Neal Mitra
Fall 2001
MBB 752a

Whole Cell Simulations:

An exponential increase in biological data and in computational power allow for an integrative approach to molecular and cellular biology

Massive amounts of biological data have been obtained in recent years from genome sequencing as well as from proteome, transcriptome, and metabolome projects, among others. The rapidly growing field of bioinformatics uses computational techniques to analyze, organize, and understand all of this vast data which comes from heterogeneous sources [1]. Once an organism's genome is fully sequenced and its gene products have been thoroughly identified and even classified into biochemical pathways, the next logical step is to use this organized set of information to understand the overall behavior of the cell[2]. After all, the ultimate goal of biology is not the generation of huge amounts of data (gene sequences, gene expression profiles, functional analyses) in itself but to use this data to understand the processes of life at the cellular and even organismal levels.

One way to use the tools of bioinformatics to understand life at more complex levels is to create computational simulations of whole cells. Just imagine the possibilities of computational models that can accurately predict the dynamic behavior of a cell. Questions could be answered about how a cell would behave if its environment were changed or if one of its genes were knocked out or over-expressed. Even more complex possibilities could be addressed *in silico*. For example, what genetic changes could be made to cause a cell to exhibit a certain behavior[2]? Although such simulations might not yield exact results, they would allow scientists to design better hypotheses for experimental investigations.

In addition to helping one elucidate more clear hypotheses, creating biological simulations would provide a theoretical framework that allows experimental observations to be corroborated with experimental conclusions[3]. Leloup and Goldbeter, who modeled circadian rhythms in *Drosophila*, indicate several other advantages to biological modeling. Such models would allow the analysis of complex situations with multiple, coupled variables, increase the ability for determining the quantitative and qualitative effects of key parameters in living systems, and allow questions to be asked which are difficult to address experimentally. In addition, such models could yield surprising predictions with counterintuitive explanations[3].

An increase in computing power has made possible the design of various kinds of simulations in recent years. A simple simulation of water molecules that once took two weeks to perform on a supercomputer in 1986 can now be run on a desktop computer in less than two days [4]. Biochemical reactions have been simulated on computers since the 1940s [5] but the increased processing speed of computers has made more complex modeling possible today. A range of metabolic pathway modeling has already been achieved (for examples: [6, 7]) but are only a first step towards whole cell simulation.

Currently, attempts are being made to integrate genomic, proteomic, and biochemical pathway information with computational tools to create simulations of entire cells [8]. Table 1 shows a list of some of the resources available from projects that are underway. These projects must address a variety of technical issues including how to take data and construct it into a model.

Table 1: Partial List of Simulation Software Tools
E-Cell: http://www.e-cell.org
Virtual Cell: http://www.nrcam.uchc.edu
DBSolve: http://websites.ntl.com/~igor.goryanin
Gepasi: http://www.gepasi.org
MCell: http://www.mcell.psc.edu
CellML: http://www.cellml.org
Adapted from [8].

Goryanin et al. created a program called DBSolve, written in C++, to respond to the growing demand of making models[9]. Their software simplifies the process of deriving and then analyzing models from metabolic reconstructions. GEPASI is another tool being developed to design models of metabolic and chemical pathways and it too includes analytical and optimization tools[8]. These tools are useful components of creating a more robust simulation of a cell, but are not actually whole cell simulators.

The E-CELL project was initiated in 1996 in Japan, after the publication (by TIGR) of the entire genome sequence of *Mycoplasma genitalium* which has the smallest known genome (580 kb) and the smallest number of genes (~480)[2]. Having such a small number of genes to work with makes the task of whole cell modeling more within grasp. Tomita et al. took up the challenge of developing an integrative model that incorporates gene regulation, metabolism and signaling [10]. Their work was a broadening of previous work which primarily dealt with subsystems of a cell. Although previous models included complicated biochemical pathways (many of which interconnect), they could not account for the influence of cellular processes such as gene regulation and the cell cycle. In addition, their models were often highly qualitative.

E-CELL, written in C++, uses a set of reaction rules and numerically integrates the differential equations described in these rules. It was used to make a model based on only 127 genes from *M. genitalium* sufficient for transcription, translation, energy production, and phopholipid synthesis. The program established a virtual self-surviving cell (SSC) that takes up glucose, metabolizes it through glycolysis and produces ATP as an energy source. The ATP is consumed for protein synthesis as the 127 genes are transcribed and translated. Proteins and the cell membrane are designed to degrade over time so more protein and phospholipds must be synthesized. There are a total of 495 reaction rules that govern the cellular processes. This simple model already led to an unexpected, biologically interesting observation about intracellular ATP levels [2].

Another whole cell simulator, called the Virtual Cell, was developed by Schaff et al. to allow biologists with little training in mathematics to use computational cell biology [8]. The software can accommodate structural information in its simulations and may prove useful in considering biological data such as localization of proteins and compartmentalization. Both Virtual Cell and E-CELL have simple and helpful graphic user-interfaces.

Although this recent generation of whole cell simulators are promising, they have many limitations. The vast amounts of data which still must be considered in accurate models of entire cells include levels of gene expression, amounts of post-transcriptional and post-translational modifications, changing cellular volumes, subcellular localization, among others. In addition, incorporating the diverse *kinds* of data that are available (from digital microscope images to gene expression profiles) provides a great challenge. Another issue which should be considered as these software tools are developed is the possibility of standardizing components of models so that simulations created by different groups can be joined together.

Despite the current limitations in whole cell simulation, a forward thinking biologist may look to the future when more complete cell and even organismal simulators have been established. James Bassingthwaighte recently wrote, "The time has arrived in biological science to put it all back together . . . the advances are so rapid, the knowledge so detailed, and the scope of the new information so broad, that the consequences of the discoveries are often obscured by the complexity of the systems being elucidated[11]." The challenge now before biologists is to assimilate diverse pieces of biological data to yield a product that is greater than the sum of its parts.

Notes

1. Luscombe, N.M., D. Greenbaum, and M. Gerstein, *What is bioinformatics? A proposed definition and overview of the field.* Methods Inf Med, 2001. **40**(4): p. 346-58.
2. Tomita, M., *Whole-cell simulation: a grand challenge of the 21st century.* Trends Biotechnol, 2001. **19**(6): p. 205-10.
3. Leloup, J.C. and A. Goldbeter, *Modeling the molecular regulatory mechanism of circadian rhythms in Drosophila.* Bioessays, 2000. **22**(1): p. 84-93.
4. Gerstein, M. and M. Levitt, *Simulating water and the molecules of life.* Sci Am, 1998. **279**(5): p. 100-5.
5. Mendes, P. and D. Kell, *Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.* Bioinformatics, 1998. **14**(10): p. 869-83.
6. de Atauri, P., A. Sorribas, and M. Cascante, *Analysis and prediction of the effect of uncertain boundary values in modeling a metabolic pathway.* Biotechnol Bioeng, 2000. **68**(1): p. 18-30.
7. Polle, A., *Dissecting the superoxide dismutase-ascorbate-glutathione-pathway in chloroplasts by metabolic modeling. Computer simulations as a step towards flux analysis.* Plant Physiol, 2001. **126**(1): p. 445-62.
8. Loew, L.M. and J.C. Schaff, *The Virtual Cell: a software environment for computational cell biology.* Trends Biotechnol, 2001. **19**(10): p. 401-6.
9. Goryanin, I., T.C. Hodgman, and E. Selkov, *Mathematical simulation and analysis of cellular metabolism and regulation.* Bioinformatics, 1999. **15**(9): p. 749-58.
10. Tomita, M., et al., *E-CELL: software environment for whole-cell simulation.* Bioinformatics, 1999. **15**(1): p. 72-84.
11. Bassingthwaighte, J.B., *Strategies for the physiome project.* Ann Biomed Eng, 2000. **28**(8): p. 1043-58.