High performance model exploration of mutation patterns in an agent-based model of colorectal cancer

Jonathan Ozik, Nicholson Collier, Justin Wozniak and Charles Macal
Argonne National Laboratory
Argonne, IL USA

Chase Cockrell, Melinda Stack and Gary An
Department of Surgery
University of Chicago
Chicago, IL USA

***Abstract*— Agent-based models (ABMs) integrate multiple scales of behavior and data to produce higher-order dynamic phenomena and are increasingly used in the study of cancer. However, the complexity of ABMs provides numerous challenges to their effective use, mostly related to the relatively high computational cost in carrying out the simulation experiments by which ABMs are developed, calibrated and used. High-performance computing (HPC) platforms can address some of these computational constraints. We have developed a framework, called Extreme-scale Model Exploration with Swift/T (EMEWS), that can leverage the computing capabilities of HPC parallel architectures by integrating model exploration (ME) modules such as machine learning and evolutionary computing methods to augment the performance of large-scale simulation experiments. EMEWS can be used to aid in the calibration, parameter estimation and model exploration of any simulation-model. Herein we provide a use case examining the factors and patterns of mutational events of oncogenesis in population level simulations of a mechanism-based ABM of colorectal cancer (CRC).**

Keywords—agent-based model; calibration; colorectal cancer; model exploration; oncogenesis; parameter estimation; workflow

I. Adaptive Model Exploration for ABMs

Agent-based modeling is an object-oriented, discrete-event, rule-based, spatially-explicit, stochastic modeling method that maps well to biology [1]–[3]. However, the use of agent-based models (ABMs) has several challenges [2]–[4]. The current development and use of ABMs typically requires the execution of many model runs to account for stochastic variation in model outputs as well as to explore the possible range of model outcomes under alternative parameter settings and experimental conditions [2]. The need for *adapting* such model exploration (ME) over time comes up in several situations: design of simulation experiments [5], [6], simulation exploration of model parameter space either locally or globally, and for simulation optimization in which the goal is to find a set of parameter values that maximize an objective or achieve satisfactory levels [7]. Adaptive ME is made more complex by the stochastic nature of the underlying simulations [8]. Currently, the "adaptive" aspect of exploring parameter/behavior space is done manually, with researchers updating how parameter sets are to be modified for the next set of batch runs. This process is a mainstay in the use of ABMs, but in practice this approach is only able to sample a small fraction of the model's total parameter space (that subset pre-selected by the user). From a practical standpoint simulations are generally aimed at finding at least one subset of parameter space that will allow sufficient calibration and "validation" of the underlying ABM. In such cases there should not be a presupposition of the uniqueness of such a solution, and it should be recognized that there may be additional, more robust or "interesting" regions of parameter space that are not being characterized. The limitations of this approach are accentuated with increasingly large and complex ABMs; practical and computational constraints invariably result in smaller and more limited sampling of the range of potentially rich behaviors from such ABMs.

In order increase the potential benefits from agent-based modeling, we have created the Extreme-scale Model Exploration with Swift/T (EMEWS) framework [9] EMEWS, which is built on Swift/T [10] offers the capability to run very large, highly concurrent ensembles of simulations of varying types while supporting a wide class of ME algorithms, including those increasingly available to the community via Python and R libraries. Furthermore, it offers a software sustainability solution, in that ME studies based around EMEWS can easily be compared and distributed. A central EMEWS design goal is to ease software integration while providing scalability to the largest scale (petascale plus) supercomputers, running millions of ABMs, thousands at a time. Initial scaling studies of EMEWS have shown robust scalability [11]. The tools are also easy to install and run on an ordinary laptop, requiring only an MPI (Message Passing Interface) implementation, which can be easily obtained from common OS package repositories. As a demonstration of its potential utility we are using EMEWS to examine the dynamic patterns of mutational events in an ABM of oncogenesis of colorectal cancer that can generate incidences of cancer that match epidemiological data from the SEER cancer database.

II. Oncogenesis: Mechanisms to Populations

We assert that cancers as generated from baseline tissue dynamics when mutational events lead to dysfunction of normal cellular processes, and that the development of cancers is an extremely rare event over the sum total of cellular divisions in an organ/tissue over the lifetime of an individual. We have developed cell-level ABMs that generate dynamically stable baseline tissue behavior, where cellular functions/agent rules are governed by the presence of a corresponding intact "gene" within an array that abstractly represents the genome of the cell.

Cells are exposed to a stochastic but bounded degree of DNA damage per time step, which the cell can repair at a certain rate assuming intact genes governing those functions. Simulations are run over decades of simulated time; genetic damage accumulates over time, leading to loss of functions, which eventually result in the development of cells that display the Hallmarks of Cancer [12]. Since each instance of the ABM represents an individual person, this approach is able to link cellular-genetic events/functions to individual patient trajectories to population scale data. We have applied this modeling approach to discover new genetic drivers of breast cancer oncogenesis [13], the transition from premalignant lesions to breast cancer [14], the selective fitness hierarchies involved in oncogenesis and inflammation [15] and the differences between sporadic CRC and CRC arising from inflammatory bowel disease [16], [17]. This latter work involves the gastrointestinal oncogenesis agent-based model (GIOABM). The GIOABM incorporates "genes" that govern functions critical for normal tissue dynamics that are also implicated in the development of colorectal cancer (APC, Beta-catenin, PIK3CA, DCC, E-cadherin, K-Ras, TGF-beta, p53, telomerase, C-src, SMAD4, and BRAF) grouped by general function to reflect those biologic behaviors regulated by each gene: dysregulated proliferation, increased DNA damage or decreased DNA repair, failure of apoptosis, and uninhibited motility. The GIOABM generated tumors in both the wild-type population and colitis population that matched known epidemiologic rates retrieved from the SEER database and extrapolated from existing epidemiologic studies on ulcerative colitis patients, respectively (Figure 1).

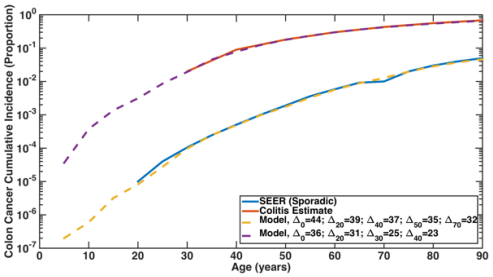

Figure 1: GIOABM matching of SEER incidences of sporadic CRC and extrapolated incidences of inflammatory bowel disease related CRC.

However, in order to generate cancer rates matching those seen in reality it was necessary to change the baseline DNA repair capacity over time. While this is expected, as it is well known (though poorly understood) that the human body's ability to repair DNA damage diminishes with age, it represented an artificial fitting parameter within the GIOABM. Therefore, the GIOABM was modified to explore patterns and network topologies of genetic events governing impaired DNA repair capacity that could generate output matching SEER data.

III. GIOABM Exploration with EMEWS

The synthetic genome of the GIOABM was modified to posit up to 10% of the genome as being able to adversely affect the ability of cells to repair DNA damage, and further posited an additional two hierarchies of controlling genes "upstream" to the effector genes. Search across the various perturbations of this network structure represented a computationally intensive task, and therefore we employed an adaptive model exploration algorithm, using the Python DEAP [18] evolutionary algorithm (EA) library, within the EMEWS framework to increase the efficiency of the search. Here we describe results from our scaling performance runs.

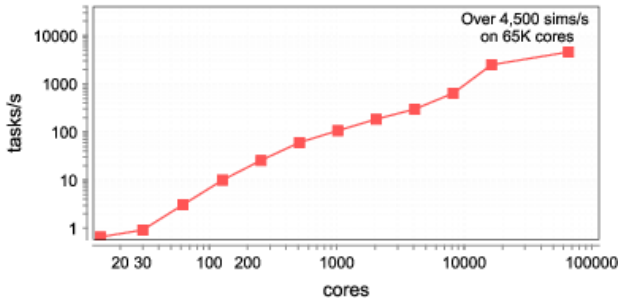

Figure 2: Scaling study results for EMEWS GIOABM calibration workflow on IBM Blue Gene/Q.

Performance results: The EMEWS framework was configured to fit model outputs with ulcerative colitis rates from the SEER database. The workflow was deployed on the IBM Blue Gene/Q Mira at the Argonne Leadership Computing Facility at Argonne National Laboratory. Results are shown in Figure 2. For each increasing core count (a power of 2, minus 2 EMEWS control cores), we increased the problem size (a product of candidate parameter combinations and number of stochastic variations) to make use of the increased processing concurrency. For each problem size, some total number N of simulation tasks were generated and completed by EMEWS, which terminated after some number of seconds T. The reported task rate is N/T. For core counts 14 - 16,382, the number of iterations within the model exploration algorithm was 4, sufficient for benchmarking. For core count 65,534, our capstone run, we ran 100 iterations, a plausible number for producing a convergent result. The capstone case ran 6,024,000 simulations in just under 22 minutes for a task rate of 4,570 simulations/second.

IV. Future Work

Based on the robust performance results, we are currently tuning the DEAP-based EA algorithm to find solution convergence in the GIOABM parameter space. We will calibrate the model across the different SEER time points to determine whether the additional network topologies and hierarchical structure introduced into the GIOABM is sufficient to generate the characteristic SEER curves. In addition, we will compare the EA approach with other ME algorithms to better understand the statistical properties of and computational requirements for exploring large-scale mechanism-based ABMs of biological systems.

Acknowledgment

This work was supported by the NIH award R01GM115839. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility. This work was completed in part with resources provided by the Research Computing Center at the University of Chicago.

REFERENCES

[1] G. An, "Dynamic knowledge representation using agent-based modeling: ontology instantiation and verification of concetual models," *Methods Mol Biol*, vol. 500, pp. 445–468, 2009.

[2] C. A. Hunt, R. C. Kennedy, S. H. Kim, and G. E. Ropella, "Agent-based modeling: a systematic assessment of use cases and requirements for enhancing pharmaceutical research and development productivity," *Wiley Interdiscip Rev Syst Biol Med*, vol. 5, pp. 461–80, Jul. 2013.

[3] M. Holcombe, S. Adra, M. Bicak, S. Chin, S. Coakley, A. I. Graham, J. Green, C. Greenough, D. Jackson, M. Kiran, S. MacNeil, A. Maleki-Dizaji, P. McMinn, M. Pogson, R. Poole, E. Qwarnstrom, F. Ratnieks, M. D. Rolfe, R. Smallwood, T. Sun, and D. Worth, "Modelling complex biological systems using an agent-based approach," *Integr. Biol.*, vol. 4, pp. 53–64, Jan. 2012.

[4] G. An, Q. Mi, J. Dutta-Moscato, A. Solovyev, and Y. Vodovotz, "Agent-based models in translational systems biology," *WIRES*, vol. 1, pp. 159–171, 2009.

[5] S. M. Sanchez, "Work Smarter, Not Harder: Guidelines for Designing Simulation Experiments," presented at the 2006 Winter Simulation Conference, 2006, pp. 47–57.

[6] R. R. Barton, "Designing Simulation Experiments," presented at the 2013 Winter Simulation Conference, 2013, pp. 342–353.

[7] M. C. Fu, "Optimization for simulation: Theory vs. practice," *Inf. J. Comput.*, vol. 14, pp. 192–215, 2002.

[8] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. New York, New York: Wiley, 2003.

[9] J. Ozik, N. T. Collier, J. M. Wozniak, and C. Spagnuolo, "From Desktop To Large-Scale Model Exploration With Swift/T.," *Winter Simul. Conf. 2016 Proc.*, In Press 2016.

[10] J. M. Wozniak, T. G. Armstrong, M. Wilde, D. S. Katz, E. Lusk, and I. T. Foster, "Swift/T: Scalable data flow programming for distributed-memory task-parallel applications," in *Proc. CCGrid 2013.*, 2013.

[11] J. Ozik, N. T. Collier, and J. M. Wozniak, "Many Resident Task Computing in Support of Dynamic Ensemble Computations," *8th Workshop Many-Task Comput. Clouds Grids Supercomput. Proc.*, 2015.

[12] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, pp. 646–674, 2011.

[13] J. Chapa, R. J. Bourgo, G. L. Greene, S. Kulkarni, and G. An, "Examining the pathogenesis of breast cancer using a novel agent-based model of mammary ductal epithelium dynamics," *PLoS One*, vol. 8, p. e64091, 2013.

[14] J. Chapa, G. An, and S. Kulkarni, "Examining the Relationship between Pre-Malignant Breast Lesions, Carcinogenesis and Tumor Evolution in the Mammary Epithelium Using an Agent-Based Model," *PLoS One*, vol. 11, no. 3, p. e0152298, Mar. 2016.

[15] G. An and S. Kulkarni, "An agent-based modeling framework linking inflammation and cancer using evolutionary principles: Description of a generative hierarchy for the hallmarks of cancer and developing a bridge between mechanism and epidemiological data," *Math. Biosci.*, vol. 260, pp. 16–24, 2015.

[16] M. E. Stack, "Investigating the Development of Ulcerative Colitis-Associated Cancer through an Agent-Based Model," *Oral Present. 11th Annu. Acad. Surg. Congr.*, Feb. 2016.

[17] R. C. Cockrell, "Supercomputing Ulcerative Colitis-associated cancer to bridge cancer mechanism with epidemiology," *Poster Present. Dig. Dis. Week*.

[18] F.-A. Fortin, F.-M. D. Rainville, M.-A. Gardner, and C. Gagne, "DEAP: Evolutionary Algorithms Made Easy," *J. Mach. Learn. Res.*, vol. 13, pp. 2171–2175, Jul. 2012.