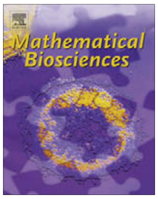
Mathematical Biosciences 221 (2009) 43–53

Contents lists available at ScienceDirect

Mathematical Biosciences

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



Sequential estimation for prescribed statistical accuracy in stochastic simulation of biological systems

Werner Sandmann

Clausthal University of Technology, Department of Mathematics, D-38678 Clausthal-Zellerfeld, Germany

ARTICLE INFO

Article history:
Received 6 January 2009
Received in revised form 4 June 2009
Accepted 25 June 2009
Available online 2 July 2009

Keywords:
Biological systems
Stochastic simulation
Statistical accuracy
Confidence intervals
Sequential estimation

ABSTRACT

Stochastic simulation of biological systems proceeds by repeatedly generating sample paths or trajectories of the underlying stochastic process, from which many relevant and important system properties can be obtained. While a great deal of research is targeted towards accelerated trajectory generation, issues concerned with the variability across trajectories are often neglected. Advanced methods for properly quantifying the statistical accuracy and determining a reasonable number of trajectories are hardly addressed formally in the context of biological system simulation, though mathematical statistics provides a large body of powerful theory. We invoke this theory and show how mathematically well-founded sequential estimation approaches serve for systematically generating enough but not too many trajectories for achieving a certain prescribed accuracy. The practical applicability is demonstrated and illustrated by numerical examples through simulation studies of an immigration-death process and a gene regulatory network.

1. Introduction

Stochastic processes are well suited for modeling various different kinds of biological systems such as population dynamics, epidemics, or chemical reaction kinetics [1–8]. In particular, branching processes, birth-and-death processes, or more general continuous-time Markov chains, where the system dynamics in terms of the state probabilities' time derivatives are described by the Kolmogorov differential equations, have a long tradition in biology.

Stochastic modeling of population dynamics, in particular by birth-and-death processes, has been pioneered by Kendall [9–11] and further advanced by [12,13]. In particular, Kendall [11] showed how to generate realizations of birth-and-death processes, that is, how to simulate them. Similarly, stochastic epidemic modeling starts at the latest in the early 1950s [14,15] and stochastic chemical reaction kinetics can be traced back yet to the 1940s and it occasionally occurred over the following three decades [16–23]. As a milestone, Gillespie [24–26] expressed the Kolmogorov differential equations in terms of the chemical master equation and proved the accordance with the general theory of thermodynamics. Hence, molecular reactions are undoubtedly subject to inherent randomness. Gillespie also proposed stochastic simulation via generating trajectories according to the stochastic dynamics of Markov processes, which is therefore often referred to as Gillespie's algorithm in the corresponding literature. This has significantly promoted the Markovian approach and in conjunction with the growing insight in the importance of accounting for random fluctuations made it widespread in physics, chemistry and biology.

In particular, biological networks consisting of mutually related and interacting components such as metabolic and signaling pathways, protein interactions, or gene regulation are constituted by huge sets of coupled chemical reactions on the molecular level. The quest for analyzing such complex networks in order to gain a system-level understanding of intra- and intercellular dynamics has culminated in the rapid emergence of systems biology, which fosters interdisciplinary research integrating cell biology, molecular biophysics, biophysical chemistry, mathematical and computational approaches [27–32].

Over the years, more and more empirical studies as well as theoretical investigations have provided evidence that randomness must not be neglected but properly taken into account, because in many cases it captures important effects that otherwise cannot be explained at all [33–37]. Since analytical results are seldom available and most of the numerical approaches for solving the Kolmogorov differential equations or the chemical master equation typically fail due to the systems' complexity, stochastic simulation is today prevalent for the analysis of many stochastic biological system models.

Stochastic simulation implies statistical uncertainty in its outcomes. For instance, although Gillespie's algorithm is termed exact, its exactness is only 'in the sense that it takes full account of the fluctuations and correlations' [25] of reactions (or, more generally,

E-mail address: werner.sandmann@tu-clausthal.de

0025-5564/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2009.06.006

state transitions according to events) within a single simulation run. Gillespie pointed out that it is 'necessary to make several simulation runs from time 0 to the chosen time t, all identical with each other except for the initialization of the random number generator'. In fact, the reliability of simulation results strongly depends on a sufficiently large number of simulation runs. The same is valid for simulations of any stochastic process. So, we can abstract away from specific applications and only need to assume that the system at any time $t \geqslant 0$ is described by a (typically multidimensional) random variable $X(t)$ and the systems evolves according to a stochastic process $(X(t))_{t \geqslant 0}$. We do not even require the stochastic process to be Markovian.

In practice, the number of simulation runs is usually chosen very large but somewhat arbitrarily. With an inappropriately chosen number of simulation runs there is a great danger of either wasting computer time or of drawing wrong conclusions from unreliable estimates. More precisely, performing many more runs than necessary means a significant waste of computer time. On the other hand, stopping simulations too early and performing too less simulation runs renders the results meaningless. Hence, it is highly desirable to find ways for avoiding both. Precisely expressing the meaning of a sufficiently large but not unnecessarily large number of simulation runs as well as determining such a suitable number calls for an appropriate formalization in terms of mathematical statistics. Approximate stochastic simulation methods do not address this issue but aim at accelerated trajectory generation for speeding up single simulation runs; repeated simulation runs are still necessary and the systematical choice of a suitable number remains crucial.

We address the statistical accuracy of stochastic simulations, which is of major importance because it is in fact the only mathematical way to investigate the reliability of simulation results. In either case, stochastic simulation tends to be computationally expensive and provides statistical estimates. Mathematically, it constitutes a statistical estimation procedure implying that the results are subject to statistical uncertainty. We consider sequential estimation for obtaining estimates with prescribed statistical accuracy in terms of confidence intervals with prescribed absolute or relative half width. Rather than fixing the number of simulation runs a priori, mathematically well-founded stopping rules apply and automatically terminate the simulation after a suitable number of simulation runs.

In the next section, we provide an appropriate mathematical framework for quantifying the statistical accuracy of stochastic simulation results and the requirements for prescribed statistical accuracy. Sequential estimation procedures as well as related computational issues are presented in Section 3. These procedures are applied to the stochastic simulation of an immigration-death process and a gene regulatory network. Numerical examples are given in Section 4. Finally, Section 5 concludes the paper and discusses topics of further research.

2. Statistical accuracy of stochastic simulation results

The effort for a stochastic simulation is the effort for generating a single trajectory (i.e. performing one simulation run) times the number of required runs in order to obtain reasonably reliable results. Our concern is the largely neglected question how many simulation runs are actually required for sufficiently reliable simulation results, which we shall cast in terms of prescribed statistical accuracy. For this purpose, the first step is to consider a suitable mathematical framework and the second is to come up with well-founded and efficiently implementable techniques such that the prescribed accuracy is safely achieved without substantially wasting computer time.

Mathematically, a stochastic simulation is a statistical estimation using computers. It generates realizations of random variables with the help of random number generators. Similarly as for observations from laboratory experiments, several properties can be derived from the realizations. Thus, from a statistical point of view repeated laboratory experiments and stochastic simulation are equivalent. The only difference is in the way realizations are generated. In a laboratory experiment they are generated within a physical real life environment whereas a stochastic simulation imitates real life environments by using appropriate probabilistic rules.

2.1. Characteristics of stochastic simulation

In practice, each simulation run is finished at some time and the outcome is a finite sequence of states where state changes are triggered by events like births, deaths, reactions, or the like, depending on the specific application. More precisely, a simulation run generates realizations of the system state at event times $t_0, t_1, \ldots, t_k$, which are itself realizations of the random event times $T_0, T_1, \ldots, T_k$ whose differences are usually independent and in the particular case of Markov processes exponentially distributed. The specific termination condition depends on the actual scope of the simulation study. In general, it is determined by a random stopping time defined in accordance with the quantity of interest to be observed in the simulation run. Therefore, also the number of events is in general random. Often, the process $(X(t))_{t \geqslant 0}$ is considered over a predefined time horizon, in which case the random stopping time becomes deterministic while the number of events and the event times are still random. But stochastic simulation is not limited to this case.

We particularly emphasize the generality and the power of stochastic simulation as it is often wrongly claimed that stochastic simulation can only provide estimates for the expectation of $X(t)$ for some t, that is, for the mean number of the quantity represented by $X(t)$ such as the population of molecules, bacteria, plants, animals, humans, or whatever is modeled by the process. This is not true. Stochastic simulation can also provide estimates for, e.g., the time until the population has reached a specific number or has been exhausted (died out), marginal probabilities and even whole probability distributions. In general, we can address any imageable property that has a suitable probabilistic representation meaning that it can be observed and appropriately extracted from trajectories. Probably it is misleading that stochastic simulation typically addresses the estimation of expectations. But in general these are expectations of an arbitrary functional of the underlying stochastic process $(X(t))_{t \geqslant 0}$. In fact, various properties and almost all quantities of potential interest and practical relevance can be mathematically described as a functional of a stochastic process, that is as a measurable function that depends on the trajectories, cf. [38–41]. As a random variable is nothing else than a measurable function, the quantity of interest can be conveniently defined as random variable Y and stochastic simulation then indeed deals with estimating its expectation. We also note that stochastic simulation is an advanced theory in itself that is applied to far more general properties than mean numbers of certain items in diverse scientific domains such as, amongst many others, computer performance evaluation, operations research, insurance risk, mathematical finance, see, e.g., [42–47].

Mathematical statistics comes into play because each time a new realization is generated, it is different in general. That is, there is variability across the trajectories. Also the quantity of interest as observed in any realization will rarely ever exactly coincide with the 'true' value. Statistical methods are required to assure that no wrong conclusions are drawn from a few accidentally untypical experiments. According to classical statistics one builds an estima-