Whole-Cell model simulations for medicine and bioengineering

A Response to the National Institute of General Medical Sciences RFI on Science Drivers Requiring Capable Exascale High Performance Computing

Professor Arthur Goldberg (arthur.goldberg@mssm.edu), Professor Jonathan Karr (karr@mssm.edu)
Icahn Institute for Genomics & Multiscale Biology, Icahn School of Medicine at Mount Sinai
1255 5th Avenue, New York, New York 10029

6 October 2015

Whole-cell modeling

Whole-cell models predict cell behaviors by modeling all molecular components and their interactions [Karr 2015, Macklin 2014, Covert 2013, Tomita 2001]. Recently, we and other developed the first whole-cell model [Karr 2012]. The model represents the functionality of all 409 characterized genes and 725 metabolites throughout one life cycle of the reduced bacterium *Mycoplasma genitalium*. This model was validated against a broad range of data and provided insights into many previously unobserved cellular behaviors.

Simulating the behavior of a single cell required modest computing resources – 1 core-day of an Intel E5520 CPU, capable of 3.3×10^{15} double-precision floating-point operations during the computation. Sampling the organism's behavior required 128 simulations. Nevertheless, we anticipate that exascale computing resources will be required to use more comprehensive and more accurate whole-cell models to personalize medicine and engineer bacteria.

Potential impacts of whole-cell modeling

Whole-cell modeling has the potential to make large impacts on both precision medicine and bioengineering:

- *Medical applications of models of human cells*: Human models could revolutionize medicine. For example, in the future, we envision that computational oncologists will use accurate, personalized whole-cell models of tumors, parameterized by each tumor's genetic variations, to find the optimal combination and dosage of drugs to treat each patient's cancer. Similar approaches could be used to personalize therapy for any patient with any disease who is being evaluated for any drug.
- *Industrial applications of genetically optimized bacteria*: Many economically transformative genetic engineering applications of bacteria are currently under investigation [Khalil 2010], including drug production [Ajikumar 2010], renewable fuel synthesis, generating energy from sunlight, and hazardous waste disposal [Lee 2012]. However, these efforts are hindered by challenges in predicting and testing the benefits of possible genetic modifications. This process could be improved by (1) using whole-cell models to rationally design genomes by optimizing *in silico* phenotypes and (2) using genome editing methods such as CRISPR [Jinek 2012] or genome synthesis methods [Gibson 2008] to implement designer genomes.

Computational landscape of whole-cell modeling in 2025

We estimate the computational costs of two representative studies that we aim to support in 2025 (Table 1):

1. A clinical trial of 100 cancer patients with drug regimens determined by personalized whole-cell models.
2. A study to engineer a bacterium to cost-effectively produce a complex drug.

As a first example, we consider the cost of using whole-cell models to select the optimal drugs for 100 patients. Compared to our *M. genitalium* model, we anticipate that human models will represent 400 times more proteins and reactions. To use whole-cell models to screen drugs, we anticipate simulating 100 drug combinations and dosages per patient, and simulating each drug combination 1,000 times to accurately model its behavior. Based on the $\approx3\times10^{15}$ potential floating point operation cost of our *M. genitalium* model simulations, we anticipate that this study will require resources providing $\approx10^{25}$ floating point operations, or about 0.3 sustained EXAFLOPS for roughly a year.

We anticipate that bacterial engineering studies will focus on more complex bacteria, such as *Escherichia coli,* that have 10 times more genes than *M. genitalium*. To predictably design bacteria, we also anticipate that whole-cell models will need to represent molecular processes with 10 times greater detail. AA challenging aspect of designing bacterial genomes will be exploring the extremely high dimensional space of possible variants. Thus, we anticipate that 10^6 model runs, each replicated 100 times to accurately sample their behavior, will be required to explore and optimize bacterial genomes. Based on the $\approx3\times10^{15}$ potential floating point operation cost of our *M. genitalium* model*,* we anticipate that whole-cell model-driven bioengineering will require resources with $\approx10^{25}$ floating point operations, also roughly 0.3 sustained EXAFLOPS for a few years.

In summary, we anticipate that future applications that employ whole-cell modeling will require significant computational resources, and that this will primarily be driven by the large numbers of simulations required to optimize the large spaces of possible drug combinations and genome sequences.

Table 1. Anticipated computational costs of future whole-cell modeling.

	Application	
Computational component	**Clinical trial of optimization of cancer drug treatment**	**Design a genetically optimized bacterium to solve a bioengineering problem**
Computational cost function	(current cost of simulating one bacterium life-cycle) x (factor increase in types of proteins) x (possible drug combinations and dosages) x (statistical replications) x (100 patients)	(current cost of simulating one bacterium life-cycle) x (factor increase in types of proteins) x (increased detail of bacterial model) x (possible DNA modifications) x (statistical replications)
Computational cost values	(3×10^{15}) x (4×10^{2}) x (1×10^{3}) x (1×10^{3}) x (1×10^{2})	(3×10^{15}) x (1×10^{1}) x (1×10^{1}) x (1×10^{6}) x (1×10^{2})
Computational cost (floating-point operations)	1.2×10^{25}	3×10^{25}
Duration of study (days)	400	1000
Sustained compute for study (EXAFLOPS)	0.3	0.3

Whole-cell modeling roadmap

The whole-cell modeling field is developing technologies to enable these applications, including:

1. Tools to collect and organize the experimental data needed to train whole-cell models;
2. Tools to design whole-cell models, including methods to partition reactions into sub-models;
3. Tools to identify the parameters of high-dimensional models, including tools that utilize model reduction and distributed optimization
4. Parallel algorithms for simulating multi-algorithm models with high numerical accuracy and performance;
5. Tools to store, visualize, and analyze of high-dimensional simulation results; and
6. Standard formats to represent whole-cell models and their simulations.

We have developed software to simulate multi-algorithm models [Karr 2012], WholeCellKB [Karr 2013] to organize experimental training data, WholeCellSimDB [Karr 2014] to store simulation results, and WholeCellViz [Lee 2013] to visualize simulations results. Other groups created the SBML and SED-ML standard representation formats, which represent systems models and their simulations, respectively.

We anticipate that the field will realize these tools in the next five years, and begin to pursue the described medical and bioengineering applications in 5-10 years.

Summary

In summary, we anticipate that applications which employ whole-cell modeling will require increasing computational resources as the field advances the technologies for building and simulating whole-cell models.

References

Ajikumar PK, Xiao W-H, Tyo KEJ, et al. Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli. Sci . 2010;330 (6000):70-74. doi:10.1126/science.1191652

Covert MW. Simulating a living cell. *Sci Am*. 2014;310(1):44–51. doi:10.1038/scientificamerican0114-44

Gibson DG, Benders GA, Andrews-Pfannkoch C et al. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*. 2008;319(5867):1215-20. doi: 10.1126/science.1151721.

Jinek M, Chylinski K, Fonfara I et al. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*. 2012;337 (6096):816–821. doi:10.1126/science.1225829.Karr JR, Phillips NC & Covert MW. WholeCellSimDB: a hybrid relational/HDF database for whole-cell model predictions. *Database*. 2014;2014(0):bau095. doi:10.1093/database/bau095.

Karr JR, Sanghvi JC, Macklin DN et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*. 2012;150(2):389–401. doi:10.1016/j.cell.2012.05.044.

Karr JR, Sanghvi JC, Macklin DN et al. WholeCellKB: Model organism databases for comprehensive whole-cell models. *Nucleic Acids Res*. 2013;41(D1):D787–D792. doi:10.1093/nar/gks1108.

Karr JR, Takahasi K & Funahashi A. The principles of whole-cell modeling. *Curr Opin Microbiol.* 2015;27:18–24.

Khalil AS, Collins JJ. Synthetic biology: applications come of age. Nat Rev Genet. 2010;11(5):367-379.

Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat Chem Biol. 2012;8(6):536-546. http://dx.doi.org/10.1038/nchembio.970.

Lee R, Karr JR & Covert MW. WholeCellViz: data visualization for whole-cell models. *BMC Bioinformatics*. 2013;14(1):253.

Macklin DN, Ruggero NA & Covert MW. The future of whole-cell modeling. *Curr Opin Biotechnol*. 2014;28:111-5.

Tomita M. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol*. 2001;19(6):205–10. doi:10.1016/S0167-7799(01)01636-5.