Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

The future of whole-cell modeling

Derek N Macklin[1,3], Nicholas A Ruggero[2,3] and Markus W Covert[1]

Integrated whole-cell modeling is poised to make a dramatic impact on molecular and systems biology, bioengineering, and medicine — once certain obstacles are overcome. From our group's experience building a whole-cell model of *Mycoplasma genitalium*, we identified several significant challenges to building models of more complex cells. Here we review and discuss these challenges in seven areas: first, experimental interrogation; second, data curation; third, model building and integration; fourth, accelerated computation; fifth, analysis and visualization; sixth, model validation; and seventh, collaboration and community development. Surmounting these challenges will require the cooperation of an interdisciplinary group of researchers to create increasingly sophisticated whole-cell models and make data, models, and simulations more accessible to the wider community.

Addresses
[1] Department of Bioengineering, Stanford University, Stanford, CA, USA
[2] Department of Chemical Engineering, Stanford University, Stanford, CA, USA
[3] These authors contributed equally to this work.

Corresponding author: Covert, Markus W (mcovert@stanford.edu)

Current Opinion in Biotechnology 2014, **28**:111–115

This review comes from a themed issue on **Systems biology**

Edited by **Christian M Metallo** and **Victor Sourjik**

http://dx.doi.org/10.1016/j.copbio.2014.01.012

Introduction

Predictive and comprehensive models of cellular physiology are critical to understanding and engineering biological systems. Such whole-cell models have the potential to guide experiments in molecular biology, enable computer-aided design and simulation in synthetic biology, and inform personalized treatment in medicine. Constructing and validating models with sufficient scope, detail, and predictive power, for a variety of cells, will be a massive undertaking.

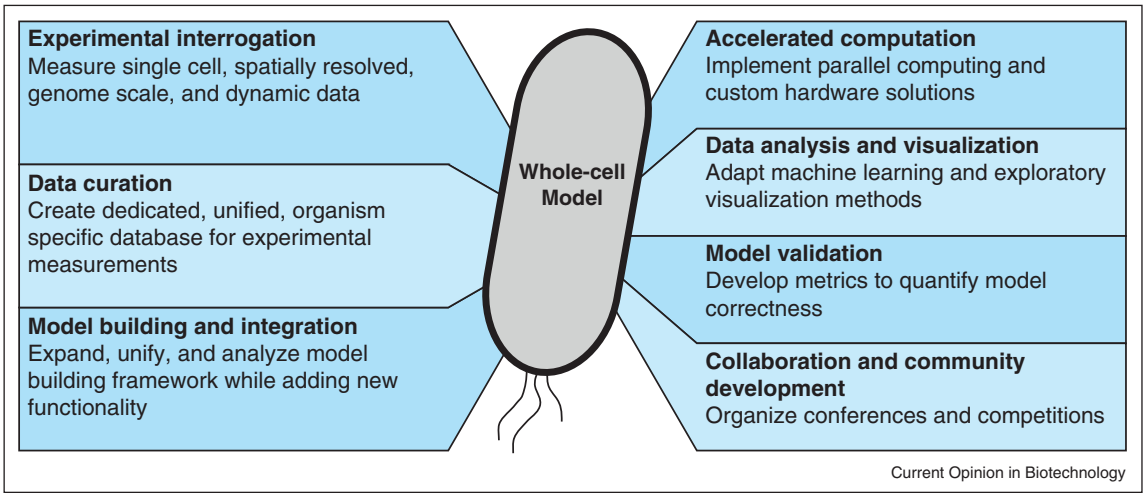
Beginning in the late 1970s [1], researchers began modeling cell physiology, primarily using ordinary differential equation (ODE) approaches, creating increasingly detailed models over the next three decades [2,3,4•]. Later, other groups introduced frameworks that generally require fewer parameters than ODE systems including constraint-based [5,6] and Boolean methods [7]. Combining these approaches for their respective benefits, our group developed a hybrid methodology: we modeled individual biological processes, each with its own mathematical representation, and merged their outputs to compute the overall state of the cell [8]. Using this approach, we simulated the life cycle of individual *Mycoplasma genitalium* cells, accounting for every molecule and representing the function of every annotated gene [9••].

Several unforeseen obstacles arose during the modeling process, which should inform any future whole-cell modeling efforts. Specifically, modeling larger cells and more complex physiology presents challenges in first, experimental interrogation; second, data curation; third, model building and integration; fourth, accelerated computation; fifth, analysis and visualization; sixth, model validation; and seventh, collaboration and community development, shown in Figure 1. No single research group can simultaneously innovate in all these areas. Rather, a broader community will need to coalesce to tackle these problems. We address this article to that community, discussing the challenges and highlighting notable progress in each area.

Experimental interrogation

Parameterizing and validating the *M. genitalium* whole-cell model was particularly challenging due to a lack of organism-specific data. Many values were estimated from measurements made in other species. Future efforts will ideally simulate well-characterized organisms, for example *Mycoplasma pneumoniae* [10–13], *Escherichia coli* [14], and *Saccharomyces cerevisiae* [15,16]. Because whole-cell models simulate the life-cycle of an individual cell, one would ideally use spatially resolved, genome-scale, dynamic, single-cell measurements to parameterize and validate the models. However, many published measurements are static ensemble averages representing a population mean at a single time point [17–21]. This lack of data ultimately presents the modeler with a dilemma: either infer missing data, or create a less detailed model of a particular phenomenon. To create the *M. genitalium* model, we necessarily inferred some degree of dynamical behavior. Faced with a similar problem, others have found ways to incorporate static spatial data in their efforts to create dynamic 3D cell-scale simulations [22•]. Promising work in advancing single-cell measurement techniques and technologies [23–26] will ultimately drive more detailed and accurate modeling. To make these efforts even more impactful and useful, the experimental community could work to establish standardized conditions and place a higher value on consistent, reproducible measurements.

Figure 1



The interdisciplinary challenges faced by future whole-cell modeling efforts. A community of scientists and engineers will need to innovate together to surmount these challenges.

Data curation

No single technology exists which can chronically measure and record the entire state of a single cell. As a result, heterogeneous data sets must be combined and unified for model parameterization and validation. While efforts such as the BioCyc databases have sought to unify genomic and metabolic pathway information [27], separate databases contain functional parameters such as kinetic rates [28,29] and expression levels [30]. To compile the data required to build the *M. genitalium* model, which we share via WholeCellKB [31], we had to download and synthesize parameters from these and other databases as well as the primary literature. For larger and more complex organisms, the sheer magnitude of data to collect, and the number of discrepancies to resolve, will present significant hurdles to parameterizing a model.

Since parameterization data increases with organism complexity and known physiology, a part-time manual curation effort will not be tenable. Researchers will need to exploit advances in natural language processing to extract information from the primary literature en masse [32], or outsource part of the effort. Formally interacting with domain experts, as has been done in the flux-balance analysis community [33], will be critical to assembling consensus data sets. Ultimately, a combination of computer-automated and human-augmented approaches will be necessary to gather and assemble the data for larger whole-cell models.

A collection of centralized, organism-specific databases similar to WholeCellKB will be required for subsequent whole-cell modeling efforts. In the best case, researchers would go beyond including raw data for each figure in a paper [34] and would deposit their results to the appropriate database in a machine-readable format. Dedicated curators would update the database schemas to incorporate new types of information as needed. In addition, the databases would alert the community to significant discrepancies between parameters and flag them as critical issues to resolve. By providing these capabilities, the databases would link experimental evidence to whole-cell models.

Model building and integration

Comprehensively representing cell physiology in a single computational model requires integrating diverse phenomena over multiple length and time scales, handling the different levels of understanding associated with each phenomenon, and representing the state of the cell in sufficient detail. Our lab's approach to meeting these requirements relies on the notion of biological modularity [35], allowing us to divide the cell into independent state variables (e.g. representing metabolite counts or the functional state of macromolecules) and cellular processes (e.g. transcription, metabolism) [9••]. We create sub-models of each cellular process using a mathematical representation informed by available data and current understanding. We assume that, over a small time step, each sub-model can independently execute and update a subset of the cell state variables. To meaningfully combine sub-models in this fashion, we must first establish and link common variables, and second, ensure that the combined behavior is consistent with physical laws and biological phenotypes.

To avoid duplicating work, it is desirable to incorporate published models of particular biological processes into a whole-cell modeling framework. This often requires that

the published models be modified to use the common whole-cell state variables, which may, for example, involve changing the published model's quantities from concentrations to counts, or linking its variables to the appropriate cell compartment in the whole-cell framework. Establishing mathematical methods for properly converting a spatially resolved variable, used in a detailed sub-model, to a bulk quantity, or even to a Boolean value, used in a less-detailed sub-model, would ease the data interconversion between sub-models. Numerical analysis of these methods could be performed to examine factors which affect stability and accuracy of the simulations, and to quantify numerical uncertainty in model predictions.

With a collection of sub-models that properly interface with cell state variables, it must further be enforced that their aggregate behavior does not violate physical laws. For example, the aggregate action of multiple sub-models should not result in the consumption of more resources than are present. To avoid this situation, we developed a method to allocate cell state variables to biological processes proportional to each process's need. In the future, this top-down approach could be replaced with one more grounded in physical laws.

Furthermore, the aggregate behavior of a collection of sub-models should be consistent with biological phenotypes. For instance, the small molecule, RNA, protein, and DNA mass fractions, must approximately double over the exponentially growing cell's life cycle. This requirement constrains certain sub-model parameters so that metabolism, for example, produces nucleotides and amino acids in the proportions needed by replication, transcription, and translation. The *M. genitalium* model performed this adjustment before simulation; however, new methods must be developed to update these loosely coupled parameters during simulation. Importantly, this will enable proper incorporation of regulatory sub-models [36,37] which modify the nucleotide and amino acid demands as the RNA and protein expression profiles change in response to perturbations.

Accelerated computation

Computational simulation is a powerful scientific and engineering tool because it enables rapid and inexpensive exploration of alternative scenarios and hypotheses, as well as design optimization. Such investigations, however, hinge on efficient computation in order to explore a sufficiently large portion of parameter space. The whole-cell simulations of *M. genitalium*, which each took approximately 10 hours to run, do not meet this criteria. We can extrapolate that, without innovation in this area, simulations of more complex organisms will take considerably longer to execute. High-performance parallelized computing technologies, such as the compute unified device architecture (CUDA) [38] or message passing interface (MPI) [39], or even custom hardware platforms [40], in the spirit of Anton [41] or Neurogrid [42], should be adapted and investigated for their abilities to speed-up the execution of whole-cell simulations.

Data analysis and visualization

Raw simulation data, like raw experimental data, typically requires extensive analysis to be adequately understood and communicated. Techniques from machine learning and dynamical systems analysis could be used to explore and interrogate simulated single-cell phenotypes. These analyses could suggest novel hypotheses about the dynamics of single cells that would not emerge from static, population-averaged data.

To complement analysis technologies, advances are needed in large-data visualization. While our group released WholeCellViz to expose a portion of the *M. genitalium* data set [43], going forward more sophisticated tools must be developed, particularly for exploration, rather than just communication, of large data sets. This requires the development of not only new visual motifs for biological data, but also improvements in data processing and retrieval to enable interactive interfaces for manipulating entire data sets. Existing tools [44] offer these interactive exploratory interfaces, but generally operate on smaller data sets [45]. Fortunately, these problems are recognized as pressing issues by the visualization community [46]. Preliminary work has begun to explore new visual motifs for biological data [47–49], and the high-performance computing community is supporting new techniques to improve data retrieval [50].

Model validation

Model predictions and experimental validation are linked by an iterative process in which each provides feedback on the other [51]. For the initial validation of the *M. genitalium* whole-cell model, we simply compared model predictions to as many heterogeneous data sets as possible that were withheld from model reconstruction. We have also used the model to predict the outcome of experiments which are performed subsequently [52•]. Nevertheless, the validation process for the *M. genitalium* model has been guided more by intuition than by a systematic methodology. Ideally, a quantitative metric would exist to specify how much of a model has been validated and would point to data sets needed to improve the coverage of validation. More subtly, methods should be developed which can differentiate novel predictions (e.g. gene essentiality in the *M. genitalium* model) from outputs arising directly from parameter fitting (e.g. biomass composition in the *M. genitalium* model). These innovations would support more widespread model adoption by building trust in the predictions.

Collaboration and community development

Whole-cell models of more complex microbes and cell types will likely become community endeavors,

particularly as the models grow in scope and detail. To facilitate interaction with the broader community, we released the entire code base for the *M. genitalium* whole-cell model under the MIT license [53], permitting open development and re-use. Going forward, we must engage the broader community in contributing to whole-cell model development. The interface between cell state variables and process sub-models must be explicitly documented in detail to lower the barrier to contribution. Furthermore, a formal plug-in system must be developed to simplify the incorporation of alternate sub-models for a particular process. At the project-management level, metrics to quantify contribution and guidelines for authorship need to be proposed and ratified. At the community level, workshops, conferences, and competitions [54] specifically focusing on whole-cell modeling need to be organized to engage the breadth of contributing researchers.

Conclusion

The need to address the aforementioned challenges provides a wealth of opportunities for interdisciplinary contribution by experimentalists, modelers, computer scientists, statisticians, bioinformaticians, and software engineers. We hope a community will form where scientists and engineers from diverse backgrounds can collaborate and innovate together to overcome these obstacles.

Whole-cell modeling can help researchers prioritize experiments by identifying knowledge gaps and by highlighting measurement discrepancies [52•]. Additionally, the comprehensive scope of a whole-cell model enables predictions of the pleiotropic effects of perturbation [55•], critical to the future of synthetic biology and personalized medicine. Addressing the issues discussed here will enable whole-cell modeling to realize its potential, and in the process make an impact on model-guided science, synthetic biology, and medicine.

Acknowledgements

We thank Elsa Birch, Ellen Casavant, Shrivats Iyer, and Jonathan Karr for their critical feedback of this manuscript, as well as members of the Covert Lab for enlightening discussions on the topic. This work was supported by an NIH Director's Pioneer Award (5DP1LM011510-05), an Allen Distinguished Investigator Award, and an award from the Stanford Bio-X Corporate Forum and Agilent to MWC, a Benchmark Stanford Graduate Fellowship and DOE CSGF Fellowship (DE-FG02-97ER25308) to DNM, and an NSF Graduate Research Fellowship to NAR.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Shuler ML, Leung S, Dick CC: **A mathematical model for the growth of a single cell**. *Ann N Y Acad Sci* 1979, **326**:35-52.

2. Domach MM, Shuler ML: **A finite representation model for an asynchronous culture of *E. coli***. *Biotechnol Bioeng* 1984, **26**:877-884.

3. Tomita M *et al.*: **E-CELL: software environment for whole-cell simulation**. *Bioinformatics* 1999, **15**:72-84.

4. • Shuler ML, Foley P, Atlas J: **Modeling a minimal cell**. *Methods Mol Biol* 2012, **881**:573-610.
An ordinary differential equation model of a 'minimal' cell with the smallest gene set required to grow and divide. Incorporates diverse aspects of cellular physiology including transcription, translation, metabolism, and replication.

5. Savinell J, Palsson BO: **Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism**. *J Theor Biol* 1992, **154**:421-454.

6. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110.**. *Appl Environ Microbiol* 1994, **60**:3724.

7. Davidson EH *et al.*: **A genomic regulatory network for development**. *Science* 2002, **295**:1669-1678.

8. Covert MW *et al.*: **Integrating metabolic transcriptional regulatory and signal transduction models in *Escherichia coli***. *Bioinformatics* 2008, **24**:2044-2050.

9. •• Karr JR *et al.*: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**:389-401.
The first whole-cell computational model. Simulates the life cycle of the human pathogen *Mycoplasma genitalium* including all functionally annotated gene products and their interactions.

10. Güell M *et al.*: **Transcriptome complexity in a genome-reduced bacterium**. *Science* 2009, **326**:1268-1271.

11. Kühner S *et al.*: **Proteome organization in a genome-reduced bacterium**. *Science* 2009, **326**:1235-1240.

12. Yus E *et al.*: **Impact of genome reduction on bacterial metabolism and its regulation**. *Science* 2009, **326**:1263-1268.

13. Maier T *et al.*: **Quantification of mRNA and protein and integration with protein turnover in a bacterium**. *Mol Syst Biol* 2011, **7**.

14. Ishii N *et al.*: **Multiple high-throughput analyses monitor the response of *E. coli* to perturbations.**. *Science* 2007, **316**:593-597.

15. Picotti P *et al.*: **A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis**. *Nature* 2013, **494**:266-270.

16. Miller C *et al.*: **Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast**. *Mol Syst Biol* 2011, **7**.

17. Hoheisel JD: **Microarray technology: beyond transcript profiling and genotype analysis**. *Nat Rev Genet* 2006, **7**:200-210.

18. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.

19. Mahmood T, Yang P-C: **Western blot: technique theory and trouble shooting**. *N Am J Med Sci* 2012, **4**:429-434.

20. Gallagher SR: **Current protocols in molecular biology**. In *Current Protocols in Molecular Biology*. Edited by Ausubel FM. 2006. (Chapter 10, Unit 10.2A).

21. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions**. *Nat Rev Genet* 2012, **13**:840-852.

22. • Roberts E *et al.*: **Noise contributions in an inducible genetic switch: a whole-cell simulation study**. *PLoS Comput Biol* 2011, **7**:e1002010.
Stochastic, single-cell, spatially-resolved kinetic model which simulates the switching of the lac operon in *Escherichia coli*. Critical steps taken in modeling more complex bacterial cell physiology spatially.

23. Taniguchi Y *et al.*: **Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells**. *Science* 2010, **329**:533-538.

24. Lee TK *et al.*: **A noisy paracrine signal determines the cellular NF-kappaB response to lipopolysaccharide**. *Sci Signal* 2009, **2**:ra65.

25. Tang Fuchou *et al.*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**:377-382.

26. Ibá nez AJ *et al.*: **Mass spectrometry-based metabolomics of single yeast cells**. *Proc Natl Acad Sci* 2013, **110**:8790-8794.

27. Caspi R *et al.*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2010, **38(Database issue)**:D473-D479.

28. Schomburg I *et al.*: **BRENDA in 2013: integrated reactions kinetic data enzyme function data improved disease classification: new options and contents in BRENDA**. *Nucleic Acids Res* 2013, **41(Database issue)**:D764-D772.

29. Wittig U *et al.*: **SABIO-RK-database for biochemical reaction kinetics**. *Nucleic Acids Res* 2011, **40**:D790-D796.

30. Barrett T *et al.*: **NCBI GEO: archive for functional genomics data sets-update**. *Nucleic acids research* 2012, **41**:D991-D995.

31. Karr JR *et al.*: **WholeCellKB: model organism databases for comprehensive whole-cell models**. *Nucleic Acids Res* 2013, **41**:D787-D792.

32. Finkel J *et al.*: **Exploring the boundaries: gene and protein identification in biomedical text**. *BMC Bioinformatics* 2005, **6(Suppl. 1)**:S5.

33. Thiele I, Palsson BO: **Reconstruction annotation jamborees: a community approach to systems biology**. *Mol Syst Biol* 2010, **6**:361.

34. *Announcement. Reducing our irreproducibility*. **Nature**2013, **496**:398.

35. Hartwell LH *et al.*: *From molecular to modular cell biology*.**Nature**1999, **402**:C47-C52.

36. Bonneau R *et al.*: *The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo*.**Genome Biol**2006, **7**:R36.

37. Carrera J, Rodrigo G, Jaramillo A: *Model-based redesign of global transcription regulation*.**Nucleic Acids Res**2009, **37**:e38.

38. CUDA Toolkit Documentation. http://docs.nvidia.com/cuda/index.html.

39. MPI Documents. http://www.mpi-forum.org/docs/.

40. Gunawardena J: *Silicon dreams of cells into symbols*.**Nat Biotechnol**2012, **30**:838-840.

41. Dror RO *et al.*: *Biomolecular simulation: a computational microscope for molecular biology*.**Annu Rev Biophys**2012, **41**:429-452.

42. Silver R *et al.*: *Neurotech for neuroscience: unifying concepts organizing principles and emerging tools*.**J Neurosci**2007, **27**:11807-11819.

43. Lee R, Karr JR, Covert MW: *WholeCellViz: data visualization for whole-cell models*.**BMC Bioinformatics**2013, **14**:253.

44. Business Intelligence and Analytics Software. http://www.tableausoftware.com/.

45. Tableau Technology — Tableau Software. http://www.tableausoftware.com/products/technology.

46. Wong PC *et al.*: *The top 10 challenges in extreme-scale visual analytics*.**IEEE Comput Graph Appl**2012, **32**:63-67.

47. Meyer M *et al.*: *MulteeSum: a tool for comparative spatial and temporal gene expression data.*.**IEEE Trans Visual Comput Graph**2010, **16**:908-917.

48. Meyer M *et al.*: *Pathline: a tool for comparative functional genomics*.**Comput Graph Forum**2010, **29**:1043-1052.

49. Meyer M, Munzner T, Pfister H: *MizBee: a multiscale synteny browser.*.**IEEE Trans Visual Comput Graph**2009, **15**:897-904.

50. Ashby S *et al.*: *The Opportunities and Challenges of Exascale Computing — Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee*. US Department of Energy Office of Science; 2010.

51. Kitano H: *Computational systems biology*.**Am J Gastroenterol**2002, **420**:206-210.

52. • Sanghvi JC *et al.*: *Accelerated discovery via a whole-cell model*.**Nat Methods**2013, **10**:1192-1195.

The *Mycoplasma genitalium* whole-cell model is demonstrated to accelerate biological discovery by guiding experiments. Turnover rates for three metabolic enzymes are identified as incorrect and subsequently correctly predicted or bounded.

53. The MIT License (MIT) — Open Source Initiative. http://opensource.org/licenses/MIT.

54. Whole-cell parameter estimation DREAM challenge — syn1876068. https://www.synapse.org/#!Synapse:syn1876068.

55. • Purcell O *et al.*: *Towards a whole-cell modeling approach for synthetic biology*.**Chaos**2013, **23**:025112.

Modifies the existing whole-cell model to allow for genome modification. Engineered a synthetic gene circuit for a Goodwin oscillator *in silico* and examined how codon usage correlates with synthetic gene expression.