Simulating Large Biochemical and Biological Processes and Reasoning about their Behaviour

Extended Abstract

M. Antoniotti† F. Park∗ A. Policriti+ N. Ugel†
B. Mishra†‡

† Courant Institute of Mathematical Sciences, NYU, New York, NY, U.S.A.
∗ Seoul National University, Seoul, S. Korea
+ Università di Udine, Udine (UD), Italy
‡ Watson School of Biological Sciences, Cold Spring Harbor, NY, U.S.A.

1. INTRODUCTION

One of the trends within the emerging fields of system biology, and its sister field of bioinformatics, focuses on creating a finely detailed and "mechanistic" picture of biology at the cellular level by combining the "*part-lists*" (genes, regulatory sequences, other objects from an annotated genome, and known metabolic pathways), with observations of both transcriptional states of a cell (using micro-arrays) and translational states of the cell (using proteomics tools). It has become evident that the mathematical foundation of these systems needs to be explored accurately, while their software implementations should trade-off usability, accuracy, and scalability in order to deal with large amounts of data. We report here about a work in progress, part of a much larger project, that aims at constructing an integrated simulation and reasoning system for Biological Systems Modeling.

We assume the following scenario. Imagine a biologist seeking to test some hypotheses against a corpus of data produced by several *in vitro*, *in vivo*, and *in silico* experiments regarding the behavior of a given biological system, *e.g.*, a regulated metabolic pathway in a given organism. A (graphical) metabolic map of the biochemical system under study, together with a specific associated S-system or GMA-system [22], is assumed available, and the number of quantities recorded is large. The biologist can access one or both of the following items:

- Raw data stored somewhere about the temporal evolution of the biological system. This data may have been previously collected by *observing* an in vivo or an in vitro system, or by *simulating* the system in silico.
- Some mathematical model of the biological system[1].

The biologist will want to formulate *queries* about the evolution encoded in the data sets. For example, the biologist may ask: *will the system reach a "steady state"?*, or *will a temporary increase in the level of a certain protein repress the transcription of another?* Clearly the set of numerical *traces* of very complex systems rapidly becomes unwieldy to wade through for increasingly larger numbers of variables.

To aid the biologist in this scenario we implemented a prototype system called simpathica/xssys. Our computational tool derives its expressiveness and flexibility by integrating in a novel manner many tools from numerical analysis, symbolic computation, temporal logic, model-checking, and visualization. A distinctive feature of our approach is the "bottom-up" construction of an *automaton* that simplifies an abstracted form of *qualitative* data analysis. Based on this automaton, we developed a *temporal query language* that allows the user to query massive sets of numerical data in an efficient and natural way. The automaton provides the "*semantic scaffold*" for the temporal query language (*cf.* [12]). Such an automaton can be constructed in several ways: we proposed elsewhere [2] a simple construction based on an *approximation* of a numerical trace (*cf.* [10] and the references contained therein). We also remark that our proposed framework is relevant to the modeling of regulatory pathways; this is the subject of future work.

To motivate our approach, we show how we applied our system to a sizable example: the purine metabolism pathway as described in [22, 7, 6].

2. AN EXAMPLE: *PURINE METABOLISM*

Let us revisit in detail the example of purine metabolism described in [22, 7, 6]. The pathway for purine metabolism is presented in Figure 1.

The main metabolite in purine biosynthesis is *5-phosphoribosyl-α-1-pyrophosphate* (PRPP). A linear cascade of reactions converts PRPP into *inosine monophosphate* (IMP). IMP is the central branch point of the purine metabolism pathway. IMP is transformed into AMP and GMP. Guanosine, adenosine and their derivatives are recycled (unless used elsewhere) into *hypoxanthine* (HX) and *xanthine* (XA). XA is finally oxidized into *uric acid* (UA). In addition to these processes, there appear to be two "salvage" pathways that serve to maintain IMP level and thus of *adenosine* and *guanosine* levels as well. In these pathways, *adenine phosphoribosyltransferase* (APRT) and *hypoxanthine-guanine phosphoribosyltransferase* (HGPRT) combine with PRPP to form ribonucleotides.

The consequences of a malfunctioning purine metabolism pathway are severe and can lead to death. The entire pathway is quite complex and contains several feedback loops, cross-activations and reversible reactions, and thus an ideal candidate for reasoning with the computational tools we have developed.

[1]We note that simulating a system *in silico* actually requires a mathematical model. However, we want to consider the case when such mathematical model is unavailable to both the biologist and the software system.

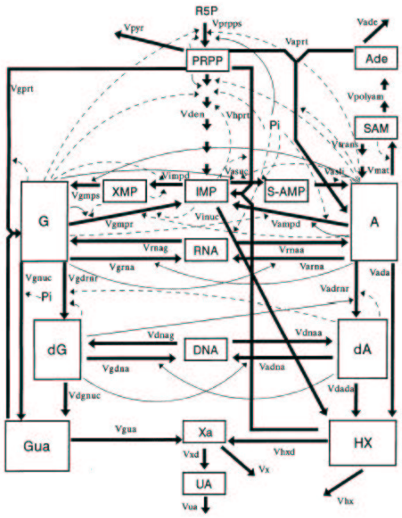

Figure 1: The metabolic scheme of purine metabolism in human. (Reprinted from [6], where a full description and further references may be found.)

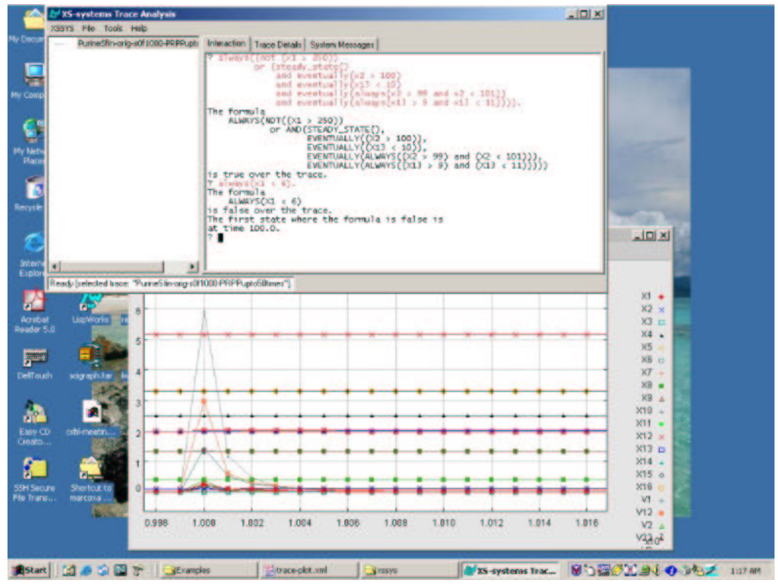

Figure 2: The in silico trace of the "final model" from [22]. We arbitrarily increased the level of PRPP (variable X1) to more than 250 at time step 100. The XSSYS system correctly answers both queries. Because of numerical effects in the floating point equality tests we had to ask a less stringent question about the steady state value of IMP (variable X2) and HX (variable X13).

We show how to formulate queries over the simulation traces of one of the mathematical models presented in [22], in order to express various desirable properties (or absence of undesirable ones) that the model should possess. Should any of these queries "fail", the model will be marked for further examination, experimentation and correction.

As an example consider the "final" model for purine metabolism presented in [22]. The in silico experiment shows that when an initial level of PRPP is increased by 50-fold, the steady state concentration is quickly absorbed by the system. The level of PRPP returns rather quickly to the expected steady state values. IMP concentration level also rises and HX level falls before returning to predicted steady state values.

Suppose that we wanted to ask the system how it will respond to a temporary (instantaneous) increase in the level of PRPP. Such request can be formulated as follows:

```
always(PRPP > 50 * PRPP1
       implies
         (steady_state()
          and eventually(IMP > IMP1)
          and eventually(HX < HX1)
          and eventually(always(IMP == IMP1))
          and eventually(always(HX == HX1))
```

an (instantaneous) increase in the level of PRPP will not make the system stray from the predicted steady state, even if temporary variations of IMP and HX are allowed. Figure 2 shows how xssys responds to the query.

3. CONCLUDING REMARKS

We have briefly presented a novel framework and software tool under which we bring together well known tools from numerical analysis, temporal logic and verification, and visualization. This is a work in progress whose initial aim is to construct an effective yet simple and testable tool to aid biologists analyze experimental results and design new ones. There are several open questions in our work that we need to address. We are now considering how to extend our automata construction by using more sophisticated *approximation theory* tools, that will allow us to take into cosideration time and frequency domain aspects of a trace. More theoretical treatments of the subject are possible as well (*cf.* [4]). Finally, we list a few of the challenges that Computational Biology needs to address in order to provide researchers with effective tools.

(1) *Hybrid Systems:* Certain interactions are purely discrete and after each such interaction, the system dynamics may change. Such a hybrid model implies that the underlying automaton must be modified for each such mode. How do these enhancements modify the basic symbolic model? (2)*Spatial Models:* The cellular interactions are highly specific to their spatial locations within the cell. How can these be modeled with richer abstractions of automata, e.g., cellular-automata? How can we account for dynamics due to changes to the cell volume? The time constants associated with the diffusion may vary from location to location; how can that be modeled? (3)*Hierarchical Models:* Finally, as we delve into more and more complex cellular processes, a clear understanding can only be obtained through modularized hierarchical models. What are the ideal hierarchical models? How do we model a population of cells with related statistics?

Acknowledgements.

The work reported in this paper was supported by grants from NSF's Qubic program, DARPA, HHMI biomedical support research grant, the US department of Energy, the US air force, National Institutes of Health and New York State Office of Science, Technology & Academic Research.

4. REFERENCES

[1] R. Alur, C. Belta, F. Ivančić, V. Kumar, M. Mintz, G. Pappas, H. Rubin, and J. Schug. Hybrid modeling and simulation of biological systems. In *Proc. of the Fourth International Workshop on Hybrid Systems: Computation and Control*, LNCS 2034, pages 19–32, Berlin, 2001. Springer-Verlag.

[2] M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra. XS-systems: extended S-Systems and Algebraic Differential Automata for Modeling Cellular Behaviour. In *Proceedings of HiPC 2002, Bangalore,* India, December 2002.

[3] U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *SCIENCE*, 283:381–387, 15 January 1999.

[4] R. W. Brockett. Dynamical systems and their associated automata. In U. Helmke, R. Mennicken, and J. Saurer, editors, *Systems and Networks: Mathematical Theory and Applications—Proceedings of the 1993 MTNS*, volume 77, pages 49–69, Berlin, 1994. Akademie-Verlag.

[5] A. Cornish-Bowden. *Fundamentals of Enzyme Kinetics*. Portland Press, London, second revised edition, 1999.

[6] R. Curto, E. O. Voit, and M. Cascante. Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunctions in man. *Biochemical Journal*, 329:477–487, 1998.

[7] R. Curto, E. O. Voit, A. Sorribas, and M. Cascante. Mathematical models of purine metabolism in man. *Mathematical Biosciences*, 151:1–49, 1998.

[8] G. Davis, S. Mallat, and M. Avellaneda. Adaptive Greedy Approximations. *Constructive Approximation*, 13:57–98, 1997.

[9] H. de-Jong, M. Page, C. Hernandez, and J. Geiselmann. Qualitative simulation of genetic regulatory networks: methods and applications. In B. Nebel, editor, *Proc. of the 17th Int. Joint Conf. on Art. Int.*, San Mateo, CA, 2001. Morgan Kaufmann.

[10] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[11] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[12] E. A. Emerson. Temporal and Modal Logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, chapter 16, pages 995–1072. MIT Press, 1990.

[13] D. Endy and R. Brent. Modeling cellular behavior. *Nature*, 409(18):391–395, January 2001.

[14] D. H. Irvine and M. A. Savageau. Efficient solution of nonlinear ordinary differential equations expressed in S-System canonical form. *SIAM Journal on Numerical Analysis*, 27(3):704–735, 1990.

[15] B. Kuipers. *Qualitative Reasoning*. MIT Press, 1994.

[16] B. Mishra. A symbolic approach to modeling cellular behavior. Submitted to *HiPC 2002*, 2002.

[17] B. Mishra and E. M. Clarke. Hierarchical Verification of Asynchronous Circuits Using Temporal Logic. *Theoretical Computer Science*, 38:269–291, 1985.

[18] M. A. Savageau. *Biochemical System Analysis: A Study of Function and Design in Molecular Biology*. Addison-Wesley, 1976.

[19] B. E. Shapiro and E. D. Mjolsness. Developmental simulation with cellerator. In *Proc. of the Second International Conference on Systems Biology (ICSB)*, Pasadena, CA, November 2001.

[20] B. Shults and B. J. Kuipers. Proving properties of continuous systmes: qualitative simulation and temporal logic. *Artificial Intelligence Journal*, 92(1-2), 1997.

[21] E. O. Voit. *Canonical Nonlinear Modeling, S-system Approach to Understanding Complexity*. Van Nostrand Reinhold, New York, 1991.

[22] E. O. Voit. *Computational Analysis of Biochemical Systems A Practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, 2000.

[23] E. O. Voit and M. Savageau. Equivalence between S-systems and Volterra systems. *Mathematical Biosciences*, 78:47–55, 1986.