Proceedings of Department of Defense High Performance Computing Modernization Program Users Group Conference 2007,
June 18-22, 2007, Pittsburgh, PA

Stochastic Simulations of Cellular Biological Processes

Yaroslav Chushak
Biotechnology HPC Software Applications Institute, HPCMP,
Wright-Patterson AFB, OH
schushak@bioanalysis.org

Brent Foy
Wright State University, Dayton, OH
brent.foy@wright.edu

John Frazier
Air Force Research Laboratory, Wright-Patterson AFB, OH
john.frazier@wpafb.af.mil

Introduction

From a systems engineering point of view, cells consist of a complex set of nested, nonlinear control systems dominated by feed-back and feed-forward loops. The more complex the system is, the more important are the issues concerning the robustness and parameter optimization, therefore modeling and simulations are important for both engineering and reverse-engineering of biosystems.

Objective

At the functional level, all biological processes in cells can be represented as a series of biochemical reactions. Since such reactions are stochastic in nature, the user must run thousands of simulations to characterize the "ensemble" behavior of biological systems. We developed a software package called Biomolecular Network Simulator (BNS) to model and simulate complex biomolecular reaction networks using High Performance Computing (HPC).

Methodology

The Biomolecular Network Simulator uses the Gillespie stochastic algorithm to simulate the evolution of a system of biochemical reactions. The BNS code is a combination of MATLAB and C-coded modules. This combination allows one to use the interactive features and visualization tools of MATLAB, while achieving high speed for the computationally intensive part of the software. The software is parallelized with the MPI library to run on multiprocessor architectures.

Result

The Biomolecular Network Simulator consists of two sets of tools: for simulations of the system and for the analysis of simulation results. The Graphical User Interface of BNS allows users to easily set parameters for the model and simulations and to select analysis method. Multiple types of post-simulation analyses are available.

Significance to DoD

The Developed software allows DoD scientists to build, simulate and analyze complex cellular biomolecular networks utilizing the capacities of HPC. It provides the foundational capability to design and integrate biological constructs into a new generation of biotechnology products.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Stochastic Simulations of Cellular Biological Processes	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory,Wright Patterson AFB,OH,45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of Department of Defense High Performance Computing Modernization Program Users Group Conference 2007, June 18-22, 2007, Pittsburgh, PA

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

1 INTRODUCTION

All chemical reactions require the interaction of the reactants with sufficient energy to overcome the activation energy barrier and, fundamentally, they are stochastic in nature. The best way to model kinetics of a system of chemical reactions is to use a stochastic approach in terms of the Chemical Master Equation (CME), with the number of molecules of each molecular species as variables. The CME describes transitions of the system from one state to another state based on probabilistic methods. Gillespie proposed a method to simulate probabilistically-correct trajectories based on the CME through the use of Monte Carlo methods [1].

Although the Gillespie stochastic algorithm is a method for exact simulations, as it explicitly counts each reaction event that occurred in the system, the accuracy of the method comes at a high computational cost. This is especially true for systems with a large number of molecular species where reactions occur numerous times in a short period of time. To accelerate discrete stochastic simulation, Gillespie proposed the tau-leaping method as an approximate simulation strategy [2]. By using Poisson random numbers, the tau-leaping method can leap over many reactions without a significant loss of accuracy.

We developed a software package – the Biomolecular Network Simulator (BNS) – that can use the Gillespie stochastic algorithm or the tau-leaping algorithm to simulate the behavior of a system of biochemical reactions. It allows scientists to build a synthetic biomolecular network and explore its performance utilizing the capacities of High Performance Computing. BNS contains tools for both simulating the system and for analyzing the results of the simulation. Since the simulations are stochastic in nature, the user must run thousands of simulations to characterize the "ensemble" behavior of biological systems. The parallelized BNS code allows users to run multiple simulations and store results on multiprocessor platforms. In this paper, we present a brief description of the Biomolecular Network Simulator software along with some examples.

2 STOCHASTIC SIMULATION ALGORITHM

Let us consider a system composed of N well mixed chemical species, S_i (i = 1,…N), in a fixed volume V, which are involved in M reactions, R_j (j = 1,…M). The dynamical state of the system can be specified by the state vector $\mathbf{X}(t) = (X_1(t), X_2(t),\ldots X_N(t))$, where $X_i(t)$ is the number of molecules of species S_i at time t.

The probability for each reaction R_j is defined by a propensity function $a_j(\mathbf{X}) = c_j \cdot H_j(\mathbf{X})$, where c_j is the stochastic probability constant and $H_j(\mathbf{X})$ represents the number of possible combinations of reactants. The probability that reaction R_j will occur in the time interval dt is defined as $a_j(\mathbf{X})dt$. Each reaction is also characterized by its state-change vector $\boldsymbol{\nu}_j = (\nu_{1j}, \nu_{2j}, \ldots \nu_{Nj})$, where ν_{ij} is the change in the number of molecules S_i caused by one R_j reaction .

To study the evolution of the state vector $\mathbf{X}(t)$, Gillespie proposed an algorithm for Monte Carlo generation of stochastic trajectories [1]. The direct simulation algorithm, which is implemented in the Biomolecular Network Simulator, answers two questions: (1) which reaction will occur next, and (2) what is the waiting time for the next reaction to occur.

To answer these questions, two random numbers uniformly distributed over the interval (0,1) – r_1 and r_2 – are generated. The first random number is used to determine the next reaction R_j, such that

$$\sum_{i=1}^{j-1} a_i < r_1 \cdot a_0 < \sum_{i=1}^{j} a_i \,, \tag{1}$$

where

$$a_0 = \sum_{j=1}^{M} a_j \,. \tag{2}$$

The distribution of the waiting time is given by following probability density function:

$$P(\tau, j) = a_j \exp(-a_0 \tau) \,. \tag{3}$$

Here, $P(\tau, j)$ is the probability that the waiting time for the reaction is τ and that it will be an R_j reaction. The waiting time for the next reaction is calculated as [1]

$$\tau = \frac{1}{a_0} \log \frac{1}{r_2} \quad . \tag{4}$$

After the next reaction and its waiting time are determined, the reaction is executed and the state of the system is changed according to the state-change vector $\boldsymbol{\nu}_j$. The simulation time is increased then by τ and the next simulation step is generated.

The Gillespie stochastic algorithm is exact for the elementary reactions (uni-uni, uni-bi and bi-uni types of reactions) with any number of molecules in the system. For the system with large numbers of molecules, the trajectories generated by stochastic Monte Carlo simulations converge to the trajectory generated by deterministic differential equations [1].

3 TAU-LEAPING ALGORITHM

The tau-leaping method calculates a time interval τ which encompasses more than one reaction event and satisfies the Leap Condition, i.e., the expected state change induced by the leap must be sufficiently small that no propensity function changes its value by a significant amount. Several methods have been proposed recently to choose the size of the time interval for tau-leaping [3-5]. We implemented the tau-leaping method proposed by Cao et al. [6]. In that method a tau selection formula is given by

$$\tau' = \min_{i \in I_{rs}} \left\{ \frac{\max\{\varepsilon x_i / g_i, 1\}}{|\mu_i(x)|}, \frac{\max\{\varepsilon x_i / g_i, 1\}^2}{\sigma_i^2(x)} \right\}, \tag{5}$$

where g_i is the highest order of reaction in which species S_i appears as reactant, ε is an error control parameter, I_{rs} is the set of indices of all reactant species, and

$$\mu_i(x) = \sum_j \nu_{ij} a_j(x) \,, \tag{6}$$

$$\sigma_i^2(x) = \sum_j \nu_{ij}^2 a_j(x) \,. \tag{7}$$

After the time interval τ' has been selected, the number of firings of each reaction channel during this time interval is approximated as a Poisson random variable. The Poisson random variable can have arbitrary large sample value and it is possible that the population of some of the molecular species can run negative. Therefore, a critical number of molecules n_c (typically in the range of 5-20) was introduced. If the number of molecular species gets less then n_c, all reaction in which that species appears as reactant are defined as critical. These reactions are simulated by the stochastic simulation algorithm. We calculate the sum of propensity functions of all the critical reactions a_0^c and generate a second candidate time τ'' according to

$$\tau'' = \frac{1}{a_o^c} \log \frac{1}{r} \,. \tag{8}$$

The actual time leap τ is selected as the smaller of τ' and τ''. If $\tau = \tau'$, a number of firings k_i is generated as a Poisson random variable with mean $a_i(x)\tau$ for all of the noncritical reaction and reactions are executed; no critical reaction is executed in this case. If $\tau=\tau''$, we generate k_i and fire noncritical reactions as in the previous case; also, another random number with uniform distribution is generated to find which critical reaction needs to be fired according to Eq. (1).

Evaluations of the performance of the tau-leaping algorithm shows a 2-3 fold speed up of simulations compared to the exact stochastic algorithm while maintaining excellent accuracy with regard to both rare and frequent events.

4 BIOMOLECULAR NETWORK SIMULATOR

The Biomolecular Network Simulator uses a combination of MATLAB and C-coded modules. The front-end, graphical user interface (GUI) and analysis tools of BNS are written in MATLAB, while the simulation core engine is written in C. Such a combination allows one to use the interactive features and visualization tools of MATLAB, while achieving high speed for the computationally intensive part of the software. The parallelization of the code is done with the help of the MPI library. The BNS can be run on any computer platform where MATLAB 6.5 or newer is installed.

4.1 Input Data

A model is a set of mathematical relationships that describe the behavior of biochemical reactions that control cellular biological processes. Each of the 'Model' directories contains one or more subdirectories with model description files and/or different sets of parameters for the same model and an 'Output' directory where the results of simulations are stored. There are two types of model directories: one for models defined in the Systems Biology Markup Language (SBML) format [5] and one for models defined as a set of MATLAB m-files, referred to as the BNS format. In addition, BNS allows one to perform simulations with multiple parameter sets, with each parameter set being run multiple times. Simulations with multiple parameter sets can be used for optimization and sensitivity analysis of the model.

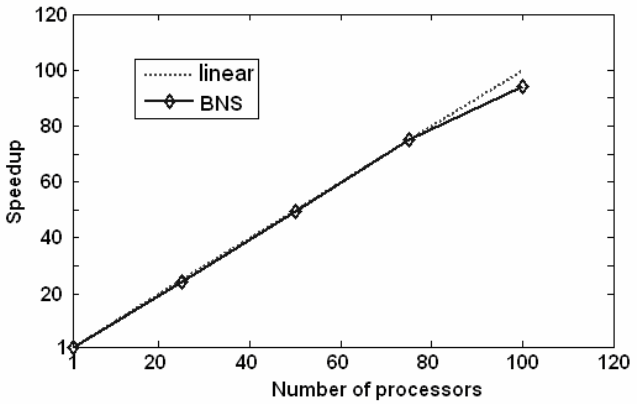

Figure 1. Scaling of BNS with the number of processors.

4.2 Output Data

There are two types of output files: snapshot data and event log data. Snapshot data files contain the state of the system (number of molecules of each molecular species) at user specified time intervals. The information stored in the snapshot files are used to create runtime interactive graphics and for *post hoc*

analysis of the data. The second type of output files – the event log files – contain the record of every discrete event that occurs during the simulation. The user should be aware that event log files may require considerable memory or hard disk space and, therefore, may create memory management problems for simulations involving a large number of long runs or for large reaction networks.

4.3 Parallelization

The parallelization of the BNS code is accomplished using the MPI library. In our parallelization scheme, the 'master' processor divides the total number of user specified runs between the available processors, sends a set of jobs to each of the 'workers' and performs some of the simulation runs itself. In this approach to parallelization, sometimes called "embarrassingly parallel" we reduce the communication between the nodes and increase the speedup of multiple simulations. To test BNS scaling with the number of processors, we ran 1000 simulations on an SGI Origin 3900 machine at ASC MSRC, which has shared memory architecture. A simulation here is defined as a single run of the mathematical model of a particular biochemical reaction network. A 92-fold speedup was observed by running BNS on 100 processors (Figure 1).

4.4 Running the Simulations

The BNS can be run either in command line mode or via a GUI. The GUI allows the user to modify model parameters at runtime and to execute simulations in the interactive or batch mode on HPC resources.

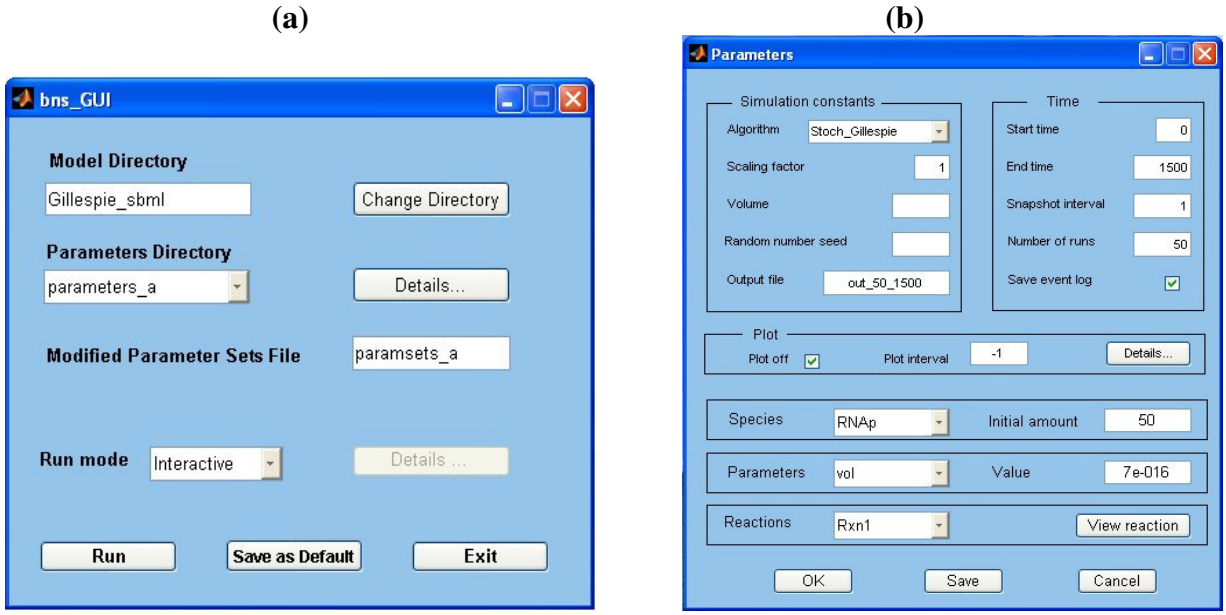

Figure 2. The screen shots of the BNS GUI dialog windows. **(a)** The main BNS dialog window. **(b)** The parameters dialog window which allows users to modify the model parameters and to set simulation parameters.

The main dialog window of the BNS GUI is shown in Figure 2. It allows the user to select the appropriate 'Model' and 'Parameters' directories and set the 'Run' mode. A click on the 'Details' button

next to the 'Parameters' directory opens a new window, shown in Figure 2(b). This dialog window allows the user to modify model parameters and to set parameters for the simulation.

If simulations are run in the interactive mode, partial results of the simulations will appear on the screen during the run. Usually, HPC centers allocate limited resources (in number of processors and running time) for interactive simulations, therefore BNS can be run in 'Batch' mode. In this mode all output data are stored on the hard drive for further analysis.

4.5 Analysis

The Biomolecular Network Simulator has a comprehensive set of tools for post-simulation analysis. A GUI for the analysis tools allows the user to easily select the data and to set conditions for the analysis. Multiple types of post-simulation analyses are available.

4.5.1 Plots of number of molecules vs. time

The most frequently used type of analysis is a plot of the number of molecules vs. time. Such plots are available in the interactive mode or as post-simulation analysis. There are two ways to create plots: each compound is plotted on a separate figure or multiple compounds are plotted on the same figure window (grouping mode). The number of molecules vs. time plots can be created with both types of output files: snapshot data or event log data.

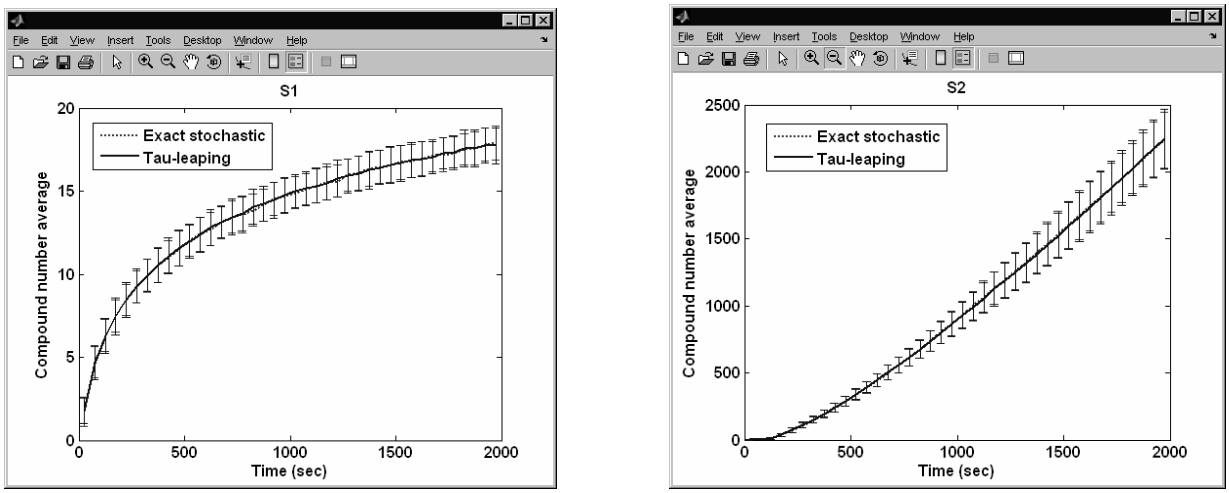

Figure 3. The averaged number of compounds S1 and S2 in the simulation interval (0, 2000) obtained using the exact Gillespie stochastic algorithm and approximate tau-leaping algorithm (ε=0.1, n_c =5).The time-weighted average was calculated using a 50-sec time-averaging interval and the individual time-weighted averages were averaged over the 200 simulation runs. Data for the mean ± SD are shown.

4.5.2 Time-weighted average analysis

A time-weighted average analysis refers to the calculations of the average number of molecules of a particular molecular species during a user selected time-averaging interval. The average is weighted according to the amount of time the compound exists in each state during the selected time-averaging interval. The time weighted average is then plotted versus time. The averaging analysis can be performed for a

single run or for a selected set of runs. When the analysis is applied to multiple runs, the plot shows the between run average (the average of each individual time-weighted average) and standard deviation.

The graphs in Figure 3 show the behavior of two molecular species, S1 and S2, over the time interval of 2000 seconds for a biomolecular reaction network containing transcription, translation and metabolic reactions. The simulations were carried out using two different algorithms implemented in BNS: Gillespie stochastic algorithm and tau-leaping algorithm with parameters $\varepsilon = 0.1$ and $n_c = 5$. Figure 3 shows the between run average of the time-weighted average number of molecules for 200 runs using a time-averaging interval of 50 sec. The average number of molecules of species S1 changed in the range of ~0-20 molecules, while the average number of molecules of species S2 changed from 0 to ~2500 molecules. Results obtained using an approximate tau-leaping algorithm are almost indistinguishable from the result of exact Gillespie algorithm for both types of species. On the other hand, using the tau-leaping algorithm speeds up simulations more than 3-fold compared to the exact stochastic algorithm.

4.5.3 Reaction frequency analysis

Complex biomolecular reaction networks usually contain reactions that occur on different time scales: some reactions have a low propensity and occur rarely; other reactions are very fast and occur frequently. The data stored in the event log files allow the user to perform various reaction frequency analyses of the simulation data to learn more about the basic nature of the system. One type of analysis creates plots of the total number of times each reaction in the network occurred during the simulation.

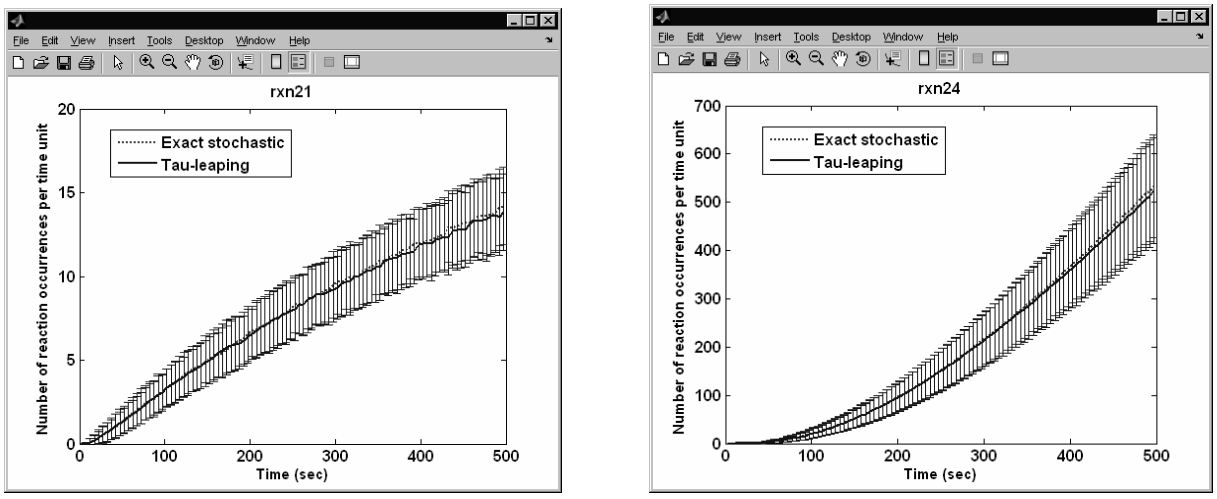

Figure 4. The averaged number of reaction occurrences per time-averaging interval for the reaction channels rxn21 and rxn24 obtained using the exact Gillespie stochastic algorithm and approximate tau-leaping algorithm ($\varepsilon = 0.1$, $n_c = 5$). The time-averaged rate was calculated using a 5 sec time-averaging interval and the individual time-averaged rates were averaged over the 200 simulation runs. Data for the mean ± SD are shown.

A second type of analysis is the average and standard deviation of the time-averaged reaction frequency in each reaction channel. Figure 4 shows an example of the time-averaged reaction frequency for two reaction channels averaged over the 200 runs obtained by running simulations with the two algorithms. The reaction rxn21 belongs to the group of "slow" reactions with the frequency of occurrences in

the range of 0-20 firings per sec. The reaction rxn24 is a "fast" reaction and reached 500-600 occurrences per sec. As in the case of average number of molecules, the tau-leaping algorithm shows an excellent agreement with the results from exact stochastic simulations for this particular biomolecular reaction network.

5 CONCLUSIONS

The Biomolecular Network Simulator allows the users to simulate the behavior of complex biological processes utilizing the capacities of high performance computers. Some of the features that distinguish BNS from similar tools are:

- usage of MATLAB and C-coded functions allows the user to combine intensive visualization of data with high speed computations;
- parallelized code for multiple simultaneous simulations allows the user to run BNS on multiprocessor machines;
- options to run the code in the interactive or batch mode;
- user friendly graphical user interface allows the user to easily set and modify parameters of the model, simulation and analysis; and
- comprehensive tool sets provide for post-simulation analysis of results.

ACKNOWLEDGMENT

The work of Yaroslav Chushak was sponsored by the US Department of Defense High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) initiative. The work of Brent Foy was made possible by a grant from the Air Force Office of Scientific Research (AFOSR) and by the Air Force Summer Faculty Fellowship Program.

DISCLAIMER

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or the US Department of Defense. This paper has been approved for public release; distribution is unlimited.

REFERENCES

[1] Gillespie, D., "Exact Stochastic Simulation of Coupled Chemical Reactions." *J. Phys. Chem*. v. 81, p. 2340 (1977) .

[2] Gillespie, D., 2001, "Approximate accelerated stochastic simulations of chemically reacting systems." *J. Chem. Phys*. v. 115, p. 1716 (2001).

[3] Gillespie, D. and Petzold, L., "Improved leap-size selection for accelerated stochastic simulations." *J. Chem. Phys*. v. 119, p.8229 (2003).

[4] Tian, T. and Burrage, K. "Binomial leap methods for simulation chemical kinetics." *J. Chem. Phys*. v. 121, p.10356 (2004).

[5] Chatterjee, A., Vlachos, D. and Katsoulakis, M. "Binomial distribution based τ-leap accelerated stochastic simulation." *J. Chem. Phys*. v. 122, 024112 (2005).

[6] Cao, Y., Gillespie, D. and Petzold, L. "Efficient step size selection for the tau-leaping simulation method." *J. Chem. Phys*. v.124, 044109 (2006).

[5] Huska, M., Finney, A., Sauro H. M., Bolouri, H. et al., 2003, "The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models." *Bioinformatics*, v. 19, no. 4: 524-531.