From Whole-Cell Simulation to WholeCellKB: A Critical Review of Whole-Cell Computational Modeling

By: Bojan Milic

Final Project/Paper

BIOC 218

Professor Doug Brutlag and Dr. Dan Davison

Department of Biochemistry

Stanford University

April 6, 2013

Introductory Overview

As Jeremy Gunawardena, Professor of Systems Biology at Harvard Medical School, wrote in a recent review, "a model is not a description of reality, [but rather] a description of our assumptions about reality".[1] In essence, the comparison of the consequences and results of scientific models, including those of computational, or *in silico*, simulations to experimentally-observed outcomes serves as a test of the validity of the existing understanding of nature and natural phenomena.[1-4] History has shown that much can be learned from reconciling models which are based on the existing understanding of a given phenomenon, but which yield results and predictions which are seemingly incompatible with the available experimental data. A notable example would be the discovery of energy quantization by Max Planck (and the subsequent development of the field of quantum mechanics), which came about as a result of efforts to develop adequate experimentally-consistent theoretical models for blackbody radiation.[5] Moreover, the development of models can contribute to identifying gaps in experimental knowledge, thereby indicating potential avenues of investigation which might lead to discovery. A recent prominent example of a scientific finding which emerged as a result of attempts to reconcile theoretical models with experimental data would be the discovery of the Higgs Boson.[6-10] Indeed, the discovery of a boson which is central to electroweak symmetry breaking, now commonly referred to as the Higgs Boson, nearly half a century after its existence was hypothesized based on the Standard Model of particle physics is certainly one of many examples in recent scientific history which serves to illustrate the notion that the predictions of models can serve as valuable guides towards discovery through experimental science.[6-10] In essence, the claim can certainly be made that the development of models, including computational simulations, is a valuable pathway of investigation towards scientific discovery.

Over the past several decades, the development of genomics and high-throughput techniques has facilitated the characterization of cells and cellular processes to the point that the transcriptome, proteome, and metabolome has been determined for a number of biologically-relevant model organisms.[11-13] Indeed, strides in fields such as proteomics and genomics, developments in bioinformatics to deal with the rapid augmentation of biological data, and the exponential increase in computing power over the past several decades, have for the first time brought within reach the tantalizing possibility of comprehensively modeling an entire cell, a possibility that has been considered at the turn of the millennium as one of the "grand challenges of the 21st century".[2,4,14] A number of collaborative projects have been undertaken over the years with the aim of eventually producing a computational whole-cell model. A joint effort between a team of Japanese scientists and The Institute for Genomic Research (now known as the J. Craig Venter institute) established in 1996 and known as the E-CELL project has been responsible for a number of early steps towards the larger goal of a producing a comprehensive whole-cell model *in silico*.[2,4] In particular, the E-CELL project achieved the first simulation of a hypothetical cell using a minimal set of 127 genes, namely 105 protein coding and 22 RNA-coding genes, which is roughly 4 times fewer than is contained by the bacterium *Mycoplasma genitalium,* which possesses the smallest number of genes found in any known organism.[2,4,14-16] Subsequent work by the E-CELL collaboration has entailed the successful development of computational models at the level of organelles and, more recently, at the level of the metabolome.[2,17-19] Nevertheless, until only recently, a comprehensive computational model of an entire cell of an existing species has been elusive, primarily due to the necessity that such a whole-cell *in silico* model integrate several quite different models for the various biological processes that occur within a cell.

Despite the early progress of the E-CELL project, the first, and thus far only, comprehensive whole-cell computational simulation of a known, existing organism, namely *M. genitalium*, was successfully produced in mid-2012 by a group at Stanford University led by Markus Covert of the Department of Bioengineering in collaboration with scientists at the J. Craig Venter Institute.[14] The *in silico* model, which will be described and discussed in further detail below, accounts for all 525 genes of *M. genitalium* and is based on a vast collection of genomic, proteomic, and metabolomic data amounting to more than 1,900 experimentally-validated parameters from over 900 publications.[1,3,14] In light of the fact that such whole-cell computational modeling necessitates the integration of vast amounts of data scattered in literally hundreds of scientific papers and dozens of databases, the development of the first whole-cell *in silico* model has highlighted the need for a novel kind of database bringing together the entire genomic, proteomic, and metabolomic data sets of a number of known organisms of particular interest (such as model organisms).[14,20] Here, we present a critical review of recent advances in the nascent field of whole-cell simulation and their impact on the emergence of model organism databases, most notably WholeCellKB, specifically tailored for the purpose of facilitating the development of whole-cell computational models.[14,20]

Computational Whole-Cell Simulation

As mentioned previously, recent work by the laboratory of Markus Covert has produced the first predictive whole-cell computational model of a living organism, namely *Mycoplasma genitalium*.[1,3,14] The choice of *M. genitalium*, a Gram-positive bacterium which functions as a human urogenital parasite, was made due to its small genome of merely 525 genes.[14] In order to model the whole cell, the entirety of the cellular functionality of *M. genitalium* was broken down into a total of 28 modules corresponding to cellular

processes.[14] Each of the 28 modules was independently modeled using the most suitable mathematical model for the requisite cellular process over a short time scale, upon which the modules were integrated into the larger, overall model.[14] It must here be noted that the whole-cell model put forth by Kerr *et al.* relies on the assumption that the sub-models used for each of the 28 modules are independent of each other on a one-second timescale.[14] A total of 16 cellular states were used to represent the complete configuration of the entire bacterium modeled at a given point in time.[14]

The whole-cell model developed is, in essence, a system of ordinary differential equations (ODEs), where the 28 modules serve as the differential equations and the 16 cellular states are the state variables of the ODEs.[14] The *in silico* whole-cell simulations were performed using an algorithm related to numerical integration methods commonly used to solve ODEs (such as the fourth-order Runge-Kutta method).[14,21] Briefly, the simulation algorithm entails the initialization of state variables, upon which the calculation of how each of the 16 cellular states evolves over a one-second time course is repeatedly performed until the cell undergoes division, which results in the termination of the particular simulation round.[14] The determination of the temporal changes of the 16 cellular variables was achieved by distributing each of the state variables between the 28 sub-model modules of cellular processes and using the results of each of the modules to update the cell state variables, thereby successfully integrating the independently-modeled modules to yield the computational whole-cell simulation.[14] A summary of the computational whole-cell model described above is provided in Figure 1.[1]

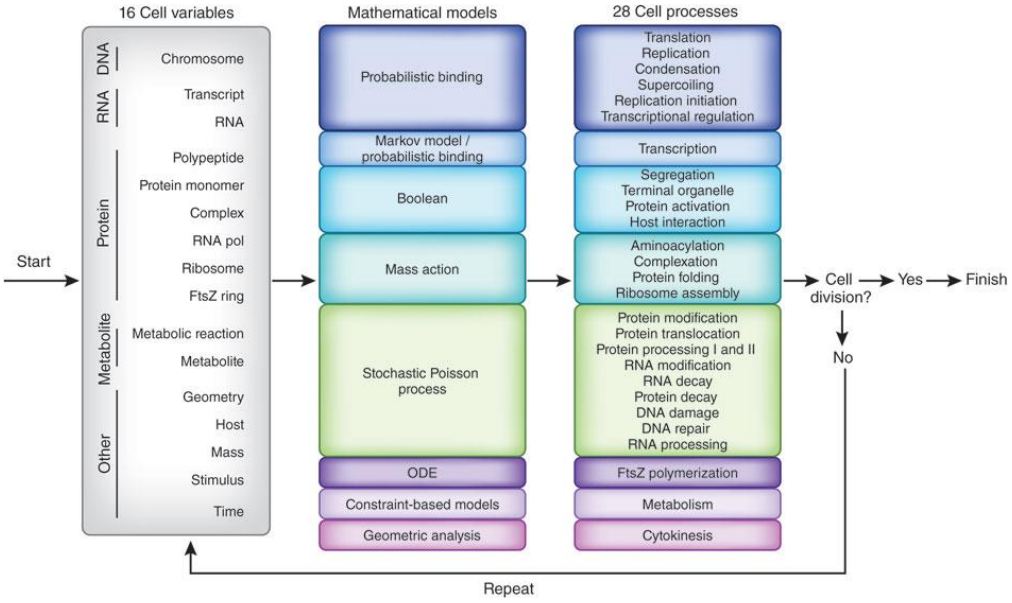

Figure 1: A visual representation of the computational whole-cell model for *M. genitalium* developed by Kerr *et al.*[1,14] Based on more than 1,900 parameters, the model involved the integration of a total of 28 independently-modeled cellular processes which were used to determine the state of the cell at a given time period through calculating the values of 16 cell state variables.[1,14]

It turns out that the model developed by Karr *et al.* is capable of recapitulating the most significant experimentally-observed aspects of *M. genitalium* with a reasonable degree of accuracy.[14] More specifically, the computational model predicts a doubling time of 8.9 hours (Figure 2B), which is entirely consistent with the experimentally-observed 9 hours (Figure 2A).[14] The predicted relative cellular composition of DNA, RNA, lipids, and protein is very close to experimentally-measured levels (Figure 2C), while the mass of the DNA, RNA, protein, membrane, and the mass of the entire contents of the cell were predicted by the model to double in approximately 9 hours, which is consistent with the doubling time of *M. genitalium* (Figure 2D).[14] Furthermore, simulations performed for single-gene deletion mutants were found to yield an accurate prediction concerning whether the particular deletion mutant would be viable for as many as 79% of all possible single-gene deletion mutants of *M. genitalium*.[14] As a whole, these results are certainly significant contributors towards evaluating the validity of the model relative to experimental findings.[14]

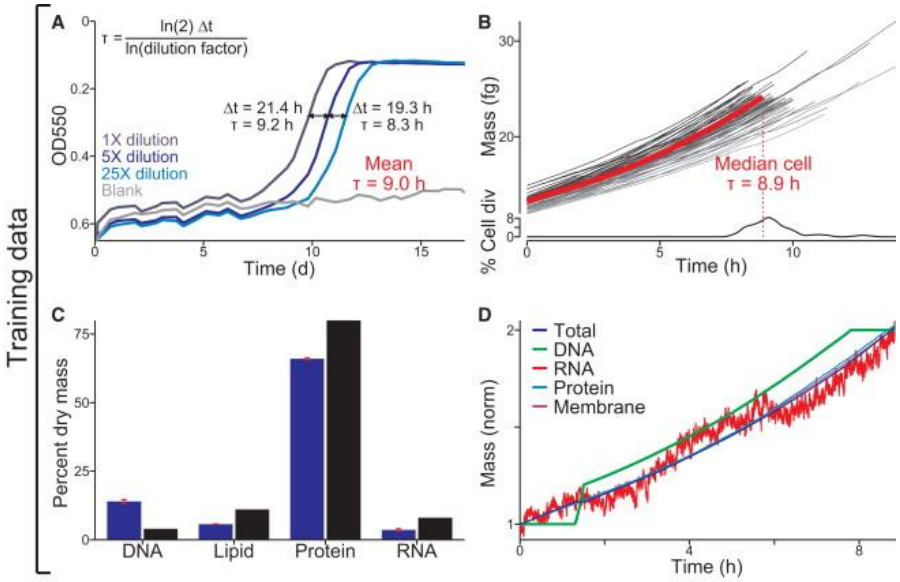

Figure 2: Shown (top) are the experimental (A) and predicted (B) doubling times of *M. genitalium*.[14] Also shown (bottom) is a comparison of the experimentally observed cellular content (black; C) to the predicted relative content of DNA, lipids, protein, and RNA (blue; C), in addition to a plot indicating the change in various cellular contents over the course of the 9-hour doubling time of *M. genitalium* (D).[14]

Incredibly, this early-stage, "first draft" of a comprehensive whole-cell computational model presented has been shown to possess substantial predictive power, ranging from being able to compute the interactions of every DNA binding protein at the level of the entire genome to predicting the collision rates of individual proteins.[1,3,14] Indeed, the predictions and consequences derived from the results of the simulations performed show potential for model-driven biological discovery.[14] For example, the computational model suggests a metabolic emergent control of the duration of the cell cycle that does not depend on genetic regulatory mechanisms.[14] Although the doubling time of *M. genitalium* has been determined to be highly consistent (Figure 2), the model predicts that the durations of both the replication initiation as well as replication are widely distributed (Figure 3).[14] Further examination of the results of 128 separate simulations of wild-type *M. genitalium* have shown that a long replication initiation results in a high concentration of dNTP available for replication, which, in turn, allows for replication to proceed faster (Figure 3).[14] In contrast, a short replication initiation means that a lower

amount of dNTP is produced during replication initiation, thereby resulting in a lower concentration of dNTP present for replication and prolonging the duration of replication (Figure 3).[14] Thus, the strong linear dependence of replication duration on the concentration of dNTP, which is itself a linear function of the length of the replication initiation, allows for the overall duration of the cell cycle to be highly consistent despite the large variability in the durations of replication and replication initiation, respectively (see Figure 3).[14]

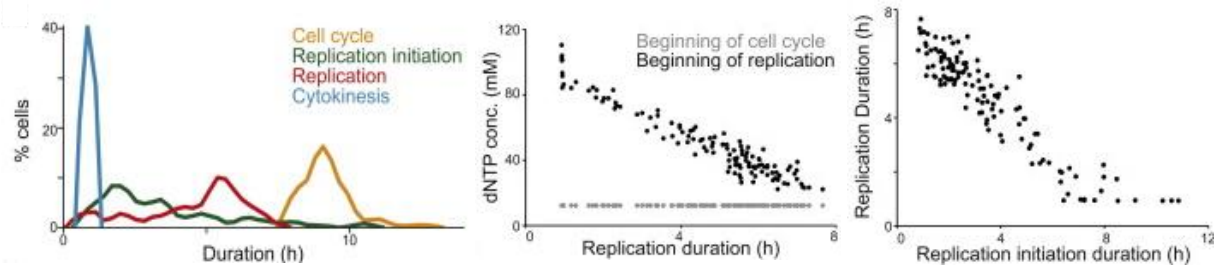

Figure 3: Shown above (left) is the distribution of durations of the cell cycle, replication initiation, replication, and cytokinesis for individual simulations of 128 wild-type *M. genitalium* cells, as predicted by the whole-cell *in silico* model put forth by Karr *et al.*[14] Also shown (center) is a plot of the dependence of the duration of replication on dNTP concentration, as well as (right) the duration of replication as a function of the duration of replication initiation.[14]

WholeCellKB

The development and implementation of computational whole-cell models based on the method presented by Karr *et al.* is reliant on the availability of comprehensive model organism databases which include detailed information regarding the genome, proteome, transcriptome, and metabolome of the desired species or model organism.[14,20] In order to facilitate future modeling efforts, Karr *et al.*, in addition to developing and implementing the first truly whole-cell *in silico* model, have produced a novel, open-source, web-based program that enables and streamlines the assembly of comprehensive model organism databases.[20] The program, aptly termed WholeCellKB, provides a structure for the organized collection of descriptions of any species of interest, including information regarding genes, proteins, macromolecular interactions, and metabolic pathways, to name a few, from a wide array of diverse sources into a single, comprehensive, and integrated database.[20] The

overall role of a database such as WholeCellKB in the process of producing whole-cell computational simulations is visually summarized in Figure 4 below:[20]

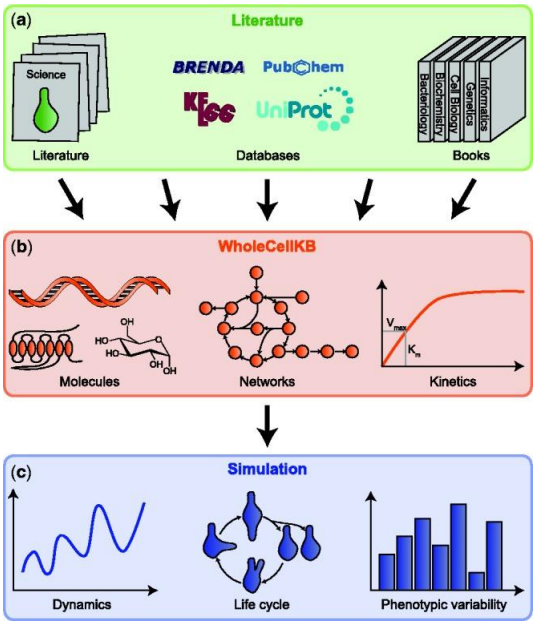

Figure 4: Shown above is a schematic representation showing the central role of WholeCellKB (b) in providing a convenient means of accumulating the comprehensive genomic, transcriptomic, proteomic, and metabolomic data (a) from a wide variety of sources (literature, databases, books) necessary for the production of a phenotypically-predictive whole-cell simulation (c).[20]

To illustrate the utility in using WholeCellKB to produce model organism databases, Karr *et al.* have implemented WholeCellKB to assemble the first such comprehensive database for the bacterium *M. genitalium*.[20] Termed WholeCellKB-MG, the database contains the complete representation of *M. genitalium* at the molecular level necessary for the production of a whole cell model, including, but not limited to: (I) organization at the subcellular level, (II) the genome sequence, (III) chromosomal features, (IV) information regarding the location, size, direction, and essentiality of each gene, (V) the organization and promoter of each transcription unit, (VI) RNA transcript expression and degradation rates, (VII) the specific RNA folding and maturation pathways, (VIII) the specific DNA folding and maturation pathways, (IX) the subunit compositions of all known *M. genitalium* macromolecular complexes, (X) the binding sites of all DNA-binding proteins known to be

present in *M. genitalium*, (XI) the structure, charge, and hydrophobicity of each metabolite, (XII) the details of all chemical reactions known to take place in *M. genitalium* (including kinetics, energetics, catalysis, stoichiometry, and information relating to any coenzymes involved), (XIII) all known transcription factors and their individual regulatory roles, (XIV) the complete chemical composition of *M. genitalium,* and (XV) the chemical compositions of laboratory growth media used during experimental data collection.[20] In essence, this list of information contained by WholeCellKB-MG represents the complete details and descriptions of the entire genome, transcriptome, proteome, and transcriptome from over 900 research papers, databases, and books, which amounts to approximately 1,900 parameters.[20] Figure 5 below offers a graphical summary of the description of *M. genitalium* cell physiology available in WholeCellKB-MG.[20]

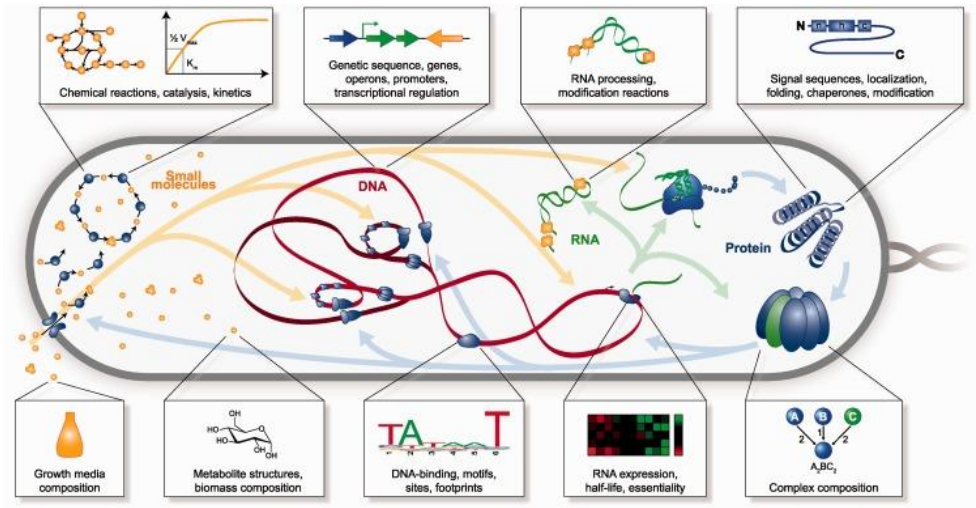

Figure 5: Shown above is a schematic representation summarizing the vast body of *M. genitalium* molecular, structural, genetic, proteomic, transcriptomic, and metabolic information accumulated and made available in WholeCellKB-MG.[20]

Using WholeCellKB-MG for the purpose of producing a whole-cell computational model offers a number of key advantages relative to the many existing databases already available. First of all, the WholeCellKB-MG database is capable of a broad representation of cell physiology, including the molecular details of each of the processes comprising the 28 modules used to produce the *M. genitalium* model described above (see Figure 1).[14,20]

Furthermore, WholeCellKB-MG explicitly defines and describes each individual participant or component of every chemical reaction and molecular interaction known to occur in *M. genitalium*.[20] Finally, each molecule and molecular reaction is not only linked to the corresponding structural information, but also with experimentally-determined quantitative descriptions, such as known kinetic parameters and reaction rate laws.[20]

Future Directions and Challenges

Although the development of the first comprehensive *in silico* whole-cell model of an existing species represents, at the very least, the first step towards the fulfillment of one of the "grand challenges of the 21st century," much work remains to be done.[1,3,6,14,20] While the *M. genitalium* model developed by Karr *et al.* is undoubtedly comprehensive, it is nevertheless not a complete and fully-accurate predictive model.[1,3,6,14] Furthermore, the fact that *M. genitalium* is difficult to work with in a laboratory setting limits the direct applicability of the model towards guiding biological discovery in widely used model organisms, such as *Escherichia coli*, or, more importantly, in human cells. However, the fact remains that the Human genome is roughly an order of magnitude larger than that of *E. coli*, whose genome is itself an order of magnitude larger than that of *M. genitalium*. Indeed, the volume genomic, transcriptomic, proteomic, and metabolomic information required to successfully model a more complex cell than *M. genitalium* (such as *E. coli*) certainly poses a formidable computational challenge.[14,20] Beyond the improvements in algorithms, techniques, and raw computational power, extensive model organism databases modeled on WholeCellKB-MG will be needed before a whole-cell *in silico* model *E. coli* is attainable.[6,14,20]

The development of more complex and complete computational whole-cell simulations has potentially far reaching practical implications beyond contributing to

fundamental biological discovery and the expansion of human knowledge. Indeed, the production of complete, whole-cell models which both match existing experimental results and are capable of predicting phenotypic outcomes based on an input genotype could potentially be used to streamline the design of bacteria to perform a specific role. For example, more detailed bacterial models would perhaps be able to aid the process of engineering of bacteria to produce a natural product drug that is impractical to generate by traditional synthetic methods, or to perhaps develop petroleum-metabolizing bacteria to aid with the clean-up of oil spills. Perhaps the ultimate goal of the emerging field of whole-cell modeling would be simulating an entire human cell. Such a predictive model would have tremendous implications on personalized and translational medicine. Beyond likely increasing the rate of medically-relevant biological discovery, a human whole-cell model with sufficient predictive power to be able to compute the consequences of the exposure of a particular cell type to a given drug-candidate compound would likely be of great practical utility in the process of drug discovery. Furthermore, the development of a human whole-cell simulation capable of predicting phenotype and cellular responses based on a given input genotype could potentially make a valuable contribution in the field of personalized medicine, particularly in light of the decreasing costs of genome sequencing and recent advances in next-generation sequencing (including the recently-reported attainment of single-cell sequencing).[22,23]

References

1) Gunawardena, J., *J. Mol. Biol. Cell*, **23**, 517 (2012).
2) Tomita, M., *Trends in Biotechnology*, **19**, 205 (2001).
3) Isalan, M., *Nature*, **488**, 40 (2012).
4) Di Ventura, B., *et al.*, *Nature*, **443**, 527 (2006).
5) McQuarrie, D. A., *Quantum Chemistry*, Second Edition. Sausalito: University Science Books, (2008).
6) ATLAS Collaboration, *Phys. Lett. B.*, **716**, 1 (2012).
7) CMS Collaboration, *Phys. Lett. B.*, **716**, 30 (2012).
8) Englert, F. and Brout, R., *Phys. Rev. Lett*, **13**, 321 (1964).
9) Higgs, P., *Phys. Rev. Lett*, **13**, 508 (1964).
10) Guralnik, G., *et al.*, *Phys. Rev. Lett*, **13**, 585 (1964).
11) Güell, M., *et al.*, *Science*, **326**, 1268 (2009).
12) Kühner, S. *et al.*, *Science*, **326**, 1235 (2009).
13) Yus, E., *et al.*, *Science*, **326**, 1263 (2009).
14) Karr, J. R., *et al.*, *Cell*, **150**, 389 (2012).
15) Tomita, M., *et al.*, *Genome Inform Ser Workshop Genome Inform*, **8**, 147 (1997).
16) Tomita, M., *et al.*, *Bioinformatics*, **15**, 72 (1999).
17) Yugi, K. and Tomita, M., *Bioinformatics*, **20**, 1795 (2004).
18) Kinoshita, A., *et al.*, *J. Biol. Chem*, **282**, 10731 (2007).
19) Yachie-Kinoshita, A., *et al.*, *J. Biomed. Biotechnol.*, **2010**, 642420 (2010).
20) Karr, J. R., *et al.*, *Nucleic Acids Res.*, **41**, D787 (2013).
21) Brannan, J. R. and Boyce, W. E., *Differential Equations*, Second Edition. Hoboken: Wiley, (2011).
22) Zong, C., *et al.*, *Science*, **338**, 1622 (2012).
23) Lu, S., *et al.*, Science, **338**, 1627 (2012).