. Transcription is terminated by unbinding of polymerase and switching off of DNA, leading to burst transcription with long intervals of constant mRNA counts.

2.2 Example: a transcriptional regulatory system

As an example of a more complex biochemical stochastic reaction system, we use a model of transcriptional regulation that was proposed by Goutsias [34]. Here, the mRNA is translated into a protein monomer M that can dimerise. The dimer D, in turn, can bind to its DNA and act as a transcription factor to autoregulate its own mRNA production. Both mRNA and protein monomers are degraded at constant rates.

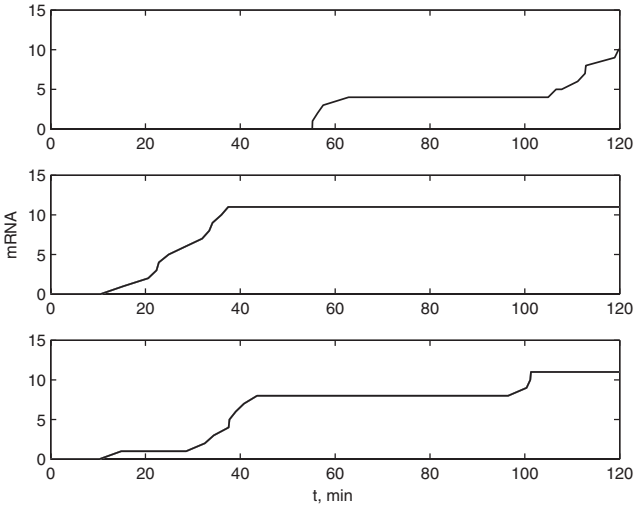

Fig. 1 *Three realisations (without measurement error) of Golding's stochastic gene expression model, showing the mRNA counts as can be measured experimentally by fluorescent tagging*

The stochastic model of our version of the Goutsias transcriptional regulation system consists of eight reactions involving five molecule species

$$\begin{aligned} R_1&: \quad \mathrm{mRNA} \rightarrow \mathrm{mRNA} + M \\ R_2&: \quad M \rightarrow \varnothing \\ R_3&: \quad \mathrm{DNA} \cdot D \rightarrow \mathrm{mRNA} + \mathrm{DNA} \cdot D \\ R_4&: \quad \mathrm{mRNA} \rightarrow \varnothing \\ R_5&: \quad \mathrm{DNA} + D \rightarrow \mathrm{DNA} \cdot D \\ R_6&: \quad \mathrm{DNA} \cdot D \rightarrow \mathrm{DNA} + D \\ R_7&: \quad 2M \rightarrow D \\ R_8&: \quad D \rightarrow 2M \end{aligned}$$

Here we neglect the double binding and unbinding of dimers to DNA in the original model, which has a very low reaction rate ($0.8765 \times 10^{-11}\ \mathrm{s}^{-1}$ by Goutsias [34]). Such reactions with such low rates are hardly observed in short time series, and consequently the associated rates cannot be estimated reliably with any algorithm. As initial values we use $M = 2$, $D = 4$, $\mathrm{DNA} = 2$, and $\mathrm{mRNA} = 0$, and dimer-bound DNA, $\mathrm{DNA} \cdot D = 0$. The parameter values are $\boldsymbol{c} = (0.043, 0.0007, 0.0715, 0.00395, 0.02, 0.4791, 0.083, 0.5)$, with all individual reaction constants in units of per second [34].

Fig. 2 shows a plot of the stochastic behaviour of M, D and mRNA. As in Golding's system, RNA is produced in a stochastic manner when the dimer is bound to DNA. The protein monomers produced by RNA translation dimerise and de-dimerise at high rates, leading to fluctuating molecule counts.

3 Parameter estimation

For estimation of reaction parameters $c_1, \ldots, c_m$, and noise strength σ from time series of molecule counts observed with error, we develop two different algorithms.

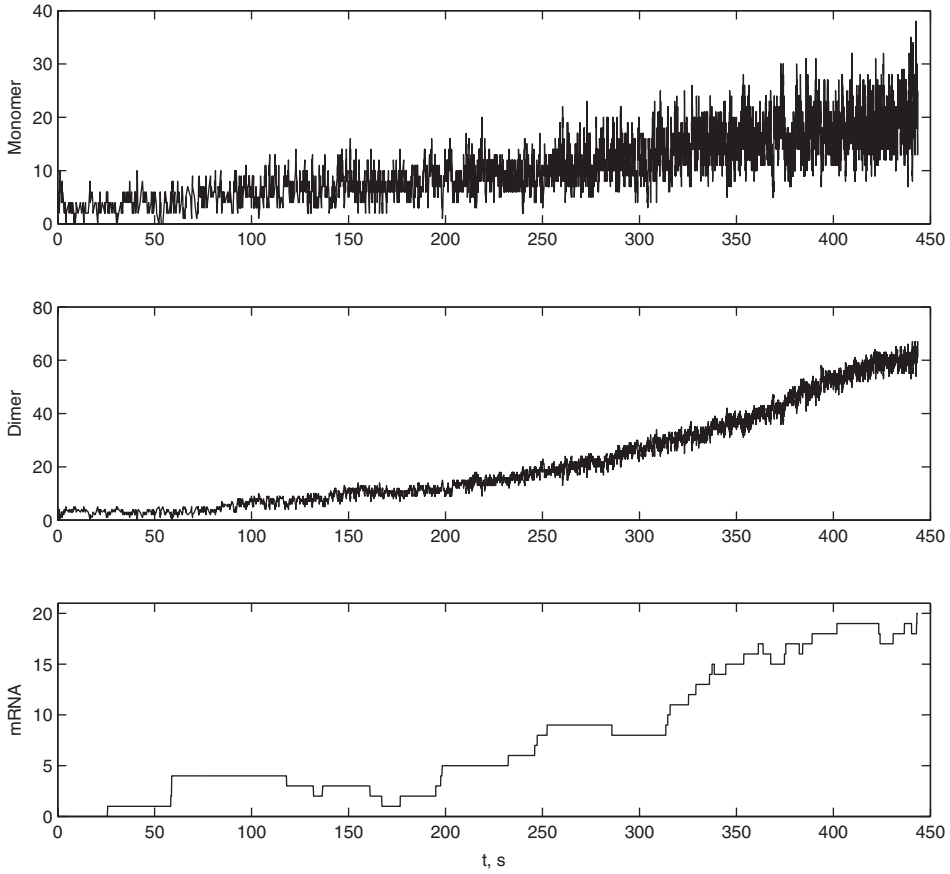

Fig. 2 *Sample trace of a realisation of the transcriptional regulatory system, measured without error*

3.1 Approximate maximum likelihood (AML) method

The likelihood of the observation given a set of parameter values can be computed as follows (using f generically to represent a density)

$$\begin{aligned}\mathcal{L} &= f(\boldsymbol{O}(t_1), \ldots, \boldsymbol{O}(t_R)) \\ &= \sum_{\boldsymbol{x}_1} \cdots \sum_{\boldsymbol{x}_R} f(\boldsymbol{O}(t_1), \ldots, \boldsymbol{O}(t_R)|\boldsymbol{X}(t_1) \\ &\quad = \boldsymbol{x}_1, \ldots, \boldsymbol{X}(t_R) = \boldsymbol{x}_R)\mathrm{P}(\boldsymbol{X}(t_1) = \boldsymbol{x}_1, \ldots, \boldsymbol{X}(t_R) = \boldsymbol{x}_R) \\ &= \sum_{\boldsymbol{x}_1} \cdots \sum_{\boldsymbol{x}_R} \left\{ \prod_{i,n} f(O_n(t_i)|X_n(t_i) = x_{in}) \right\} \mathrm{P}(\boldsymbol{X}(t_1) = \boldsymbol{x}_1) \\ &\quad \times \prod_{i=2}^{R} \mathrm{P}(\boldsymbol{X}(t_i) = \boldsymbol{x}_i|\boldsymbol{X}(t_{i-1}) = \boldsymbol{x}_{i-1})\end{aligned}$$

This summation is over a prohibitively large number of terms. However, the special structure of the HMM allows us to simplify this expression. In particular, letting $\pi_{\boldsymbol{x}_1} = \mathrm{P}(\boldsymbol{X}(t_1) = \boldsymbol{x}_1)$ and $P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} = \mathrm{P}(\boldsymbol{X}(t_i) = \boldsymbol{x}_i|\boldsymbol{X}(t_{i-1}) = \boldsymbol{x}_{i-1})$, we can write

$$\begin{aligned}\mathcal{L} = &\sum_{\boldsymbol{x}_1} \pi_{\boldsymbol{x}_1} \prod_n f(O_n(t_1)|X_n(t_1) = x_{1n}) \\ &\times \sum_{\boldsymbol{x}_2} P_{\boldsymbol{x}_1,\boldsymbol{x}_2} \prod_n f(O_n(t_2)|X_n(t_2) = x_{2n}) \cdots \\ &\sum_{\boldsymbol{x}_R} P_{\boldsymbol{x}_{R-1},\boldsymbol{x}_R} \prod_n f(O_n(t_R)|X_n(t_R) = x_{Rn}) \qquad (4)\end{aligned}$$

This expression is equivalent to a product of matrices, and hence is easy to compute if N is not too large and the state space of $\boldsymbol{X}(t_i)$ is limited.

In general, $\boldsymbol{X}(t_i)$ can take on any positive integer value (i.e. its state space is infinite), so evaluation and maximisation of the likelihood is more challenging. For this reason, we make two approximations:

1. We take $f(O_n(t_i)|X_n(t_i) = x_{in}) = 0$ for $|O_n(t_i) - x_{in}| > s\sqrt{\sigma}$ for some integer number s.
2. We assume that a fixed maximum number reactions can occur per time interval.

Both approximations effectively reduce the dimension of the summation over $\boldsymbol{x}_i$. In particular, the first restricts the range of $\boldsymbol{X}(t_i)$, given $\boldsymbol{O}(t_i)$. This is equivalent to assuming that the measurement error follows a truncated Gaussian, rather than a Gaussian distribution. The second approximation relies on the assumption that $P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} = 0$ for most pairs $(\boldsymbol{x}_{i-1}, \boldsymbol{x}_i)$. In other words, under this approximation, the matrices in (4) are sparse so that their product is easy to compute.

Then we can compute $P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i}$ by determining which reactions could have occurred in the interval $(t_{i-1}, t_i]$. Under the assumption of no more than two reactions per time interval, for example, we apply the following results:

Case 1: No reaction takes place in the interval $(t_{i-1}, t_i]$. Then $\boldsymbol{X}(t_i) = \boldsymbol{X}(t_{i-1})$ and

$$\begin{aligned}\mathrm{P}(\text{no reaction in } (t_{i-1}, t_i]) &= 1 - \sum_{\mu} \int_0^{\Delta t} a_\mu \mathrm{e}^{-a_0\tau}\,\mathrm{d}\tau \\ &= \mathrm{e}^{-a_0\Delta t}\end{aligned}$$

where $\Delta t = t_i - t_{i-1}$.

Case 2: Exactly one reaction of type μ occurs in the interval $(t_{i-1}, t_i]$. Then $\boldsymbol{X}(t_i) = \boldsymbol{X}(t_{i-1}) + \nu_\mu$ and

$$\begin{aligned}
&\text{P(only reaction } \mu \text{ in } (t_{i-1}, t_i]) \\
&= \int_0^{\Delta t} a_\mu e^{-a_0\tau} \text{P(no reaction in } (\tau, t_i])\, d\tau \\
&= \int_0^{\Delta t} a_\mu e^{-a_0\tau} e^{-a_0'(\Delta t-\tau)}\, d\tau \\
&= \frac{a_\mu}{a_0 - a_0'}\left[e^{-a_0'\Delta t} - e^{-a_0\Delta t}\right] \qquad (5)
\end{aligned}$$

Here, a_0' denotes the changed propensity after a reaction of type μ.

Case 3: Exactly two reactions of types μ_1 and μ_2 occur, in that order, in the interval $(t_{i-1}, t_i]$. Then, $\boldsymbol{X}(t_i) = \boldsymbol{X}(t_{i-1}) + \nu_{\mu_1} + \nu_{\mu_2}$ and

$$\begin{aligned}
&\text{P(reaction } \mu_1 \text{ and then } \mu_2 \text{ in } (t_{i-1}, t_i]) \\
&= \int_0^{\Delta t} a_{\mu_1} e^{-a_0\tau} \int_\tau^{\Delta t} \text{P(only reaction } \mu_2 \text{ in } (s, t_i])\, d\tau \\
&= \int_0^{\Delta t} a_{\mu_1} e^{-a_0\tau} \frac{a'_{\mu_2}}{a_0' - a_0''}\left[e^{-a_0''(\Delta t-\tau)} - e^{-a_0'(\Delta t-\tau)}\right] d\tau \\
&= \frac{a_{\mu_1} a'_{\mu_2}}{a_0' - a_0''}\left[\frac{e^{-a_0''\Delta t} - e^{-a_0\Delta t}}{a_0 - a_0''} - \frac{e^{-a_0'\Delta t} - e^{-a_0\Delta t}}{a_0 - a_0'}\right]
\end{aligned}$$

Here, a'_{μ_2} and a_0' again denote the changed propensities after reaction μ_1, and a_0'' denotes the overall reaction propensity after reactions μ_1 and μ_2.

In the case that a_0 does not change during a reaction, these formulas simplify considerably. For example, the formulas for one and two reactions in an interval reduce to $a_\mu \exp(-a_0\Delta t)\Delta t$ and $a_{\mu 1} a_{\mu 2} \exp(-a_0\Delta t)\Delta t^3/6$, respectively.

We can now compute

$$\begin{aligned}
P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} &= \text{P(no reaction in } \Delta t)\mathbf{1}(\boldsymbol{x}_i = \boldsymbol{x}_{i-1}) \\
&\quad + \sum_\mu \text{P(only reaction } \mu \text{ in } \Delta t)\mathbf{1}(\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \nu_\mu) \\
&\quad + \sum_{\mu_1}\sum_{\mu_2} \text{P(reaction } \mu_1 \text{ and then } \mu_2 \text{ in } \Delta t) \\
&\quad \times \mathbf{1}(\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \nu_{\mu_1} + \nu_{\mu_2})
\end{aligned}$$

where $\mathbf{1}(\cdot)$ is an indicator function equal to 1 if the argument is true and 0 otherwise.

These approximations allow us to evaluate the likelihood. Underflow errors may arise in the case of long time series. For this reason, we store intermediate computational results in scientific notation. We then use a numerical maximisation routine (e.g. the quasi-Newton routine) to find the maximum likelihood estimates (MLEs).

It should be noted that our assumption of a fixed limited number of reactions per time interval is valid only when the sampling rate is high relative to the reaction rates. Typically, this condition will be satisfied only for small systems. The assumed maximum number of reactions per time interval can be adjusted depending on the sampling and reaction rates of the system of interest; however, the cost is increased complexity in the computation of $P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i}$.

In the implementation of our algorithm, as the reaction rate parameters and σ are all non-negative, we transform the parameters and maximise the log-likelihood over these transformed values to simplify computations. We define $c_i^* = \log c_i$ and $\sigma^* = \log \sigma$, and re-write the log-likelihood in terms of c_i^* and σ^*. As the range of c_i^* and σ^* is the entire real line, maximising over these parameters is easier than maximising over the original parameters. By the invariance property of the MLEs, $\hat{c}_i = e^{\hat{c}_i^*}$ and $\hat{\sigma} = e^{\hat{\sigma}^*}$.

One advantage of the AML method is that, if our simplifying assumptions are valid, the MLEs are guaranteed to be consistent and asymptotically normal, and the diagonal entries of the negative inverse of the observed Fisher information matrix (the matrix of second derivatives of the maximised log-likelihood) provide approximate variances [35].

3.2 Singular value decomposition likelihood (SVDL) method

In this section, we develop an algorithm that works for large systems, when many reactions occur in each time interval. This algorithm works for some small systems as well. The key idea is that the observations are treated as exact molecule counts without error, so that the number of reactions in each time interval can be approximated by solving a linear equation. Then, parameter estimates can be obtained by maximising the likelihood based on these values.

First, consider the case where the observations truly are measured without error, that is $\boldsymbol{O}(t_i) = \boldsymbol{X}(t_i)$. Then, the difference between $\boldsymbol{X}(t_i)$ and $\boldsymbol{X}(t_{i+1})$ is necessarily a linear combination of the ν_μs, $\mu = 1, \ldots, M$, with non-negative integer coefficients. The coefficients indicate how many times each type of reaction has occurred during the interval.

Let ν be the matrix with jth row given by ν_j, $j = 1, \ldots, M$. We then solve the linear equation

$$(\boldsymbol{X}(t_{i+1}) - \boldsymbol{X}(t_i))^{\mathrm{T}} = \nu^{\mathrm{T}} \boldsymbol{C}^i \qquad (6)$$

for $\boldsymbol{C}^i \in \mathbb{N}^M$, where $\boldsymbol{C}^i$ is the M-dimensional column vector of coefficients with jth entry C_j^i.

The uniqueness of $\boldsymbol{C}^i$ cannot be assumed automatically. For example, if both a reaction and its reverse can occur, the number of reactions will not be uniquely determined by $\boldsymbol{X}(t_i)$ and $\boldsymbol{X}(t_{i+1})$, as multiple forward and backward reactions may have occurred. In such a system, the rates of these reactions may be underestimated. Even if the matrix does not have full rank, so that the solution is not uniquely determined, a singular value decomposition will give a solution of minimum norm. Because only non-negative numbers of reactions are possible, we may have to add an appropriate vector from the nullspace of ν in order to obtain the minimum norm non-negative solution.

From the solution $\boldsymbol{C}^i$, $P_{\boldsymbol{x}_i,\boldsymbol{x}_{i+1}}$ can be computed. Note that the observed data do not give information on the order of the reaction events, so all different permutations of the reaction order must be considered. Then exact computation of the probability of such events becomes complicated as the number of reactions in the sampling interval increases.

One way of circumventing this problem is to assume that a_μ, $\mu = 1, \ldots, M$, is constant during each sampling interval. (See Gillespie [36] for a similar assumption about the forward problem.) This assumption is reasonable for systems with many molecules. We can now calculate the probability that certain reactions occur in a sampling interval using this singular value decomposition approach.

Case 1: $\boldsymbol{X}(t_i) = \boldsymbol{X}(t_{i-1})$. We assume that no reaction takes place in the interval $(t_{i-1}, t_i]$, and hence

$$\begin{aligned} P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} &= \mathrm{P}(\text{no reaction in } (t_{i-1}, t_i]) \\ &= \mathrm{e}^{-a_0\Delta t} \end{aligned}$$

as above.

Case 2: $\boldsymbol{X}(t_i) = \boldsymbol{X}(t_{i-1}) + \nu_\mu$ for some μ. We assume that only one reaction of type μ has occurred in $(t_{i-1}, t_i]$, and hence

$$\begin{aligned} P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} &= \mathrm{P}(\text{only reaction } \mu \text{ in } (t_i, t_{i+1}]) \\ &= \int_0^{\Delta t} a_\mu \mathrm{e}^{-a_0\tau}\, \mathrm{P}(\text{no reaction in } (\tau, t_{i+1}])\, \mathrm{d}\tau \\ &= \int_0^{\Delta t} a_\mu \mathrm{e}^{-a_0\tau} \mathrm{e}^{-a_0(\Delta t-\tau)}\, \mathrm{d}\tau \\ &= a_\mu \mathrm{e}^{-a_0\Delta t}\Delta t \end{aligned}$$

Note the difference between this formula and (5), where we included the changed overall reaction rate a_0'.

In general, the probability that reactions $\mu_1, \ldots, \mu_K$ occur (in that order) in a sampling interval is

$$\begin{aligned} &\mathrm{P}(\text{reactions } \mu_1, \ldots, \mu_K \text{ in } (t_i, t_{i+1}]) \\ &\quad = \left(\prod_{l=1}^{K} a_{\mu_l}\right) \mathrm{e}^{-a_0\Delta t} \frac{\Delta t^{2K-1}}{(2K-1)!} \end{aligned} \tag{7}$$

as can easily be proven by induction (see the Supplementary). The solution to (6), $\boldsymbol{C}^i$, specifies how many reactions of each type have occurred in the sampling interval, but does not identify the order of these reactions. Therefore if, according to $\boldsymbol{C}^i$, reactions $\mu_1^i, \ldots, \mu_{K^i}^i$ have occurred (in some unknown order) in $(t_{i-1}, t_i]$, then

$$P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i} = \frac{K^i!}{\prod_{l=1}^{M} C_l^i!} \left(\prod_{l=1}^{K^i} a_{\mu_l^i}\right) \mathrm{e}^{-a_0\Delta t} \frac{\Delta t^{2K^i-1}}{(2K^i-1)!}$$

where the first factor is the number of possible permutations of these K^i reactions.

The log-likelihood based on the observations $\boldsymbol{X}(t_i)$ can then be computed as

$$\begin{aligned} \log \mathcal{L} &= \log(\mathrm{P}(\boldsymbol{X}(t_1) = \boldsymbol{x}_1, \ldots, \boldsymbol{X}(t_R) = \boldsymbol{x}_R)) \\ &= \log\left(\pi_{\boldsymbol{x}_1} \prod_{i=2}^{R} P_{\boldsymbol{x}_{i-1},\boldsymbol{x}_i}\right) \\ &= -a_0\Delta t(R-1) + \sum_{i=2}^{R}\left((2K^i - 1)\log(\Delta t)\right. \\ &\quad \left. + \sum_{l=1}^{K^i} \log(a_{\mu_l^i}) - \sum_{l=K^i+1}^{2K^i-1} \log(l) - \sum_{l=1}^{M} \log C_l^i!\right) \end{aligned} \tag{8}$$

This SVDL (8) can then be maximised using a numerical optimisation method.

We return now to the more realistic scenario where the molecule counts are measured with error so that $\boldsymbol{O}(t_i) \neq \boldsymbol{X}(t_i)$. In this setting, the exact number of reactions can no longer be computed. However, we simply ignore the measurement error and solve (6) using the non-integer observations, $\boldsymbol{O}(t_i)$. We round the values obtained in order to obtain an approximation to the number of each type of reaction, which can be used to evaluate and maximise (8).

The log-likelihood (8) corresponds to a Markov chain model rather than a HMM, and hence is much easier to compute than the log-likelihood in the AML method (4). In particular, (8) is a simple sum, whereas (4) is a product of matrices. For this reason, the SVDL method is much faster than the AML method, and hence makes estimating parameters of complex reaction systems more feasible.

4 Results

In this section, we present simulation studies of the simple gene transcription and the more complex transriptional regulation systems to demonstrate the performance of the AML and the SVDL algorithms.

The systems are simulated using the Gillespie algorithm to generate time series of molecule counts. These are sampled at regular time intervals of length Δt to mimic experimental measurements. We then add Gaussian white noise with standard deviation σ to these counts to simulate measurement error. The focus of this investigation is how the performances of the algorithms vary with the parameters σ, Δt and R, and for each parameter choice we perform 100 simulation and estimation runs.

Table 1: Average (standard deviation) of the estimates of Golding's stochastic gene expression model parameters using the AML algorithm

R	Δt	σ	Parameter estimates			
			c_1	c_2	c_3	σ
300	1.0	1.0	0.0276(0.0158)	0.1494(0.0866)	0.3708(0.1611)	1.0012(0.0869)
300	1.0	0.1	0.02930(0.0168)	0.1875(0.1145)	0.3792(0.1789)	0.1004(0.0077)
300	1.0	3.0	0.03197(0.0205)	0.1201(0.0703)	0.3023(0.1565)	3.0189(0.2465)
300	0.5	1.0	0.04247(0.0277)	0.2415(0.1662)	0.4595(0.2213)	0.9952(0.1023)
300	0.5	0.1	0.03321(0.0232)	0.2560(0.2790)	0.4908(0.3373)	0.1014(0.0083)
300	0.5	3.0	0.03257(0.0276)	0.1951(0.1557)	0.4498(0.3276)	2.9969(0.2436)
600	0.5	1.0	0.02990(0.0160)	0.1875(0.1496)	0.3834(0.1601)	1.0002(0.0658)
600	0.5	0.1	0.0322(0.0143)	0.2262(0.1806)	0.4393(0.2117)	0.1005(0.0058)
600	0.5	3.0	0.0341(0.0189)	0.2072(0.1308)	0.4211(0.2112)	3.0463(0.1762)

The true parameter values were $c_1 = 0.027$, $c_2 = 0.1667$, $c_3 = 0.4$, and the initial values were $X_1(0) = 1$, $X_2(0) = 0$, $X_3(0) = 0$

Table 2: Average (standard deviation) of the estimates of Golding's stochastic gene expression model using the AML algorithm with five independent time series in each run

R	Δt	σ	Parameter estimates			
			c_1	c_2	c_3	σ
300	1.0	1.0	0.0257(0.0054)	0.1409(0.0402)	0.3461(0.0630)	1.0025(0.0504)
300	1.0	0.1	0.0268(0.0061)	0.1523(0.0424)	0.3741(0.0557)	0.1012(0.0031)
300	1.0	3.0	0.0250(0.0065)	0.1140(0.0337)	0.3160(0.0674)	3.0292(0.1393)
300	0.5	1.0	0.0282(0.0110)	0.1890(0.0960)	0.3752(0.0909)	1.0013(0.0411)
300	0.5	0.1	0.0264(0.0079)	0.1642(0.0716)	0.3863(0.1036)	0.1004(0.0038)
300	0.5	3.0	0.0310(0.0116)	0.1695(0.0897)	0.3891(0.1425)	3.0378(0.1452)

The true parameter values are as in Table 1

4.1 Simple gene expression system

For the AML algorithm, we limit the likelihood calculation to a maximum of two reactions per sampling interval, and assume that $X_n(t)$ falls within 3σ of $O_n(t)$. These assumptions are justified by the experimental conditions used by Golding *et al.* [17] on this system. We also assume that the initial state of the system is known, so that $\pi_{(1,0,0)} = 1$. The initial probabilities could also be estimated with the AML method, although at the cost of increased computational time. We use the true parameter values as starting values in the quasi-Newton routine. Because in Golding's experimental setup only mRNA rates were measured, the estimation is performed using only the mRNA measurements, whereas DNA_{ON} and DNA_{OFF} are hidden.

Table 1 gives the averages and sample standard deviations of the parameter estimates obtained using the AML algorithm. In general, the algorithm delivers good estimates of the original parameter values with reasonably small standard deviations. Increased measurement error naturally leads to less precision in the parameter estimates, but interestingly the error is not much higher under more observational noise. In addition, we see that for a constant number of sampling points, R, the parameters are estimated better when the sampling rate, Δt, is lower (i.e. the observation period is longer). This result is to be expected, as over a longer time period we would see more reactions, and hence have more information about the reaction propensities. Consequently, more sampling points, reflecting longer time series, also improve estimation accuracy. The noise strength σ is estimated quite accurately as well.

The performance of parameter estimation can be improved if a number of traces from one or multiple cells are lumped together and the total likelihood is maximised (see Golding *et al.* [17] for a similar procedure on experimental data). Table 2 gives the averages and sample deviations of the parameter estimates in the case of estimation using five traces. The standard deviations are noticeably lower than in Table 1, reflecting decreased variation in individual estimates. Whether such estimation from multiple traces is possible in practice will depend on the system under investigation, the experimental conditions and the constancy of reaction rates among traces.

Table 3 shows parameter estimates obtained with the SVDL algorithm. We use random starting values for the quasi-Newton optimisation routine. In all estimation runs, data from all molecule species are included (in contrast to the AML method) because the SVDL method cannot incorporate a state space search over hidden variables. As this method implicitly assumes exactly measured molecule counts, we do several runs with $\sigma = 0$. In this case, the parameters are estimated with good accuracy for longer time series. For shorter time series and higher sampling rates, the standard deviation increases, reflecting the lack of information about the parameters contained in the data. The best estimation results are obtained from long time series with high sampling rates.

When observational error is added to the data, the SVDL algorithm is still able to generate estimates of the reaction rates, albeit with higher standard deviations. For $\sigma = 1.0$, as in the experiments, the precision of the parameter estimates is decreased, especially of c_1 (note that in this system, c_1 is the smallest reaction rate, and hence R_1 occurs least frequently). Again, estimation from a number of combined data time series greatly improves estimation accuracy and decreases the standard deviations, as shown in Table 4. In systems with few reactions over the time course of one data time series, as in this simple model (see also Fig. 1 [17]), such a combination of multiple time

Table 3: Average (standard deviation) of the estimates of the simple gene expression parameters using the SVDL algorithm

R	Δt	σ	Parameter estimates		
			c_1	c_2	c_3
300	0.5	0.0	0.0299(0.0166)	0.1934(0.1271)	0.3612(0.2564)
300	1.0	0.0	0.0280(0.0114)	0.1775(0.0765)	0.3712(0.1133)
600	0.5	0.0	0.0271(0.0107)	0.1826(0.0865)	0.3891(0.1036)
600	1.0	0.0	0.0263(0.0076)	0.1678(0.0433)	0.3708(0.0672)
300	1.0	0.1	0.0236(0.0105)	0.1576(0.1457)	0.4660(0.3993)
300	1.0	0.5	0.0287(0.0214)	0.1182(0.1305)	0.3541(0.4231)
300	1.0	1.0	0.0447(0.1306)	0.1164(0.1447)	0.3381(0.5421)

Parameter values are as in Table 1

Table 4: Average (standard deviation) of the estimates from five data time series in the simple gene transcription model using the SVDL algorithm

R	Δt	σ	Parameter estimates		
			c_1	c_2	c_3
300	0.5	0.0	0.0275(0.0073)	0.1680(0.0444)	0.3852(0.0556)
300	1.0	0.0	0.0275(0.0047)	0.1564(0.0229)	0.3721(0.0333)
600	1.0	0.0	0.0273(0.0035)	0.1593(0.0166)	0.3730(0.0299)
600	0.5	0.0	0.0264(0.0045)	0.1625(0.0270)	0.3909(0.0359)
300	1.0	0.1	0.0229(0.0041)	0.1573(0.0691)	0.4594(0.1923)
300	1.0	0.5	0.0242(0.0062)	0.1390(0.0722)	0.3979(0.1979)
300	1.0	1.0	0.0295(0.0102)	0.1321(0.0787)	0.3842(0.2140)

Parameter values are as in Table 1

series may be necessary to obtain a good parameter estimate.

These results demonstrate that the approximate reaction numbers, calculated with the singular value decomposition method, are sufficiently close to the real reaction numbers, to provide good estimates when sufficient sampling points are considered.

In terms of the computation time, C implementations of both algorithms on an Intel Pentium 4 processor with 3.00 GHz took ~68 min in the case of the AML method and 20 s for the SVDL method (100 runs with $\sigma = 0.1$, $\Delta t = 1.0$, $R = 500$). Increased σ requires substantially longer estimation run times.

4.2 Transcriptional regulation system

In the higher dimensional model of transcriptional regulation, the AML method is not practical because the large state space leads to prohibitively large computation times. In such a complex system, the SVDL method is at an advantage because it does not search the state space but rather simply solves a systems of linear equation with dimensions given by the number of species and reactions.

Table 5 shows the average and sample standard deviations of the SVDL parameter estimates in this larger stochastic reaction system. The parameter estimates are reasonably close to the true values in most cases. Again, longer time series and higher sampling rates give better estimates. Especially c_7 and c_8 are estimated better at higher sampling rates. This can be explained by the complementarity of reactions 7 and 8. Because one reaction 7 and consequent reaction 8 (and vice versa) lead to a zero net change of molecule counts, discretely sampled data contain only information about the net numbers of these reaction pairs. As many such reactions take place during longer sampling intervals, any method will have difficulty estimating the associated reaction rates under these conditions. Added measurement error decreases measurement accuracy and actually often leads to failed convergence of the algorithm.

Parameter estimation in this model system took ~4 min for 300 data points and 100 repetitions of the SVDL algorithm.

4.3 Confidence intervals for the parameters

In practice, often only one time series of molecule counts will be available for parameter estimation. In this case, it is useful to compute confidence intervals for the parameters.

In the case of the AML method, we use the negative inverse of the matrix of second derivatives of the maximised log-likelihood to derive such confidence intervals. The quasi-Newton routine provides this matrix as a by-product of the optimisation algorithm. As long as the assumption of at most two reactions per time interval is satisfied, the diagonal entries of the matrix can be used as estimates of the asymptotic variances of the parameter estimates, $\hat{c}_i^*$ and $\hat{\sigma}^*$ (see Section 3.1).

In order to estimate the confidence intervals for the parameters c_i and σ from the approximations of the log parameters (see Section 3.1) we can use the Delta method. This method is based on the fact that if a random variable Z is normally distributed with mean μ_z and variance σ_z, then $g(Z)$ is approximately normally distributed with mean $g(\mu_z)$ and variance $\sigma_z[g'(\mu_z)]^2$.

In order to demonstrate the validity of the asymptotic standard errors (ASEs) computed in this way, we compare them with the empirical standard errors (ESEs) in 100 parameter simulation and estimation runs of the simple gene transcription system (using only one time series per run).

The ASEs in Table 6 are mostly very close to the observed ESEs, demonstrating that the ASEs provide an adequate idea about the precision of the estimates, and confidence intervals based on these values will be valid.

Table 5: Average (standard deviation) of the estimates of the transcriptional regulation system parameters using the SVDL method

R	Δt	σ	Parameter estimates							
			c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
500	0.5	0.0	0.0666(0.0251)	0.0022(0.0035)	0.0808(0.0253)	0.0364(0.0362)	0.0279(0.0104)	0.2359(0.1491)	0.0473(0.0143)	0.1600(0.0722)
500	0.1	0.0	0.0992(0.0143)	0.0037(0.0078)	0.1268(0.0930)	0.0639(0.0358)	0.0659(0.0174)	0.2757(0.4110)	0.1646(0.0529)	0.6574(0.2629)
500	1.0	0.0	0.0477(0.0155)	0.0006(0.0004)	0.0645(0.0190)	0.0110(0.0195)	0.0159(0.0107)	0.2646(0.0761)	0.0149(0.0143)	0.0615(0.0332)
1000	0.5	0.0	0.0462(0.0134)	0.0006(0.0004)	0.0697(0.0159)	0.0091(0.0166)	0.0176(0.0088)	0.3323(0.0862)	0.0194(0.0123)	0.1024(0.0522)
1000	1.0	0.0	0.0429(0.0046)	0.0004(0.0002)	0.0630(0.0086)	0.0043(0.0058)	0.0080(0.0036)	0.1817(0.0327)	0.0036(0.0047)	0.0167(0.0153)
1000	0.1	0.0	0.0851(0.0202)	0.0025(0.0052)	0.1254(0.0519)	0.0440(0.0353)	0.0450(0.0165)	0.3974(0.4251)	0.1750(0.0738)	0.6953(0.2461)
1000	0.5	0.5	0.0342(0.0465)	0.0021(0.0104)	0.0786(0.0698)	0.1156(0.2334)	0.0401(0.1034)	0.1356(0.2356)	0.01342(0.0764)	0.2456(0.3435)

The true parameter values were $c_1 = 0.043$, $c_2 = 0.0007$, $c_3 = 0.0715$, $c_4 = 0.00395$, $c_5 = 0.02$, $c_6 = 0.4791$, $c_7 = 0.083$, $c_8 = 0.5$

Table 6: Comparison of the ASEs and ESEs for the estimates of the parameters of the simple gene transcription system using the AML algorithm

R	Δt	σ	ESE				ASE			
			$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{\sigma}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{\sigma}$
300	1.0	1.0	0.0158	0.0866	0.1611	0.0869	0.0161	0.1111	0.1654	0.0848
300	1.0	0.1	0.0168	0.1145	0.1789	0.0077	0.0169	0.1336	0.1650	0.0092
300	1.0	3.0	0.0205	0.0703	0.1565	0.2465	0.0217	0.1047	0.1485	0.2553
300	0.5	1.0	0.0277	0.1662	0.2213	0.1023	0.0328	0.2490	0.2908	0.0835
300	0.5	0.1	0.0232	0.2790	0.3373	0.0083	0.0257	0.2525	0.2992	0.0084
300	0.5	3.0	0.0276	0.1557	0.3276	0.2436	0.0293	0.2454	0.3344	0.2481
600	0.5	1.0	0.0161	0.1496	0.1601	0.0658	0.0197	0.1559	0.1838	0.0651
600	0.5	0.1	0.0143	0.1806	0.2117	0.0058	0.0171	0.1424	0.1774	0.0059
600	0.5	3.0	0.0189	0.1308	0.2112	0.1762	0.0238	0.1739	0.2223	0.1784

In the case of the SVDL algorithm, the Fisher information matrix can be computed analytically. However, quantile plots of the parameter estimates (not included) show marked deviations from the normal distribution, even for large values of R. Hence, we cannot assume that the distribution of these estimates is asymptotically normal. Further insight into the distribution of the SVDL estimates is required in order to be able to form valid confidence intervals.

5 Conclusion

In this article, we presented two different algorithms for the estimation of stochastic reaction parameters from discrete time series of molecule counts measured with error. The main challenge associated with this problem is the complexity of the likelihood when many reactions may occur in a sampling interval. Therefore both algorithms make simplifying assumptions in order to make parameter estimation feasible.

In the AML algorithm, because of the expensive state space search and computations involved, we assume a finite maximum on the number of reactions that can occur during a sampling interval. This assumption restricts the use of this method to systems with high sampling rates, where few reactions take place in a sampling interval. A major advantage of this method is that it can handle incomplete data when one or more of the involved molecule species cannot be measured.

The algorithm gives good estimates for the three-dimensional gene transcription systems, and has the advantage of providing asymptotically normal parameter estimates. The estimated standard errors generated as a by-product of the optimisation routine are very accurate. Hence, we can form valid confidence intervals for the parameters using this algorithm.

In the SVDL algorithm, we treat the observations as exact molecule counts, and assume constant reaction propensities over a sampling interval. These assumptions are reasonable for systems with large numbers of molecules and small measurement errors. Our simulation studies demonstrate that this method can reliably estimate the kinetic parameters of stochastic reaction systems, even from data with measurement error. However, confidence intervals for the parameters cannot be generated using the SVDL algorithm. In addition, computation speed is faster by about two orders of magnitude, compared with the AML algorithm. The computational efficiency of the SVDL algorithm is a clear advantage of this method.

Our simulations stress the importance of the sampling rate on parameter identifiability, a problem that is known in other stochastic parameter identification problems as well [37]. Discretely sampled data do not include information about forward and backward reactions that occur in the same time interval, and hence the associated reaction constants may be biased for small samples. The exact likelihood is not available in closed form, and hence theoretical analysis of the magnitude of this bias is challenging.

Our methods, like the Gillespie framework, ignore spatial effects in molecular dynamics. However, it is known that in cells, scaffolding proteins and metabolic complexes can localise substrates and enzymes, so that effective local concentrations are much higher at the reaction sites [38]. Because we neglect spatial effects, our estimates will reflect overall reaction rates in the cellular environment averaged over the whole cell. Thus, for comparison of reaction rates between different experimental conditions and cell lines, our estimates should be considered as mean rates in the cell.

Inside cells, various compartments restrict the intracellular space, making it more tortuous and impeding diffusion [39, 40]. This can lead to nonlinear and even fractal-like diffusion behaviour. When spatially resolved data with high accuracy are available from experiments, our techniques could be extended to estimate parameters in spatial versions of the Gillespie algorithm [41]. Tackling the full complexity of cellular dynamics, including inhomogeneities and transport effects, is still beyond the scope of current systems biology research, but progress in this direction is made [8, 42].

To obtain experimental data on biochemical reactions, the latest experimental techniques, especially imaging methods of fluorescent tagged proteins, can monitor biological processes at microscopic levels, as required in our context. For example, the stochastic expression dynamics of a reporter gene with a fluorescent tag has been monitored successfully in different settings [17, 29], and resolution of gene expression is in the range of tens or even single molecules [17, 42, 43]. Fluorescence measurement techniques also give data of sufficient time and concentration resolution to enable comparison and validation of stochastic models and experimental data [44]. Zlokarnik *et al.* [31] used a FRAP technique to measure stochastic transcription and obtain time series of molecule counts. Fluorescent protein tags of different colours are available to monitor multiple proteins simultaneously (see, for example, Becskei *et al.* [8] and Elowitz *et al.* [11]). However, time-resolved

simultaneous measurement of a number of reacting species have, to our knowledge, not yet been performed. Technical difficulties, including the reliable tagging of multiple molecules, calibration and the data acquisition speed of microscope cameras, need to be overcome in order to generate multi-species times series.

With the recent development of highly accurate methods for the measurement of mRNA, protein and metabolite concentrations in living cells, biochemical processes can be followed with unprecedented precision. Modern fluorescence methods are on the brink of resolving molecular reactions at the individual molecule level, and will provide data that require methods such as ours for analysis.

6 Supplementary

We can prove (7) using induction, assuming that it is correct for K reactions. The probability that reactions $\mu_1, \ldots, \mu_{K+1}$ occur in the sampling interval is then

$$
\begin{aligned}
&\mathrm{P}(\text{reactions } \mu_1, \ldots, \mu_{K+1} \text{ in } (t_i, t_{i+1}]) \\
&= \int_0^{\Delta t} \mathrm{P}(\text{reactions } \mu_1, \ldots \mu_K \text{ in } (0, \tau]) \\
&\quad \times \mathrm{P}(\text{reaction } \mu_{K+1} \text{ in } (\tau, \Delta t])\, \mathrm{d}\tau \\
&= \int_0^{\Delta t} \left(\prod_{l=1}^{K} a_{\mu_l} \right) \mathrm{e}^{-a_0 \tau} \frac{\Delta t^{2K-1}}{(2K-1)!} a_{\mu_{K+1}} \mathrm{e}^{-a_0(\Delta t - \tau)} (\Delta t - \tau)\, \mathrm{d}\tau \\
&= \left(\prod_{l=1}^{K+1} a_{\mu_l} \right) \mathrm{e}^{-a_0 \Delta t} \frac{1}{(2K-1)!} \int_0^{\Delta t} \tau^{2K-1} (\Delta t - \tau)\, \mathrm{d}\tau \\
&= \left(\prod_{l=1}^{K+1} a_{\mu_l} \right) \mathrm{e}^{-a_0 \Delta t} \frac{1}{(2K-1)!} \left[\frac{\Delta t \tau^{2K}}{2K} - \frac{\tau^{2K+1}}{2K+1} \right]_0^{\Delta t} \\
&= \left(\prod_{l=1}^{K+1} a_{\mu_l} \right) \mathrm{e}^{-a_0 \Delta t} \frac{\Delta t^{2K+1}}{(2K-1)!\, 2K(2K+1)} \\
&= \left(\prod_{l=1}^{K+1} a_{\mu_l} \right) \mathrm{e}^{-a_0 \Delta t} \frac{\Delta t^{2(K+1)-1}}{[2(K+1)-1]!}
\end{aligned}
$$

Therefore (7) is true for all K.

7 References

1 Arkin, A., Ross, J., and McAdams, H.H.: 'Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells', *Genetics*, 1998, **149**, pp. 1633–1648

2 Levin, M.D.: 'Noise in gene expression as the source of non-genetic individuality in the chemotactic response of *Escherichia coli*', *FEBS Lett.*, 2003, **550**, pp. 135–138

3 Korobkova, E., Emonet, T., Vilar, J.M.G., Shimizu, T.S., and Cruzel, P.: 'From molecular noise to behavioural variability in a single bacterium', *Nature*, 2004, **428**, pp. 574–578

4 Paulsson, J., Berg, O.G., and Ehrenberg, M.: 'Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation', *PNAS*, 2000, **97**, pp. 7148–7153

5 Samoilov, M., Plyasunov, S., and Arkin, A.P.: 'Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations', *PNAS*, 2005, **102**, pp. 2310–2315

6 Blake, W.J., Kærn, M., Cantor, C.R., and Collins, J.J.: 'Noise in eukaryotic gene expression', *Nature*, 2003, **422**, pp. 633–637

7 McAdams, H.H., and Arkin, A.: 'Stochastic mechanisms in gene expression', *PNAS*, 1997, **94**, pp. 814–819

8 Becskei, A., Kaufmann, B.B., and van Oudenaarden, A.: 'Contributions of low molecule number and chromosomal positioning to stochastic gene expression', *Nat. Genet.*, 2005, **37**, pp. 937–944

9 Paulsson, J.: 'Summing up the noise in gene networks', *Nature*, 2004, **427**, pp. 415–418

10 Swain, P.S., Elowitz, M.B., and Siggia, E.D.: 'Intrinsic and extrinsic contributions to stochasticity in gene expression', *PNAS*, 2002, **99**, pp. 12795–12800

11 Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S.: 'Stochastic gene expression in a single cell', *Science*, 2002, **297**, pp. 1183–1186

12 Wolf, D.M., and Arkin, A.P.: 'Motifs, modules, and games in bacteria', *Curr. Opin. Microbiol.*, 2003, **6**, pp. 125–134

13 Orrell, D., and Bolouri, H.: 'Control of internal and external noise in genetic regulatory networks', *J. Theor. Biol.*, 2004, **230**, pp. 301–312

14 Thattai, M., and van Oudenaarden, A.: 'Intrinsic noise in gene regulatory networks', *PNAS*, 2001, **98**, pp. 8614–8619

15 Kierzek, A.M., Jolanta, Z., and Zielenkiewicz, P.: 'The effect of transciption and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression', *J. Biol. Chem.*, 2001, **276**, pp. 8165–8172

16 Kepler, T.B., and Elston, T.C.: 'Stochasticity in transcriptional regulation: origins and consequences and mathematical representations', *Biophys. J.*, 2001, **81**, pp. 3116–3136

17 Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C.: 'Real-time kinetics of gene activity in individual bacteria', *Cell*, 2005, **123**, pp. 1025–1036

18 Barkai, N., and Leibler, S.: 'Biological rhythms: circadian clocks limited by noise', *Nature*, 2000, **403**, pp. 267–268

19 Raser, J.M., and O'Shea, E.K.: 'Control of stochasticity in eukaryotic gene expression', *Science*, 2004, **304**, pp. 1811–1814

20 Fraser, H.B., Hirsh, A.E., Giaever, G., Kumm, J., and Eisen, M.B.: 'Noise minimization in eukaryotic gene expression', *PLoS Biol.*, 2004, **2**, pp. 834–838

21 Swain, P.S.: 'Efficient attenuation of stochasticity in gene expression through post-transcriptional control', *J. Mol. Biol.*, 2004, **344**, pp. 965–976

22 Shibata, T., and Fujimoto, K.: 'Noisy signal amplification in ultrasensitive signal transduction', *PNAS*, 2005, **102**, pp. 331–336

23 Moles, C.G., Mendes, P., and Banga, J.R.: 'Parameter estimation in biochemical pathways: a comparison of global optimization methods', *Genome Res.*, 2003, **13**, pp. 2467–2474

24 Swameye, I., Müller, T.G., Timmer, J., Sandra, O., and Klingmüller, U.: 'Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling', *PNAS*, 2003, **100**, pp. 1028–1033

25 Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P.: 'Causal protein-signaling networks derived from multiparameter single-cell data', *Science*, 2005, **308**, pp. 523–529

26 Perkins, T.J., Hallett, M., and Glass, L.: 'Inferring models of gene expression dynamics', *J. Theor. Biol.*, 2004, **230**, pp. 289–299

27 Puchalka, J., and Kierzek, A.M.: 'Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks', *Biophys. J.*, 2004, **86**, pp. 1357–1372

28 Gillespie, D.T.: 'Exact stochastic simulation of coupled chemical reactions', *J. Phys. Chem.*, 1977, **81**, pp. 2340–2361

29 Ozbudak, E.M., Thattai, M., Kurtser, I., Grossmann, A.D., and van Oudenaarden, A.: 'Regulation of noise in the expression of a single gene', *Nat. Genet.*, 2002, **31**, pp. 69–73

30 Wolkenhauer, O., Ullah, M., Kolch, W., and Cho, K.-H.: 'Modelling and simulation of intracellular dynamics: choosing an appropriate framework', *IEEE Trans. Nanobioscience*, 2005, **3**, pp. 200–207

31 Zlokarnik, G., Negulescu, P.A., Knapp, T.E., Mere, L., Burres, N., Feng, L., Whitney, M., Roemer, K., and Tsien, R.Y.: 'Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter', *Science*, 1998, **279**, pp. 84–88

32 Golightly, A., and Wilkinson, D.J.: 'Bayesian inference for stochastic kinetic models using a diffusion approximation', *Biometrics*, 2005, **61**, pp. 781–788

33 Boys, R.J., Wilkinson, D.J., and Kirkwood, T.B.L.: 'Bayesian inference for a discretely observed stochastic kinetic model'. Statistics Preprint STA04, University of Newcastle, UK, 2004

34 Goutsias, J.: 'A hidden Markov model for transcriptional regulation in single cells', *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2006, **3**, pp. 57–71

35 Douc, R., and Matias, C.: 'Asymptotics of the maximum likelihood estimator for general hidden Markov models', *Bernoulli*, 2001, **7**, pp. 381–420

36 Gillespie, D.T.: 'Approximate accelerated stochastic simulation of chemically reacting systems', *J. Phys. Chem.*, 2001, **115**, pp. 1716–1733

37 Timmer, J.: 'Parameter estimation in nonlinear stochastic differential equations', *Chaos Solitons Fractals*, 2000, **11**, pp. 2571–2578

38 Kholodenko, B.N.: 'Cell-signalling dynamics in time and space', *Nature Revies Molecular Cell Biology*, 2006, **7**, pp. 165–176

39 Ahn, J., Kopelman, R., and Argyrakis, P.: 'Hierarchies of nonclassical reaction kinetics due to anisotropic confinements', *J. Chem. Phys.*, 1999, **110**, pp. 2116–2121
40 Schnell, S., and Turner, T.E.: 'Reaction kinetics in intracellular environments with macro-molecular crowding: simulations and rate laws', *Progress in Biophys. Mol. Biol.*, 2004, **85**, pp. 235–260
41 Turner, T.E., Schnell, S., and Burrage, K.: 'Stochastic approaches for modelling in vivo reactions', *Comput. Biol. Chem.*, 2004, **28**, pp. 165–178
42 Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B.: 'Gene regulation at the single-cell level', *Science*, 2005, **307**, pp. 1962–1965
43 Wu, J.Q., and Pollard, T.D.: 'Counting cytokinesis proteins globally and locally in fission yeast', *Science*, 2005, **310**, pp. 310–314
44 Isaacs, F.J., Hasty, J., Cantor, C.R., and Collins, J.J.: 'Prediction and measurement of an autoregulatory genetic module', *PNAS*, 2003, **100**, pp. 7714–7719