Computers and Chemical Engineering 90 (2016) 111–120

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Regularized maximum likelihood estimation of sparse stochastic monomolecular biochemical reaction networks☆

Hong Jang[a], Kwang-Ki K. Kim[b], Richard D. Braatz[c], R. Bhushan Gopaluni[d], Jay H. Lee[a,*]

[a] Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
[b] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta 30332, GA, USA
[c] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge 02139, MA, USA
[d] Department of Chemical and Biological Engineering, University of British Columbia, 2360 East Mall, Vancouver BC V6T1Z3, Canada

ARTICLE INFO

Article history:
Received 5 November 2015
Received in revised form 21 March 2016
Accepted 21 March 2016
Available online 26 April 2016

Keywords:
Sparse parameter estimation
Regularized maximum likelihood estimation
Mono-molecular biochemical reaction network
Chemical master equation
Stochastic simulation algorithm

ABSTRACT

A sparse parameter matrix estimation method is proposed for identifying a stochastic monomolecular biochemical reaction network system. Identification of a reaction network can be achieved by estimating a sparse parameter matrix containing the reaction network structure and kinetics information. Stochastic dynamics of a biochemical reaction network system is usually modeled by a chemical master equation (CME) describing the time evolution of probability distributions for all possible states. This paper considers closed monomolecular reaction systems for which an exact analytical solution of the corresponding chemical master equation can be derived. The estimation method presented in this paper incorporates the closed-form solution into a regularized maximum likelihood estimation (MLE) for which model complexity is penalized. A simulation result is provided to verify performance improvement of regularized MLE over least-square estimation (LSE), which is based on a deterministic mass-average model, in the case of a small population size.

1. Introduction

Stochastic dynamics of biological systems have lately received increased attention from researchers in the field of biological engineering (Raj and van Oudenaarden, 2009). In the past, such studies were greatly hampered by lack of qualitative measurement data. Nowadays, quantitative but noisy data can be obtained using the microarray technology. Recent developments in sensing techniques that can provide real-time observations of intrinsic stochastic dynamics at small length scales have motivated many scientific investigations (Raj et al., 2010; Xie et al., 2008). One such sensing technique for biological systems is bio-imaging using fluorescent proteins (García-Parajó et al., 2001). By grafting a fluorescent protein into gene expression, proteins and mRNA expressions originating from targeted DNA can be detected quantitatively in real time. Specifically, yellow fluorescent protein (yfp) (Elowitz et al., 2002; Li and Xie, 2011; Yu et al., 2006) has been widely used for detecting changes with single-macromolecule sensitivity in individual live cells.

In the real-time data reported in the aforementioned papers, strongly stochastic behavior has been observed. For example, bursts of transcribed protein molecules from the cell, controlled by an identical messenger RNA molecule, have different copy numbers (Elowitz et al., 2002). The total population number of species in the system within the detectable range is small, ranging from tens to thousands of copies. Stochastic dynamics of systems with discrete states can be modeled by the chemical master equation (CME) (Feinberg, 1979; Fichthorn and Weinberg, 1991),

$$\frac{\partial P(\boldsymbol{\sigma}, t)}{\partial t} = \sum_{\boldsymbol{\sigma}'} W\left(\boldsymbol{\sigma}', \boldsymbol{\sigma}\right) P\left(\boldsymbol{\sigma}', t\right) - \sum_{\boldsymbol{\sigma}'} W\left(\boldsymbol{\sigma}, \boldsymbol{\sigma}'\right) P(\boldsymbol{\sigma}, t) \tag{1}$$

where $P(\boldsymbol{\sigma}, t)$ is the probability of the system being in discrete state $\boldsymbol{\sigma}$ at time t, and $W(\boldsymbol{\sigma}', \boldsymbol{\sigma})$ is the transition rate from state $\boldsymbol{\sigma}'$ to state $\boldsymbol{\sigma}$. The CME describes the time evolution of the probability distri-

☆ Preliminary results in this manuscript were published in conference papers "H. Jang, K. K. K. Kim, J. H. Lee, and R. D. Braatz, Regularized Maximum Likelihood Estimation of Sparse Stochastic Monomolecular Biochemical Reaction Networks, Proceedings of the IFAC World Congress, Cape Town, South Africa, 2014" and "K. K. K. Kim, H. Jang, R. B. Gopaluni, J. H. Lee and R. D. Braatz, Sparse Identification in Chemical Master Equations for Monomolecular Reaction Networks, Proceedings of the American Control Conference, Portland, Oregon, 2014."

* Corresponding author. Fax: +82 42 350 3910.
E-mail address: jayhlee@kaist.ac.kr (J.H. Lee).

http://dx.doi.org/10.1016/j.compchemeng.2016.03.018
0098-1354/

bution among all possible configurations, and can be written as (MacNamara et al., 2008; Munsky and Khammash, 2006)

$$\frac{d\boldsymbol{P}(t)}{dt} = \boldsymbol{A}\left(t;\theta\right)\boldsymbol{P}(t) \tag{2}$$

where $\boldsymbol{P}(t)$ is the state vector containing all the state probability variables and $\boldsymbol{A}\left(t;\theta\right)$ is a matrix containing all the transition rate constants with dependence on the model parameter vector θ related with physicochemical phenomena, e.g., biochemical reactions. Many numerical algorithms have been developed for solving the matrix ordinary differential Eq. (2), which can be divided into direct methods (MacNamara et al., 2008; Munsky and Khammash, 2006) and indirect methods (Gibson and Bruck, 2000; Gillespie, 1977). Direct methods, such as the finite state projection (FSP) algorithm, attempt to evaluate the matrix exponential directly. In practice, given the large size of the state space, indirect methods that use stochastic simulation algorithms (SSAs) to generate approximate probability distributions have been more popular.

A reaction network is one of fundamental models for describing biological mechanisms. Different types of biological reaction networks exist and interact to accomplish needed functions including gene regulatory networks, metabolic networks, and signaling networks. Network identification is an important aspect of studying biological network systems. Given time series experimental data from sensors, a parameter estimation method can be used to identify the parameters of a given reaction network model. Typically, the estimation is formulated to find parameter values minimizing the difference between the experimental data and their model predictions. Most of the literature has employed least-squares estimation (LSE) approaches, which fit stochastic data to a deterministic continuum model (Gennemark and Wedelin, 2009).

The LSE method based on deterministic models, however, can provide poor parameter estimates for highly stochastic systems (Tian et al., 2007). Several stochastic parameter estimation methods based on the stochastic differential equation of type (1) have shown improved estimation performance (Munsky et al., 2012). These methods attempt to solve for the probability density functions (PDFs) of the CME and use them for estimation. Previous studies employed the moment-based method (Munsky et al., 2009; Zechner et al., 2012), the Bayesian method (Boys et al., 2008; Golightly and Wilkinson, 2011; Lillacci and Khammash, 2010), the maximum likelihood estimation (MLE) method (Daigle et al., 2012; Tian et al., 2007), and the density function distance (DFD) method (Lillacci and Khammash, 2013; Poovathingal and Gunawan, 2010). However, these published methods are based on approximated PDFs of the CME. In many cases, PDFs are approximated using the SSA approach (Gillespie, 2007), which typically demands a very large number of simulations to be performed for an accurate estimation.

In many cases, the exact network structure of a biological system is not known. An important problem in biological systems is to use measurement data to identify the topology of reaction networks and estimate associated parameters at the same time. Craciun and Pantea, 2008 discussed the issue of identifiability of reaction networks with the fact that there might exist more than one parameter set that fits the measurement data with the same level of accuracy. To cope with such challenges, model discrimination/invalidation of reaction networks has been studied (Conradi et al., 2005; Kremling et al., 2004). Non-uniqueness of an estimation algorithm can be overcome by using the principle of parsimony and choosing simple models over complex models (Jefferys and Berger, 1991). An approach to reduce model complexity is to use penalization term for the number of parameters such as ridge and ℓ_1 regularization (Hesterberg et al., 2008).

Taking the above issues into consideration altogether, this paper considers stochastic monomolecular reaction systems in a sparse parameter matrix estimation problem. Stochastic monomolecular reaction system, described by an exact probability distribution solution of the CME (Jahnke and Huisinga, 2007), can represent realistic gene regulatory networks or metabolic networks. The exact solution enables formulation of a regularized MLE method rather than LSE. Improved performance over the LSE method is verified with a simulation study involving a small scale reaction network system. Applicability of the proposed MLE method to a more stochastic and larger scale reaction network system is also tested.

2. Biochemical reaction network system

In a living organism, most biological functions arise from complex interactions between numerous components such as genes, metabolites, and proteins. The interactions form a large biochemical reaction network system involving thousands of components. Gene expression is the main mechanism by which cells regulate the interaction webs to perform functions. The gene expression process occurs in two steps: (1) transcription of genes into mRNAs initiated by specialized proteins and (2) translation of mRNAs into biochemically active proteins. By detecting the gene expression time-profiles, we can construct a model of the gene–gene interaction which can be a subset of a larger network (Fig. 1) (Gardner et al., 2003). This gene network identification has important potential applications, for example, in drug discovery to identify candidate pathways to be targeted (Schreiber, 2000).

Gene regulatory networks as well as other complex biochemical reaction networks, such as metabolic networks (Feist et al., 2008; Jeong et al., 2000) and signal networks (Hyduke and Palsson, 2010), can be described by a combination of monomolecular reactions (Radulescu et al., 2012). Possible monomolecular reactions can be categorized with conversion, inflow, and outflow reactions with a set of $n \in \mathbb{N}$ different species or complexes denoted by S_i, $i = 1, \ldots, n$. The reactions are given by

$$\begin{aligned} &S_i \xrightarrow{k_{ij}} S_j \quad \text{Conversion}(i \neq j) \\ &S_o \xrightarrow{k_{0j}} S_j \quad \text{Inflow} \\ &S_i \xrightarrow{k_{i0}} S_* \quad \text{Outflow} \end{aligned} \tag{3}$$

where S_o and S_* are pseudo-species outside the system and k_{ij} is a nonnegative rate constant for reaction from S_i to S_j for $i \neq j$ and can be time-varying. The monomolecular conversion reaction excludes catalytic or splitting reactions of the types

$$\begin{aligned} &S_i \xrightarrow{k} S_i + S_j \quad \text{Catalytic}(i \neq j) \\ &S_i \xrightarrow{k} S_j + S_r \quad \text{Splitting}(i \neq j \neq r) \end{aligned} \tag{4}$$

If the number of species in the system is sufficiently large, the dynamics of the system can be described by the deterministic ordinary differential equations

$$\frac{\mathrm{d}C_i(t)}{\mathrm{d}t} = \sum_{j \neq i} k_{ji}(t)\, C_j(t) - \sum_{j \neq i} k_{ij}(t)\, C_i(t), \; i = 1, \ldots, n \tag{5}$$

where $C_i(t)$ is the population density or concentration of the species S_i and continuous variable. Eq. (5) can be written in a vector form as

$$\frac{\mathrm{d}\boldsymbol{C}(t)}{\mathrm{d}t} = \boldsymbol{A}(t)\,\boldsymbol{C}(t) \tag{6}$$

$$[\boldsymbol{A}]_{ij}(t) = k_{ji}(t), \; j \neq i \tag{7}$$

$$[\boldsymbol{A}]_{ii}(t) = -\sum_{j \neq i} k_{ij}(t) \tag{8}$$