Toward Scalable Whole-Cell Modeling of Human Cells

Arthur P. Goldberg
Icahn School of Medicine at Mount Sinai
1 Gustave L. Levy Pl, NY, NY 10029
Arthur.Goldberg@mssm.edu

Yin Hoon Chew
Icahn School of Medicine at Mount Sinai
1 Gustave L. Levy Pl, NY, NY 10029
YinHoon.Chew@mssm.edu

Jonathan R. Karr
Icahn School of Medicine at Mount Sinai
1 Gustave L. Levy Pl, NY, NY 10029
Karr@mssm.edu

ABSTRACT

Whole-cell (WC) models comprehensively predict cellular phenotypes by simulating the biochemistry in individual cells. WC models have the potential to enable bioengineers and physicians to rationally design microorganisms and medical therapies. WC models are developed by combining multiple mathematically distinct pathway sub-models into a single multi-algorithm model. The only existing WC model represents a small bacterium. However, to enable medical therapy, new scalable methods are needed to model human cells that contain 100 times more molecular species and 10,000–100,000 times more molecules. We describe the design of a novel system for building and simulating WC models, including an expressive sequence- and rule-based modeling language and a multi-algorithm simulator that employs optimistic parallel discrete event simulation.

Keywords

Whole-cell modeling; Modeling human cells; Systems biology; Optimistic parallel discrete event simulation; Time Warp.

1. INTRODUCTION

A central goal of biology is to understand how genotype and environment influence phenotype. However, despite decades of research, a wealth of quantitative data, and extensive knowledge, we still do not understand these causal relationships [10].

Our long-term goal is to create whole-cell (WC) computational models that accurately predict how genotype influences phenotype by representing all of the biochemical processes inside cells. WC models have the potential to accelerate biological discovery by enabling unprecedented computational experiments. These models could transform microbial bioengineering and medicine. Microbial WC models could enable bioengineers to rationally design genomes to perform practical tasks, such as efficiently produce biofuels and drugs, or sequester carbon. Human WC models could enable physicians to personalize medical therapy for individual patients. For example, an analytical oncologist would use omics analyses of a patient's tumor to construct a personalized WC model, and then use the model to identify the patient's optimal drug treatment plan (Figure 1). WC models could also help scientists identify new drug targets.

Recently, we and our collaborators created the first WC model, which analyzes the bacterium *Mycoplasma genitalium* [5]. The model is composed of 28 mathematically distinct pathway sub-models. It describes the function of every biologically understood gene and predicts the dynamics of every molecular species over the cell cycle of a single *M. genitalium*. The model was extensively validated against independent experimental data. We have used the model to discover novel biological insights, calculate the metabolic costs of synthetic circuits, and reposition antibiotics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SIGSIM-PADS '16, May 15 - 18, 2016, Banff, AB, Canada
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3742-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2901378.2901402

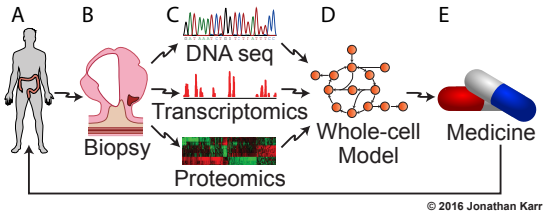

Figure 1. WC models could inform medicine. Patients (A) could be biopsied (B), tumors could be analyzed by omics techniques (C), this data could personalize WC models (D), and oncologists could use these models to design therapy (E).

However, the model does not represent several cell functions or predict certain phenotypes, and the methods used to build the model were inefficient and not systematic. The model was developed over 4 years by manually curating hundreds of databases and scientific papers and by writing 3,000 pages of MATLAB.

WC modeling must be systemized in order to achieve models of human cells that have 40 times more genes and 10^4–10^5 times more molecules than *M. genitalium*.

To address the challenges above, we are developing a systematic and scalable six-step process for WC modeling: 1) comprehensively curate experimental data about the cell being modeled, and store the data in a database; 2) design and program the model; 3) simulate the model with high accuracy and speed; 4) estimate the model's parameters; 5) verify and validate the model; and 6) analyze model predictions to gain new biological insights, design genomes, or personalize medicine. This process will be iterated to improve the model.

This paper describes our designs for steps 2 and 3: a language for describing WC models and a multi-algorithmic, Time Warp parallel discrete event simulator for simulating WC models. In other work, we are developing new methods to accelerate steps 1 and 4-6. We motivate our designs with examples of the challenges presented by modeling human cells.

2. EXISTING MODELING METHODS AND THEIR LIMITATIONS

Multiple modeling formalisms have been developed to predict the dynamics of biochemical pathways. Here we discuss some of the most common approaches and their limitations.

Ordinary differential equations (ODEs) are frequently used to model biochemical systems. This method assumes that a cell is a well-mixed container of molecules. ODEs have been used to model several well-studied signaling pathways. However, ODEs cannot represent stochastic processes and cannot be used to model

entire cells because they require more kinetic data than is available for some pathways.

The Stochastic Simulation Algorithm (SSA) is widely used to predict stochastic processes [2]. However, SSA also requires more kinetic data than is currently available for some pathways.

Flux Balance Analysis (FBA) [7] is commonly used to model cellular metabolism. Given a cell's metabolic biochemical reactions and the cell's chemical composition, FBA predicts the steady-state flux of each reaction. FBA does not require kinetic data. However, FBA cannot be used to model entire cells because it relies on assumptions which are only satisfied by metabolic pathways and it does not predict cellular dynamics.

Numerous other useful methods are also used to model various biochemical pathways in cells. These include rule-based modeling, partial differential equations, logical modeling, agent-based modeling, and Petri Nets.

2.1 Multi-algorithm WC modeling

No existing modeling formalism is suitable on its own for building fine-grained models of entire cells because the current fine-grained modeling methods require pathways to be modeled at the same level of granularity, and we do not have sufficient experimental data to finely describe every pathway.

Recently, we and others pioneered a *multi-algorithmic* approach to WC modeling [5]. This approach enables modelers to represent each pathway using the most appropriate mathematical representation driven by data availability. Separate sub-models are built for each pathway and combined into a single model. We used this approach to manually build a WC model of *M. genitalium* which is composed of 28 sub-models.

However, the WC modeling approach taken by this previous work is inadequate for efficiently developing WC models, especially of human cells. In particular, the biological properties of the *M. genitalium* model are difficult to understand because the model was described by a 3,000 page program, and the multi-algorithm simulation software was slow because it is single-threaded.

3. SCALING TO HUMAN WC MODELS

We aim to develop WC models of human cells which are orders of magnitude bigger and more complex than *M. genitalium* (Table 1). By comparison with *M. genitalium*, typical human cells contain 42 times more genes and approximately 100 times more protein types. In addition, their genomes are 6,000 times larger, and they contain qualitatively more biological compartments. Thus, human WC models will be far more complex and computationally expensive than any prior model.

Table 1. Relative sizes of *M. genitalium* and human cells.

	M. genitalium	*Homo sapiens*	Scale factor
Genes	525	21,983	42
Protein types	525	~50,000	~100
Volume	0.02 μm^3	500–5,000 μm^3	$2.5 \times 10^4 - 2.5 \times 10^5$

3.1 Computational complexity

Most human cells are 10^4–10^5 times more voluminous than *M. genitalium*. The computational cost of simulating larger cells increases linearly with cellular volume. This occurs because the SSA modeling formalism, which is used to model many well-characterized pathways and is the most computationally costly formalism, has a cost that scales linearly with the number of reactions it models. The number of reactions modeled with SSA scales linearly with the number of molecules being modeled, which grows linearly with cellular volume because molecular size varies little between organisms. Thus, we expect the computing cost of simulating WC models to grow linearly with cell volume.

Based on the 1 core-day cost of simulating the *M. genitalium* model and the 10^4–10^5 times greater size of human cells, we estimate that it will take 10^4–10^5 core-days to simulate one cell-cycle of a human cell. Assuming a pragmatic maximum acceptable execution time of 10 days, human WC model simulations must therefore be parallelized on at least 10^3–10^4 cores.

4. SYSTEMIZING WC MODELING

To enable WC models of substantially larger and more complex cells, including models of human cells, we are developing new methods and software tools to formalize and accelerate every step of the WC modeling process.

WC models are primarily composed of chemical species and biochemical reactions which transform these species. In addition, WC models represent the cell wall and compartments inside the cell. These model components should be described concisely and comprehensibly so that WC models can be easily developed, understood, reused, and modified. To enable descriptions of WC models with these properties, we are developing a WC model description language.

To quickly and accurately simulate WC models, we are developing a new parallel multi-algorithm model simulator. The simulator will be a parallel discrete event simulation application (PDES) [4]. To enable highly parallel simulations, the species and reactions which compose WC models must be partitioned into large numbers of *modules* that interact infrequently with each other to provide adequate parallelism (Figure 2). The natural spatial locality of biological processes in cells and an analysis of the *M. genitalium* model (not shown) indicate that this clustering will be feasible because most pathways only interact with a small subset of all of the modeled species.

The simulation maintains the state of species in the cell. It stores the population of each species in each compartment. It also stores the configuration details of many individual macromolecules, such as DNA, RNA and proteins. To allow parallel access to the species state, it will be partitioned into *species modules* (Figure 2). Each species module will store the state of many species.

Simulation time will synchronize interactions between reaction modules and species modules.

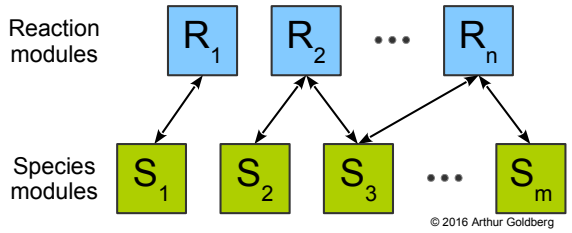

Figure 2. WC model partitioning. To enable parallel simulation, WC models will be partitioned into species and reaction modules. Each module will run in a PDES logical process. Each reaction module will interact (arrows) with a small set of species modules.

5. DESCRIBING WHOLE-CELL MODELS

To provide maximum flexibility, the domain-specific WC model description language will be implemented as a software library. The language offers two novel features. First, the language will access a database of experimental data needed to build a WC model. This database is created by the curation in step 1 of the WC modeling process summarized in Section 1. These data contain extensive information about the cell, including its genome, metabolite concentrations, RNA and protein population counts, RNA and protein half-lives, biochemical reactions, protein-protein interactions, and kinetic reaction rates. Second, the language will support concise and powerful rules for using this data to describe species and reactions in a model. Thus, the language will enable WC modelers to seamlessly integrate genomics with large-scale dynamical modeling.

The modeling language will provide several critical innovations to allow scaling to human WC models: 1) The language will support multi-algorithmic modeling by enabling the modeler to specify the modeling algorithm of sets of reaction. 2) To describe the combinatorial complexity of biological systems, the language will support *data-based modeling*, or the definition of species and reaction patterns in terms of patterns based on biochemical, genomic, and other experimental data. Data-based modeling will generalize rule-based modeling and enable WC models to explicitly combine genomics with large-scale dynamical modeling. 3) The language will implement species as typed objects. This will enable the language to efficiently handle genomic and other specialized biological data.

5.1 Specifying species

Modelers will define species by instantiating objects with experimental data from the curated database. To help modelers efficiently develop models, the language will provide an extensive set of types of biological molecules, such as proteins and nucleic acids like DNA and RNA. The language will include many of the *species types* used in typical models. Each species type will incorporate attributes to represent the structural and functional properties of that biomolecule, such as the sequences and half-lives of RNA and proteins. Modelers will also be able to create new species types or extend existing ones.

Additionally, each species type will support an associated *species pattern* that will enable convenient retrieval of instances of the species that have attributes specified in the pattern.

The species types will be implemented as a hierarchy. For example, the messenger RNA (mRNA) and ribosomal RNA (rRNA) species types will be implemented as subtypes of the RNA species type. We will use object composition to support compound types such as complexes composed of RNA and protein subunits.

5.2 Specifying reactions as rules

The language will support the description of reactions as patterns that encode biological principles which generalize across many individual reactions. Because each biological principle can encompass numerous reactions, describing reactions as patterns can avoid a combinatorial explosion of reaction descriptions that would make models infeasible. For example, proteins that bind to DNA recognize specific DNA sequences known as *motifs*. The new language will enable modelers to store observed DNA binding motifs in the attributes of protein objects and then define a single reaction pattern that represents all of the reactions in which a protein binds to a chromosomal DNA region that has an observed sequence motif.

As discussed in Section 6 below, our new WC simulator will dynamically evaluate and expand these species and reaction patterns to determine the set of active reactions.

Figure 3 illustrates the instantiation of a reaction pattern that describes the binding of proteins to DNA. The modeling language will support analogous textual descriptions of reaction patterns.

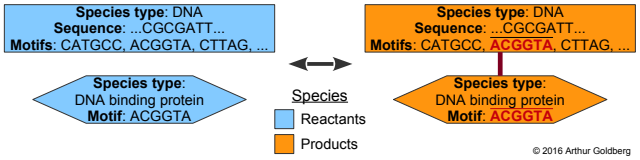

Figure 3. Instantiation of a reaction pattern. We visualize the instantiation of a specific reaction that matches a reaction pattern describing protein (lozenges) binding to DNA (rectangles). The reaction is bidirectional (arrow) with reactants on the left by convention and products on the right. The motif in the protein matches a motif in the DNA (red text in product) so they can bond (red line).

6. SIMULATING WHOLE-CELL MODELS

Most pathway simulation tools [3, 6, 8, 9] use only one modeling algorithm at a time. However, as discussed in Section 2.1, to represent all of the pathways in a cell, WC models must simultaneously employ multiple modeling algorithms. Furthermore, as discussed in Section 3, to scale up to WC models of human cells, WC models must be simulated in parallel. To achieve speedup using many cores and their memory we are developing a parallel WC simulator that will be implemented as a PDES application that runs on an optimistic PDES system, such as ROSS [1]. The simulation consists of both reaction modules and species modules. It will execute each reaction module and each species module inside a separate PDES logical process. The processes will communicate via PDES event messages (Figure 4).

Each *reaction module* will run in a PDES logical process, and use one biochemical modeling method. We will develop custom wrappers to interface continuous and static simulation algorithms such as ODEs and FBA with PDES. Multi-algorithm techniques for using continuous methods like ODEs in optimistic PDES processes are under study, but outside the scope of this paper.

Each reaction module will include the logic needed to retrieve the relevant species values from the species modules, run its modeling algorithm, and update the pertinent species values in the species modules.

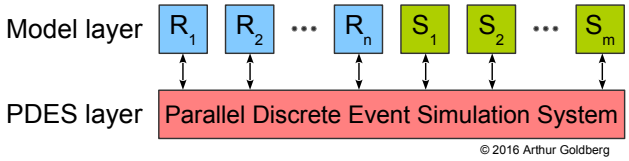

Figure 4. WC model as a PDES application. Each component of the model – reaction modules (blue) and species modules (green) – will be executed inside a PDES logical process. Reaction modules and species modules interact with each other via PDES event messages (arrows).

The simulator will support the most common biochemical modeling algorithms, including SSA, ODEs, and FBA. Wherever possible, we will reuse existing simulation libraries that support these modeling algorithms, such as libRoadrunner [9]. SSA will integrate naturally into a PDES application because SSA is a discrete event algorithm. In fact, PDES is well-suited for WC

simulation because SSA is directly compatible with PDES and SSA is one of the most useful biochemical simulation algorithms.

6.1 Partitioning reactions into modules

As discussed in Section 3, to achieve practical simulation execution times and to effectively utilize PDES, we estimate that human WC models must be partitioned into at least 10^3–10^4 modules. Given this scale, the reactions in WC models must be partitioned algorithmically.

The partitioning algorithm's objective is to create a partition that minimizes simulation run-time. However, given the complexity of WC models, the run-time cannot be directly estimated. Instead, we use an alternative computable heuristic which minimizes the interactions between reaction modules.

This approach constructs a graph in which the nodes represent reaction patterns, edges connect reactions that share at least one common species, and edge weights represent the total frequency of the connected reactions. Reaction frequencies will be estimated from the reaction rate laws and rate parameters. Clustering algorithms will be used to partition the graph into loosely-connected reaction modules.

We anticipate that this clustering will improve simulation performance by reducing PDES network messages and Time Warp rollbacks [4].

6.2 Partitioning species into modules

We will use a similar approach to partition species into modules. The approach associates each species with the reaction module that most heavily uses the species (Figure 5). We will then co-locate these pairs of reaction and species modules to minimize network traffic and Time Warp rollbacks.

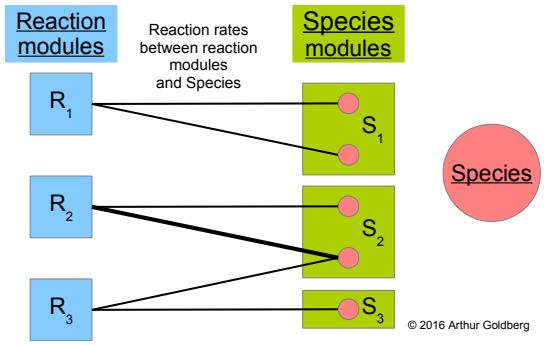

Figure 5. Heuristic for partitioning a WC model species state. First, reactions will be partitioned into modules (blue) as discussed in Section 6.1. Second, one species module (green) will be associated with each reaction module. Third, each species (salmon) is assigned to the species module associated with the reaction module that most heavily uses the species. The thickness of the black lines connecting reaction modules and species indicates the total rate at which modules use species.

7. SUMMARY

Our long-term objective is to use WC models of bacteria and human cells to advance biological science, bioengineering, and medicine. We envision using these models for a wide range of applications, including rationally designing microorganisms for a variety of industrial, agricultural and medical applications; predicting new drug targets; interpreting personal omics data; and identifying an optimal drug or drug combination for individual patients.

Significant research and engineering is needed to achieve these goals. We are beginning by developing algorithms and software for systematically building and simulating WC models. To ensure that the new tools meet all of the requirements for WC modeling and are practical, we plan to pilot the tools in conjunction with building the first human WC model. Thus, we anticipate that the tools will enable vastly more comprehensive and accurate WC models, including models of human cells.

We plan to publish the software tools open-source, along with extensive documentation, tutorials, and examples. In addition, we plan to integrate the software tools into a comprehensive, user-friendly platform for building, simulating, and analyzing WC models.

This platform will contain numerous innovations, including a WC modeling language that modelers can use to create compact and comprehensible multi-algorithm models, and an accurate, high-performance PDES whole-cell model simulator.

We anticipate that the methods and software outlined in this paper will dramatically advance WC modeling, thereby enabling routine use of whole-cell models in bioengineering and medicine.

8. ACKNOWLEDGMENTS

This work was supported by ERASynBio/NSF Grant 1548123 to JRK, by James S. McDonnell Foundation Postdoctoral Fellowship Award in Studying Complex Systems 220020377 to JRK, and by the Icahn Institute for Genomics & Multiscale Biology. We thank Amelia Goldberg for copy editing the manuscript.

9. REFERENCES

[1] Carothers, C.D. et al. 2000. ROSS: a high-performance, low memory, modular time warp system. *Proceedings Fourteenth Workshop on Parallel and Distributed Simulation*. 62, (2000), 53–60.

[2] Gillespie, D.T. 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*. 81, 25 (1977), 2340–2361.

[3] Hoops, S. et al. 2006. COPASI—a COmplex PAthway SImulator. *Bioinformatics* . 22 , 24 (Dec. 2006), 3067–3074.

[4] Jefferson, D. et al. 1987. Time warp operating system. *ACM SIGOPS Operating Systems Review*. 21, 5 (1987), 77–93.

[5] Karr, J.R. et al. 2012. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*. 150, 2 (2012), 389–401.

[6] Lopez, C.F. et al. 2013. Programming biological models in Python using PySB. *Molecular Systems Biology*. 9, 1 (2013), 646.

[7] Orth, J.D. et al. 2010. What is flux balance analysis? *Nat Biotech*. 28, 3 (Mar. 2010), 245–248.

[8] Sekar, J.A.P. and Faeder, J.R. 2012. Rule-Based Modeling of Signal Transduction: A Primer. *Computational Modeling of Signaling Networks SE - 9*. X. Liu and M.D. Betterton, eds. Humana Press. 139–218.

[9] Somogyi, E.T. et al. 2015. libRoadRunner: A High Performance SBML Simulation and Analysis Library. *Bioinformatics*. (Jun. 2015).

[10] Tomita, M. 2001. Whole-cell simulation: a grand challenge of the 21st century. *Trends in biotechnology*. 19, 6 (2001), 205–210.