Solving the Chemical Master Equation Using Creeping Windows

Frédéric Didier,[1] Thomas A. Henzinger,[1] Maria Mateescu,[1] and Verena Wolf[1,2]

[1]*EPFL, Switzerland*
[2]*Saarland University, Germany*

Short Abstract – The chemical master equation (CME) is a system of ordinary differential equations that describes the evolution of a network of chemical reactions as a stochastic process. Its solution yields the probability density vector of the system at each point in time. Solving the CME numerically is in many cases computationally expensive or even infeasible as the number of reachable states is huge.

We introduce the creeping window method, which computes an approximate solution of the CME by performing a sequence of local analysis steps. In each step, only a manageable subset of states is considered, representing a "window" in the state space. In subsequent steps, the window follows the direction in which the probability mass moves until the time period of interest has elapsed. We construct the window based on a discretization of the process and add/neglect states on the fly according to their likelihoods.

In order to show the effectiveness of our approach, we apply it to examples of biochemical reaction networks with up to 6 chemical species and 10 reactions. The experimental results show that the proposed method speeds up the analysis considerably, compared to stochastic simulation.

Motivation. The traditional approach for a dynamical model of cellular reaction networks is based on the assumption that the concentrations of the chemical species change continuously and deterministically in time. During the last decade, however, stochastic models with discrete state spaces have seen growing interest [1–8]. The reason is that they take into account the effects of molecular noise in the cell. Molecular noise has a significant influence on important processes such as gene expression [9–14], decisions of the cell fate [15–17], and circadian oscillations [18–20].

The most appropriate modeling approach for systems that are subject to molecular noise is a discrete-state continuous-time Markov process, also called *continuous-time Markov chain* (CTMC). This is particularly evident in the presence of *intrinsic noise* arising from random microscopic events in the cell, such as the location of molecules or the order of the reactions. As opposed to continuous models, the discrete-state stochastic model is able to capture the discreteness of the random events in the cell.

The evolution of the CTMC is given by a master equation that is derived according to Gillespie's theory of stochastic chemical kinetics [21]. Since the state space grows exponentially in the number of involved chemical species, the state space of the CTMC is large, which renders its analysis difficult. Besides the computation of cumulative measures such as expectations and variances of the copy numbers of certain chemical species, the computation of event probabilities is important for several reasons. First, cellular process may decide probabilistically between several possibilities, e.g., in the case of developmental switches [2, 15, 22]. In order to verify, falsify, or refine the mathematical model based on experimental data, the likelihood for each of these possibilities has to be calculated. But also full distributions are of interest, such as the distribution of switching delays [12], the distribution of the time of DNA replication initiation at different origins [23], and the distribution of gene expression products [24]. Finally, many parameter estimation methods require the computation of the posterior distribution because means and variances do not provide enough information to calibrate parameters [25].

Analysis Methods. Two different families of computational approaches have been proposed and used to estimate event probabilities and approximate probability distributions. The first kind of approach is based on numerical simulation, i.e., the generation of many sample trajectories (or *simulation runs*) of the system. The second kind of approach is based on numerical reachability analysis, i.e., the propagation of the probability mass through the state space. The former approach is well-known as *Gillespie simulation* [26], in which pseudo-random numbers are used to simulate molecular noise. Measures of interest are obtained via statistical output analysis. The main advantage of simulation is that it is easy to implement and the generation of trajectories is not limited by the size of the state space. Moreover, the precision level of the method can be easily adjusted by performing more or fewer simulation runs. For the computation of the probability of certain events, however, simulative approaches become computationally expensive, because a large number of runs have to be carried out to bound the statistical error. For estimating event probabilities, a higher precision level is necessary than for estimating cumulative measures such as expectations, and simulation becomes expensive because doubling the precision requires four times more simulation runs.

In contrast, approaches based on a numerical reachability analysis approximate probability distributions of the CTMC. As opposed to a statistical estimation of probabilities, which yields an indirect solution, the master equation is numerically solved by integrating the system's behavior over time. Standard numerical techniques are impractical for many systems because of the enor-

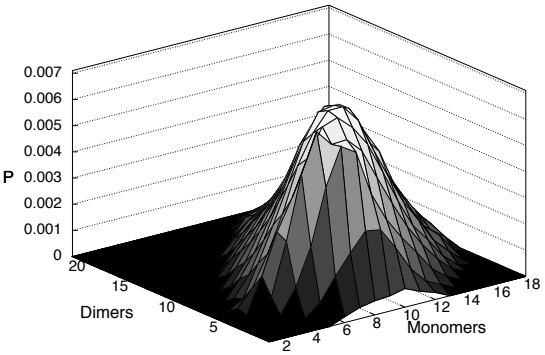

FIG. 1: Probability distribution of monomers and dimers in the phage λ model.

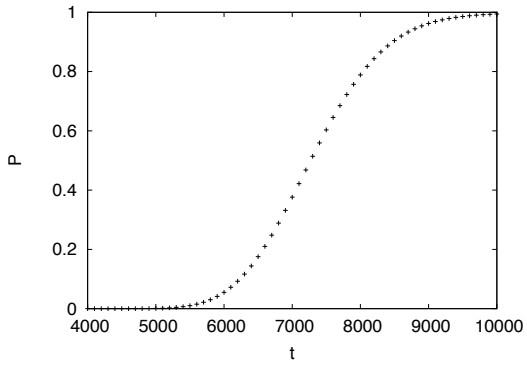

FIG. 2: Cumulative probability distribution of the time until the number of proteins reaches 500 for the first time in the gene expression example.

mous size of the state space. Recently, however, more sophisticated numerical approximation methods have been proposed, which solve the system in an iterative fashion and consider only subsets of the state space during any given time interval [27–29]. They are significantly more efficient than global analysis because they use localization optimizations (such as "sliding windows") and dynamic adaptation ("on-the-fly" generation of windows). These methods efficiently compute the probability distribution of large CTMC at several time instances up to a small approximation error.

Creeping Window Algorithm. We propose the *creeping window algorithm* (CWA), which performs a sequence of local analysis steps on dynamically constructed abstractions of the system. We use adaptive uniformization [30] to discretize the underlying process. We iteratively solve the discrete model and in each step we add/neglect states on the fly depending on their likelihood. Other approaches for an approximate numerical solution of the underlying Markov chains can be found in [28, 29]. They differ from our approach in that they compute a finite projection of the state space that is based solely on the structure of the underlying graph. In our method, we add and neglect states in an on-the-fly fashion based on the stochastic properties of the Markov chain. Therefore, we consider a significantly smaller set of states during a certain time interval, without being less accurate. The projection algorithms include all states that are reachable within a fixed path depth. In our algorithm, for each single state, we dynamically decide if it significantly contributes to the overall solution or not. We found that this dynamic adaptation of the analysis to be essential for efficiency.

Experimental Results For our experimental results, we consider two examples from biology and compare the running times of our algorithm with Gillespie simulation. In order to achieve an appropriate statistical accuracy with simulation, we assume a confidence level of 95% and bound the relative width of the confidence interval by 0.2. By assuming that the smallest event probability that has to be estimated is γ all results of the simulation have a "precision" of at least γ. Intuitively, we simulate often enough to reason about events that occur with a probability of at least γ. We therefore refer to γ as the *single event error*. As opposed to that, the *total approximation error* of the numerical method is the sum of the approximation errors of all states considered during the computation.

Our first example is a model of the transcription regulation of a repressor protein in bacteriophage λ [31]. This protein is responsible for maintaining lysogeny of the λ virus in E. coli [15]. We compute the full probability distribution at several time instances for different precision levels. The length of the time horizon is 300. Fig. 1 shows a plot of the distribution of dimers and monomers at time instant $t = 300$. Below we list the running times of our numerical method as well as the running time of the simulation.

CWA		simulation	
running time	total approx. error	running time	single event error
55 min 5 sec	3×10^{-6}	> 6000 h	10^{-8}
39 min 16 sec	2×10^{-5}	> 500 h	10^{-7}
25 min 2 sec	2×10^{-4}	67 h 22 min	10^{-6}
15 min 41 sec	1×10^{-3}	6 h 44 min	10^{-5}
6 min 33 sec	7×10^{-3}	40 min	10^{-4}
3 min 12 sec	4×10^{-2}	4 min	10^{-3}

Our second example is a model for transcription of a gene into messenger RNA (mRNA), and subsequent translation of the latter into proteins [14]. We calculate the distribution of the time until the number of produced proteins exceeds 500. We compute the probability that at least 500 proteins are in the system at 100 equidistant time instances. Fig 2 shows the cumulative probability distribution of the time until the number of proteins reaches 500 for the first time (note that eventually the threshold of 500 is reached with probability one). Below,

we list the results for the gene expression example.

CWA		simulation	
running time	total approx. error	running time	single event error
4.2 sec	5×10^{-6}	> 500 h	10^{-7}
3.6 sec	5×10^{-5}	> 50 h	10^{-6}
3.0 sec	5×10^{-4}	5 h 3 min	10^{-5}
2.4 sec	4×10^{-3}	30 min 18 sec	10^{-4}
1.9 sec	4×10^{-2}	3 min sec	10^{-3}

Discussion. Even if we consider the total approximation error δ as a rough bound for the single error of each state probability, thus favoring simulation, the speed-up factor of the numerical approximation is large, especially if the precision increases. The necessary precision level up to which probability distributions are approximated may depend on the system under study. It is, however, important to note that the occurrence of rare biochemical events can have important effects. For instance, the spontaneous, epigenetic switching rate from the lysogenic state to the lytic state in phage λ-infected E. coli is experimentally estimated to be in the order of 10^{-7} per cell per generation [32].

Conclusion We have demonstrated that, for the computation of event probabilities, the creeping window algorithm provides an efficient alternative to simulation-based methods.

Even though simulation is widely used, the advantages of numerical methods increase as more sophisticated techniques become available. They reduce the computational effort, especially if accurate results are desired. Moreover, for the calibration of parameters many instances of the model have to be solved and in this case short running times for a single solution are necessary.

[1] H. H. McAdams and A. Arkin, Trends in Genetics **15**, 65 (1999).
[2] C. Rao, D. Wolf, and A. Arkin, Nature **420**, 231 (2002).
[3] R. Srivastava, L. You, J. Summers, and J. Yin, Journal of Theoretical Biology **218**, 309 (2002).
[4] N. Fedoroff and W. Fontana, Science **297**, 1129 (2002).
[5] J. Paulsson, Nature **427**, 415 (2004).
[6] P. S. Swain, M. B. Elowitz, and E. D. Siggia, PNAS, USA **99**, 12795 (2002).
[7] T. E. Turner, S. Schnell, and K. Burrage, Computational Biology and Chemistry **28**, 165 (2004).
[8] D. J. Wilkinson, *Stochastic Modelling for Systems Biology* (Chapman & Hall, 2006).
[9] A. Kierzek, J. Zaim, and P. Zielenkiewicz, Journal of Biological Chemistry **276**, 8165 (2001).
[10] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, Nature **422**, 633 (2003).
[11] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, Nature Genetics **31**, 69 (2002).
[12] H. H. McAdams and A. Arkin, PNAS, USA **94**, 814 (1997).
[13] M. B. Elowitz, M. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).
[14] M. Thattai and A. van Oudenaarden, PNAS, USA **98**, 8614 (2001), ISSN 0027-8424.
[15] A. Arkin, J. Ross, and H. H. McAdams, Genetics **149**, 1633 (1998).
[16] H. Maamar, A. Raj, and D. Dubnau, Science **317**, 526 (2007).
[17] R. Losick and C. Desplan, Science **320**, 65 (2008).
[18] D. Gonze, J. Halloy, and A. Goldbeter, PNAS, USA **99**, 673 (2002).
[19] N. Barkai and S. Leibler, Nature **403**, 267 (2000).
[20] D. Gonze, J. Halloy, and A. Goldbeter, Quantum Chemistry **98**, 228 (2004).
[21] D. T. Gillespie, *Markov Processes* (Academic Press, N. Y., 1992).
[22] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins, PNAS USA **97**, 2075 (2000).
[23] P. Patel, B. Arcangioli, S. Baker, A. Bensimon, and N. Rhind, Mol Biol Cell **17**, 308 (2006).
[24] A. Warmflash and A. Dinner, PNAS **105**, 17262 (2008).
[25] D. A. Henderson, R. J. Boys, C. J. Proctor, and D. J. Wilkinson, in *Handbook of Applied Bayesian Analysis*, edited by A. O'Hagan and M. West (Oxford University Press, 2009).
[26] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[27] T. Henzinger, M. Mateescu, and V. Wolf, in *Proc. CAV* (Springer, 2009), LNCS, to appear.
[28] B. Munsky and M. Khammash, J. Chem. Phys. **124**, 044144 (2006).
[29] K. Burrage, M. Hegland, F. Macnamara, and R. Sidje, in *Proc. of the Markov 150th Anniversary Conference* (Boson Books, 2006), pp. 21–38.
[30] A. van Moorsel and W. Sanders, ORSA Communications in Statistics: Stochastic Models **10**, 619 (1994).
[31] J. Goutsias, J. Chem. Phys. **122**, 184102 (2005).
[32] J. W. Little, D. P. Shepley, and D. W. Wert, The EMBO Journal **18**, 4299 (1999).