BioSystems 162 (2017) 135–146

Contents lists available at ScienceDirect

BioSystems

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

Slow update stochastic simulation algorithms for modeling complex biochemical networks

Debraj Ghosh, Rajat K. De*

Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 26 October 2016
Received in revised form 17 June 2017
Accepted 20 October 2017

Keywords:
Gillespie algorithm
B cell receptor signaling network
FcϵRI signaling network
Stochastic modeling

ABSTRACT

The stochastic simulation algorithm (SSA) based modeling is a well recognized approach to predict the stochastic behavior of biological networks. The stochastic simulation of large complex biochemical networks is a challenge as it takes a large amount of time for simulation due to high update cost. In order to reduce the propensity update cost, we proposed two algorithms: slow update exact stochastic simulation algorithm (SUESSA) and slow update exact sorting stochastic simulation algorithm (SUESSSA). We applied cache-based linear search (CBLS) in these two algorithms for improving the search operation for finding reactions to be executed. Data structure used for incorporating CBLS is very simple and the cost of maintaining this during propensity update operation is very low. Hence, time taken during propensity updates, for simulating strongly coupled networks, is very fast; which leads to reduction of total simulation time. SUESSA and SUESSSA are not only restricted to elementary reactions, they support higher order reactions too.

We used linear chain model and colloidal aggregation model to perform a comparative analysis of the performances of our methods with the existing algorithms. We also compared the performances of our methods with the existing ones, for large biochemical networks including B cell receptor and FcϵRI signaling networks.

1. Introduction

The dynamic behavior of large complex biochemical networks introduces difficulties in understanding the state space of entities and their interactions. In order to overcome these difficulties, some mathematical modeling techniques including ordinary differential equations (ODEs) have been introduced. ODEs have extensively been used for capturing the time evolution of a biochemical system. However, ODE based methods do not exhibit the system's stochastic behavior which has been proven to be an important feature depending on the nature of the system. The stochastic behavior can exhibit the mechanism of complex biochemical processes, for example, cell to cell variations of a system depending on the copy numbers of the species in the system (Cao et al., 2004). Stochastic simulation algorithms (SSAs), including direct method (DM) and first reaction method (FRM), have been introduced for capturing the stochastic behavior of a biochemical network (Gillespie, 1976, 1977). In DM and FRM, a biochemical network is assumed to be generated through a well stirred (spatially homogeneous) biochemical system in which each reaction is associated with a parameter termed as propensity. The propensity of each reaction is proportional to the probability of occurrence of that reaction. Based on the parameter propensity, DM and FRM decide which reaction to occur and when to occur. The computational complexities of these two methods are extremely high for simulating large complex biochemical networks. In order to get rid of this issue, several algorithms have been developed.

A priority queue based algorithm, called next reaction method (NRM), has been developed, which is an improved version of FRM (Gibson and Bruck, 2000). In NRM, the computational cost has been reduced by introducing a dependency graph for propensity updates. Another improved version, named as optimized direct method (ODM), has been developed in Cao et al. (2004). ODM sorts the reactions of the network, based on their rates, by performing a pre-simulation. The computational costs towards propensity updates in ODM are less compared to NRM, as NRM needs extra cost of maintaining the priority queue. Sorting direct method (SDM), another version of DM, has been introduced in McCollum et al. (2006). It does not need any pre-simulation, rather it maintains an array holding the reaction indexes and sorts it gradually, based on the reaction rates, during simulation. Some approximation methods, including τ-leaping, R-leaping, K-leap, L-leap,

* Corresponding author.
E-mail address: rajat@isical.ac.in (R.K. De).

https://doi.org/10.1016/j.biosystems.2017.10.011
0303-2647/

slow-scale method, k_α-leaping and implicit τ-leaping methods, have been introduced in Gillespie (2001), Gillespie (2007), Gillespie and Petzold (2003), Cao et al. (2005a,b, 2006), Peng et al. (2007), Auger et al. (2006), Peng and Wang (2007), Cai and Xu (2007), and Rathinam et al. (2003). These approximation methods reduce the duration of the total simulation time by jumping over the sequences of less critical reactions in the reaction firing event.

The methods including ODM and SDM have computational complexities of the order of the number of reactions, and they are basically suitable for modeling weakly coupled networks, e.g., the linear chain model. However, the biochemical networks are mostly complex and strongly coupled in which the number of reactions are much higher than the number of species. Therefore, the methods having computational complexities of the order of the number of reactions take longer simulation times for modeling strongly coupled networks. In order to overcome this issue, Ramaswamy et al. have introduced the partial-propensity direct method (PDM) and sorting partial-propensity direct method (SPDM) in Ramaswamy et al. (2009), which have reduced the computational scaling upto the order of the number of species, and thereby, achieved a significant speed up in execution time for simulating strongly coupled networks. Some attempts on delay stochastic simulation algorithm (DSSA) which is used for describing transcription and translation processes of gene regulatory systems have been made in Barrio et al. (2006, 2013), Leier et al. (2014), and Leier and Marquez-Lago (2015).

A tree-based search algorithm has been introduced in Thanh and Zunino (2014). This algorithm is based on Huffman tree which has minimized the search depth of finding the next reaction to occur. A composition-rejection based algorithm, viz., SSA-CR, has been introduced, for which computational cost is independent of network size (Slepoy et al., 2008). A partial-propensity method combined with composition rejection (PSSA-CR) has been developed in Ramaswamy and Sbalzarini (2010). In weakly coupled networks, SSA-CR and PSSA-CR execute faster than PDM and SPDM, whereas these composition-rejection based methods take longer execution time in the case of strongly coupled networks. An exact rejection based stochastic simulation algorithm (SSA), called RSSA, has been developed in Thanh et al. (2014). RSSA executes the reactions based on pre-computed propensity bounds, and it avoids frequent propensity updates without compromising the exactness of SSA. Some modified versions of RSSA, called RSSA-Binary, RSSA-Lookup and SRSSA, have been introduced in Thanh et al. (2015). A tree based search operation is maintained in RSSA-Binary, whereas RSSA-Lookup is based on look up table search for finding the reaction to be executed next. SRSSA creates several independent trajectories simultaneously in a single execution run. Another approach involving infrequent propensity updates, called lazy updating method, has been introduced, which has been applied to the sorting direct method (Ehlert and Loewe, 2014).

In this paper, we have developed two stochastic simulation algorithms, viz., slow update exact stochastic simulation algorithm (SUESSA) and slow update exact sorting stochastic simulation algorithm (SUESSSA) for modeling strongly coupled networks. SUESSA and SUESSSA do not rely on partial propensities like PDM and SPDM which are limited to elementary reactions, and they can model networks with higher order reactions. The main advantage of these two algorithms is that they are equipped with fast propensity update operations. For this purpose, we have applied propensity bound infrequent propensity update technique for reducing the number of propensity updates. We have used efficient data structures for reducing their propensity update costs. The search operation for finding the next reaction has been improved by applying cache-based linear search technique. We have used linear chain model, colloidal aggregation model and two large biochemical networks, including B cell receptor signaling network and FcϵRI signaling network for comparing the simulation times of our methods with other SSAs.

The organization of the paper is as follows. We discuss our proposed algorithms in Section 2. The comparative performance analysis and validation of our methods with the existing ones have been discussed in Section 3 and it is followed by discussion in Section 4.

2. Methodology

Let us consider a biochemical system of M molecular species $C_1, \ldots, C_M$, with state vector $\boldsymbol{X}(t) = [X_1(t), \ldots, X_M(t)]^T$ representing the number of molecules (populations) of the species at time t. These species form a biochemical network through their conversions/reactions. There are N reactions $R_1, R_2, \ldots, R_N$ in the network. The system is well stirred (spatially homogeneous) in order to focus only on the populations of the chemical species rather than their individual positions in the system. Each reaction R_i is associated with a parameter, called stochastic rate constant k_i and the corresponding stoichiometry. The probability of occurrence of the reaction R_i at time t within a small duration dt is defined by a parameter $a_i(\boldsymbol{X}(t))dt$, where $\boldsymbol{X}(t) = \boldsymbol{x}$ (Cao et al., 2004; McCollum et al., 2006). The parameter a_i is called the propensity of the reaction R_i, which is the product of the substrate populations and the stochastic rate constant.

The system dynamics is characterized by the chemical master equation (CME) (Gillespie, 1976, 1977) which is given by

$$\frac{\partial P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)}{\partial t} = \sum_{i=1}^{N} [a_i(\boldsymbol{x} - \boldsymbol{v_i})P(\boldsymbol{x} - \boldsymbol{v_i}, t|\boldsymbol{x}_0, t_0) - a_i(\boldsymbol{x})P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)] \quad (1)$$

Here, $\boldsymbol{v_i} = [v_{1i}, \ldots, v_{Mi}]^T$ is the stoichiometry associated with each reaction R_i. That is, $\boldsymbol{v_i}$ is the ith column vector, with the same unit as $\boldsymbol{x}$, of $M \times N$ stoichiometric matrix. $P(\boldsymbol{x}, t|x_0, t_0)$ is the conditional probability that $\boldsymbol{X}(t)$ will be $\boldsymbol{x}$, given the initial species population $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$. It is very difficult to solve the chemical master equation for large networks, and therefore, we use some practical simulation techniques, e.g., stochastic simulation algorithms (SSAs).

SSAs work by performing the following two tasks: (a) finding the index (μ) of the reaction to occur, and (b) finding the time (τ) of occurrence of the reaction having index μ. Let us calculate the total propensity function as

$$a_0(\boldsymbol{x}) = \sum_{i=1}^{N} a_i(\boldsymbol{x}) \quad (2)$$

The time τ of occurrence of the reaction having index μ can be defined by the exponentially distributed random variable and is given by Cao et al. (2004)

$$p(\tau = s) = a_0(\boldsymbol{x}) \exp[-a_0(\boldsymbol{x})s] \quad (3)$$

There is enough justification in literature (Gillespie, 1976, 1977) that $P(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)$ can equivalently be considered as the probability of occurrence of the reaction having index μ. Thus the problem boils down to finding (searching) the reaction indexes, which will be executed next. This search may be done in various ways. Here, we developed a cache-based linear search technique for finding reactions to be executed. Next we describe our proposed algorithms SUESSA and SUESSSA along with their correctness and computational complexities.