Smart systems for model exploration with application in computational systems biology

Fredrik Wrede* and Andreas Hellander

Division of Scientific Computing, Department of Information Technology
Uppsala University, SE-75105 Uppsala, Sweden
email: {fredrik.wrede, andreas.hellander}@it.uu.se

Large computational experiments involving parameter sweep applications (PSAs), where a simulator acts much like a "black box", can be used for e.g. robustness analysis of underlying models, uncertainty quantification, computational design, and model exploration. Parameter sweeps involves scanning large regions in parameter space and with complex models involving many parameters, such sweeps can become massive. This makes it impossible for the modelers to manually analyze the results and organize the data. To deal with this problem we need to develop efficient infrastructures and frameworks which can process PSAs in a more automatic fashion. This involves intelligent ways to deploy, process and store the results from sweeps. To make sense of the vast amount of unsupervised output we need to structure the realizations into informative clusters so that the modelers can comprehend and analyze the data with higher efficiency. We propose a methodology involving machine learning technologies and associated workflows where the modelers act as a central unit by interactively directing the PSA based on patterns learned from the data. By providing the modelers with examples of realizations from clusters, they will be able to associate them with labels according to their own preferences (figure 1). This kind of active learning is a common approach within unsupervised problems[1, 5]. Further, as the modelers discover interesting patterns the PSA can be guided towards particular parameter regions, i.e going from a coarse initial design of a parameter sweep to a finer one in certain regions, according to the modelers interests. Importantly, this guidance is based on the dynamic behavior in feature space, not distance in input space. This process is referred to as active sampling within the area of surrogate modeling [2].

Our main application is exploration of quantitative, dynamic models of biochemical networks, i.e. with discrete, temporal data. Such models are typically modeled as stochastic processes governed by the chemical master equation and simulated by the Stochastic Simulation Algorithm (SSA) [4]. In practical modeling, exploration of the parameters associated with the model are frequently used in order to discover unknown properties of the underlying system and to assess the robustness of a hypothesis to parameter variations. In projects where modeling is used exploitively to generate new hypotheses of a pathway's inner workings, one often starts off with an assumed set of macromolecular interactions but little prior knowledge of kinetic parameters. Due to the high levels of uncertainty in those parameters, the effort for modelers to explore the parameter space can become overwhelming. For stochastic, highly nonlinear and high-dimensional models one often have to resort to naive global parameter sweeps. Apart from the large computational requirements, the sheer amount of data generated by such sweeps makes it hard to analyze and explore the results.

*Presenting author.

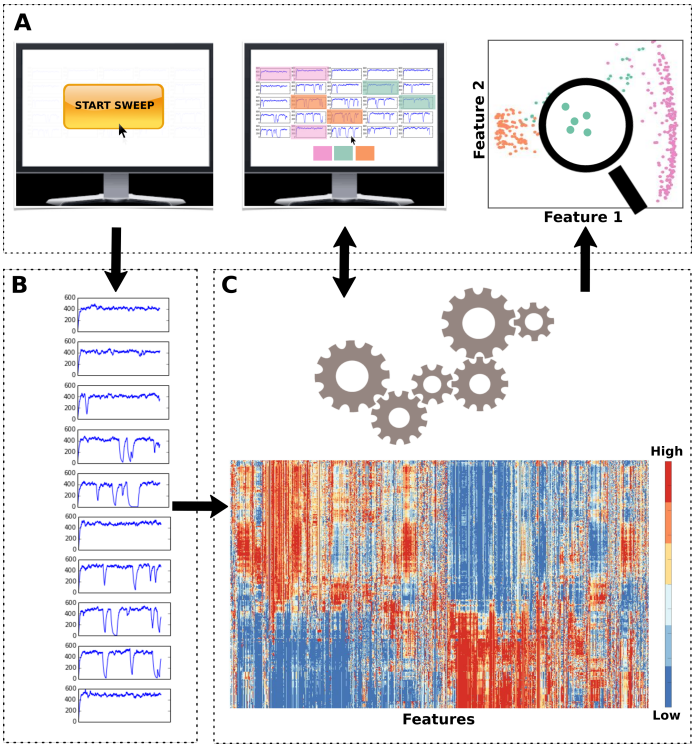

Figure 1: Workflow of the envisioned smart parameter sweep applications (PSA). In the user front-end (A), modelers can initiate a parameter sweep of their model by automatically generating time series from the corresponding simulator. As the sweep progresses, time series data (B) pass through the machine learning module (C) where features are being extracted, followed by clustering of the newly represented data. Modelers are then provided with examples from the clusters with the possibility to label the data. Information exchange occur between (A) and (B) to further learn interesting patterns, for example by finding discriminative features between different labels.

Our introductory approach for smarter parameters sweeps involves simulated temporal data passing through a feature-layer to generate a collection of features as an alternative representation of the data (figure 1 C). These features are generated from implementations associated with e.g time series analysis and distribution statistics taken from scientific literature[3]. For stochastic underlying models, the features itself will be interpreted as stochastic variables. By selecting a metric to measure distance between vectors in the feature set we are the able to group similarly behaving realizations after performing an agglomerative hierarchical clustering. Using features opposed to the raw simulated temporal data have several advantages; it allows for a more descriptive way of representing time series and it enables the presentation of discriminative sub-features between different clusters of times series.

We suggest that this approach will allow the modeler to more quickly learn about their system under study by sorting the data in parameter space according to the essential behavior of the dynamics rather than the distance between parameter points in Euclidean space. The modeler can then, according to their own preferences, label clusters that are presented. This information can then as a next step be used in a supervised algorithm, such as Support Vector Machines (SVM), to build a machine learning model which can predict group membership of newly generated data.

References

[1] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145, 1996.

[2] A. I. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1):50–79, 2009.

[3] B. D. Fulcher, M. a. Little, and N. S. Jones. Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society, Interface / the Royal Society*, 10(83):20130048, 2013.

[4] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[5] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55):11, 2010.