SIMULATING THE GROWTH OF VIRUSES

LINGCHONG YOU AND JOHN YIN*

Department of Chemical Engineering, University of Wisconsin-Madison
1415 Engineering Drive, Madison, WI 53706-1691 USA

To explore how the genome of an organism defines its growth, we have developed a computer simulation for the intracellular growth of phage T7 on its *E. coli* host. Our simulation, which incorporates 30 years of genetic, biochemical, physiological, and biophysical data, is used here to study how the intracellular resources of the host, determined by the specific growth rate of the host, contribute toward phage development. It is also used to probe how changes in the linear organization of genetic elements on the T7 genome can affect T7 development. Further, we show how time-series trajectories of T7 mRNA and protein levels generated by the simulation may be used as raw data to test data-mining strategies, specifically, to identify partners in protein-protein interactions. Finally, we suggest how generalization of this work can lead to a knowledge-driven simulation for the growth of any virus.

Introduction

Because of their small size and the central role they played in setting the foundations of molecular genetics, viruses are ideal model systems to explore how minimal genomes define processes of growth and development. The genomes of most viruses, which may be composed of single- or double-stranded DNA or RNA, range from five to 200 kilobases in length and encode from five to 300 genes. For well-studied viruses such as phage lambda, phage T7, or human immunodeficiency virus (HIV-1), all genes that are essential for growth are known and key functions have been characterized.

The T7 model

We study the growth of phage T7 on *E. coli* because of its 30-year foundation of genetic, physiological, biochemical, and biophysical data [1-3]. The double-stranded DNA genome of T7 is 39,937 base-pairs long, and it contains 56 genes encoding 59 proteins, five *E. coli* RNA polymerase (RNAP) promoters, 17 T7 RNAP promoters, two transcriptional terminators, and ten RNase III splicing sites (Figure 1). Using existing mechanisms and parameters for the constituent processes of wild-type T7 development, we have developed a computer simulation for the intracellular growth cycle of wild-type phage T7 in a single *E. coli* cell [4-6]. By defining and numerically solving a system of coupled differential and algebraic equations, the model accounts for all the major steps of T7 infection: *E. coli* RNAP- and T7 RNAP-mediated entry of the T7 genome into its host, transcription of T7 mRNAs,

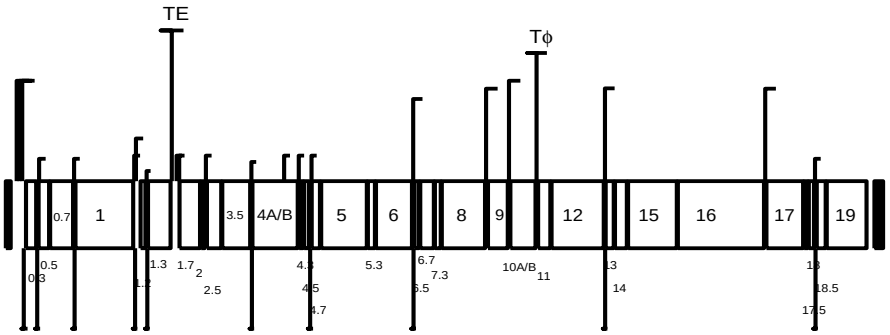

Figure 1. The wild-type phage T7 genome. Boxes represent genes. Vertical lines with half bars above the genes represent EcRNAP and T7RNAP promoters; the heights of these lines represent the relative activities of the promoters. Vertical lines with full bars above the genes represent EcRNAP (TE) and T7RNAP (Tϕ) terminators. Vertical lines below the genes represent RNAaseIII processing sites. The two black bars at both ends represent the left end and the right end of the genome. Adapted from reference [18]

regulation of transcription by promoter strengths and protein-protein interactions, translation of T7 proteins, replication of the T7 genome, and assembly and maturation of T7 progeny (Figure 2). As detailed elsewhere [6], the current version of the T7 model (T7v2.5)[α] is formulated using an object-oriented approach.

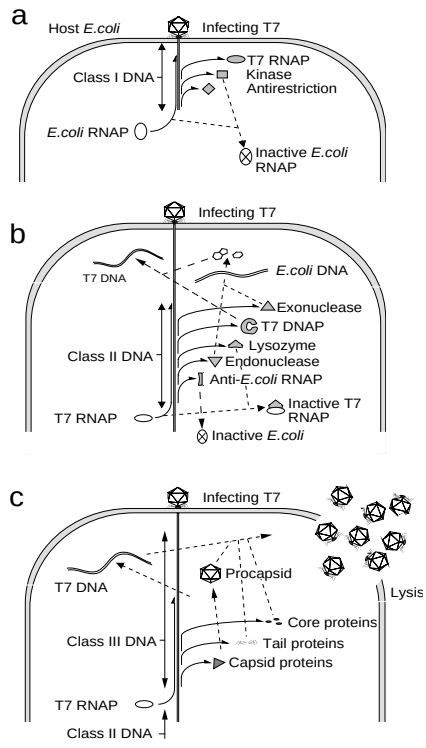

Figure 2. Intracellular growth cycle of phage T7. The solid lines with half arrows indicate transcription and translation, the dashed lines represent reactions, and the solid lines with full arrows indicate the three classes of T7 DNA. (a) Infection initiation, class I gene expression. (b) Class II gene expression, phage DNA replication. (c) Class III gene expression, procapsid assembly, phage maturation, and lysis. RNAP: RNA polymerase; DNAP: DNA polymerase.

Simulating how host physiology affects T7 growth

To evaluate the global behavior of the simulation, we compare the predicted with measured intracellular one-step growth (OSG) curve of phage T7 (Figure 3). To facilitate comparison, we dissect each OSG curve into three stages (Figure 3a). When no parameter is adjusted, the simulation (μ = 1.2) appears to match the observed eclipse time, but it underestimates the rise rate significantly and overshoots the burst size by almost

[α] The source code of T7v2.5 can be obtained from http://virus.che.wisc.edu; the later versions will be available on request. A simulation can also be performed through the web interface of the model (http://virus.molsci.org/~t7).

two-fold. The mismatch for the burst size is most likely due to the absence of a mechanistic implementation for the lysis of the host cell, a process that remains to be characterized mechanistically. The cause of the mismatch for the rise rate is more complex. Inaccuracy in any parameters that may affect the rate of protein synthesis, the rate of procapsid assembly, or the rate of DNA packaging can be responsible for this mismatch. For simplicity, here we assume the mismatch is due to a systematic mismatch between the model "cell" implemented in the simulation and the actual host cell used in the experiment, and attempt to improve the agreement between simulation and experiment by changing the host growth rate alone. Because the simulation result for μ =1.5 1/hr overall matches the experiment the best for the pre-plateau region (Figure 3b), we use this setting as our default for other simulations unless otherwise indicated.

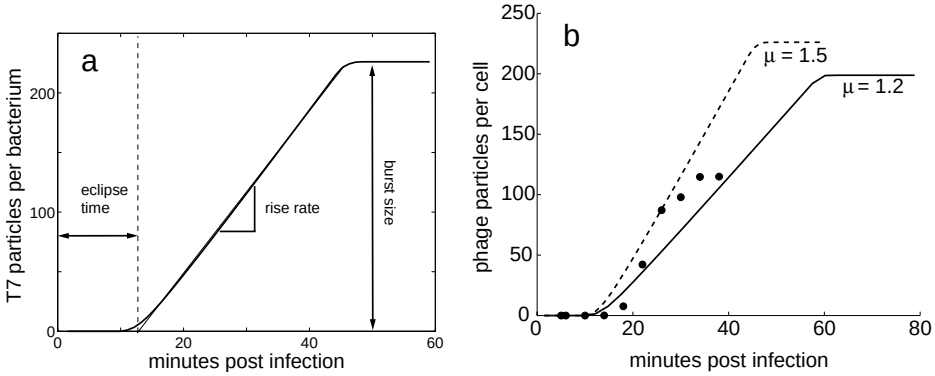

Figure 3. Intracellular one-step growth (OSG) of phage T7. (a) Characterization of an intracellular OSG curve by three parameters: the *eclipse time* is the time period between infection initiation and the time point when phage progeny begin to appear, the *rise rate* is the slope of the straight line starting from the end of the eclipse period and ending at the end of lysis, and the *burst size* is the final number of phage progeny being produced from a single infection. (b) Comparison of the simulated (lines) and experimental (dots) intracellular OSG curves. Experimental data were the average of two separate experiments, both performed with host cells growing in LB at 30°C with a measured growth rate (μ) of 1.2 doublings per hour. Simulated results are shown for μ =1.2 and 1.5 doublings per hour. In the experiment, host cells were artificially lysed using chloroform to liberate the intracellular phage particles. The simulated eclipse time was calculated as the time required for the first viral progeny to appear; the experimental eclipse time was estimated as the time point corresponding to the intersection of the straight line (fitted to the data points corresponding to the rise phase) crossing the time axis.

Experiments have shown that phage growth proceeds more rapidly on faster growing hosts [7-12]. However, it is not known which specific resources of physiologically active hosts have the greatest influence. From the fundamental perspective, identifying key resources may provide insight into phage growth at sub-optimal conditions, such as those that exist in nature [11]. From an applied view, such understanding may assist our evaluation of phage therapies against

antibiotic-resistant bacteria [13].

To study how host physiology affects phage growth we developed empirical expressions or employed existing mechanistic expressions [14, 15] that coupled the specific growth rate of the host with the pool sizes and elongation rates of its RNA polymerases and ribosomes, pool sizes of amino acids and NTPs, the cellular content of host genomic DNA, and the host-cell volume. These functions enabled us to evaluate the sensitivity of the eclipse time and the rise rate to the host-cell growth rate (Figure 4). They fell and rose, respectively, as the growth rate of the host increased, in qualitative agreement with experimentally observed trends for other phage [12]. More extensive sensitivity analysis on the uncoupled host parameters has revealed that T7 growth is limited by the level and processivity of the host translation machinery [6].

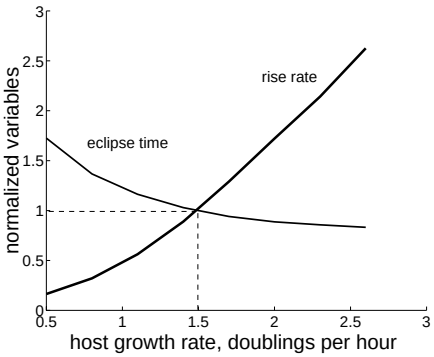

Figure 4. Effect of the host growth rate on T7 growth. The eclipse time and rise rate as the function of the growth rate of the host cell growing at exponential phase on T7 infection. Both the eclipse time and the rise rate are normalized to the corresponding values for a host growth rate of 1.5 doublings per hour.

Genome organization affects T7 growth

To develop safe live-virus vaccines, one seeks to retain the immune-system provoking virus attributes while irreversibly attenuating their disease-associated replicative abilities. Repositioning the nucleocapsid gene of vesicular stomatitis virus reduced virus productivity and lethality, supporting the idea that the reorganization of viral genomes may serve as the basis for the development of safe attenuated vaccines [16, 17]. To better understand fundamentally how genome rearrangements may affect virus growth, we generalized our simulation to calculate the behavior of the process defined by any linear order of T7 genes [18]. Here we compute the effect on growth of repositioning the gene *1* element, which encodes the T7 RNAP, at each possible element position on an otherwise wild-type genome. Gene *1* is chosen for this analysis because it plays a central role in T7 development. T7 RNAP is responsible for the entry of about 85% of the T7 genome and the transcription of class II and class III genes [2], and it is needed for T7 DNA replication as well [3]. The current simulation differs from the published one [18] in two aspects. First, we assumed unlimited host resources for phage development in the previous simulation, but used a host environment with limited resources in the

current simulation[β]. Second, as the objective functions to characterize intracellular T7 growth, we choose here to study the eclipse time and the rise rate, instead of the doubling rate [18], to better dissect the effects of repositioning of the different attributes of the growth curve.

The simulation result is shown along with available experimental data on three ectopic gene *1* mutants (Figure 5), whose experimental construction is detailed elsewhere [18]. When placed close to the left end of its genome, gene *1* cannot be expressed because it is upstream of all promoters. For these mutants, the eclipse times are infinite (not shown) and the rise rates are zero (Figure 5b). The simulation predicts that T7 growth is nearly optimal when gene *1* is located at its wild-type position. At the wild-type position T7 achieves nearly its shortest eclipse time (1.01 times the minimum) and largest rise rate (0.99 times the maximum).

When gene *1* is placed downstream of its wild-type position, its location has a dramatic impact on the eclipse time (Figure 5a) but a less obvious effect on the rise rate (Figure 5b). The nearly linear correlation between the eclipse time and the gene *1* position arises because positioning gene *1* further downstream delays its entry into the cell and its expression. Since T7 DNA replication also requires T7 RNAP, DNA replication is delayed and the time required to manufacture the first phage progeny is increased. The lack of an overarching trend for the rise rate dependence on gene *1* position indicates that once progeny are produced they can be produced with rise rates approaching wild type. Interestingly, in some cases a minor change in gene *1* position can cause a sudden change in the rise rate. For example, when gene *1* is moved from upstream to downstream of the T7 RNAP transcription terminator (Tϕ), the rise rate sharply increases from below 0.6 to nearly 1.0 (Figure 5b). This is consistent with the notion that T7 development is limited by the translation machinery. For example, when gene *1* is located just upstream of Tϕ, it is regulated by promoters ϕ*9* and ϕ*10*, which are among its strongest. Such a configuration creates unnatural autocatalytic loops for gene *1* expression leading to high levels of gene *1* mRNA that divert ribosomes away from the translation of the major capsid protein (gp10A) and the scaffold protein (gp9), components of the phage virion that are needed in large quantity. In contrast, positioning gene *1* after Tϕ still allows adequate T7 RNAP to be expressed without diversion of the ribosomes from the synthesis of the particle proteins. Hence, the rise rate returns to a nearly wild-type level.

The simulation matches the observed rise rates of ectopic mutants ecto*1.7* and ecto*3.8*, but it underestimates their eclipse times (Figure 5). The interruption by ecto*1.7* of gene *1.7*, a non-essential gene, may nevertheless influence growth [18].

[β] With the host parameters used in this simulation, the host cell represents the average of an *E. coli* population that grows at 1.5 doublings per hour prior to T7 infection.

Further, the simulation predicts well the eclipse time for ecto*12*, but it significantly overestimates the rise rate (Figure 5). This discrepancy may be due to our assumption that the host cell provides a constant environment during the entire infection cycle, an assumption that may not be valid for ecto*12*, where the growth cycle lasts 80 minutes [18], more than two-fold longer than the wild-type growth cycle.

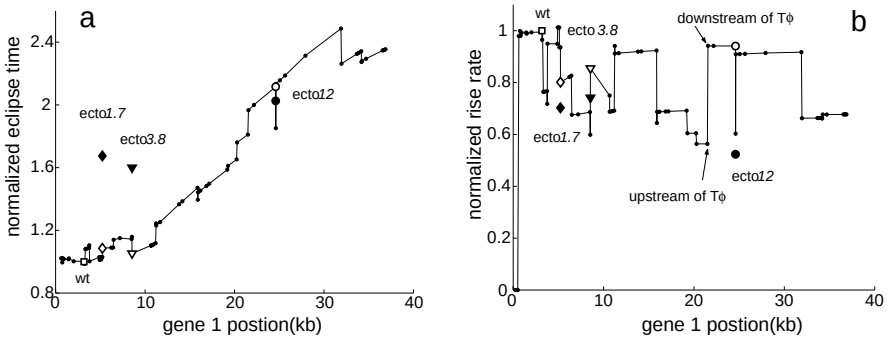

Figure 5. Effect of gene*1* position on (a) the eclipse time and (b) the rise rate of T7 infection. The experimental data on three ectopic gene*1* mutants are shown for comparison: square, wild type; diamond, ecto*1.7*; triangle, ecto*3.8*; circle, ecto*12*. The dots along the solid line and the open symbols represent the simulation predictions, and the filled symbols represent experimental data. Simulated results are normalized to the corresponding simulated values for the wild-type T7: 10.65 (minutes) for the eclipse time and 6.13 (particles/minute) for the rise rate. Experimental data are normalized to the experimental data for wild-type T7: 14.81 (minutes) for the eclipse time and 6.35 (particles/minute) for the rise rate.

Inferring protein-protein interactions from mRNA and protein time series

Our simulation of wild-type T7 growth calculates as a byproduct the time series of all T7 mRNAs and protein levels, which we have employed as raw data to test data-mining strategies, specifically, to identify partners of potential protein-protein interactions [19]. The need for such methodologies arises owing to the rapid growth of genomic databases and development of technologies for global monitoring of mRNA [20, 21] and protein levels [22].

We summarize here our methodology and provide details elsewhere [19]. Given the initial information about mRNA levels and the processing rates and distribution of ribosomes and activated tRNAs that constitute the translation resources, one can estimate how cytoplasmic levels of the corresponding proteins will change. However, factors other than translation may influence the actual protein levels observed. For example, processes of protein modification or degradation, transport from the cytoplasm, and process that form protein-protein complexes will tend to reduce a protein's level. When coupled with translation, different modes of depletion

will produce patterns of expression that may reflect aspects of the protein's function. A strong zero-time-lag correlation between depletions of two proteins further suggests a potential interaction between these two proteins.

Table 1. The essential concepts and their definitions of the correlated deviation algorithm.

Concept	Definition	Notes
Protein rate	$v_i \equiv dP_i / dt$	P_i is the level of protein i.
Translation rate	$v_{Ti} \equiv k_T R_i$	R_i is the level of the mRNAs encoding protein i.
Dynamic deviation factor (DDF)	$D_i \equiv (v_i - v_{Ti})/v_{Ti}$	DDF represents the relative deviation of the protein rate from the translation rate. It is equal to (when the protein is not depleted) or smaller than 0.
Protein correlation coefficient (PCC)	$C_{ij} \equiv \frac{\int_0^T D_i D_j dt}{\left(\int_0^T D_i^2 dt \int_0^T D_j^2 dt\right)^{1/2}}$	C_{ij} indicates the likelihood of two proteins being associating with each other: the larger C_{ij} is, the higher the likelihood. Since both D_I and D_j are negative, C_{ij} is between 0 and 1. T is the length of the time course for the measurement.

To demonstrate the basic idea of this methodology, we simplified our analysis by assuming translation to be dependent only on message levels, uniform ribosome elongation rates for all messages, and no limitations on ribosomes [19]. The essential components of this correlated-deviation algorithm are presented in Table 1.

When the mRNA and protein time-series are available, the protein rate and the translation rate can be calculated following the equations in Table 1. In the absence of depletion effects on protein i, v_i is equal to v_{Ti}, and plotting it against R_i yields a line through the origin with slope k_T. In the presence of depletion effects, however, trajectories in this plot deviate from the linear behavior ($D_i < 0$) and may lead to patterns that are suggestive of different classes of protein functions [19]. By computing the pair-wise PCCs between all the proteins, we can construct a protein-correlation matrix (PCM) that identifies potential protein-protein interactions.

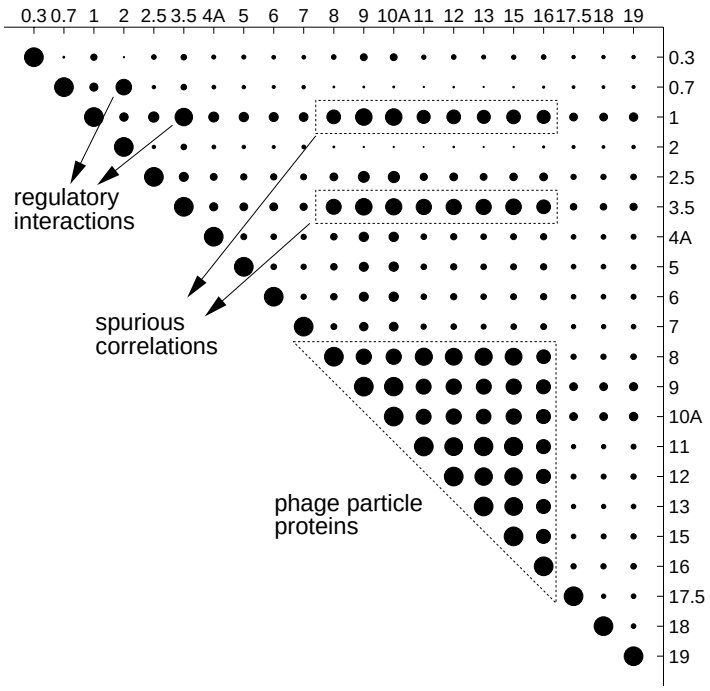

Figure 6. A protein correlation matrix (PCM) for 21 essential T7 proteins. Each off-diagonal matrix element represents the correlation coefficient, propotional to the diameter of the filled circle, between the DDFs of two proteins. Diagonal elements, which indicate the self-correlation of the proteins, are by definition always equal to 1.

A PCM is shown for 21 essential T7 proteins based on the simulation output for the time series of these proteins and their mRNAs (Figure 6). It is overall consistent with the known functions for these proteins. The deviations for gp1 and gp3.5 are strongly correlated because they interact with each other, and those of gp0.7 and gp2 are strongly correlated because both proteins are depleted in a similar fashion through their interactions with *E. coli* RNAP. Moreover, gp8 through gp16, including especially gp9 and gp10A, are highly correlated with each other because they are depleted in a stoichiometric manner during phage particle formation [23]. With the exception of gp9, gp8 through gp16 are components of the final phage particle. Although gp9 is not present in the final particle, it is required for and consumed by the particle assembly process. Since the PCM provides correlations, we caution against attempting to infer mechanisms from it alone. Some spurious associations for which no interactions were implemented in the simulation, such as those between gp1 (or gp3.5) and particle proteins, are evident. Extension of the analysis to allow for time-lag correlations may improve discrimination [24].

Toward E-virus

Approximately 4000 viruses that infect bacteria, plants, and animals have been identified and assigned to 71 families, 9 subfamilies, and 164 genera [25]. Several

computer models [26-30], including the current T7 model, have been developed to study the intracellular growth of only a handful of them. As the utility of such simulations to experimentalists grows, so will the demand to develop new simulations. To meet this need, we are currently developing a knowledge-driven program for viral-growth simulation which we call "Electronic virus" or E-virus, by analogy with the E-CELL project by Tomita and coworkers [31].

We envision E-virus will work in tandem with a virus knowledge-base, where each virus is as thoroughly documented as our current knowledge allows (Figure 7). As a one-for-all modeling platform, E-virus will seek to enable investigators to focus

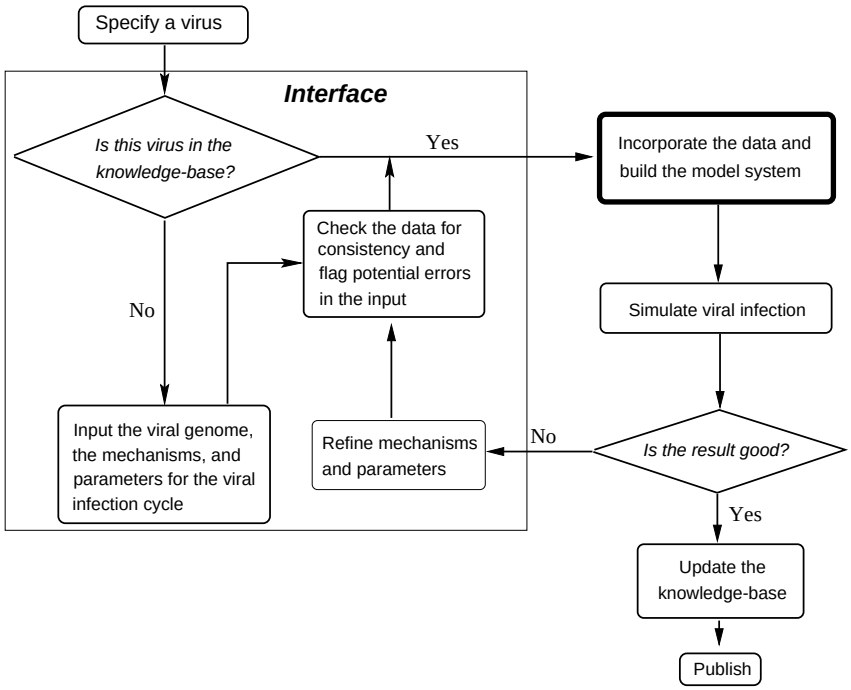

Figure 7. E-virus: a knowledge-driven general simulation of viral growth. To use E-virus, the user needs to specify the virus name. E-virus will then search the knowledge-base for information on the specified virus. If it locates the information, it will continue to the next step; if not, it will prompt the user to input the detailed information about the virus and then continue to the next step. To complete a simulation request, the user will also need to specify what kind of simulation he or she wants to perform. By incorporating the information from the knowledge-base or directly from the user input, E-virus will subsequently generate the appropriate differential and algebraic equations on the fly and then perform the computation. Next, the user can judge the quality of the simulation, probably by comparing it with experimental data, and then decide whether to accept the output and update the knowledge-base, or to reject it and to perform a further simulation using refined mechanisms or parameters.

more on systems-level issues than on the modeling details. In addition, it will provide a platform for comparative study of the infection dynamics of related viruses from the systems perspective.

E-virus is being developed as a generalization of T7v2.5, which itself has been designed using an objected-oriented approach with C++ [6]. Genetic elements, genomes, RNAs, proteins, capsids, and viral particles constitute the most essential collection of entities that one may encounter during a viral growth cycle. Moreover, the number of rules governing the dynamics of these entities is also small [32, 33]. Because these players and rules are shared by all viruses, many C++ classes (blueprints of objects) designed for and used in the T7 model can be used in E-virus, with only minor modifications. Like the T7 model, E-virus will be developed using a deterministic approach, where reactions will be represented and solved using differential and algebraic equations, but its framework will be flexible enough to conveniently accommodate stochastic formulations [34].

Acknowledgments

We thank the Office of Naval Research (Grant# N00014-98-1-0226) and the National Science Foundation Presidential Early Career Award to J.Y. (Grant# BES-9896067) for supporting this work.

References

1. Studier, F.W. and J.J. Dunn, *Organization and expression of bacteriophage T7 DNA.* CSH Quant. Biol., 1983. **47**: p. 999-1007.
2. Dunn, J.J. and F.W. Studier, *Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of T7 Genetic Elements.* J. Mol. Biol., 1983. **166**: p. 477-535.
3. Molineux, I.J., *Bacteriophage T7*, in *Encyclopedia of Molecular Biology*, T.E. Creighton, Editor. 1999, New York: Wiley. p. 2495-2507.
4. Endy, D., D. Kong, and J. Yin, *Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7.* Biotech. and Bioeng., 1997. **55**: p. 375-389.
5. Endy, D., *Development and Application of a Genetically Structured Simulation For Bacteriophage T7*, 1997, Dartmouth College.
6. You, L. and J. Yin, *Effects of E.coli physiology on bacteriophage T7 growth in silico.* In preparation, 2000.
7. Delbrück, M., *Bacterial viruses or bacteriophages.* Biol. Revs. Cambridge Phil. Soc., 1946. **21**: p. 30-40.
8. Cohen, S.S., *Growth requirements of bacterial viruses.* Bact. Rev., 1949. **13**: p. 1-24.
9. Hedén, C.-G. *Studies of the infection of E. coli B with the bacteriophage T2*. in

Acta Pathologica et Microbiologica Scandinavica. Supplementum LXXXIX. 1951.
10. Adams, M.H., *Bacteriophages*. 1959, New York: Interscience.
11. Kutter, E., *et al.*, *Effects of Bacterial Growth Conditions and Physiology on T4 Infection*, in *Molecular Biology of Bacteriophage T4*, J.D. Karam, Editor. 1994, Washington, DC: ASM Press. p. 406-418.
12. Hadas, H., *et al.*, *Bacteriophage T4 development depends on the physiology of its host Escherichia coli.* Microbiology, 1997. **254**: p. 179-185.
13. Carlton, R.M., *Phage therapy: past history and future prospects.* Arch. Immunol. Ther. Exp., 1999. **47**: p. 267-274
14. Bremer, H. and P.P. Dennis, *Modulation of Chemical Composition and other Parameters of the Cell by Growth Rate*, in *Escherichia coli and Salmonella: Cellular and Molecular Biology II*, F.C. Neidhardt, *et al.*, Editors. 1996, Washington, DC: ASM Press. p. 1553-1569.
15. Donachie, W.D. and A.C. Robinson, *Cell division: parameter values and the process*, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, J.L. Ingraham, *et al.*, Editors. 1987, Washington, DC.: ASM Press. p. 1578-1593.
16. Ball, L.A., *et al.*, *Phenotypic consequences of rearranging the P, M, and G genes of vesicular stomatitis virus.* J. Virol., 1999. **73**(6): p. 4705-4712.
17. Wertz, G.W., V.P. Perepelitsa, and L.A. Ball, *Gene rearrangement attenuates expression and lethality of a nonsegmented negative strand RNA virus.* Proc. Natl. Acad. Sci. USA, 1998. **95**(7): p. 3501-3506.
18. Endy, D., *et al.*, *Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes.* Proc. Nat. Acad. Sci. USA, 2000. **97**(10): p. 5375-5380.
19. You, L. and J. Yin, *Patterns of regulation from mRNA and protein time-series.* Metabolic Engineering. In press, 2000.
20. Velculescu, V.E., *et al.*, *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-487.
21. Schena, M., *et al.*, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-470.
22. O'Farrell, P.H., *High resolution two-dimensional electrophoresis of proteins.* J. Biol. Chem., 1975. **250**(10): p. 4007-4021.
23. Steven, A.C. and B.L. Trus, *The structure of bacteriophage T7*, in *Electron microscopy of proteins*, J.R. Harris and R.W. Horne, Editors. 1986, London: Academic Press. p. 1-35.
24. Arkin, A. and J. Ross, *Statistical Construction of Chemical Reaction Mechanism from Measured Time-Series.* J. Phys. Chem, 1995. **99**: p. 970-979.
25. Martelli, G.P., M.A. Mayo, and M.D. Summers, eds. *Virus Taxonomy Classification and Nomencalture of Viruses: Sixth Report of the International Committee on Taxonomy of Viruses*. 1995, Springer-Verlag:

Vienna, Austria.
26. Shea, M.A. and G.K. Ackers, *The O_R Control System of Bacteriophage Lambda. A Physical-Chemical Model for Gene Regulation.* J. Mol. Biol., 1985. **181**: p. 211-230.
27. Buchholtz, F. and F.W. Schneider, *Computer simulation of T3/T7 phage infection using lag times.* Biophys. Chem., 1987. **26**(2-3): p. 171-179.
28. Hammond, B.J., *Quantitative study of the control of HIV-1 gene expression.* J. Theor. Biol., 1993. **163**(2): p. 199-221.
29. McAdams, H.H. and L. Shapiro, *Circuit simulation of genetic networks.* Science, 1995. **269**(5224): p. 650-656.
30. Reddy, B. and J. Yin, *Quantitative intracellular kinetics of HIV type 1.* AIDS Res. Hum. Retroviruses, 1999. **15**(3): p. 273-283.
31. Tomita, M., *et al.*, *E-CELL: software environment for whole-cell simulation.* Bioinformatics, 1999. **15**(1): p. 72-84.
32. Savageau, M.A., *Rules for the evolution of gene circuitry.* Pac. Symp. Biocomput., 1998: p. 54-65.
33. Flint, S.J., *et al.*, *Principles of Virology. Molecular Biology, Pathogenesis, and Control*. 2000, Washington, DC: ASM Press.
34. Gillespie, D.T., *Exact stochastic simulation of coupled chemical reactions.* J. Phys. Chem., 1977. **81**(25): p. 2340-2361.