Chapter 20

Modeling a Minimal Cell*

Michael L. Shuler, Patricia Foley, and Jordan Atlas

Abstract

One important aim of synthetic biology is to develop a self-replicating biological system capable of performing useful tasks. A mathematical model of a synthetic organism would greatly enhance its value by providing a platform in which proposed modifications to the system could be rapidly prototyped and tested. Such a platform would allow the explicit connection of genomic sequence information to physiological predictions. As an initial step toward this aim, a minimal cell model (MCM) has been formulated. The MCM is defined as a model of a hypothetical cell with the minimum number of genes necessary to grow and divide in an optimally supportive culture environment. It is chemically detailed in terms of genes and gene products, as well as physiologically complete in terms of bacterial cell processes (e.g., DNA replication and cell division). A mathematical framework originally developed for modeling *Escherichia coli* has been used to build the platform MCM. A MCM with 241 product-coding genes (those which produce protein or stable RNA products) is presented. This gene set is genomically complete in that it codes for all the functions that a minimal chemoheterotrophic bacterium would require for sustained growth and division. With this model, the hypotheses behind a minimal gene set can be tested using a chemically detailed, dynamic, whole-cell modeling approach. Furthermore, the MCM can simulate the behavior of a whole cell that depends on the cell's (1) metabolic rates and chemical state, (2) genome in terms of expression of various genes, (3) environment both in terms of direct nutrient starvation and competitive inhibition leading to starvation, and (4) genomic sequence in terms of the chromosomal locations of genes.

Key words: Minimal cell, Systems biology, Synthetic biology, Cell model, Minimal gene set, Dynamic cell model, Bacterial cell model, Differential algebraic equation cell model

1. Introduction

1.1. Synthetic Cells

Efforts to develop synthetic systems generally fall into two categories: (1) the development of biological components and systems that can be combined to produce a preprogrammed outcome in a biological system and (2) the generation of a complete, self-replicating biological system capable of performing some useful

*Portions of this chapter are reused, with permission, from J. C. Atlas, "Simulation of a Whole-Cell With the Minimum Number of Genes Necessary for Sustained Replication." PhD thesis, Cornell University, 2010.

Ali Navid (ed.), *Microbial Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 881,
DOI 10.1007/978-1-61779-827-6_20, © Springer Science+Business Media, LLC 2012

task. Several recent review articles discuss efforts toward and the potential impacts of the former goal (1–4). Alternatively, we and others aim to address the latter goal—to design, characterize, and ultimately synthesize a living cell *de novo*.

The concept of a synthetic cell is over a 100 years old (5). Even then, Loeb described experimental abiogenesis as “the goal of biology.” However, as the field of synthetic biology grows, it is clear that what constitutes a “synthetic cell” is in many ways defined by the goals of the researchers involved in producing the cell. Nonetheless, some basic criteria have been defined. A “living” cell must be capable of metabolic homeostasis, cellular reproduction, and Darwinian evolution (6, 7). These properties can be achieved by various strategies, and attempts to construct a synthetic cell can be classified as either bottom-up or top-down. Bottom-up approaches generally seek insights about the origin of life, and therefore utilize basic chemical compounds that could plausibly self-assemble into biological entities (8–10). Conversely, top-down approaches aim to utilize modern cellular machinery, including DNA genomes, transcription and translation machinery, and phospholipid bilayers, in the design of a synthetic cell (11, 12). As engineers, we envision an ideal synthetic cell as a platform system that is chemically, physically, and importantly, mathematically defined, and will facilitate future optimization of the organism for a variety of predefined tasks.

A key factor that drives the development of a synthetic organism is the value of a completely defined system. Natural microbes have been utilized for decades as the workhorses for bioprocessing facilities, and one may wonder whether the development of a synthetic organism will have advantages beyond genetically manipulated natural organisms. Because a synthetic organism would be a completely defined system, it would be unique in that a complete mathematical model of the cell could be developed. An *in silico* counterpart of a synthetic cell would allow for quick and inexpensive optimization of new properties that are to be incorporated into the cell to achieve some predefined goal (e.g., to incorporate a drug synthesis pathway into the cell’s metabolic pathway). As a first step toward development of a synthetic cell, attempts are underway to define a minimal cell, in both physical and computational terms.

1.2. Minimal Gene Sets and Minimal Cells

The minimal cell concept can be traced back to the 1950s when Harold Morowitz and colleagues began to seek the smallest, autonomous, self-replicating entity (13). Because the genetic material of an organism defines its characteristics, what most succinctly defines a minimal cell is the makeup of its chromosome. Based on Morowitz’s original concept, a minimal cell is defined as one possessing a minimal gene set, defined as the minimum number of genes that are both necessary and sufficient to promote sustained growth and division of a bacterial cell in some optimally supportive culture environment.

Establishing a minimal gene set, or minimal gene sets, is an important step in synthetic biology. Various comparative genomic, genetic, and biochemical approaches have been used to estimate hypothetical minimal gene sets. However, a reductionist approach that only considers each gene in the minimal gene set independently will be insufficient. It is necessary to evaluate how these cell systems functionally integrate (14).

1.2.1. Synthesis of Minimal Cells

The J. Craig Venter Institute has been actively pursuing the goal of synthesizing a cell using a top-down approach. Toward this end, they successfully transplanted a complete *Mycoplasma mycoides* chromosome into a *Mycoplasma capricolum* cell which had its own genome removed (15). They also constructed a synthetic *Mycoplasma genitalium* genome *de novo* (16). Next, they took the entire genome from *M. mycoides*, modified it in yeast using yeast genetic systems, and then transplanted the modified chromosome into *M. capricolum* (17). Finally, they reported the creation of a bacterial cell containing only the chemically synthesized genome (18). Although the Venter Institute has developed the technical procedures necessary for synthetic cell construction, another important step toward minimal cell synthesis is defining precisely what is in its genome. Furthermore, there are no examples of an experimental test of whether a proposed gene set is sufficient for driving cellular life.

1.2.2. Natural Examples of Minimized Gene Sets

There are some natural analogs of the hypothetical minimal cell that have evolutionarily reduced genome sizes. All known small-genome bacteria are associated with specialized lifestyles in stable environments, e.g., obligate symbiosis or specialized ecological niches (14). The two largest forces pushing a bacterial species toward genome reduction are symbiosis and resource economization, so it is not surprising that the smallest genomes in nature are all in prokaryotes living in symbiosis with other cells (14). Notable examples include: *Nanoarchaeum equitans*, a symbiotic archaeon with 536 protein-coding genes (19); *Buchnera aphidicola*, an endosymbiont of aphids with 480 genes (20); and *Pachypsylla venusta*, an endosymbiont of *P. venusta* with 182 predicted ORFs (21). Because it can be grown in pure cultures and has an extremely small-genome size (580 kbp, 470 genes), *M. genitalium* is considered the best living example of a minimal cell (22); its genome represents a significant reduction from that of other well-studied bacteria such as *Escherichia coli*, which has approximately 4,300 genes (23). The *M. genitalium* genome developed through "top-down" genomics, where genes are removed from an existing organism to provide a metabolically simpler cell (24). Thus, it exemplifies natural selection for a minimized genome. These naturally minimized cells show that evolution (a "bottom-up" approach) has suggested many forms of a minimal cell (24), but all of them can survive the disruption of one or more genes (i.e., gene knockouts), and are therefore not truly minimal. Estimates based on observation of naturally occurring bacteria suggest minimal gene sets in the range of 200–500 genes (25–30).

1.2.3. Experimental Estimates of Minimal Gene Sets

There are genetic (26, 30), comparative genomic (25, 27, 31, 32), and biochemical (11, 33) approaches to establishing an *in vivo* minimal cell (11). Taken together, these techniques go beyond naturally occurring minimization to propose minimal gene sets in the range of 200–400 genes.

Genetic approaches identify essential genes by large-scale gene disruption Kobayashi et al. (28) estimated 271 genes as the minimal gene set by systematically inactivating single genes in *Bacillus subtilis* using transposon mutagenesis. Similar genetic methods have been used to estimate 1,490 essential genes in *Mycobacterium tuberculosis* (34), 254 essential genes in *B. subtilis* (35), and 382 essential genes in *M. genitalium* (26, 30). There have been other efforts to determine gene essentiality using gene inactivation (36, 37). However, this approach can lead to falsely labeling required genes as dispensable, which can derail any effort to create a minimal gene set (11, 38). Conversely, a genetic approach could also overestimate the minimal set substantially because genome scale knockouts may identify genes as essential even when the deletion only slows growth (32). Estimates have also been made using comparative genomics. Mushegian and Koonin estimated a set of about 250 genes as a minimal gene set after comparing the full genome sequences of *Haemophilus influenzae* and *M. genitalium* (25). In 2000, Koonin reviewed advances since their 1996 paper that demonstrate the complexity in using comparative genomics to establish a minimum gene set (25, 27). For example, of the 256 genes identified as essential in 1996, 15% were found to be dispensable in knockout experiments (27). Although many other computational analyses have yielded important insights (31, 39–44), comparative genomic approaches suffer limitations that could lead to either an over- or underestimation of minimal gene sets (11). They are particularly prone to missing unrelated proteins with the same activity, which is referred to as nonorthologous gene displacement (NOGD). Therefore, it is critical to develop a methodology for distinguishing among proposed minimal gene sets. Finally, parallel efforts have examined the minimal set of cellular reactions or functions, as opposed to specific genes. Forster and Church described the main biochemical pathways that are necessary for essential bacterial functions, as well as an *in vitro* plan to synthesize a minimal cell (11, 45). They proposed a minimal genome with 151 genes for cellular information processing but omitted genes involved in major metabolic pathways (11). Azuma and Ota (46) determined the "minimal pathway maps," or the minimal set of autonomous pathways maps that could synthesize all required biomass components, for *E. coli* and *B. subtilis.* They found that pathways maps from the

Kyoto Encyclopedia of Genes and Genomes (KEGG) were more likely to be conserved if they were involved in cellular information processing. This approach, while still computational, avoids the possibility of NOGD because a cellular function can be accepted into the minimal set regardless of NOGD.

The various approaches to determine a minimal gene set have been compiled and summarized in literature reviews (11, 14, 29). Forster and Church conclude that the biochemical approach is still more promising than genetics or comparative genomics (11). They and others outline the steps necessary for synthesizing a minimal cell, primarily from genes found in *E. coli* (11, 12, 33). Forster lists the five gaps in our current knowledge that should be filled for the production of a synthetic minimal cell. The fourth among these is the lack of "biochemical parameters and computational models sufficiently detailed to predict the effects of alterations (in a near-minimal cell)" (11). Similarly, Foley and Shuler list five essential characteristics of a biotechnological synthetic cell, the fifth being "mathematically defined interactions and predictable kinetics of (the) system" (47). These claims illustrate the importance of the minimal cell model (MCM) approach.

In 2004, Gil et al. (29) presented an enhanced review of all the previously proposed strategies for establishing a minimal gene set and proposed what they called the "core" of a minimal bacterial gene set (29). They started with a computational comparison of five sequenced endosymbionts: *Blochmannia floridanus*; *Wigglesworthia glossinidia*; and *B. aphidicola*, strains BAp, BSg, and BBp (41). To that, they added in genes that had functional, but not sequence, similarity amongst the bacteria considered. They compared their gene set with the essential genes for *B. subtilis* (28) and *E. coli* (37), as well as the computationally and experimentally derived minimal gene sets for *M. genitalium* (25, 26). Genes that were present in all five endosymbionts and that appeared to be essential in *Mycoplasmas* were considered essential even if they were determined to be nonessential in bacteria with larger genomes (29). Finally, they analyzed the gene list to fill in gaps in metabolic pathways that are assumed to be essential. This resulted in a gene set with 206 protein coding genes (29). The total was later corrected to 207 protein coding genes to account for a step missing from the pentose phosphate pathway (43). The gene set proposed by Gil has the following features (29):

1. A virtually complete DNA replication machinery, composed of one nucleotide DNA binding protein, single-stranded binding protein (SSB), DNA helicase, primase, gyrase, polymerase III, and ligase.
2. A simple DNA repair system.
3. A virtually complete transcriptional machinery, including the three subunits of the RNA polymerase, a σ factor, an RNA helicase, and four transcriptional factors.

4. A nearly complete translational system.
5. Protein-processing, folding, secretion, and degradation.
6. Cell division driven by FtsZ only.
7. Two substrate transporters (PTS for glucose and PitA for inorganic phosphate).
8. ATP production via substrate-level phosphorylation.
9. Four enzymes from the nonoxidative branch of the pentose phosphate pathway.
10. Biosynthesis of phosphatidylethanolamine from dihydroxyacetone phosphate and activated fatty acids.
11. Nucleotide biosynthesis from phosphoribosyl pyrophosphate (PRPP) and free bases adenine, guanine, and uracil, which are obtained from the environment.
12. Cofactor biosynthesis from precursors obtained from the environment.
13. No pathways for amino acid biosynthesis.
14. No protein transport systems for amino acids or inorganic ions (with the exception of phosphate).
15. No genes for stable RNA products (i.e., tRNA or rRNA), although they do define their proposed gene set as a minimal set of "protein-coding" genes.

Gil et al. (29) argue that there may be several possible minimal gene sets, saying "we should accept that there is no conceptual or experimental support for the existence of one particular form of minimal cell." In this work, one potential mechanism for distinguishing amongst minimal gene sets through computer modeling is presented.

1.3. Coarse-Grained Cell Models

Computational models have made significant contributions to our understanding of bacterial metabolism. Some models take advantage of detailed genomic information (48), while others are based primarily on flux-balance analysis, metabolic control theory, and mathematical techniques for optimization (49–53). These modeling techniques are all intrinsically static, and so have limited ability to predict aspects of cell regulation and dynamic response. Other investigators have proposed methods to directly incorporate dynamic (kinetic) information into models of central metabolism (54). While some models have attempted to describe whole cells (31, 55), these neglect important, nonmetabolic aspects of cell growth (e.g., control of chromosome replication or cell division) because there is no formalism to handle such "events" in the context of a cell model.

Constraint-based models, including flux-balance analysis, have a large representation in the literature. Under the time scale of minutes, metabolite concentrations in cells are generally at steady levels and remain constant as long as environmental conditions do not change. Therefore, a modeler can use the law of conservation of mass to constrain the synthesis and consumption rates of those metabolites. This is expressed as a stoichiometric constraint based on the stoichiometric relation proposed by each reaction in the system under study (53). The stoichiometric constraints are supplemented with restrictions regarding reaction reversibility and maximum reaction rate. The construction and applications of these models are reviewed in Durot et al. (53), and there are several interesting applications available (49–52).

These studies, and many other similar ones, make important contributions toward our perception of systems biology. However, all of these approaches neglect the coupling between cell physiology and cell growth that is prevalent in physiological events such as chromosome replication. Descriptions that neglect this coupling may yield misleading conclusions because they implicitly assume that the output of each pathway cannot influence any input into the same pathway (56). Further, many of the models referenced above assume an objective function, which typically maximizes the growth rate. While such a function can be justified in the context of a specific short-term situation, the real objective function (e.g., survival of the organism) is more complex and involves issues such as the ability to grow robustly and in a variety of environmental conditions.

The Shuler group has previously developed a whole-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-grained" model serves as the basis for our efforts to build a MCM. The Shuler group first described a mathematical model of a single *E. coli* cell in 1979 (57). While the *E. coli* model summarizes the physiological functionality required for a minimal cell, it does not capture explicitly the physical chemistry that supports those functions. It is unique in its natural coupling of metabolism, transport, and cellular events. At that time, it was the only model of an individual cell that did not dictate timing of cell division (e.g., growth rate) and cell size; instead, those aspects were outputs of the simulation. Also, it responded explicitly to concentrations of nutrients in the environment (58). The base model presented by Domach and Shuler (59) has been embellished with additional biological details to allow prediction of a wide-range of responses to environmental and genetic manipulations (60). The initial model included only 18 pseudochemical species that represented large groups of related chemical species. Figure 1 lists the components of the *E. coli* model and graphically depicts their relationships.

The mathematical description of cellular functions that comprise the model was based on time-variant mass balances for each component. Each mass balance accounted for the component's synthesis

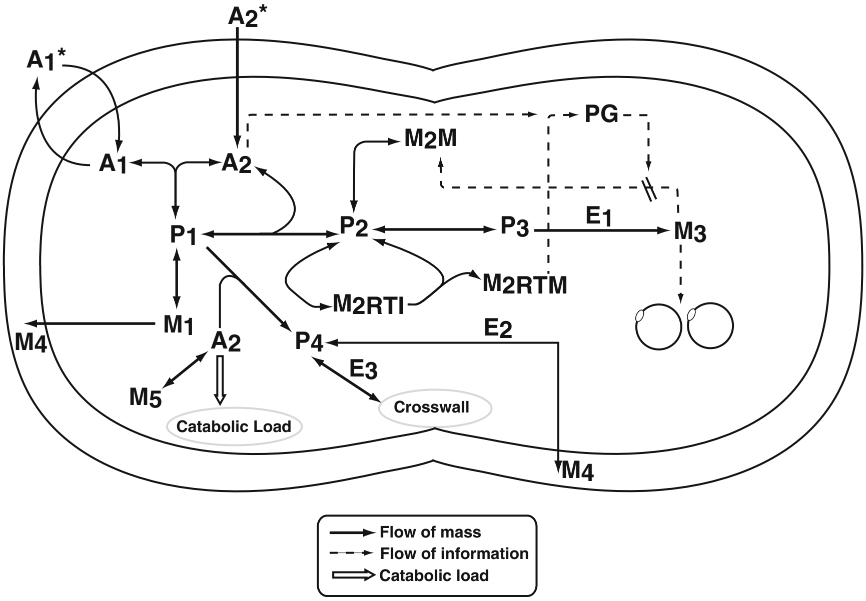

Fig. 1. A schematic representation of the single cell model and the modular approach to cell modeling. Labels indicate pseudochemical groups which are defined below. *Solid arrows* represent pseudochemical reactions that govern the rate and stoichiometry with which the pseudochemicals interconvert. *Hollow arrows* represent catabolic loads that account for energy metabolism. *Dashed arrows* represent the flow of information. The approach illustrated with this figure is used as the basis for the MCM presented in this chapter. The labels in pathways represent lumped pseudospecies defined as: A_1 ammonium ion; A_2 glucose; P_1 amino acids; P_2 ribonucleotides; P_3 deoxyribonucleotides; P_4 membrane precursors; M_1 protein; M_{2M} mRNA; M_{2RTI} immature stable RNA; M_{2RTM} mature stable RNA; M_3 DNA; M_4 cell envelope; M_5 glycogen; *PG* ppGpp; E_1 enzymes for conversion of P_2 to P_3; E_2 and E_3 enzymes for cross-wall formation and cell envelope synthesis. *Asterisk* indicates species that are external to the cell (61, 66, 95).

(as a function of availability of precursors, energy, and relevant enzymes), utilization, and degradation. Stoichiometric coefficients for relating components through mass balances were derived primarily from published research, and in some cases, from experimental data. It is important to note that the model did not contain 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 the cell, the model accurately predicted changes in cell composition, size, and shape, as well as the timing of chromosome synthesis as a function of changes in external glucose and ammonium concentration (61–65). The model also addressed important issues such as energy generation and the maintenance of the electropotential and chemical potential gradients across the cytosolic membrane by including a description of the cell's energy accounting process and the movement of H^+ ions (leaky protons) along the membrane (57, 62, 63). Two examples of stoichiometric mass balances for formation of precursors (amino acid) and macromolecules (RNA) are given in Eqs. 1 and 2.

$$\alpha_1 A_1 + \beta_1 A_1 + \cdots \rightarrow P_1 \tag{1}$$

$$\gamma_2 P_2 \rightarrow M_2 + \cdots \tag{2}$$

In Eqs. 1 and 2, α_1, β_1, and γ_2 are stoichiometric coefficients, and A_1, A_2, P_1, P_2, and M_2 are the masses of ammonium ion, glucose, amino acids, ribonucleotides, and total RNA, respectively. Chemical concentrations are measured in mass per cell, and stoichiometric balances are based on carbon and nitrogen. Equation 3 shows the corresponding requirements for phosphate energy coupled with the biosynthetic reactions.

$$\delta_{P1}\mathrm{ATP} \rightarrow \delta_{P1}(\mathrm{ADP} + P_i) \tag{3}$$

In Eq. 3, δ_{P1} is a stoichiometric coefficient representing the average amount of ATP hydrolysis that must occur to supply the energy required for synthesis of a specific amount of amino acids (P_1) per cell. Also, the chemical reducing potential generated and utilized is included in the accounting system. The change in mass of a substance per cell per unit time can be found from a dynamic mass balance accounting for synthesis, import, export, and consumption. Note that this is not the same as concentration because the cell volume is changing. Equation 4 is an example mass balance for deoxyribonucleotides:

$$\begin{aligned}\frac{\mathrm{d}P_3}{\mathrm{d}t} = k_3 \cdot \left(\frac{K_{P3}}{K_{P3} + (P_3/V_\mathrm{C})}\right)\left(\frac{P_2/V_\mathrm{C}}{K_{P3P2} + (P_2/V_\mathrm{C})}\right) \\ \times \left(\frac{A_2/V_\mathrm{C}}{K_{P3A2} + (A_2/V_\mathrm{C})}\right) \cdot E_1 - \gamma_3 \cdot \left(\frac{\mathrm{d}M_3}{\mathrm{d}t}\right)\end{aligned} \tag{4}$$

In Eq. 4, k_3 is the maximum rate of synthesis for deoxyribonucleotides formation (time^{-1}), K_{P3}, K_{P3P2}, and K_{P3A2} are saturation constants (mass/volume), γ_3 is a stoichiometric coefficient, and E_1 is the mass of enzyme E_1 per cell (the rate limiting enzyme for conversion of ribonucleotides into deoxyribonucleotides). The first term in brackets on the right hand side shows dependency based on deoxyribonucleotide concentration (P_3/V_C where V_C is cytosolic cell volume), the second term represents feedback inhibition of synthesis by ribonucleotide concentration (P_2/V_C), the third term indicates saturation-type dependence on glucose primarily for ability to supply energy (A_2/V_C), and the last term represents consumption to form DNA (M_3).

The original model explicitly described discrete events that had typically been ignored in other models (66). For example, in the *E. coli* model changes in gene dosage (the number of copies of a gene in a cell at a given time) depended on the replication fork position, and the completeness of cross-wall formation depended on the cell size and amount of cell membrane components synthesized. Other biochemical details were added in subsequent studies. For example, in one study, amino acids were differentiated into five families (67). In another study, the synthesis of ribosomes was incorporated in greater detail (68). The model was utilized extensively to improve the use of plasmids for recombinant protein production, e.g., (68–72). Bailey reviewed the importance of these contributions to the whole field of mathematical modeling in biochemical engineering (58); now the approach serves as the basis for the MCM modeling framework.

1.4. A Minimal Cell Model

We present here the construction of a MCM based on the gene set proposed by Gil et al. (29). However, previous work to establish a prototype MCM attempted to identify a minimal gene set independently. In 2001, the Cornell *E. coli* model was first used by the Shuler group as a basis to construct a prototype MCM that simulates a hypothetical bacterial cell with the minimum number of genes necessary to grow and divide in an optimal environment (73). The prototype MCM has also been posed as a generalized model of chemoheterotrophic bacteria. The strategy for transitioning from the original Cornell single-cell model into the prototype MCM was to sequentially replace "pseudochemical" and "pseudoreaction" components of the model with distinct chemicals and detailed reactions (74, 75).

It is our belief that a detailed model of *E. coli* would not be computationally tractable because of its large number of gene products (73). While it was not chemically detailed, the prototype MCM was complete in terms of physiological function and was modular in its structure. A modular species is one that can be deconstructed into individual components while still maintaining the essential connectivity to other functions in the cell (74). Adding detail to different modules allowed us to recombine those submodels into a functioning whole. The concept of modularity was demonstrated by the inclusion of genomically/chemically detailed nucleotide and lipid biosynthesis modules (74, 75). Additionally, detailed genomic information about the location of DnaA binding boxes on the *E. coli* chromosome was incorporated into the coarse-grained model to predict key features of DNA replication (65). Hence, the prototype MCM is a functionally complete, system-level model formed by modification of a coarse-grained model of a single cell of *E. coli* (73–75).

The new MCM described here goes beyond these prior models to describe explicitly all genes in the cell, all chemical species, and incorporates mechanisms for most cellular processes. The MCM focuses on essential functions while finding examples of gene products that can perform those functions. While the postulated set of minimal genes may change (e.g., if a new multifunctional protein is found), the set of essential functions is expected to stay relatively constant. Further, the technical difficulties associated with generating an experimental minimal cell and the ambiguities in interpretation of comparative genomic data promote 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.

The efficacy of constructing a MCM has been demonstrated in various proof of concept and validation studies (64, 73–75). It has been also demonstrated that it is not the exact values of model parameters that determine behavior, but that their values relative to one another is critical (73). This suggests that the lessons from a hypothetical general cell model will be broadly applicable to chemoheterotrophic bacteria.

2. Materials

2.1. Python Framework for a Minimal Cell Model

The MCM is a differential algebraic equation (DAE) system with discontinuities due to discrete physiological events (e.g., cell division). The full set of equations and parameters in Systems Biology Markup Language (SBML) format as well as instructions for download and simulation are available online at http://minimalcell.bme.cornell.edu. The DAE is integrated numerically using SloppyCell, a Python software package for simulation and analysis of biomolecular networks (76).

SloppyCell automatically compiles the structures listed in Table 1 and creates a Reaction Network object which can be integrated to obtain time course data for any variable in the model. All model simulation results presented here are generated by integrating the model from an initial condition until a stable cell-division limit cycle is reached. It is common to study how bacterial behavior changes at different steady-state growth rates, which is controlled by varying the external nutrient concentration. While we have done preliminary exploration of response to reduced glucose levels, only growth at saturating levels of glucose is necessary for a minimal cell.

2.2. Model Testing Framework

A model must meet certain requirements to be considered representative of a minimal organism, and several of these requirements are testable computationally. Using the Python Unittest framework (http://docs.python.org/library/unittest.html), a set of automated tests was implemented to verify that updates to the model did not violate any testable minimal cell requirements. Most importantly, we aimed to automatically verify that every version of the MCM met the following requirements:

1. Genome minimality—Every gene in a minimal cell is essential, by definition. Therefore, the elimination of any gene from the model should result in model failure. A series of tests were

Table 1
Model structures used in the minimal cell model

Model structure	Count	Examples
Compartments	4	Cytoplasm, cell membrane, whole cell, medium
Chemical species	408	Glucose-6P, alanine, mRNAs, proteins
Reactions	570	Fructose-6P synthesis, CTP synthesis
Rate parameters	570	Mass action or Michaelis–Menten rate constants
Saturation parameters	581	Michaelis–Menten-like saturation parameters
Inhibition parameters	25	Michaelis–Menten-like inhibition parameters
Rate rules	1	Methylation state of chromosome
Algebraic rules	1	Cell width (CW)
Events	36	DNA replication initiation, cell division
Constraints	408	Each species must have mass >0
Genes	241	Protein and stable RNA coding genes
Single coding genes	102	*dnaB*, *pgi*, etc.
Gene clusters	19	*replisome*, etc.
Genes in clusters	139	Ribosomal proteins, *dnaE*, etc.

With the exception of genes and gene clusters, all the modeling structures are analogous to their SBML counterparts (96). Rate, saturation, and inhibition parameters are can be set to values from the literature, or estimated using the procedures described in this chapter. While there are 241 identified coding loci in the model, only 102 are modeled as single genes. The remaining 139 are lumped into groups that have closely coupled function and dynamics. These lumped groups are here named "gene clusters." Table reused with permission from ref. (95)

implemented that sequentially removed each gene in the model, and verified that the loss caused model failure.

2. Resource minimality—While the minimal cell does live in an optimally supportive culture environment, it should not have unnecessary nutrients in the medium. The presence of an unnecessary nutrient indicates a logical error in the assumptions about which genes are essential, because those nutrients are likely participating in one or more reaction pathways that may not be required. These tests removed each nutrient in turn from the medium to ensure that its loss causes model failure.
3. Structure tests—A third set of tests ensured that rules, events, and other model structures worked as expected in the MCM. For example, for all times, the sum of all individual protein masses in the cell should equal the total mass of protein in the cell (M_1). Similarly, the total mass of the cell should equal the mass of the membrane plus the mass of the cytoplasm.

3. Methods

3.1. Minimal Gene Set

The MCM implements a whole-cell dynamic model of a single cell that contains the minimal gene set described by Gil et al. (29). The authors break their minimal gene set into five major categories:

1. Information storage and processing.
2. Protein processing, folding, and secretion.
3. Cellular processes.
4. Energetic and intermediate metabolism.
5. Poorly characterized genes.

There are key differences between the gene set presented in Gil et al. (29) and what is included in the base MCM. In particular, the minimal gene set proposed by Gil et al. (29) only considers protein-coding genes (it does not include tRNA or rRNA species). Furthermore, the authors assumed that the cell could import amino acids and inorganic ions (e.g., K^+ and Mg^{2+}) from the environment through diffusion, but it is likely that transporters will be required. Finally, the authors suggest that the cell will synthesize ATP exclusively through substrate-level phosphorylation via lactate fermentation, but they provide no mechanism for synthesized lactate to exit the cell. Therefore, genes coding for three rRNA species, 20 tRNA species, 14 protein components of amino acid transport systems, four protein components for transport of inorganic ions, and one protein corresponding to a lactate transporter has been added to the MCM. These, together with the genes identified in Fraser et al. (22), account for the 241 genes included in the MCM (see http://minimalcellmodel.bme.cornell.edu for a detailed listing). Figure 2 shows an overview of the metabolic features of the MCM. Table 2 shows a summary of how many genes fall into each functional category in the MCM.

3.1.1. Information Storage and Processing

DNA Metabolism

The DNA replication and repair systems are less complex in Mycoplasma species than in bacteria with larger genomes (77), and similarly we expect that a minimal bacterium would retain a simple DNA replication system. Gil et al. (29) state that the four basic steps of DNA replication are:

1. Recognition of the origin of replication by protein components.
2. Recruitment of initiator proteins to the origin to promote initiation of replication.
3. DNA synthesis along two forks on the circular chromosome.
4. Replication termination and the separation of the daughter chromosomes.

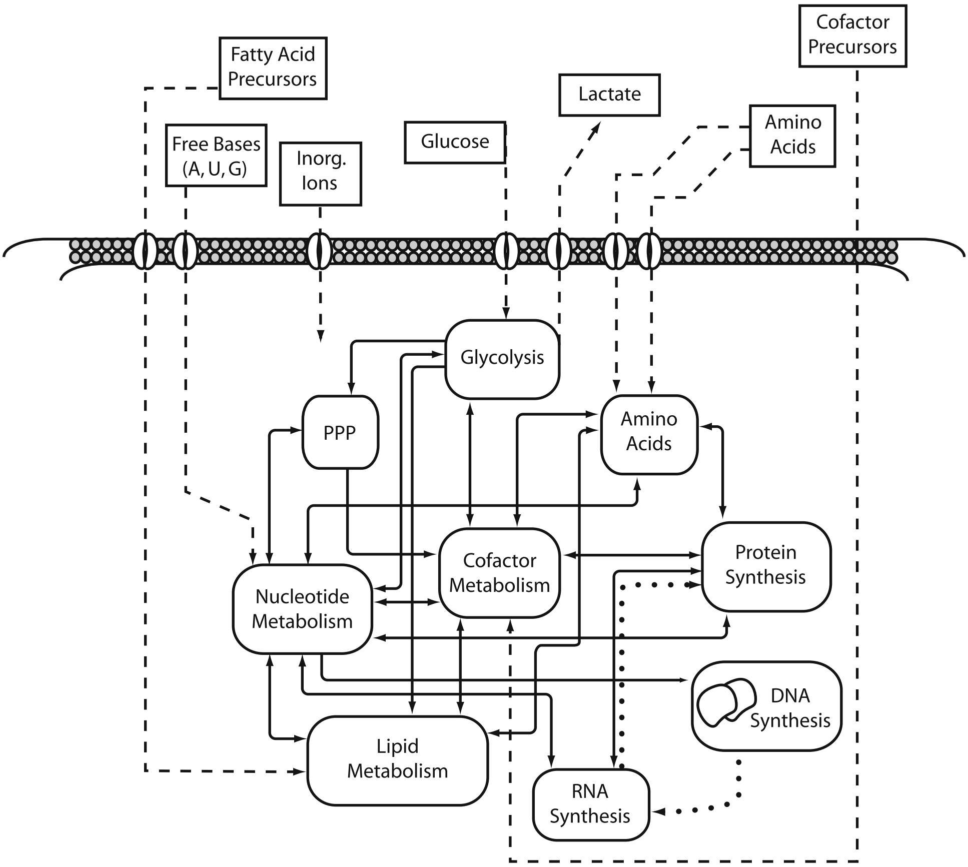

Fig. 2. Overview of metabolic processes included in the MCM. External nutrients for the MCM include glucose, amino acids, inorganic ions, cofactor precursors, fatty acid precursors, and free bases. Boxes in the cytoplasm are subsets of metabolism described by the MCM. *PPP* pentose phosphate pathway, *solid lines*—flow of mass within the cell, *dashed lines*—transport processes.

DNA replication initiation mechanisms vary widely in different bacteria. The MCM combines concepts proposed by Gil et al. (29) and those used in a DNA replication model in *E. coli* (64, 65). Gil et al. include 13 genes in the minimal gene set for DNA replication (29). Of those, three (*dnaB*, *dnaG*, and *hupA*) are modeled explicitly as initiators of DNA replication, while the remaining 10 are included in the replisome gene cluster.

Gil et al. (29) also include three genes in the minimal gene set for DNA repair, restriction, and modification. It is debatable whether a minimal cell would require these functions. Because the MCM exists in a totally benign environment the extent of DNA damage would be minimized. However, because single strand breaks during DNA replication are common in natural bacterial species, we would expect that the absence of these genes in a hypothetical minimal cell would result in severely reduced cell viability based on studies done in *E. coli* (78). Note that an average cell viability of less than 50% would result in an unsustainable

Table 2
Summary of genes used in the minimal cell model, listed by category

Category	No. genes
Basic DNA replication machinery	14
Basic transcription machinery	8
Biosynthesis of cofactors	12
Biosynthesis of nucleotides	15
Cell division	1
DNA repair, restriction, and modification	3
Glycolysis	10
Lipid metabolism	7
Pentose phosphate pathway	4
Protein folding	5
Protein post-translational modification	3
Protein translocation and secretion	5
Protein turnover	3
Proton motive force generation	9
Ribosomal RNA (rRNA)	3
Transfer RNA (tRNA)	20
Translation factors	12
Translation: amino-acyl-tRNA synthesis	21
Translation: ribosomal proteins	50
Translation: ribosome function, maturation, and modification	7
Translation: tRNA maturation and modification	6
Transport	23

Table reused, with permission, from ref. (95)

cell culture. Therefore, the three genes suggested by Gil et al. (29) (*nth*, *polA*, *ung*) have been included. However, because the MCM does not include a mechanism for DNA damage, the protein products of these genes have no mathematical impact on the cell behavior. Currently, their only impact is via the energy burden the cell experiences in their synthesis. It is possible that this model could be modified to account for relevant DNA damage, and in that case the three genes included for DNA repair would have a mathematical function.

RNA Metabolism and Translation

Gil et al. list eight genes as being necessary for the basic transcription machinery (29). Of these, seven are included in an RNA polymerase gene cluster. The remaining gene, *nusA*, is used in transcription/translation coupling, and is therefore included in the gene cluster for translation factors. In addition to these eight, the MCM explicitly includes 19 of the 21 proposed amino-acyl-tRNA synthesis genes. The remaining two, *pheS* and *pheT*, are the α and β subunits of a single amino-acyl-tRNA synthetase, so are included as a single gene cluster. The six genes for tRNA maturation and modification are included in the MCM as a single gene cluster. There are 50 ribosomal proteins included in the Gil et al. gene set (29). All 50 of these are included in a single gene cluster called *ribO*, the largest gene cluster by far. In the absence of a detailed mechanistic model for ribosome assembly and function, these genes must remain in a single cluster with a single product corresponding to ribosomal protein. Seven genes responsible for ribosome function and maturation are included in the MCM as a single gene cluster called *ribM*. The product of this gene cluster catalyzes RNA maturation and ribosome synthesis reactions in the MCM. All 12 genes listed as translation factors in the Gil et al. (29) gene set and *nusA* are included as a single "translation factor" gene cluster called *transF*. There are two genes that participate in RNA degradation in the Gil et al. (29) gene set, *pnp* and *rnc*. They are included in the MCM as a single gene cluster called *degRNA*.

3.1.2. Protein Processing, Folding, and Secretion

The minimal gene set proposed by Gil et al. (29) includes two genes related to post-translational modification. One of these, *pepA*, was omitted from the MCM gene set because it is unclear how its product, aminopeptidase A/I, would be used in the minimal cell. Gil et al. (29) included *pepA* because it was present in all of the genomes they considered. However, it is nonessential in both *E. coli* and *B. subtilis* (29). The other gene dedicated to post-translational modification in the proposed minimal gene set is *map*, which codes for methionine aminopeptidase, has been included in the MCM (29). Five genes for protein folding, *dnaJ*, *dnaK*, *groEL*, *groES*, and *grpE*, are included in the Gil et al. (29) gene set. Because protein folding is required in all cells, we have included these genes in the MCM as a single gene cluster. However, the MCM does not contain a protein folding submodel, so the products of the protein folding gene cluster do not impact the model simulation. Finally, the three "protein turnover" genes proposed by the Gil et al. (29) gene set, *gcp*, *hflnB*, and *ion* are included as a single gene cluster that catalyzes protein degradation.

3.1.3. Cell Division

Gil et al. (29) propose that the only gene necessary for cell division in their minimal cell is *ftsZ*, and this gene is explicitly included in the MCM. At the time of DNA replication termination, FtsZ catalyzes the transfer of membrane material to the midcell region, promoting cell division. Bacterial cells with the *ftsZ* gene typically have between

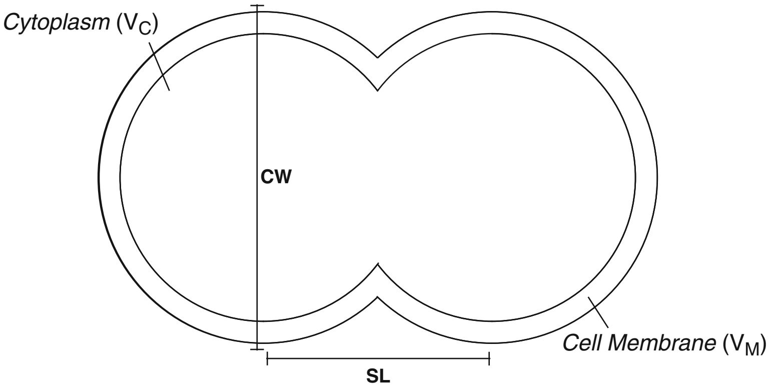

Fig. 3. The spherical minimal cell model. *CW* cell width. The two labeled compartments, cytoplasm (V_C) and cell membrane (V_M), together comprise the volume of the whole cell, V. This illustration shows the cell after septum formation as started. When the septum is complete (i.e., SL = CW·2), division occurs. Figure reused, with permission, from ref. (95).

5,000 and 20,000 FtsZ molecules (79). When termination of DNA replication completes and the cell division process starts, FtsZ recruits membrane material to the septum. This results in a "figure-eight" shaped cell where the connecting region gets thinner and thinner until the cell divides, as in Fig. 3.

3.1.4. Transport

Gil et al. (29) include four genes related to transport of nutrients into the cell. An inorganic phosphate transporter, *pitA*, is included explicitly in the MCM. The three genes coding for the phosphotransferase system (PTS), *ptsG*, *ptsH*, and *ptsI*, are included as a single gene cluster.

3.1.5. Energetic and Intermediate Metabolism

Metabolic processes are straightforward to represent in the coarse-grained modeling framework, as these reactions are the main basis for the previous cell models (61). All 10 genes listed by Gil et al. (29) for glycolysis are included explicitly in the MCM. The nine genes included as part of the ATP synthase machinery are included as a single gene cluster in the MCM. It is presumed that the ATP synthase can extrude protons from the cell and thereby maintain the proton gradient by catalyzing the ATP synthesis reaction in reverse. This is common behavior amongst lactic acid bacteria (80). The four genes included for the pentose phosphate pathway are included explicitly in the MCM (29, 43). The minimal gene set contains genes for synthesizing ATP through substrate-level phosphorylation only. Specifically, the cell does not have an electron transport chain. It does contain the F1ATPase in the cell membrane, but Gil et al. (29) proposed it would participate principally in proton gradient maintenance. The Gil et al. (29) gene set does not explicitly address the issue of cellular use of NAD^+ vs. $NADP^+$ in terms of reducing power. A review of the reactions catalyzed by the minimal proteome reveals that in principle NAD^+ coupled with NADH should be sufficient. The single exception is that TrxB (thioredoxin reductase) does prefer $NADP^+$, but there is some evidence that a similar enzyme could function with NAD^+ (81), so we follow the assumption of Gil et al. (29) and Gabaldón et al. (43) and use NAD^+/NADH for redox reactions. Importantly, the metabolic rates in the MCM are able to balance NAD^+ and NADH so that there is sufficient reducing power generated. Of the seven genes listed for lipid metabolism, four (*cdsA*, *gpsA*, *psd*, and *pssA*) are included explicitly as single genes. The remaining three genes (*plsB*, *plsC*, and *fadD*) are included as a single gene cluster involved in lipid biosynthesis. *plsB* and *plsC* have been proposed as the basis for lipid membrane synthesis in semisynthetic minimal cells (82). All 15 genes listed for nucleotide biosynthesis by Gil et al. (29) are included explicitly as single genes in the MCM. The 12 genes identified by Gil et al. (29) for cofactor biosynthesis are also explicitly included in the MCM.

3.1.6. Additional Genes

The Gil et al. (29) gene set contains only four genes related to transport of nutrients into the cell, as the authors proposed that the cell could obtain essential nutrients from the environment by diffusion (29). This may suffice for some nutrients, but it is likely that protein transporters will be necessary for many others. Therefore, the gene set proposed by Gil et al. (29) is supplemented with an additional 19 genes dedicated to the transport of chemicals such as amino acids. The MCM has a total of 23 genes related to transport. The Gil et al. (29) gene set also does not include coding regions for tRNA or rRNA species as they are not protein-coding genes. These genes, however, are clearly essential parts of the minimal genome for a modern chemoheterotrophic bacterium. The MCM computer chromosome was supplemented with coding regions corresponding to 20 tRNA species. In cases where multiple tRNA alleles correspond to a single amino acid, we assumed that the tRNA region represented a gene cluster coding for all of those alleles. The genome was also supplemented with genes for three rRNA species. We found that the MCM generated large amounts of lactate because while the Gil et al. (29) gene set includes lactate dehydrogenase (which consumes pyruvate and NADH), it does not include a mechanism to consume lactate. We propose the addition of the *lctP* gene for export of lactate to the external environment.

3.1.7. Other Departures from the Proposed Minimal Gene Set

There are other genes that, while necessary for a minimal cell, have no mathematical model available for their interaction with the whole-cell. In these cases, we have elected to include the genes to account for their metabolic burden on the cell, but their genes and gene-products currently have no connection to the rest of the cell. The mathematical model could be adjusted to reflect their function as more detailed descriptions of these components become available. These genes include those whose gene products degrade macromolecules (*degM1* and *degRNA*), act solely on ions in the cell (*kup*, *mgtA*, *mntH*, *nhaB*, *pitA*, *pmf*, and *ppa*), or catalyze processes for which the MCM lacks mechanistic detail (*dnarep*, *protfold*, *map*). The proposed minimal gene set includes the *pepA* aminopeptidase. However, there is no clear function for this gene in the minimal cell, so we choose not to include it. Eight "poorly" characterized genes are included in the gene set proposed by Gil et al. (29). Most of these have no known function, but were included because they were present in all of the genomes considered in the study. Of these eight, only *mraW* is included in the MCM. MraW is a methyltransferase which is assumed to be necessary for DNA methylation and chromosome replication. However, the rest have no clear function for a minimal cell, and are therefore not included in the MCM. The full list of genes from the gene set proposed by Gil et al. (29) which have been excluded in the MCM is presented at the project website at http://minimalcellmodel.bme.cornell.edu.

3.1.8. Analysis of the Minimal Gene Set

The minimal gene set proposed by Gil et al. (29) has been analyzed in subsequent work by Gabaldón et al. (43). To perform a structural analysis, Gabaldón et al. (43) eliminated many of the 206 protein-coding genes from the minimal gene set proposed by Gil et al. (29). Specifically, they removed polymerization reactions and any reactions involving macromolecules. Furthermore, they only considered reactions represented in the pathway maps of the KEGG database, which eliminates many reactions involving cofactors. Finally, the authors also only considered reactants and products that had at least one carbon atom in common on each side of the reaction. A metabolic reaction network was thus constructed by comparing the gene functions from Gil et al. (29) to the new reaction database created in Gabaldón et al. (43). The connection degree distribution, clustering coefficient, average path length, and network diameter, were measured for the metabolic reaction network (43). It was found that the average path length and network diameter tended to decrease with the size of the network (n) rather than with the size of the genome. An average path length and network diameter of 5.34 and 18, respectively, were reported for the minimal gene set when they considered a network with 165 nodes by applying the eliminations discussed above (43). Gabaldón et al. (43) also found that a random network had a much smaller clustering coefficient than the natural or minimal gene sets ($C = 0.031$ for the minimal gene set compared to $C_r = 0.00977$ for a random network of the same size). However, the ratio C/C_r increases linearly with the number of nodes in a network, so smaller networks (including the minimal gene set) have less clustering. Most importantly, the results from Gabaldón et al. (43) show that the minimal gene set and its corresponding reaction network behaved as one would expect for a natural genome of the same

Table 3
Characteristics of the minimal cell model genome

Characteristic	MCM value	Lit. value	Reference
Genome size (kbp)	233	580	Value from *M. genitalium* (22)
GC content (%)	40	27.73	Median value for mollicutes (83)
Gene density	100	81–92	Various *Mycoplasma* sp. (83)

Table reused, with permission, from ref. (95)

size. Gabaldón et al. (43) also considered a reduced theoretical reaction network containing only 39 genes with 50 enzymatic steps for stoichiometric analysis. Their stoichiometric analysis did not include cofactor metabolism because, they argued, coenzymes play a catalytic function and do not affect the stoichiometric analysis. The reduced theoretical reaction network also assumes lactate to be a "sink" chemical whose concentration is essentially buffered. Using the reduced theoretical reaction network, they investigated the robustness of the minimal gene set. They found that most mutations had a limited effect on the topology of the network, but that the removal of a few key enzymes had drastic effects. At the same time, the network was sensitive to sustained random attacks. This analysis, however, did not imply that the minimal gene set could be further reduced because maintaining the topology of a network is different than maintaining its viability (43).

The minimal gene set used in the MCM is a modified and supplemented version of that presented by Gil et al. (29). This genome's characteristics can be compared to those of some naturally occurring small-genome bacteria as in Table 3 (22, 83). The mollicutes, a category of bacteria that tend to have small size and small genome, do not have a common general organization to their genomes (83), but some of their features could be used as organizational baselines for the MCM. For example, some mollicutes display bias in the GC skew near the chromosomal replication origin and DNA replication initiation loci. Table 3 lists a gene density of 100% for the MCM. This is because the MCM has no noncoding regions of DNA. If one or more noncoding regions are deemed necessary to bacterial survival, they can be added to the MCM as genetic loci. For example, the origin of replication, *ori*, is included as a genetic locus.

3.2. Reaction Network Construction

3.2.1. Genome Construction

Once the gene set is assembled, the reaction network for the MCM is constructed within our modeling framework. The genes in the minimal bacterial gene set are not necessarily present in all bacterial species (due to nonorthologous gene displacement), nor is the sequence for a gene always known. The genomic sequences for

Table 4
Distribution of source genomes for finding sequences for the genes in the minimal gene set

Organism	KEGG abbreviation	Number genes used
Mycoplasma genitalium	mge	162
Escherichia coli	eco	59
Bacillus subtilis	bsu	10
Wigglesworthia brevipalpis	wbr	3
Synechococcus elongatus	syc	4
Cytophaga hutchinsonii	chu	1
Bacillus pumilus	bpu	1
Rhodobacter sphaeroides	rsp	1

The organisms are listed in the order in which they were searched. Table reused, with permission, from ref. (95)

the MCM's gene set were almost exclusively downloaded from the KEGG website (http://www.genome.jp/kegg/). For each gene in the minimal gene set, we searched the KEGG database gene bank for the following list of organisms, in the order shown in Table 4.

3.2.2. RNA and Protein Synthesis

After we identified an appropriate DNA and protein sequence for each gene in the MCM, sequence-dependent stoichiometries were constructed for the mRNA and protein synthesis/degradation reactions. Furthermore, the stoichiometry of DNA synthesis was based on the DNA sequence. Thus, the actual consumption of amino acids and nucleotides in the MCM depended on gene-level sequence information. Rate laws for the synthesis of RNA species were constructed according to the coarse-grained templates in Eqs. 5 and 6.

$$\left(\frac{\mathrm{dRNA}}{\mathrm{d}t}\right)_{\mathrm{S}} = v_{\mathrm{RNA}i} \cdot \frac{\mathrm{GD}_i}{\mathrm{GD}_{\mathrm{sum}}} \cdot \left(\frac{\mathrm{d}M_2}{\mathrm{d}t}\right)_{\mathrm{S}} \tag{5}$$

$$\left(\frac{\mathrm{d}M_2}{\mathrm{d}t}\right)_{\mathrm{S}} = \mu_{M2\mathrm{S}} \cdot P2\mathrm{min}_{\mathrm{sat}} \cdot M_3 \cdot \mathrm{RNA}_{\mathrm{pol}} \tag{6}$$

In Eq. 5 $v_{\mathrm{RNA}i}$ is a synthesis rate specific to RNA_i that is biologically related to a promoter strength (pg RNA_i/pg M_2), $\mathrm{GD}_i/\mathrm{GD}_{\mathrm{sum}}$ is the fraction of total gene dosage represented by gene i, and $\mathrm{d}M_2/\mathrm{d}t_{\mathrm{S}}$ is the overall RNA synthesis rate for the cell. The gene dosage term appears for all mRNA synthesis equations by default, but if it is not required it can be optionally removed (i.e., when a gene's transcription is not regulated this way).

In Eq. 6, μ_{M2S} is the overall RNA synthesis rate constant (pg M_2/h/pg M_3/pg RNA_{pol}), $P2\text{min}_{\text{sat}}$ is a dimensionless saturation term for the scarcest ribonucleotide precursor, M_3 is the mass of DNA (pg), and RNA_{pol} is the lumped mass of enzymes involved in RNA synthesis (pg). Note that due to the promoter strength constant in Eq. 5, the sum of all RNA synthesis rates will not sum to $\mathrm{d}M_2/\mathrm{d}t_S$. Equation 6 is therefore supposed to represent a base capacity for RNA synthesis, the apportionment of which is determined for each RNA species by Eq. 5.

Gene dosage for each gene is monitored automatically as a function of the replication fork position on the chromosome. If there is a single, nonreplicating chromosome, in the cell, then the dosage for each gene is equal to the gene copy number. Once DNA replication begins, the gene dosage for each gene becomes a calculable function of fork position (fork position is constrained by the mass of DNA that has been synthesized since the most recent DNA replication initiation). There are two ways to calculate gene dosage. It can be updated via events each time the replication fork passes through a coding locus. For many genes, this tends to be a slow method because many events will fire as soon as the chromosome begins replicating. Alternatively, gene dosage can be calculated using a smooth function that approximates a step function. We use a smooth exponential function to calculate the gene dosage (see Note 1).

Real cells require RNA degradation so they can reuse nutrients over the course of the cell cycle as different gene functions become necessary. For a minimal cell cultured under constant benign environment, the need for RNA turnover is far less compelling than for a cell that has a plethora of genes to choose from. Therefore, the MCM has relatively low degradation rate constants. Finally, it is assumed that "stable" RNA species such as ribosomal RNA (rRNA) have no degradation reactions.

Protein synthesis rates are calculated using a similar coarse-grained template inspired by our previous efforts in bacterial cell modeling (61, 66).

3.2.3. Metabolic Reactions

Metabolic reactions corresponding to the genes in the MCM genomes were assembled with the aid of the KEGG database as well as knowledge of microbiological biochemistry. Developing a model of this scale is complicated by lack of kinetic information for most of the proposed reactions. At the same time, parameter analysis research has revealed that in many biological models, the specific values of parameters are not as critical as their ratios to one another (73, 84, 85).

Saturation constants for activation terms in saturation-type rate laws were estimated by applying a general rule of thumb that postulates that a reasonable value for an unknown saturation constant is one 25th of its normal intracellular concentration (NIC)

(61). Similarly, inhibition constants for inhibition terms in rate laws are estimated by applying a heuristic that the constant will be equal to ten times that chemicals NIC. In the MCM, the NIC is set to the predicted average concentration of each chemical species. This rule has been applied in previous models (61, 67).

We also present here a method to quickly estimate rate constants for coarse-grained models of single cells growing at steady. The goal of developing this procedure is to rapidly obtain a reasonable set of parameters that can be used to help test the plausibility candidate minimal gene sets. This method is based on the assumption that in a single cell growing and repeatedly dividing at steady-state, each chemical species' mass will double in the time that it takes for the cell to divide, τ_D. This assumption is certainly true in an exponentially growing population of bacterial cells experiencing balanced growth, and applying the assumption to the single-celled model allows us to calculate rate constants for the reactions in the model.

We begin by using the doubling assumption for species X_i (i.e., $X_i(\tau_d) = 2X_i(0)$) to write Eq. 7

$$\int_0^{t_d} \frac{dX_i}{dt} dt = X_i(t_d) - X_i(0) = X_i(0) \tag{7}$$

The rate dX_i/dt is not constant, but for most chemical species the mass X_i will increase monotonically until it doubles in a nearly linear fashion. We can take advantage of this to calculate a set of approximate rate constants that are likely to result in a cell model that will achieve a stable cell division cycle. Specifically, it is assumed that the rate of production of a species X_i is linear in the rate constants v_j, and that the nonlinear portions of the rate laws are known functions of the set all chemical species masses X. Furthermore, it is assumed that each species creates a constraint on some of the rate constants as in Eq. 8.

$$\sum_{j=0}^{N_R} v_j \cdot \alpha_{i,j} \cdot f_j(X) \geq ss_i \cdot \frac{X_i(0)}{t_d} \tag{8}$$

Specifically, Eq. 8 says that the sums of all the reaction rates acting on species i are constrained to being greater than $X_i(0)$, the mass of species i at time 0, divided by the desired doubling time. While the assumption of linearity is not true (because $f_j(X)$ is nonlinear), by applying this assumption to the initial conditions for the MCM, linear constraints on the rate constants for the model are obtained. This results in a system of constraint equations on all the rate constants in the model, which can be expressed as a matrix **A**. We define an objective function f_{opt} as

$$f_{opt} = \sum_{i=1}^{N_R} v_i \tag{9}$$

where N_R is the number of reactions, and v_i is the rate constant for rate constant i, is introduced to frame the problem as a Linear Programming (LP) problem with constraints **A** and objective function f_{opt}, which is minimized to obtain a starting set of rate constants (see Note 2).

3.3. Geometry

The model cell is composed of two compartments: a cytoplasm and a membrane. The shape of the cell is assumed to be constrained to a sphere, but a cylindrical model has been tested. Cell size is determined automatically from the volume of its compartments (i.e., a constant density is assumed for each compartment). It is assumed that the cell shape is spherical, and that septum formation at the mid-cell region (Fig. 3). The two parameters describing the shape of the cell are the length of the cylindrical cell body (CL) and the width of the cell body (CW). For a spherical cell CL is always zero. The length of a dividing cell's dividing region (the septum) is referred to as SL.

3.4. Demands

Cellular processes such as DNA replication, transcription, and translation, consume various reactants to create long biological polymers (i.e., DNA, RNA, and protein, respectively). While it is possible to model a dependence on multiple substrates using a combination of Michaelis–Menten like saturation terms, the combination of many such terms leads to unreliable models. This is because the combination of many fractional terms can lead to greatly reduced reaction rates, even if all the reactants are in excess in the cytoplasm. For example, there are 20 reactants in the pseudoreaction that produces a particular protein product. Even at high concentrations, the cumulative effect of 20 saturation terms in a rate law could greatly decrease the calculated rate if they were all included. Instead, we hypothesize that at any given time, a single reactant will have the highest "demand" in a reaction. We propose that synthesis of biological polymers depend on single reactants in a Michaelis–-Menten fashion. For example, translation will only depend on a single, limiting amino acid. During growth and development, the limiting amino acid may change to reflect the changing demands of the cell. To address that phenomenon, a "Demand" class was created for the MCM. Each Demand object creates the parameters, equations, and events necessary to track the limiting reagent for a particular reaction. To create each Demand, we specify the species that could act as limiting reagents for a reaction, as well as their saturation constant for that particular reaction. The mass of each species was used to determine the limited chemical (i.e., the species with the lowest mass has the highest demand). This could later be updated to use the number of moles or molar concentration, but such an update is left as future work. The potential for demands to impact the cell behavior are illustrated in Fig. 4, which shows an example of how the "in demand" species for a reaction could change over the cell cycle, and how that change affects the model equations.

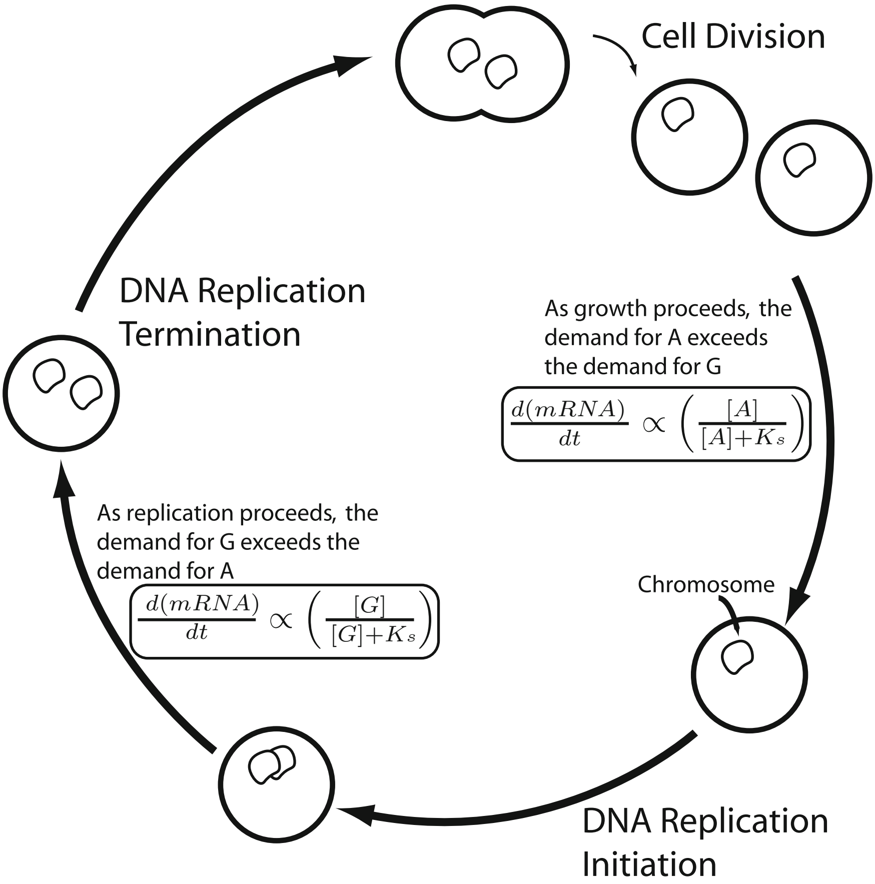
Cell Division
As growth proceeds, the demand for A exceeds the demand for G
$\frac{d(mRNA)}{dt} \propto \left(\frac{[A]}{[A]+K_s}\right)$
Chromosome
DNA Replication Initiation
As replication proceeds, the demand for G exceeds the demand for A
$\frac{d(mRNA)}{dt} \propto \left(\frac{[G]}{[G]+K_s}\right)$
DNA Replication Termination

Fig. 4. Chemical species demands over the course of the cell cycle. During the course of the cell cycle, changes in gene dosage can cause changing requirements for nucleotides. In this illustration, the demand is initial for ATP, and then switches to GTP. Figure reused, with permission, from ref. (95).

Note that at the beginning of the simulation, one (and only one) of the demand species in a Demand object can be limiting (i.e., the species associated with a particular Demand cannot all initially be equal). If they were, the system could not select an initially limiting reagent. The purpose of tracking demand during the simulation is to calculate which reactant is limiting the reaction at a given time. A high demand corresponds to a low concentration of a species, and a low demand corresponds to a high concentration. When the demand for species A surpasses the demand for species B, the reaction in question will automatically start using the mass of the species B in the calculation of the reaction rate.

3.5. Events

Events describe instantaneous, discontinuous changes in the state of the model, and an implementation of events based on SBML is used here (86). Because they cause discrete changes in the cell structure or behavior that occur instantaneously when the cell reaches some predefined condition, events require special mathematical treatment during a simulation. For example, the "initiation of DNA replication" event occurs when a threshold number of DnaA molecules are bound to the DNA *OriC*. In the MCM, an event could, e.g., describe instantaneous changes in the masses of the chemical species in the cell (i.e., at cell division). There are a total of 36 events in the base model. The names and trigger functions for all 36 events are presented at http://minimalcellmodel.bme.conell.edu. Here, we present as examples a generic event, as well as the "DNA Initiation" and "DNA Termination" events from the MCM.

3.5.1. Generic Event Example

Imagine an event where the concentration of a metabolite (elicitor) activates the synthesis of a species in a secondary metabolic pathway. When the concentration of the elicitor is above a threshold, the event is triggered, i.e., when $[elicitor] > threshold$. Once the trigger function's value changes from false to true, the event "fires," and the cell responds by executing a number of event assignments. In the case of the elicitor, one might expect a number of reaction pathways to be activated or augmented. For example, we could write the following two event assignments:

$$v_x \rightarrow 1e^6$$

$$\text{flag}_e \rightarrow 1$$

where v_x is some reaction rate constant that is increased to a new level by the presence of the elicitor, and flag_e represents that some other physiological process has been activated.

3.5.2. DNA Initiation DNA

Initiation is the start of chromosome synthesis. The trigger function for DNA Initiation is shown in

$$(\text{DnaG}_{\text{bound-to-Ori}} \geq \text{init}_{\text{threshold}})||(\text{flag}_{\text{meth}} = 1) \qquad (10)$$

In short, the replication process is triggered when the mass of DnaG bound to the origin of replication (Ori) exceeds threshold $\text{init}_{\text{threshold}}$. There are currently 21 event assignments associated with DNA replication initiation (see Note 3).

3.5.3. DNA Termination

The simple trigger function for DNA replication termination becomes true when the replication fork reaches the terminus of replication.

$$\text{ForkPos}_0 \geq 1.0 \qquad (11)$$

After DNA replication ends, 11 variables are updated in the MCM. For example, C_{period}, the length of chromosome replication, is updated to reflect the total time during which chromosome replication was active.

3.6. Estimation of Initial Conditions

A chemically detailed model of a bacterial cell must have an initial mass equal to the sum of all its chemical species. For many chemical species, even average cell cycle values are not known, let alone detailed concentration information as a function of the cell cycle progression. To obtain initial conditions for the MCM, we used data for groups of chemical species published for *E. coli* and made assumptions about how these groups would be subdivided into the hypothetical cell (87). Because no experimental analog for a minimal cell exists, we propose that using composition data measured in *E. coli* is a valid first-approximation because a minimal cell would have a similar chemical make-up to other chemoheterotrophic bacteria.

The average component masses used to calculate initial conditions are summarized in Table 5 (87). These proportions agree with the *E. coli* data from which they were derived. Once the component masses were estimated, the masses of individual chemical species were initialized using a procedure we developed for the MCM (see Note 4). The initial conditions for all species in the base MCM are available for download from http://minimalcellmodel.bme.cornell.edu.

This estimate of initial conditions for each chemical species is instrumental in determining the reaction rate constants in the MCM. The final simulated birth composition is found by letting the cell establish steady-state replication and differs from this initial estimate. The initial estimate must be sufficiently realistic to yield a stable behavior in the model cell.

3.7. Simulating a Repeating Cell Cycle

To demonstrate that the current proposed minimal gene set is capable of supporting cellular life, we show now that it is capable of simulating a repeated cell division cycle. Once the initial conditions and parameter values for the model are all set, we perform a numerical integration of the model DAE system using SloppyCell. Typical results from such an integration are presented in Fig. 5, which shows the mass of ATP over time for a nascent MCM integration. It is of note that the trajectory is not initially steady. Rather, the mass of ATP increases sharply over the first several hours of simulation time and then dips again before reaching a stable, repeating state, showing that the MCM dynamically approaches a steady-state rather than arbitrarily being forced into one.

3.8. Calculation of Growth Parameters

Part of the utility of a chemically detailed cell model is that an engineer can design experiments that probe its behavior in response to various environmental and genetic manipulations. The MCM can serve as a platform to evaluate and test the plausibility of candidate minimal gene sets, as it does in the work presented here. One way to perform such a test is to compare the model predictions to those for general chemoheterotrophic bacteria.

Table 5
Initial conditions of groups of macromolecules in the minimal cell model

Class	Parameter	Symbol	*E. coli*	MCM
I	Deoxyribonucleotide residues per genome kbp/genome	kbo/genome	4,700	233
	Ribonucleotide residues per 70S ribosome	nucl/rib	4,566	4,546
	Amino acid residues per 70S ribosome	aa/rib	7,336	6,856
	Ribonucleotide residues per tRNA	nucl/tRNA	80	77
	Amino acid residues per RNA polymerase core	aa/pol	3,407	3,010
II	Fraction of total RNA that is stable RNA	f_{sRNA}	0.98	0.96
	Fraction of stable RNA that is tRNA	f_{tRNA}	0.14	0.15
	Fraction of active ribosomes	$frac_{rt}$	0.921	0.797
III	Fraction of total protein that is r-protein	α_r	0.09–0.22	0.12
	Fraction of total protein that is RNA polymerase	α_p	0.009–0.01	0.03
IV	Peptide chain elongation rate	C_p	12–22 aa/s	23 aa/s
	DNA chain elongation rate	C_d	500–830 nucl bp/s	184 nucl bp/s
V	Time to replicate the chromosome	C	40–67 min	21.1 min
	Time between termination of replication and division	D	20.2 min	19.5 min

The average masses from *E. coli* are based on values reported in Neidhardt et al. (87). The average mass in the MCM is calculated by assuming that each component accounts for the same mass percentage in *E. coli* and the minimal cell, but that the total average mass of the minimal cell is 0.2 pg. Note that the actual average value of DNA used in the MCM is based on its genome sequence, not on the data from *E. coli* presented in this table. In the current model the mass of the chromosome is $M_{CHR} \sim 3.77 \times 10^{-4}$ pg. Table reused, with permission, from ref. (95)

While there is not a biological analog of the MCM, it is comparable to a generalized chemoheterotrophic bacterial cell (73, 74). Table 6 contains calculated growth and molecular composition parameters obtained using the MCM. These values are compared to values for *E. coli* (88). In Table 6, genomic sequence measurements are based on values from *Mycoplasma* and other organisms listed in the KEGG database (89). Parameters in class I are inputs to the model (e.g., the number of deoxyribonucleotide residues per genome is fixed by the sequences of the genes in the minimal gene set). Parameters in classes II–V are outputs from the model simulation, except for C_p, which is an input constant based on our previous model of *E. coli* (61). The five classes in Table 6 are defined as:

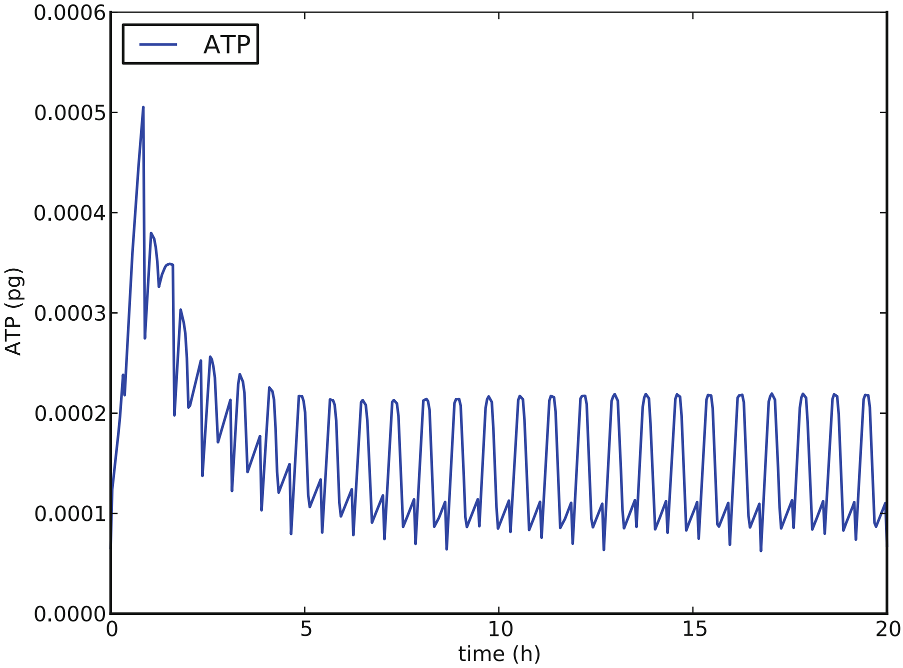

Fig. 5. The approach to steady-state for a MCM. The trajectory shown is for the mass of ATP over time, but any chemical defined in the MCM can be output. The sudden periodic halving in the mass of ATP corresponds to the moment of cell division, when all masses in the cell are instantaneously halved.

Table 6
Parameters related to the growth and molecular composition of the minimal cell model

Component	Avg. mass in *E. coli* (pg)	Avg. mass in MCM (pg)
Protein	1.56×10^{-1}	1.20×10^{-1}
rRNA	4.77×10^{-2}	3.68×10^{-2}
tRNA	6.33×10^{-3}	6.33×10^{-3}
mRNA	2.10×10^{-3}	1.62×10^{-3}
DNA	9.00×10^{-3}	6.95×10^{-3}
Lipid	2.60×10^{-2}	2.01×10^{-2}
Metabolites	1.00×10^{-2}	7.72×10^{-3}

This table is modeled after Table 20.1 from ref. (88). See the main text for a definition of parameter classes I–V. Table reused, with permission, from ref. (95)

1. Structural parameters that do not vary with growth rate. These parameters are calculated from the genome/proteome sequence of the minimal cell.
2. Partition parameters are essentially invariant. The values presented are typical values for the model and are close to those for *E. coli* presented by (88).
3. Other partition parameters expected to vary with the growth rate. The values presented here are for a minimal cell with growth rate equal to 0.86 h^{-1}.
4. Kinetic parameters describing functional activities. The peptide chain elongation rate, C_p, is a constant parameter of the model, which we chose to match the value used by (61). The DNA chain elongation rate, C_d, is calculated by dividing the chromosome length by the period of time it takes to replicate the chromosome during the simulation (the *C* period).
5. Chromosome replication and cell division parameters calculated by the simulation.

There are many common features between the *E. coli* data and the MCM (e.g., fraction of active ribosomes, or DNA chain elongation rate). However, some calculations from the MCM do not match the data from *E. coli* due to the nature of a minimal cell. In class I, e.g., the deoxyribonucleotide residues per genome will be lower in the MCM because it is a model of a cell defined by its low number of genes. Slight differences in the sequence lengths for ribosomes, tRNAs, and RNA polymerase occur due to sequence differences between *E. coli* and the source organisms used for the MCM. The partition factors (classes II and III) show strong agreement between *E. coli* and the MCM, and one would expect these features to hold constant amongst many bacterial species. The peptide chain elongation rate, C_p, is in agreement with the high-end of the values for *E. coli*, but this quantity is actually an input to the model based on data for *E. coli* (59), so it is unsurprising that they concur. The DNA chain elongation rate, C_d, falls significantly below that of *E. coli*. *Mycoplasma* species tend to have slow DNA replication rates, e.g., 100 bp/s in *M. capricolum* (90), so it is not unexpected that a minimal cell would also have slower DNA replication rates. However, because of its minimized chromosome, the MCM actually exhibits a shorter *C*-period (24–25 min) than *E. coli*. Finally, the *D*-period, the time between replication termination and cell division, for the MCM and for *E. coli* is similar (20.2 min for *E. coli* vs. 19.6 min for the MCM).

3.9. Response to Environmental Conditions

The MCM connects the physiology of the minimal cell directly to its environment. The MCM could be used to guide the development of an appropriate nutrient media for synthetic cells. Except for inorganic ions, which are not tracked in the MCM, removing any of the external nutrients causes the cell to fail. To further study the effect of environmental nutrient modifications, model cells growing at steady-state were exposed to step-changes in the external concentration of arginine, a competitive inhibitor of transport for other amino acids. Transport systems with multiple substrates are subject to competitive inhibition (91). To reduce the total number of genes as much as possible, several transporters with broad specificity were included in the MCM. For example, the Bgt transport system, an ATP-binding-cassette (ABC) dimer found in *Synechocystis* sp., is known to transport alanine, glutamine, glycine, leucine, proline, and serine (92). The MCM accounts for multiple substrate inhibition using Michaelis–Menten competitive inhibition terms. Each transport rate law has one inhibition term for each alternative substrate. For example, a transporter that carries four substrates will have three external inhibition multipliers for each of its transport rate laws.

Thus, the concentrations of some substances cannot be arbitrarily increased because at some level they inhibit growth by causing the cell to be starved of another nutrient. To exemplify the effect of competitive substrate inhibition on the viability of the MCM, the external concentration of arginine was increased 5×, 10× and 15× (Fig. 6). Arginine is transported into the cell by the Nat transport system of *Synechocystis* sp., which also transports histidine and lysine (92). The rate of histidine uptake is described in Eqs. 12 and 13.

$$R_{\mathrm{His}} = v_{\mathrm{R-His}} \cdot K_{\mathrm{sat-His-ext}} \cdot K_{\mathrm{sat-ATP}} \cdot K_{\mathrm{i-His}} \cdot K_{\mathrm{i-R-His}} \cdot T_{\mathrm{Nat}} \tag{12}$$

$$K_{\mathrm{i-R-His}} = \frac{K_{\mathrm{i-R-His-Arg-ext}}}{K_{\mathrm{i-R-His-Arg-ext}} + \mathrm{Arg}_{\mathrm{ext}}} \cdot \frac{K_{\mathrm{i-R-His-Lys-ext}}}{K_{\mathrm{i-R-His-Lys-ext}} + \mathrm{Lys}_{\mathrm{ext}}} \tag{13}$$

In Eq. 12, R_{His} describes the rate of histidine uptake (pg/h), $v_{\text{R-His}}$ is the rate constant for histidine uptake (pg His/h/pg T_{Nat}), $K_{\text{sat-His-ext}}$ and $K_{\text{sat-ATP}}$ are dimensionless Michaelis–Menten saturation terms for external histidine and cellular ATP, respectively, $K_{\text{i-His}}$ is a dimensionless Michaelis–Menten product inhibition constant for cellular histidine, $K_{\text{i-R-His}}$ is a dimensionless competitive inhibition term defined in Eq. 13, and T_{Nat} is the mass of transporter T_{Nat} (pg). In Eq. 13, $K_{\text{i-R-His-Arg-ext}}$ and $K_{\text{i-R-His-Lys-ext}}$ are inhibition constants (g/mL) that describe transport inhibition by arginine and lysine, respectively on the histidine transport reaction. Based on these equations, it is expected that the transport rate for histidine will drop as either arginine or lysine is introduced into the medium. Figure 6 demonstrates such an effect, with arginine values becoming inhibitory somewhere between the 10× and 15× increase of the default concentration (Fig. 6c, d). This shows that there is an intermediate transition nutrient concentration where the cell transitions between life and death.

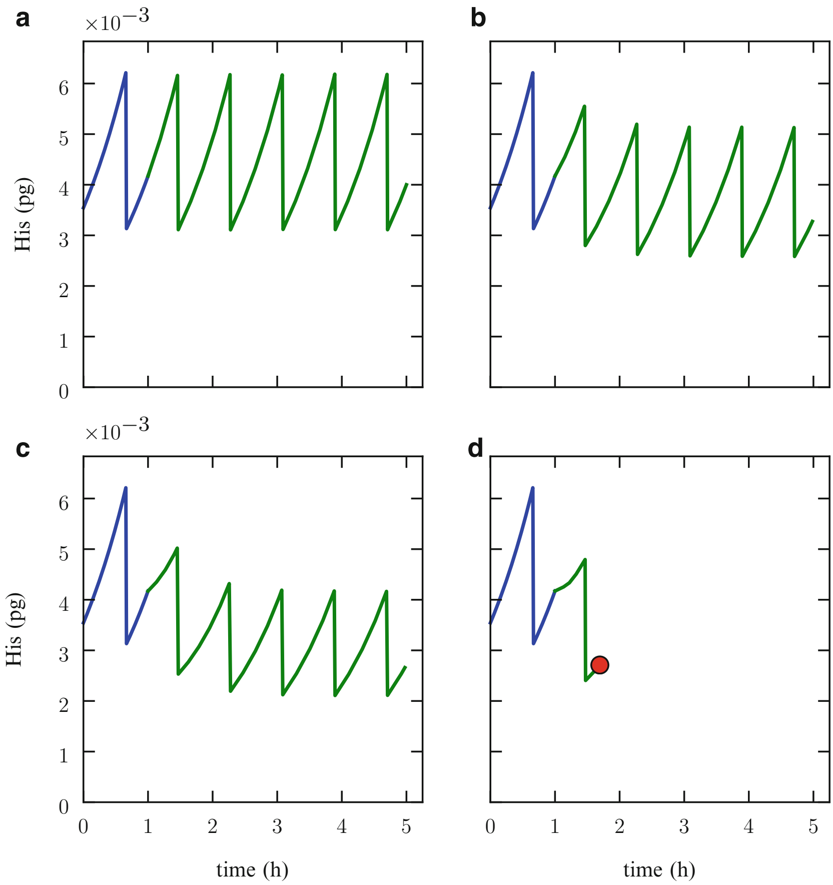

Fig. 6. Effect of amino acid inhibition on histidine (His) mass and cell viability in response to increases in extracellular arginine (Arg). *Blue* trajectories are the unaltered histidine mass over time, while the *green* trajectories represent the histidine mass after changes (**a**–**d**). *Red dots* represent the time and state of cell death. (**a**) Default trajectory. (**b**) 5× increase in the external concentration of arginine. (**c**) 10× increase in the external concentration of arginine. (**d**) 15× increase in the external concentration of arginine.

4. Notes

1. The gene dosage for each gene in the MCM was calculated using functions of the form:

$$\mathrm{HF(FP, gp)} = \frac{1}{\left(1 + \mathrm{e}^{-200\cdot(\mathrm{FP}-\mathrm{gp})}\right)}$$

where HF is the heavy-step function, FP is the fork position (a function of time), and gp is the position of the gene on the chromosome (from 0 to 1). This function approximates a discrete change in gene dosage without slowing down the integration for the firing of many events.

2. The space of possible rate constant choices is a many dimensional space and there can be infinitely many sets of constants that would satisfy the given constraints. The objective function is minimized because the constraints placed on the reaction rate constants (doubling all chemical species masses) tend to force the system to have higher rate constants. To balance these constraints and estimate reasonably sized rate constants, their sum is minimized. The LP system is solved using the Python lpsolve package (93). A wrapper class for lpsolve is included with the MCM code.

3. Every event is associated with event assignments that can both specify the physiological effect of the event and set tracking parameters to measure statistics about the cell cycle progress (e.g., time for chromosome replication). Some event assignments associated with DNA initiation are listed here as examples.

   After DNA replication commences, it is assumed that the proteins bound to the Ori are rapidly forced off by the opening of the chromosome replication fork. Thus, we include, e.g., an event assignment for unbinding of DnaG protein from the origin of replication,

   $$\mathrm{DnaG}_{\mathrm{boundto-Ori}} \rightarrow 0,$$

   as well as an event assignment for renewal of the cytosolic DnaG pool.

   $$\mathrm{DnaG} \rightarrow \mathrm{DnaG} + \mathrm{DnaG}_{\mathrm{boundto-Ori}} \cdot \mathrm{Ori}_{\mathrm{GD}}$$

   Some event assignments reflect changes in the cell's state. For example, setting a flag variable to indicate that the chromosome is no longer methylated indicates that the chromosome is not immediately ready to start another round of initiation.

   $$\mathrm{flag}_{\mathrm{meth}} \rightarrow 0$$

   Other event assignments are updates of bookkeeping parameters. For example, $t_{\mathrm{DNA\text{-}init}}$ tracks when chromosome replication initiation occurs.

   $$t_{\mathrm{DNA-init}} \rightarrow \mathrm{time}$$

4. To derive initial values for chemical masses, the following procedure was used (M. Domach, Carnegie Mellon University, personal communication, October 17, 2007):

   (a) The minimal cell is assumed to have an average dry mass of about 0.2 pg, which is about 75% of the dry weight of *E. coli* (87).

   (b) Data for the average composition of protein, mRNA, tRNA, rRNA, DNA, lipids, and metabolites in *E. coli*, was gathered (87). These weight fractions were assumed to be the same for the MCM.

(c) Cell age is defined as age $= t/\tau_{\mathrm{D}}$, where t is the time since the last division and τ_{D} is the steady-state doubling time. A steady-state growth rate μ_{g} is also defined. The age distribution, $\phi(\mathrm{age})$, for a culture in continuous steady-state growth with a constant τ_{D} was derived by (94) as

$$\phi(\mathrm{age}) = 2\mu_{\mathrm{g}} e^{-\ln(2)\cdot \mathrm{age}}$$

We find the average age of a culture (i.e., the 50th percentile), by solving the following equation for age_{50}.

$$\int_0^{\mathrm{age}_{50}} \phi(\mathrm{age})d(\mathrm{age}) = 0.5$$

This yields that the average age of a synchronized, exponentially growing cell population (i.e., age_{50}) is approximately $0.415{}^{*}t_{\mathrm{D}}$.

(a) Assuming the cell is in balanced growth, the population weighted average mass of a chemical species X in the cell will correspond to when the cell is 41.5% of the way through the division cycle. Thus, the initial mass can be calculated from the average mass using the following relations:

$$X = X_0 e^{(\ln(2)\cdot 0.4_{15})}$$

$$X = 1.33 \cdot X_0$$

(b) The average mass of each of the protein, mRNA, tRNA, rRNA, and metabolites groups, was set to be equal to the mass fraction calculated in step b times the total mass selected in step a. Then, the mass at the start of the cell cycle was assumed to be the average value divided by 1.33.

(c) The initial mass of DNA was set to the mass of one complete chromosome, which was based on the mass of the sequence of the minimal gene set.

(d) The initial mass of membrane lipids was set to be adequate to “envelope” the cytoplasm of the cell.

Acknowledgements

JA gratefully acknowledges funding from the DOE Computational Science Graduate Fellowship Program (CSGF) of the Office of Science and National Nuclear Security Administration in the DOE under contract DE-FG02-97ER25308.

References

1. Agapakis CM, Silver PA (2009) Synthetic biology: exploring and exploiting genetic modularity through the design of novel biological networks. Mol Biosyst 5(7):704–713. doi: 10.1039/b901484e, http://dx.doi.org/10.1039/b901484e
2. Drubin DA, Way JC, Silver PA (2007) Designing biological systems. Genes Dev 21 (3):242–254. doi: 10.1101/gad.1507207, http://dx.doi.org/10.1101/gad.1507207
3. Purnick PEM, Weiss R (2009) The second wave of synthetic biology: from modules to systems. Nat Rev Mol Cell Biol 10 (6):410–422. doi: 10.1038/nrm2698, http://dx.doi.org/10.1038/nrm2698
4. Leonard E et al (2008) Engineering microbes with synthetic biology frameworks. Trends Biotechnol 26(12):674–681. doi: 10.1016/j.tibtech.2008.08.003, http://dx.doi.org/10.1016/j.tibtech.2008.08.003
5. Loeb J (1906) The dynamics of living matter. Macmillan, New York, NY
6. Pohorille A, Deamer D (2002) Artificial cells: prospects for biotechnology. Trends Biotechnol 20(3):123–128
7. Rasmussen S et al (2004) Evolution. Transitions from nonliving to living matter. Science 303(5660):963–965. doi: 10.1126/science.1093669, http://dx.doi.org/10.1126/science.1093669
8. Hanczyc MM, Szostak JW (2004) Replicating vesicles as models of primitive cell growth and division. Curr Opin Chem Biol 8(6):660–664. doi: 10.1016/j.cbpa.2004.10.002, http://dx.doi.org/10.1016/j.cbpa.2004.10.002
9. Luisi PL, Ferri F, Stano P (2006) Approaches to semi-synthetic minimal cells: a review. Naturwissenschaften 93(1):1–13. doi: 10.1007/s00114-005-0056-z, http://dx.doi.org/10.1007/s00114-005-0056-z
10. Segré D et al (2001) The lipid world. Orig Life Evol Biosph 31(1–2):119–145
11. Forster AC, Church GM (2006) Towards synthesis of a minimal cell. Mol Syst Biol 2:45
12. Zimmer C (2003) Genomics—Tinker, tailor: can Venter stitch together a genome from scratch? Science 299(5609):1006–1007
13. Morowitz HJ (1984) The completeness of molecular-biology. Isr J Med Sci 20(9):750–753
14. Moya A et al (2009) Toward minimal bacterial cells: evolution vs. design. FEMS Microbiol Rev 33(1):225–235. doi: 10.1111/j.1574-6976.2008.00151.x, http://dx.doi.org/10.1111/j.1574-6976.2008.00151.x
15. Lartigue C et al (2007) Genome transplantation in bacteria: changing one species to another. Science 317(5838):632–638. doi: 10.1126/science.1144622, http://dx.doi.org/10.1126/science.1144622
16. Gibson DG et al (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. Science 319 (5867):1215–1220. doi: 10.1126/science.1151721, http://dx.doi.org/10.1126/science.1151721
17. Lartigue C et al (2009) Creating bacterial strains from genomes that have been cloned and engineered in yeast. Science 325 (5948):1693–1696. doi: 10.1126/science.1173759, http://dx.doi.org/10.1126/science.1173759
18. Gibson DG et al (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329(5987):52–56. doi: 10.1126/science.1190719, http://dx.doi.org/10.1126/science.1190719
19. Waters E et al (2003) The genome of nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci USA 100(22):12984–12988
20. Gil R et al (2002) Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA 99(7):4454–4458
21. Nakabachi A et al (2006) The 160-kilobase genome of the bacterial endosymbiont carsonella. Science 314(5797):267. doi: 10.1126/science.1134196, http://dx.doi.org/10.1126/science.1134196
22. Fraser CM et al (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270(5235):397–403
23. Blattner FR et al (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277 (5331):1453–1474
24. Maniloff J (1996) The minimal cell genome: "on being the right size". Proc Natl Acad Sci USA 93(19):10004–10006
25. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93(19):10268–10273
26. Hutchison CA et al (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 286(5447):2165–2169
27. Koonin EV (2000) How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet 1:99–116

28. Kobayashi K et al (2003) Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA 100 (8):4678–4683
29. Gil R et al (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68(3):518–537
30. Glass JI et al (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103 (2):425–430
31. Tomita M et al (1999) E-CELL: software environment for whole-cell simulation. Bioinformatics 15(1):72–84
32. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1 (2):127–136
33. Luisi PL (2002) Toward the engineering of minimal living cells. Anat Rec 268(3):208–214
34. Lamichhane G et al (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to mycobacterium tuberculosis. Proc Natl Acad Sci USA 100(12):7213–7218
35. Itaya M (1995) An estimation of minimal genome size required for life. FEBS Lett 362 (3):257–260
36. Forsyth RA et al (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. Mol Microbiol 43 (6):1387–1400
37. Gerdes SY et al (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol 185(19):5673–5684
38. Peterson SN, Fraser CM (2001) The complexity of simplicity. Genome Biol 2(2):1–8. http://genomebiology.com/2001/2/2/comment/2002
39. Nesbø CL, Boucher Y, Doolittle WF (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol 53(4–5):340–350. doi: 10.1007/s002390010224, http://dx.doi.org/10.1007/s002390010224
40. Harris JK et al (2003) The genetic core of the universal ancestor. Genome Res 13 (3):407–412. doi: 10.1101/gr.652803, http://dx.doi.org/10.1101/gr.652803
41. Gil R et al (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci USA 100(16):9388–9393
42. Pál C et al (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667–670. doi: 10.1038/nature04568, http://dx.doi.org/10.1038/nature04568
43. Gabaldón T et al (2007) Structural analyses of a hypothetical minimal metabolism. Philos Trans R Soc Lond B Biol Sci 362(1486):1751–1762. doi: 10.1098/rstb.2007.2067, http://dx.doi.org/10.1098/rstb.2007.2067
44. Carbone A (2006) Computational prediction of genomic functional cores specific to different microbes. J Mol Evol 63(6):733–746. doi: 10.1007/s00239-005-0250-9, http://dx.doi.org/10.1007/s00239-005-0250-9
45. Forster AC, Church GM (2007) Synthetic biology projects *in vitro*. Genome Res 17 (1):1–6. doi: 10.1101/gr.5776007, http://dx.doi.org/10.1101/gr.5776007
46. Azuma Y, Ota M (2009) An evaluation of minimal cellular functions to sustain a bacterial cell. BMC Syst Biol 3:111. doi: 10.1186/1752-0509-3-111, http://dx.doi.org/10.1186/1752-0509-3-111
47. Foley PL, Shuler ML (2010) Considerations for the design and construction of a synthetic platform cell for biotechnological applications. Biotechnol Bioeng 105(1):26–36. doi: 10.1002/bit.22575, http://dx.doi.org/10.1002/bit.22575
48. Karp PD et al (2004) The *E-coli* ecocyc database: no longer just a metabolic pathway database. ASM News 70(1):25–30
49. Burgard AP, Maranas CD (2001) Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. Biotechnol Bioeng 74(5):364–375
50. Burgard AP, Vaidyaraman S, Maranas CD (2001) Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. Biotechnol Prog 17(5):791–797
51. Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97(10):5528–5533
52. Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. Environ Microbiol 4(3):133–140
53. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev 33(1):164–190. doi: 10.1111/j.1574-6976.2008.00146.x, http://dx.doi.org/10.1111/j.1574-6976.2008.00146.x
54. Chassagnole C et al (2002) Dynamic modeling of the central carbon metabolism of *Escherichia coli*. Biotechnol Bioeng 79(1):53–73
55. Tomita M (2001) Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol 19(6):205–210

56. Schlosser PM, Bailey JE (1990) An integrated modeling-experimental strategy for the analysis of metabolic pathways. Math Biosci 100 (1):87–114

57. Shuler ML, Dick C (1979) A mathematical model for the growth of a single bacterial cell. Ann N Y Acad Sci 326:35–55

58. Bailey JE (1998) Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. Biotechnol Prog 14(1):8–20. doi: 10.1021/bp9701269, http://dx.doi.org/10.1021/bp9701269

59. Domach MM (1983) Refinement and use of a structured model of a single cell of *Escherichia coli* for the description of ammonia-limited growth and asynchonous population dynamics. Ph.D. thesis. Cornell University

60. Shuler ML (1999) Single-cell models: promise and limitations. J Biotechnol 71(1–3):225–228

61. Domach MM, Shuler ML (1984) Testing of a potential mechanism for *Escherichia coli* temporal cycle imprecision with a structural model. J Theor Biol 106(4):577–585

62. Lee AL, Ataai MM, Shuler ML (1984) Double-substrate-limited growth of *Escherichia coli*. Biotechnol Bioeng 26(11):1398–1401

63. Shuler ML, Domach MM (1983) Mathematical-models of the growth of individual cells—tools for testing biochemical-mechanisms. ACS Symp Ser 207:93–133

64. Browning ST, Castellanos M, Shuler ML (2004) Robust control of initiation of prokaryotic chromosome replication: essential considerations for a minimal cell. Biotechnol Bioeng 88(5):575–584. doi: 10.1002/bit.20223, http://dx.doi.org/10.1002/bit.20223

65. Atlas JC et al (2008) Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of *Escherichia coli*: application to DNA replication. IET Syst Biol 2 (5):369–382. doi: 10.1049/iet-syb:20070079, http://dx.doi.org/10.1049/iet-syb:20070079

66. Nikolaev E, Atlas J, Shuler ML (2006) Computer models of bacterial cells: from generalized coarse-grained to genome-specific modular models. J Phys Conf Ser 46:322–326

67. Shu J, Shuler ML (1991) Prediction of effects of amino-acid supplementation on growth of *Escherichia coli* B/r. Biotechnol Bioeng 37 (8):708–715

68. Laffend L, Shuler ML (1994) Ribosomal-protein limitations in *Escherichia coli* under conditions of high translational activity. Biotechnol Bioeng 43(5):388–398

69. Laffend L, Shuler ML (1994) Structured model of genetic-control via the lac promoter in *Escherichia coli*. Biotechnol Bioeng 43 (5):399–410

70. Kim BG et al (1987) Growth-behavior and prediction of copy number and retention of cole1-type plasmids in *Escherichia-coli* under slow growth-conditions. Ann N Y Acad Sci 506:384–395

71. Kim BG, Shuler ML (1990) A structured, segregated model for genetically modified *Escherichia coli* cells and its use for prediction of plasmid stability. Biotechnol Bioeng 36 (6):581–592

72. Kim BG, Shuler ML (1991) Kinetic-analysis of the effects of plasmid multimerization on segregational instability of cole1 type plasmids in *Escherichia coli* B/R. Biotechnol Bioeng 37 (11):1076–1086

73. Browning ST, Shuler ML (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(3):187–192

74. Castellanos M, Wilson DB, Shuler ML (2004) A modular minimal cell model: purine and pyrimidine transport and metabolism. Proc Natl Acad Sci USA 101(17):6681–6686. doi: 10.1073/pnas.0400962101, http://dx.doi.org/10.1073/pnas.0400962101

75. Castellanos M et al (2007) A genomically/chemically complete module for synthesis of lipid membrane in a minimal cell. Biotechnol Bioeng 97(2):397–409. doi: 10.1002/bit.21251, http://dx.doi.org/10.1002/bit.21251

76. Gutenkunst RN et al (2007) Extracting falsifiable predictions from sloppy models. Ann N Y Acad Sci 1115:203–211. doi: 10.1196/annals.1407.003, http://dx.doi.org/10.1196/annals.1407.003

77. Labarère J (1992) DNA replication and repair. In: Maniloff J, McElhaney R, Finch L, Baseman J (eds) Mycoplasmas molecular biology and pathogenesis. American Society for Microbiology, Washington, DC, pp 23–40

78. Capaldo-Kimball F, Barbour SD (1971) Involvement of recombination genes in growth and viability of *Escherichia coli* k-12. J Bacteriol 106(1):204–212

79. Bramhill D (1997) Bacterial cell division. Annu Rev Cell Dev Biol 13:395–424. doi: 10.1146/annurev.cellbio.13.1.395, http://dx.doi.org/10.1146/annurev.cellbio.13.1.395

80. Hutkins RW, Nannen NL (1993) pH homeostasis in lactic acid bacteria. J Dairy Sci 76:2354–2365

81. Reynolds CM, Meyer J, Poole LB (2002) An NADH-dependent bacterial thioredoxin reductase-like protein in conjunction with a glutaredoxin homologue form a unique peroxiredoxin (AhpC) reducing system in *Clostridium pasteurianum*. Biochemistry 41(6):1990–2001
82. Kuruma Y et al (2009) A synthetic biology approach to the construction of membrane proteins in semi-synthetic minimal cells. Biochimica et Biophysica Acta 1788(2):567–574. doi: 10.1016/j.bbamem.2008.10.017, http://dx.doi.org/10.1016/j.bbamem.2008.10.017
83. Sirand-Pugnet P et al (2007) Evolution of mollicutes: down a bumpy road with twists and turns. Res Microbiol 158(10):754–766. doi: 10.1016/j.resmic.2007.09.007, http://dx.doi.org/10.1016/j.resmic.2007.09.007
84. Brown KS, Sethna JP (2003) Statistical mechanical approaches to models with many poorly known parameters. Phys Rev E 68(2)
85. Gutenkunst RN et al (2007) Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 3(10):1871–1878. doi: 10.1371/journal.pcbi.0030189, http://dx.doi.org/10.1371/journal.pcbi.0030189
86. Hucka M et al (2008) Systems biology markup language (SBML) level 2: structures and facilities for model definitions. Nat Proc. doi: doi.org/10.1038/npre.2008.2715.1, http://dx.doi.org/10.1038/npre.2008.2715.1
87. Neidhardt FC, et al (1996) Chemical Composition of *Escherichia coli*, in Escherichia coli and Salmonella: cellular and molecular biology, 2nd edn., vol. 1 ASM Press, Washington, D.C., pp 13–16
88. Bremer H, Dennis P (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington
89. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
90. Seto S, Miyata M (1998) Cell reproduction and morphological changes in *Mycoplasma capricolum*. J Bacteriol 180(2):256–264
91. Cheng Y, Prusoff WH (1973) Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem Pharmacol 22(23):3099–3108
92. Quintero MJ et al (2001) Identification of genes encoding amino acid permeases by inactivation of selected ORFs from the synechocystis genomic sequence. Genome Res 11 (12):2034–2040
93. Berkelaar M, Eikland K, Notebaert P (2010) lpsolve—open source (mixed-integer) linear programming system, version 5.1.0.0. http://lpsolve.sourceforge.net/
94. Powell EO (1956) Growth rate and generation time of bacteria, with special reference to continuous culture. J Gen Microbiol 15 (3):492–511
95. Atlas JC (2010) Simulation of a whole-cell with the minimum number of genes necessary for sustained replication. Ph.D. thesis. Cornell University
96. Hucka M et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19 (4):524–531