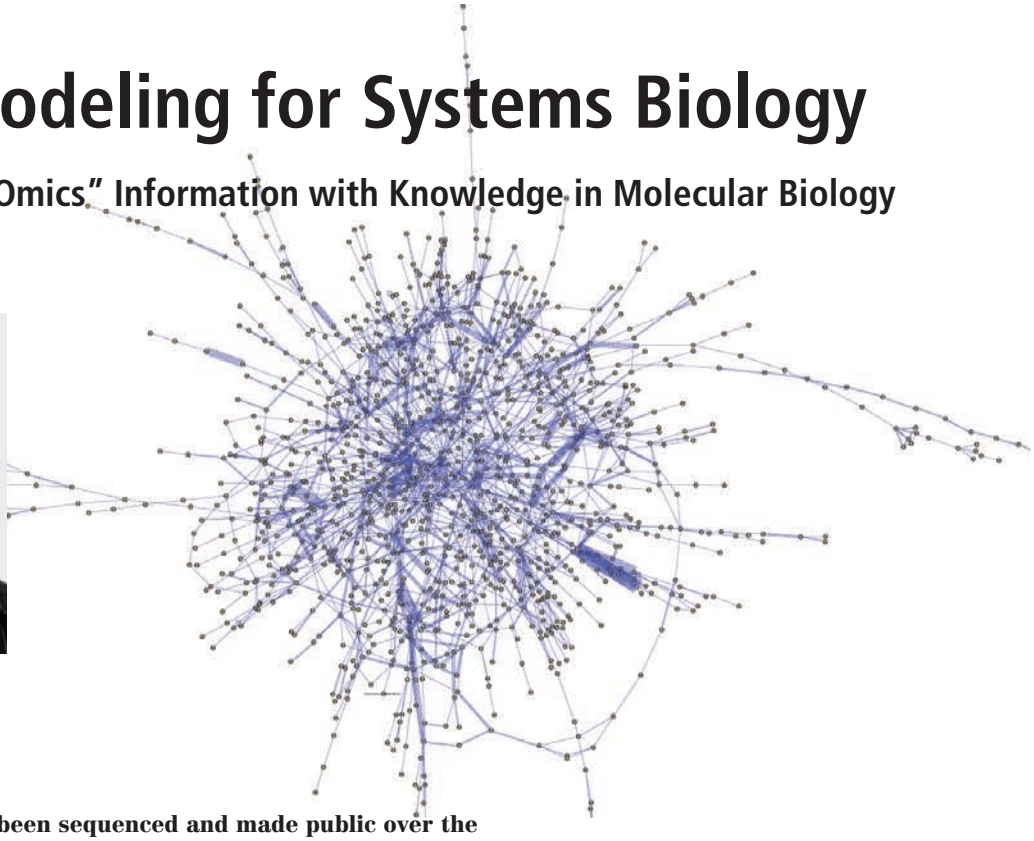
Large-scale Modeling for Systems Biology

Systematic Integration of "Omics" Information with Knowledge in Molecular Biology

Prof. Masaru Tomita,
Director, Keio University

Kazuharu Arakawa,
PhD, Keio University



Hundreds of complete genomes have been sequenced and made public over the last decade, with a deluge of comprehensive and quantitative data on intracellular molecules through the recent advances in high-throughput measurement technologies. Grand challenge of life sciences is now primarily in the integration of these "omics" information with the reservoir of knowledge in molecular biology to discover how these components concert to form the complex and dynamic behaviors of living systems.

Dynamic Simulation of the Whole Cell

Introduction of the genome sequence data and the subsequent high-throughput "omics" data in the fields of transcriptome, proteome, metabolome, and interactome has quickly transformed molecular biology into a data rich, and in some aspects, data-driven discipline. On the other hand, omics approach nonetheless remains inherently reductive and descriptive. Systems biology approach is therefore anticipated to enhance our understanding of cells as living systems, through the integration of comprehensive data of the catalogues of molecular components to observe how the multitudes of interactions shape the complex behaviors of the whole cell. For this purpose, computational modeling and simulation are invaluable tools for the observation and experimentation of the systematic behaviors resulting from the nonlinear interactions. The E-Cell Project (www.e-cell.org/) is dedicated to this challenge, by developing a cell simulation software environment that supports object-oriented modeling and multi-algorithm/timescale simulation, and by constructing and analyzing dynamic simulation models for several types of cells and cellular processes such as the mitochondrion, neuron, erythrocyte, circadian rhythms, and central carbohydrate metabolism of *Escherichia coli*. Cellular systems are commonly modeled using simultaneous ordinary differential equations based on mass action and enzyme reaction kinetics such as the derivations from Henri-Michaelis-Menten type equations, and modeling with this approach requires large amount of time and labor-intensive manual effort to search through the literature to obtain accurate parameters and equations necessary for the quantitative mathematical formulation. In many cases it is difficult to obtain the complete set of kinetic information from literature and databases, and therefore parameter estimation using computational methods such as the genetic algorithms is further required.

Automatic Prototyping of Large-scale Models with Omics Data

Although it is still not possible to cover the entire modeling processes for systems biology, automation is possible to a certain extent of model prototyping procedures to allow the construction of large-scale cell-wide models by utilizing the wealth of omics information. Here the key challenge for automatic modeling centrally resides in computational database integration of the myriad of data formats and identifiers, which are spread among specific public biological information resources. Moreover, integration of biological information requires semantic mapping of terms in different biomolecular layers (i.e. genes, mRNAs, proteins, metabolites) with various synonyms for each entity. For example, pyruvate kinase I gene in *Escherichia coli* may also be called PK-1, PK-I, pykF, pyruvate kinase F, or b1676, and this gene exists in the forms of DNA (gene), mRNA, and protein. Meta-database integration therefore depends on nontrivial tasks of text mining and curation of a set of controlled vocabularies and dictionaries of synonyms. However, knowledge in molecular biology is mostly structured based on the central dogma, and the data entries represented in the databases are primarily linked to the corresponding nucleotide sequences regardless of the biological layer. Therefore, we have taken a gene-centric approach for massive data

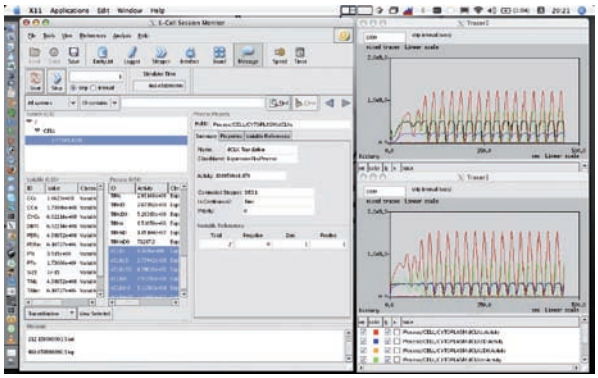


Fig. 1: Screenshot of E-Cell Simulation Environment version 3. E-Cell System allows dynamic simulation of cellular systems *in silico*, and has several advantages over other cell simulation platforms for its object-oriented modeling capabilities, multi-algorithm/time-scale simulation, and high extensibility with Python programming language. The software is actively developed and distributed with the open-source GNU General Public License at www.e-cell.org/.

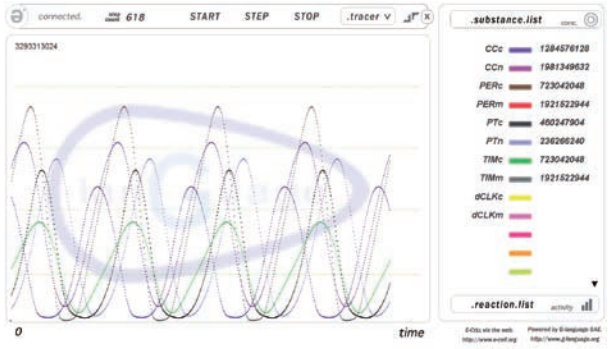


Fig. 2: Simulation of circadian clock machinery of *Drosophila* with E-Cell System. Dynamic behavior of cells results from the non-intuitive complex interactions among the cellular components. Here the feedback loops of PER and CLK proteins produce 24-hour oscillation.

integration, by using the sequence information as the primary dictionary and querying by similarity-based tools such as BLAST to integrate information from various resources.

Another advantage of this gene-centric approach is the possibility to use the genome information as the starting point. Since the complete set of genes in the genome defines the maximal number of intracellular components, we can confine the search space for whole cell modeling and gather related information from the gene sequences. We have developed a software system designated the Genome-based Modeling (GEM) System, which converts a complete genome into a metabolic pathway model of the organism by matching all composite genes to bipartite graphs of stoichiometric enzyme reactions automatically through similarity and orthology searches combined with database integration techniques. Although the models automatically generated by GEM System does not contain kinetic information *per se*, resulting stoichiometric matrix or the model in the standard Systems Biology Markup Language (SBML) can be readily applied to flux-based analyses and graph theoretical researches to observe the systematic properties of the entire metabolic pathway. Since this pathway model already limits the maximum search space based on the complete genome, it is also suitable as a prototype for top-down modeling to add kinetic information.

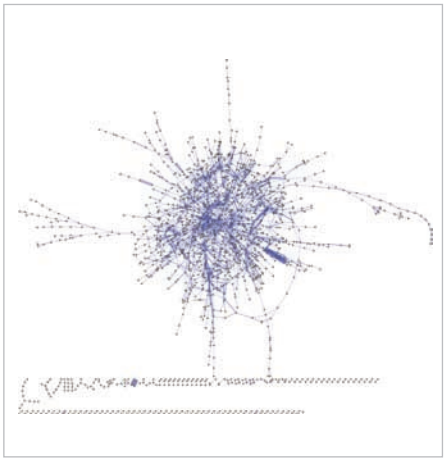


Fig. 3: Metabolic pathway of *Escherichia coli* automatically generated by GEM System and visualized with Cytoscape. This model contains 968 reactions of 1195 metabolites, achieving 100% KEGG coverage.

To validate the accuracy of GEM System, we have automatically generated the pathway models for over 90 bacterial genomes and compared them with the KEGG pathway database. Most models contained more than 500 enzymes and over 800 metabolites, and achieved 90~100% KEGG coverage. Model of *Escherichia coli* had the best accuracy, comprised of 968 reactions of 1195 metabolites, with 100% KEGG coverage and 92% EcoCyc coverage. Complete listings of the generated downloadable models are available at www.g-language.org/gem/models/static.cgi.

While GEM System is currently limited to the static stoichiometric representation of the metabolic pathway, the gene-centric approach is scalable in its application to almost all types of biomolecular layer, including gene regulation and signal transduction pathways. Quantitative data from high-throughput time-series measurements especially in the fields of metabolomics will enhance the dynamic aspects of the cellular network as a system, coupled with our knowledge of biochemical pathways.

References

[1] Arakawa K. *et al.*: BMC Bioinformatics 7, 168 (2006)
[2] Arakawa K. *et al.*: Journal of Pesticide Science 31, 282-288 (2006)
[3] Arakawa K. *et al.*: Bioinformatics 19, 305–306 (2003)
[4] Arakawa K. *et al.*: In Silico Biology 5, 0039 (2005)
[5] Tomita M. *et al.*: Bioinformatics 15, 72–84 (1999)
[6] Tomita M.: Trends in Biotechnology 19, 205–210 (2001)
[7] Tomita M.: Bioinformatics 17, 1091–1092 (2001)
[8] Arita M. *et al.*: Current Opinion in Biotechnology 16, 344–349 (2005)
[9] Takahashi K. *et al.*: Bioinformatics 20, 538–546 (2004)

For further information on the research referred to in this article, publications, detailed references, and activities of the research groups, see www.iab.keio.ac.jp.

▸ **www.eMagazineBIOforum.com**

CONTACT:
Kazuharu Arakawa, PhD
Prof. Masaru Tomita
Institute for Advanced Biosciences
Keio University, Japan
Tel.: +81 466 47 5099
Fax: +81 466 47 5099
mt@sfc.keio.ac.jp