Methodology article

Bayesian inference of biochemical kinetic parameters using the linear noise approximation

Michał Komorowski*[1,2], Bärbel Finkenstädt[1], Claire V Harper[4] and David A Rand[2,3]

Address: [1]Department of Statistics, University of Warwick, Coventry, UK, [2]Systems Biology Centre, University of Warwick, Coventry, UK, [3]Mathematics Institute, University of Warwick, Coventry, UK and [4]Department of Biology, University of Liverpool, Liverpool, UK

Email: Michał Komorowski* - M.Komorowski@warwick.ac.uk; Bärbel Finkenstädt - B.F.Finkenstadt@warwick.ac.uk; Claire V Harper - claire.harper@liverpool.ac.uk; David A Rand - d.a.rand@warwick.ac.uk

* Corresponding author

Abstract

Background: Fluorescent and luminescent gene reporters allow us to dynamically quantify changes in molecular species concentration over time on the single cell level. The mathematical modeling of their interaction through multivariate dynamical models requires the deveopment of effective statistical methods to calibrate such models against available data. Given the prevalence of stochasticity and noise in biochemical systems inference for stochastic models is of special interest. In this paper we present a simple and computationally efficient algorithm for the estimation of biochemical kinetic parameters from gene reporter data.

Results: We use the linear noise approximation to model biochemical reactions through a stochastic dynamic model which essentially approximates a diffusion model by an ordinary differential equation model with an appropriately defined noise process. An explicit formula for the likelihood function can be derived allowing for computationally efficient parameter estimation. The proposed algorithm is embedded in a Bayesian framework and inference is performed using Markov chain Monte Carlo.

Conclusion: The major advantage of the method is that in contrast to the more established diffusion approximation based methods the computationally costly methods of data augmentation are not necessary. Our approach also allows for unobserved variables and measurement error. The application of the method to both simulated and experimental data shows that the proposed methodology provides a useful alternative to diffusion approximation based methods.

Background

The estimation of parameters in biokinetic models from experimental data is an important problem in Systems Biology. In general the aim is to calibrate the model so as to reproduce experimental results in the best possible way. The solution of this task plays a key role in interpreting experimental data in the context of dynamic mathematical models and hence in understanding the dynamics and control of complex intracellular chemical networks and the construction of synthetic regulatory circuits [1]. Among biochemical kinetic systems, the dynamics of gene expression and of gene regulatory networks are of particular interest. Recent developments of fluorescent microscopy allow us to quantify changes in protein

concentration over time in single cells (e.g. [2,3]) even with single molecule precision (see [4] for review). Therefore an abundance of data is becoming available to estimate parameters of mathematical models in many important cellular systems.

Single cell imaging techniques have revealed the stochastic nature of biochemical reactions (see [5] for review) that most often occur far from thermodynamic equilibrium [6] and may involve small copy numbers of reacting macromolecules [7]. This inherent stochasticity implies that the dynamic behaviour of one cell is not exactly reproducible and that there exists stochastic heterogeneity between cells. The disparate biological systems, experimental designs and data types impose conditions on the statistical methods that should be used for inference [8-10]. From the modeling point of view the current common consensus is that the most exact stochastic description of the biochemical kinetic system is provided by the chemical master equation (CME) [11]. Unfortunately, for many tasks such as inference the CME is not a convenient mathematical tool and hence various types of approximations have been developed. The three most commonly used approximations are [12]:

> 1. The macroscopic rate equation (MRE) approach which describes the thermodynamic limit of the system with ordinary differential equations and does not take into account random fluctuations due to the stochasticity of the reactions.

> 2. The diffusion approximation (DA) which provides stochastic differential equation (SDE) models where the stochastic perturbation is introduced by a state dependant Gaussian noise.

> 3. The linear noise approximation (LNA) which can be seen as a combination because it incorporates the deterministic MRE as a model of the macroscopic system and the SDEs to approximatively describe the fluctuations around the deterministic state.

Statistical methods based on the MRE have been most widely studied [8,13-15]. They require data based on large populations. The main advantages of this method are its conceptual simplicity and the existence of extensive theory for differential equations. However, single cells experiments and studies of noise in small regulatory networks created the need for statistical tools that are capable to extract information from fluctuations in molecular species. Few methods used CME to address this. Algorithm, proposed by [16], approximated the likelihood function, the other, suggested by [17] simulated it using Monte Carlo methods. Recently, also a method based on the exact likelihood [18] has been developed. Although, substantial progress has been made in numerical methods for

solving CME, inference algorithms based on the CME are computationally intensive and difficult to apply to problems of realistic size and complexity [19]. Another group of methods is based on the DA [9,20]. This uses likelihood approximation methods (e.g. [21]) that are computationally intensive and require sampling from high dimensional posterior distributions. Inference about the volatility process becomes difficult for low frequency data that are not directly measured at the molecular level [10,20]. The aim of this study is to investigate the use of the LNA as a method for inference about kinetic parameters of stochastic biochemical systems. We find that the LNA approximation provides an explicit Gaussian likelihood for models with hidden variables and measurement error and is therefore simpler to use and computationally efficient. To account for prior information on parameters our methodology is embedded in the Bayesian paradigm. The paper is structured as follows: We first provide a description of the LNA based modeling approach and then formulate the relevant statistical framework. We then study its applicability in four examples, based on both simulated and experimental data, that clarify principles of the method. Additional file 1 contains details of mathematical and statistical modeling, particularly comparison of the proposed method with an algorithm based on the DA.

Methods

The chemical master equation (CME) is the primary tool to model the stochastic behaviour of a reacting chemical system. It describes the evolution of the joint probability distribution of the number of different molecular species in a spatially homogeneous, well stirred and thermally equilibrated chemical system [11].

Even though these assumptions are not necessarily satisfied in living organisms the CME is commonly regarded as the most realistic model of biochemical reactions inside living cells. Consider a general system of N chemical species inside a volume Ω and let $\mathbf{X} = (X_1,..., X_N)^T$ denote the number and $\mathbf{x} = \mathbf{X}/\Omega$ the concentrations of molecules. The stoichiometry matrix $\mathbf{S} = \{S_{ij}\}_{i = 1,2...N; \ j = 1,2...R}$ describes changes in the population sizes due to R different chemical events, where each S_{ij} describes the change in the number of molecules of type i from X_i to $X_i + S_{ij}$ caused by an event of type j. The probability that an event of type j occurs in the time interval $[t, t + dt)$ equals $\tilde{f}_j(\mathbf{x}, \Omega, t)\Omega dt$.

The functions $\tilde{f}_j(\mathbf{x}, \Omega, t)$ are called *mesoscopic transition rates*. This specification leads to a Poisson birth and death process where the probability $h(\mathbf{X}, t)$ that the system is in the state $\mathbf{X}$ at time t is described by the CME [12] which is given in Additional file 1. It is straightforward to verify

that the first order terms of a Taylor expansian of the CME in powers of $1/\sqrt{\Omega}$ are given by the following MRE

$$\frac{d\phi_i}{dt} = \sum_{j=1}^{R} S_{ij} f_j(\varphi, t) \quad i = 1, 2, ..., N; \qquad (1)$$

where $\phi_i = \lim_{\Omega\to\infty,\ X\to\infty} X_i/\Omega$, $= \varphi(\phi_1,..., \phi_N)^T$ and $f_j(\varphi, t) = \lim_{\Omega\to\infty} \tilde{f}_j(\mathbf{x}, \Omega, t)$.

Including also the second order terms of this expansion produces the LNA

$$\mathbf{x}(t) = \varphi(t) + \Omega^{-\frac{1}{2}} \xi(t) \qquad (2)$$

which decomposes the state of the system into a deterministic part φ as solution of the MRE in (1) and a stochastic process ξ described by an Itô diffusion equation

$$d\xi(t) = \mathbf{A}(t)\xi dt + \mathbf{E}(t)dW(t), \qquad (3)$$

where $W(t)$ denotes R dimensional Wiener process, $[\mathbf{A}(t)]_{ik} = \sum_{j=1}^{R} S_{ij}\partial f_j/\partial\phi_k, [\mathbf{E}(t)]_{ik} = S_{ik}\sqrt{f_k(\varphi, t)}$ and $f_i = f_i(\varphi)$ (see Additional file 1 for derivation).

The rationale behind the expansion in terms of $1/\sqrt{\Omega}$ is that for constant average concentrations relative fluctuations will decrease with the inverse of the square root of volume [22]. Therefore the LNA is accurate when fluctuations are sufficiently small in relation to the mean (large Ω). Hence, the natural measure of adequacy of the LNA is the coefficient of variation i.e. ratio of the standard deviation to the mean (see Additional file 1). Validity of this approximation is also discussed in details in [22,23]. In addition it can be shown that the process describing the deviation from the deterministic state $\Omega^{\frac{1}{2}}$ converges weakly to the diffusion (3) as $\Omega \to \infty$ [24]. In order to use the LNA in a likelihood based inference method we need to derive transition densities of the process **x**.

Transition densities
The LNA provides solutions that are numerically or analytically tractable because the MRE in (1) can be solved numerically and the linear SDE in (3) for an initial condition $\xi(t_i) = \xi_{t_i}$ has a solution of the form [25]

$$\xi(t) = \Phi_{t_i}(t - t_i)\left(\xi_{t_i} + \int_{t_i}^{t} \Phi_{t_i}(s - t_i)^{-1}\mathbf{E}(s)dW(s) \right), \qquad (4)$$

where the integral is in the Itô sense and $\Phi_{t_i}(s)$ is the fundamental matrix of the non-autonomous system of ODEs

$$\frac{d\Phi_{t_i}}{ds} = \mathbf{A}(t_i + s)\Phi_{t_i}, \quad \Phi_{t_i}(0) = I. \qquad (5)$$

The Itô integral of a deterministic function is a Gaussian random variable [26], therefore equations (4), (5) imply that the transition densities of the process ξ are Gaussian [26] (throughout the paper we use 'Gaussian' or 'normal' shortly to denote either a univariate or a multivariate normal distribution depending on the context)

$$\mathbf{p}(\xi_{t_i} \mid \xi_{t_{i-1}}, \Theta) = \psi(\xi_{t_i} \mid \mu_{i-1}, \Xi_{i-1}) \qquad (6)$$

where Θ denotes a vector of all model parameters, $\psi(\cdot \mid \mu_{i-1}, \Xi_{i-1})$ is the normal density with mean μ_{i-1} and covariance matrix Ξ_{i-1} specified by

$$
\begin{aligned}
\mu_{i-1} &= \Phi_{t_{i-1}}(\Delta_{i-1})\xi_{t_{i-1}}, \quad \Delta_{i-1} = t_i - t_{i-1}, \\
\Xi_{i-1} &= \int_{t_{i-1}}^{t_i} (\Phi_s(t_i - s)E(s))(\Phi_s(t_i - s)E(s))^T ds.
\end{aligned}
\qquad (7)
$$

It follows from (2) and (6) that the transition densities of **x** are normal

$$\mathbf{p}(\mathbf{x}_{t_i} \mid \mathbf{x}_{t_{i-1}}, \Theta) = \psi(\mathbf{x}_{t_i} \mid \varphi(t_i) + \Omega^{-\frac{1}{2}}\mu_{i-1}, \Omega^{-1}\Xi_{i-1}). \qquad (8)$$

The properties of the normal distribution allow us to construct a convenient inference framework that is reminiscent of the Kalman filtering methodology (see e.g. [27]).

Inference
It is rarely possible to observe the time evolution of all molecular components participating in the system of interest [28]. Therefore, we partition the process $\mathbf{x}_t$ into those components $\mathbf{y}_t$ that are observed and those $\mathbf{z}_t$ that are unobserved.

Let $\bar{\mathbf{x}} \equiv (\mathbf{x}_{t_0}, ..., \mathbf{x}_{t_n})$, $\bar{\mathbf{y}} \equiv (\mathbf{y}_{t_0}, ..., \mathbf{y}_{t_n})$ and $\bar{\mathbf{z}} \equiv (\mathbf{z}_{t_0}, ..., \mathbf{z}_{t_n})$ denote the time-series that comprise the values of processes **x**, **y** and **z**, respectively, at times $t_0, ...,$

t_n. Here and throughout the paper we use the same letter to denote the stochastic process and its realization.

Our aim is to estimate the vector of unknown parameters Θ from a sequence of measurements $\bar{\mathbf{y}}$. The initial condition $\varphi(t_0)$ is parameterized as an element of Θ. Given the Markov property of the process $\mathbf{x}$ the augmented likelihood $P(\bar{\mathbf{y}}, \bar{\mathbf{z}} | \Theta)$ is given by

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}} | \Theta) = \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_{t_i} | \mathbf{x}_{t_{i-1}}, \Theta) \mathbf{p}(x_{t_0} | \Theta), \quad (9)$$

where $\mathbf{p}(\mathbf{x}_{t_i} | \mathbf{x}_{t_{i-1}}, \Theta)$ are Gaussian densities specified in (8), and $\mathbf{p}(\mathbf{x}_{t_0} | \Theta)$ is an initial density assumed to be normal for mathematical convenience. It can then be shown that (see Additional file 1) $\bar{\mathbf{x}}$ is Gaussian. Therefore

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}} | \Theta) = \psi(\bar{\mathbf{x}} | \varphi(t_0), ..., \varphi(t_n), \Sigma), \quad (10)$$

where $\psi(\cdot | \varphi(t_0), ..., \varphi(t_n), \hat{\Sigma})$ is Gaussian density with mean vector $(\varphi(t_0), ..., \varphi(t_n))$ and covariance matrix $\hat{\Sigma}$ whose elements can be calculated numerically in a straightforward way (see Additional file 1). Since the marginal distributions are also Gaussian it follows that the likelihood function $P(\bar{\mathbf{y}} | \Theta)$ can be obtained from the augmented likelihood (10)

$$P(\bar{\mathbf{y}} | \Theta) = \psi(\bar{\mathbf{y}} | \varphi_\gamma(t_0), ..., \varphi_\gamma(t_n)), \Sigma), \quad (11)$$

where the covariance matrix $\Sigma = \{\Sigma^{(i,j)}\}_{i,j=0,...,n}$ is a submatrix of $\hat{\Sigma}$ such that $\Sigma^{(i,j)} = Cov(\mathbf{y}_{t_i}, \mathbf{y}_{t_j})$ and φ_γ is the vector consisting of the observed components of φ.

Fluorescent reporter data are usually assumed to be proportional to the number of fluorescent molecules [29] and measurements are subject to *measurement error*, i.e. errors that do not influence the stochastic dynamics of the system. We therefore assume that instead of the matrix $\bar{\mathbf{y}}$ our data have the form $\bar{\mathbf{u}} \equiv \lambda \bar{\mathbf{y}} + (\epsilon_{t_0}, ..., \epsilon_{t_n})$. The parameter λ is a proportionality constant (it is straightforward to generalize for the case with different proportionality constants for different molecular components) and ϵ_{t_i} denotes a random vector for additive measurement error. For mathematical convenience we assume that the joint distribution of the measurement error is normal with mean 0 and known covariance matrix Σ_ϵ, i.e. $(\epsilon_{t_0}, ..., \epsilon_{t_n}) \sim N(0, \Sigma_\epsilon)$. If measurement errors are independent with a constant variance σ_ϵ^2 then $\Sigma_\epsilon = \sigma_\epsilon^2 I$. Equation (11) implies that the likelihood function can be written as

$$P(\bar{\mathbf{u}} | \Theta) = \psi(\bar{\mathbf{u}} | \lambda(\varphi_\gamma(t_0), ..., \varphi_\gamma(t_n)), \lambda^2 \Sigma + \Sigma_\epsilon). \quad (12)$$

Since for given data $\bar{\mathbf{u}}$ the likelihood function (12) can be numerically evaluated, any likelihood based inference is straightforward to implement. Using Bayes' theorem, the posterior distribution $P(\Theta | \bar{\mathbf{u}})$ satisfies the relation [30]

$$P(\Theta | \bar{\mathbf{u}}) \propto P(\bar{\mathbf{u}} | \Theta) \pi(\Theta). \quad (13)$$

We use the standard Metropolis-Hastings (MH) algorithm [30] to sample from the posterior distribution in (13).

Results and Discussion

In order to study the use of the LNA method for inference we have selected four examples which are related to commonly used quantitative experimental techniques such as measurements based on reporter gene constructs and reporter assays based on Polymerase Chain Reaction (e.g. RT-PCR, Q-PCR). For expository reasons, all case studies consider a model of single gene expression.

Model of single gene expression

Although gene expression involves various biochemical reactions it is essentially modeled in terms of only three biochemical species (DNA, mRNA, protein) and four reaction channels (transcription, mRNA degradation, translation, protein degradation) [31-33]. The stoichiometry matrix has the form

$$S = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad (14)$$

where rows correspond to molecular species and columns to reaction channels. Let $\mathbf{x} = (r, p)$ denote concentrations of mRNA and protein, respectively. For the reaction rates

$$\tilde{\mathbf{f}}(\mathbf{x}) = (k_R(t), \gamma_R r, k_P r, \gamma_P p)^T \quad (15)$$

we can derive the following macroscopic rate equations

$$\dot{\phi}_R = k_R(t) - \gamma_R \phi_R, \quad \dot{\phi}_P = k_P \phi_R - \gamma_P \phi_P. \quad (16)$$

For the general case it is assumed that the transcription rate $k_R(t)$ is time-dependent, reflecting changes in the reg-

ulatory environment of the gene such as the availability of transcription factors or chromatin structure.

Using (14), (15) and (16) in (3) we obtain the following SDEs describing the deviation from the macroscopic state (see section 3.1.4 of Additional file 1 for derivation)

$$dξ_R = -γ_R ξ_R dt + \sqrt{k_R(t) + γ_R φ_R(t)} dW_R$$
$$dξ_P = (k_P ξ_R - γ_P ξ_P)dt + \sqrt{k_P φ_R(t) + γ_P φ_P(t)} dW_P.$$

(17)

We will refer to the model in (16) and (17) as the *simple model* of single gene expression.

In order to test our method on a nonlinear system we will also consider the case of an autoregulated network where the transcription rate of the gene is a function of the concentration of the protein that the gene codes for and where the protein is a transcription factor that inhibits the production of its own mRNA. This is parameterized by a Hill function [31] $k_R(t, p) = k_R(t)/(1 + (p/H)^{n_H}$ where $k_R(t)$ now describes the maximum rate of transcription, H is a dissociation constant and n_H is a Hill coefficient.

Thus, the nonlinear autoregulatory model the system is described by the MRE

$$\dot{φ}_R = k_R(t, φ_P) - γ_R φ_R, \quad \dot{φ}_P = k_P φ_R - γ_P φ_P \quad (18)$$

and the SDEs

$$dξ_R = (k'_R t)ξ_P - γ_R ξ_R)dt + \sqrt{k_R(t) + γ_R φ_R(t)} dW_R$$
$$dξ_P = (k_P ξ_R - γ_P ξ_P)dt + \sqrt{k_P φ_R(t) + γ_P φ_P(t)} dW_P$$

(19)

where $k'_R(t) ≡ ∂k_R(t, φ_P)/∂φ_P$. We refer to this model as the *autoregulatory model* of single gene expression. The two models constitute the basis of our inference studies below.

Inference from fluorescent reporter gene data for the simple model of single gene expression

To test the algorithm we first use the simple model of single gene expression. We generate data according to the stoichiometry matrix (14) and rates (15) using Stochastic Simulation Algorithm (SSA) [34] and sample it at discrete time points. We then generate artificial data that are proportional to the simulated protein data with added normally distributed measurement error with known variance $σ_ε^2$. Furthermore we assume that mRNA levels are unobserved. The volume of the system $Ω$ is unknown

and we put $Ω = 1$ so that concentration equals the number of molecules. Thus the data are of the form

$$\bar{\mathbf{u}} = (u_{t_0}, ..., u_{t_n})^T, \quad (20)$$

where $u_{t_i} = λp_{t_i} + ε_{t_i}$, p_{t_i} is the simulated protein concentration, $λ$ is an unknown proportionality constant and $ε_{t_i}$ is measurement error. For the purpose of our example we model the transcription function by

$$k_R(t) = \begin{cases} b_0 \exp(-b_1(t - b_3)^2) + b_4 & t ≤ b_3 \\ b_0 \exp(-b_2(t - b_3)^2) + b_4 & t > b_3. \end{cases} \quad (21)$$

This form of transcription corresponds to an experiment, where transcription increases for $t ≤ b_3$ as a result of being induced by an environmental stimulus and for $t > b_3$ decreases towards a baseline level b_4.

We assume that at time t_0 ($t_0 << b_3$) the system is in a stationary state. Therefore, the initial condition of the MRE is a function of unknown parameters $(φ_R(t_0), φ_P(t_0)) = (b_4/γ_R, b_4 k_P/γ_R γ_P)$.

To ensure identifiability of all model parameters we assume that informative prior distributions for both degradation rates are available. Priors for all other parameters were specified to be non-informative. To infer the vector of unknown parameters

$$Θ = (γ_R, γ_P, k_P, λ, b_0, b_1, b_2, b_3, b_4)$$

we sample from the posterior distribution

$$P(Θ | \bar{\mathbf{u}}) ∝ P(\bar{\mathbf{u}} | Θ)π(Θ)$$

using the standard MH algorithm. The distribution P($\bar{\mathbf{u}}$ |$Θ$) is given by (12).

The protein level of the simulated trajectory is sampled every 15 minutes and a sample size of 101 points obtained. We perform inference for two simulated data sets: estimate 1 is based on a single trajectory while estimate 2 represents a larger data set using 20 sampled trajectories (see Figure 1A). All prior specifications, parameters used for the simulations and inference results are presented in Table 1A. Estimate 1 demonstrates that it is possible to infer all parameters from a single, short length time series with a realistically achievable time resolution. Estimate 2 shows that usage of the LNA does not seem to result in any significant bias. A bias has not been detected despite the very small number of mRNA molecules (5 to 35 - Figure 2A in Additional file 1) and protein molecules (100 to 500 - Figure 1A). The coefficient of variation var-
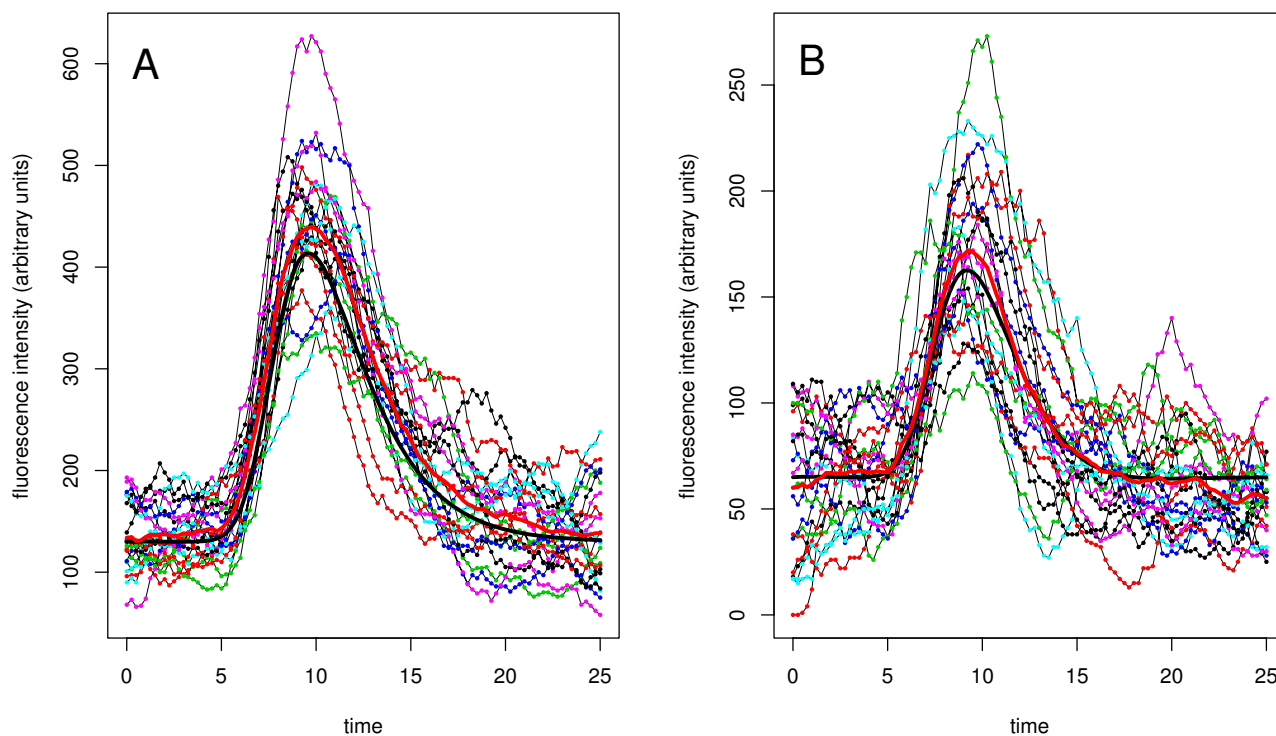
Figure 1

Protein timeseries generated using Gillespie's algorithm for the simple A and autoregulatory B models of single gene expression with added measurement error (σ_ϵ^2 = 9). Initial conditions for mRNA and protein were sampled from Poisson distributions with means equal to the stationary means of the system with equal constant transcription rate b_4. In the autoregulatory case we set $H = b_4 k_P / 2\gamma_R\gamma_P$. In each panel 20 time series are presented. The deterministic and average trajectories are plotted in bold black and red lines respectively. Corresponding mRNA trajectories (not used for inference) are presented in Additional file 1.

ied between approximately 0.15 and 0.4 for both molecular species (Figure 1 in the Additional file 1).

Inference for this model required sampling from the 9 dimensional posterior distribution (number of unknown parameters). If instead one used a diffusion approximation based method it would be necessary to sample from a posterior distribution of much higher dimension (see Additional file 1). In addition, incorporation of the measurement error is straightforward here, whereas for other methods it involves a substantial computational cost [20].

Inference from fluorescent reporter gene data for the model of single gene expression with autoregulation
The following example considers the autoregulatory system with only a small number of reacting molecules. Using SSA we generated artificial data from the single gene expression model with autoregulation. The protein time

courses were then sampled every 15 minutes at 101 discrete points per trajectory (see Figure 1B). As before we assume that the mRNA time courses are not observed and that the protein data are of the form given in (20), i.e. proportional to the actual amount of protein with additive Gaussian measurement error. As in the previous case study we estimate parameters from two simulated data sets, a single trajectory and an ensemble of 20 independent trajectories. The inference results summarized in Table 1B show that despite the low number of mRNA (0-15 molecules, see Figure 2 in Additional file 1) and protein (10-250 molecules, see Figure B) all parameters can be estimated well with appropriate precision.

Inference for PCR based reporter data
In the case of reporter assays based on Polymerase Chain Reaction (e.g. RT-PCR, Q-PCR) measurements are obtained from the extraction of the molecular contents
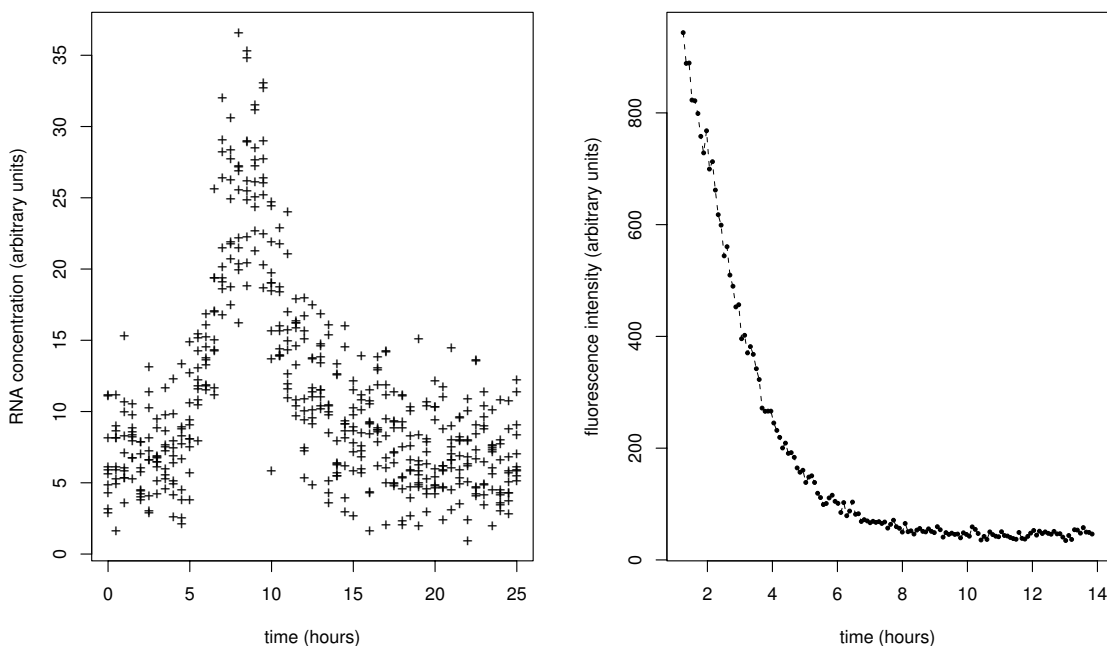
Figure 2
Left: PCR based reporter assay data simulated with Gillespie's algorithm using parameters presented in Table 2 and extracted 51 times (n = 50), every 30 minutes with an independently and normally distributed error (σ_ϵ^2 = 9). Each cross correspond to the end of simulated trajectory, so that the data drawn are of form (22). Since number of RNA molecules is know upto proportionality constant y-axis is in arbitrary units. Time on x-axis is expressed in hours. Estimates inferred form this data are shown in column *Estimate 1* in Table 2. **Right**: Fluorescence level from cycloheximide experiment is plotted against time (in hours). Subsequent dots correspond to measurements taken every 6 minutes.

from the inside of cells. Since the sample is sacrificed, the sequence of measurements are not strictly associated with a stochastic process describing the same evolving unit. Assume that at each time point t_i ($i = 0,..n$) we observe l measurements that are proportional to the number of RNA molecules either from a single cell or from a population of s cells. This gives a $(n + 1) \times l$ matrix of data points

$$\bar{\mathbf{u}} \equiv \{u_{t_i,j}\}_{i=0,\dots n;j=1,\dots,l} \qquad (22)$$

where $u_{t_i,j} = \lambda r_{t_i,j} + \epsilon_{t_i,j}$, $r_{t_i,j}$ is the actual RNA level, λ is the proportionality constant, $\epsilon_{t_i,j}$ is a Gaussian independent measurement error indexed by time t_i. $j = 1,..., l$ indexes the l measurements that are taken at time t_i. Note that $r_{t_i,j}$ and $r_{t_{i+1},j}$ are independent random variables as they refer to different cells. We assume that the dynamics

of RNA is described by the simple model of single gene expression with LNA equations (16) and (17). Let Υ_t denote the distribution of measured RNA at time t ($u_t \sim \Upsilon_t$). In order to accommodate for the different form of data we modify the estimation procedure as follows. For analytical convenience we assumed that the initial distribution is normal $\Upsilon_{t_0} = N(u_{t_0}, \sigma_{t_0}^2)$. This together with eq. (8) and normality of measurement error implies that $\Upsilon_t = N(u_t, \sigma_t^2)$. Simple explicit formulae for μ_t and σ_t^2 are derived in Additional file 1. Since all observations u_{t_i} are independent we can write the posterior distribution as

$$\pi(\Theta \mid \bar{\mathbf{u}}) \propto \prod_{i=0}^{n} \prod_{j=1}^{l} \psi(u_{t_i,j} \mid \mu_{t_i}, \sigma_{t_i}^2)\pi(\Theta), \qquad (23)$$

Table 1: Inference results for (A) the simple model and (B) autoregulatory model of single gene expression

(A) Param.	Prior	Value	Estimate 1	Estimate 2
γ_R	$\Gamma\,(0.44,10^{-2})$	0.44	0.43 (0.27-0.60)	0.49 (0.38-0.61)
γ_P	$\Gamma\,(0.52,10^{-2})$	0.52	0.51 (0.35-0.67)	0.49 (0.38-0.61)
k_P	Exp(100)	10.00	21.09 (3.91-67.17)	11.41 (7.64-16.00)
λ	Exp(100)	1.00	1.42 (0.60-2.57)	1.08 (0.76-1.36)
b_0	Exp(100)	15.00	6.80 (0.97-18.43)	12.78 (9.80-15.33)
b_1	Exp(1)	0.40	0.79 (0.05-3.02)	0.29 (0.18-0.43)
b_2	Exp(1)	0.40	0.77 (0.08-2.79)	0.77 (0.32-1.59)
b_3	Exp(10)	7.00	6.13 (4.41-7.85)	7.25 (6.79-7.55)
b_4	Exp(100)	3.00	0.94 (0.11-2.88)	2.87 (2.11-3.52)

(B) Param.	Prior	Value	Estimate 1	Estimate 2
γ_R	$\Gamma\,(0.44,10^{-2})$	0.44	0.44 (0.27-0.60)	0.42 (0.32-0.54)
γ_P	$\Gamma\,(0.52,10^{-2})$	0.52	0.49 (0.33-0.65)	0.49 (0.36-0.61)
k_P	Exp(100)	10.00	14.86 (3.18-47.97)	9.35 (5.87-13.15)
λ	Exp(100)	1.00	1.26 (0.48-2.30)	1.15 (0.81-1.50)
b_0	Exp(100)	15.00	5.99 (0.20-23.06)	13.47 (9.24-17.13)
b_1	Exp(1)	0.40	0.59 (0.01-2.75)	0.27 (0.14-0.53)
b_2	Exp(1)	0.40	0.92 (0.05-2.92)	0.83 (0.21-3.52)
b_3	Exp(10)	7.00	6.53(0.74-14.69)	7.27 (6.44-7.79)
b_4	Exp(100)	3.00	2.85 (0.35-7.19)	2.64 (1.82-3.32)

Parameter values used in simulation, prior distribution, posterior medians and 95% credibility intervals. Estimate 1 corresponds to inference from single time series, Estimate 2 relates to estimates obtained from 20 independent time series. Data used for inference are plotted in Figure A for case **A** and Figure B for case **B**. Variance of the measurement error was assumed to be known = 9. Rates are per hour. Parameters are n_H = 1, H = 61.98 in case **B**.

where $\psi(\,\cdot\mid u_{t_i},\ \sigma_{t_i}^2\,)$ is the normal density with parameters $u_{t_i},\ \sigma_{t_i}^2$. In order to infer the vector of the unknown parameters $\Theta = (\gamma_R,\ \lambda,\ b_0,\ b_1,\ b_2,\ b_3,\ b_4,\ u_{t_0},\ \sigma_{t_0}^2)$ we sample from the posterior using a standard MH algorithm. To test the algorithm we have simulated a small (l = 10, n = 50, plotted in Figure 2) and a large (l = 100, n = 50) data set using SSA algorithm with parameter values given in Table 2. The data were sampled discretely every 30 min-

utes and a standard normal error was added. Initial conditions were sampled from the Poisson distribution with mean b_4/γ_R. The estimation results in Table 2 show that parameters can be inferred well in both cases even though the number of RNA molecules in the generated data is very small (about 5-35 molecules). Since subsequent measurements do not belong to the same stochastic trajectory, estimation for the model presented here is not straightforward with the diffusion approximation based methods.

Table 2: Inference results for PCR based reporter assay simulated data

Parameter	Prior	Value	Estimate 1	Estimate 2
γ_R	Exp(1)	0.44	0.45 (0.35-0.60)	0.46 (0.42-0.50)
λ	Exp(100)	1.00	1.07 (0.90-1.22)	1.01 (0.95-1.05)
b_0	Exp(100)	15.00	13.13 (10.20-15.87)	14.91 (13.86-15.77)
b_1	Exp(1)	0.40	0.29 (0.14-0.51)	0.43 (0.32-0.54)
b_2	Exp(1)	0.40	0.32 (0.12-0.93)	0.32 (0.21-0.43)
b_3	Exp(10)	7.00	7.05 (6.39-7.63)	6.99 (6.76-7.15)
b_4	Exp(100)	3.00	2.97 (2.00-4.18)	3.10 (2.76-3.43)
u_0	Exp(100)	6.76	6.90 (5.79-7.69)	6.55 (6.14-6.85)
σ_0^2	Exp(100)	6.76	3.52 (0.01-8.99)	7.59 (5.44-9.49)

Parameter values used to generate data, prior distributions used for estimation, posterior median estimates together with 95% credibility intervals. Estimate 1, Estimate 2 columns relate to small (l = 5, n = 50) and large (l = 100, n = 50) sample sizes. Variance of the measurement was assumed to be known σ_ϵ^2 = 4. Estimated rates are per hour.

Estimation of gfp protein degradation rate from cycloheximide experiment

In this section the method is applied to experimental data. After a period of transcriptional induction, translation of gfp was blocked by the addition of cycloheximide (CHX). Details of the experiment are presented in Additional file 1. Fluorescence was imaged every 6 minutes for 12.5 h (see Figure 2). Since inhibition may not be fully efficient we assume that translation may be occurring at a (possibly small) positive rate k_P. The model with the LNA is

$$\begin{aligned} \dot{\phi}_P &= k_P - \gamma_P \phi_P, \\ d\xi_P &= -\gamma_P \phi_P dt + \sqrt{k_P + \gamma_P \phi_P} dW_P. \end{aligned} \quad (24)$$

The observed fluorescence is assumed to be proportional to the signal with proportionality constant λ. For comparison we also consider the diffusion approximation for which an exact transition density can be derived analytically (see Additional file 1 for derivation)

$$dp = (k_P - \gamma_P p)dt + \sqrt{k_P + \gamma_P p} dW_P. \quad (25)$$

Since incorporation of measurement error for the diffusion approximation based model is not straightforward, we assume that measurements were taken without any error to ensure fair comparison between the two approaches. Table 3 shows that estimates obtained with both methods are not very different.

Conclusion

The aim of this paper is to suggest the LNA as a useful and novel approach to the inference of biochemical kinetics parameters. Its major advantage is that an explicit formula for the likelihood can be derived even for systems with unobserved variables and data with additional measurement error. In contrast to the more established diffusion approximation based methods [9,20] the computationally costly methods of data augmentation to approximate transition densities and to integrate out unobserved model variables are not necessary. Furthermore, this method can also accommodate measurement error in a straightforward way.

The suggested procedure here is implemented in a Bayesian framework using MCMC simulation to generate posterior distributions. The LNA has previously been studied in the context of approximating Poisson birth and death processes [22-24,35] and it was shown that for a large class of models the LNA provides an excellent approximation. Furthermore, in [35] it is shown that for the systems with linear reaction rates the first two moments of the transition densities resulting from the CME and the LNA are equal. Here we propose using the LNA directly for inference and provide evidence that the resulting method can give very good results even if the number of reacting molecules is very small. In our previous study [10] we have presented differences between fitting deterministic and stochastic models, where we used diffusion approximation based method. Our experience from that work and from study [20] is that implementation of diffusion approximation based methods is challenging especially for data that are sparsely sampled in time because the need for imputation of unobserved time points leads to a very high dimensionality of the posterior distribution. This usually results in highly autocorrelated traces affecting the speed of convergence of the Markov chain. Our method considerably reduces the dimension of the posterior distribution to the number of unknown parameters of a model only and is independent of the number of unobserved components (see Additional file 1). Nevertheless it can only be applied to the systems with sufficiently large volume, where fluctuations around a deterministic state are relatively close to the mean.

Authors' contributions

MK proposed and implemented the algorithm. CVH performed the cycloheximide experiment. MK wrote the paper with assistance from BF and DAR, who supervised the study.

Additional material

Additional file 1
Supplemental information. Supplementary information contains derivation of the theoretical results, details about algorithm implementation and comparison with the inference method based on the diffusion approximation.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-343-S1.PDF]

Table 3: Inference results for CHX experimental data

Param.	Prior	Estimate LNA	Estimate DA
γ_P	Exp(1)	0.45 (0.31-0.62)	0.53 (0.39-0.67)
k_P	Exp(50)	0.32(0.10-1.75)	0.43 (0.16-1.07)
λ	Exp(50)	22.79(13.79-36.92)	23.85(16.31-36.54)
$\tilde{\phi}_P(0)$	$N(u_{t_0}, u_{t_0})$	889.03(831.44-945.34)	-

Priors, posterior mean and 95% credibility intervals obtained from CHX experimental data using the LNA approach and diffusion approximation approach. Estimation with the LNA involved one more parameter $\tilde{\phi}_P(0) = \lambda \phi_P(0)$. Estimated rates are per hour.

Acknowledgements

References

1. Ehrenberg M, Elf J, Aurell E, Sandberg R, Tegner J: **Systems Biology Is Taking Off.** *Genome Res* 2003, **13(11):**2377-2380.
2. Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic Gene Expression in a Single Cell.** *Science* 2002, **297(5584):**1183-1186.
3. Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, *et al.*: **Oscillations in NF-kappaB Signaling Control the Dynamics of Gene Expression.** *Science* 2004, **306(5696):**704-708.
4. Xie SX, Choi PJ, Li GW, Lee NK, Lia G: **Single-Molecule Approach to Molecular Biology in Living Bacterial Cells.** *Annual Review of Biophysics* 2008, **37:**417-444.
5. Raser JM, O'Shea EK: **Noise in Gene Expression: Origins, Consequences, and Control.** *Science* 2005, **309(5743):**2010-2013.
6. Keizer J: *Statistical Thermodynamics of Nonequilibrium Processes* Springer, New York; 1987.
7. Guptasarma P: **Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli?** *Bioessays* 1995, **17(11):**987-97.
8. Moles CG, Mendes P, Banga JR: **Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods.** *Genome Res* 2003, **13(11):**2467-2474.
9. Golightly A, Wilkinson DJ: **Bayesian Inference for Stochastic Kinetic Models Using a Diffusion Approximation.** *Biometrics* 2005, **61(3):**781-788.
10. Finkenstadt B, Heron E, Komorowski M, Edwards K, Tang S, Harper C, Davis J, White M, Millar A, Rand D: **Reconstruction of transcriptional dynamics from gene reporter data using differential equations.** *Bioinformatics* 2008, **24(24):**2901.
11. Gillespie DT: **A Rigorous Derivation of the Chemical Master Equation.** *Physica A* 1992, **188(1-3):**404-425.
12. Van Kampen N: *Stochastic Processes in Physics and Chemistry. North Holland* 2006.
13. Mendes P, Kell D: **Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.** *Bioinformatics* 1998, **14(10):**869-883.
14. Ramsay JO, Hooker G, Campbell D, Cao J: **Parameter estimation for differential equations: a generalized smoothing approach.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007, **69(5):**741-796.
15. Esposito W, Floudas C: **Global Optimization for the Parameter Estimation of Differential-Algebraic Systems.** *Industrial and Engineering Chemistry Research* 2000, **39(5):**1291-1310.
16. Reinker S, Altman R, Timmer J: **Parameter estimation in stochastic biochemical reactions.** *Systems Biology, IEE Proceedings* 2006, **153(4):**168-178.
17. Tian T, Xu S, Gao J, Burrage K: **Simulated maximum likelihood method for estimating kinetic rates in gene expression.** *Bioinformatics* 2007, **23:**84.
18. Boys R, Wilkinson D, Kirkwood T: **Bayesian inference for a discretely observed stochastic kinetic model.** *Statistics and Computing* 2008, **18(2):**125-135.
19. Wilkinson D: **Stochastic modelling for quantitative description of heterogeneous biological systems.** *Nature Reviews Genetics* 2009, **10(2):**122-133.
20. Heron EA, Finkenstadt B, Rand DA: **Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study.** *Bioinformatics* 2007, **23(19):**2596-2603.
21. Elerian O, Chib S, Shephard N: **Likelihood Inference for Discretely Observed Nonlinear Diffusions.** *Econometrica* 2001, **69(4):**959-993.
22. Elf J, Ehrenberg M: **Fast Evaluation of Fluctuations in Biochemical Networks With the Linear Noise Approximation.** *Genome Res* 2003, **13(11):**2475-2484.
23. Lars F, Per L, Andreas H: **A Hierarchy of Approximations of the Master Equation Scaled by a Size Parameter.** *Journal of Scientific Computing* 2007, **34(2):**127-151.
24. Kurtz TG: **The Relationship between Stochastic and Deterministic Models for Chemical Reactions.** *The Journal of Chemical Physics* 1972, **57(7):**2976-2978.
25. Arnold L: *Stochastic differential equations: theory and applications* Wiley-Interscience; 1974.
26. Oksendal B: *Stochastic differential equations: an introduction with applications* 3rd edition. Springer; 1992.
27. Brockwell P, Davis R: *Introduction to time series and forecasting* Springer New York; 2002.
28. Ronen M, Rosenberg R, Shraiman BI, Alon U: **Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(16):**10555-10560.
29. Wu JQ, Pollard TD: **Counting Cytokinesis Proteins Globally and Locally in Fission Yeast.** *Science* 2005, **310(5746):**310-314.
30. Gamerman D, Lopes HF: *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference* 2nd edition. Chapman & Hall/CRC; 2006.
31. Thattai M, van Oudenaarden A: **Intrinsic noise in gene regulatory networks.** *Proceedings of the National Academy of Sciences* 2001. 151588598
32. Chabot JR, Pedraza JM, Luitel P, van Oudenaarden A: **Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock.** *Nature* 2007, **450:**1249-1252.
33. Komorowski M, Miekisz J, Kierzek A: **Translational Repression Contributes Greater Noise to Gene Expression than Transcriptional Repression.** *Biophysical Journal* 2009, **96(2):**.
34. Gillespie DT: **Exact stochastic simulation of coupled chemical reactions.** *Journal of Physical Chemistry* 1977, **81(25):**2340-2361.
35. Ryota T, Hidenori K, J KT, Kazuyuki A: **Multivariate analysis of noise in genetic regulatory networks.** *Journal of Theoretical Biology* 2004, **229(4):**501-521.