Emerging whole-cell modeling principles and methods

Arthur P. Goldberg[1,2,3], Balázs Szigeti[1,2,3], Yin Hoon Chew[1,2,3], John A. P. Sekar[1,2,3], Yosef D. Roth[1,2], Jonathan R. Karr[1,2,*]

[1]*Icahn Institute for Genomics and Multiscale Biology*
[2]*Department of Genetics and Genomic Sciences*
Icahn School of Medicine at Mount Sinai, New York, New York, 10029, USA
[3]*These authors contributed equally to the work.*

Abstract

Whole-cell computational models aim to predict cellular phenotypes from genotype by representing the entire genome, the structure and concentration of each molecular species, each molecular interaction, and the extracellular environment. Whole-cell models have great potential to transform bioscience, bioengineering, and medicine. However, numerous challenges remain to achieve whole-cell models. Nevertheless, researchers are beginning to leverage recent progress in measurement technology, bioinformatics, data sharing, rule-based modeling, and multi-algorithmic simulation to build the first whole-cell models. We anticipate that ongoing efforts to develop scalable whole-cell modeling tools will enable dramatically more comprehensive and more accurate whole-cell models, including models of human cells.

Introduction

Whole-cell (WC) computational models aim to predict cellular phenotypes from genotype and the environment by representing the function of each gene, gene product, and metabolite [1]. WC models could unify our understanding of cell biology and enable researchers to perform *in silico* experiments with complete control, scope, and resolution [2, 3]. WC models could also help bioengineers rationally design microorganisms that can produce useful chemicals and act as biosensors, and help physicians design personalized therapies tailored to each patient's genome.

Despite their potential, there is little consensus on how WC models should represent cells, what phenotypes WC models should predict, or how to achieve WC models. Nevertheless, we and others are beginning to leverage advances in measurement technology, bioinformatics, rule-based modeling, and multi-algorithmic simulation to develop WC models [4–8]. However, substantial work remains to achieve WC models [9].

To build consensus on WC modeling, we propose a set of key physical and chemical mechanisms that a WC model should aim to represent, and a set of key phenotypes that a WC model should aim to predict. We also summarize the experimental and computational advances that are making WC models feasible, and outline several technologies that would help accelerate WC modeling.

Note, our proposals focus on defining WC models that are needed for research studies and applications such as bioengineering and personalized medicine which depend on understanding the

---

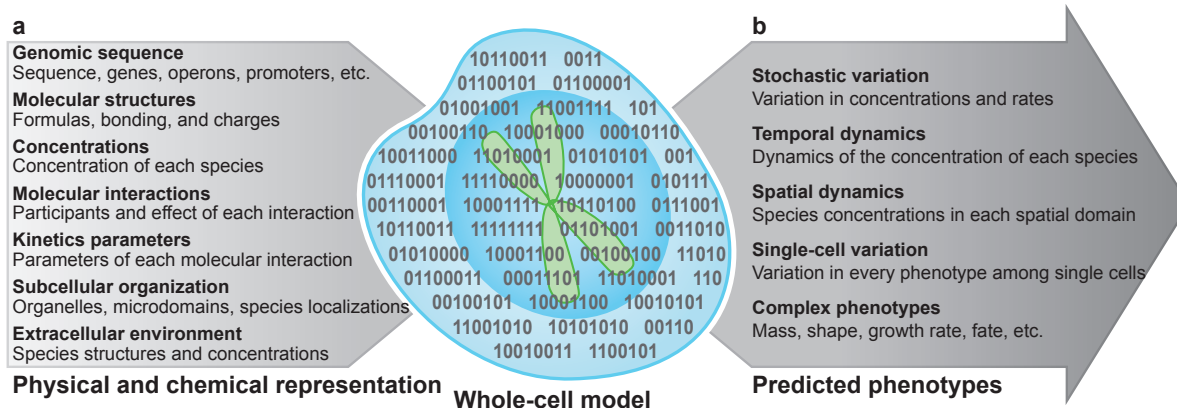*Corresponding author: Karr, Jonathan R (karr@mssm.edu)

Figure 1. The physical and chemical mechanisms that WC models should aim to represent (a) and the phenotypes that WC models should aim to predict (b).

molecular details of the majority of intracellular processes. Many research studies would be better served by smaller, more focused models.

Physics and chemistry that WC models should aim to represent

We propose that WC models aim to represent all of the chemical reactions in a cell and all of the physical processes that influence their rates (Figure 1a). This requires representing (a) the **sequence** of each chromosome, RNA, and protein; the location of each feature of each chromosome such as genes, operons, promoters, and terminators; and the location of each site of each RNA and protein; (b) the **structure** of each molecule, including atom-level information about small molecules, the domains and sites of macromolecules, and the subunit composition of complexes; (c) the **subcellular organization** of cells into organelles and microdomains; (d) the participants and effect of each **molecular interaction**, including the molecules that are consumed, produced, and transported, the molecular sites that are modified, and the bonds that are broken and formed, (e) the **kinetic parameters** of each interaction; (f) the **concentration** of each species in each organelle and microdomain; and (g) the concentration of each species in the **extracellular environment**. Together, these would enable WC models to describe how cellular behavior emerges from the combined function of each gene and genetic variant, and capture how cells respond to changes in their internal and external environments.

Phenotypes that WC models should aim to predict

We also propose that WC simulations aim to predict the trajectory of a single cell over its entire lifecycle (Figure 1b) including (a) the **stochastic dynamics** of each molecular interaction; (b) the **temporal dynamics** of the concentration of each species; (c) the **spatial dynamics** of the concentration of each species in each organelle and microdomain; (d) the **single-cell variation** of cellular behavior; and (e) **complex phenotypes** such as the cell shape, growth rate, and fate. Together, these would enable WC models to capture how stochastic and single-cell variation can generate phenotypic diversity; how cells respond to external cues such as nutrients, growth factors and drugs; and how cells coordinate critical events such as the G1/S transition.

Available resources

Achieving WC models will require extensive data to constrain every parameter. Fortunately, measurement technology is rapidly advancing. Here, we review the latest methods for generating data for WC models, and highlight repositories and other resources that contain useful data for WC modeling.

Measurement methods

Advances in single-cell and genomic measurement are rapidly generating data that is useful for WC models [10–12] (Table S1). For example, Meth-Seq can assess epigenetic modifications [13], Hi-C can determine chromosome structures [14], ChIP-seq can determine protein-DNA interactions [15], fluorescence microscopy can determine protein localizations, mass-spectrometry can quantitate metabolite and protein concentrations, FISH [16] and scRNA-seq [17] can quantitate the dynamics and single-cell variation of RNA abundances, and fluorescence microscopy and mass cytometry [18] can quantitate the dynamics and single-cell variation of protein abundances.

Data repositories

Researchers are also rapidly aggregating much of the data needed for WC modeling into public repositories (Table S2). For example, UniProt contains a multitude of information about proteins [19]; BioCyc contain extensive information about interactions [20]; ECMDB [21], ArrayExpress [22], and PaxDb [23] contain metabolite, RNA, and protein abundances; and SABIO-RK contains kinetic parameters [24]. Furthermore, meta-databases such as *Nucleic Acid Research*'s Database Summary contain lists of repositories.

Prediction tools

For certain types of data, accurate prediction tools can be a useful alternative to direct experimental evidence which may have incomplete coverage or may be limited to a small number of genotypes and environments. Currently, many tools can predict properties such as operons, RNA folds, and protein localizations (Table S3). For example, PSORTb predicts the localization of bacterial proteins [25]. However, many current prediction tools are not sufficiently accurate for WC modeling.

Published models

WC models can also be built by combining separately published models of individual pathways. Currently, there are several model repositories which contain numerous cell cycle, circadian rhythm, electrical signaling, signal transduction, and metabolism models (Table S4–S5). However, most pathways such as RNA degradation do not yet have genome-scale dynamical models; many reported models are not publicly available; and it is difficult to merge most published models because they often use different assumptions and representations.

Emerging methods and tools

Recent advances in data aggregation, model design, model representation, and simulation (Table S6) are also rapidly making WC models feasible. We expect that ongoing efforts to adapt and combine these advances will accelerate WC modeling [26] (Figure 2). Here, we summarize the most important emerging methods and tools for WC modeling.

Data aggregation and organization

To make the large amount of publicly available data usable for modeling, researchers are developing automated methods for extracting data from publications [27], building central public repositories [28], and creating tools for programmatically accessing repositories [29]. Pathway/genome database (PGDB) tools such as Pathway Tools [30] are well-suited to organizing this data because they
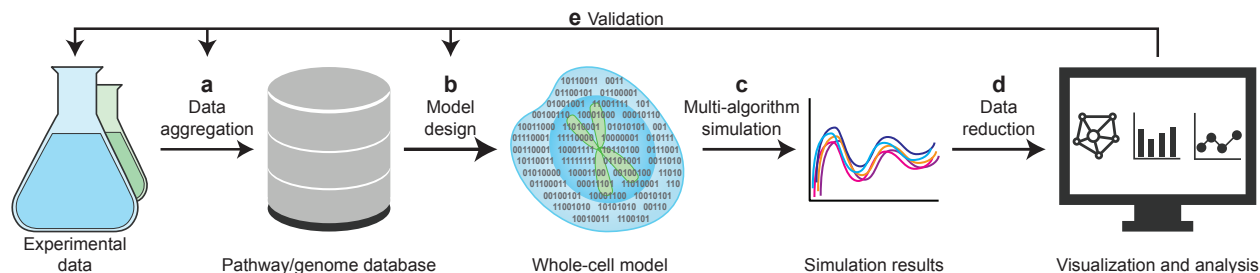
Figure 2. Emerging WC modeling methodology. (a) Data should be aggregated from thousands of publications, repositories, and prediction tools and organized into a PGDB. (b) Models should be designed, calibrated, and validated from PGDBs and described using rules. (c) Models should be simulated using parallel, network-free, multi-algorithmic simulators and their results should be stored in a database. (d) Simulation results should be visualized and analyzed. (e) Results should be validated by comparison to experimental measurements. Importantly, all of these steps should be collaborative.

support structured representations of metabolites, DNA, RNA, proteins, and their interactions. However, they provided limited support for non-metabolic pathways and quantitative data. To overcome these limitations, we developed the WholeCellKB tool to organize data for WC modeling [31].

Scalable model design
Several new tools can help researchers develop large models. For example, the Cell Collective facilitates collaborative model design [32], MetaFlux facilitates the design of constraint-based models from PGDBs [33], PySB facilitates programmatic model construction [34], SEEK facilitates model design from data tables [35], and Virtual Cell facilitates model design from KEGG and SABIO-RK [36].

Model languages
Researchers have developed several languages for representing biochemical models. SBML can represent several types of models including flux balance analysis models, deterministic dynamical models, and stochastic dynamical models [37]. Rule-based languages such as BioNetGen can efficiently describe the combinatorial complexity of protein-protein interactions [38].

Simulation
Numerous tools can simulate biomodels. For example, COPASI [39] and Virtual Cell [36] support deterministic, stochastic, hybrid deterministic/stochastic, network-free, and spatial simulation; CO-BRApy supports constraint-based simulation [40]; and E-Cell supports multi-algorithmic simulation [41].

Calibration
New tools such as saCeSS [42] support distributed calibration of large biochemical models. In addition, aerospace and mechanical engineers have developed methods for using reduced models to efficiently calibrate large models [43].

Verification
Researchers have begun to adapt formal model checking techniques to biomodeling. For example, BioLab [44] and PRISM [45] can verify BioNetGen-encoded and SBML-encoded models, respectively.

Simulation results analysis
Tools such as COPASI [39] and Virtual Cell [36] can visualize simulation results. We have developed the WholeCellSimDB [46] simulation results database to help researchers organize, search, and share WC simulation results. We have also developed the WholeCellViz [47] simulation results dashboard to help researchers visualize WC simulation results in their biological context.

Technological challenges
Beyond these emerging tools, several technological advances are needed to enable WC models. Here, we summarize the most critically needed technologies.

Experimental measurement
While substantial data about cellular populations already exists, additional data would enable better WC models. In particular, we need metabolome-wide and proteome-wide measurement technologies that can quantitate the dynamics and single-cell variation of each metabolite and protein. Additionally, we need technologies that can measure kinetic parameters at the interactome scale and technologies that can measure cellular phenotypes across multiple genetic and environmental conditions.

Prediction tools
While existing tools can predict many properties of metabolites, DNA, RNA, and proteins, additional tools are needed to accurately predict the molecular effects of insertions, deletions, and structural variants. Such tools would help WC models design microbial genomes and predict the phenotypes of individual patients.

Data aggregation
As described above, extensive data is now available for WC modeling. However, this data spans a wide range of data types, organisms, and environments; the data is often not annotated and normalized; it is scattered across many repositories and publications and it is described using inconsistent identifiers and units. To make this data more usable for modeling, we are developing a framework for aggregating data from repositories; merging data from multiple specimens, environmental conditions, and experimental procedures; standardizing data to common units; and identifying the most relevant data for a model.

Scalable, data-driven model design
To scale WC modeling, we need tools for collaboratively building large models directly from experimental data, recording how data is used to build models, and identifying gaps and inconsistencies in models. As described above, several tools support each of these functions. To accelerate WC modeling, the field must develop a single tool that supports all of these functions at the scale required for WC modeling.

Rule-based model representation
Several languages can represent individual biological processes, but no existing language supports all of the biological processes that WC models must represent [48, 49]. To overcome this, we are developing a rule-based language that can represent each molecular species at multiple levels of granularity (for example, as a single species, as a set of sites, and as a sequence); the combinatorial complexity of each molecular species and interaction; composite, multi-algorithmic models; and the data used to build models.

Scalable multi-algorithmic simulation
Simulating WC models requires a simulator that supports both network-free interpretation of rule-based model descriptions and multi-algorithmic co-simulation of submodels that are described using different simulation algorithms. However, no existing simulator supports both network-free and multi-algorithmic simulation. To scalably simulate WC models, we are using parallel discrete event simulation and graph transformation to develop a parallel, network-free, multi-algorithmic simulator [26].

Calibration and verification
Scalable tools are needed to calibrate and verify WC models. Although we and others have begun to explore surrogate strategies for efficiently calibrating and validating WC models [50], further work is needed to formalize these methods.

Simulation analysis
We and others have developed tools for organizing and visualizing simulation results, but they provided limited support for large datasets or custom visualizations such as pathway maps. To visualize WC simulation results, researchers should use distributed database and data processing technologies to search and reduce simulation results, standard visualization grammars to enable flexible and custom visualizations, and high-performance visualization toolkits to handle terabyte-scale simulation results.

Collaboration
Ultimately, achieving WC models will require extensive teamwork. To facilitate collaboration, the field must develop collaborative model design tools, version control systems for models, standards for annotating and verifying submodels, and protocols for merging separately developed submodels.

Conclusion
WC models have great potential to advance bioscience, bioengineering, and medicine. However, significant challenges remain to achieve WC models. To advance WC modeling, we have proposed how WC models should represent cells and the phenotypes that WC models should predict, and summarized the best emerging methods and resources. We have also outlined several technological solutions to the most immediate WC modeling challenges. Specifically, we must develop new tools for scalably and collaboratively designing, simulating, calibrating, validating and analyzing models. We must also develop new methods for measuring the dynamics and single-cell variation of the metabolome and proteome and for measuring kinetic parameters at the interactome scale. Despite these challenges, we and others are building the first WC models, developing the first WC modeling tools, and beginning to form a WC modeling community [48, 50]. We anticipate these efforts will enable comprehensive predictive models of cells.

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

   •• Describes the principles of WC modeling including mechanistic representation of each gene, molecular species, and molecular interaction over the lifecycle of an organism.

2. Tomita M: **Whole-cell simulation: a grand challenge of the 21st century.** *Trends Biotechnol* 2001, **19**:205–210.

   • Describes the need to develop WC models to understand biology and personalize medicine.

3. Carrera J, Covert MW: **Why build whole-cell models?** *Trends Cell Biol* 2015, **25**:719–722.

   •• Describes several potential uses of WC models including data integration, knowledge evaluation, phenotype prediction, hypothesis generation, and genome design.

4. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC *et al.*: **E-CELL: software environment for whole-cell simulation.** *Bioinformatics* 1999, **15**:72–84.

   • Describes one of the earliest efforts to build a model that represents the function of each gene.

5. Atlas J, Nikolaev E, Browning S, Shuler M: **Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of Escherichia coli: application to DNA replication.** *IET systems biology* 2008, **2**:369–382.

   • Describes an early effort to build WC models.

6. Roberts E, Magis A, Ortiz JO, Baumeister W, Luthey-Schulten Z: **Noise contributions in an inducible genetic switch: a whole-cell simulation study.** *PLoS computational biology* 2011, **7**:e1002010.

7. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype.** *Cell* 2012, **150**:389–401.

   •• Describes a novel approach for combining heterogeneous data and models, and the first model that represents the function of each characterized gene of an organism.

8. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO: **Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics.** *Cell Syst* 2015, **1**:283–292.

   • Describes the development of genome-scale kinetic models of cellular metabolism.

9. Macklin DN, Ruggero NA, Covert MW: **The future of whole-cell modeling.** *Curr Opin Biotechnol* 2014, **28**:111–115.

   •• Summarizes the challenges to achieve WC models including aggregating the data needed to build WC models; merging pathway submodels; simulating, calibrating, and validating large models; and coordinating large modeling teams.

10. Macaulay IC, Voet T: **Single cell genomics: advances and future perspectives.** *PLoS Genet* 2014, **10**:e1004126.

11. Altelaar AM, Munoz J, Heck AJ: **Next-generation proteomics: towards an integrative view of proteome dynamics.** *Nat Rev Genet* 2013, **14**:35.

12. Fuhrer T, Zamboni N: **High-throughput discovery metabolomics.** *Curr Opinion Biotechnol* 2015, **31**:73–78.

13. Laird PW: **Principles and challenges of genome-wide DNA methylation analysis.** *Nat Rev Genetics* 2010, **11**:191.

14. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.** *Nat Rev Genet* 2013, **14**:390.

15. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nat Rev Genet* 2009, **10**:669.

16. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R *et al.*: **Highly multiplexed subcellular RNA sequencing in situ.** *Science* 2014, **343**:1360–1363.

17. Saliba AE, Westermann AJ, Gorski SA, Vogel J: **Single-cell RNA-seq: advances and future challenges.** *Nucleic Acids Res* 2014, **42**:8845–8860.

18. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK: **A deep profiler's guide to cytometry.** *Trends Immunol* 2012, **33**:323–332.

19. Consortium TU: **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res* 2017, **45**:D158–D169.

20. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA *et al.*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2016, **44**:D471–D480.

21. Sajed T, Marcu A, Ramirez M, Pon A, Guo AC, Knox C, Wilson M, Grant JR, Djoumbou Y, Wishart DS: **ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli.** *Nucleic Acids Res* 2016, **44**:D495–D501.

22. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T *et al.*: **ArrayExpress update–simplifying data submissions.** *Nucleic Acids Res* 2015, **43**:D1113–D1116.

23. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, Mering C: **Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines.** *Proteomics* 2015, **15**:3163–3168.

24. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, Algaa E, Weidemann A, Sauer-Danzwith H, Mir S *et al.*: **SABIO-RK–database for biochemical reaction kinetics.** *Nucleic Acids Res* 2012, **40**:D790–D796.

25. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ *et al.*: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26**:1608–1615.

26. Goldberg AP, Chew YH, Karr JR: **Toward scalable whole-cell modeling of human cells.** in: Proc 2016 Annu ACM Conf SIGSIM Princip Adv Discrete SimulACM; 2016:259–262
    • Proposes a parallel algorithm for simulating rule-based WC models.

27. Cohen PR: **DARPA's Big Mechanism program.** *Phys Biol* 2015, **12**:045008.

28. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, Goebelbecker HJ, Gundlach J, Schirmbacher P, Dierolf U: **Making research data repositories visible: The re3data.org Registry.** *PloS One* 2013, **8**:e78080.

29. Cokelaer T, Pultz D, Harder LM, Serra-Musach J, Saez-Rodriguez J: **BioServices: a common Python package to access biological web services programmatically.** *Bioinformatics* 2013, **29**:3241–3242.

30. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, Kothari A, Weaver D, Lee T, Subhraveti P *et al.*: **Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology.** *Brief Bioinform* 2016, **17**:877–890.

31. Karr JR, Sanghvi JC, Macklin DN, Arora A, Covert MW: **WholeCellKB: model organism databases for comprehensive whole-cell models.** *Nucleic Acids Res* 2013, **41**:D787–D792.
    • Describes a tool for organizing data for WC modeling.

32. Helikar T, Kowal B, Rogers J: **A cell simulator platform: the Cell Collective.** *Clin Pharmacol Ther* 2013, **93**:393–395.

33. Latendresse M, Krummenacker M, Trupp M, Karp PD: **Construction and completion of flux balance models from pathway databases.** *Bioinformatics* 2012, **28**:388–396.

34. Lopez CF, Muhlich JL, Bachman JA, Sorger PK: **Programming biological models in Python using PySB.** *Mol Syst Biol* 2013, **9**:646.

35. Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D *et al.*: **SEEK: a systems biology data and model management platform.** *BMC Syst Biol* 2015, **9**:33.

36. Resasco DC, Gao F, Morgan F, Novak IL, Schaff JC, Slepchenko BM: **Virtual Cell: computational tools for modeling in cell biology.** *Wiley Interdiscip Rev Syst Biol Med* 2012, **4**:129–140.

37. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al.*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19**:524–531.

38. Harris LA, Hogg JS, Tapia JJ, Sekar JA, Gupta S, Korsunsky I, Arora A, Barua D, Sheehan RP, Faeder JR: **BioNetGen 2.2: advances in rule-based modeling.** *Bioinformatics* 2016, **32**:3366–3368.

39. Mendes P, Hoops S, Sahle S, Gauges R, Dada J, Kummer U: **Computational modeling of biochemical networks using COPASI.** *Methods Mol Biol.*

40. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR: **Cobrapy: Constraints-based reconstruction and analysis for python.** *BMC systems biology* 2013, **7**:74.

41. Dhar PK, Takahashi K, Nakayama Y, Tomita M: **E-Cell: Computer simulation of the cell.** *Rev Cell Biol Mol Med.*

    • Describes a multi-algorithmic simulator.

42. Penas DR, González P, Egea JA, Doallo R, Banga JR: **Parameter estimation in large-scale systems biology models: a parallel and self-adaptive cooperative strategy.** *BMC Bioinformatics* 2017, **18**:52.

43. Forrester AI, Keane AJ: **Recent advances in surrogate-based optimization.** *Progress in Aerospace Sciences* 2009, **45**:50–79.

44. Clarke EM, Faeder JR, Langmead CJ, Harris LA, Jha SK, Legay A: **Statistical model checking in BioLab: Applications to the automated analysis of T-cell receptor signaling pathway.** in: Int Conf Comput Meth Syst BiolSpringer; 2008:231–250

45. Kwiatkowska M, Norman G, Parker D: **PRISM 4.0: Verification of probabilistic real-time systems.** in: Computer Aided VerificationSpringer; 2011:585–591

46. Karr JR, Phillips NC, Covert MW: **WholeCellSimDB: a hybrid relational/HDF database for whole-cell model predictions.** *Database.*

    • Describes a tool for organizing WC simulation results for visualization and analysis.

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

48. Waltemath D, Karr JR, Bergmann FT, Chelliah V, Hucka M, Krantz M, Liebermeister W, Mendes P, Myers CJ, Pir P *et al.*: **Toward community standards and software for whole-cell modeling.** *IEEE Trans Biomed Eng* 2016, **63**:2007–2014.

    • Describes the need for a new language to represent WC models.

49. Medley JK, Goldberg AP, Karr JR: **Guidelines for reproducibly building and simulating systems biology models.** *IEEE Trans Biomed Eng* 2016, **63**:2015–2020.

50. Karr JR, Williams AH, Zucker JD, Raue A, Steiert B, Timmer J, Kreutz C, Wilkinson S, Allgood BA, Bot BM *et al.*: **Summary of the DREAM8 parameter estimation challenge: toward parameter identification for whole-cell models.** *PLoS Comput Biol* 2015, **11**:e1004096.