E-CELL: Software Environment for Whole Cell Simulation

Masaru Tomita [1]
mt@sfc.keio.ac.jp
Kenta Hashimoto [1]
kem@sfc.keio.ac.jp
Kouichi Takahashi [1]
t94249kt@sfc.keio.ac.jp
Tom Shimizu [1]
tom@sfc.keio.ac.jp
Yuri Matsuzaki [1]
t94402ym@sfc.keio.ac.jp
Fumihiko Miyoshi [1]
t95894fm@sfc.keio.ac.jp
Kanako Saito [1]
t95401ks@sfc.keio.ac.jp
Sakura Tanida [1]
t94613st@sfc.keio.ac.jp
Katsuyuki Yugi [1]
t95980ky@sfc.keio.ac.jp
J. Craig Venter [2]
venter@tigr.com
Clyde A. Hutchison [2]
clyde@tigr.com

[1] Laboratory for Bioinformatics
Keio University
5322 Endo, Fujisawa, 252, Japan

[2] The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850, USA

Abstract

We present E-CELL, a generic computer software environment for modeling a cell and conducting experiments *in silico*. The E-CELL system allows a user to define functions of proteins, protein-protein interactions, protein-DNA interactions, regulation of gene expression and other features of cellular metabolism, in terms of a set of reaction rules. The system then executes those reactions iteratively, and the user can observe, through a computer display, dynamic changes in concentrations of proteins, protein complexes and other chemical compounds in the cell.

Using this software, we constructed a model of a hypothetical cell with only 127 genes sufficient for transcription, translation, energy production and phospholipid synthesis. Most of the genes are taken from *Mycoplasma genitalium*, the organism having the smallest known chromosome, whose complete 580kb genome sequence was determined at TIGR in 1995.

We discuss future applications of the E-CELL system with special respect to genome engineering.

Keywords computer simulation, Mycoplasma, genome engineering, bioinformatics, experiment *in silico*

1 Introduction

Complete genomes of a dozen microbes have been sequenced, and it has led to the emergence of the next phase of genome biology: *proteome analysis*. Systematic analyses of genes/proteins are now under way, and catalogues of all the protein functions of those microbial species will be constructed.

The challenge we face is to understand how those proteins work collectively as a cellular system. If we can successfully understand the proteome system, we should be able to predict consequences of changes introduced into the cell and/or its environment, such as knocking out a gene or altering ingredients of the culture medium. Possible consequences of such intervention include cell death, changes in growth rate, and increase or decrease in expression of specific genes.

Computer simulations are essential in understanding such complex systems. Many attempts have been made to simulate important molecular processes in both cellular and viral systems. Several groups have proposed and analyzed gene regulation and expression models by simulation (Koile and Overton 1989, Karp 1993, Arita *et al.* 1994, McAdams and Shapiro 1995). The cell division cycle has also been an active area of research for biological modeling and simulation (Tyson 1991, Novak and Tyson 1995). Signal transduction mechanisms have also been simulated. (Bray *et al.* 1993).

Many of these simulations utilize qualitative models to deal with the general lack of information in molecular biology, especially quantitative kinetic data. However, while qualitative models are generally useful when information is incomplete (Kuipers 1986), it often generates ambiguous results(Kuipers 1985) the behaviors of which are difficult to predict due to combinatorial explosion. For a review on computer simulations in biology, see Galper and Brutlag (1993).

We present E-CELL, a computer software environment for modeling and simulation of the cell. The E-CELL system is a generic object-oriented environment for simulating molecular processes in user-definable models, equipped with interfaces that allow observation and interaction. Thus, the E-CELL system can be thought of as a framework for conducting experiments *in silico.*

The E-CELL system allows the user to utilize qualitative, quantitative or hybrid qualitative/quantitative models. This facilitates the incorporation of quantitative information where possible, while alleviating much of the limitations of qualitative models (deKleer 1977, Koile and Overton 1989).

The hypothetical cell we have modeled using E-CELL uptakes glucose from the culture medium using a phosphotransferase system, generates ATPs by catabolizing glucose to lactate by glycolysis and fermentation pathways, and exports lactate out of the cell. Since enzymes and other proteins are modeled to degrade spontaneously over time, they must be constantly synthesized in order for the cell to sustain "life". The protein synthesis is implemented by modeling the molecules necessary for transcription and translation, namely RNA polymerase, ribosomal subunits, rRNAs, tRNAs and tRNA ligases. The cell also uptakes glycerol and fatty acid and produces phosphatidyl glycerol for membrane structure using a phospholipid biosynthesis pathway.

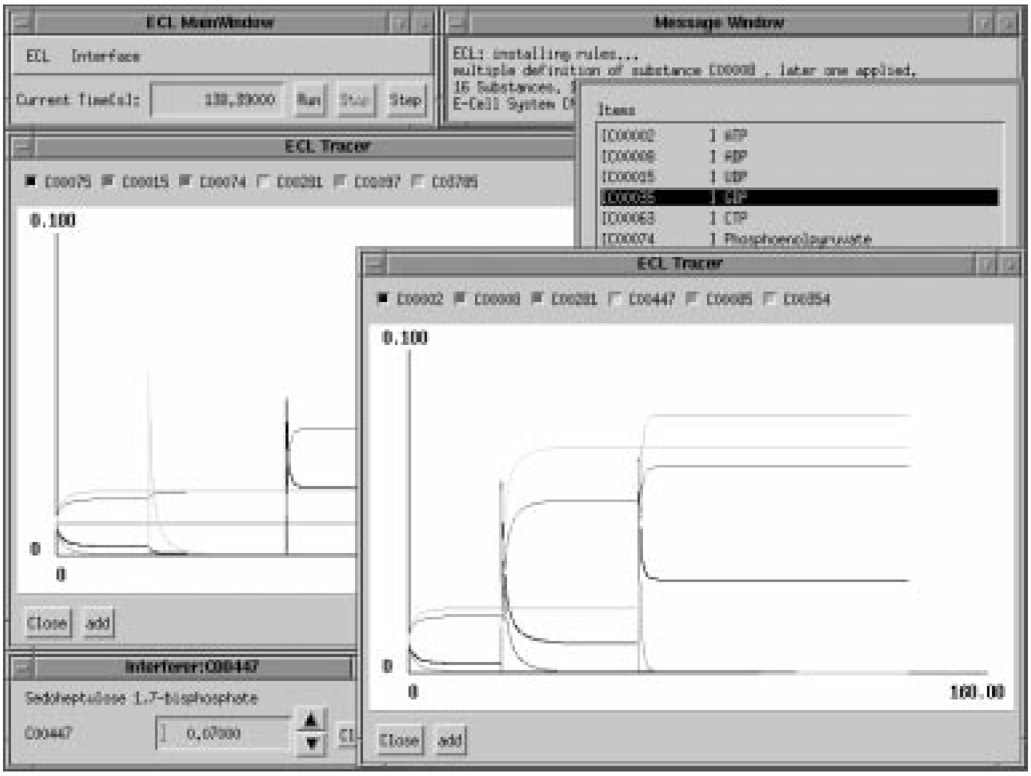
The genome of the cell consists of 127 genes including 20 tRNA genes and 2 rRNA genes. Out of the 127 genes, 120 are taken from *Mycoplasma genitalium*, the organism having the smallest known chromosome, whose complete 580kb genome sequence was determined at TIGR (Fraser *et al.* 1995). Since 7 genes which are essential for the model cell are not found in *Mycoplasma genitalium*, 2 are taken from *E. coli* and 5 are from no specific organism. We have been utilizing the wealth of information on metabolic pathways now available through knowledgebases such as EcoCyc (Karp *et al.* 1996) and KEGG(Kanehisa 1996). As an example, the pathway for phospholipid biosynthesis in the model cell is illustrated in figure 1.

2 Modeling a Cell

We define all possible objects in the cell and culture medium as $O1, O2, ..., On$. Objects which make up a cell are proteins, RNAs, DNAs, and small molecules. Multi-subunit complexes, such as RNA polymerase and the ribosome, are also defined as separate objects. Since we are dealing with a procaryotic species in the present work, small organelles such as mitochondria are not considered. We then represent a state of the cell as a list of real numbers $C1, C2, ..., Cn$, where Ci is the approximate concentration of Oi (or the number of molecules) in the cell, along with a few global values such as cell volume, pH and temperature. Operators to change a state into another state are called reactions, and are represented as $R1, R2, ..., Rn$. Each reaction consists of a list of substrates, a list of products, a list of catalysts, and a function to compute the rate of reaction from substrate and catalyst concentrations. Substrates, catalysts and products are defined as objects.

2.1 Object Definitions

Proteins Each protein/polypeptide is defined as an object. A separate object is defined for a modified version of the protein, such as a phosphorylated protein.


ECL MainWindow
ECL Interface
Current Time[s]: 130.29000
Run
Stop
Step
Message Window
ECL Tracer
■ C00075 ☒ C00015 ☒ C00074 □ C00281 ☒ C00097 □ C03785
0.100
0
0
Close
add
Items
Interferer:C00447
Sedoheptulose 1,7-bisphosphate
C00447
0.07000
ECL Tracer
■ C00002 ☒ C00008 ☒ C00281 □ C00447 ☒ C00085 □ C00354
0.100
0
0
160.00
Close
add

```
ECL: installing rules...
multiple definition of substance C00008 , later one applied.
16 Substances,
E-Cell System
```

```
IC00002      ] ATP
IC00008      ] ADP
IC00015      ] UDP
IC00035      ] GDP
IC00063      ] CTP
IC00074      ] Phosphoenolpyruvate
```

Figure 1: The phospholipid biosynthesis pathway.

DNAs Each gene is defined as an individual object. Positive and negative control regions (protein binding sites) of each gene are also defined as separate objects.

RNAs A messenger RNA for each gene is defined as an object. Other RNAs such as ribosomal RNAs and transfer RNAs are also defined as objects.

Small Molecules Other small molecules in the cell defined as objects include sugars, lipids, amino acids, ATPs, etc.

2.2 Reaction Rules

Enzyme reaction A typical reaction in a metabolic pathway is transformation of a small molecule into another molecule catalyzed by an enzyme which remains unchanged. An enzyme e transforming a substrate s into a product p can be represented as:

```
    [e]
  s ---> p
```

An object within square brackets is a catalyst. If there are two substrates, then:

```
          [e]
 s1 + s2 ---> p.
```

In this way, a reaction with any number of substrates and products can be dealt with. If the reaction requires energy, then it can be represented as:

```
         [e]
s + ATP ---> p + ADP.
```

Phosphorylation Protein phosphorylation can be represented as:

```
         [e]
s + ATP ---> ps + ADP
```

where *s* is an unphosphorylated protein, *ps* is a phosphorylated protein, and *e* is a protein kinase.

Dimer formation Two monomers forming a dimer can be represented as

```
2 x <--> d.
```

Similarly, reactions of two different subunits forming a complex can be represented as:

```
x + y <--> c.
```

Oligomer formation Trimer formation can be represented by a single rule as

```
3 x <--> t.
```

Alternatively, the following two rules would give a more precise model for trimer formation:

```
2 x <--> d
x + d <--> t.
```

Protein-DNA Interaction Protein-DNA interactions can be represented as:

```
p + d <--> pd
```

where *d* is a protein binding region of DNA, *p* is a DNA binding protein, and *pd* is a protein-DNA complex.

Translocation and transport Besides quantitative information (concentration) of each object, information concerning the location of an object is sometimes important. For example, transmembrane proteins are synthesized in the cytoplasm and then transported to the cell membrane in order to function in their assigned roles. If transportation is blocked, then they cannot function.

We can deal with those phenomena by defining the same objects at two different locations (x and y), as two different objects, Ox and Oy. Translocation of an object can then be represented by a simple reaction rule:

```
Ox --> Oy.
```

It is necessary to define in advance a finite set of locations or subregions of the cell.

The transport of nutrients from exterior medium into the cell can be modeled in this way.

Transcription and translation The synthesis of mRNAs can be modeled as a series of many detailed reactions, including: binding of RNA polymerase to the DNA promoter region, initiation of the RNA chain, chain elongation by addition of ribonucleoside triphosphates, and release of polymerase and completed RNA chain.

Gene regulation and cell cycle By modeling transcription and translation, along with DNA-protein interactions of regulatory factors, various gene regulation networks can be implemented.

Similarly, reactions related to the cell cycle such as DNA replication and cell division can be modeled as a series of a large number of primitive reactions, or a small number of abstracted reactions.

2.3 Reaction Rates

The rate (velocity) of each reaction can be defined by a mathematical equation which specifies the amount of change (the number of molecules) in a single time unit. One such equation for enzyme reaction is the *Michaelis-Menten equation*:

$$v = \frac{V_{max} * [s]}{[s] + K_m}$$

where $[s]$ is substrate concentration, V_{max} is maximal velocity, and K_m is the Michaelis constant.

Some reactions, such as dimer formation and DNA-protein binding, reach equilibrium within a single time unit. In those cases, dissociation constants can be used to directly compute the concentration of each molecule at the equilibrium. For a reaction such as

```
a + b <--> ab
```

the following equation holds at the equilibrium.

$$K_d = \frac{[a][b]}{[ab]}$$

where K_d is the dissociation constant of the reaction, $[a]$, $[b]$, and $[ab]$ are concentrations of a, b and ab, respectively.

3 Implementation of the E-CELL System

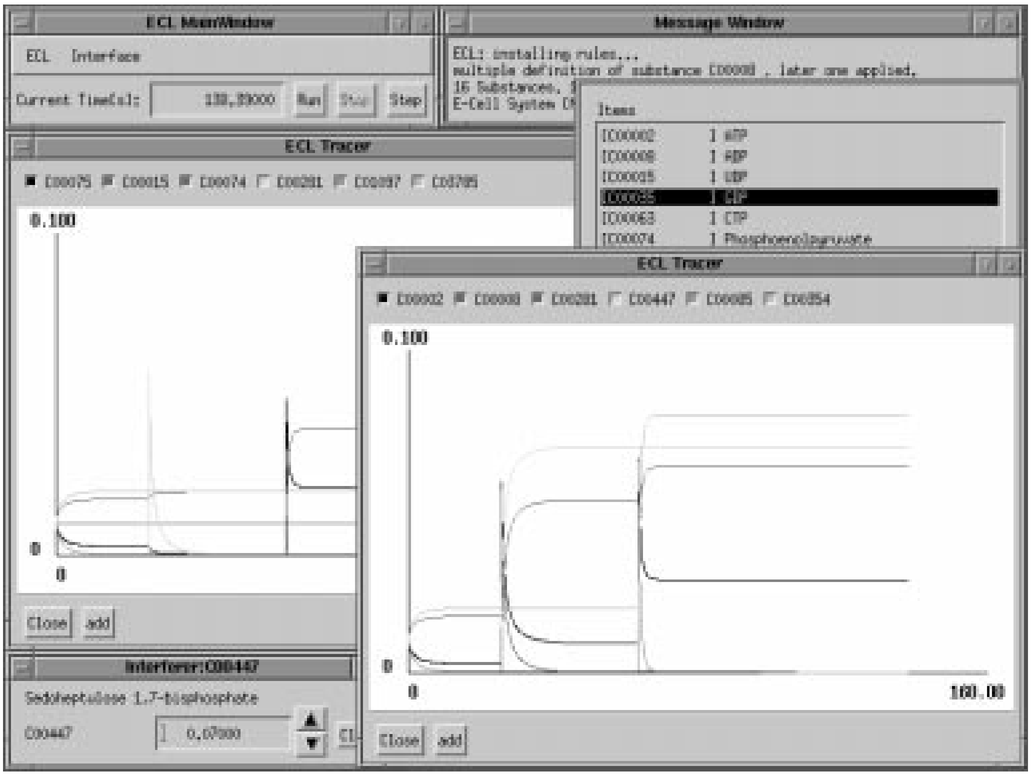
3.1 System Architecture and Data Representation

The E-CELL system is implemented as a rule based simulation system and is written in C++, an object oriented programming language. It allows the use of qualitative, quantitative or hybrid qualitative/quantitative models for simulation. The model consists of two lists, and is loaded at runtime. The *substance list* defines all objects which make up the cell and the culture medium. The *rule list* defines all of the reactions which can take place within the cell. The state of the cell at each time frame is expressed as a list of concentration values of all substances within the cell, along with global values for cell volume, pH and temperature. The simulator engine generates the next state in time by pseudo-parallel computation of all of the functions defined in the reaction rule list. In addition to using the sample qualitative models provided with the system, the user can create user-defined models by writing original substance and rule lists which may be either qualitative or quantitative. Graphical interfaces are provided to allow observation and interaction throughout the simulation process.

A substance can be a substrate, catalyst or product of a reaction. Typical substances include proteins, protein complexes, DNA(genes), RNA, and small molecules. The list of substance concentrations is updated with the new values computed by the simulator engine after each time unit.

Each rule in the rule list is called upon by the simulator engine to compute the concentration of each substance in the next time unit. The summation of all changes in concentrations of a substance resulting from each reaction occurring in a time unit are added to the concentrations in the present state to generate the next state of the cell.

In qualitative simulations where the rates of the defined reactions can serve as a measure for assessing time, it is not necessary to define an absolute value for the size of the time unit, Δt. However, the E-CELL system provides the option of defining an absolute value for Δt, allowing the user to assess quantitative time, when there is sufficient information to quantify the rate of the reactions.



ECL MainWindow
ECL Interface
Current Time[s]: 130.29000
Run
Stop
Step
Message Window
ECL Tracer
■ C00075 ☒ C00015 ☒ C00074 □ C00281 ☒ C00097 □ C03785
0.100
0
0
Close
add
Items
ECL Tracer
■ C00002 ☒ C00008 ☒ C00281 □ C00447 ☒ C00085 □ C00354
0.100
0
0
100.00
Close
add
Interferer:C00447
Sedoheptulose 1,7-bisphosphate
C00447
0.07000
▲
▼
Cl

```
ECL: installing rules...
multiple definition of substance C00008 , later one applied.
16 Substances, 1
E-Cell System (
```

```
IC00002      ] ATP
IC00008      ] ADP
IC00015      ] UDP
IC00035      ] GDP
IC00063      ] CTP
IC00074      ] Phosphoenolpyruvate
```

Figure 2: A screen dump of the E-CELL system

3.2 User Interfaces

The E-CELL system provides a graphical interface which allows the user to select substances of interest and observe dynamic changes in their concentration (figure 2). The interface is implemented as a window displaying a two dimensional plot in which line graphs represent changes in the concentration of selected substances. Each window can display up to 6 substances at once for comparison, and multiple windows may be opened when it is necessary to observe more than 6 substances at a time. The interface also allows the user to conduct experiments *in silico* by changing the concentration values at will during the simulation process.
The graphical user interface of the E-CELL System allows the user to observe the dynamic changes in concentrations of substances.

3.3 Future Enhancements

Our model cell's gene set of 127 genes is much smaller than the "minimal gene set" derived through sequence comparison by Musheginan and Koonin (1996). This is not surprising since our model lacks several important features present in all real living cells. It has no ability to proliferate; we are currently modeling cell growth, DNA replication, chromosome segregation and cell division.

Furthermore, the present cell model relies on unrealistically favorable environmental conditions. All of the amino acids and nucleotides must exist, and pH and osmolarity must be kept at physiologically stable levels at all times. The model also lacks cell structure proteins, which would be indispensable in any natural environment because of physical pressure.

To address these problems, we are currently modeling amino acid and nucleotide biosynthesis pathways. We also plan to model homeostasis of pH and osmolarity, and structure proteins for membrane structure and cytoskeleton.

4 Application to genome engineering

One of our ultimate goals is to model the real cell of *M. genitalium*, the organism having the smallest known chromosome. Because of the small number of genes (470 proteins, 37 RNAs), *M. genitalium* is a prime candidate for exhaustive functional (proteome) analysis. Because there are still many genes whose function is not yet known, it will probably be necessary to hypothesize putative proteins to complement missing metabolic functions, in order for the model cell to work *in silico*.

4.1 Metabolic requirements

The evaluation of metabolic requirements of the cell provides an excellent example of a potential application for the E-CELL. At present *M. genitalium* is grown in a complex medium containing several chemically undefined components including fetal bovine serum, and also extracts of yeast and beef. At TIGR experiments are in progress to produce a chemically defined medium which supports growth of *M. genitalium*. This problem could be attacked from a purely empirical point of view, however a more interesting approach is one that is informed by knowledge of the complete genome sequence. By combining knowledge of the metabolic enzymes present in the cell with information concerning protein transporters of metabolites across the cell membrane, it should be possible to evaluate whether a particular defined medium can support growth, by using the E-CELL model. The main difficulty in this approach is that identification of gene function solely on the basis of sequence is uncertain. Comparison of laboratory results with E-CELL predictions should help to overcome this difficulty. Agreement between the model and laboratory growth experiments will be evaluated for a large number of different chemically defined media. Differences between experimental observations and the E-CELL predictions will be used to refine the model. This could lead to the identification of new enzymes or transporters among genes with previously unassigned roles, or to the removal of a questionable role assignment based on a marginal level of sequence similarity.

4.2 Gene expression

Another area in which we plan to apply the E-CELL model concerns the control of gene expression. Gene expression patterns of *M. genitalium* are currently being determined at TIGR under a variety of growth conditions. We expect that these results will suggest specific mechanisms for control of transcript levels which can be modeled by rules in the E-CELL system. We will conduct parallel experiments in the laboratory and *in silico* with the E-CELL system; given an appropriate model of the cell, we can change initial values of ingredients of the culture medium and observe increases and decreases of messenger RNA levels. The results of those *in silico* experiments should be consistent with results of biological and biochemical experiments. The computer model will then be refined as necessary.

4.3 Minimal gene set

We expect that the E-CELL system will be useful in defining the minimal set of genes required for a self replicating cell under a specific set of laboratory conditions. At TIGR work is underway to identify the genes of *M. genitalium* which are non-essential, by gene disruption experiments using transposons. If the E-CELL model is sufficiently detailed and accurate, then these gene disruption experiments can be modeled *in silico* to predict a minimal gene set. The laboratory experiments will lead to the prediction of a reduced gene set which should be a close approximation to the truly minimal Mycoplasma genome. Alternative predictions of a minimal gene set can also be proposed on theoretical grounds, or by deducing a core set of genes conserved between *M. genitalium* and other microbial genomes. The E-CELL system should be useful in modeling cells based on these alternative proposals for a minimal cellular genome.

We expect that a combination of laboratory experiments and *in silico* modeling using the E-CELL system, will lead to a more reliable prediction of the minimal gene complement for a self-replicating cell than could be obtained by either method alone.

5 Concluding Remarks

We have constructed a hypothetical cell using the first version of E-CELL, and have developed hundreds of reaction rules for a partial set of metabolic pathways of *M. genitalium*, including glycolysis, lactate fermentation, glycose uptake, glycerol and fatty acid uptake, phospholipid biosynthesis, gene transcription, protein synthesis, polymerase and ribosome assembly, protein degradation and mRNA degradation.

Its application to genome engineering has just begun. The approaches to defining a minimal gene set, described in section 4, are testable in principle. At TIGR a longer term goal of this work is the engineering of the genome to produce living cells with substantially reduced genomes. This will allow us to test proposals for minimal gene sets directly. It will be interesting to compare real cells so created with their computer models. Comparison of the models with the results of laboratory experiments will allow further refinement of the computer models. This in turn will lead to a better understanding of the biological and biochemical results, and hence a better understanding of the essential requirements of a minimal living cell. In this way we can think of the E-CELL system as the first step toward a tool for the computer assisted design of novel cells.

Acknowledgments

We would like to thank Peter Karp and Doug Brutlag for useful comments on the early draft of this paper. Masanori Arita kindly helped us in reviewing literature on biological simulations. Many of the ideas presented in this paper were inspired by discussions with other members of Tomita Lab. at Keio University, including Junko Shinada, Keiko Miura, Hisako Nakano and Daisuke Kamiyoshikawa. We also thank for useful comments Scott Peterson and Karen Ketchum of The Institute for Genomic Research, as well as Nobuyoshi Shimizu and his colleagues at Department of Molecular Biology, Keio University.

This work was supported by Eizai Research Institute and also in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Science' from The Ministry of Education, Science, Sports and Culture of Japan.

References

Arita, M., Hagiya, M., and Shiratori, T. 1994. GEISHA SYSTEM: An Environment for Simulating Protein Interaction. In *Proceedings, Genome Informatics Workshop 1994*. Bunkyou-ku, Tokyo: Universal Academy Press. 81–89.

Bray, D., Bourret, R.B., and Simon, M.I. 1993. Computer Simulation of the Phosphorylation Cascade Controlling Bacterial Chemotaxis. *Molecular Biology of the Cell* 4:469–482.

deKleer, J. 1977. Multiple Representations of Knowledge in a Mechanics Problem-Solver. In *Proceedings of IJCAI-77* 299–304.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, III, C.A., and Venter, J.C. 1995. The Minimal Gene Complement of *Mycoplasma genitalium. Science* 270:397–403.

Gaasterland, T., and Selkov, E. 1995. Reconstruction of Metabolic Networks Using Incomplete Information. In *Proceedings, Third International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, CA: AAAI Press. 127-135.

Galper, A., and Brutlag, D. 1993. Computational Simulations of Biological Systems. In Smith, D., ed., *Biocomputing: Informatics and Genome Projects.* San Diego, CA: Academic Press.

Kanehisa, M. 1996. Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan* 59:34–38.

Karp, P.D. 1993. A Qualitative Biochemistry and Its Application to the Regulation of the Tryptophan Operon. In Hunter, L., ed., *Artificial Intelligence and Molecular Biology.* Menlo Park, CA: AAAI Press. 289–324.

Karp, P.D., Riley, M., Paley, S.M., and Pelligrini-Toole, A. 1996. EcoCyc: Encyclopedia of *E. coli* Genes and Metabolism. *Nucleic Acids Research* 24(1):32–40.

Kuipers, B. 1985. The Limits of Qualitative Simulation. In *Proceedings of IJCAI-85* 128–136.

Kuipers, B. 1986. Qualitative Simulation. *Artificial Intelligence* 29:289–338.

Koile, K., and Overton, G.C. 1989. A Qualitative Model for Gene Expression. In *Proceedings of the 1989 Summer Computer Simulation Conference* 415–421.

McAdams, H.H., and Shapiro, L. 1995. Circuit Simulation of Genetic Networks. *Science* 269:650–656.

Mushegian, A.R. and Koonin, E.V. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes; *Proc Natl Acad Sci USA* 1996 Sep 17;93(19):10268–10273

Novak, B., and Tyson, J.T. 1995. Quantitative Analysis of a Molecular Model of Mitotic Control in Fission Yeast. *Journal of theoretical Biology* 173:283–305.

Tyson, J.T. 1991. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Science USA* 88:7328–7332.