IS 1-2

Computer Models of Bacterial Cells To Integrate Genomic Detail with Cell Physiology

Michael L. Shuler

Department of Biomedical Engineering and School of Chemical and Biomolecular Engineering
Cornell University, Ithaca, NY 14853

Abstract

While genomics (the set of experimental and computational tools that allows the blueprints of life to be read) opens the doors to a more rational approach to the design and use of living cells to bring about desirable chemical transformations, genomics is, by itself, insufficient. We need tools that allow us to relate genomic and molecular information to cellular physiology and then to the response of a population of cells. We propose the development of hybrid computer cellular models. In such models genomics and chemical detail for a cellular subsystem (e.g. pathogenesis) is embedded in a coarse-grain cell model. Such a construct allows the quantitative and explicit linkage of genomic detail to cell physiology to the extracellular environment. To illustrate the principles involved we are constructing a model for a minimal cell. A minimal cell is a bacterial cell with the fewest number of genes necessary to sustain life as a free living microbe.

I. Overall Vision

Our vision is to develop a framework using computer and experimental models to quantitatively and explicitly link genomic/molecular level insights to the physiology of whole organisms. We are particularly interested in a new approach: "hybrid" models. We believe this approach will be our best opportunity to build models of complex organisms that are both realistic, tractable, and useful. Hybrid models start with a genomically/molecularly detailed model of a subsystem of interest. We then insert this detailed submodel into a cellular model with pseudo-molecular detail (i.e. "coarse grain" models), and then the cellular model into a system model. For bacteria the system is a population of cells in a bioreactor or a natural environment such as a biofilm.

II. Building Bacterial Cell Models: The "Minimal Cell" as a Foundation

While our overall aim is to build a model for real organisms using this hybrid approach, we believe a

fully-detailed model of a hypothetical "minimal cell" is critical to both testing fundamental concepts about microbial physiology and to building the methodology necessary to construct hybrid models of real cells quickly and efficiently.

A minimal cell is a hypothetical cell defined by the essential functions required for life [Castellanos *et al.*, 2004]. The model seeks to identify a minimum number of genes necessary and sufficient for the cell to divide and grow continuously in a rich environment with preformed nutrients and relatively constant temperature and pH. The model, which contains kinetic and thermodynamic constraints as well as stoichiometric constraints, can be used as a tool to identify the organizing principles which relate the dynamic non-linear functioning of the cell to the static linear sequence information of the genome.

Systems biology [Kitano, 2002] investigates the "behavior and relationships of all of the elements in a particular biological system while it is functioning". The emphasis in our project is on modeling the complete functionality of a cell and its explicit response to perturbations in its environment [Browning & Shuler, 2001] and to build hybrid models starting with a genomically/molecularly detailed model of a subsystem of interest, inserting that submodel into a cellular model with pseudo-molecular detail (i.e. "coarse grain" models), and then the cellular model into a system model. Our attempt to generate "complete" and hybrid models that predict time-dependent responses of a cell differentiates this project from others.

Many investigators have made significant contributions to our understanding of bacterial metabolism, particularly central carbon metabolism. The studies have taken advantage of detailed genomic information and some models are based primarily on stoichiometry and techniques involving flux balance analysis, metabolic control theory, and mathematical techniques for optimization eg. [Burgard *et al.*, 2001; Edwards & Palsson, 2000]. Since these models are intrinsically static, they have limited ability to predict aspects of cell regulation and dynamic response (although by the addition of constraints, such as uptake rates of a nutrient, these models provide some insight into the dynamic state that can be achieved). Others have proposed methodology to incorporate more directly dynamic (kinetic) information into models of central metabolism eg. [Chassagnole *et al.*, 2002]. Others have attempted to model whole cells [Tomita, 2001], but those models neglect non-metabolic aspects of cell growth (eg. control of chromosome replication or spatial issues associated with position of septa). These approaches are "incomplete" descriptions.

Incomplete descriptions may lead to conclusions that are inaccurate as there is an implicit assumption in such studies. The assumption is most easily illustrated by considering the metabolic flux analysis of an isolated pathway. As shown by Schlosser and Bailey [1990], such analysis is correct only if the output of the pathway cannot influence any input into the pathway. Any "cell" model that is "incomplete" assumes that no output of the model either directly or indirectly can influence any input or state within the model within the timescale of interest.

A "complete" model of a real organism is a daunting task [McAdams & Shapiro, 2003; Bailey, 2001], but we believe our goal of a hypothetical minimal cell model is both achievable and will provide insights into biological questions of immediate importance. McAdams & Shapiro [2003] write "··· to develop "whole-cell" models··· major, perhaps insurmountable, difficulties must be overcome··· Problems include

lack of quantitative data on molecular concentrations and kinetic parameters as well as only piecemeal characterization of the cell's regulatory circuitry". While we agree with the problems they identify, we are optimistic and we agree with Alon [2003] that a "reverse engineering" approach that takes advantage of the natural characteristics of biological systems: modularity, robustness, and use of recurring circuits elements can succeed. This is the basis of the approach we will describe. Just as an engineer will design an airplane based on functional constrains and make use of prior designs we will design a cell (using guidance from existing cells) to perform the essential tasks necessary to survive indefinitely and translate that design into a hypothetical genome, just as the airplane design is translated into blueprints and construction documents. We describe in this paper methodology to rapidly estimate a credible set of kinetic parameters overcoming one of the key limitations suggested by McAdams & Shapiro [2003].

Although a minimal cell is hypothetical, the applicability of such a detailed model is enormous. The proposal model can lead to a better understanding of the behavior of chemoheterotrophic bacteria. While a minimal cell model will suggest the essential components of regulation, a deeper insight into the logic of cell regulation can be achieved in future studies by perturbing the environment with large changes until the model cell fails ("dies") and then finding regulatory approaches that allow survival. In essence, we wish to understand how selective pressure relates to microbial evolution. A more complete understanding of essential cellular structure and regulation is important for bioprocess engineers to metabolically engineer cells for production of desirable metabolites and/or to design improved operating strategies for bioprocesses.

Additionally, we can use the minimal cell as a basis to learn to build hybrid models of real cells. The key requirements for such hybrid modules is "modularity" and the ability to construct species specific coarse grain models rapidly. Using the minimal cell we demonstrate modularity and also techniques to evaluate kinetic parameters rapidly.

III. The Minimal Cell Concept

The minimal cell concept can be traced back to the 1950's when Harold Morowitz and colleagues began to seek the smallest, autonomous, self-replicating entity. They correctly identified *Mycoplasma* as the best living example of a minimal organism. Morowitz proposed that it should be possible to build a computer model of such a complete cell. He wrote that with *Mycoplasma* "Their existence with all the properties of life says the 'logic of life' is finite, relatively simple, and subject to full exploration" [Morowitz, 1984].

By the mid-1990's the issue of a minimal gene set began to attract increased attention. In 1995, Itaya [1995] used random gene knockouts in *Bacillus subtilis* to estimate that 254 genes are essential. Mushegian and Koonin [1996] compared the full genome sequences of *Haemophilus influenzae* and *Mycoplasma genitalium* and proposed a set of about 250 genes as a minimal gene set. A large project was begun shortly afterwards to create a minimal cell. The ultimate goal was the experimental construction of an artificial minimal genome. Hutchinson *et al* [1999] used transposon knockouts of *M. genitalium* to predict that about 265 to 350 genes (about 100 with unknown function) were essential. The so named E-CELL model

was created as a component of this minimal cell project [Tomita, 2001]. However, the project was abandoned [Peterson & Fraser, 2001]. A very similar project under Venter's leadership has been restarted to develop a synthetic chromosome as the first step toward making a self-replicating organism [Zimmer, 2003]. The technical difficulty of fitting the minimal genome with a working cell structure is acknowledged as a major challenge [Zimmer, 2003].

A minimal gene set derived by comparative genomics approach is likely to be an underestimate (due to non-orthologous gene displacement). Additionally an experimental approach can overestimate the minimal set substantially (genome scale knockouts could identify genes as essential even when the deletion only slows growth [Koonin, 2003]). Computer simulations offer an alternative to comparative genomics and experiments to identify a minimal gene set.

A computer model approach should result in a set of minimal functions that will correspond to real genes, which exist in nature. The model we are developing focuses on essential functions while finding examples of gene products that can perform those functions. While the set of minimal genes we postulate may change (eg. if a new multifunctional protein is found), we believe we can find a set of essential functions. Further, the technical difficulties associated with generating an experimental minimal cell and the ambiguities in interpretation of comparative genomic data argue for the establishment of a theoretical computer model of a minimal cell. This model must be explicit about minimal functions and include a realistic set of proteins to accomplish these functions. This is, we believe, the most practical route to a minimal cell.

IV. The Cornell Single Cell

We have previously developed a "complete" cell model of *E. coli* that contains all of the functional elements for the cell to grow, divide, and respond to a wide variety of environmental perturbations. All chemical species are included, but lumped into pseudochemical groups. This "coarse-grain" model serves as the basis for our efforts to build a minimal cell model. Basically, the *E.coli* model is a good summary of the functionality required for a minimal cell, but it does not capture explicitly the physical chemistry that supports those functions. We described our first mathematical model of a single *E. coli* cell in 1979 [Shuler *et al.*, 1979]; at that time, it was the only model of an individual cell that did not include artificially-imposed constraints on aspects such as mode of growth, timing of cell division (eg. growth rate), and cell size. Also, it was unique in its ability to respond explicitly to concentrations of nutrients in the environment [Bailey, 1998]. This base model [Domach *et al.*, 1984] has been embellished with additional biological details to allow prediction of a wide-range of responses to environmental and genetic manipulations. The initial model included only 18 pseudochemical species that represented large groups of related chemical species; Fig. 1 lists the components and graphically depicts the relationships between components. The mathematical description of cellular function, which is the core of the model, is based on time-variant mass balances for each component. Each mass balance takes into account the component's synthesis (as a function of availability of precursors and energy, relevant enzymes), utilization, and

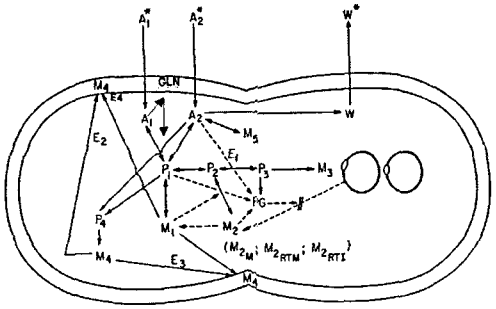

Fig. 1 An idealized sketch of the model of E. coli B/rA growing in a glucose-ammonium salts medium with glucose or ammonia as the limiting nutrient. At the time shown, the cell has just completed a round of DNA replication and initiated cross-wall formation and a new round of DNA replication. Solid lines indicate the flow of material, while dashed lines indicate flow of information.

A_1= ammonium ion
A_2= glucose (and associated compounds)
W= waste products (CO_2, H_2O, and acetate)
P_1= amino acids
PG= *ppGpp*
P_2= ribonucleotides
P_3= deoxyribonucleotides
P_4= cell envelope precursors
M_1= protein (both cytoplasmic and envelope)
M_{2RTI}= immature "stable" RNA
*— the material is present in the external environment
M_{2RTM}=mature *r*-RNA and *t*-RNA
M_{2M}=messenger RNA
M_3=DNA
M_4=non-protein part of cell envelope
M_5=glycogen
E_2, E_3=molecules involved in cross- wall formation and cell envelope synthesis
GLN=glutamine
E_1=enzyme in the conversion of P2 to P3
E_4 = glutamine synthetase

degradation. Stoichiometric coefficients for relating components through mass balances were derived primarily from published research, and in some cases, from our own experiments. It is important to note that the model was **NOT** developed by using adjustable parameters to fit model predictions to experimental results, nor did the stoichiometric mass balances assume a steady state (i.e. the amount of each component was allowed to vary with time). Despite the simplifications that were made in describing cell composition and relationships, the model can accurately predict changes in cell composition, size, and shape, and the timing of chromosome synthesis as a function of changes in external glucose and ammonium concentration [Domach *et al.*, 1984].

The dynamic mass balances are solved by forward integration from an initial condition (both extracellular concentrations of nutrients and intracellular estimates of all pseudochemical species) using a Runge-Kutta type technique. A variable step size is used as the equations become very stiff when chromosome replication is initiated (i.e. large flux through a small pool P_3). Conditional statements are included for fork position, gene dosage, potential for initiation of chromosome replication, cell shape and position and completeness of cross-wall formation.

Other biochemical details have been added in subsequent studies that allow the study of the effects of amino acid supplementation [Shu & Shuler, 1991] and of competition between recombinant mRNA and ribosomal mRNA in the context of high translational activity [Laffend & Shuler, 1994a]. The model has also been utilized extensively to improve the use of plasmids for recombinant protein production; (eg.[Kim & Shuler, 1990, 1991; Laffend & Shuler, 1994b]). The calculations have proven to be quite robust and results are reproducible.

We have used the structure of the *E.coli* model to build a coarse-grain minimal cell model that is a generalized model of chemoheterotropic bacteria. By using dimensionless concentrations and growth rates, we demonstrate that a generalized chemoheterotroph can be constructed that is consistent with a wide range of experimental data (Browning & Shuler, 2001). The robustness of a molecularly realistic mechanism to control initiation of chromosome replication has been tested and can be built into the coarse grain cell model framework (Browning et al., 2004).The coarse-grain minimal cell model is "complete" in terms of function and is "modular". By modular we mean that we can "delump" a pseudochemical species into individual components while still maintaining the essential connectivity to other functions in the cell [Castellanos *et al.*, 2004]. This allows us to add detail in parallel efforts on different "modules" and then have confidence that they can be recombined into a functional and functioning whole. This strategy is our basic approach to constructing a genomically and chemically detailed minimal cell model.

V. Demonstration of Modularity of Basic Minimal Cell Model

We have tested the hypothesis that it is not the exact values of parameters in the model that determine function, but that the values relative to one and another is critical (Browning & Shuler, 2001). We tested this hypothesis by varying all kinetic rates by a scaling factor (or kinetic ratio). The growth rate scales directly with the kinetic ratio over about two orders of magnitude. At low values of growth rate, membrane energization becomes important and linearity is lost. Cell composition (eg. protein/cell, RNA/cell, etc.) remains constant for a wide range of kinetic ratios. Further, relative growth rate changes for models with different kinetic ratios is essentially the same for a wide variety of perturbations to cell function (which also confirms the computational robustness of the model). Also the general physiological behavior of a variety of common bacteria (based on experiment) scales with a dimensionless growth rate, suggesting that the lessons from a hypothetical cell model will be broadly applicable to chemoheterotrophic bacteria.

Our minimal cell model for nucleotide metabolism [Castellanos *et al.*, 2004] confirms the concept of modularity by testing a functional nucleotide subsystem model with significantly fewer gene-encoded functions (12) than estimated previously. In the *M. genitalium* genome sequence [Fraser, 1995] 25 genes can be associated with nucleotide transport and metabolism. Mushegian and Koonin [1996] estimated that the minimal gene set includes 23 genes for nucleotide metabolism. Hutchinson *et al* [1999] concluded that only 18 genes were essential for transport and metabolism of ribonucleosides. Kobayashi *et al* [2003] include 10 genes in the nucleotide category but they point out that due to single gene inactivation, the number of genes in their minimal gene set is likely to be underestimated. Their list of essential genes appears incomplete (supplemented information to Kobayashi *et al.*, 2003). Our minimal cell pathway with 11 functions (12 genes) permitting growth from preformed ribonucleosides precursors is the most efficient (fewest genes) of any study with a complete pathway. Our minimal cell pyrimidine nucleotide biosynthesis pathway includes: uracil phosphoribosyltransferase, cytidylate kinase, ribonucleotide reductase, thymidylate synthetase, deoxyuridine triphosphatase, adenylate kinase, and thymidylate kinase. Our minimal cell purine nucleotide biosynthesis pathway includes: adenine phosphoribosyltransferase, guanine

phosphoribosyltransferase, adenylate kinase, ribonucleotide reductase, and guanylate kinase.

A key achievement is the demonstration that we can delump a module, insert genomic and chemical details, and maintain a fully functional complete cell model. In essence, we have a hybrid "coarse-grain" whole cell model in which a genomically detailed model is embedded within the coarse grain minimal cell model. Thus, we have established the concepts of "modularity" and "connectivity".

VI. Demonstration of Approach to Rapid Estimation of Kinetic Parameters

The construction of the coarse-grain model with detailed nucleotide modules required a tedious search through the literature to estimate parameters and requires considerable biological insight by the modeler. To build such hybrid modules for real cells, it would be helpful to have a methodology to take a general coarse-grain structure and estimate parameters using growth data that can be obtained quickly.

Sethna's research group has developed a generalizable approach to extract falsifiable predictions from biological models using statistical mechanical type models [Brown & Sethna, 2003; Brown *et al.*, 2002]. The method involves using a specific cost function using all experimental data along with error values, and a corresponding model output evaluated with a parameter set p. For a minimal cell a "data" set is the required "design performance" of the minimal cell model and is obtained from the generalized behavior of chemoheterotrophic bacteria. This cost function is optimized to find a "*best*" parameter fit with the lowest cost. This parameter set becomes the starting point to generate an ensemble of parameters using the Monte Carlo method.

We have validated this approach with our base coarse grain model (*E. coli* model), for which we have significant experimental data. The optimization routine has provided a parameter set that allows an excellent match of model predictions to the experimentally observed behavior of *E.coli* in glucose-limited chemostats. Additionally we have found that most of the parameters previously estimated from mass balances, stoichiometry and the literature and from applying the Sethna approach to a series of chemostat data gave nearly the same value ($\pm$ 10 %) for almost all the parameters. Only 6 parameters (68 parameters were studied) differed more than 10% from original parameter estimate with a maximum difference of 25%.

VII. Application to Microbes

Consider how we might approach modeling a microbial pathogen. The first step requires construction of the coarse-grain model. For microbes that can be cultured we believe we can rapidly estimate all of the essential parameters. Relatively high throughput chemostat systems using mini reactors are available. Using a variety of steady state flow rates, nutrient levels and types, and flow perturbations as inputs and measuring cell composition, size, residual nutrient levels, and by-product levels it is possible to form a significant database. Microarray data from perturbation experiments would be useful but not essential. We can then apply the approach described in Section VI to estimate the basic parameters. Here we assume that

the general chemoheterotrophic behavior applies to the microbe of interest.

The second step is to abstract a proposed mechanism, say of pathogensis, into a detailed model incorporating all suspected genes and known regulatory features. Of particular importance will be the connection of this mechanism to extracellular cues and to the physiologic state of the cell.

The third stage is to place the cell in the context of environments of interest. For validation purposes experiments with low density cultures can be used to compare to predictions of response to predetermined perturbations. A more sophisticated set of experiments would be to examine microbe to microbe interaction (e.g. quorum factors) and microbe-tissue interactions. Proposed mechanisms of interaction would have to be placed in the model to make experimental to model prediction comparisons.

Acknowledgements

This work was supported in part by NSF, DOE/GTL, and USDA.

References

1. Alon, U. 2003 *Biological networks: The tinkerer as an engineer.* Science. 301:1866-1867.
2. Bailey, J.E. 1998. *Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities.* Biotechnol Prog. 14:8-20.
3. Bailey, J.E. 2001. *Complex biology with no parameters.* Nature Biotechnol. 19:503-514.
4. Brown, K.S. and J.P. Sethna. 2003. *Statistical mechanical approaches to models with many poorly known parameters.* Physical Review E. 68 021904.
5. Brown, K.S., et al. 2002. *Integrated approaches to signal transduction: PC12 differentiation.* Biophysical Journal. 82:219A.
6. Browning, S.T. and M.L. Shuler. 2001. *Towards the development of a minimal cell model by generalization of a model of Escherichia coli: Use of dimensionless rate parameters.* Biotechnol Bioeng. 76:187-192.
7. Browning, S.T., M. Castellanos, and M.L. Shuler. 2004. *Robust control of initiation of prokaryotic chromosome replication: essential considerations for a minimal cell.* Biotechnol Bioeng. *88*(5):575-584.
8. Burgard, A.P., S. Vaidyaraman, and C.D. Maranas. 2001. *Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments.* Biotechnol Prog. 17:791-797.
9. Castellanos, M., D.B. Wilson, and M.L. Shuler. 2004. *A modular minimal cell model: Purine and pyrimidine transport and metabolism.* PNAS(USA). 101(17):6681-6686.
10. Chassagnole, C., et al. 2002. *Dynamic modeling of the central carbon metabolism of Escherichia coli.* Biotechnol Bioeng. 79:53-73.
11. Domach, M.M., et al. 1984. *Computer-Model for Glucose-Limited Growth of a Single Cell of Escherichia-Coli B/R-A.* Biotechnol Bioeng. 26:203-216.
12. Edwards, J.S. and B.O. Palsson. 2000. *The Escherichia coli MG1655 in silico metabolic genotype: Its*

definition, characteristics, and capabilities. PNAS(USA). 97:5528-5533.

13. Hutchison, C.A., et al. 1999. *Global transposon mutagenesis and a minimal Mycoplasma genome.* Science. 286:2165-2169.
14. Itaya, M. 1995. *An estimation of minimal genome size required for life.* FEBS Letters. 362:257-260.
15. Kim, B.G. and M.L. Shuler. 1990. *A structured, segregated model for genetically modified Escherichia-coli-cells and its use for prediction of plasmid stability.* Biotechnology and Bioengineering, 1990. 36(6): p. 581-592.
16. Kim, B.G. and M.L. Shuler. 1991. *Kinetic-analysis of the effects of plasmid multimerization on segregational instability of ColE1 type plasmids in Escherichia-coli B/R.* Biotechnol Bioeng. **37**:1076-1086.
17. Kitano, H. 2002. *Systems biology: a brief overview.* Science, 2002. 295(5560): p. 1662-1674.
18. Kobayashi, K., et al. 2003. *Essential Bacillus subtilis genes.* PNAS(USA). 100:4678-4683.
19. Koonin, E.V., *How many genes can make a cell: the minimal-gene-set concept.* 2000. 1: p. 99-116.
20. Koonin, E.V. 2003. *Comparative genomics, minimal gene-sets and the last universal common ancestor.* Nature Reviews Microbiology. 1:127-136.
21. Laffend, L. and M.L. Shuler. 1994a. *Ribosomal-protein limitations in Escherichia-coli under conditions of high translational activity.* Biotechnol Bioeng. 43:388-398.
22. Laffend, L. and M.L. Shuler. 1994b. *Structured model of genetic-control via the lac promoter in Escherichia-coli.* Biotechnol Bioeng, 43:399-410.
23. Morowitz, H.J. 1984. *The completeness of molecular biology.* Israel J Medical Sci. 20:750-753.
24. Mushegian, A.R. and E.V. Koonin. 1996. *A minimal gene set for cellular life derived by comparison of complete bacterial genomes.* PNAS(USA). 93:10268-10273.
25. Peterson, S.N. and C.M. Fraser. 2001. *The complexity of simplicity.* Genome Biology. 2:1-8.
26. Schlosser, P.M. and J.E. Bailey. 1990. *An integrated modeling-experimental strategy for the analysis of metabolic pathways.* Mathematical Biosc. 100:87-114.
27. Shu, J. and M.L. Shuler. 1991. *Prediction of effects of amino-acid supplementation on growth of Escherichia-coli B/R.* Biotechnol Bioeng. 37:708-715.
28. Shuler, M.L., L. S., and D. CC., *A mathematical model for the growth of a single bacterial cell.* Ann. NY Acad. Sci., 1979. 326: p. 35-55.
29. Tomita, M. 2001. *Whole-cell simulation: a grand challenge of the 21st century.* Trends in Biotechnology. 19:205-210.
30. Zimmer, C. 2003. *Genomics. Tinker, tailor: Can Venter stitch together a genome from scratch?* Science. 299:1006-1007.