Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

An optimal algorithm for enumerating state space of stochastic molecular networks with small copy numbers of molecules

Youfang Cao and Jie Liang

Abstract— **Stochasticity plays central role in molecular networks of small copy numbers, including those important in protein synthesis and gene regulation. The combination of copy numbers of molecular species defines the microscopic state of molecular interactions. With this formulation, nonlinear reactions can be effectively modeled through chemical master equations. However, currently little is known about the state space associated with stochastic networks, other than the defeatist admission that it is exponentially large. There is neither closed-form solution nor computational algorithm that can effectively characterize the state space of molecular networks. Such a characterization is a prerequisite for directly solving the chemical master equation. In this study, we describe an algorithm that can exhaustively characterize all possible states of a molecular networks with small copy numbers of species for a given initial condition. Our algorithm works for networks of arbitrary stoichiometry, and is optimal in both storage and time complexity. It allows the approach of solving chemical master equation to be applicable to a larger class of stochastic molecular networks. We show an example of application of our method to the MAPK cascade network.**

I. INTRODUCTION

Networks of interacting biomolecules are at the heart of the regulation of cellular processes. The temporal dynamics of molecular networks are often modeled using coupled ordinary differential equations (ODEs) based on macroscopic reaction rates. These models can effectively account for behavior of average concentrations of molecules, and have found wide applications in networks where concentrations of interacting molecules are large, and fluctuations are negligible.

However, there are many situations where proteins and mRNAs in a cell have low copy numbers. For example, the regulation of transcriptions depends on the binding of a single protein to a promoter site. The synthesis of protein peptides on ribosome also involve small copy numbers of molecular species. In such biological processes where nanomolar concentrations of molecules interact, fluctuations due to the stochastic behavior intrinsic in low copy number events play important and essential roles [3]. For these processes, ODEs are inappropriate and stochasticity needs to be considered.

The importance of stochasticity in cellular functions is well recognized. Studies of simple genetic switches and cascade models show that stochasticity plays central roles in magnifying signal, sharpening discrimination, and in inducing bistability [1, 6–8]. Noise is also found to actively facilitate molecular communication in cells. Therefore, understanding the stochastic nature and its consequences of cellular processes involving molecular species of small copy numbers is a fundamental problem in studying molecular networks.

The chemical master equation provides the framework that account for full stochasticity. By treating microscopic states of reactants explicitly, this formulation can model non-linear reactions effectively. The challenging problem is to study a realistic system beyond simple toggles and switches that involve a nontrivial number of species. To approximate the master equation, Fokker-Planck or Langevin equations can be obtained by adding Gaussian stochastic terms to a deterministic equation. Although they do not account for full stochasticity, they are applicable when a modest number of molecules are involved. A different approach is not to solve the master equation directly, but to carry out Monte Carlo simulations using the Gillespie algorithm [2, 7]. This approach has found wide applications, although it cannot guarantee an account of full stochasticity. It is also challenging to sample adequately when the network becomes complex. As a single simulation follows high probability path, this method is not efficient to explore rare events. It is also difficult to determine whether a simulation is extensive enough to obtain accurate statistics.

Another approach is to solve the chemical master equation directly. This approach can account for full stochasticity in small copy number events. However, a critical issue is the lack of efficient computational methods. A challenging problem is that a method realistic, and detailed enumeration and characterization of the state space of molecular interactions is currently lacking. Here the combination of the copy numbers of all molecular species define a state of the reaction system. Since the state space is usually thought to be

Y. Cao is with the Shanghai Center for Systems Biomedicine, Shanghai Jiaotong University, 800 Dongchuan Road, 200240 Shanghai, China. yfcao@sjtu.edu.cn

J. Liang is with the Department of Bioengineering, SEO MC-063, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, Chicago, IL 60607–7052, USA, and the Shanghai Center for Systems Biomedicine, Shanghai Jiaotong University, 800 Dongchuan Road, 200240 Shanghai, China. jliang@uic.edu

Correspondence should be addressed to: jliang@uic.edu.

exponentially large, the application of directly solving the chemical master equations is precluded for many realistic systems. In this paper, we study the problem of enumerating the state space of networks of molecular interactions.

Several routes can be followed to enumerate and characterize the state space of molecular networks. One obvious route is through simulation. In the spirit of the Gillespie algorithm [2], one simply follows explicitly simulated reaction events to whatever state of copy numbers one reaches. However, although one often reaches the most frequented states, this approach cannot guarantee that all important reachable states will be explored, therefore cannot guarantee the full characterization of rare events.

Another simple route is to predefine the maximum copy number of the reactants. The state space will then be bounded by the product of a maximum number. The size of state spaces produced with this method will be inflated and will be enormous. For example, if there are 10 species, and there is a total maximum of 6 molecules at any time in the system. This naive method will not take into consideration of the details of the network, and the state space will have $(6)^{10} = 60,466,176 \approx 6.05 \times 10^7$ states, as we cannot rule out *a priori* the possibility that any species at some time may reach the maximum copy number of 6. This method is intrinsically inefficient. First, there may be many states which may never be visited. For some states, no reactions may occur and therefore are not needed. For others, no reactions can lead to them. With this approach, the size of the state space rapidly becomes unnecessarily the bottleneck for computation. As a result, one can only study a very small network with small copy numbers of molecular species.

In this study, we address the problem of defining the state space of reactions involving a small copy number of molecular species in a molecular network. We describe an optimal algorithm that gives description of the state space and the set of transitions optimal in both space and time complexity. That is, all states reachable from an initial condition will be accounted for, and no irrelevant states will be included. All possible transitions will be recorded, and no infeasible transitions will be encountered. As a result, the state-transition matrix used in formulating a chemical master equation obtained by this algorithm is compact and efficient, with no redundant information, and is of the minimal size. In addition, the computational time is also optimal up to a constant.

II. METHODS

A. The Algorithm

Suppose we have a biological model, which contains m molecular species and can have n reactions. Given an initial condition, namely, the copy numbers of each of the m molecular species, we aim to calculate all states that the biological system can reach starting from this initial condition. These states collectively constitute the state space of the network under this initial condition.

Formally, we have a biological model $\boldsymbol{M} = (\boldsymbol{S}, \boldsymbol{R})$, with m number of molecular species: $\boldsymbol{S} = (S_1, \ldots S_m)$, whose copy numbers is specified as $S_1, \ldots$, and S_m, and n reactions: $\boldsymbol{R} = \{R_i | i = 1, \ldots, n\}$. Here an reaction can involve an arbitrary number (≥ 1 and $\leq m$) of molecular species, with any arbitrary nonzero positive integer coefficient (*i.e.*, arbitrary stoichiometry). The state space $\boldsymbol{X}$ is $\boldsymbol{X} = \{\boldsymbol{S}\}$, namely, the set of m-tuples of copy numbers for each of the m molecular species. The set of allowed transitions are $\boldsymbol{T} = \{t_{ij}\}$. We are given with an initial condition: $\boldsymbol{S}^{t=0} = (S_1^0 = s_1, S_2^0 = s_2, \ldots, S_m^0 = s_m)$, where s_i is the initial copy number of the i-th molecular species at time $t = 0$.

The algorithm is written as Algorithm 1.

Algorithm 1 State Enumerator

```
Biological model M ← (S, R);
Initial condition: S^{t=0} ← (s_1, s_2, ..., s_m)
Initialize the state space: X ← ∅;
Initialize the set of transitions: T ← ∅;
Stack ST ← ∅;    Push(ST, S_0);
while ST ≠ ∅ do
  S_i ← Pop(ST);
  for j = 1 to n do
    if reaction R_j occur under condition S_i then
      generate state S_{i+R(j)} that is reached by following
      reaction R_j from S_i;
      if t_{i,i+R(j)} ∉ T then
        T ← T ∪ t_{i,i+R(j)};
      end if
    end if
    if (S_{i+R(j)} ∉ X) then
      X ← X ∪ S_{i+R(j)};
      Push(ST, S_{i+R(j)})
    end if
  end for
end while
Output X and T.
```

The algorithm performs the following computation. After initialization, we start with the initial state $\boldsymbol{S}^{t=0}$. We examine each reaction in turn to determine if this reaction can occur for this state. If so, we generate the state that this reaction leads to. If it was not encountered before, we add it to our collection of states for the state space, and declare this as a new state. We repeat this for all new states, which is maintained by a stack data structure. The algorithm terminates when all new states are exhausted.

B. Correctness and Optimality

The algorithm will terminate. The state space and the transitions under a given initial condition can be considered as a directed graph $G = (\boldsymbol{S}, \boldsymbol{T})$, in which

vertices are the state vectors, *i.e.*, the set of reachable states $\boldsymbol{S}$, or the m-tuples of copy numbers of the m molecular species. Edges are the set of allowed transitions $\boldsymbol{T}$ between the states, *i.e.*, reactions connecting two state vertices: Two vertices $s_i \in \boldsymbol{S}$ and $s_j \in \boldsymbol{S}$ are connected by a directed edge $t_{i,j} \in \boldsymbol{T}$ if and only if s_i can be transformed to s_j through a reaction $R_k \in \boldsymbol{R}$. Any reachable state can be transformed from the initial state by one or more steps of reactions, and the directed graph G is a connected graph.

Our algorithm implicitly generates this graph G. Assume the algorithm will not terminate in finite number of steps. Because the set of reactions $\boldsymbol{R}$ is finite, G has a limited tree-width at any steps away from the initial condition. Then G must have an unlimited depth. That is, there must exist a path p in the graph G that start from the initial state and extend to infinite. According to the algorithm, each state in the path p only appears once. Therefore, p must contain an infinite number of different states. However, this is impossible for a given initial condition, as each molecular species has a limited copy number, and the atomic mass is conserved in the system. We conclude that the algorithm will terminate.

The algorithm gives correct answers, because all states visited can be reached from the initial condition, and all visited states is actually reached as each is brought to by a chemical reaction, except the trivial case of the initial state. In addition, all reachable states will be visited, as the algorithm test at each state all possible reactions, and will only terminates when all new states are exhausted. It is easy to see all possible transitions between states will be recorded.

The time complexity of our algorithm is optimal. Since only unseen state will be pushed onto the stack, every state is pushed and popped at most once. As access of each state and push/pop operations take $O(1)$ time, the total time required for the stack operations is $O(|\boldsymbol{S}|)$. As the algorithm examines each of the n reaction for each reached state, the complexity of total time required is $O(n|\boldsymbol{S}|)$, where n is usually a small constant (*e.g.* < 50).

Using the same reasoning, it is also easy to see that the algorithm is optimal in storage, as only valid states and transitions are recorded. It is also optimal in time complexity, as each state will be generated/visited at most twice before it is popped from the stack.

C. Molecular network model.

We apply our algorithm to the MAPK cascade model (BIOMD28 in BioModels database at EBI (`http://www.ebi.ac.uk/biomodels`) [5]. The SBML (Systems Biology Markup Language) model file is parsed and the molecular species and reactions are extracted. This network contains 16 molecular species with 17 reactions [5]. Abbreviations used in this model are listed in Table I. Fig 1 shows the topology of the model. All 16 molecular species are labeled with numbers from 1 to 16. MEK and MKP3 are the key enzymes catalyzing phosphorylation and dephosphorylation reactions in this network. The rest of the molecular species are substrates, intermediates, and products of MEK and MKP3 induced reactions. Most of the reactions in this model (14 of 17) are second-order.

III. RESULTS

a) Simple initial conditions: We generated the state spaces of the MAPK cascade for different initial conditions and record their sizes. In the first set of calculations, we increase the copy number for one species from 1 to 20, and record the size of resulting state space, while keeping the copy numbers of all other species to 0. We repeat this process for each of the 16 molecular species in turn. Altogether, we have $16 \times 20 = 320$ data points of size of state space. In Fig 2, the x-axis lists the labels of the 16 molecular species, the y-axis the copy number of each species taken in turn, and the z-axis the computed size of the state spaces.

It is clear that different molecular species in this model affect the size of the state space differently. Increasing the copy number of M-MEK-Y, M-MEK-T, and Mpp-MKP3 molecules (species 9, 10 and 11) lead to large state spaces (size $888,030$ at 20 copies), while the initial conditions of 20 copies of any other species result in modest state spaces. For example, species 7, 8, 15 and 16 when given 20 copies have a state-space size of 231. For species 1-6, no reactions can occur at these initial conditions, and the state space contains only the the initial state.

The computing time increases with the copy number, but the state space for each of the 320 initial conditions can be computed within one minute. We found that when any of S_9, S_{10}, or S_{11} has an initial copy of 28 and all others 0 copies, namely, with 28 copies of one of these molecular species initially in the network, the state spaces increases to 6,724,520, and the computing time also increase, although all can be computed within 10 minutes on a Linux workstation.

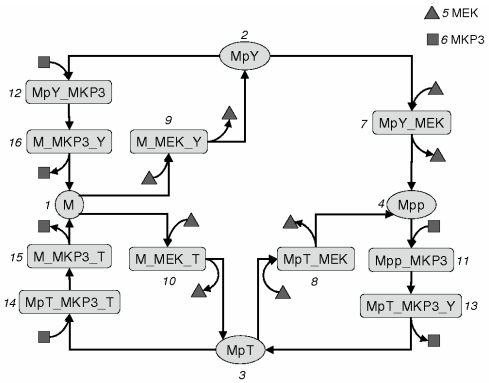

Fig. 1. Illustration of the model BIOMD28, labeled with species number.

TABLE I

ABBREVIATIONS USED IN BIOMD28 NETWORK.

#	Abbreviation	Description
1	M	ERK, extracellular signal-regulated kinase
2	MpY	ERK with Y phosphorylated
3	MpT	ERK with T phosphorylated
4	Mpp	ERK with dual phosphorylated
5	MEK	ERK kinase
6	MKP3	ERK phosphatase
7	MpY_MEK	Binding of MpY and MEK
8	MpT_MEK	Binding of MpT and MEK
9	M_MEK_Y	Binding of M and MEK at Y site
10	M_MEK_T	Binding of M and MEK at T site
11	Mpp_MKP3	Binding of Mpp and MKP3
12	MpY_MKP3	Binding of MpY and MKP3
13	MpT_MKP3_Y	Binding of MpT and MKP3 at Y
14	MpT_MKP3_T	Binding of MpT and MKP3 at T
15	M_MKP3_T	Binding of M and MKP3 at T site
16	M_MKP3_Y	Binding of M and MKP3 at Y site

b) Biological initial conditions: We further test biologically more reasonable initial conditions, in which species M, MEK and MKP3 are all given an equal number of i copies, while all the other species start with zero copies. We increase i from 1 to only 11 due to the limitation of our linux workstation. These initial conditions correspond to a total of $3 \times 1 = 3$ copies to $3 \times 11 = 33$ copies of molecules of three species in the network. The results are shown in Table II.

IV. DISCUSSION

Stochasticity plays important roles in molecular networks for processes involving small copy numbers of molecular species. By considering the state space of the combination of copy numbers of all molecular species, stochasticity and nonlinearity of molecular networks can be studied in details by directly solving the chemical master equation.

A prerequisite for studying full stochasticity by solving master equation is a full and accurate characterization of the state space of molecular networks. Such a characterization can also provide the basis for a global probabilistic picture of the nature of molecular interactions at low copy numbers.

We have developed in this work an algorithm to enumerate the state space of chemical master equation that is optimal in storage and time. It can also find all

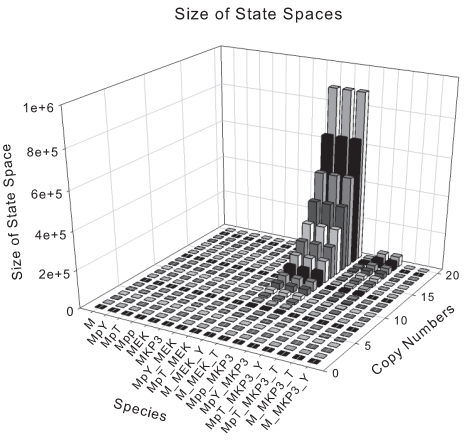

Fig. 2. Sizes of state spaces for a model of the MAPK cascades under the initial condition of 1 to 20 copies of each of the 16 species and 0 other species. Altogether the size of state space for $16 \times 20 = 320$ initial conditions are shown here.

TABLE II

SIZE OF STATE SPACES WITH DIFFERENT M, MEK AND MKP3 COPY NUMBERS.

M	MEK	MKP3	Sizes of state spaces
1	1	1	14
2	2	2	105
3	3	3	560
4	4	4	2,380
5	5	5	8,568
6	6	6	27,132
7	7	7	77,520
8	8	8	203,490
9	9	9	497,420
10	10	10	1,144,066
11	11	11	2,496,144

possible transitions between states, and can be further used to compute transition rates. We show it can generate the full state space for selected initial conditions of MAPK cascade, a network of nontrivial size that is far beyond toggles and switches usually studied for full stochasticity using master equation.

In general, the state space of molecular networks is necessarily large, but our work shows that with an optimal algorithm, and perhaps with judicious choice of biologically motivated initial conditions, the state space of many networks can be fully characterized, and perhaps the full stochasticity can also be studied by solving the chemical master equation using techniques such as in [4] for a wide class of molecular networks.

V. ACKNOWLEDGEMENT

We thank Dr. Bhaskar DasGupta and Hsiao-Mei Lu for helpful discussions. This work is supported by a phase II 985 Project (Sub-Project:T226208001) at Shanghai Jiaotong University, Shanghai, China.

REFERENCES

[1] L. Chen, R. Wang, T. Zhou, and K. Aihara, "Noise-induced cooperative behavior in a multicell system." *Bioinformatics*, vol. 21, pp. 2722–9, 2005.

[2] D.T. Gillespie, "Exact stochastic simulation of coupled chemical reactions." *The Journal of Physical Chemistry*, vol. 81, 1977, pp2340-2361.

[3] J. Hasty, J. Pradines, M. Dolnik, and J. Collins, "Noise-based switches and amplifiers for gene expression." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, pp. 2075–80, 2000.

[4] S. Kachalo, H. Lu, and J. Liang, "Protein folding dynamics via quantification of kinematic energy landscape." *Phys Rev Lett*, vol. 96(5), p. 058106, 2006.

[5] N. Markevich, J. Hoek, and B. Kholodenko, "Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades." *J. Cell. Biol.*, vol. 164, pp. 353–9, 2004.

[6] J. Mettetal, D. Muzzey, J. Pedraza, E. Ozbudak, and A. van Oudenaarden, "Predicting stochastic gene expression dynamics in single cells." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, pp. 7304–9, 2006.

[7] Y. Morishita, T. Kobayashi, and K. Aihara, "An optimal number of molecules for signal amplification and discrimination in a chemical cascade." *Biophys. J.*, vol. 91, pp. 2072–81, 2006.

[8] T. Zhou, L. Chen, and K. Aihara, "Molecular communication through stochastic synchronization induced by extracellular fluctuations." *Phys. Rev. Lett.*, vol. 95, p. 178103, 2005.