THE EXTENSION, APPLICATION, AND GENERALIZATION OF A PHAGE T7 INTRACELLULAR GROWTH MODEL

by

Lingchong You

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Chemical Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2002

In memory of Shiqin Xu, my late grandmother

THE EXTENSION, APPLICATION, AND GENERALIZATION OF A PHAGE T7 INTRACELLULAR GROWTH MODEL

Lingchong You

Under the supervision of Professor John Yin

at the University of Wisconsin-Madison

ABSTRACT

I use bacteriophage T7, a lytic virus that infects bacterium *E. coli*, as a model system to explore how the genetic information encoded in a genome determines the phenotype of an organism in a given environment. By incorporating the existing experimental data and mechanisms, our group previously developed a genetically structured model of T7 intracellular growth. I extended and improved the model by recasting it in an objected-oriented framework, by incorporating a simple model to account for the host physiology, and by implementing a more mechanistic description of several steps of T7 infection.

By using the revised T7 model, I explored several biological questions that have broad relevance. I examined the effects of host physiology on T7 growth. Consistent with experimental data, I found that T7 growth was sensitive to the physiological conditions of its *E. coli* host. Simulations further suggested that T7

growth was limited by the capacity and efficiency of the host translation machinery. This work may provide insight into the interplay between the genome of an organism and its growth environment. Next I probed the design features of T7 by investigating the response of phage T7 to perturbations in its parameters and genomic structure. My results indicated that phage T7 was nearly optimal for environments with limited resources. These results may lead to a better understanding of the design principles of a biological system in the context of its environment. In addition to studies at the single-cell level, I employed the T7 model to investigate the genetic interactions among deleterious mutations at the population level. Such interactions have profound implications in many important biological phenomena, such as the evolution of sex. Simulations suggested that the nature and degree of genetic interactions depended on the growth environment and the severity of mutations. This result provided an intuitive explanation for the experimental controversy over the nature of genetic interactions.

The T7 model generates as a byproduct the time-series of T7 mRNAs and proteins, which I employed to evaluate a novel algorithm for inferring gene functions. The algorithm highlighted the function of several T7 regulatory proteins, and established a protein correlation matrix that correctly reflected the interactions among selected T7 proteins. This algorithm is potentially useful to

interpret the gene expression data that are produced from high-throughput techniques, such as DNA microarrays and protein 2D gel electrophoresis.

To facilitate the modeling of biological systems in general, I have developed a software package called Dynetica – a biologist-friendly simulator of dynamic networks. Dynetica provides an intuitive environment for constructing, visualizing, and analyzing mathematical models of biological systems, such as metabolic, signaling and genetic networks. I expect that Dynetica, along with other modeling tools, will synergistically benefit biological research at large by serving as a computational framework for creating and sharing mathematical models.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor Dr. John Yin for his guidance and support during the last four and a half years. Encouraging and open-minded, he has always been willing to entertain my craziest ideas while rigorously challenging and critiquing them, and helping to transform them into manageable research projects. Without his mentoring, I would not have gone so far in my research – I am eternally grateful for what he has done for me.

I owe many thanks to many current and past Yin group members for stimulating and inspirational interactions. I would like to especially thank Dr. Drew Endy, who patiently guided me into his digital phage world and encouraged me to revamp the whole thing – which has turned into a major part of my thesis work. Thanks go to Patrick F. Suthers for wonderful times, socially and scientifically, that we have shared as officemates. In addition, I would like to thank several undergraduate students, Apirak Hoonlor in particular, for their assistance in improving Dynetica.

I would like to thank Dr. Ian J. Molineux in the University of Texas at Austin for his genuine and constructive help in sharing with me his expertise in phage T7 biology. I also thank him for the opportunity to learn microbiology experiments in his lab during my first year in graduate school.

The department of Chemical Engineering and UW-Madison as a whole have provided an excellent environment for my research. I am extremely grateful to each of my other thesis committee members: Dr. Fred Blattner for his enthusiasm and encouragement in my developing Dynetica and for inviting me to his group meetings, Dr. Richard Gourse for teaching me Prokaryotic Molecular Biology and interesting discussions on modeling ribosome synthesis; Dr. Sean Palecek for his excellent advice and encouragement in my career development, and Dr. James Rawlings for numerous discussions on stochastic and deterministic modeling and also for his advice and help in my job search. In addition, I am also deeply indebted to them for spending their precious time in reading and critiquing my dissertation.

Dr. Nick Abbott and Dr. Juan de Pablo have also been extremely helpful in serving as my references. I am fortunate to have served as a TA for a course taught by Dr. Abbott and taken a very intensive course taught by Dr. de Pablo; both experiences have proven tremendously rewarding. I have benefited greatly by learning from other faculty members in the Chemical Engineering department through classes or extracurricular interactions. I would also like to thank all the staff members in the Department of Chemical Engineering for their patience and help. They have made my daily life in the department much easier and more enjoyable.

My life would be largely meaningless without family and friends. I owe all my achievements to their support during the last thirty years or so. I would like to particularly thank my Mother and my wife for their love and faith. My uncle served as my private mentor during my high school and, as a physics teacher, guided me into the realm of science. I am grateful to my younger brother for taking up much of the responsibility in taking care of my family back in China, so that I could pursue my goals in doing research.

I have been fortunate to have two great friends, Jing Liu and Ronghua Yu, who have always been understanding and supportive no matter how far apart we are in the world. While in Madison, I have greatly enjoyed friendship with Fan Ding and his wife Bo Zhan, Guang Yao, Qiang Wang, Hangsheng Wang, Shouhong Guang, Weidong Xia, Xurong Kong, as well as many graduate students in Chemical Engineering. In addition, I am particularly grateful to David Arfa and his wife Sandra for their friendship, support and mentoring.

Finally I would like to thank the UW-Madison Graduate School for granting me the Marie Christine Kohler Knapp Fellowship, as part of which I am fortunate to live in the former Wisconsin governor's mansion with 11 wonderful graduate students with diverse backgrounds. Interactions with these people have greatly expanded my perspective on research and life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

C h a p t e r 1

INTRODUCTION

> *"If we hope to understand biology, instead of looking at one little protein at a time, which is how biology works, we will need to understand the integration of thousands of proteins in a dynamically changing environment."*
>
> Craig Venter

1.1 Modeling for integrated understanding of biological systems

1.1.1 Expansion of biological information

Amid hope, awe and controversies[1], two draft human genome sequences were simultaneously released at the dawn of the new millennium (Lander et al 2001; Venter et al 2001). This milestone highlights the exponential growth of biological information in the last several decades. Counting the human genome, about 1000 genomes of natural plasmids, organelles, viruses and viroids, bacteria, animals and plants have been sequenced to date (Lander et al 2001). Other evidence of the growth in biological information is vividly illustrated by the increase of the total number of gene or genome sequences stored in GenBank. This number has increased exponentially from 600 or so to more than 15 million over a period of merely twenty years (http://www.ncbi.nih.gov/Genbank/), and the rate of growth has been gradually

[1] Controversies over sequencing of human genome are both social and technological. The Human Genome Project raises many concerns about potential abuse of the sequence information in recruitment, health insurance, and racial profiling. The reported two sequences were obtained using two different approaches. This has sparked many heated discussions as for which method is more appropriate and whether the sequencing team from the private sector led by Celera has actually made any significant contribution at all (Butler, 2001; Green, 2002;Myers et al., 2002; Waterston et al., 2002).

accelerating (Benson et al 2002). Meanwhile, as the technologies of X-ray and NMR spectroscopy become more mature, the number of protein structures that are solved every year has enjoyed steady increase (Westbrook et al 2002).

The rapid expansion in biological information has moved us one step closer to a complete understanding on how the genetic information stored in a genome determines the behaviors or characteristics (i.e., the phenotype[2]) of an organism or a cell in a given environment. Understanding the relationship between genotype and phenotype is important not only for our conceptual satisfaction, but also for many practical issues, such as finding cures for human diseases. These diseases may be caused by deleterious changes in human genes or by infectious agents such as viruses or bacteria. If we understand mechanistically the causes of a disease, we may be able to more rationally develop appropriate pharmaceuticals.

As has been demonstrated in the sequencing of numerous genomes, including very complex ones, such as those of human, and more recently of rice (Goff et al 2002; Yu et al 2002) [3], advances in biotechnology have probably made the sequencing of any genome a routine work[4]. Even with the sequence data available, however, much remains to be done to really make the link between a genome and a resulting phenotype. The

---

[2] Phenotype is the appearance or other characteristics of an organism, resulting from the interaction of its genetic constitution with the environment. This contrasts with the genotype, which is genetic constitution of an organism (Lewin, 1997). Note that the phenotype depends on the environment but the genotype does not.

[3] Here the genome complexity can be roughly characterized by genome size. For example, the human genome is composed of about 3 billion base pairs, and the rice genome is about 470 million base pairs in size.

[4] This point is certainly not to diminish the amount of capital and effort that need be invested into such mega-projects. For example, the sequencing of human genome has involved collaborative efforts by hundreds of researchers from around the world.

next step is to identify the genes in a genome and determine their functions. With the advances in sequence analysis, predicting potential genes in a genome has been relatively successful for simple organisms such as *E. coli* and yeast, but remains a major challenge for higher organisms including the human being. However, being able to predict genes does not mean that we know how they function. For example, about one third of the 4000 or so genes in *E. coli* have unknown functions (Blattner et al 1997). Major efforts are now being made to characterize the function of genes using comparative sequence analysis (Bork & Koonin 1998) or by studying whole-genome scale gene expression at the mRNA level, using DNA microarrays (Schena et al 1995) or the method of serial analysis of gene expression (Velculescu et al 1995), and at the protein level using protein 2D gel electrophoresis (Karger et al 1995; O'Farrell 1975) or mass spectrometry (Eckerskorn et al 1992; Henzel et al 1993; Mann et al 2001). Such studies have spawned the so-called transcriptome and proteome[5], which represent complete sets of RNAs and proteins, respectively, that correspond to the genome. The transcriptome and the proteome are distinguished from the genome in that they are context dependent (Oliver 2000); for example, levels of RNAs and proteins in a cell usually vary with environmental conditions.

---

[5] There are too many new "-ome" and "-omics" terms to completely list here. Interested readers are directed to http://www.genomicglossaries.com/content/omes.asp for a very detailed list of these terms and their definitions. Other popular "-ome" terms include interactome (the complete set of interactions, often referring to protein interactions in the cell), metabolome (the complete set of metabolites in the cell), and physiome (the quantitative description of the physiological dynamics or functions of the intact organism). In reflecting this trend, the journal of *Microbial and Comparative Genomics* was recently renamed to *Omics: a Journal of Integrative Biology* (http://www.catchword.com/titles/15362310.htm).

In addition to the characterization of the function of individual genes, endeavors have been undertaken to establish the interaction map between gene products, particularly proteins. To date, construction of large-scale interaction maps among proteins has primarily relied on the yeast two-hybrid analysis method, which identifies protein-protein interactions by the activation of reporter gene expression in yeast (Fields & Song 1989). This method has been applied to probe genome-wide protein-protein interactions in phage T7 (Bartel et al 1996), hepatitis C virus (Flajolet et al 2000), vaccinia virus (McCraith et al 2000), and *Saccharomyces cerevisia* (Ito et al 2001; Uetz et al 2000). It has also been used to characterize interactions among selected groups of proteins (not genome-wide) in *Drosophila*, *C. elegans*, and mouse (for a review, see (Uetz 2002)). The yeast two-hybrid method has been generalized to establish the so-called n-hybrid systems to detect DNA-protein (one-hybrid), protein-protein (two-hybrid), and RNA-protein or small molecule-protein (three-hybrid) interactions (Vidal & Legrain 1999). N-hybrid systems have also been established using bacterial hosts (Hu 2001). As a powerful tool to identify and quantify proteins, mass spectrometry has also been used to identify protein complexes in *Saccharomyces cerevisia* (Gavin et al 2002; Ho et al 2002). These tools complement one another in revealing different aspects of protein interactions (Uetz 2002), and are beginning to provide a comprehensive view of how the individual cellular parts are functionally connected to one another.

1.1.2 Modeling for knowledge integration

If one day we have characterized the function of all genes by constructing the transcriptome, the proteome, and the interaction map between gene products in a cell

(or an organism), will we have completed the link between the genome and cellular behaviors? Partly. By reaching that step, we will have the complete part-lists of the cell. However, knowing what a cell is composed of and how each individual component works does not necessarily mean understanding how the cell as a whole works. For example, understanding what parts a car is composed of and how each part works does not mean that we would understand how the car itself works. To really claim that we understand how the car works, we should be able to put the parts back together and demonstrate that the car works. Similarly, to understand how a cell function as a whole, we will need to integrate our understanding of the parts and see whether or to what extent these pieces of understanding will coherently predict cellular behaviors. This integration process will be greatly facilitated by *modeling*, particularly mathematical modeling. Such a systems or integrated approach for biology research contrasts the reductionist approach, which has been the predominant approach for accumulating our knowledge about biological systems.

Depending on its objective, a model may involve details at different levels. One of the most profound biological models is the so-called "central dogma" of molecular biology (Crick 1970), which describes the following information transfer process:

$$\text{DNA} \rightarrow \text{mRNA} \rightarrow \text{protein}$$

While omitting many details involved in the individual steps of this process, such as binding of RNA polymerases to the DNA sequence during transcription, the central

dogma epitomizes decades of endeavor by biologists that led to the elucidation of this process. It elegantly reduces many complex biological processes into one dimension, and has served as the theoretical foundation for the entire field of genetic engineering.

The central dogma as illustrated above exemplifies the probably most intuitive models that biologists have been using – diagrams. Diagrams have been widely used and will probably continue to prevail in biology textbooks and literature. Recently, there have been significant efforts in formalizing the conventions of drawing diagram models (Kohn 1999). In general, diagrams serve as tremendous visual aid for understanding the biological processes and for formulating new hypotheses to be tested experimentally. However, diagrams are descriptive only, and cannot predict in a quantitative and sometimes not even in a qualitative fashion behaviors of a given system. A more sophisticated approach, mostly in characterizing gene networks, is to use a Boolean representation of a system (Glass 1975; Glass & Kauffman 1973; Thomas 1973). In this paradigm, each player (oftentimes a gene) has two states, ON and OFF; the system of interest is overall represented as a logic network, and the dynamics describes how genes interact to change one another's states over time (Hasty et al 2001). Although a Boolean model can provide insight into the qualitative behavior of the underlying system, it is usually overly simplified and tends to give ambiguous predictions (Kuipers 1986).

To achieve deeper understanding and greater predictive power, more detailed information need to be incorporated. This leads to the next level of modeling, the stoichiometric models (Clarke 1988; Fell 1992). A stoichiometric model represents the

underlying system as a set of coupled chemical reactions[6]. It is more detailed than previous approaches in that it requires the stoichiometry of each reaction. For example, to describe the process of A being turned into B, we will also need the information on how many molecules of A are turned into B. Coupled with a technique called metabolic flux analysis (Fell 1992), stoichiometric models have played an instrumental role in shaping the field of metabolic engineering, by providing theoretic guidance for experimental manipulation of metabolic networks (Stephanopoulos et al 1998). Recently, stoichiometric models have proven powerful in characterizing the underlying structure of metabolic networks by determining the elementary flux modes (Schuster et al 2000) or the null space base vectors (Schilling & Palsson 1998) and in predicting steady-state metabolic capabilities of several model organisms, such as *E. coli* (Edwards et al 2001; Schilling et al 1999) and *H. influenzae* (Edwards & Palsson 1999). But their application is limited by their inability to predict the temporal evolution of biological systems. To make such predictions, the stoichiometric structure of the reaction networks needs be supplemented with detailed kinetic information, resulting in kinetic models.

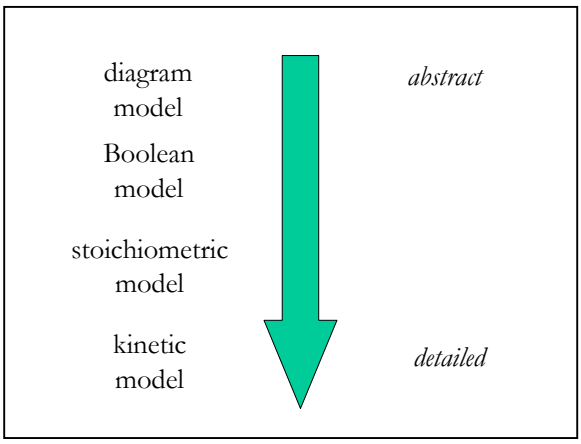

Figure 1.1 Models of varying details

[6] While not very obvious, many biological processes, particularly cellular processes, can indeed be easily formulated as chemical reactions.

From diagrams to kinetic models, more detailed information will be needed, but the resulting models will be more powerful in their predictive power (Figure 1.1). In early phases of biological research, much of the knowledge integration exercise relied on the use of simpler and more descriptive approaches, particularly diagram models. Thanks to the more detailed data and mechanisms that can be obtained from experiments, more sophisticated models are now becoming increasingly preferable. For some well-characterized systems, even the more information-intensive approach – kinetic modeling – has become a realistic goal. Kinetic models have recently been applied to the analysis of a wide variety of biological systems, including bacterial chemotaxis signaling networks (Barkai & Leibler 1997; Spiro et al 1997), developmental pattern formation in *Drosophila* (von Dassow et al 2000), aggregation stage network of *Dictyostelium* (Laub & Loomis 1998), viral infection (Endy et al 1997; McAdams & Shapiro 1995; Reddy & Yin 1999; You et al 2002), circadian rhythms (Barkai & Leibler 2000; Smolen et al 2001), single cell growth (Shuler et al 1979), and physiological processes (Noble 2002; Quick & Shuler 1999; Winslow et al 2000). In chemical engineering, kinetic modeling is not new; it has long become a routine approach employed to characterize a wide variety of chemical processes, some of which, such as combustion, may involve thousands of coupled chemical reactions. The above-mentioned examples have demonstrated that kinetic modeling can also be employed to describe many biological processes.

Kinetic models may be further coupled with equations describing mass transport processes, leading to more complicated mathematical models not covered in Figure 1.1.

For example, in modeling some biological processes, it is necessary and feasible to account for not only reaction kinetics, but also transport of interacting components by diffusion (Schaff et al 2000; Yin & McCaskill 1992; You & Yin 1999). Such models often involve using partial differential equations (PDE) in addition to ordinary differential equations (ODEs) [7]. Owing to the higher computational cost (for the same number of interacting components) and the lack of parameters on diffusion processes, particularly for intracellular processes, however, most mathematical models in biology have so far ignored spatial heterogeneity in a system. Even when it is essential to describe some transport processes, these processes may still be approximated by first-order reactions, which in turn can be described using ODEs (Reddy & Yin 1999; von Dassow et al 2000).

1.2 Phage T7

1.2.1 A genomic model system

It remains a daunting task to explore the genotype-phenotype relationship at a whole genome level for most organisms because of the sheer number of components involved in the functioning of these organisms. Although, as briefly discussed in the previous section, a lot of data are being generated as part of the effort to bridge a genome and the resulting phenotype, the amount of data is still relatively small if we consider the large number of variables that are being dealt with (Lauffenburger 2001). However, it is now feasible to do so for well-studied simple organisms, such as

[7] ODE models account for the majority of kinetic models. However, as shall be seen in Chapter 7, ODEs are not the only format of kinetic models; the latter can be also formulated in a stochastic framework.

bacteriophage T7. Thanks to its rich knowledge base that has been accumulated since its first characterization nearly half a century ago (Demerec & Fano 1944), phage T7 serves as a good genomic model system for exploring many fundamental and applied biological questions regarding the genotype-phenotype relationship. It should be noted that there are some limitations in resorting to phage T7 as a model system. As a virus, phage T7 depends on the host cell for its survival, and as such does not possess many features that are present in a self-sustaining organism. For example, there are no genes coding for tRNAs or rRNAs, and the gene expression process is overall much simpler than in an animal cell. Nevertheless, because of the availability of rich data and mechanisms in the literature for phage T7, it provides an opportunity to explore several key questions on modeling biological systems: how and to what extent can we build mathematical models for biological systems, particularly at the whole-organism scale? How are such models useful for answering biological questions? And what questions can be answered by employing an integrated model?

Phage T7 is a lytic virus that infects *E. coli* and produces approximately 100 progeny per infected cell within thirty minutes at 30 °C. The 39, 937 base-pair T7 genome contains 56 known or potential genes (Molineux 1999) (see also http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=genome&gi=10461). T7 genes are numbered in order from left to right, according to their position on the genetic map. Overlapping genes are given numbers according to the relative positions of their left ends. T7 genes are expressed coordinately in three classes based on their position and function. The early, or class I genes code for functions to overcome host restriction

and convert the metabolism of the host cell to the production of T7 components. Class II genes, the next to be expressed, are responsible for functions involved in T7 DNA replication. The last to be expressed, class III genes code for proteins of the phage particle and functions for maturation and packaging of T7 DNA.

T7 infection is remarkably rapid and efficient, and highly regulated (Figure 1.2). The linear molecule of T7 DNA enters into the cell gradually from left to right, which is also the orientation of all T7 genes and their transcription. Although other phages like lambda inject their DNA within one minute of binding to the host, the entry of T7 DNA takes about 10 minutes (Garcia & Molineux 1995a; Garcia & Molineux 1995b; McAllister et al 1981). After the complete entry of T7 DNA, T7 transcription is mostly shut off by the accumulated T7 lysozyme (gp3.5), a class II gene product (Zhang & Studier 1997). The remaining transcription capacity switches to the stronger class III T7 promoters, which direct the transcription of class III genes. T7 DNA replication begins at about the same time when most transcription is shut off (Summers 1969; Summers 1970). The replicated T7 DNA exists in concatemeric forms and the mature form of T7 DNA with unique ends is regenerated from concatemers during packaging of the DNA into phage particles (Studier 1972). The infection cycle ends with the abrupt lysis of the host cell and progeny phages are released.

1.2.2 A phage T7 model and its applications

By incorporating the existing experimental data and mechanisms on T7 biology, our group previously developed a genetically-structured model (Endy et al 1997) to

account for the major steps of T7 intracellular growth: entry of the T7 DNA into the host, transcription of the T7 genes, translation of the resulting mRNAs, procapsid assembly, DNA replication, and the formation of phage progeny. This model distinguished itself in its ability to simulate the entire process from genetic information to phenotype by accounting for most of the known functions of T7 genes. As part of my Ph.D. thesis work, I improved the model by recasting it in an object-oriented framework, which can be easily generalized to model other biological systems. I further extended it by incorporating a simple model to account for the host physiology, and by implementing a more mechanistic description of several steps of T7 infection. These extensions led to overall better agreement with the experiments in predicting T7 intracellular growth, and better computational performance (You et al 2002; You & Yin 2001). These modifications and improvements are described in detail in Chapter 2.

The host-cell model in the revised T7 model, along with the high efficiency of computer simulations, presents an opportunity to explore on a large scale how interactions between the T7 genome and its environment (the host cell) determine various characteristics of phage T7 development. This aspect has been exploited to investigate a number of fundamental biological questions that may have broad relevance. As a first step to better understand genome-environment interactions, I investigated how the physiology of *E. coli* would affect T7 growth. This study may lead to a better understanding of the phage growth at sub-optimal conditions, such as those that exist in nature (Kutter et al 1994), and assist our evaluation of phage therapy strategies against antibiotic-resistant bacteria (Carlton 1999).

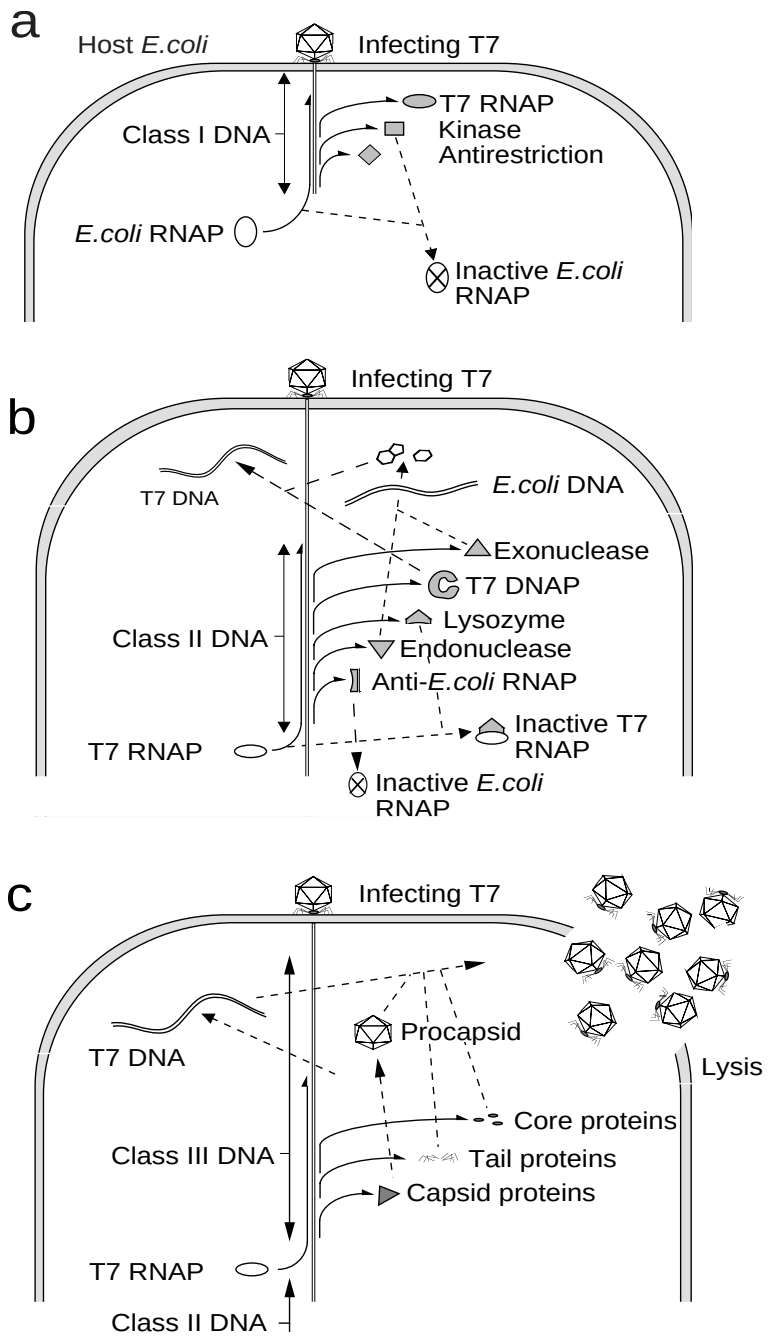

Figure 1.2 Intracellular growth cycle of phage T7. The solid lines with half arrows indicate transcription and translation, the dashed lines denote reaction, and the solid lines with full arrows indicate the three classes of T7 DNA. (a)Infection initiation, class I gene expression. (b) Class II gene expression, phage DNA replication. (c) Class III gene expression, procapsid assembly, phage maturation, and lysis. RNAP: RNA polymerase; DNAP: DNA polymerase. Adapted from (Endy et al 1997).

To probe design features of phage T7 as a complex system, I investigated how T7 growth would respond to perturbations in its kinetic parameters and its genomic structure[8] by creating and evaluating more than 150,000 *in silico* mutants. My results suggest several important features of T7 design. First, phage T7 is almost optimally designed for environments having limited resources. Second, phage T7 is overall robust to perturbations in its parameters but very fragile to perturbations in its genomic structure. Third, the robustness of phage T7 is primarily achieved by having redundancy and negative feedback embedded in the genome design. This work highlights the importance of the environment in shaping the design of an organism through evolution.

In addition to applications at the single-cell level, I systematically characterized the genetic interactions among deleterious mutations at the population level, by using the T7 model to generate 90,000 T7 *in silico* mutants and to efficiently evaluate their fitness. Such genetic interactions play a major role in a variety of fundamental biological phenomena, such as the evolution of recombination, the dynamics of fitness landscapes, and the buffering of genetic variations, but their experimental characterization has been hindered by the difficulty in generating and quantifying a large number of mutants. This work illustrated the importance of the growth environment for an organism in determining the nature of genetic interactions. Conclusions from this chapter provide an intuitive explanation for the controversy over the exact nature of genetic interactions.

---

[8] In this work, I use the term "genomic structure" to describe the ordering of genetic elements, such as genes, promoters, and transcription terminators.

As a byproduct, the T7 model generates the time-series of all T7 mRNAs and proteins, which may be considered as simulated gene expression profiling data. Taking advantage of these data, I have developed and evaluated a novel data-mining algorithm for inferring gene functions and potential protein-protein interactions. This algorithm is potentially useful for interpreting the large volume of data generated from high-throughput technologies such as DNA microarrays and protein 2D gel electrophoresis.

1.3 Modeling beyond phage T7

As has been demonstrated by examples listed above, mathematical modeling and computer simulation may deepen our understanding of complex biological systems by testing the validity and consistency of experimental data and mechanisms, by generating experimentally testable hypotheses, and by providing new insight into the behaviors of these systems. However, the application of this integrated approach in biology has been hindered by the lack of software tools to build and analyze models, particularly for researchers unfamiliar with programming and numerical methods. To meet this need, I have developed Dynetica – a simulator of dynamic networks. With Dynetica, the user can easily build models for systems that can be formulated as a coupled network of chemical reactions. A particularly distinguishing feature of Dynetica is that it facilitates easy construction of models for genetic networks, where the dominant reactions are the expression of genes and the interactions among gene products. In addition, Dynetica provides users the flexibility of performing time-course simulations using deterministic or stochastic algorithms. A deterministic algorithm is appropriate when the system variables can be treated as continuous or the model is highly lumped, whereas a

stochastic algorithm is appropriate when the system is homogeneous, numbers of interacting molecules are small, and fluctuations in these numbers are significant. Finally, since it is written in Java, Dynetica is platform-independent, allowing models to be run on most modern computers and easily shared among researchers. I anticipate that Dynetica will dramatically speed up the process of model construction and analysis for a wide variety of biological systems. Dynetica is described in detail in Chapter VII.

1.4 Overview of thesis

Chapter II details construction of the current model of phage T7 intracellular growth cycle.

Chapter III describes the application of the phage T7 model to explore the effects of host physiology on growth of phage T7. This work illustrates the interplay between the genome of an organism and its growth environment.

Chapter IV describes the investigation of the design of phage T7 by perturbing its kinetic parameters and genomic structure.

Chapter V presents the quantification of genetic interactions between deleterious mutations at the population level by using *in silico* mutagenesis.

Chapter VI illustrates an application of the model in developing and validating a new methodology to infer protein functions and protein-protein interactions from time courses of mRNAs and proteins. This methodology suggests a means to infer how

genetic networks are constructed given the large volume of data being generated by gene expression profiling at the mRNA level and the protein level.

Chapter VII presents a detailed description of construction and usage of Dynetica – a biologist-friendly simulator of dynamic networks.

Chapter VIII highlights important findings of the previous chapters and suggests future research directions.

C h a p t e r 2

CONSTRUCTION OF AN IMPROVED MODEL OF PHAGE T7 INTRACELLULAR GROWTH

"All models are wrong, but some are useful."

George E. P. Box

"To improve is to change; to be perfect is to change often."

Winston Churchill

The present T7 model, T7v2.5, recasts previous T7 models developed in our group (Endy 1997; Endy et al 1997) using an object-oriented approach. In addition, it incorporates a host-cell model that accounts for the empirical relationships between the host growth rate and the other host physiological parameters such as the numbers of *E. coli* RNA polymerases (EcRNAPs) and ribosomes, the pool sizes of nucleoside triphosphates (NTPs) and amino acids, and the cell volume; it accounts for the stoichiometric relation of the T7 helicase/primase (gp4A) and the DNA polymerase (gp5) in forming replication complexes, or replisomes, as well as the stoichiometric balance between the number of replication complexes and the maximum number of replication forks that can form on the newly synthesized T7 genomes; it determines the allocation of host and T7 RNA polymerases (T7RNAPs) among different mRNAs by the relative transcription capacities of the mRNAs; it incorporates a module that

facilitates the sensitivity analysis of the model with respect to single parameters or two parameters simultaneously.

With the new design, the model is significantly improved in several aspects. First, with an object-oriented design, T7v2.5 presents a framework that can be easily generalized to model the growth of other kinds of viruses. Second, by incorporating a host cell model and updating the mechanisms of several viral infection steps, T7v2.5 gives overall better agreement with the experiments for the base simulation, in particular in the prediction of the eclipse time and the rise rate. Third, the computation speed of T7v2.5 is over ten-fold faster than its direct precursor -- T7v2 (Endy 1997). The source code of T7v2.5 is available on request.

2.1 Design of the base model

By taking into consideration the host, the viral infection process can be treated as a complex reaction system formed by various components from both the virus and its host. We define such a system as a viral infection system (VIS). As shown in Figure 2.1, a VIS consists of at least a virus and a host cell, which in turn are composed of components of various levels. In the so-called object-oriented paradigm, many of these components can be programmed as objects. Each object has its own state defined by some attributes, which may change through the object's "communications" with itself or other objects. For example, a protein, if treated as an object, may have such attributes as its name, its length in amino acids, the position of its gene along the genome, and its concentration in the cell. The protein concentration changes when it decays and when it

is translated; the decay process can be considered as the result of the "communication" between the protein and itself, and the translation process can be thought of as the translation machinery "telling" the protein to change its concentration. The entire system is essentially a web of objects telling each other what to do. The object-oriented modeling approach usually yields well-defined reusable components and an easily maintainable source code.

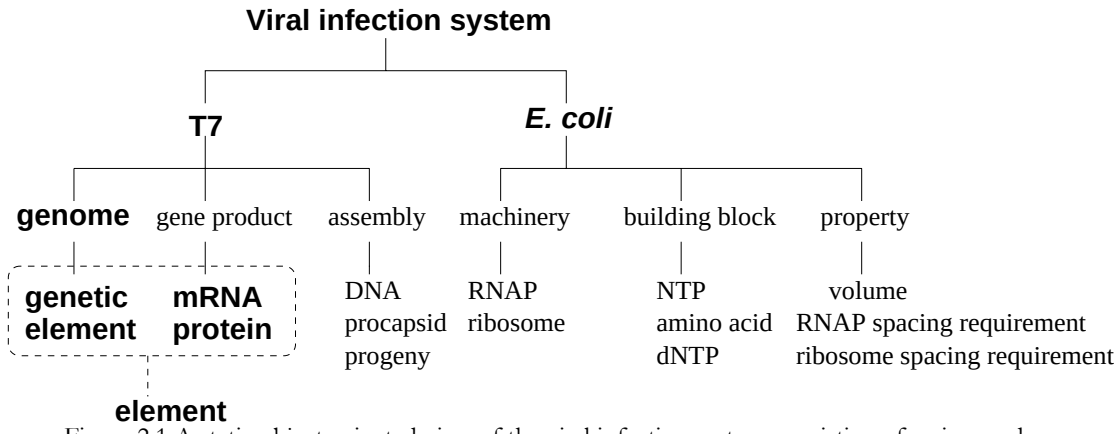

Figure 2.1 A static object-oriented view of the viral infection system consisting of a virus and a host cell. Entities shown in bold are defined as C++ classes. The others are represented as simple variables. The classes of higher levels are derived from classes and variables from the immediately lower level. Interactions among various components are omitted from this figure. Because of some common features they share, the classes genetic element, mRNA, and protein are all derived from a more abstract class element.

2.1.1 *Representation of the host cell*

The current T7 model incorporates an empirically based host-cell model that accounts for experimentally observed correlations between the *E. coli* growth rate and resources such as the numbers of EcRNAPs and ribosomes, the pool sizes of NTPs and amino acids, and the cell volume, shown in Table 2.1 and reviewed elsewhere (Bremer & Dennis 1996). In setting up the host-cell model, we treat the host cell as a spatially homogeneous resource reservoir, where the levels of the resources are defined at the

initiation of the T7 infection. Moreover, by employing the equations in Table 2.1, we implicitly assume: (1) cell volume is constant over the course of infection, (2) negligible exchange of metabolites between the infected cell and its extracellular environment during infection, (3) the model host cell represents an average of a cell population that grows at the exponential phase immediately prior to phage infection, (4) the initial NTP pool size is equivalent to the total RNA content of the cell, (5) NTPs are not consumed as the energy source for the reactions, (6) the initial amino acid pool size is equivalent to the total protein content of the cell. These assumptions are made based on our still limited understanding of the interplay between phage T7 and its *E. coli* host; some of them are open to debate. For example, the value for the NTP pool size is chosen mainly because of the lack of mechanistic understanding on the exact level of NTP in the host cell that is available for phage development. Nevertheless, these assumptions serve as a starting point toward representing the host cell, and they can be refined when more experimental data and mechanisms become available.

Table 2.1 Dependence of *E. coli* physiological parameters on the cell growth rate (μ, doublings/hr)[a]

Parameter	Equation
E. coli cell volume (liter)	[b] $V_c = 8^{\mu} \times 10^{-16}$
EcRNAP number (molecule)	$N_P = 192.2\,\mu^3 - 155.2\,\mu^2 + 549.0\,\mu - 103.6$
EcRNAP elongation rate (base/s)	$k_{PE} = 108.7 - 63.6\,\mu^{0.1797}$
Ribosome number (molecule)	$N_R = 0.8\times(13.1\mu^2 - 6.89\mu + 6.46) \times 10^3$
Ribosome elongation rate (residue/s)	$k_E = 102.9 - 55.5\mu^{0.3609}$

Amino acid number (molecule) [c]	[d] $P = (-0.73\mu^2 + 2.74\mu + 1.48) \times 2^{(C+D)\mu/60} \times 10^8$
NTP number (molecule) [c]	$R = 1.30\ \mathrm{P}\ \mu / k_E$
DNA content (genome equivalent)	[d] $G_e = 60 / (C \ln 2)\ (2^{(C+D)\mu/60} - 2^{\mu/60})$

[a] All the equations are directly taken from reference (Bremer & Dennis 1996) or derived from the data therein unless otherwise indicated.
[b] From reference (Donachie & Robinson 1987).
[c] The NTP level represents the total RNA content of the host cell, and the amino acid level represents the total protein content of the host cell.
[d] C and D are the lengths of the C and D periods, respectively; they are related to the growth rate as follows: C = 42.4 + 132.7 $e^{-2.812\mu}$ (min); D = 22.3 + 15.4$e^{-1.16\mu}$ (min).

2.1.2 Representation of the viral genome

The viral genome is represented as an array of genetic elements; each genetic element is defined as a segment of DNA in the viral genome, with a unique function (genes, promoters, RNA splicing sites, etc.) or no known function (spacer DNAs). The wild-type T7 genome consists of 145 genetic elements, including two genome ends, 59 genes [9], seven EcRNAP promoters, 17 T7RNAP promoters, two transcription terminators, ten ribonuclease (RNase) III processing sites and 48 spacer DNAs (Figure 2.2). The viral genome is specified in an input file, where each genetic element is assigned five attributes: a name, a starting position and an ending position along the genome, type, and efficiency. Shown below is part of the input file specifying the wild-type T7 genome:

Name	Start	End	Type	Efficiency
…	…	…	…	…
A3	750	815	2	1
R0.3	816	890	6	0

[9] Here every protein-coding gene is counted as a separate one, even if two genes may overlap.

Blk6	891	924	9	0
gp0.3	925	1278	0	0
…	…	…	…	…

For example, A3 is the third major promoter for EcRNAP, it starts at 750 base pairs (bp) relative to the left genome end and ends at 815 bp, its type is 2 (0 = non-essential gene, 1 = essential gene, 2 = EcRNAP promoter, 3 = T7RNAP promoter, 4 = EcRNAP transcription terminator, 5 = T7RNAP transcription terminator, 6 = RNase III site, 9 = spacer DNA, 10 = genome end), and its relative activity (indicated as Efficiency) is 1.

The current model is general enough to deal with *in silico* T7 mutants with genomes in which the genetic elements are reordered. In these mutants, some or all elements can be relocated to positions different from their wild-type positions. During the relocation, a group of overlapping elements is treated as a block to be moved together. Since a spacer DNA does not have any defined functions, it is treated as part of the proceeding block. Therefore, in cases where we need to rearrange T7 genetic elements, we end up having 74 blocks. For example, one such block consists of gene 3.8 (start = 11225 bp, end = 11590 bp), gene 4A (11565, 13265), gene 4.1 (11635, 11757), gene 4B (11754, 13265), ϕ4.1 (12671, 12706), gene 4.2 (12988, 13326), and spacer DNA 23 (13327, 13340). For simplicity, in this dissertation, the term genetic element is used to represent a non-overlapping genetic element or a genetic block unless stated otherwise.

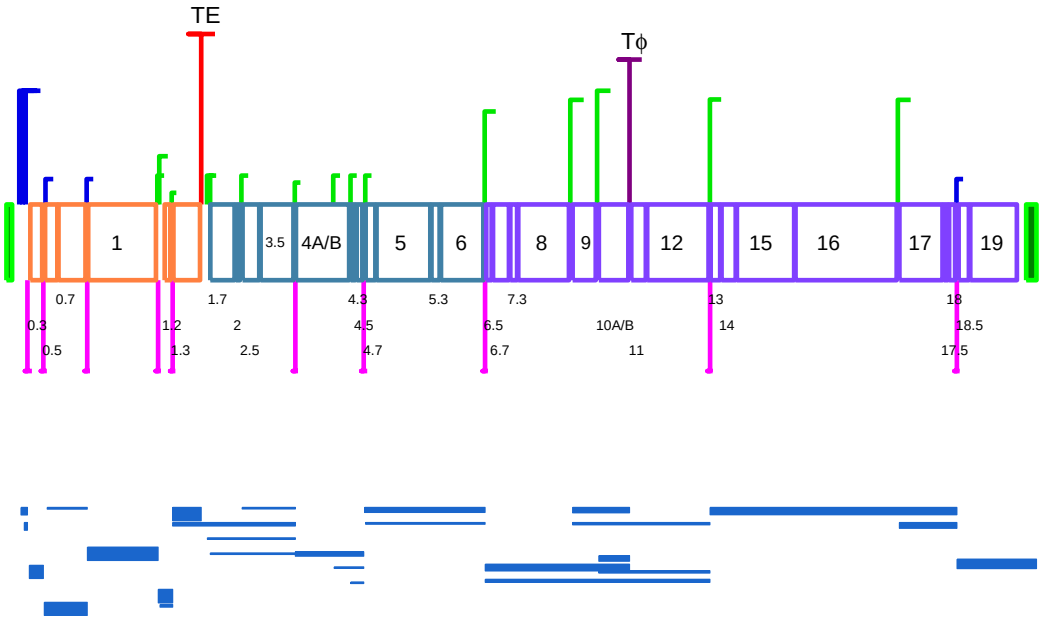

Figure 2.2 The wild-type T7 genome. Boxes represent genes. Vertical lines with half bars above the genes represent EcRNAP (blue) and T7RNAP (green) promoters; the heights of these lines represent the relative activities of the promoters (not to the scale). Vertical lines with full bars above the genes represent EcRNAP (red) and T7RNAP (dark purple) terminators. Vertical lines below the genes represent RNase III processing sites. The two green bars on both ends represent the left end (LE) and right end (RE) of the genome. In altering the genetic-element order, these two elements are always fixed. The horizontal bars below the genome represent the various processed T7 mRNAs; the thickness of a bar represents the transcriptional capacity allocated to the mRNA (not to the scale).

2.1.3 Translocation of T7 genome into the host cell

The translocation of the T7 genome into the host is simulated by starting with a 90-second delay representing the time between the addition of phage to a shaker flask of host cells and the beginning of genome entry. Then it is modeled as a three-stage process based on experimental results (Garcia & Molineux 1995b; Garcia & Molineux 1996; Struthers-Schlinke et al 2000). The first stage involves the efficient ejection of approximately 850 bp of the leading end of the T7 genome at a rate of 70 bp/s. This section of T7 genome contains three major EcRNAP promoters (A1, A2 and A3). As

these promoters initiate the transcription of class I genes, the T7 genome is pulled into the host at 40 bp/s. When T7RNAP is expressed, it recognizes the class II promoters and begins the transcription of class II genes, and pulls the T7 genome into the host at 200 bp/s. The length of the internalized T7 DNA is calculated at each time point until the entire genome enters the cell at about seven minutes post infection.

The modes of genome entry may change when T7 genetic elements are reordered. For example, other EcRNAP promoters (B, C, or E) may be moved into the first 850bp, and they can initiate EcRNAP-mediated entry even if the three major EcRNAP promoters are not present in this region. However, because the second stage of the genome entry is coupled with the transcription by EcRNAP, T7 will fail to deliver its DNA, thus fail to grow, if none of the six EcRNAP promoters are located within the first 850bp of the genome.

2.1.4 Transcription of T7 genes

The model accounts for the synthesis of unique mRNA species from the promoters that have entered the host cell. Most T7 primary mRNAs are efficiently cleaved at the ten RNase III sites. For simplicity, we assume the cleavage at these sites is instantaneous and 100% efficient. Further, the model accounts for the termination and read-through of RNA polymerases at terminators TE (early terminator, for EcRNAP) and Tϕ (phage terminator, for T7RNAP). The mRNA species are generated on the fly during the simulation, based on the organization of the promoters, terminators, and

RNase III sites on the genome. Twelve categories of mRNA species will be generated during the simulation (Table 2.2).

The rate of concentration change in each mRNA species is accounted for using the following equation.

$$\frac{d[\mathrm{mRNA}_i]}{dt} = k_{PE} S_{Ei} + k_{PT7} S_{T7i} - k_{dm}[\mathrm{mRNA}_i] \qquad 2.1$$

where [mRNA$_i$] is the concentration of the ith mRNA as a function of time, k_{PE} and k_{PT7} are the elongation rates of EcRNAP and T7RNAP respectively, S_{Ei} and S_{T7i} are the densities of EcRNAP and T7RNAP along the internalized region of the genome that codes for the mRNA, and k_{dm} is the decay rate constant of the mRNA.

Table 2.2 Categories of mRNAs that can be generated during the simulation*

Starting from	Ending at	Transcribed by
EcRNAP promoter	TE RNase III site Genome end	EcRNAP
T7 RNAP promoter	Tϕ RNase III site Genome end	T7 RNAP
	TE RNase III site	EcRNAP EcRNAP

	Genome end	EcRNAP
RNase III site	Tϕ	T7 RNAP
	RNase III site	T7 RNAP
	Genome end	T7 RNAP

* For the wild-type genome 28 mRNA species will be generated

In calculating S_{Ei} and S_{T7i}, we assume that EcRNAP and T7RNAP molecules are allocated to the portions of the genome that code for the mRNAs according to the relative transcription capacities of these mRNAs, which in turn are calculated from the relative activities of the promoters (Tables 2.3 & 2.4). The maximum density of either EcRNAP or T7RNAP along any mRNA is determined from the RNAP spacing requirement. Currently the same spacing requirement data are used for both the EcRNAP and T7RNAP due to the lack of data for the latter.

Table 2.3 *E. coli* RNA polymerase promoter data used in the model

Promoter	Relative Activity
A1[a]	1
A2[a]	1
A3[a]	1
B[b]	0.05
C[b]	0.05
E[b]	0.05

[a] Some in vitro data indicate that A2 and A3 are less active that A1 (Dunn 1976).
[b] Promoters B, C, and E are shown to be utilized but their activities are much weaker than the major promoters (Dayton et al 1984).

2.1.5 Translation of T7 mRNAs

We assume that the rate of translation is limited by ribosome elongation, until all amino acid residues are exhausted, at which point the translation is abruptly stopped. We further ignore possible effects of the mRNA secondary structures and codon usage on translation, and assume that the translation rate of a protein is proportional to the total concentrations of the mRNAs that code for it. With these assumptions, Eq. 2.2 is used to track the concentration of protein *i*:

$$\frac{d[\text{protein}_i]}{dt} = k_E R_d \sum_k [\text{mRNA}_k] - k_{dpi}[\text{protein}_i] - v_d \qquad 2.2$$

where [protein$_i$] is the concentration of the *i*th protein, k_E is the elongation rate of the ribosome, R_d is the ribosomal density along mRNAs, k_{dpi} is the decay rate constant of the protein, v_d accounts for the consumption of the protein by assembly of the procapsids or the packaging of the progeny particles, and the summation accounts for all the mRNA species that code for the *i*th protein. k_{dpi} is assumed to be the same constant for all T7 proteins.

Table 2.4 T7 RNA polymerase promoter data used in the model (Endy 1997) *

Promoter	Relative Strength**	Initiation Efficiency	Relative Activity***
ϕ1.1A	0.15[a]	0.296[c]	0.044
ϕ1.1B	0.34	0.361	0.123
ϕ1.3	0.045	0.163	0.007
ϕ1.5	0.15[a]	0.296[c]	0.044
ϕ1.6	0.15[a]	0.296[c]	0.044
ϕ2.5	0.15[a]	0.296[c]	0.044
ϕ3.8	0.07	0.364	0.025
ϕ4C	0.15[a]	0.296[c]	0.044
ϕ4.3	0.15[a]	0.296[c]	0.044
ϕ4.7	0.15[a]	0.296[c]	0.044
ϕ6.5	0.61	0.748	0.456
ϕ9	0.80[b]	0.721[d]	0.577
ϕ10	1.00	0.681	0.681
ϕ13	0.79	0.734	0.580
ϕ17	0.80[b]	0.721[d]	0.577

* Relative strength and initiation efficiency data taken from (Ikeda & Bailey 1992).

** All promoter strengths are scaled relative to ϕ10.

*** Relative activity is the product of the relative strength and the initiation efficiency.

[a] 0.15 is the average strength of the three known class II promoters.

[b] 0.80 is the average strength of the three known class III promoters.

[c] 0.296 is the average initiation efficiency of the three known class II promoters.

[d] 0.721 is the average initiation efficiency of the three known class III promoters.

In calculating R_d, we assume that the ribosomes are uniformly allocated to all the available mRNAs based on their levels and lengths, and that all the ribosomes are directed to the translation of T7 mRNAs once infection starts. That is, $R_d = \min(N_R / \sum_i L_{\mathrm{mRNA}_i}[\mathrm{mRNA}_i], \quad 1/d_R)$, where the function min(*a*, *b*) returns the smaller of *a* and *b*, N_R is the total number of ribosomes, d_R is the minimum distance between neighboring elongating ribosomes, $L_{\mathrm{mRNA}i}$ and [mRNA_i] are the length and the level of mRNA *i*, respectively, and the summation accounts for all T7 mRNA species. According to this equation, R_d will change with the total mRNA level when the latter is high, but will reach its maximum and become a constant ($1/d_R$) when the latter is low. This dependence of R_d on the mRNA level marks a major difference between T7v2.5 and its precursors where ribosomes were never limiting and R_d was thus always constant.

2.1.6 Protein-protein interactions

We treat the following protein-protein interactions as rapidly equilibrated reactions:

$$\mathrm{EcRNAP} + \mathrm{gp0.7} \Leftrightarrow \mathrm{EcRNAP\text{-}gp0.7}, \quad K_1$$

$$\mathrm{EcRNAP} + \mathrm{gp2} \Leftrightarrow \mathrm{EcRNAP\text{-}gp2}, \quad K_2$$

$$\mathrm{T7RNAP} + \mathrm{gp3.5} \Leftrightarrow \mathrm{T7RNAP\text{-}gp3.5}, \quad K_3$$

where K_1, K_2 and K_3 are the association constants for the corresponding interactions. K_1 and K_2 are derived from published data (Endy 1997), and K_3 is directly taken from

published data (Kumar & Patel 1997). These interactions can be formulated in the same form: A + B ⇔ AB, with an association constant of *K*. They are implemented as constraints on the concentrations of the involved species. At each time point, the concentrations of the free and associated proteins are updated using the following algebraic equations:

$$\begin{cases} [AB] = \left\{ ([A]_0 + [B]_0 + 1/K) - [([A]_0 + [B]_0 + 1/K)^2 - 4[A]_0[B]_0]^{1/2} \right\} / 2 \\ [A] = [A]_0 - [AB] \\ [B] = [B]_0 - [AB] \end{cases} \quad 2.3$$

where the subscript 0 indicates the total concentration of A or B, including both free and bound forms.

2.1.7 Replication of the T7 genome

During the T7 infection cycle, approximately 85% of the host genome is efficiently digested by T7 endonuclease (gp3) and exonuclease (gp6) into acid-soluble fragments during about 7.5 to 15 minutes after infection initiation (Sadowski & Kerr 1970)[10]. We model the digestion of the host genome with a 0th-order reaction and assume instantaneous conversion of the acid-soluble fragments into deoxyribonucleoside triphosphates (dNTPs), which are the direct precursors for DNA synthesis.

[10] These data were measured at 37°C; they may be significantly delayed at 30 °C

In addition to dNTPs, T7 DNA replication *in vivo* requires T7RNAP, DNA polymerase (gp5), primase/helicase (gp4A/B) (Egelman et al 1995; Kusakabe et al 1998; Kusakabe & Richardson 1997; Mendelman et al 1992), single-stranded DNA binding protein (gp2.5) (Kim & Richardson 1993), and probably T7 lysozyme (gp3.5) (Zhang 1995). Gp5 is only processive when it forms a 1:1 complex with thioredoxin, a host protein (Huber et al 1987; Tabor et al 1987). We assume thioredoxin is always in excess and all gp5 molecules are processive. We further assume that a replisome consists of one gp5 molecule and six gp4A molecules and replisomes form instantaneously given stoichiometric amounts of these components. Finally, we assume that up to two replication forks can form per complete T7 genome, given sufficient numbers of replisomes. The roles of T7RNAP, gp2.5, and gp3.5 are accounted for by incorporating "hard" switches such that DNA synthesis occurs only when the concentrations of gp2.5 and the T7RNAP:gp3.5 complex are above zero.

Having met these constraints, we treat replisome elongation as the rate-limiting step in DNA synthesis. This rate is set as a function of dNTP concentrations using Michaelis-Menten kinetics:

$$\frac{d[\mathrm{DNA}]}{dt} = \frac{k_{PD} N_r [\mathrm{dNTP}]}{2([\mathrm{dNTP}] + K_m) L_{\mathrm{DNA}}} - v_{T7} \qquad 2.4$$

$$\frac{d[\mathrm{dNTP}]}{dt} = k_{\mathrm{dNTP}}(t - t_0) - \frac{k_{PD} N_r [\mathrm{dNTP}]}{[\mathrm{dNTP}] + K_m} \qquad 2.5$$

where [DNA] is the number of T7 progeny DNA, [dNTP] is the concentration of free dNTPs, k_{PD} is the elongation rate of the replisome, N_r is the number of elongating replisomes and N_r = min (the number of replisomes, 2[DNA]), K_m is the half-maximum velocity constant, L_{DNA} is the length of a T7 DNA in base pairs, v_{T7} is the production rate of phage progeny, k_{dNTP} is the rate for the release of dNTPs from digestion of the host genome, and t_0 (= 7.5 min) is the time point for the initiation of host DNA digestion (Sadowski & Kerr 1970).

2.1.8 Procapsid assembly and phage particle formation

Procapsid assembly is simulated with a 4.8th-order nucleation-limited reaction derived from data for phage P22 (Prevelige et al 1993), which has similar shape and size as T7. The P22 kinetic data for procapsid assembly is the most comprehensive for any phage and allows the development of a procapsid rate expression. This representation for the formation of procapsids includes the requirement that the major capsid protein concentration exceeds a nucleating concentration (C_n) before assembly starts. The consumption of procapsids as progeny are formed requires complete procapsids, T7 DNAs, and enough other structural proteins to complete the phage. Further, we assume that packaging of DNA into the procapsid is the rate-limiting step for T7 progeny formation, given DNA, procapsids, and non-zero concentrations for all particle proteins and the DNA maturation proteins, gp18 and gp19. Based on these assumptions, we use the following equations to simulate the procapsid assembly and progeny formation:

$$\frac{d[\text{procapsid}]}{dt} = \frac{k_a[\text{gp10A}]^{4.8}}{N_c} - v_{T7} \quad 2.6$$

$$v_{T7} \equiv \frac{d[\text{T7}]}{dt} = k_{pack} P_R \quad 2.7$$

where [procapsid] is the number of procapsids, k_a is the procapsid assembly rate constant, N_c is the number of gp10A molecules per procapsid, k_{pack} is the packaging rate of T7 DNA into the procapsids, and P_R = min ([procapsid], [DNA]).

As procapsids and progeny phage particles are assembled the model accounts for the depletion of various T7 proteins. The model also accounts for the utilization of the scaffold protein during procapsid assembly (Roeder & Sadowski 1977). Based on Eqs. 2.6 & 2.7, we can write out the equation for the consumption terms of different proteins (v_d of Eq 2.2):

$$v_d = N_s \frac{k_a[\text{gp10A}]^{4.8}}{N_c} \text{ for gp9} \quad 2.8$$

$$v_d = k_a[\text{gp10A}]^{4.8} \text{ (if [gp10A] > } C_n\text{) for gp10A} \quad 2.9$$

$$v_d = N_i v_{T7} \text{ for other particle proteins} \quad 2.10$$

where N_s is the number of scaffolding proteins (gp9) consumed for each procapsid produced, N_i is the number of protein i molecules per progeny phage. The stoichiometric data for phage particle formation are listed in Table 2.5.

Table 2.5 Functions and phage particle stoichiometry data of selected T7 proteins (Steven & Trus 1986; Studier & Dunn 1983)

Protein	Function or stoichiometry data
gp0.3	Anti-restriction protein
gp0.7	Protein kinase; also inhibits EcRNAP
gp1	T7 RNA polymerase
gp2	Inhibitor of EcRNAP
gp3	Endonuclease
gp3.5	Lysozyme, inhibitor of gp1
gp4A/B	Primase/helicase
gp5	DNA polymerase
gp6	Exonuclease
gp8	Capsid-tail connection protein (12)*
gp9	Capsid assembly protein (137)*
gp10A	Major capsid protein (415)*
gp11	Tail protein (18)*
gp12	Tail protein (6)*
gp13	Core protein (33)*
gp14	Core protein (18)*
gp15	Core protein (12)*
gp16	Core protein (3)*
gp17	Tail fiber protein (18)*

gp17.5	Lysis protein
gp18	DNA maturation
gp19	DNA maturation

* Number of protein molecules needed for each phage particle.

Table 2.6 The default setting for T7-specific parameters

Parameter	Value	References
T7 DNA translocation rate (bp/s)	70, 40, and 200	(Garcia & Molineux 1995b; Garcia & Molineux 1996; Struthers-Schlinke et al 2000)
Promoter activities		Tables 2.3 and 2.4
TE efficiency	$_eTE = 0.99$	(Dunn & Studier 1983)
Tϕ efficiency	$_eT\phi = 0.85$	(Macdonald et al 1993)
T7 mRNA decay rate constant	$^*k_{dm} = 2 \times 10^{-3}$ s^{-1}	(McCarron & McAllister 1978; Pfennig-Yeh et al 1978; Yamada & Nakada 1976; Yamada et al 1974)
T7 protein decay rate constant	$k_{dp} = 2.85 \times 10^{-5}$ s^{-1}	(Lee & Bailey 1984)
EcRNAP and gp0.7 association constant	$K_1 = 5.5 \times 10^6$ M^{-1}	(Endy et al 1997)
EcRNAP and gp2 association constant	$K_2 = 5.0 \times 10^7$ M^{-1}	(Endy et al 1997)
T7RNAP and gp3.5 association constant	$K_3 = 1.087 \times 10^7$ M^{-1}	(Kumar & Patel 1997)
Replisome elongation rate	$k_{PD} = 370$ bp/s	(Rabkin & Richardson 1990)
Procapsid assembly rate constant	$k_a = 5.085 \times 10^{15}$ (M$^{-3.8}$/s)	(Endy et al 1997)
Procapsid assembly nucleation concentration	$C_n = 6.23 \times 10^{-6}$ M	(Endy et al 1997)
DNA packaging rate constant	467 bp/s	(Son et al 1993)

* The functional half-lives reported for T7 mRNAs vary from 5 to 7 minutes. For simplicity, we assume all the mRNAs have the same decay rate constant ($k_{dm} = 2 \times 10^{-3}$ s^{-1}), which corresponds to a half-life of 5.77 minutes.

2.2 Simulation output

The output of the model, which can be easily configured by the user, includes the simulated state of genome entry and the concentrations of T7 mRNAs, proteins, DNA, procapsids, and progeny, each as a function of the time. The tabulated output files can then be used for detailed analysis. An additional module is designed to wrap the base model to enhance flexibility and simulation capacity. The wrapper takes care of the input/output to the base model and provides functionality for performing sensitivity analysis on the model system (Chapters 3 and 4), simulating the T7 variants with alternative gene orderings (Chapter 4, also in (Endy et al 2000)), either by sliding a genetic element along the genome or randomly shuffling the entire genome, and simulating the optimization of T7 gene ordering or parameter setting using a Monte Carlo algorithm (unpublished).

2.3 Summary of assumptions in the current T7 model

The default parameter set used by the model is summarized in Tables 2.1, 2.3, 2.4 & 2.6; other assumptions follow:

1. The host cell grows at the exponential phase at the time immediately prior to T7 infection; it is a spatially homogeneous resource reservoir with a constant volume

during the infection process; the only change to the host during infection is consumption of the resources by T7.

2. The T7 DNA enters at a constant rate (different for different stages) during each of the three stages of DNA entry -- injection of the first 850 base pairs, EcRNAP-mediated translocation, and T7RNAP-mediated translocation.

3. Active EcRNAP and T7RNAP (a) have constant elongation rates over time, (b) can instantaneously recognize their promoters and effect translocation and transcription, and (c) are instantaneously allocated (and redistributed at each time step) to their promoters that have entered the host cell based on their relative activities

4. All T7 primary mRNAs are instantaneously cleaved at all RNase III sites with 100% efficiency.

5. EcRNAP and T7RNAP do not block each other on the T7 DNA; EcRNAP and T7RNAP do not block the replisomes and vice versa.

6. The number of active ribosomes remains constant during T7 infection.

7. Ribosomes are allocated among available T7 mRNAs based on the levels and the lengths of these mRNAs.

8. The inhibition of EcRNAP by gp0.7 and that by gp2 are non-competitive.

9. gp0.7 inhibits EcRNAP by forming a 1:1 complex.

10. The elongation of the replisomes is the rate-limiting step of T7 DNA replication, and this process follows Michaelis-Menten kinetics with dNTP as the substrate.

11. The concentration of gp4B does not affect replication kinetics.

12. Digestion of the host DNA is a zeroth-order reaction; it will occur if and only if gp3 and gp6 are present and only during 7.5 and 15 minutes after the initiation of infection.

13. Nucleotides released from the digestion of the host DNA are instantaneously converted into dNTPs for incorporation into T7 DNAs.

14. The host protein thioredoxin is always in excess compared with gp5, so that all gp5 molecules are in the processive form.

15. DNA replication occurs only if T7RNAP, gp2.5, and gp3.5 molecules are present.

16. A replisome consists of one gp5 and six gp4A molecules.

17. Up to two replication forks can form on each T7 genome.

18. T7 procapsid assembly follows the same kinetics as that of phage P22.

19. Unlimited energy is available for T7 DNA metabolism.

C h a p t e r 3

EFFECTS OF *E. COLI* PHYSIOLOGY ON PHAGE T7 GROWTH *IN SILICO* AND *IN VIVO*[11]

> *"You are a product of your environment. So choose the environment that will best develop you toward your objective. Analyze your life in terms of its environment. Are the things around you helping you toward success – or are they holding you back?"*
>
> W. Clement Stone

Abstract

Phage development depends not only upon phage functions, but also on the physiological state of the host, characterized by levels and activities of host cellular functions. We established *Escherichia coli* at different physiological states by continuous culture under different dilution rates and then measured their production of phage T7 during a single cycle of infection. We found that the intracellular eclipse time decreased and the rise rate increased as the growth rate of the host increased. To develop mechanistic insight we extended a computer simulation for the growth of phage T7 to account for the physiology of its host. Literature data were used to establish mathematical correlations between host resources and the host growth rate; host resources included the amount of genomic DNA, pool sizes and elongation rates of

[11] The content of this chapter has been published in You, Suthers, and Yin (2002) *J. Bacteriology*. 184: 1888-1894.

RNA polymerases and ribosomes, pool sizes of amino acids and nucleoside triphosphates, and the cell volume. The *in silico* dependence of the phage intracellular rise rate on the host growth rate gave quantitatively good agreement with our *in vivo* results, increasing five-fold for a 2.4-fold increase in host doublings/hr, and the simulated dependence of eclipse time on growth rate agreed qualitatively, deviating by a fixed delay. When the simulation was used to numerically uncouple host resources from the host growth rate, phage growth was found to be most sensitive to the host translation machinery, specifically, the level and elongation rate of the ribosomes. Finally, the simulation was used to follow how bottlenecks to phage growth shift in response to variations in host or phage functions.

3.1 Introduction

Bacteriophage studies played a key role in setting the foundations of molecular biology (Cairns et al 1992). As a result, phage ranks among the best-characterized organisms at the molecular level. Mathematical models or computer simulations can add value to this wealth of phage information by showing how the molecular components and interactions, when taken together, can define developmental processes. For example, in recent years simulations have shown how the physicochemical interactions that govern gene regulation in lambda phage correlate with its lysis-lysogeny decision (Arkin et al 1998; McAdams & Shapiro 1995; Reinitz & Vaisnys 1990; Shea & Ackers 1985), how the coupling of RNA and replicase production enables rapid takeover of the host during phage Qβ infections (Eigen et al 1991), and how the

intracellular development of phage T7 depends on the organization of its genome (Endy 1997; Endy et al 2000; You & Yin 2001).

The relative simplicity of phage developmental processes, compared with those of microbes or higher organisms, is balanced in part by the complexity of the resources they need for growth. Phage require at least nucleic-acid precursors, protein precursors, and translation machinery from their hosts. Consequently, phage infection processes depend not only on the physicochemical characteristics of their genome-encoded functions, but also on the intracellular resources of their hosts, which depend further on the physiological state of their hosts. Studies spanning 60 years have demonstrated this dependence (Adams 1959; Cohen 1947; Cohen 1949; Cohen 1953; Delbrück 1946; Ellis & Delbrück 1939; Hadas et al 1997; Hedén 1951; Kutter et al 1994). Different stages of phage growth, including the attachment of the phage particle to its host, the penetration of phage DNA into its host, and the synthesis of the phage components, have been found to be sensitive to the physiological state of the host, which has been modulated by its growth medium (Cohen 1949; Hadas et al 1997; Hedén 1951), temperature (Ellis & Delbrück 1939; Kutter et al 1994), oxygen tension (Kutter et al 1994), and pretreatment using chemical agents (Adams 1959; Cohen 1949; Hadas et al 1997; Hedén 1951). These studies have shown that the faster the host grows at the time of phage infection, the faster the phage will grow, corresponding to a shorter latent time (or a shorter intracellular eclipse time), faster progeny production rate, and larger burst size.

How is phage production influenced by the levels of different essential intracellular resources? The answer is not obvious because of the coupling of resources to host growth. As the growth rate of the host increases, so does its cell size (Donachie & Robinson 1987), as well as its intracellular levels of genomic DNA, RNA polymerase, ribosomes, nucleoside triphosphates (NTPs), and amino acids (Bremer & Dennis 1996). These are all essential resources that, to unknown extents, affect phage growth. In addition, the effects of these factors may be convoluted: the increase in the cell volume as the host cell grows faster will affect the concentrations of various resources, which in turn may affect phage growth. Mutant hosts may help uncouple these effects. For example, the effects of DNA levels can be uncoupled from growth rate by growing a thy mutant under thymine limitation (Zaritsky & Woldringh 1973). The one-step growth (OSG) behavior of phage T4 on this mutant was used to infer the key role of the protein synthesis machinery in phage growth (Hadas et al 1997). Such results, while useful, do not exclude possible confounding effects of the mutation on other host resources.

Using phage T7 as the model system, we explored the effect of the host physiology on phage growth using both experimental and computational approaches. Phage one-step growth experiments were conducted by infecting *E. coli* cells growing at different rates, achieved by using a chemostat. These experiments suggested a similar dependence of phage growth rate on the host growth rate as observed for other phage studied: the faster the host grows, the faster the phage infection. To better quantitatively understand this dependence, we employed the T7 model to probe the effect of the host physiology at the molecular level. In particular, we explored the effects of host resources

on phage growth by accounting for the coupling or uncoupling of the host resources to its growth rate.

3.2 Materials and methods

3.2.1 Experiment

Strains and growth conditions. *Escherichia coli* BL21 (Gal$^-$ λ^s *hsdS*) and wild-type bacteriophage T7 were generously provided by I. J. Molineux (University of Texas, Austin, TX). BL21 cultures were grown aerobically on Luria-Bertani (LB) broth (Difco) containing 0.4g/L glucose at 30ºC and pH 7 in a 3-liter KLF 2000 Fermenter (Bioengineering AG). The fermenter was inoculated with 20mL of an overnight of BL21 grown on LB at 30ºC. A 500 Series Fermentation Controller (Valley Instrument Company) kept the pH, the temperature, and the impeller rotation rate (300 rpm) at set points. The feed rate to the fermenter was continuous and controlled at various rates, and the working volume was kept constant at 1740 mL using an overflow tube. A MasterFlex pump (ColeParmer) maintained the flows in both the feed and the outflow lines. PharMed MasterFlex tubing (ColeParmer) was used for the feed and overflow tubing, CFlex tubing (ColeParmer) for the 4N H_2SO_4 acid and the 4M NaOH base feeds, and platinum-cured silicone (ColeParmer) for the air flow tubing.

Intracellular one-step growth experiments. Established methods were used in the preparation, preservation, and assay of the phage (Adams 1959; Miller 1992; Studier 1969). Bottom agar for plates and soft agar for overlayers were LB broth containing 1.5% and 0.7% Bacto-agar (Difco), respectively. The host cells went through at least 6-8

doublings after reaching the new steady state for each change in flow rate, as determined by OD_{600}. Sampling of the chemostat was performed under sterile conditions. Ten mL of the sample was added to a shaker flask already at 30ºC, which was then infected with phage T7. The multiplicity of infection was 0.05, and all phage growth was carried out at 30ºC. At three minutes post infection, a sample from the infected shaker flask was diluted 500-fold into 30mL of cell-free LB medium to minimize further binding of phage to bacteria. Samples of the diluted infected cell culture flask were treated with chloroform (Sigma) to release intracellular phage and titered on BL21 grown in shaker flasks on LB. The number of infective centers was determined as the difference between the chloroform-treated sample and a sample that was not chloroform-treated, both taken immediately after dilution. Phage dilutions for titering were performed in buffer containing 10 *mM* Tris–HCl (pH 7.5), 1 *mM* $MgCl_2$, 0.1 *M* NaCl, 10 mg/L gelatin, and 10 *mM* $CaCl_2$. Phage growth curves were generated in triplicate at each host growth rate.

Parameter estimation. The objective function to extract eclipse time, rise rate, and burst size from the intracellular one-step growth curves was defined as

$$\phi = \begin{cases} 0, & 0 < t < a \\ r(t-a), & a < t < \dfrac{B}{r} + a \\ B, & t > \dfrac{B}{r} + a \end{cases}$$

where ϕ is the number of phage progeny as a function of time, a is the eclipse time, r is the rise rate, and B is the burst size. The non-linear fit function (nlinfit) of Matlab 6.0 (The Mathworks, Inc.) was used to estimate the three parameters along with their 95% confidence intervals.

3.2.2 Simulation

We performed a sensitivity analysis on all the *E. coli* physiological parameters that affect the growth of T7 (Table 2.1). Here we use the term "sensitivity analysis" in its broad sense, i.e., the determination of changes in the value of model variables in responding to the change in a parameter value. The base case value of the host growth rate was set to 1.5 doublings/hr, and the corresponding base case values of the other parameters were calculated using the equations in Table 2.1. In each sensitivity analysis, one parameter was varied over a range from 0.1 to 10 times its base case value while the other parameters were kept constant, and a simulation was conducted for each value of the selected parameter. For scenarios where host physiology was defined by a single growth rate, we coupled host resources to the growth rate using the equations in Table 2.1. To provide a basis for comparison with the experiments we conducted, the host growth rate was varied from 0.7 to 1.7 doublings/hr. Each simulation was conducted over a time period of 60 minutes with a time step of 0.02 seconds for numerical integration, except for those where the host-cell volume was less than half its base case value. In this case, a time step of 0.001 seconds was used to avoid computational errors resulting from the integration of stiff equations (e.g., the assembly of procapsid is 4.8th-order in terms of the concentration of major capsid protein; significant computational

errors may occur when this concentration is high, which may arise for small volumes and large time steps). The sensitivity of phage growth to both EcRNAP and ribosomes was studied by generating a mesh of 400 (or 20×20) nodes ranging from 0.1-fold to 100-fold of the base case values, and the rise rate was determined at each node. Derivatives with respect to EcRNAP number and ribosome number were then determined at each node, and local changes in these derivatives were used to demarcate transitions in resource limitation. Specifically, the boundary between the EcRNAP-limited and ribosome-limited growth was defined between nodes where the derivative of the rise rate with respect to the EcRNAP numbers changed from positive to either negative or zero. Further, the boundary between growth limitations by protein synthesis and DNA synthesis was defined between nodes where the derivative of the rise rate with respect to both EcRNAP numbers and ribosome numbers changed from positive to either negative or zero.

3.3 Results

3.3.1 T7 growth was sensitive to E. coli growth rate

The host *E. coli* cell can affect phage T7 growth in many ways. From the perspective of the intracellular processes comprising the phage infection cycle, the *E. coli* cell serves at least as a set of material resources. Additionally, the cell volume may affect viral growth by affecting the concentrations of interacting molecular species. The physiological parameters that characterize the levels of these cellular resources and the cell volume are all closely correlated with the cell growth rate (Bremer & Dennis 1996). Therefore, changes in the cell growth rate will change the physiological parameters; in

turn, these changes may affect the rate of phage development. We found from our experiments and simulations that phage infections indeed became increasingly productive as the growth rate of the host *E. coli* increased from 0.7 to 1.7 doublings/hr (Figure 3.1).

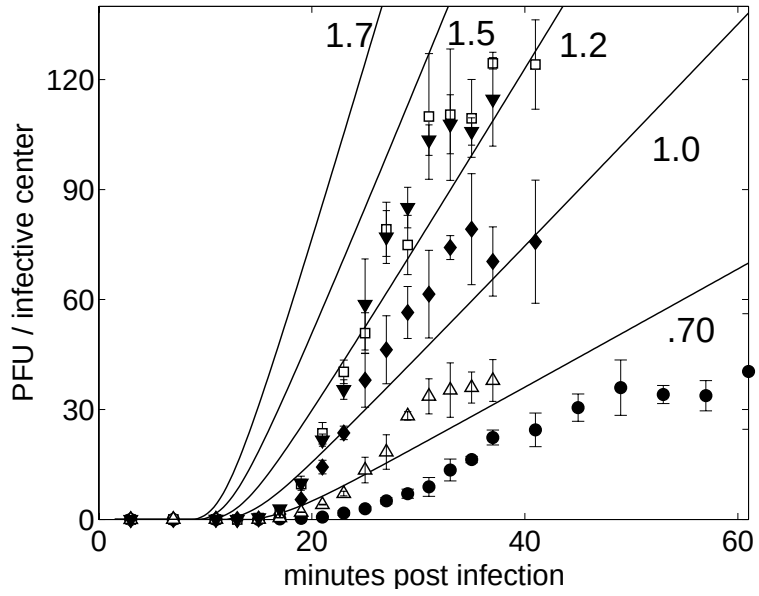

Figure 3.1 Intracellular one-step growth of phage T7 on *E. coli* BL21 growing at different rates. The host cells were grown at 0.7, 1.0, 1.2, 1.5 and 1.7 doublings/hr. Experimental data of the intracellular plaque forming units (PFU) for each host growth rate are from three separate infections indicated by ●, △, ♦, ▼, □, in the order of the growth rates above; output from the simulation is shown by solid lines.

To better compare simulations and experiments, we characterize each intracellular one-step growth curve using three variables: the time required by the infected host to produce the first phage progeny (*eclipse time*), the rate of intracellular phage progeny production (*rise rate*), and the total number of viable progeny per infected host (*burst size*), as shown in Figure 3.2A. Since the mechanism for the phage-induced lysis of the host cell is not well understood and not included in our simulation, we focus here on the

behaviors of the eclipse time and the rise rate, which are lysis-insensitive. The simulation agreed well with experiments in predicting the dependence of the eclipse time and the rise rate on the *E. coli* growth rate (Figure 3.2B-C): the rise rate monotonically increased, and the eclipse time monotonically decreased, with increasing *E. coli* growth rates. The agreement for the rise rate was better overall than for the eclipse time, where there was a systematic mismatch between simulation and experiment (Figure 3.2C, solid line). Also note from Figure 3.2 that the rise rate was much more sensitive to the changes in the *E. coli* growth rate than was the eclipse time. As the *E. coli* growth rate increased from 0.7 to 1.7 doublings/hr, the rise rate increased approximately five-fold, whereas the eclipse time decreased by less than 30 percent. This sensitivity analysis gives an overall picture of the dependence of T7 growth on the *E. coli* physiological state.

3.3.2 Simulated T7 growth most strongly depended on the host translation machinery

To investigate *in silico* the effect of individual *E. coli* physiological parameters on T7 growth, we uncoupled them from the *E. coli* growth rate and from each other, and analyzed the sensitivity of the rise rate and the eclipse time to each, independently. As in the case of the sensitivity analysis on the *E. coli* growth rate (Figure 3.2), the rise rate was overall much more sensitive to the host parameter changes than was the eclipse time. In addition, the rise rate and the eclipse time always changed in inverse directions with any parameter changes; therefore, we focus here on the rise rate.

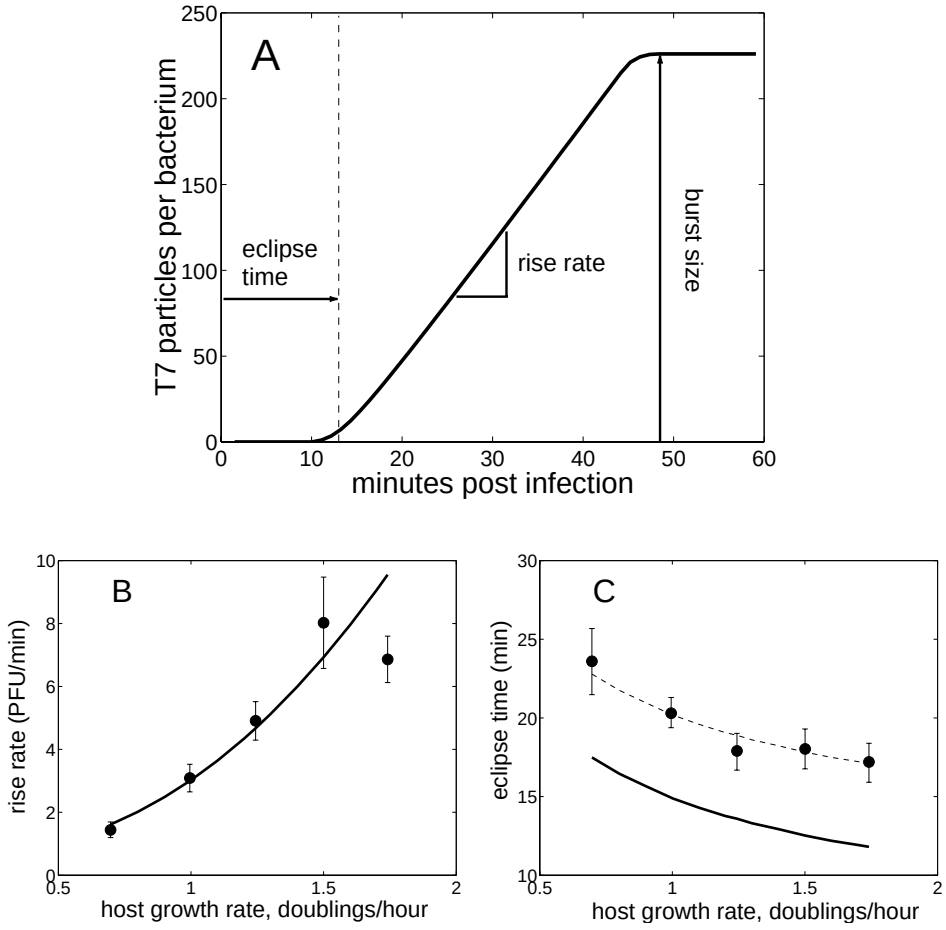

Figure 3.2 Phage T7 growth dependence on host growth rate. (A) An intracellular one-step growth curve can be characterized by three variables: eclipse time is the time period between infection initiation and the time point when phage progeny first appear, the rise rate is the slope of the straight line starting from the end of the eclipse period, and the burst size is the final number of phage progeny produced from a single infection. (B) The intracellular rise rate and (C) the eclipse time, both extracted using a three-parameter model from the data in Figure 3.1, as a function of *E. coli* growth rate. The experimental results are shown in ● and 95% confidence intervals are indicated. Results of processing the computer simulation growth curves are shown by solid lines. A one-parameter adjustment to the eclipse time is shown by a dashed line. This adjustment incorporated a constant delay in the initiation of phage adsorption to the host cell.

The number and the elongation rate of ribosomes had the strongest effect on the rise rate (Figure 3.3). They had virtually the same effect until they were large, where the rise rate leveled off at different levels. The increase in ribosome elongation rates increased rise rate more than did the increase in the ribosome numbers. The dependence of the rise rate on the cell volume was overall moderate, and interestingly, biphasic; the rise rate first increased and then decreased as the cell volume increased from 0.1 to 10 times its base case value (Figure 3.3). The optimal rise rate occurred at the base-case cell volume.

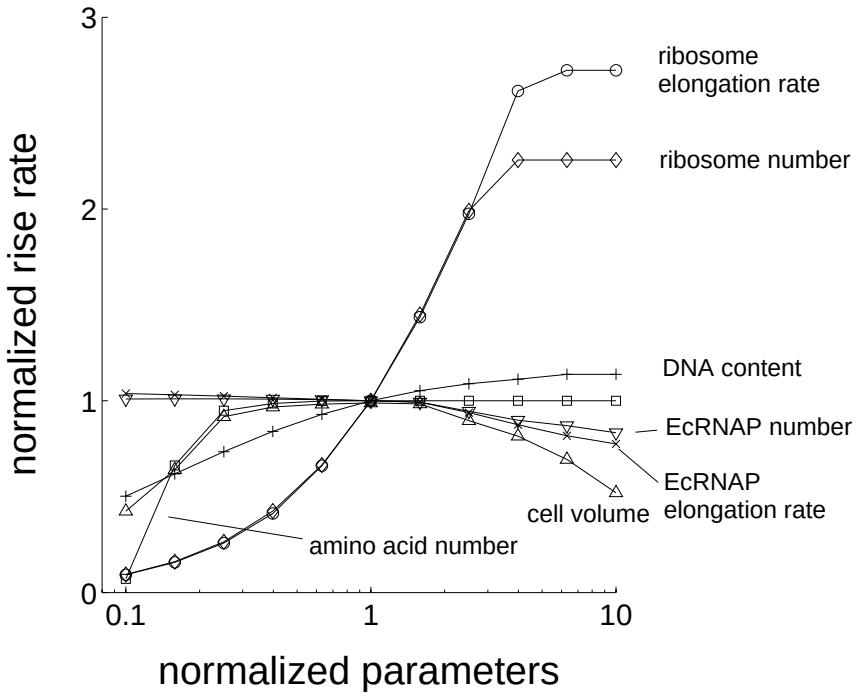

Figure 3.3 Sensitivity of the intracellular rise rate to host physiological parameters. The parameters were normalized to their base case values, which were calculated based on an *E. coli* growth rate of 1.5 doublings/hr using equations from Table 2.1. The rise rate was normalized to the value calculated from the base case parameters. Because NTP number did not affect the rise rate over the parameter range examined, it is omitted from the figure.

The dependence of T7 growth on the other host parameters was less significant (Figure 3.3). An increase in the DNA content had a slightly positive effect on the rise rate. The amino acid pool size did not have any significant effect on the rise rate until it was smaller than about 0.25 times its base case value, when the rise rate became sensitive to the parameter: as the amino acid pool size dropped from 0.25 to 0.1 times its base case value, the rise rate decreased from close to its base case value to nearly zero. The number and the elongation rate of EcRNAPs had similar effects on T7 growth within the range of parameter values examined: the rise rate decreased to a similar degree as either increased (Figure 3.3). In the range of parameter values examined, the variation in the NTP pool size did not have any effect on the rise rate (data not shown).

To better understand how an increase in the EcRNAP number could slow down T7 growth, we examined the effect of the increase in the EcRNAP number on the production of several phage components. We found that, with the other parameters kept constant at their base case values, an increase in the EcRNAP number reduced the rate of procapsid synthesis (Figure 3.4). This reduction corresponded with significant increases in the number of ribosomes allocated to the mRNAs for the early T7 genes, such as 0.7 (T7 kinase gene) and 1 (T7 RNAP gene), and reductions in the number of ribosomes allocated to the mRNAs for the late T7 genes, such as 9 (scaffold protein gene), 10A (major capsid protein gene) and 19 (DNA maturation protein gene) (Figure 3.4B).

3.4 Discussion

We employed experiments and computer simulations of the phage T7 intracellular growth cycle to investigate how the host physiological conditions affect phage growth. Although previous studies have shifted host physiologies in shaker cultures by using media based on different carbon sources (Hadas et al 1997), we found that chemostat cultures using constant growth media but different dilution rates yielded greater flexibility, control, and reproducibility of results. Further, by using chemostat-cultured hosts for our one-step growth studies we provided a better basis for comparison with our simulations, which employed host resource parameters measured from continuous cultures (see Table 2.1).

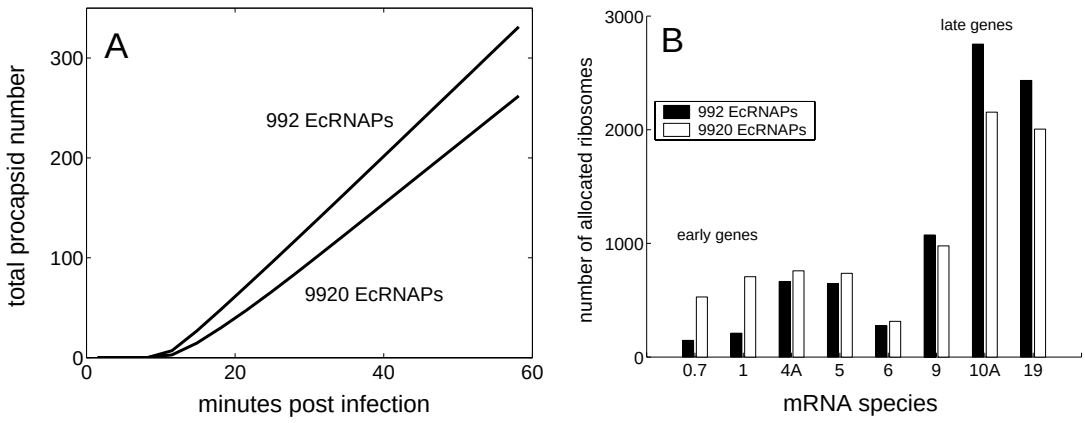

Figure 3.4 Effect of the EcRNAP number on (A) the procapsid assembly process and (B) the allocation of ribosomes to different mRNAs (a snapshot taken at 21 minutes post infection initiation). The total procapsid number accounts for both mature capsids and procapsids.

Consistent with our experimental results, the simulations predicted that faster growing host cells supported faster T7 growth (Figure 3.1, Figure 3.2B-C). The

simulations give overall better predictions on the rise rate than on the eclipse time. The systematic mismatch between predicted and experimental values for eclipse time (Figure 3.2) may result from the instantaneous nature of several "hard-switches" that we implement in the simulation, where data or mechanisms are not available. These switches include the initiation of phage DNA entry, the initiation of T7RNAP-modulated translocation and transcription, and the initiation of host DNA digestion. For instance, we assume a 90-second delay for the initiation of phage infection to account for the time required for phage adsorption, which can take longer. Increasing this delay by 5.3 minutes can eliminate the mismatch (Figure 3.2, dashed line). Nonetheless, this correction does not exclude potential contributions from other factors. Although the intracellular OSG curves cannot distinguish between these mechanisms, available experimental data on the intracellular processes of T7 infection at the molecular level may provide some hints. A close examination of earlier data on T7 protein expression (Endy et al 2000) suggests that the mismatch is unlikely due to a delay in phage absorption because T7 proteins begin to appear around two minutes after infection. Further, since expression of class II proteins initiates in less than two minutes after the appearance of T7RNAPs (Endy et al 2000), the second mechanism cannot completely account for the mismatch either. The last mechanism seems to contribute the major part of this mismatch. In fact, we assumed that the host DNA is digested between 7.5 minutes and 15 minutes after infection. These time points were based on experiments that measured the degradation of host DNA by T7 infection at 37ºC (Sadowski & Kerr 1970), and could be significantly delayed at 30ºC, the temperature

used in the current study. Similar experiments conducted at 30ºC will help to explain the mismatch and improve the simulation by providing more accurate parameters.

Previous efforts to identify key host constraints to phage growth have been hindered by the experimental challenge of independently modulating and studying the effects of different host resources on phage growth (Hadas et al 1997; Hedén 1951). We have addressed this challenge by using the simulation to numerically uncouple the host intracellular resources from the host growth rate and then independently examine their effects on phage growth. While such studies may appear to distort relationships among the host resources, they are useful for two reasons. First, they allow us to better understand how uncertainties in parameters may affect the behavior of the simulation. While empirically determined cell compositions and activities are well defined, it is not always clear what may be accessible for phage growth. For example, NTP pool size in the simulation is based on the total host cell RNA content, which includes both stable and unstable RNAs (Bremer & Dennis 1996). However, we do not know for certain what fraction of the total NTPs per cell are accessible to T7 processes during infection. On the one hand, since stable RNAs constitute parts of ribosomes that are needed by T7 for protein translation, they are improbable precursors for producing T7 mRNAs. On the other hand, the unstable RNAs alone are probably inadequate to support phage growth. Further, the host cell may have means of providing NTPs for phage growth, for example, by continuing to synthesize NTPs during infection; to our knowledge there has been no evidence that T7 infection shuts down host NTP synthesis. Because of the resultant uncertainty in the accessible NTP pool size, it is useful to explore the effects

on phage growth of a broad range of values. Second, studying the sensitivity of the simulated phage growth to extreme changes in single host parameters may reveal aspects of phage development in cells growing in complex environments or in highly distorted hosts like anucleate minicells (Ponta et al 1977). By varying host growth rates in a chemostat we have considered only a narrow range of conceivable host resource distributions. Hosts growing in complex environments, where spatial gradients and dynamic variations in nutrients, temperatures, and pH can be significant, may well have intracellular resource distributions that drastically differ from our experiments. While we lack detailed information about how resources might be correlated in these cases, determining the sensitivity of phage growth to variations in single or pairs of host parameters can begin to provide insights into potential constraints under such conditions.

The sensitivity analysis with the individual host physiological parameters indicates that T7 growth most strongly depends on the number and the elongation rate of the ribosomes, suggesting that the rate-limiting step of phage growth is the synthesis of phage proteins. However, since the ribosome elongation rate only increases slightly as the cell growth rate increases, its contribution is probably less important than the ribosome number, which increases much more rapidly with the cell growth rate (Table 2.1). Thus our simulation results support the hypothesis that the amount per cell of the host translation machinery is one of the most crucial factors in determining the rate of phage growth (Hadas et al 1997). This result is also consistent with the weaker dependence of phage growth on the number and elongation rate of EcRNAPs, the sizes

of the NTP and amino-acid pools, and the DNA content, although all these factors also play indispensable roles during T7 growth.

Because of the complex nature of the viral infection dynamics, increasing the amount of an essential component may actually slow down viral growth, as demonstrated by the counter-intuitive observation that an increase in the EcRNAP number can decrease the T7 growth rate (Figure 3.3). One may expect that increasing the EcRNAP number would increase mRNA levels, protein synthesis rates, and finally the T7 growth rate. This expectation holds when the EcRNAP number is sufficiently small or when the ribosome number is sufficiently high (data not shown). In either scenario, T7 growth is limited by transcription and this limitation can be relieved by an increase in the EcRNAP number. When the EcRNAP number is too large, however, T7 growth will be limited by translation. In this case, increasing the number of EcRNAPs, which are mainly responsible for the transcription of T7 class I genes, will result in a higher level of early mRNAs and divert the ribosomes away from the translation of several late proteins that are needed in large quantities for phage particle formation (Figure 3.4B), particularly the scaffold protein (gp9) and the major capsid protein (gp10A). Gene product 10A is especially important for two reasons: first, compared with other T7 proteins, it is needed in the largest quantity for each T7 particle (Steven & Trus 1986); second, the rate of procapsid assembly is assumed to be proportional to $[gp10A]^{4.8}$ (Endy 1997; Endy et al 1997). Hence, even a slight decrease in gp10A level may significantly decrease the procapsid assembly rate (Figure 3.4A) and in turn significantly decrease the rise rate. In summary, when translation is already limited by the

ribosome number, an increase in EcRNAP number causes ribosomes to distribute in a manner that is unfavorable for the expression of important late proteins, thus reducing the rise rate (Figure 3.3).

Our simulations indicate that T7 growth is more sensitive to the processivity of T7 RNAPs than to the processivity of EcRNAPs, although the effect of either factor is much weaker than that of the ribosomes (see Figure 4.2). This is to be expected because the transcription of T7 genes is primarily dependent on the T7 RNA polymerase, which is more efficient than EcRNAP in internalizing and transcribing T7 DNA. This point further suggests that T7 growth limitation by the host translation machinery may well be due to the fact that the phage provides its own highly efficient RNA polymerase, thus creating a bottleneck for processing of its mRNAs.

The biphasic effect of the cell volume on T7 growth is another demonstration of the complexity of the infection dynamics. A large cell volume may slow down the intracellular interactions by causing a decrease in the concentrations of interacting species. This is particularly true for the assembly of procapsids from major capsid proteins (gp10A). An increase in the cell volume can result in a decrease in the gp10A concentration, which in turn can cause a major decrease in the procapsid assembly rate. When the cell volume is too small, however, other effects may dominate: the inhibition of EcRNAP by gp0.7 and gp2, and the inhibition of T7RNAP by gp3.5 will be enhanced because of the increased total concentrations for these proteins. Consequently, under such conditions fewer free T7RNAP and EcRNAP molecules are

available for the transcription of viral genes, and the rate of viral protein expression decreases, which in turn reduces the rate of phage progeny formation. This idea is supported by the observation that the numbers of active EcRNAPs and T7RNAPs decrease with a decreasing cell volume (data not shown). In summary, extremes in cell volume can, through different mechanisms, have detrimental effects on T7 growth.

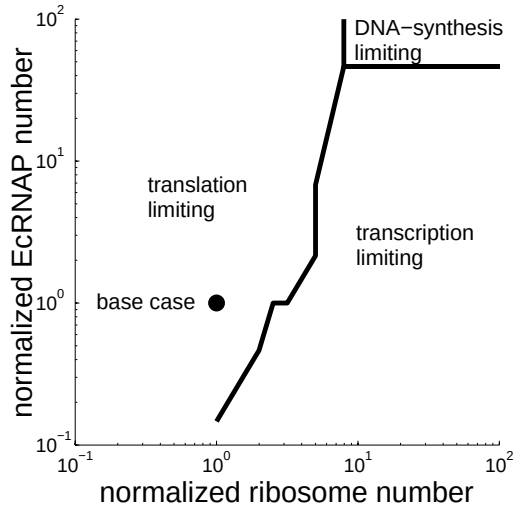

Figure 3.5 A "bottleneck landscape" for phage T7 growth with respect to two host parameters, the levels of the host RNA polymerase (EcRNAP) and ribosome. At any point on this figure, phage growth is limited by the rate of translation by the host ribosomes, transcription by the EcRNAP, or by the rate of phage DNA synthesis. The base case setting is labeled by the filled circle.

As the activity of host or phage functions vary, so too may bottlenecks to phage production, as shown in the "bottleneck landscape" of Figure 3.5. For the base case parameters T7 growth is bottlenecked or limited by the rate of protein synthesis, determined by the level of ribosomes. If one relieves this bottleneck by increasing the ribosome number, then the phage growth becomes limited by the host transcription rate, determined by the level of EcRNAP. When levels of both ribosomes and

EcRNAPs are large, then the bottleneck to phage growth can become the rate of its own DNA synthesis. While such a scenario may be unlikely for wild-type phage growth on a wild-type host, due to the greater than 10-fold increases that would be required for both the EcRNAP and ribosome levels, phage mutants that expressed reduced DNA synthetic capacities could alter the landscape by expanding the DNA-synthesis limiting region. In summary, specific features of the bottleneck landscape will change with genetic or environmental modifications to phage or host functions, but so long as the simulation accounts for the modifications, the corresponding landscapes can be created.

Although simulations can facilitate the analysis of complex processes like the growth of phage T7 in its host cell, what the results of simulations truly reflect is the behavior of the model. How well the model predictions apply to the real system depends on the accuracy and completeness of the knowledge-base and the validity of the simplifying assumptions implemented in the model. The knowledge-base in turn has been accumulated and will be enriched through continued laboratory experiments. To this end, simulations may facilitate the learning process by revealing key deficiencies or inconsistencies in the knowledge-base and by making experimentally testable predictions. For example, the prediction that the rate of T7 growth was limited primarily by the synthesis of the late proteins, especially the major capsid protein (gp10A), could be tested by designing specialized ribosomes (Hui & de Boer 1987) that preferentially translated gp10A mRNA. If the simulation is correct, then enhanced expression of gp10A should increase the rate of T7 progeny formation. If experimental results do not match the prediction, then the mismatch serves as a foundation for refining the model.

It is the iterative experimental testing and refinement of simulation by experiments that gradually promotes a systems-level integration of data, and aims ultimately to shed light on relationships that would be otherwise difficult to observe.

Acknowledgments

We thank Jeremy Francken and William Wodicka for their technical assistance and Dr. Ranjan Srivastava for reading and commenting on the manuscript. We also thank Dr. Ian J. Molineux for helpful discussions, and two anonymous reviewers for their comments and suggestions. This work was supported by 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). P.S. was supported by a National Science Foundation Graduate Fellowship.

C h a p t e r 4

PROBING THE DESIGN OF PHAGE T7 *IN SILICO*[12]

> *"Designers must do two seemingly contradictory things at the same time: They must design for perfection, and they must design as though errors are inevitable. And they must do the second without compromising the first."*
>
> Bob Colwell

Abstract

We aim to better understand how evolution of organisms depends on constraints imposed by their environments. Insights may be gained by probing how the growth of viruses depends on the intracellular resources provided by their host cells. We employed an *in silico* approach to study how phage T7 growth depends on the intracellular resources of its *Escherichia coli* host. We explored how T7 growth would be affected by a broad range of *in silico* mutations, spanning from perturbations in the transcriptional activity from single promoters to vast transformations that were attained by randomly rearranging genetic elements in the T7 genome. We generated more than 150,000 T7 mutants and evaluated their growth in two host environments: (1) a limited environment that corresponds to transcription, translation, and DNA replication resources of an *E. coli* cell growing at 1.0 doublings/hr, and (2) an unlimited environment where all host resources are unbounded. We found that in the unlimited environment many mutants

[12] The content of this chapter is being prepared for publication (with John Yin).

grew faster than wild-type T7. However, most of these winners grew more slowly than the wild type in the limited environment. The wild-type growth was indeed nearly optimal in the limited environment. Our results suggest that limited host environments played an important role in the evolution of both the genome organization and regulatory activities of phage T7.

4.1 Introduction

All naturally existing organisms are winners in the competition for survival during evolution. Why have they succeeded? One may argue that the design of these winners, for example the organization of underlying reaction networks, *somehow* guarantee their success. With increasingly detailed understanding of many prototype biological systems, we are beginning to understand, at the molecular level, *how* the interaction between individual components provides key properties of the functioning of the overall system (Alon et al 1999; Barkai & Leibler 1997; von Dassow et al 2000). For example, both computer simulations (Barkai & Leibler 1997) and experiments (Alon et al 1999) illustrated that the structure of the *E. coli* chemotaxis signaling network leads to robust behaviors in response to perturbations in the biochemical parameters of the network. Yi and coworkers analyzed the mathematical model by Barkai & Leibler (Barkai & Leibler 1997) from a system control perspective, and concluded that the robustness of the system can be attributed to an integral feedback loop imbedded in the network (Yi et al 2000). Further, computer simulations suggested that the segmentation network in *Drosophila* embryonic development is a robust functional module that ensures the correct patterning of the *Drosophila* embryo (von Dassow et al 2000).

These pioneering studies provide an integrated view of how components of a biological system interact with one another to ensure its proper functioning. Moreover, they provide guidance for further work to elucidate how the design of an organism ensures the survival of an organism, although these studies focused on biochemical networks isolated from a broader context – the organisms that these networks are part of. It remains a major challenge to characterize the growth of a whole cell at the genetic level. However, simpler and better-characterized organisms, such as bacteriophage T7, may provide an opportunity for exploring how the design of an organism facilitates its survival. Because of its nearly 50-year foundation of genetic, physiological, biochemical, and biophysical data (Dunn & Studier 1983; Molineux 1999; Studier & Dunn 1983), phage T7 serves as an excellent model organism to explore the genotype-phenotype relationship mechanistically. By incorporating the existing experimental data and mechanisms on bacteriophage T7 biology, we previously developed a genetically structured model of T7 intracellular growth. The model accounts for all the major steps of T7 infection, including entry of the viral genome into the host cell, transcription of viral genes, translation of mRNAs, DNA replication, procapsid assembly, and finally, formation of viral progeny.

In essence, the T7 model bridges the genotype of a phage with its phenotype. This feature offers an opportunity to examine the design principles of this virus by computer simulations. Previously, we explored this issue by testing the dependence of phage T7 growth on the ordering of genetic elements (Endy et al 2000). One interesting observation was that close to three percent of T7 mutants with random genomes grew

faster than the wild type. In light of the significant percentage of faster growing mutants, a question naturally follows: why has evolution selected the wild type instead of these mutants? Here we argue that the presence of a high percentage of faster growing mutants is primarily due to the model assumption that the virus has access to unlimited levels of ribosomes, amino acids, and nucleoside triphosphates (NTP) provided by the host cell. In reality the host cell will have limited amounts of all these resources; thus it seems plausible that phage T7 may have been optimized for such limited environments after millions of years of evolution. In this work we tested this hypothesis by investigating how phage T7 would respond to various perturbations in two contrasting environments: one has limited resources but the other has unlimited resources. If our hypothesis were correct, we would anticipate that the wild-type T7, compared with mutants, would grow faster in a limited environment than in an unlimited one.

4.2 Methods

4.2.1 Computer simulations

We explored the effects on T7 growth of four types of perturbations as detailed below. All these analyses were carried out in two contrasting environments: the limited and unlimited environments. The host physiological parameters that define the limited environment correspond to a host cell growing at 1.0 doublings/hr (Table 2.1). The unlimited environment is the same as the limited one except that it has infinite levels of EcRNAPs, ribosomes, NTPs, amino acids, and DNA contents in the cell.

In analyzing the results, the growth rate was used as the measure of T7 intracellular growth (Figure 4.1). Although alternative growth measures, for example the maximum doubling rate (Endy et al 2000) and the rise rate (You et al 2002), could be used instead, conclusions of this study would remain the same. When comparing the performance of a T7 mutant in two environments, we focused on the growth rate relative to the wide-type growth rate in the corresponding environments. By doing this, we were able to evaluate the performance of the wild-type T7 relative to mutants in either environment. If we were to compare the absolute growth rates directly, the same mutant would always grow faster in the unlimited environment than in the limited environment.

Single-parameter sensitivity analysis. We performed a sensitivity analysis on 30 model parameters. These parameters included relative activities of six EcRNAP promoters (A1, A2, A3, B, C, and E), relative activities of 15 T7RNAP promoters (1.1A, 1.1B, 1.3, 1.5, 1.6, 2.5, 3.8, 4C, 4.3, 4.7, 6.5, 9, 10, 13 and 17), T7RNAP elongation rate (k_{PT7}), T7 DNA polymerase elongation rate (k_{PD}), T7 procapsid assembly rate constant (k_a), T7 DNA packaging rate (k_{pack}), gp0.7-EcRNAP association constant (K_1), gp2-EcRNAP association constant (K_2), T7RNAP-gp3.5 association constant (K_3), degradation rate constants of T7 mRNAs (k_{dm}) and T7 proteins (k_{dp}). Default values of these parameters are listed in Chapter 2. In each sensitivity analysis, one parameter was varied over a range from 0.1 to 10 times its base-case value while the other parameters were kept constant, and a simulation was conducted for each value of the selected parameter.

Random mutagenesis in T7 promoters. We generated 50,000 T7 mutants that each had random values for 28 out of the 30 parameters listed above. Each random value was selected within the range from 0.1 to 10 times the base value of the corresponding parameter following a uniform distribution on a logarithmic scale. Since relative activities of promoters were used as weighting factors for distributing EcRNAPs or T7RNAPs, it is necessary, for either RNAP, to keep one promoter activity constant while changing those of others. In this work, we chose to hold constant the relative activities of promoter 10 (for T7RNAP) and promoter A1 (for EcRNAP).

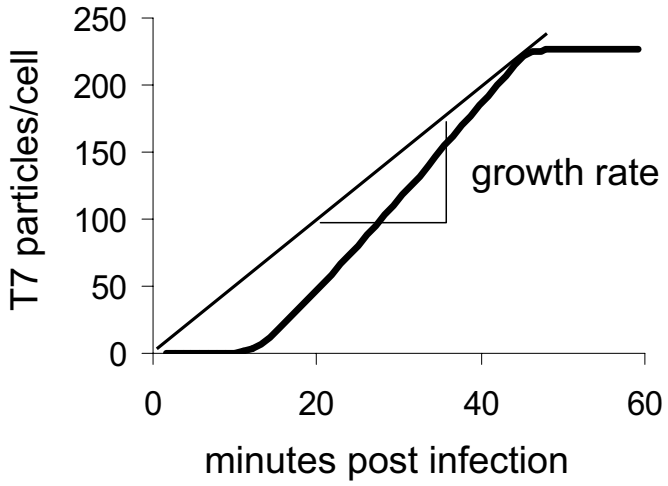

Figure 4.1 A simulated T7 intracellular growth curve. The growth rate is defined as the slope of the straight line starting from the origin and in tangent with the growth curve.

Sliding mutations. In a sliding mutation, an internal T7 genetic element was moved from its wild-type position along the genome to a different position. Here an internal genetic element is any element other than the left or right T7 genome end. The relocation of an element may affect T7 growth by affecting the timing of expression for

some genes because the latter is coupled with entry of T7 genome (Studier & Dunn 1983). To test the sensitivity of T7 growth with respect to the position of selected elements, we moved each selected element to 72 possible positions (including its wild-type position), and at each location we evaluated the corresponding T7 growth rate by simulating the growth of the strain.

Genomic permutations. We generated 100,000 T7 mutants by randomly permuting the 72 internal elements of the T7 genome. The permutation space is astronomically huge: the total number of possible permuted genomes is $72! \approx 6 \times 10^{103}$, and the 100,000 mutants accounted for only an infinitesimally small sampling of the entire space. However, they seemed to be sufficient to give a representative view: the distribution of the growth rates remained overall similar for sample sizes above 10,000.

4.3 Results

4.3.1 Effects of parametric perturbations

The sensitivity of the T7 growth rate with respect to 30 T7 parameters in two contrasting environments is shown in Figure 4.2. If we define "robustness" as the ability to grow in a diversity of environments, T7 seemed to be robust: it grew in both environments despite up to 10-fold change in these parameters from their base value. This is particularly true for the unlimited environment, where the decrease in the growth rate was within 50% for a 100-fold change in the activity of any promoter (Figure 4.2A). Also, in the unlimited environment, there was much room for T7 to grow much faster: decrease in decay rate constant of mRNAs (k_{dm}) and increase in T7RNAP elongation

rate (k_{PT7}) both led to significant increase in T7 growth rate (Figure 4.2C). Overall, the wild-type T7 growth rate was nearly in the middle among mutants in the unlimited environment. In the limited environment, T7 growth responded more negatively to perturbations in the parameters. The growth rate could decrease to below 30% of the wild-type value when any of promoters 10, 13, 17 was too strong (Figure 4.2B). In contrast with the case of the unlimited environment, the growth rate for the wild type was close to optimal compared with those of mutants in the limited environment (Figure 4.2B, D).

Random mutagenesis with 28 parameters led to similar observations (Figure 4.3). All 50,000 randomly created mutants grew in the unlimited environment and 99.95% (49976) grew in the limited environment. In the unlimited environment, 24% (11983) of the mutants grew faster than the wild type (Figure 4.3A, shaded area), with an average growth rate of 0.98 and a maximum growth rate of 43.7. However, merely 5.3% (2657) outperformed the wild type in the limited environment (Figure 4.3B, shaded area). The average and maximum growth rates were 0.51 and 1.7, respectively, again highlighting the relative optimality of the wild type in this scenario. The comparison of relative growth rates within the two environments can be more clearly seen in Figure 4.3C. The growth rates of these mutants distributed broadly in the unlimited environment (Y-axis of Figure 4.3C), and the wild-type value was close to the minimum on a linear scale. The distribution was much narrower in the limited environment, especially if measured on a linear scale, with the wild type close to the maximum. Interestingly, some mutants clustered into an apparent straight line in parallel with the parity line (Figure 4.3C),

which indicates that these mutants grew proportionally slower in the unlimited environment.

4.3.2 Effects of perturbations in genomic structure

Next we analyzed the sensitivity of the simulated growth rate to the position of T7 genetic elements along the genome (Figure 4.4). Genes 1, 10 and promoters 9, 10 were selected for detailed analysis. The two genes were chosen because they play a central role in T7 development. The product of gene 1, T7RNAP, is responsible for the entry of about 85% of the T7 genome and the transcription of class II and class III genes (Dunn & Studier 1983), and T7RNAP is needed for T7 DNA replication as well (Molineux 1999). The product of gene 10, gp10A (the major capsid protein), is needed in large quantities (415 molecules per procapsid) for assembling the viral procapsid. Another reason for choosing gene 10 is that the previous simulations have suggested the synthesis of gp10A was limiting T7 growth in a natural host (You et al 2002). For similar reasons, promoter 9 (ϕ9) and T7RNAP terminator (Tϕ) were selected because they are important in regulating the expression of gp10A.

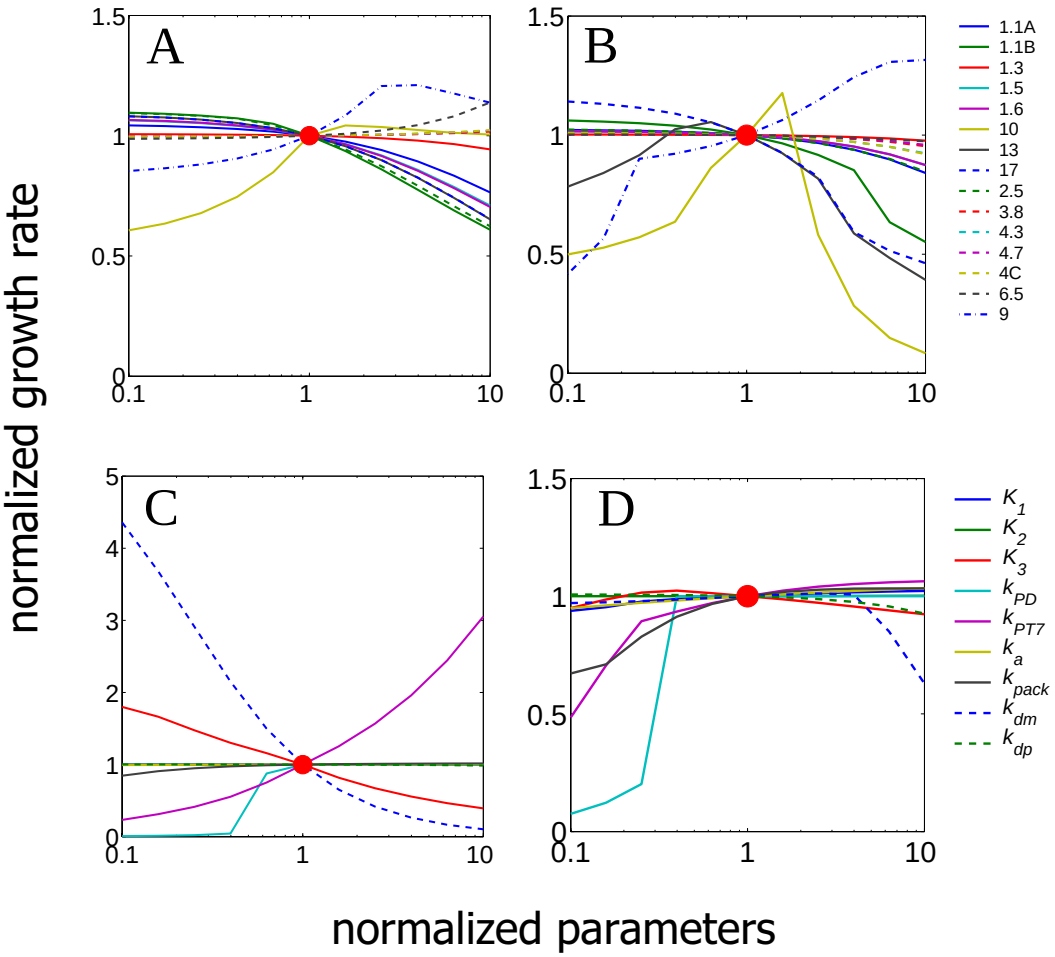

Figure 4.2 The sensitivity of T7 growth rate with respect to T7 promoter activities (A & B) and several other parameters (C & D) for unlimited host environment (A & C), and limited host environment (B & D). The x-axis represents parameters normalized with respect to their default values. The y-axes indicate the corresponding growth rate calculated for the chosen parameters. Growth rates were normalized with respect to the values calculated from the base-case parameter setting for either environment. Along each curve one parameter is varied while the others were kept constant. Filled red circles represent the wild-type T7. Since changes in the activities of six *E. coli* promoters did not have significant effect on the T7 growth rate, they are omitted from this figure.

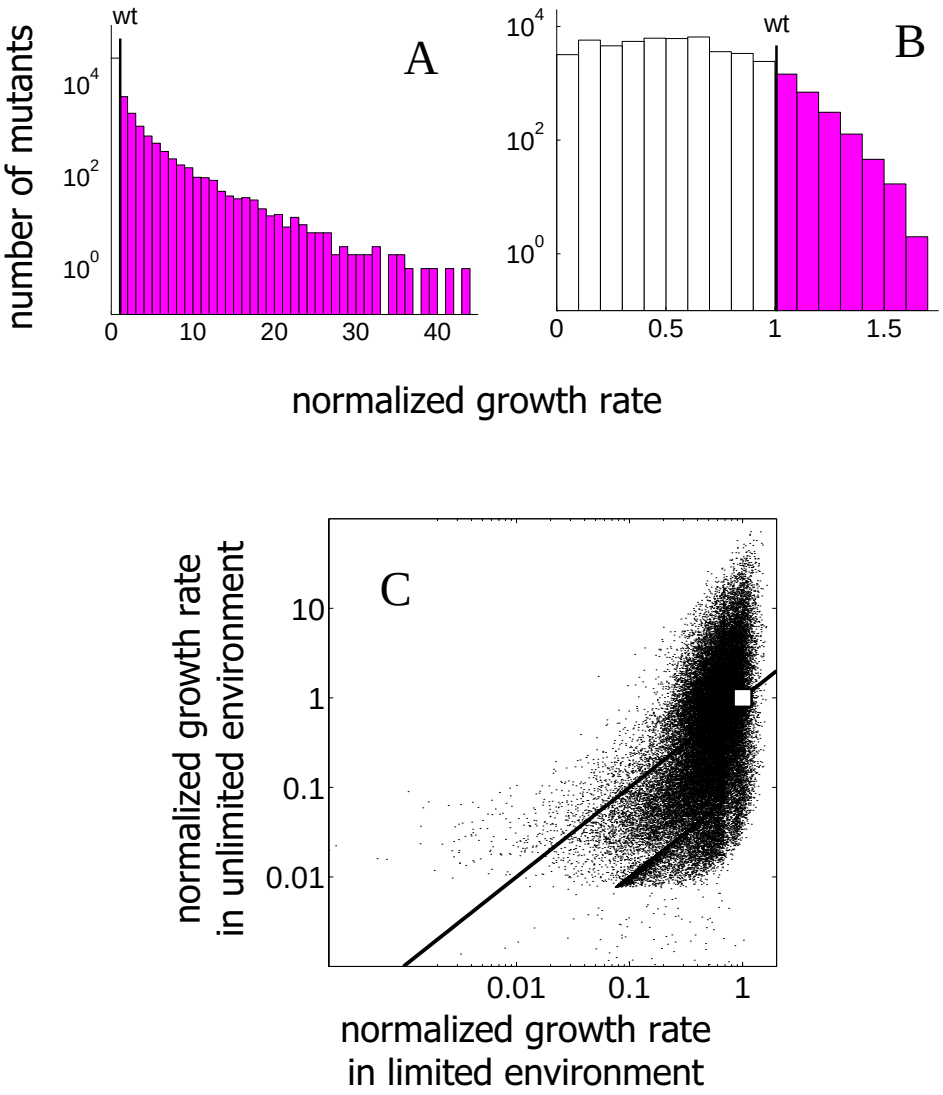

Figure 4.3 Growth of 50,000 T7 mutants with random parameters. Figure shows the distribution of normalized growth rates in (A) the unlimited environment, and (B) the limited environment, as well as (C) the parity plot of the normalized growth rates in the two environments, where each dot represents a mutant and its coordinates correspond to its normalized growth rates in the two environments. In (A) and (B), the x-axes represent growth rates normalized to the wild-type value in either environment, and the y-axes represent the number of mutants with growth rates falling into the corresponding bin. The wild-type growth rate is shown by the vertical red lines in (A) and (B), and by the open square in (C). Along the straight line in (C), T7 mutants have the same growth rates relative to the wild-type growth rates in the two environments.

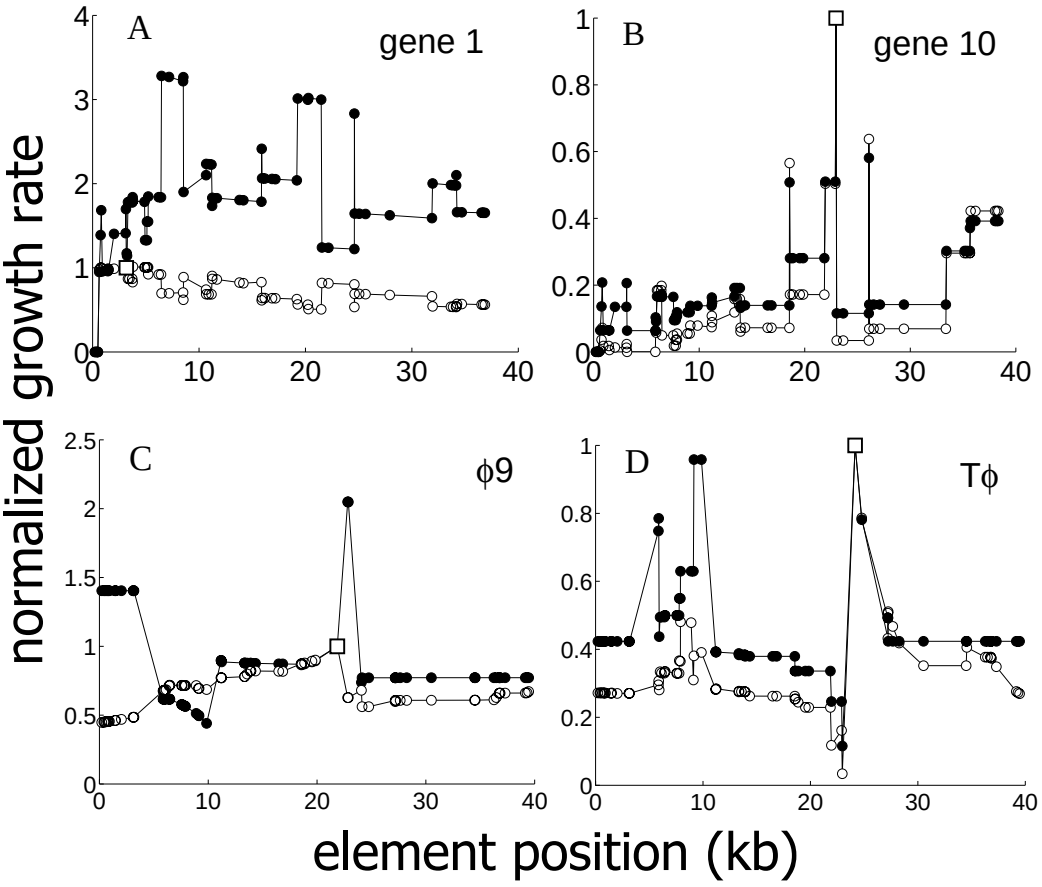

Figure 4.4 Growth of selected T7 slider mutants in the unlimited environment (filled circles) and the limited environment (open circles). The elements selected for sliding are (A) gene 1, (B) gene 10, (C) promoter 9, and (D) T7RNAP terminator. The x-axes represent the position of the elements along the genome, and the y-axes represent growth rates normalized to the wild-type value in the corresponding environment. The wild-type position of each element is indicated by an open square.

For gene 1 (Figure 4.4A), at the wild-type position T7 achieved nearly its largest growth rate (0.99 times the maximum) in the limited environment, suggesting the optimality of the wild type genome arrangement in this environment. In addition, sliding gene 1 along the genome did not have a strong effect on the growth rate. In sharp

contrast, the wild-type gene 1 position led to nearly the lowest growth rate in the unlimited environment (Figure 4.4A, filled circles). Placing gene 1 downstream of its wild-type position actually gave rise to significantly increased growth rate in the unlimited environment. The growth rate was also more sensitive to the gene 1 position in the unlimited environment. Note that, in both scenarios T7 failed to grow when gene 1 was placed close to the left end of the genome because gene 1 was then upstream of all promoters and not expressed. This point is also true for other essential T7 genes, including gene 10 (Figure 4.4B). Gene 10 contrasts with genes 1 in that its wild-type position was the distinctively the best in both environments, and the growth rate had almost the same dependence on the gene 10 position in the two environments; moving gene 10 to any position other than its wild-type one would dramatically slow down T7 growth (Figure 4.4B).

The wild-type position of ϕ9 was optimal in the limited environment but far from optimal in the unlimited environment (Figure 4.4C), the position immediately downstream the wild-type position, that of gene 9, led to the distinctively larger growth rate. In other words, switching positions of gene 9 and its promoter in the unlimited environment led to more T7 productive growth. Placing ϕ9 in any position upstream of gene 1 (around 5kb) also led to significantly elevated T7 growth in the unlimited environment. Finally, the wild-type position of Tϕ, like gene 10, gave maximum growth in both environments, although the superiority of the wild type was much more distinctive in the limited environment. In the unlimited environment, a couple of

adjacent positions before gene 3.5 (around 10 kb) led to growth rates close to the wild-type value.

All cases in Figure 4.4 suggest that the wild-type position for the selected genetic elements resulted in optimal or close to optimal growth in the limited environment, but often had better alternative positions in the unlimited environment. Although we show here only four cases, sliding mutations with other elements yielded the same conclusion (not shown).

Compared with other types of perturbations, random permutations of the genome had much stronger effects on T7 growth. More than 80% of 100,000 T7 mutants with random genomes failed to grow in either the limited environment (83663 dead) or the unlimited environment (82414 dead). Among the viable mutants, 4915 grew faster than the wild type in the unlimited environment, in which the maximum growth rate was nearly seven times greater than the wild-type value (Figure 4.5A). In contrast, no mutants grew faster than the wild type in the limited environment, and the maximum growth rate was only about 97% the wild-type value. Moreover, the vast majority (99%) of the viable mutants grew comparatively faster in the unlimited environment than in the limited environment (Figure 4.5C).

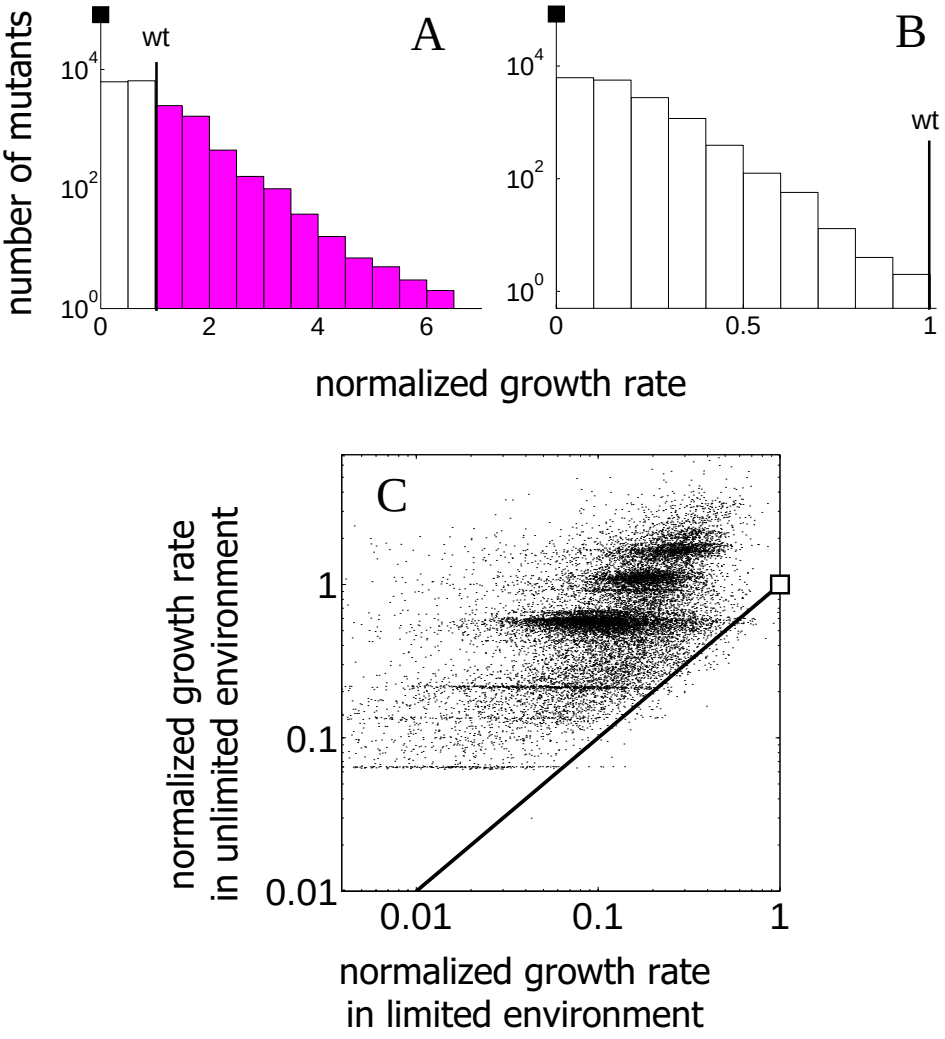

Figure 4.5 Growth of 100,000 T7 mutants with permuted genomes. Figure shows the distribution of normalized growth rates in (A) the unlimited environment, and (B) the limited environment, as well as (C) the parity plot of the normalized growth rates in the two environments. The wild-type growth rate is shown by the vertical red lines in (A) and (B), and by the open square in (C). Along the straight line in (C), T7 mutants have the same growth rates relative to the wild-type growth rates in the two environments. In (A) and (B), the filled squares indicate the number of dead mutants in either environment. Since (C) is plotted in log scale along both axes, all the dead mutants (growth rate = 0) are omitted.

4.4 Discussion

The response of T7 growth with respect to four types of perturbations in two contrasting host environments may provide insight into the design principles of phage T7 as a complex biological system in the context of its environment.

4.4.1 Robustness and fragility

As evidenced in Figures 4.2 and 4.3, T7 is robust to perturbations in its parameter values. The relative insensitivity of T7 growth on individual promoter activities (Figures 4.2A & B) is primarily due to the redundancy of these promoters (Figure 2.2). A close inspection of the T7 genome suggests that none of the promoters for the class II genes is absolutely essential. Another reason for the apparent insensitivity is that promoter activities come into play only by affecting the distribution of T7RNAPs to different genes, in turn affecting the production rate of different proteins. Since resources are always abundant in the unlimited environment, changes in the promoter activities will normally have smaller effects on T7 growth. In the limited environment, effects of such perturbations are more profound because the redistribution of transcription resources will lead to the redistribution of translation machinery as well. For instance, an increase in ϕ9 activity from 0.1 to 10 times its base value results in about a 3-fold increase in the growth rate. The reason for this increase is that more mRNAs will be produced from genes 9 and 10 when ϕ9 is stronger, resulting in increased production rate for gp9 (scaffold protein) and gp10A. This observation supports the notion that, in a limited environment, T7 growth is primarily limited by the synthesis of gp9 and gp10A (You et al 2002; You & Yin 2001). The same argument can be used to explain the dependence

of T7 growth rate on ϕ10 activity. The growth rate increases significantly when the activity of ϕ10 is increased from 0.1 times to about two fold its base value, but sharply decreases with further increase in ϕ10 activity. The initial increase clearly results from the increased production rate for gp10A, but the subsequent decrease is due to the overproduction of mRNAs from gene 10, which divert ribosomes from producing other proteins that are needed for progeny formation, particularly gp9.

Changes in several other parameters have moderate to strong effects on T7 growth, particularly in the unlimited environment, because they significantly affect key processes of viral infection (Figures 4.2C & D). The growth rate is highly sensitive to changes in the DNA polymerase elongation rate (k_{PD}) in both environments when kPD is small, because DNA synthesis becomes rate limiting in this scenario. Further, for T7 mutants with k_{PD} between 0.1 to 0.3 times its base-case value, the normalized growth rates in the limited environment are proportionally faster than the corresponding values in the unlimited environment (not shown). Because T7 growth is highly sensitive to k_{PD} values in this range, the effect of k_{PD} variations will dominate the effect of changes in other parameters, which in turn leads to the apparent clustering of mutants into a straight line (Figure 4.3C). An increase in the T7RNAP causes an increase in the growth rate in both environments by facilitating the transcription process. However, while this resulting increase is nearly quadratic in the unlimited environment, it quickly flattens out in the limited environment, indicating additional control in the limited environment.

This saturation is indeed due to the limitation in translation rate in the limited environment where only finite numbers of ribosomes are available.

In addition to its robustness to parametric perturbations, T7 is robust with respect to sliding mutations because the majority of the slider mutants are viable (Figure 4.4). Considering the central role that gene 1 plays in the phage metabolism, it is interesting to note the rather insensitive dependence of the growth rate on the gene 1 position, particularly in the limited environment (Figure 4.4A). In the unlimited environment, the growth rate increases significantly when gene 1 is moved downstream of its wild-type position; this results from the autocatalytic loop for the production of T7RNAPs when gene 1 is placed at the downstream of a T7RNAP promoter. In fact, any position downstream of the wild-type position will result in such a configuration. Note that the autocatalytic loop is absent in the wild-type genome, where gene 1 is upstream of all T7RNAP promoters (Figure 2.2). The same mechanism is responsible for the elevated T7 growth when ϕ9 is placed upstream of gene 1 (Figure 4.4C). While enjoying significant increase, the growth rate does not skyrocket as expected for an autocatalytic reaction loop (the virus should be able to afford to fast production of any viral components in the unlimited environment). The stability of the growth rate in this case turns out to be maintained by a negative feedback loop that regulates the level of T7RNAP: T7RNAP expresses gene 3.5, whose product, the T7 lysozyme, inhibits T7RNAP (Zhang & Studier 1997) (see also Figure 6.5). In the limited environment, an additional negative feedback loop further regulates the level of T7RNAP. The protein product of gene 2 (another gene transcribed by T7RNAP) inhibits EcRNAP, which in

turn is responsible for the transcription of T7RNAP (Figure 6.5). This second feedback loop does not function in the unlimited environment, because finite amount of gp2 will not affect the transcription by infinite amount of EcRNAP. Therefore the T7RNAP level is more stringently regulated in a limited environment, and as a result the T7 growth rate is less dependent on the position of gene 1. The presence of these loops provides further evidence for the ubiquity of negative feedback control in maintaining proper functioning of biological systems (Hartman et al 2001; Rao & Arkin 2001; Yi et al 2000).

The superiority of wild-type gene 10 position in both environments underscores the requirement for high production rate for gp10. The T7 genome seems to have evolved to meet this requirement. The two strongest T7 promoters (ϕ9 and ϕ10) are located immediately before gene 10 (Figure 2.2), with ϕ9 also covering gene 9 whose product is needed in large quantities. In addition, most transcripts originating from 13 (out of 15) T7 promoters will end immediately after gene 10, so the total level of the transcripts for gene 10 will be much higher than that for other genes. Following the same reasoning, we can explain the dependence of the growth rate on the positions of ϕ9 (Figure 4.4C) and Tϕ (Figure 4.4D). Switching the position of ϕ9 with gene 9 facilitates the production of gp10, thus resulting in an increase in T7 growth in the unlimited environment. This configuration causes a decrease in T7 growth in the limited environment for a similar reason as when ϕ10 activity is increased too much (Figure 4.2B). It directs too many ribosomes into producing gp10, resulting a bottleneck in

producing other proteins. Placing Tϕ upstream of gene 10 generally leads to significantly decreased T7 growth because of the reduced gp10 production rate. However, this decreased T7 growth can be compensated in the unlimited environment when Tϕ is placed between a promoter and gene 3.5 (Figure 4.4D). In this configuration, the negative feedback loop that keeps T7RNAP at low level is partially inhibited owing to reduced production rate of gp3.5, thus the overall transcription rate is increased. Again in the limited environment, because T7RNAP is more stringently regulated with an additional negative feedback loop and because increased levels of early to middle transcripts will divert ribosomes away from translating gene 10 transcripts, T7 growth rate will not be improved even when gp3.5 production is inhibited (Figure 4.4D).

According to the Highly Optimized Tolerance (HOT) theory (Calson & Doyle 2000), engineered or natural complex systems should be robust with respect to anticipated, designed-for perturbations but fragile to unanticipated perturbations. In this context, perturbations in single or multiple parameters, and in the position of a single element, would probably fall in to the category of "anticipated" perturbations because they cause, in most cases, quantitative rather than qualitative changes in the virus. By contrast, random perturbations of the T7 genome would qualify as unanticipated perturbations. The fragility of T7 with respect to such perturbations (Figure 4.5) makes intuitive sense. Since T7 genome entry is mediated by EcRNAP after the first 850bp enters the host cell (Garcia & Molineux 1995b; Garcia & Molineux 1996), at least one EcRNAP promoter needs to be present within 850bp towards the entering end of the

genome. Otherwise T7 will fail to grow. With only six EcRNAP promoters out of 72 internal genetic elements, it is very probable to place all these promoters after 850bp. The probability for such fatal configurations is approximately $[(39937 - 850)/39937]^6 \approx 88\%$, which agrees well with our simulation results (~82%). Therefore, failure to enter the cell accounts for the majority of dead T7 mutants. In addition, even if T7 enters the cell successfully, it may fail to grow if at least one of its 23 essential genes is upstream all promoters.

4.4.2 Optimality

Our simulation results suggest that the wild-type T7 is almost optimally designed for a limited-resource environment. Phage T7 responds more negatively to perturbations in the limited environment. Compared with mutants, the wild-type T7 performs overall much better in a limited environment than in the unlimited environment. This point is clear from the distribution of growth rates of T7 mutants with random parameters (Figure 4.3) or permuted genomes (Figure 4.5). Most mutants that grow faster than the wild type in the unlimited environment grow much slower than the wild type in the limited environment. This observation highlights the potential importance of the growth environment in shaping the design of an organism. In the context of evolution, this point seems to be reasonable: phage T7 relies on the host cell for its growth, and each realistic host cell will have only limited resources; through evolution, phage T7 should have been adapted for these environments, resulting in the apparent optimality as revealed by our simulation.

The optimality of the wild-type T7 is particularly distinct in terms of its genomic structure. In the unlimited environment, more than 20% of mutants with random parameters grow faster than the wild type, with the maximum growth rate about 40-fold greater than the wild-type value. In the same environment, only 4.9% of mutants with random genomes grow faster than the wild type, with a maximum growth rate of seven times the wild-type value. Further supporting this notion, the corresponding percentages are 5.3% and <0.001% respectively in the limited environment. This observation seems to highlight an important design principle in engineered or natural systems: the structure of a system, be it the structure of a metabolic network, that of an airplane, or the process for producing a chemical, is of fundamental importance in ensuring the proper functioning of the system. In T7, the proper arrangement of the genetic elements seems to lead to proper sequence and timing of various reactions, which in turn result in nearly balanced production of viral components for the limited environment, and thus nearly optimal growth. In general, we may argue that desirable key properties of any given system rely on a proper underlying structure. For example, negative feedback or redundancy, or both are usually required for achieving robustness (see above, and references (Alon et al 1999; Barkai & Leibler 1997; Yi et al 2000)), and positive feedback is often necessary for generating oscillations or for spontaneous pattern formation in spatial domain if coupled with diffusion (Becskei et al 2001).

Although it grows much faster than most mutants, clearly the wild type is still not always *the* best even in the limited environment (Figure 4.2). Faster growing mutants do appear. Why? There are several potential explanations. First, the host composition we

adopt in this work represents a host cell growing under standard lab conditions (You et al 2002). Such a host cell could be very different from one growing in nature, and phage T7 could have been more optimized for a naturally existing host cell. Second, it is also possible that phage T7 is actually not designed for optimal performance in a single environment, but rather for near-optimal performance under different environments, each of which has limited resources. To test this notion, further computational work can be conducted to evaluate the relative performance of wild-type phage T7 under a wide spectrum of realistic growth environments, for example host cells with different growth rates. Third, the data and mechanisms incorporated in the model might be inaccurate. In constructing the computer model, we have been faithful to literature data and have not attempted to adjust any parameters in order to fit existing experimental data (Endy et al 2000). However, potential uncertainties in the parameters may lead to biased view of the "wild type". In other words, the wild-type phage T7 as defined by the model parameters may be a mutant that is close to the wild type that the model is supposed to represent. This gap between model and reality, however, is difficult to completely fill because of the experimental difficulty in obtaining accurate parameter values. Finally, as argued in the "quasi-species" theory, the variant that is selected by evolution need not be the fittest; it can be sub-optimal and still win the race if it is "supported" by its mutational neighbors with which it gives highest average fitness (Eigen 1971a; Eigen 1971b; Wilke et al 2001). Our simulation results seem to agree with this notion: most subtle mutations with respect to parameters or genomic structure lead to mutants with growth rates comparable to the wild-type value.

C h a p t e r 5

QUANTIFYING GENETIC INTERACTIONS[13]

> *"Macroscopic properties often do not result from static structures, but from dynamic interactions playing both within the system and between the system and its environment"*
>
> Eirch Jantsch

Abstract

Understanding how interactions among deleterious mutations affect fitness may shed light on a variety of fundamental biological phenomena, including the evolution of sex, the buffering of genetic variations, and the topography of fitness landscapes. It remains an open question under what conditions and to what extent such interactions may be synergistic or antagonistic. To address this question, we employed a computer model for the intracellular growth of bacteriophage T7. We created *in silico* 90,000 mutants of phage T7, each carrying from one to 30 mutations, and evaluated the fitness of each by simulating its growth cycle. The simulations sought to account for the severity of single deleterious mutations on T7 growth, as well as the effect of the resource environment on our fitness measures. We found that mildly deleterious mutations interacted synergistically in poor-resource environments but antagonistically in rich-resource environments. However, severely deleterious mutations always interacted antagonistically, irrespective of environment. These results suggest that

[13] The content of this chapter has been published in You & Yin (2002), *Genetics*. **160**: 1273-1281

synergistic epistasis may be difficult to experimentally distinguish from non-epistasis because its effects appear to be most pronounced when the effects of mutations on fitness are most challenging to measure. Our approach demonstrates how computer simulations of developmental processes can be used to quantitatively study genetic interactions at the population level.

5.1 Introduction

The interaction among mutations in their effects on fitness, known as epistasis, plays a major role in evolutionary processes (Wolf et al 2000). It affects the mutation load of a population (Crow 1970; Kimura & Maruyama 1966), the drift and fixation of deleterious mutations (Phillips et al 2000), and the topography of fitness landscapes (Phillips et al 2000; Whitlock et al 1995). Epistasis is also an essential component of models on the evolution of sex (Kondrashov 1993; Peters & Lively 2000). In small populations Muller's ratchet (Muller 1964) offers sexual organisms an advantage over their asexual counterparts even in the absence of interactions among deleterious mutations (Haigh 1978). However, the mutational deterministic hypothesis, which addresses the advantages to sex for any population size, requires synergistic epistasis (Kondrashov 1993); in this case the collective effect of two or more deleterious mutations is more severe than their combined effect, had they acted independently. Following Lenski et al. (Lenski et al 1999), one may distinguish among different forms of epistasis using a power model:

$$\log(w) = -\alpha n^{\beta} \qquad 5.1$$

where w is the fitness of the organism relative to wild type, n is the number of deleterious mutations the organism carries, and α and β are parameters. For deleterious mutations, $\alpha > 0$ and α is larger for more severe mutations. The type and strength of the epistasis among the mutations are determined by β, where $\beta = 1$ for independent interactions (non-epistasis), $\beta > 1$ for synergistic epistasis, and $0 < \beta < 1$ for antagonistic epistasis (Figure 5.1). The degree of epistasis increases as $|\log(\beta)|$ increases. The power model has an advantage over the more commonly used log-quadratic model (Charlesworth 1990), $\log(w) = -(\alpha' n + \beta' n^2)$, which can erroneously predict an increase in fitness for large n in the case of antagonistic epistasis ($\beta' < 0$).

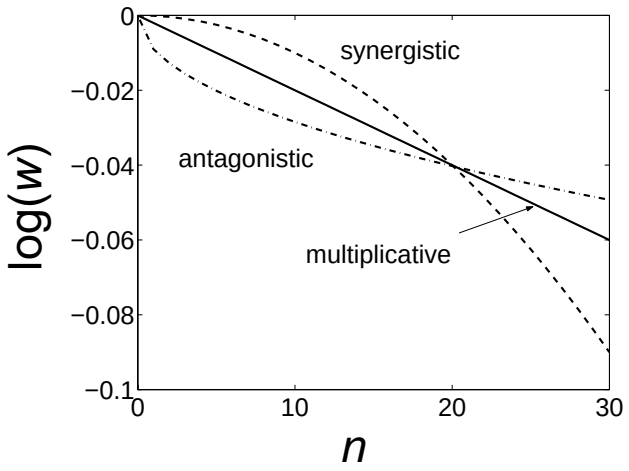

Figure 5.1 Three forms of epistasis as described by the power model (Eq 5.1) Multiplicative: $\alpha = 0.002$, $\beta = 1$; synergistic: $\alpha = 0.0001$, $\beta=2$; antagonistic: $\alpha = 0.009$, $\beta = 0.5$.

Numerous experimental studies have been conducted to determine the dependence of fitness on the number of deleterious mutations, in particular, whether synergistic epistasis is ubiquitous in nature. Several of them have examined directly the

variation in epistasis, and found that both synergistic and antagonistic interactions are prevalent among individual sets of mutations (de Visser et al 1997b; Elena & Lenski 1997; Whitlock & Bourguet 2000). However, the nature of epistasis across populations is all but clear. Some studies find synergistic epistasis (Mukai 1969; Spassky et al 1965), while others find multiplicative interactions (de Visser et al 1997b; Elena & Lenski 1997). Further complications follow as different forms of epistasis may occur for different fitness components, such as the longevity or the productivity of an organism (de Visser et al 1997a; de Visser & Hoekstra 1998; Peters & Keightley 2000; Whitlock & Bourguet 2000).

The confusion results from several factors. First, it is difficult to accurately estimate the fitness because of the complex life cycles of the model organisms (de Visser & Hoekstra 1998; de Visser et al 1997b; Peters & Keightley 2000; Whitlock & Bourguet 2000). Second, although the Darwinian fitness is the appropriate measure for w in Eq 5.1(West et al 1998), the Malthusian parameter, which is the natural logarithm of the Darwinian fitness, has been used instead in some studies (de Visser et al 1997a). This is problematic because synergism in the Malthusian scale can correspond to multiplicative or antagonistic interactions in the Darwinian scale. Third, theories on the advantages of sex have suggested that the most relevant interactions are those among mildly deleterious mutations, because deleterious mutations with large effects can easily be eliminated, even in small populations (Keightley et al 1998). The effect of deleterious mutations, however, has yet to be adequately quantified or controlled. Fourth, mutants with many deleterious mutations are difficult to construct systematically in experiments

even for a relatively simple organism like bacterium *E. coli*; the resulting data are thus insufficient to distinguish at the population level among different types of epistasis (Elena & Lenski 1997). Efforts to accumulate many mutations, for instance, by applying mutagens (de Visser et al 1996; Peters & Keightley 2000) or by growing the organism under minimum pressure of natural selection (Mukai 1969; Wloch et al 2001a), replace this difficulty with that of accurately estimating the number of deleterious mutations. Finally, two related methods that detect epistasis by comparing the mean log-fitness of the parents with that of the offspring after a cross (de Visser et al 1996) or by testing the skewness of the log-fitness distribution of these offspring (de Visser et al 1997a; Wloch et al 2001a) suffer several methodological limitations that may obscure their conclusions (West et al 1998).

Given current limitations in generating, characterizing, and quantifying the effects of mutations on the fitness of organisms in the laboratory or in the field, we chose to study how simulated mutations affect the development of bacteriophage T7 in a computer model of its life cycle. The construction of this model is detailed elsewhere (Endy et al 1997; Endy et al 2000; You et al 2002; You & Yin 2000). Briefly, it incorporates biochemical, genetic, and mechanistic data accumulated on phage T7 over the last four decades, and it uses ordinary differential equations and algebraic equations to describe the major molecular processes of T7 development: entry of T7 DNA into the host, transcription and translation of T7 genes, protein-protein interactions that regulate transcription, procapsid assembly, DNA replication and progeny formation. It predicts, as a function of time post-infection, the intracellular levels of mRNAs,

proteins, DNA and, eventually, phage progeny. In essence, the model bridges the genotype of a phage with its growth phenotype, a feature we exploit here to predict how simulated phage mutations affect fitness.

Our use of the phage T7 model here to study epistasis shares advantages with approaches based on artificial-life programs (Lenski et al 1999; Wilke & Adami 2001). Both approaches allow for the efficient creation and appraisal of thousands of *in silico* mutants. However, the T7 model is based on established biochemical mechanisms while the artificial-life programs are not. Simulated T7 mutations correspond to changes in molecular functions, such as binding constants between interacting proteins, promoter strengths, or terminator efficiencies, just as mutations alter molecular functions in laboratory mutants of T7. Mutations in any function can quantitatively span a broad parameter range and produce a correspondingly broad range of effects on phage growth, a unique feature of the T7 model that enables us to probe, at the level of individuals or populations, the severity of mutation effects on growth.

By incorporating the results of extensive experimental studies on T7, we have sought with our simulation to create a faithful quantitative representation of its intracellular infection dynamics. Nevertheless, gaps in our knowledge remain. Functions and mechanisms for many T7 genes are lacking (Endy et al 2000; Molineux 2001; You et al 2002) and only sparse data exists for the effects of host-cell physiology on phage growth (You et al 2002). Our imperfect knowledge was most evident when we employed the simulation to predict the behavior of T7 mutants carrying permutated

genomes (Endy et al 2000); while the simulation captured qualitatively the detrimental effects on growth of repositioning an important early gene to locations downstream of its wild-type position, it was unable to account quantitatively for the resulting highly perturbed protein expression profiles. In contrast, under conditions that are closer to wild type, the simulation has performed well. Specifically, it has enabled us to better understand how a mutant carrying a single-gene deletion grows faster than the wild type (Endy et al 1997), and more recently, the simulation has, together with laboratory experiments, provided a means for us to identify likely host-resource limitations to T7 growth (You et al 2002). Our current work here employs no permutated genomes or other large excursions from wild type. Instead, we study how slightly deleterious mutations interact against a backdrop of essentially wild-type T7 behavior. As a result, we expect that deficiencies or imperfections in our detailed knowledge and simulated implementation of T7 biology will have little if any effect on the outcomes of this study.

5.2 Materials and methods

5.2.1 Definition of fitness

The infection of an *E. coli* cell by phage T7 is often characterized by a one-step growth curve, from which we can extract parameters to define its fitness (Figure 5.2). Fitness is essentially determined by the interplay between the genotype of an organism and the environment in which it grows; so one genotype can exhibit different fitnesses in different environments. Here we consider two extreme scenarios. First, if a phage grows in an environment that permits only one cycle of infection, the fecundity of the phage, characterized by its burst size (*Y*), or the number of progeny produced per

infected host bacterium, will be the most crucial parameter in determining its fitness. Phage that maximize their burst size in such poor-resource environments maximize their chances of survival. Therefore, an appropriate fitness measure (W_{poor}) in this poor-resource environment is the burst size, or $W_{poor} = Y = \max\{N(t)\}$, where t is the time after infection initiation and $N(t)$ is the number of phage particles at t (Figure 5.2a). In the second scenario, if the phage grows in a rich environment that allows an infinite number of infection cycles, then both the burst size (Y) and the burst time (τ), the time when the burst occurs, will contribute to its fitness. For example, let us start with a single phage at time zero, and ask how many phage there will be at time t_m, assuming infinite host resources. For large t_m, the number of phage will be approximately $Y^{t_m/\tau} = [Y^{1/\tau}]^{t_m}$, where t_m/τ corresponds approximately to the number of phage generations elapsed at time t_m. This expression suggests a different measure of fitness: $Y^{1/\tau}$. Since the phage with the highest fitness will burst at a time that maximizes $Y^{1/\tau}$, we define $W_{rich} = \max\{[N(t)]^{1/t}\}$ (Figure 5.2b).

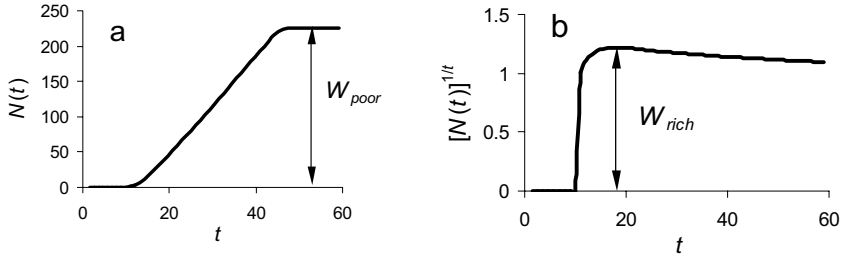

Figure 5.2 Definition of fitness. (a) The fitness measure for a poor-resource environment (W_{poor}) is defined from the simulated one-step growth curve as the maximum value of $N(t)$, where t is the time (minutes) after infection initiation, and $N(t)$ is the number of phage progeny at t. (b) The fitness measure for a rich-resource environment (W_{rich}) is defined as the maximum value of $N(t)^{1/t}$. In the mid-section of the growth curve, $N(t)$ is approximately linear in t, thus the function $N(t)^{1/t}$ will always have a maximum. The rationale for these definitions is provided in the text.

We further denote the relative fitness values of a mutant phage for the cases above with $w_{poor} = W_{poor}/W_{poor,wt}$, and $w_{rich} = W_{rich}/W_{rich,wt}$, respectively, where the subscript *wt* indicates the fitness values for the wild-type phage. Note that both w_{poor} and w_{rich} are fitness measures in the Darwinian scale.

5.2.2 Constructing T7 mutants in silico

Although we cannot yet predict how mutations at the DNA, RNA or protein level influence molecular function, we do know that such changes can quantitatively alter function. By altering function they change molecular properties that are typically described by parameters such as enzymatic rates, equilibrium binding constants that characterize interactions between molecular components, promoter strengths, or extents of regulatory inhibition or enhancement of other molecular functions. Taking this perspective, we simulated the effects of mutations on specific T7 functions by altering T7 parameters from their wild-type default values. We then used one or more altered parameters in our simulation to calculate how one or more mutations could affect the intracellular development of the phage. Further, we defined a *deleterious mutation* as a single-parameter change that would reduce the T7 relative fitness, w_{rich}, to a value below one. Different fitness metrics can be used to characterize phage growth, so the same mutation may be deleterious for one metric but not for another. However, for all cases examined, single mutations that were deleterious in w_{rich} were also deleterious or at least neutral in w_{poor}

Table 5.1 The list of parameters investigated in this study.

Parameter	Relative value range *	
	minimum	Maximum
T7 RNAP elongation rate	0	1
T7 DNAP elongation rate	0	1
gp0.7-EcRNAP association constant	0	1
gp2-EcRNAP association constant	0	1
T7RNAP-gp3.5 association constant	0	0.4
Procapsid assembly rate constant	0	1
DNA packaging rate constant	0	1
EcRNAP terminator efficiency	0	1
T7 RNAP terminator efficiency	0	1
A1	1	1.58
A2	0.3	1
A3	0.1	1
B	1	10
C	0.1	1
E	0.1	1
ϕ1.1A	1	15.4
ϕ1.1B	1	5.5
ϕ1.3	1	97.3

	Relative value range *	
ϕ1.5	1	15.4
ϕ1.6	1	15.4
ϕ2.5	1	15.4
ϕ3.8	1	27.2
ϕ4C	1	15.4
ϕ4.3	1	15.4
ϕ4.7	1	15.4
ϕ6.5	1	1.5
ϕ9	0	1
ϕ10	0	1
ϕ13**	0	0.174
ϕ13**	1	1.2
ϕ17	1	1.2

* In the given range of values, each single parameter change alone is deleterious for T7 growth.

** When the relative activity of promoter 13 is within the range of (0.3, 1), T7 growth is enhanced.

Thirty T7 parameters were specified as potential targets for mutations; for each parameter, a range was specified from which random selected values corresponded to random deleterious mutations. The parameters and their ranges normalized to the

default values are listed in Table 5.1**.** The default values of these parameters are listed in Chapter 2. These parameters were identified and their ranges were determined based on a single-parameter sensitivity analysis on all T7 parameters. In the sensitivity analysis, each parameter was varied from zero to 100-fold its default value and the resulting change in w_{rich} was examined for deleterious effect. For instance, if we reduce the T7 RNAP elongation rate from its default value (200 bp/s) to 0, the simulated w_{rich} will decrease from 1 to 0; thus the deleterious range for this parameter is (0, 1). To control the magnitude of deleterious mutations, we further partitioned each parameter range into five equal-width sub-ranges. Each sub-range was labeled by an index from 1 to 5, based on its deviation from the default parameter value, 1 for the least deviant and 5 for the most deviant. Again consider the elongation rate of the T7 RNA polymerase as an example. Its complete range (0, 1) was partitioned into the five sub-ranges: sub-range 1, (0.8, 1.0); sub-range 2, (0.6, 0.8); sub-range 3, (0.4, 0.6); sub-range 4, (0.2, 0.4); and sub-range 5, (0, 0.2). We call a mutation in sub-range k as a class k mutation. For a given parameter, the deleterious effect of a mutation increases with its class index. A special sub-range – (0.9, 1.0) – was created to represent mutations with very mild effects, which we called class 0.5 mutations.

A T7 mutant with n random class k deleterious mutations was constructed by randomly selecting n parameters and then setting each selected parameter to a value randomly sampled from the class k sub-range following a uniform distribution.

5.2.3 Simulation and statistical analysis

For each n, where $1 \leq n \leq 30$, 500 T7 mutants carrying the same class of random deleterious mutations were constructed, and a simulation was performed for each mutant to compute its fitness using two measures, w_{poor} and w_{rich}. The means and the standard deviations of $\log(w_{poor})$ and $\log(w_{rich})$ were calculated for each n; the means were then fitted by least squares against n using Eq 5.1 to obtain α and β values. For each fitted curve, R^2 was calculated as the ratio of the difference between the corrected total sum of squares and the residual sum of squares to the corrected total sum of squares. The magnitudes of β values obtained for w_{poor} and w_{rich} may be compared; however, because the α values from which curves depend arbitrarily on the dimensions of time t, comparisons between α values obtained for w_{poor} and w_{rich} will not be meaningful. A sample size of 500 appeared to be sufficient; sampling 1000 did not yield significantly different results. Further, all the simulations were conducted assuming a host growth rate of 1.5 doublings/hr. The same conclusion as presented here was reached when assuming other host growth rate values (LY, unpublished). The statistical analysis was conducted using Matlab and Mathematica.

5.3 Results and discussion

5.3.1 Effects of environment and mutation severity on epistasis

The power model matched well the fitness loss for phage strains carrying up to 30 class 1 mutations in poor- and rich-resource environments (Figure 5.3). Moreover, these strains exhibited either synergistic or antagonistic epistasis, when tested in a poor-

or rich-resource environment, respectively. When the analysis was extended to other mutation classes, the power model also served well to capture data trends (Figure 5.4). By extracting the α and β values from each curve fit and then plotting β versus $\log(\alpha)$ we were able to probe how the form of epistasis, measured by β, depended on the mutation severity, measured by α, in different environments (Figure 5.5). In a poor environment β was greater than one for mildly deleterious mutations (classes 0.5, 1 and 2), indicating synergism, but it rapidly decreased with increasing α; it was nearly one for class 3 mutations, and then became less than one for the most severe mutations (classes 4 and 5), reflecting antagonism (Figure 5.5a). These results show that the form of epistasis can depend on the severity of the mutations. By contrast, in a rich environment the epistasis was always antagonistic, but the antagonism decreased with decreasing severity of mutations (Figure 5.5b). Note the almost linear dependence of β on $\log(\alpha)$ in either case. This relationship was stronger in the poor environment, where a least-square linear fit between β and $\log(\alpha)$ yielded an R^2 of 0.9971 compared with an R^2 of 0.9655 in the rich environment. Further, the rich environment deviation of the data from linearity suggests that β may asymptotically approach unity instead of crossing it (Figure 5.5b), a trend that is further discussed below.

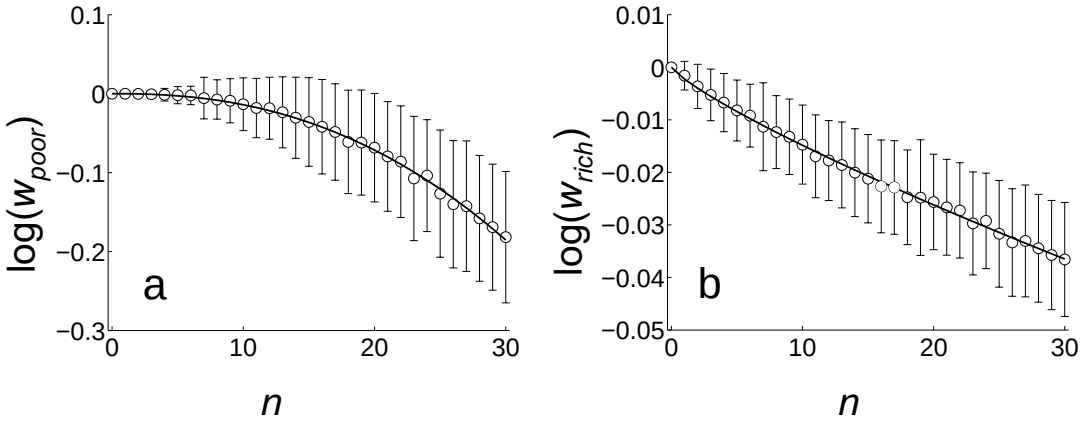

Figure 5.3 The dependence of the forms of epistasis on the fitness metrics. For mildly deleterious mutations (class 1), epistasis is (a) synergistic for w_{poor}: $\alpha = 6.25 \times 10^{-5}$ ($p = 4.25 \times 10^{-8}$), $\beta = 2.35$ ($p < 1 \times 10^{-10}$ for β - 1), $R^2 = 0.9970$, and (b) antagonistic for w_{rich}: $\alpha = 2.25 \times 10^{-3}$ ($p < 1 \times 10^{-10}$), $\beta = 0.818$ ($p < 1 \times 10^{-10}$ for β - 1), $R^2 = 0.9972$. Each circle represents the mean of log(fitness) values of 500 mutants; each vertical line represents the corresponding standard deviation.

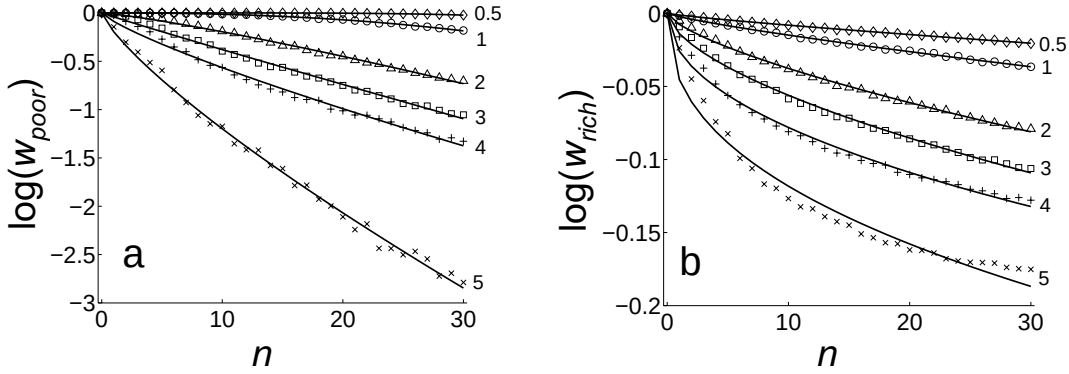

Figure 5.4 The dependence of the two different fitness measures on the number of deleterious mutations with varying severity. Each symbol represents the mean of log(fitness) values of 500 mutants. For clarity, standard deviations of the simulated fitness values are not shown. Good fit between the power model and simulated data is found in all cases ($R^2 > 0.99$).

Table 5.2 The dependence of epistasis on growth environment and the magnitude of deleterious mutations.

		ENVIRONMENT	
		poor	rich
MUTATIONS	mild	**synergism**	weak antagonism
	severe	no epistasis or weak antagonism	**antagonism**

From these results we draw two main conclusions, summarized in Table 5.2. First, mildly deleterious mutations tend to interact synergistically in a poor-resource environment, where fecundity is the primary determinant of fitness. This result is consistent with the notion that synergistic epistasis can emerge from competition for food or limited resources (Peck & Waxman 2000). Second, deleterious mutations tend to interact antagonistically in a rich-resource environment, and the degree of antagonism increases as the mutation severity increases. This conclusion may provide an intuitive explanation to the controversy over the nature of epistasis: perhaps previous studies reached different conclusions because they focused on different environments or were based on mutations of differing severity.

5.3.2 Correlation between epistasis and mutation severity

Inverse correlations between β and α (Figure 5.5) have also recently been observed by Wilke and Adami in self-replicating computer programs and models for RNA folding (Wilke & Adami 2001). To probe the relationship between β and α they considered sequences composed of a finite fixed number of monomers, where the fitness of mutants was defined as either neutral or lethal. Assuming that the fitness always decreased with the number of mutations, following a power model (Eq 5.1) they suggested that β must be inversely correlated with α due to a conservation law that the total number of neutral mutants in the genetic space is constant (Wilke & Adami 2001).

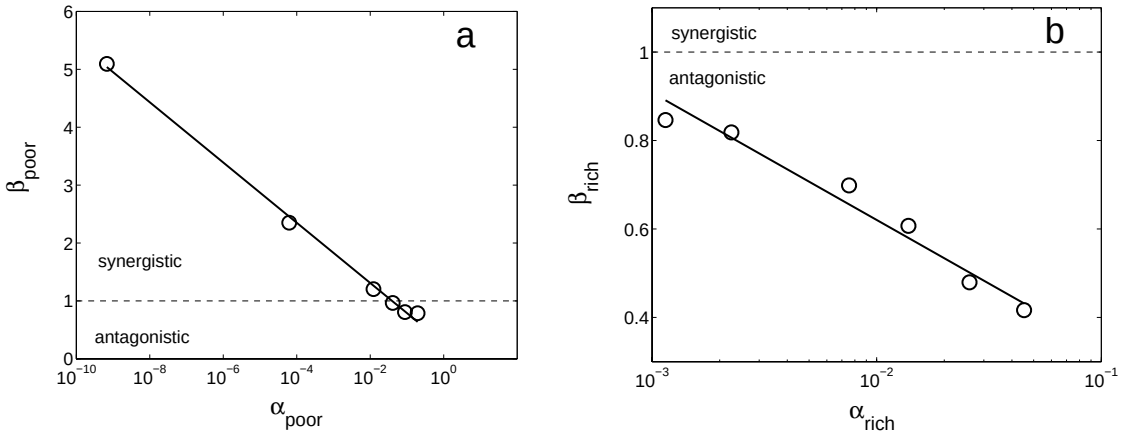

Figure 5.5 The dependence of β on α for (a) the poor environment and (b) the rich environment. Dashed lines indicate multiplicative interactions among mutations (β = 1, non-epistasis). A linear equation fits well the dependence of β on log(α) in each case, with R^2 = 0.9971 for the poor environment, and R^2 = 0.9655 for the rich environment. All β values are significantly different from 1.0, with $p < 0.01$ for β-1, except the β value for w_{poor} and class 3 mutations, where $p = 0.087$ for β-1.

However, this argument cannot be directly mapped to our study because of the infinite diversity of mutational effects on fitness in our system. Furthermore, the apparent linear dependence of β on log(α) in Figure 5.5 suggests an alternative mechanism. If β and log(α) followed an exactly linear relation, then the different log(*w*)-

vs-n curves, each specified by a different pair of α and β values, would share two points of intersection, one at $n = 0$ (wild type) and the other at some large number of mutations ($n >> 30$, for our examples). The reasoning is as follows. At the large n intersection point, we have $\log(w_N) = -\alpha N^{\beta}$, where N is the number of mutations and w_N is its corresponding fitness. This equation defines a linear relation between β and $\log(\alpha)$, since $\log(-\log(w_N)) = \log(\alpha) + \beta \log N$. We assume that the power model is valid if and only if $n \leq N$; otherwise, for $n > N$ it would predict that mutants carrying severe mutations would have a higher average fitness than mutants carrying the same number of mild mutations. Within this framework mutations that yield a fitness w_N are effectively lethal since a fitness equal to zero is not defined. Further, the dependence of fitness on the accumulation of mutations will, for differing degrees of mutation severity, all originate from wild type but follow different paths that ultimately converge to the same fitness w_N of the effectively lethal mutants. Hence, to reach this fitness with N mutations, mild mutations will tend to reinforce each other, leading to synergistic epistasis, whereas severe mutations will tend to buffer each other, resulting in antagonistic epistasis.

5.3.3 Antagonism in rich environments

In light of the above argument for a linear correlation between β and $\log(\alpha)$, it appears contradictory that mildly deleterious mutations should still exhibit antagonistic epistasis in rich environments. We found that the relative fitness in a rich environment, $w_{rich,}$ is generally more sensitive to mutations than the measure in a poor environment,

w_{poor}; mutations that changed the latter also changed the former, but the inverse was not true. It would be thus conceivable that the same mutations could have a greater deleterious effect on w_{rich} than on w_{poor}. (Note that the higher sensitivity of w_{rich} to mutations is not reflected by the α values, because the absolute value of α_{rich} depends on the time units we use in the definition of W_{rich}). Perhaps if mutations were sufficiently mild, their interactions in a rich environment would become synergistic, following the pattern of epistasis in a poor environment. To test this possibility we examined interactions among mutations that were more than 100-fold milder than our 0.5 class mutations. These very mild mutations exhibited very slight antagonism in the rich environment (not shown) and behavior that was indistinguishable from wild-type in the poor environment. This result further confirms the overall trend of changes in β in a rich environment (Figure 5.5b): as severity of mutations decreases, the interaction among these mutations asymptotically approaches multiplicativity (β=1). Therefore, the difference in the forms of epistasis between our metrics do not merely reflect differences of degree to which these metrics are affected by mutations, but rather intrinsic differences in the nature of their responses to deleterious mutations.

5.3.4 Limits to observability

Besides its implication that epistasis and mutational effect could only evolve in a coordinated fashion (Wilke & Adami 2001), the correlation between β and $\log(\alpha)$ further leads to a dilemma for any attempts to distinguish synergistic epistasis ($\beta > 1$) from non-epistasis ($\beta = 1$): if synergistic epistasis is present (for example, in a poor

environment), it will be highest under conditions where the effects of mutations on fitness are minimal and most challenging to accurately measure (mutation severity class 0.5). This challenge may be better understood by considering a quantitative example. Weak synergistic epistasis is apparent for class 2 mutations, where 30 mutations decrease the fitness by about one half (Figure 5.4a). The average deleterious effect of each mutation should thus be no greater than about $1 - (0.5)^{1/30} = 0.023$. That is, each mutation on average should decrease the fitness by less than three percent.

Although fitness effects of around two percent might be experimentally established in competition experiments using microbes, the resulting synergistic epistasis would be mild (β is only slightly smaller than 1.0). In order to measure such synergistic epistasis, one would need to quantify the fitness of a large number of mutants, ranging from those with single mutations to those with a large number of mutations. This would be a daunting experimental task. If the mutations are overall mild enough to demonstrate high degree of synergistic epistasis, the effects of individual mutations may be too small and fall within the experimental variability of most fitness measures. For example, most experimental studies to date have measured only mutations with average selection coefficients greater than 0.01 (Elena & Lenski 1997; Fry et al 1999; Mukai et al 1972; Spassky et al 1965; Wloch et al 2001b), and it has been suggested that the majority of deleterious mutations have effects that are immeasurably small under laboratory conditions (Davies et al 1999).

5.3.5 Extensions of in silico mutagenesis

We have focused here on probing the effects of mutational severity and resource environment on the form and extent of epistasis in the simulated intracellular growth of a well-studied bacteriophage. This work may serve as a foundation to test the consequences of additional mechanisms or assumptions. For example, to study the effects of mutation severity on fitness we have assumed the mutations carried by each strain are uniformly distributed across each class of mutation severity. More natural distributions could be implemented by making mild mutations more frequent than severe mutations. From this perspective, our simulations of low severity (class 0.5 and class 1) mutations more likely reflect the effects of natural distributions than those involving severe mutations. Our current study has also neglected pleiotropic effects, where a mutation in one gene may affect more than one phenotypic trait. For example, a mutation that altered the processivity of the T7 RNA polymerase could at the same time influence the strength of its association with the T7 lysozyme, which down-regulates the polymerase activity. To account for such effects, one would need to obtain data that quantitatively described the nature of each pleiotropic effect or assume and implement a mathematical model for its form.

Acknowledgments

We thank J.F. Crow, S.F. Elena, R. Kishony, and R.E. Lenski for helpful comments and suggestions, and H. Wang for assistance with the statistical analysis. Support was provided by the National Science Foundation.

C h a p t e r 6

PATTERNS OF REGULATION FROM MRNA AND PROTEIN TIME-SERIES[14]

> *"Nature uses only the longest threads to weave her patterns, so each small piece of her fabric reveals the organization of the entire tapestry."*
>
> Richard Feynman

Abstract

The rapid advance of genome sequencing projects challenges biologists to assign physiological roles to thousands of unknown gene products. We suggest here that regulatory functions and protein-protein interactions involving specific products may be inferred from the trajectories over time of their mRNA and free protein levels within the cell. The level of a protein in the cytoplasm is governed not only by the level of its mRNA and the rate of translation, but also by the protein's folding efficiency, its biochemical modification, its complexation with other components, its degradation, and its transport from the cytoplasmic space. All these co- and post-translational events cause the concentration of the protein to deviate from the level that would result if we only accounted for translation of its mRNA. The dynamics of such deviations can create patterns that reflect regulatory functions. Moreover, correlations among deviations highlight protein pairs involved in potential protein-protein interactions. We explore and

[14] The content of this chapter has been published in You & Yin, 2000, *Metabolic Engineering.* **2**: 210-217.

illustrate these ideas here using a genetically-structured simulation for the intracellular growth of bacteriophage T7.

6.1 Introduction

The rapid growth of genomic databases and development of technologies for global monitoring of mRNA (Schena 1996; Velculescu et al 1995) and protein levels (Eckerskorn et al 1992; Henzel et al 1993; O'Farrell 1975) are creating a need for efficient data-mining methods. In the last few years, reverse engineering methods have been proposed to infer the structure of chemical reaction networks near a steady state (Arkin & Ross 1995; Arkin et al 1997) or, more abstractly, the architecture of Boolean networks defined by simple logic rules (Akutsu et al 1999; Liang et al 1998). However, little effort has been directed toward inferring relationships among elements of real genetic networks. To address this need, we suggest here a framework for organizing dynamic mRNA and protein data, with the aim of identifying characteristic patterns of function and potential protein-protein interactions. Given initial information about mRNA levels, and the processing rates and distribution of the ribosomes and activated tRNAs that constitute the translation resources, one can estimate how levels of corresponding proteins will change. Other factors, in addition to translation, will influence the actual protein levels observed: processes of protein modification or degradation, protein transport to or from the cytoplasm, regulatory processes, and the formation of protein-protein complexes. When coupled with translation, different modes of protein appearance or depletion will produce patterns of expression that may reflect aspects of the protein's function. To explore this notion, we employ the T7

model to generate dynamic mRNA and protein data. The simulation provides as a byproduct the concentration-versus-time trajectories of all T7 mRNAs and proteins (Figure 6.1), which we employ in this section as raw data to explore new modes of inference. In particular, we analyze these simulated data using a simple algorithm and attempt to identify patterns of regulation and infer protein-protein interactions. Finally, we evaluate the algorithm by comparing its results with the known mechanisms implemented in the model.

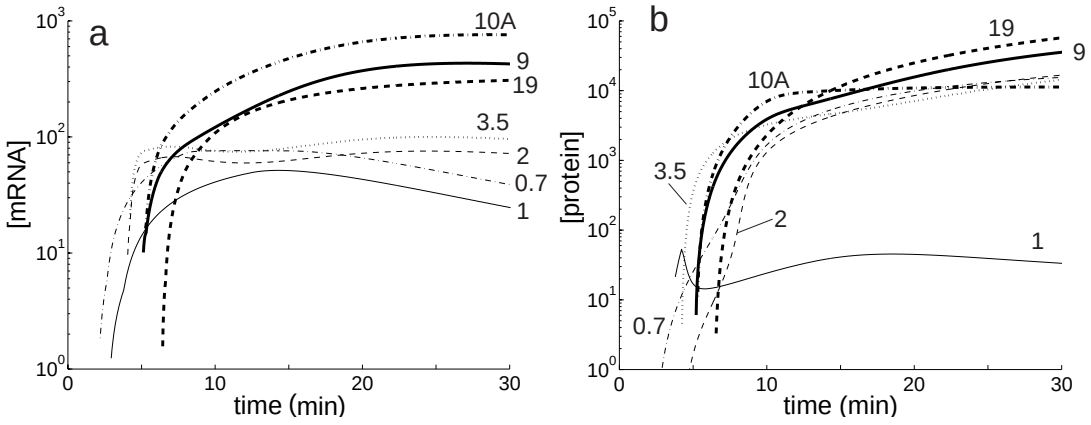

Figure 6.1 The time series of (a) mRNAs and (b) free proteins for seven selected T7 genes as labeled.

6.2 Methodology

6.2.1 Protein rate

We define the overall rate change in protein concentration, or the protein rate, as follows:

$$v_i(t) \equiv \frac{dP_i}{dt} = v_{Ti}(t) - v_{Di}(t) \qquad 6.1$$

where v_i is the protein rate for protein i, P_i is the concentration of protein i, v_{Ti} is the rate of change in protein i concentration due to translation, and v_{Di} is the protein i depletion rate. All these variables are the functions of time t. v_i is determined from levels of free protein taken over time.

We assume the protein translation rate, v_{Ti}, is proportional to the concentration of the mRNA:

$$v_{T\,i}(t) = k_{Ti} R_i \qquad 6.2$$

where R_i is the concentration of mRNA encoding protein i, and k_{Ti} is the translation rate coefficient for protein i. We assume k_{Ti} is a constant for all proteins that has been measured or can be estimated. For simplicity the index i is dropped in subsequent analyses.

In the absence of any depletion effects v_i plotted versus R_i yields a line through the origin with slope k_T. In the presence of depletion effects trajectories on this plot may deviate from the linear behavior in revealing ways. In this work the protein rate is approximated by taking the first-order finite difference of protein concentrations provided by the simulation.

6.2.2 Dynamic deviation factor (DDF)

In order to quantitatively investigate and compare the deviations of different proteins we introduce a dimensionless parameter, the dynamic deviation factor (DDF) for protein i:

$$D_i(t) \equiv \frac{v_i(t) - v_{Ti}(t)}{v_{Ti}(t)} = -\frac{v_{Di}(t)}{v_{Ti}(t)} \qquad 6.3$$

where D_i is the PDF as a function of time. Further, we define a time-averaged DDF as a measure of the overall deviation of the protein rate from the translation rate:

$$\overline{D_i} = \frac{1}{N}\sum_{k=1}^{N} D_{ik} \qquad 6.4$$

where D_{ik} is the DDF of protein i at the kth time point of a discretized time course that spans N time points.

6.2.3 Protein correlation coefficient (PCC)

If pairs of proteins associate, their DDFs should be highly correlated. We assess potential pairwise associations by defining a protein correlation coefficient:

$$C_{ij} \equiv \frac{\sum_{k=1}^{N} D_{ik} D_{jk}}{\sqrt{\sum_{k=1}^{N} D_{ik}^2 \sum_{k=1}^{N} D_{jk}}} \qquad -1 \le C_{ij} \le 1 \qquad 6.5$$

The larger C_{ij} is, the more likely two proteins are associating with each other. The extreme case C_{ij} means that the protein rates of the two proteins deviate proportionally from the translation rate in the same direction, i.e., $D_{ik}=\lambda D_{jk}$ for $k=1,...,N$, where λ is a positive constant. A protein correlation matrix (PCM) can be constructed by calculating the pairwise PCCs for all the proteins of interest.

6.3 Results

The time series of mRNA and free protein concentrations generated by the simulation for different T7 gene products shown in Figure 6.1 reflect little of the diverse enzymatic, structural, and regulatory roles the proteins play during T7 development. Only gp1 (T7RNAP) provides a hint of its regulated activity through its unusual trajectory relative to other proteins (Figure 6.1 (b)). Plotting protein rate versus mRNA concentration for the T7 gene products reveals a diversity of trajectories, as shown in Figure 6.2(a). For clarity, a representative subset of the total trajectories is shown. A time course is implicit for each trajectory, with an initial condition at the origin where neither mRNAs nor proteins have been synthesized. All trajectories coincide with or lie below the reference line (labeled "Linear"), corresponding to translation without any depletion.

The trajectories in Figure 6.2(a) may be coarsely classified as linear or non-linear. The simplest mechanistic rationale for the linear behavior of gp19, which coincides with the reference line, is that it is generated by translation and lacks any depletion effects. Other gene products gp0.3, gp2.5, gp4A, gp5, gp6, gp7, gp17.5, and gp18 exhibited similar behavior (not shown).

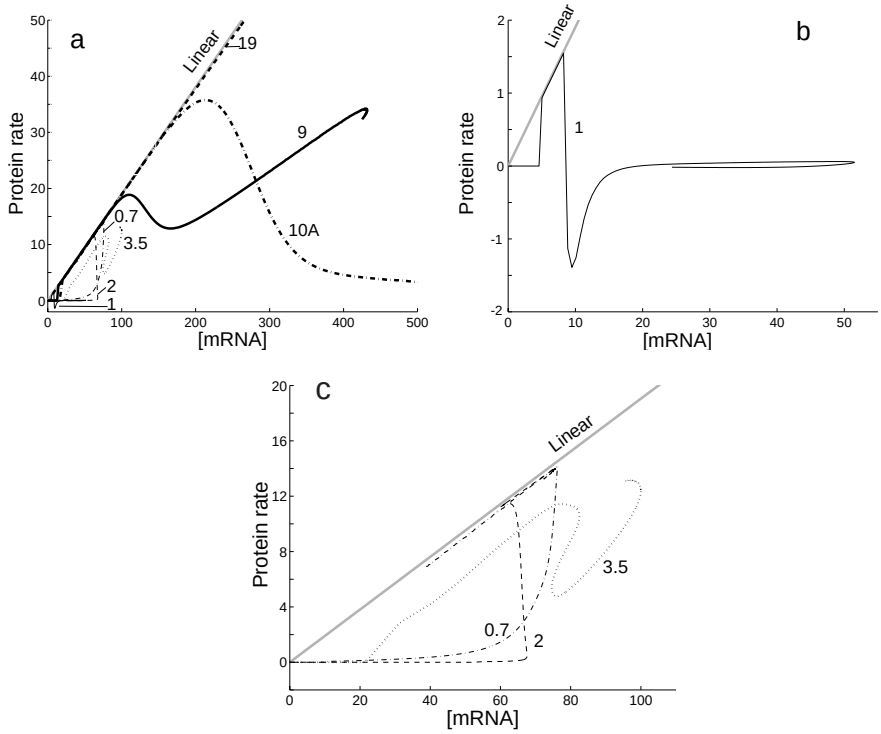

Figure 6.2 Protein rate as a function of the corresponding mRNA concentration for (a) seven representative gene products, (b) gp1, expanded axes, and (c) gp0.7, gp2, and gp3.5, expanded axes. The protein rate is calculated by taking the first order finite difference of the corresponding protein concentration, i.e., $v_i \approx {}^{\Delta P_i}/_{\Delta t}$. The straight line is a plot of v_{Ti} vs R_p with slope k_T determined directly from the parameters used in the simulation, viz, $k_T = k_E R_d$, where k_E is the ribosomal elongation rate, and R_d is the density of ribosomes along the mRNAs.

The non-linear trajectories may be subdivided into two groups based on their shapes. Trajectories for gp9 and gp10A follow the reference line early and deviate later (Figure 6.2 (a)). Other gene products, gp8, gp11, gp12, gp13, gp15, and gp16 followed similar trends (not shown). Gp10A is noteworthy because its maximum of 500 mRNA molecules per cell is the highest achieved by any message, and its protein rate drops to near zero as it approaches this concentration, indicating a strong depletion effect. The

other set of non-linear trajectories including gp0.7, gp1, gp2 and gp3.5 are more complex. The trajectory for gp1, which is the most complex, is shown on expanded axes in Figure 6.2(b). We know from the simulation that the sudden jump in protein rate from zero to the reference line at low mRNA levels observed for gp1 and other gene products is due to the finite time required by the ribosomes to complete synthesis of the first proteins. The segment of linear increase coinciding with the reference line indicates a growth in mRNA and proteins without any apparent depletion. Then the dramatic drop to negative protein rates, even as mRNA is increasing, suggests the appearance of some factor, which could be another component or process that significantly depletes gp1. After passing through a minimum, the protein rate gradually returns to nearly zero, where it remains, even as the mRNA level goes through a maximum. This pattern suggests that the translation of gp1 is nearly balanced by its depletion during the later stages of development. The trajectories for gp0.7, gp2, and gp3.5 are defined by their extended initial deviation from the reference line, shown in Figure 6.2(c). Protein rates are near zero well beyond the time required by the ribosomes to complete initial synthesis of each protein, indicating that these proteins are immediately depleted when they otherwise would have begun to appear. The eventual rise and fall of mRNA concentrations along with protein rates trace complex paths that loop back or move along the reference line as gp0.7 and gp2, respectively, or always deviate from the reference line as gp3.5. The different trajectory patterns for 21 essential T7 proteins are summarized in Figure 6.3 using time-averaged DDFs.

Using 21 essential T7 proteins, we calculated a protein association matrix for all pairwise interactions. As shown in Figure 6.4, high PCCs were found for gp0.7 and gp2, gp1 and gp3.5, and among gene products 8 through 16.

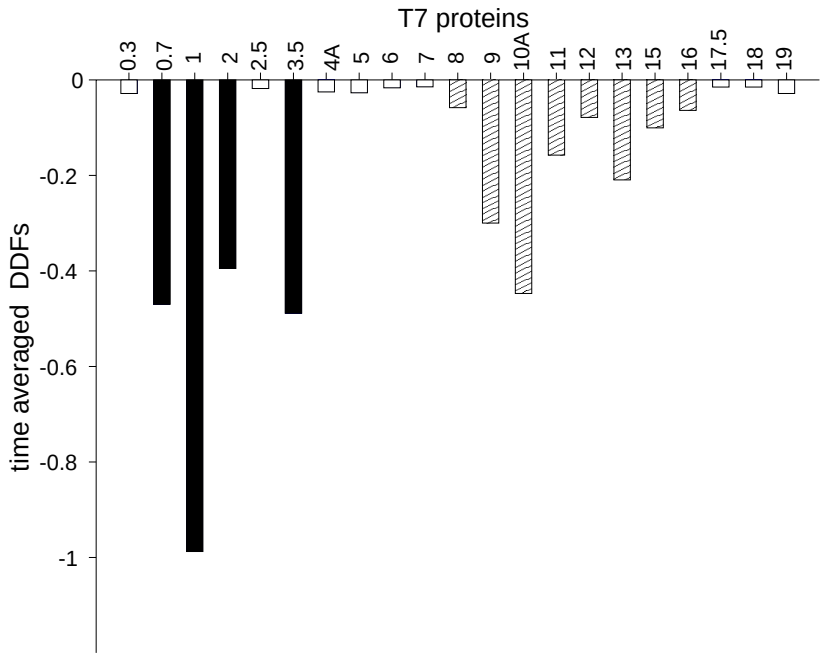

Figure 6.3 The time-averaged DDFs for 21 essential T7 proteins. Proteins that exhibit linear (non-shaded) and non-linear (hatched or black) protein-rate versus mRNA trajectories are shown. Gene products that started off linear, but later deviated (hatched) are distinguished from those that started off with large deviations (black)

6.4 Discussion

We have suggested a correlated deviation algorithm (CDA) for identifying potential patterns of protein function given time series data for mRNAs and proteins. Using data from a genetically-structured simulation for the growth of phage T7 we found a variety of expression trajectories, which we classified by their general features, from linear, to non-linear, to highly non-linear trajectories that looped back on themselves.

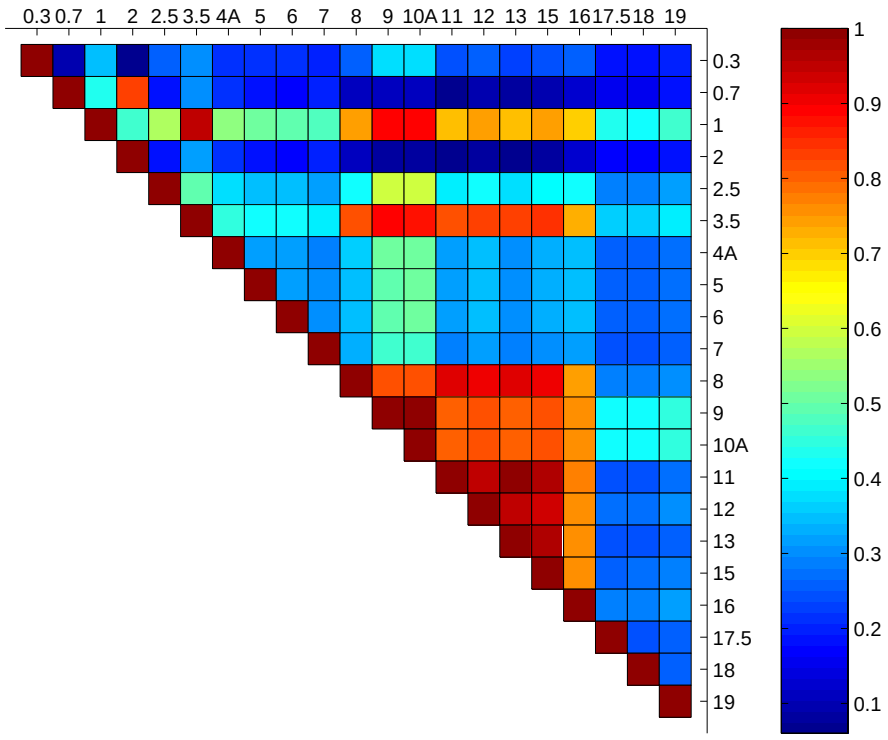

Figure 6.4 The T7 protein correlation matrix (PCM). Each off-diagonal matrix element represents the correlation coefficient between the DDFs of two proteins,ranging from high (red) to low (blue) correlation. Diagonal elements,which indicate the self correlation of the proteins, are by definition always equal to one.

The observed trajectory patterns reflect known functions of the T7 proteins. The complex trajectories of Figures 6.2(b) and (c) describe proteins that all indirectly control their own syntheses, as illustrated by the three negative feedback loops in Figure 6.5. Gp0.7 is a protein kinase that is transcribed by EcRNAP, but it also inhibits EcRNAP; increased levels of gp0.7 result in a down-regulation of its own transcription. Gp2 is an inactivator of EcRNAP, and EcRNAP is required for transcription of T7 RNAP. Gp2, however, is predominately transcribed by T7 RNAP, so gp2 ultimately down-regulates its own transcription. Gp3.5 is a lysozyme that associates with and inhibits T7RNAP

(gp1), but it is also transcribed by T7 RNAP, so it also down-regulates its own transcription. Hence, gp0.7, gp2, and gp3.5 share the feature that they all down-regulate their own transcription by inhibiting their own RNA polymerases. Gp1 is still more involved because it not only down-regulates its own transcription through the effect of gp2 on EcRNAP, but also transcribes its own inhibitor in gp3.5.

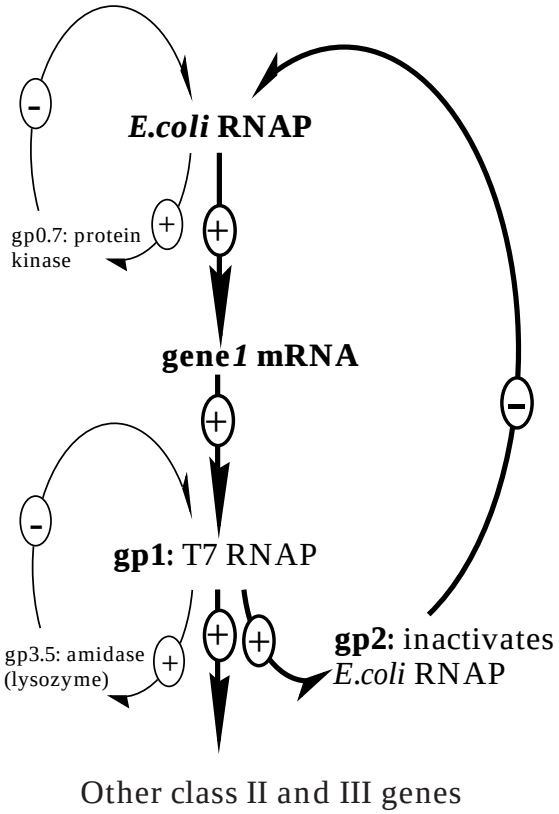

Figure 6.5 Negative feedback loops in the early stages of T7 infection. Gp0.7 inhibits the EcRNAP via an unknown mechanism, but it is assumed in the simulation that gp0.7 and EcRNAP form 1:1 complex. Gp2 inhibits the EcRNAP by forming a 1:1 complex with the polymerase that prevents transcription. Gp3.5 inhibits gp1 by binding the polymerase and increasing the rate of aborted transcript production.

The protein correlation matrix (PCM) (Figure 6.4) is overall consistent with the known functions. The deviations for gp1 and gp3.5 are correlated because they interact with each other, and those of gp0.7 and gp2 are correlated because both proteins are depleted through their interactions with EcRNAP. 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, as shown in Table 2.5. The gp9 and gp10A trajectories exhibited few if any early deviations (Figure

6.2(a)) because the particle formation process that depletes these proteins occurs late in phage development. With the exception of gp9, gp8 through gp16 are components of the final phage particle. Although gp9 if not present in the final particle, it is required for and consumed by the particle assembly process. Since the matrix provides correlations, we caution against attempting to infer mechanisms from it. Spurious associations for which no interactions were implemented in the simulation, such as those between gp1 or gp3.5 with particle proteins, are evident. Extension of the correlation analysis to allow for time-lags may improve discrimination.

The context of the CDA is summarized in Figure 6.6. The CDA assumes that the time series of mRNAs and proteins for a given multi-gene process have sufficient accuracy and resolution for analysis. The protein rate versus mRNA trajectories may reveal patterns that reflect the function of the proteins. The time-averaged DDFs serve as a qualitative measurement of the non-linear behavior of proteins. The PCM further provides a global picture of the potential associations among the proteins of interest. The CDA aims to facilitate the data-mining process of the large volume of output generated from emerging high-throughput experimental techniques, while focusing subsequent experiments on the proteins that demonstrate interesting behaviors.

Reverse engineering approaches have been developed to deduce the underlying wiring of Boolean logic networks, given sufficient state transition (input-output) pairs (Akutsu et al 1999; Liang et al 1998). Such approaches may eventually be useful for

identifying biological regulatory functions, but to date their application has been limited to the analysis of non-biological networks.

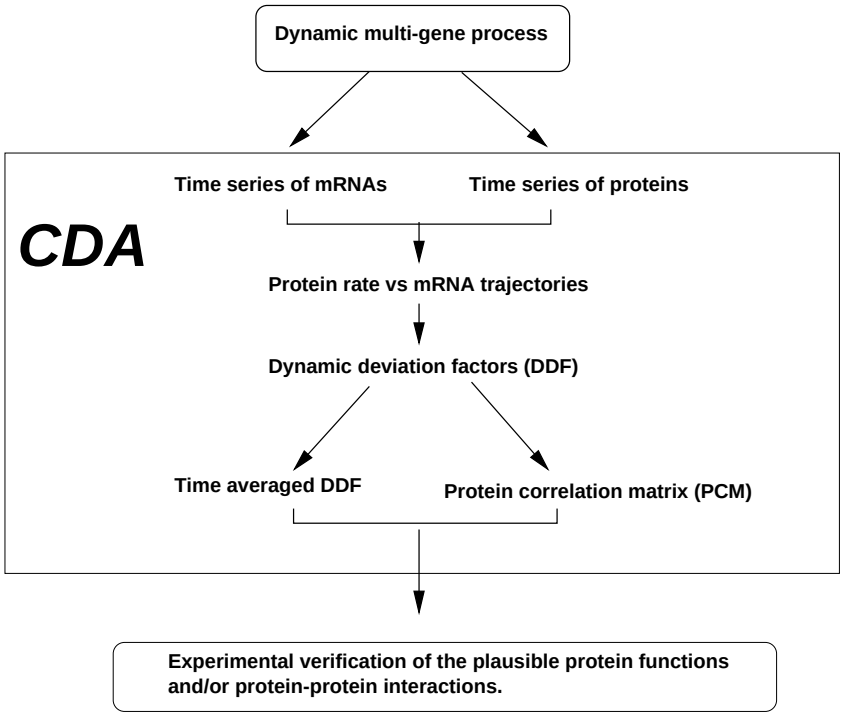

Figure 6.6 The context of the correlated deviation algorithm (CDA). By organizing and presenting the data in the forms of protein rate vs mRNA trajectory, DDF, and PCM, the CDA can highlight proteins that demonstrate interesting behavior. These analyses may facilitate the processing of the output from the emerging high-throughput experimental techniques.

An important feature of the CDA lies in its use of protein rate versus mRNA trajectories to infer mechanisms. It is an intriguing question to ask whether there exists a "mapping" between particular protein functions and some unique patterns evident in the protein rate versus mRNA trajectories. Our analysis with the simulated T7 system suggests this might be the case. We are currently exploring this question by testing the CDA with some simple artificial genetic networks.

The CDA resembles the correlation matrix construction approach proposed by (Arkin & Ross 1995) in its use of a correlation matrix, but differences are apparent. Although the PCC is analogous, both in its meaning and form, to the conventional correlation coefficient with zero time-lags, there is significant difference between the two concepts: the conventional correlation coefficient emphasizes the deviation of a variable from its mean when the system is near a steady state or dynamic equilibrium (Arkin & Ross 1995), but the PCC focuses on the deviation of a protein rate from its "expected" value, and it does not matter whether the system is near a steady state or not.

6.5 Concluding remarks

Several issues need to be addressed before one can apply the CDA to real data sets. In order to infer functions from mRNA and protein time series we have used a model system and employed many simplifying assumptions. In general, the magnitude of the translation rate for a given protein species, v_{Ti}, will depend on the distribution of translation resources among messages, the efficiency of the ribosomes, the level of message, the concentration of activated tRNAs, biases in codon usage, and mRNA structure. We simplified our analysis by making translation dependent only on message levels, but assumed uniform translation rates for all messages and no limitations on ribosomes. We have explored the effects of ribosome limitations and obtained similar results, while assuming that the limited ribosomes were uniformly distributed across all messages (not shown). We have neglected protein degradation, modification, or transport out of the cytoplasm, leaving the formation of protein-protein interactions as the lone mode for depleting a protein. These assumptions are reasonable for the phage

T7 model system, but they will need to be addressed for more complex systems. Finally, we have neglected stochastic effects due to the indeterminate behaviors of small numbers of molecules, effects that may be especially magnified when transcripts and proteins are just beginning to appear (Gillespie 1977; McAdams & Arkin 1997).

This work has focused on concepts rather than practice. We have saved for later questions regarding the data quality or frequency of data sampling needed to discriminate among different classes of trajectories. Moreover, although DNA-array technologies are beginning to provide global profiles of mRNA over time (DeRisi et al 1997; Spellman et al 1998), such profiles for free protein present both a significant technical challenge and an opportunity. Global protein profiles obtained by two-dimensional gel electrophoresis are typically carried out under conditions that destroy non-covalent protein-protein interactions (Fichmann & Westermeier 1999), which also destroy information about the free protein levels that we have used in our analysis. Current technologies to detect and analyze protein-protein interactions focus explicitly on forming and studying the protein-protein complex (Phizicky & Fields 1995). Our analysis suggests an opportunity to indirectly infer protein-protein interactions under physiological conditions if free proteins can be isolated from the cell, separated, and measured without disrupting existing complexes.

Acknowledgments

We thank the Office of Naval Research and the National Science Foundation Presidential Early Career Award to J.Y. for supporting this work.

C h a p t e r 7

TOWARDS GENERIC MODELING OF BIOLOGICAL SYSTEMS USING DYNETICA[15]

工 欲 善 其 事, 必 先 利 其 器

(Good tools are prerequisite to the successful execution of a job)

Chinese proverb

Abstract

Mathematical modeling and computer simulation may deepen our understanding of complex systems by testing the validity and consistency of experimental data and mechanisms, by generating experimentally testable hypotheses, and by providing new insight into the integrated behaviors of these systems. However, the application of this approach in biology has been hindered by the lack of software tools to build and analyze models. To meet this need, we have developed Dynetica – a simulator of dynamic networks – to facilitate model building for systems that can be expressed as reaction networks. A distinguishing feature of Dynetica is that it facilitates easy construction of models for genetic networks, where many reactions are the expression of genes and the interactions among gene products. In addition, it provides users the flexibility of performing time-course simulations using either deterministic or stochastic algorithms. Finally, since it is written in Java, Dynetica is platform-independent, allowing models to

[15] The content of this chapter has been submitted to Bioinformatics for publication (with Apirak Hoonlor and John Yin).

be easily shared among researchers. We anticipate that Dynetica will dramatically speed up the process of model construction and analysis for a wide variety of biological systems.

7.1 Introduction

Over the past several decades, mathematical modeling has arguably become an important tool in biological research. Owing to the lack of detailed information for many biological systems, past efforts in modeling have relied on relatively simple approaches, such as Boolean network modeling (Glass 1975; Glass & Kauffman 1973; Thomas 1973) and stoichiometric modeling (Clarke 1988; Fell 1992). In Boolean representations of gene networks, each gene is treated as having two states, ON or OFF, and the dynamics describes how genes interact to change one another's states over time (Hasty et al 2001). Although a Boolean model can provide insight into the qualitative behavior of the underlying system, it is usually overly simplified and tends to give ambiguous predictions (Kuipers 1986). A stoichiometric model represents the underlying system as a series of coupled chemical reactions. It does not require any information on the kinetics of the reactions, and as such is particularly attractive for systems where only sparse kinetic data are available or when steady-state assumptions can be justified (Bailey 2001; Varner & Ramkrishna 1999). Coupled with a technique called metabolic flux analysis (Fell 1992), stoichiometric models have played an instrumental role in shaping the field of metabolic engineering, by providing theoretic guidance for experimental manipulation of metabolic networks (Stephanopoulos et al 1998). Recently, stoichiometric models have proven powerful in characterizing the

underlying structure of metabolic networks by determining the elementary flux modes (Schuster et al 2000) or the null space base vectors (Schilling & Palsson 1998) and in predicting steady-state metabolic capabilities of several model organisms, such as *E. coli* (Edwards et al 2001; Schilling et al 1999) and *H. influenzae* (Edwards & Palsson 1999). But their applications are limited by their inability to predict the temporal evolution of these networks. To make such predictions, the stoichiometric structure of the reaction networks needs be supplemented with detailed kinetic information, resulting in kinetic models. Thanks to the rapid expansion of our knowledge in biology, kinetic modeling has become a realistic goal, particularly for the experimentally well-characterized systems. For example, kinetic models have recently been successfully applied to the analysis of a wide variety of biological systems, including bacterial chemotaxis signaling networks (Barkai & Leibler 1997; Spiro et al 1997), developmental pattern formation in *Drosophila* (von Dassow et al 2000), aggregation stage network of *Dictyostelium* (Laub & Loomis 1998), viral infection (Eigen et al 1991; Endy et al 1997; McAdams & Shapiro 1995; Reddy & Yin 1999; Shea & Ackers 1985; You et al 2002), circadian rhythms (Barkai & Leibler 2000; Smolen et al 2001), single cell growth (Shuler et al 1979), and physiological processes (Noble 2002; Quick & Shuler 1999; Winslow et al 2000).

A kinetic model essentially represents a mathematical integration of existing data and mechanisms on a particular system, and may be useful in a number of ways. By providing a global view of the underlying system, a kinetic model can be used to test the consistency in the experimental data or mechanisms (von Dassow et al 2000) or provide mechanistic explanations for counter-intuitive observations (Fallon & Lauffenburger

2000), to facilitate the formulation of experimentally testable hypotheses (Abouhamad et al 1998; Endy et al 2000; You et al 2002) or to test hypotheses that are difficult, expensive, or even impossible to explore experimentally with current technology (You & Yin 2002), and to provide insight into emergent properties, such as robustness (Alon et al 1999; Barkai & Leibler 1997; von Dassow et al 2000), which may be otherwise difficult to grasp intuitively. As models become more "realistic" by incorporating more detailed data and mechanisms, they may be treated as *in silico* organisms and used to explore applied or fundamental questions that are beyond the underlying system per se. For example, a phage T7 model has been employed to explore anti-viral strategies using anti-sense mRNAs (Endy & Yin 2000), to elucidate the nature of genetic interactions by *in silico* mutagenesis at the population level (You & Yin 2002), and to test data-mining strategies for identifying potential protein-protein interactions from gene expression data (You & Yin 2000). Moreover, advances in high-throughput biotechnologies for genome-wide gene expression profiling at the transcription and translation level provide additional challenges and opportunities for mathematical modeling, which may accelerate the characterization of whole organisms by allowing the understanding of gene expression data (at the mRNA level or the protein level) in their natural context. This point is demonstrated in a recent work where kinetic formulation of DNA microarray data was used to determine the timing of transcriptional onsets and cessation in *Dictyostelium* (Iranfar et al 2001).

Despite its potential benefits for fundamental and applied biological research, broader application of kinetic modeling has been hindered by the lack of powerful and

easy-to-use software tools for model construction and analysis. This is particularly true for experimental biologists who are often unfamiliar with numerical methods and programming. This aspect is probably best evidenced by the fact that the majority of mathematical models of biological systems have been developed by researchers trained in disciplines other than biology. Further, because of the lack of such tools, most published models were developed from scratch, which can be a tedious and error-prone process.

To address this issue, a number of programs that aim to facilitate the model construction and analysis have been developed in the last several years. These programs include Gepasi (Mendes 1993; Mendes 1997), DBsolve (Goryanin et al 1999), E-Cell (Tomita 2001; Tomita et al 1999), SCAMP (Sauro 1993), Virtual Cell (Schaff et al 1997; Schaff & Loew 1999; Schaff et al 2000), StochSim (Morton-Firth & Bray 1998), and STOCKS (Kierzek 2002). It would go beyond the scope of this current work to give a detailed account of these tools. Briefly, Gepasi, DBsolve, and SCAMP focus on the analysis of biochemical and metabolic networks. In addition to basic time-course simulations, these programs provide additional modules to explore the properties of metabolic networks. E-Cell aims to construct whole-cell models, and it has been applied to model a self-sustaining hypothetic cell (Tomita et al 1999) and a human erythrocyte (Tomita 2001). Virtual Cell is advantageous in that it accounts for the diffusion of molecules in addition to their reactions in describing cellular processes. Distinct from other programs, StochSim and STOCKS simulate the system dynamics using stochastic algorithms instead of deterministic algorithms. These two differ in that StochSim

employs a semi-empirical algorithm, while STOCKS uses the Gillespie algorithm (Gillespie 1977), which is rigorous for spatially homogenous systems. More extensive discussion of recent progress in the development of modeling tools may be found in excellent recent reviews (Arkin 2001; Loew & Schaff 2001).

We present here a unique, general-purpose computational framework for creating, visualizing, and analyzing mathematical models of biological networks, including biochemical, metabolic, signaling, and genetic networks. We call this program Dynetica, or a simulator of *dy*namic *net*works. Dynetica is distinct from other software packages in three aspects: (1) it facilitates the construction of kinetic models of genetic networks where most reactions are expression of genes; (2) it provides a visual representation of each model for interactive manipulation and interrogation; (3) it allows time-course simulations using both deterministic and stochastic algorithms. Furthermore, because it is written in Java, a platform-independent, object-oriented programming language, Dynetica can be run on most modern computers, which will facilitate the sharing of models among researchers. We anticipate that Dynetica will contribute significantly to advancing broader application of kinetic modeling in biological systems.

7.2 Modeling in Dynetica

7.2.1 Representation of generic reaction networks

A reaction network in Dynetica consists of a list of substances that interact with one another via a list of reactions. Kinetics of these reactions may be specified by a list of parameters (Figure 7.1). In addition to a tree structure, Dynetica provides a graphic representation of each reaction network. Figure 7.2 shows a hypothetical reaction network in Dynetica that consists of two reactions (Table 7.1). Each reaction is characterized by two basic attributes: its stoichiometry, which specifies the quantitative relationship between the substances in a reaction, and its kinetics, which specifies how fast (for non-equilibrated reactions) or to what extent (for equilibrated reactions) the reaction occurs.

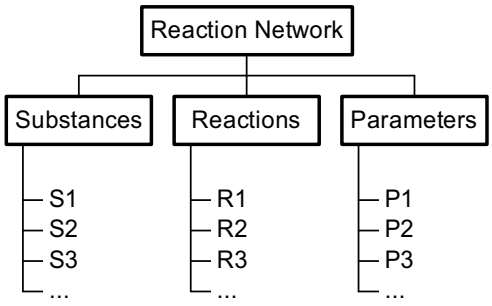

Figure 7.1 Representation of reaction networks in Dynetica. Each reaction network is represented as three lists: substances, reactions through which substances interact with one another, and parameters that specify the kinetics of the reactions.

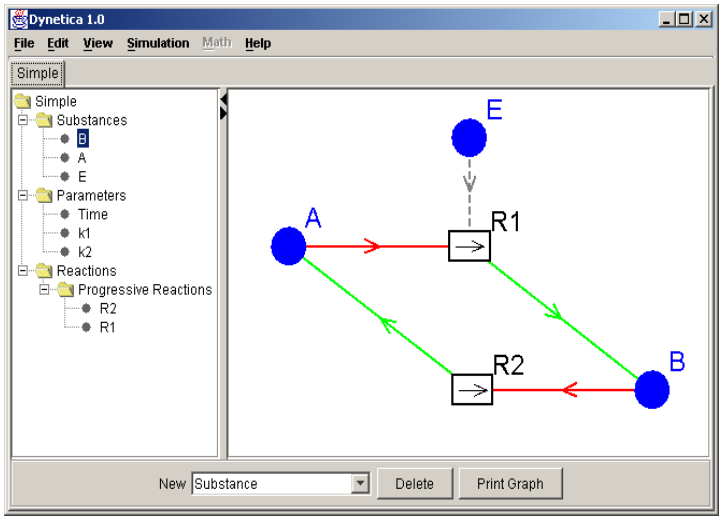

Figure 7.2 Screenshot of a hypothetical reaction network in Dynetica. The left panel shows the tree-structure view of the network, and the right panel gives a graphic representation. In the graph a green line indicates the production of the connected substance by the connected reaction, a red line represents the consumption of the connected substance by the connected reaction, and a gray dashed line indicates that the connected substance affects the kinetics of the connected reaction. See text for details of the reactions.

Table 7.1 The reactions in the simple reaction network shown in Figure 7.2.

Reaction	Stoichiometry	Kinetics
R1	$A \rightarrow B$	k_1 [A] [E] [a]
R2	$B \rightarrow A$	k_2 [B]

[a] The rate expression is actually written as k_1 [A] * [E] in Dynetica.

Dynetica employs two modules to describe generic reaction networks: a reaction parser and a mathematical expression parser. The reaction parser can interpret conventional chemical reaction formulas (using "→" as the separator between reactants and products), which specify the stoichiometry of reactions. The mathematical expression parser is used to interpret conventional mathematical expressions, which describe the kinetics of reactions. In Dynetica expressions both substances and parameters have values associated with them. The expression parser distinguishes between these entities by enclosing substance names with brackets. For example, the rate expression for reaction R1 in Table 7.1 is k1 [A] [E], which means the value of parameter k1 times the level of substance A and the level of substance E. The expression parser can interpret mathematical expressions composed of the operations and functions shown in Table 7.2. The kinetics of most chemical reactions can be formulated easily within this framework.

7.2.2 Representation of genetic networks

Genetic networks can be loosely defined as reaction networks involving gene expression processes, such as transcription of genes and translation of mRNAs. In Dynetica, a genetic network is treated as a special reaction network that contains one or more genomes (Figure 7.3A). Here a genome is defined as an entity composed of an

array of genetic elements, such as genes, promoters, and transcription terminators. Examples of genomes include genomes of cells and viruses, as well as plasmids.

Each genetic element is characterized by two attributes, namely, its starting and ending positions (in base-pair number) along the genome. A gene in Dynetica is a special genetic element characterized by several additional attributes: the RNA polymerase responsible for its transcription, the ribosome responsible for its translation, the name of its RNA, and the name of its protein (if the gene is to be translated), the relative transcription activity, and the relative translation activity. The relative transcription activity is essentially a weighting factor by which RNA polymerases are allocated to different genes, and the relative translation activity is the weighting factor by which ribosomes are allocated to different genes (more precisely, to different mRNAs). Genetic reactions can easily be formulated in Dynetica. Figure 7.3B demonstrates the Dynetica formulation of the central dogma of molecular biology. Essentially, the information transfer process from gene to mRNA to protein can be represented by two reactions. The transcription reaction specifies the conversion of nucleoside triphosphates (NTP) into mRNA, and is catalyzed by the gene and RNA polymerase (RNAP). The translation reaction specifies the conversion of amino acids (AA) into the protein, and is catalyzed by the mRNA and the ribosome.

Table 7.2 The mathematical operations and functions that are supported by Dynetica

	Symbols or expressions	Notes
Basic operations	+, -, *, /, ^	'^' represents to the power of.
Basic functions [a]	sin(*a*), cos(*a*), tan(*a*), sqrt(*a*), log(*a*)	log(*a*) returns the natural logarithm value of *a*
Special functions [a]	step(*a*, *b*)	returns 1 if $a \geq b$, and 0 otherwise
	compare(*a*, *b*)	returns 1 if $a > b$, 0 if $a = b$, and –1 if $a < b$
	pulse(*a*, *x*, *b*)	returns 1 if $a < x < b$, 0 otherwise
	random(*a*, *b*)	returns a random value between *a* and *b*
	rand()	returns a random value between 0 and 1
	min(*a*, *b*, *c*, …)	returns the minimum value from the list of arguments
	max(*a*, *b*, *c*,…)	returns the maximum value from the list of arguments

[a] Each of the symbols (*a*, *b*, *c* and *x*) may represent a simple variable or a mathematical expression.

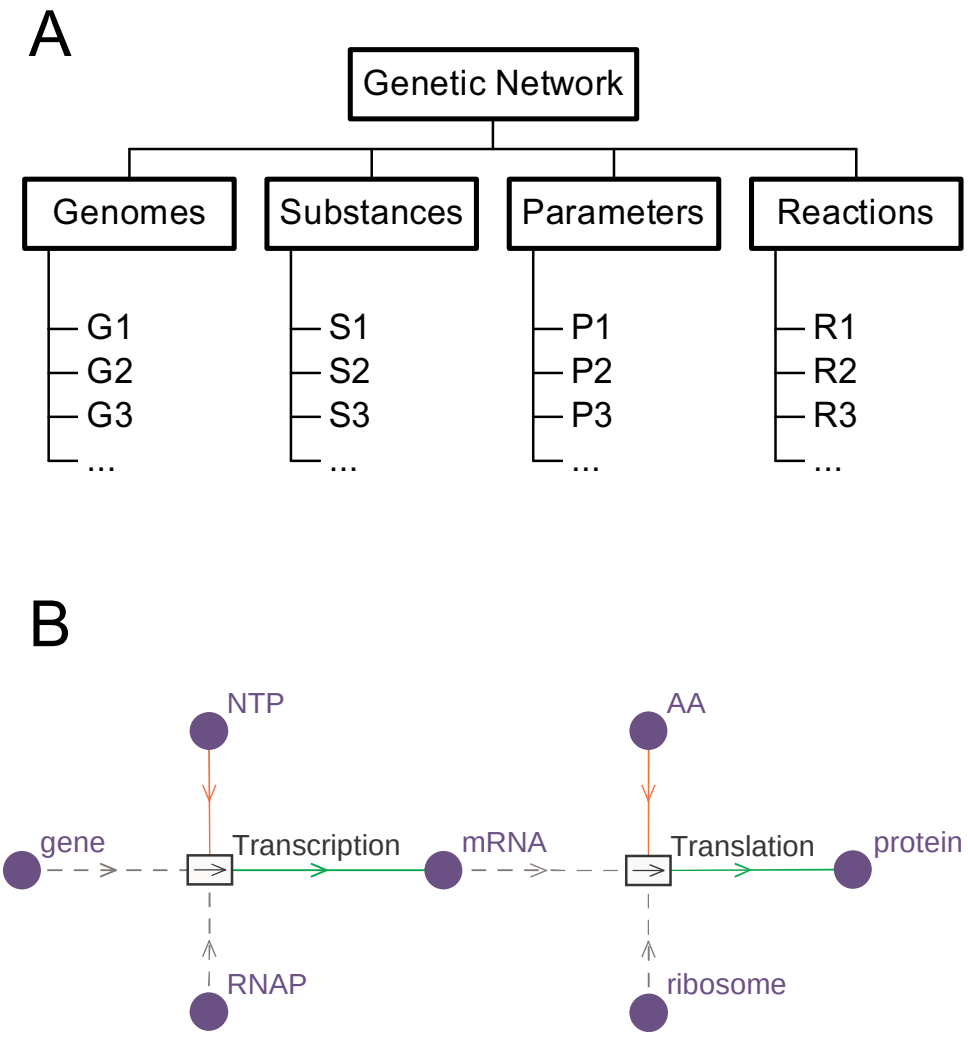

Figure 7.3 Formulation of genetic networks in Dynetica. (A) A genetic network in Dynetica is represented as a special reaction network that contains one or more genomes. (B) The central dogma represented in Dynetica.

Because expression of most genes follows the pattern as specified by the central dogma, Dynetica automatically creates a transcription reaction and a translation reaction for each gene that the user specifies in a genome. In addition, it also generates two reactions to represent the degradation of the gene products, the mRNA and the protein. In setting up the transcription reaction, we assume that the limiting step is the

elongation of the RNAP, and the transcription follows Michaelis-Menten kinetics with NTP as the substrate. For the translation reaction, we assume that the limiting step is the elongation of the ribosome, and the reaction follows Michaelis-Menten kinetics with AA as the substrate. Note that these automatically generated reactions are essentially "first-order approximations" by the program based on the genetic information provided by the user. These approximations are useful because they provide an initial estimate of gene expression dynamics. The user can then refine the stoichiometry and kinetics of such reactions as needed.

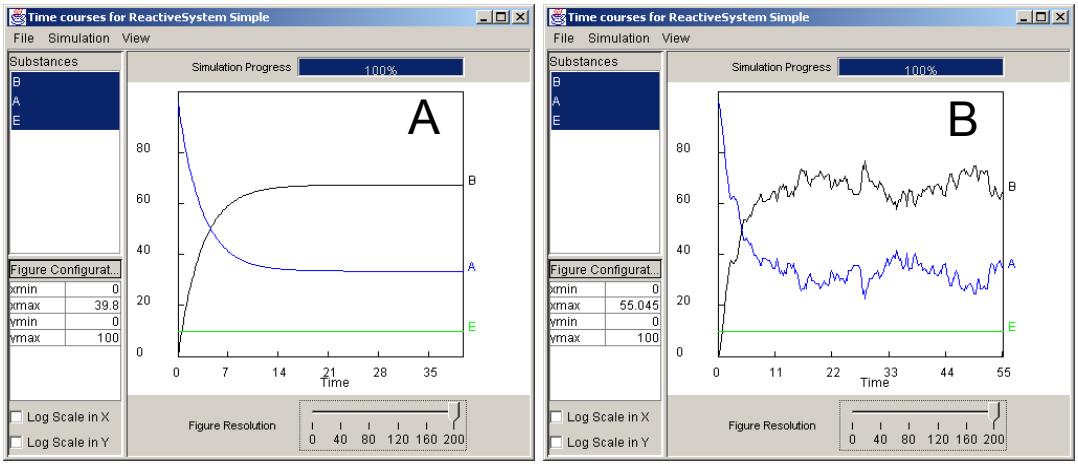

Figure 7.4 The simulation results from the reaction network in Figure 7.2 using both (A) deterministic and (B) stochastic algorithms

7.2.3 Simulation

A model in Dynetica gives a schematic representation of the corresponding system, but it does not specify how the system evolves over time. The latter will be determined by an algorithm. Here, an algorithm is defined as the scheme by which the system represented by the model will be updated as a function of time. It can be either

deterministic or stochastic. Deterministic algorithms include all the traditional numerical algorithms that are designed to solve coupled differential equations, such as fixed or variable time-step Runge-Kutta algorithms. A deterministic algorithm is appropriate when the continuity of the system can be justified. .

Stochastic algorithms focus on updating reactions in the system. For example, a widely used stochastic algorithm proposed by Gillespie (Gillespie 1977) updates a reactive system by determining, at each step, which and when the next reaction will occur. A stochastic algorithm is appropriate for a spatially homogeneous system where the interacting molecules are few that fluctuations in their numbers are significant. A number of researchers have strongly advocated the use of stochastic algorithms for modeling biological systems, especially for intracellular processes (Arkin et al 1998; Goss & Peccoud 1998; Kierzek 2002; Morton-Firth & Bray 1998).

The structure of a reaction network model in Dynetica is flexible enough to allow simulations by either deterministic or stochastic algorithms. Currently we have implemented three different algorithms: a fixed time-step 4th order Runge-Kutta algorithm, a variable time-step 4th order Runge-Kutta algorithm, and Gillespie's algorithm. By applying an algorithm to a model, we can generate the dynamics of the underlying system. Shown in Figure 7.4 are the results of deterministic and stochastic simulations with the model in Figure 7.2. In this particular case, both approaches generate qualitatively the same result: substance A is gradually converted into substance B until equilibrium is reached, whereas the level of substance E remains constant over

time. However, the details of the dynamics generated from these different approaches are quite different. For instance, there are no fluctuations in the substance concentrations as predicted by the deterministic simulation, but fluctuations are evident in the result from the stochastic simulation. In addition, because of the stochastic aspect of the Gillespie algorithm, every new simulation starting from the same initial condition will generate different dynamics (Gillespie 1977).

In addition to simulating the temporal evolution of a reaction network, Dynetica provides the basic functionality to explore how the dynamics of the network responds to the perturbations to the network, in terms of variations in parameter values or the initial levels of substances. This feature is desirable for simulating dosage curves and for identifying key system parameters that are important in determining overall behaviors of the system.

7.3 Applications

To demonstrate the application of Dynetica we use it to build two models: one for the *Dictyostelium* aggregation stage network, and the other for the intracellular growth cycle of phage T7. The aggregation stage network model is shown here as an example of a general reaction network. The phage T7 model shown as an example of a genetic network.

7.3.1 A Dictyostelium aggregation stage network model

Amoebae of *Dictyostelium discoideum* grow as independent cells in the soil, but aggregate and develop as a multicellular organism under starvation. It has been proposed that the aggregation stage network, which consists of seven interacting components, is responsible for regulating the expression of developmental genes in homogeneous populations of *Dictyostelium* shortly after starvation (Loomis 1998; Soderbom & Loomis 1998). Previously, a kinetic model was developed to analyze the dynamics of this signaling network (Laub & Loomis 1998). The model accounted for the interactions among seven molecular species, and was shown to be able to predict the oscillations in the enzyme activities during *Dictyostelium* development.

Based on (Laub & Loomis 1998), we used Dynetica to reconstruct the aggregation stage network model (Figure 7.5A, Table 7.3). Figure 7.5B shows a representative simulation result demonstrating stable oscillations in levels of the interacting components.

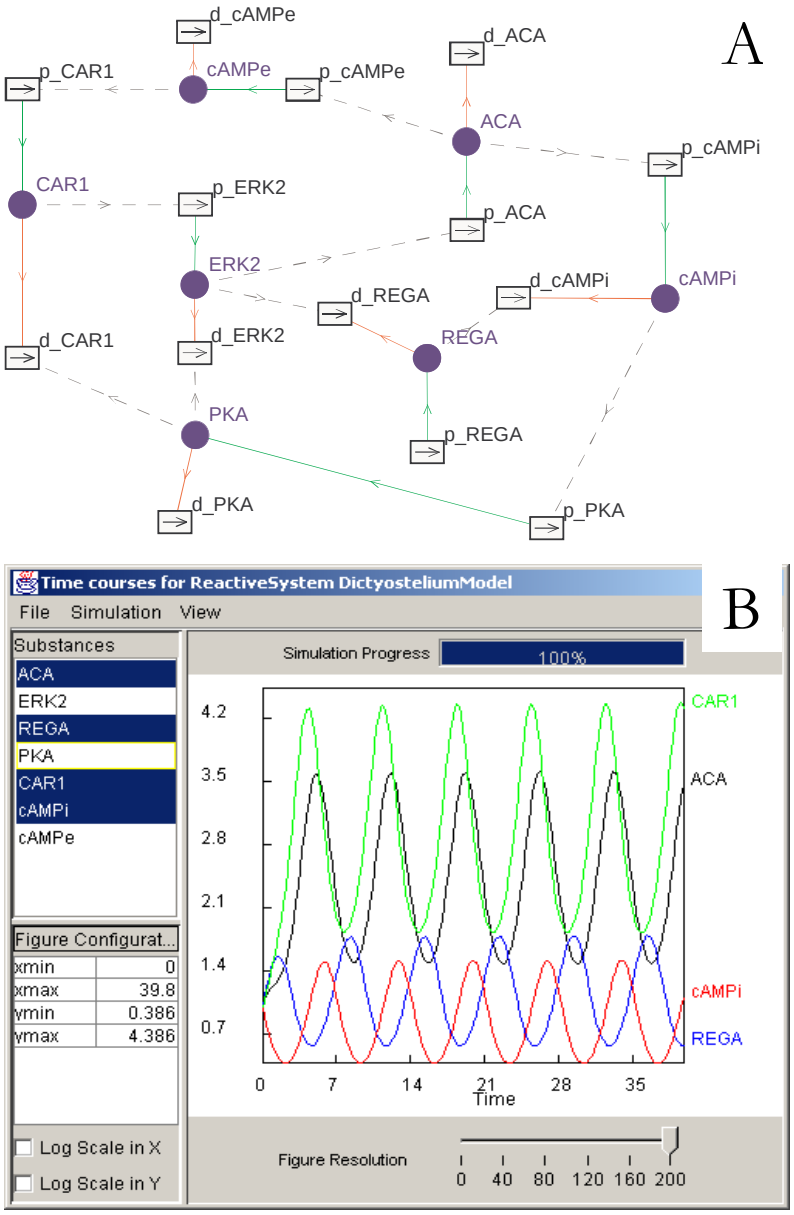

Figure 7.5 Aggregation stage network model. (A) The graphic representation of the reaction network. (B) A representative simulation result. The network was constructed based on the reference (Laub & Loomis 1998). The reactions involved in this network are shown in Table 7.3. The parameter values for the simulation are: k1 =1.4, k2 = 0.9, k3 = 2.5, k4 = 1.5, k5 = 0.6, k7 = 2.0, k8 = 1.3, k9 = 0.3, k10 = 0.8, k11 = 0.7, k12 = 4.9, k13 = 18, k14 = 1.5 (W. Loomis, personal communication). The initial levels of all substances were set to be 1.0, and the variable time-step 4th order Runge-Kutta algorithm was used for the simulation.

Table 7.3 The production reactions in the aggregation stage network [a]

Reaction	Stoichiometry	Kinetics	Notes
p_ACA	→ ACA	k_1 [ERK2]	Activation of ACA by ERK2
d_ACA	ACA →	k_2 [ACA]	degradation of ACA
p_PKA	→ PKA	k_3 [cAMPi]	Activation of PKA by cAMPi
d_PKA	PKA →	k_4 [PKA]	degradation of PKA
p_ERK2	→ ERK2	k_5 [CAR1]	Activation of ERK2 by CAR1
d_ERK2	ERK2 →	k_6 [ERK2] [REGA]	degradation of ERK2 (catalyzed by REGA)
p_REGA	→ REGA	k_7	constant production of REGA
d_REGA	REGA→	k_8 [REGA] [ERK2]	degradation of REGA (catalyzed by ERK2)
p_cAMPi	→ cAMPi	k_9 [ACA]	Activation of cAMPi by ACA
d_cAMPi	cAMPi →	k_{10}[REGA][cAMPi]	degradation of cAMPi (catalyzed by REGA)
p_cAMPe	→ cAMPe	k_{11} [ACA]	activation of cAMPe by ACA
d_cAMPe	cAMPe→	k_{12} [cAMPe]	degradation of cAMPe
p_CAR1	→ CAR1	k_{13} [cAMPe]	activation of CAR1 by cAMPe
d_CAR1	CAR1→	k_{14} [CAR1][PKA]	degradation catalyzed by PKA

[a]Although recent studies have suggested a slightly revised reaction network (http://www.biology.ucsd.edu/labs/loomis/network/laubloomis.html), the published model suffices to illustrate the usage of Dynetica.

7.3.2 A phage T7 model

The model presented here (Figure 7.6) is a simplified version of the full T7 model described in previous chapters. The major difference between the current model and the full model is that a simplified genome is used here (Figure 7.6A). This simplified genome contains 20 essential T7 genes. The regulatory effect of promoters and transcription

terminators is accounted for by specifying the relative transcription activity of each gene. As a result, RNA polymerases are allocated to different genes based on their relative transcription activities, whereas in the complete model RNA polymerases are allocated based on the relative strengths of promoters (Chapter 2). The resulting T7 reaction network contains 91 reactions and 55 substances, excluding genes (Figure 7.6B). In this network, the reactions describing expression of genes and degradation of gene products are automatically generated by Dynetica. Although the network diagram is overall complex, it highlights several features of the system. First, most substances are involved in two reactions, one for production (green line) and the other for consumption (red line). Second, several nodes (as labeled) are highly connected. For example, the nodes for amino acid and NTP are highly connected because these two substances are used as precursors for transcription and translation reactions, respectively. Likewise, the nodes for T7 RNAP and ribosome are highly connected because they are used as catalysts for transcription and translation reactions, respectively.

Like the full model, the current model accounts for the major steps of T7 infection: transcription of viral genes, translation of the resulting mRNAs, interactions between regulatory proteins, host DNA degradation and T7 DNA replication, procapsid assembly, and eventually production of phage progeny. A representative simulation result showing the time courses of three viral components is presented in Figure 7.6C. It illustrates the synthesis of T7 DNAs and procapsids, and the packaging of T7 DNAs into procapsids to form viral progeny. Overall, this simplified model captures the main features of viral growth as predicted by the more comprehensive model.

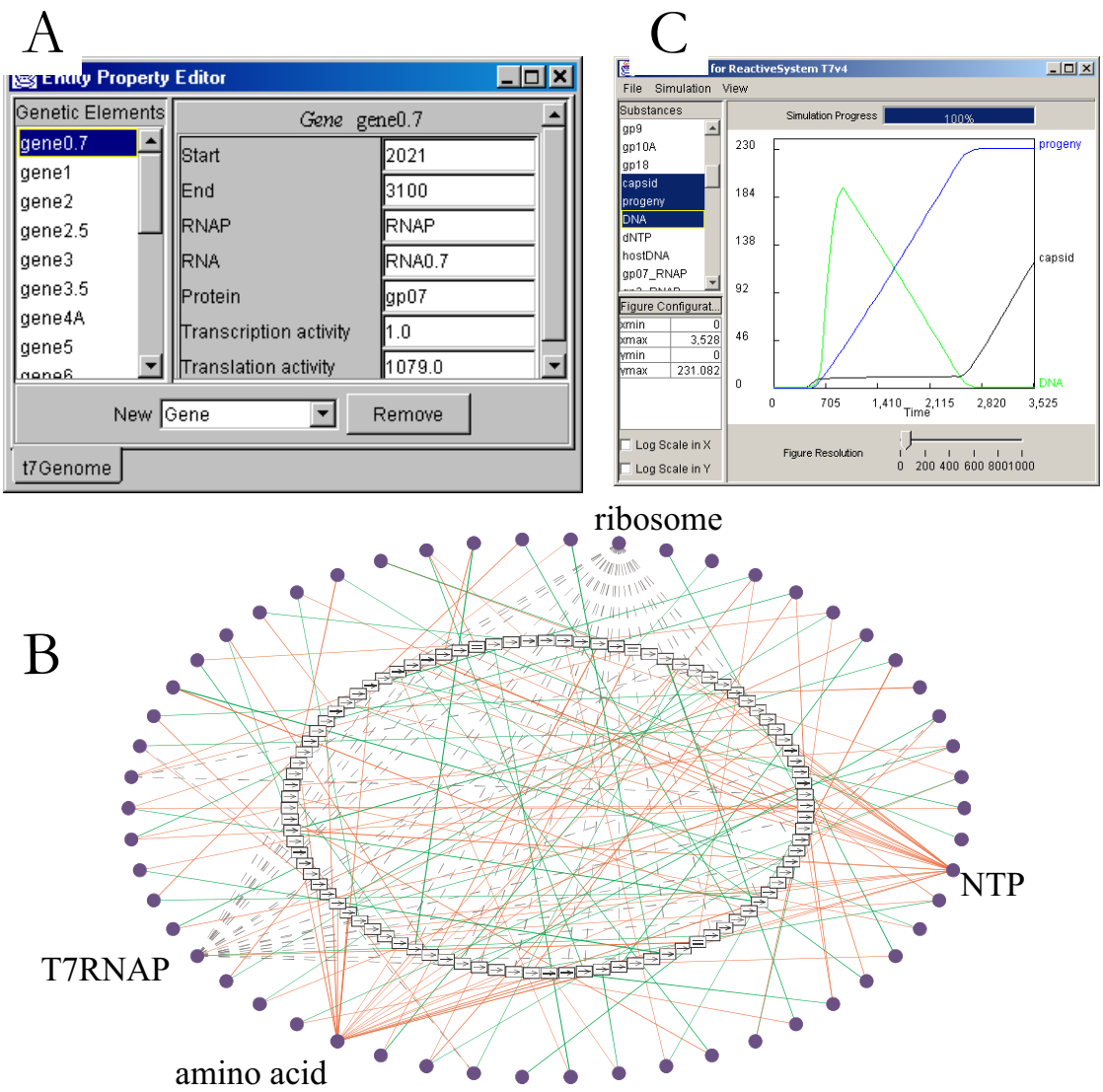

Figure 7.6 A simplified phage T7 model. (A) The simplified T7 genome. The left panel shows a list of genes in the genome (not all genes are shown); the right panel shows the attributes of the currently selected gene. (B) The graphic representation of the reaction network. The reactions describing transcription and translation of genes were automatically generated by Dynetica. (C) A representative simulation result showing the time courses of three viral components.

7.4 Discussion

We have developed Dynetica to facilitate the construction, visualization and analysis of mathematical models for biological systems that can be formulated as a coupled system of reactions. With Dynetica, the user need only specify the chemistry of this system, that is, what components are in the system and how they interact. Throughout the model-building process, the user need not write any differential equations, or formulate numerical algorithms to conduct simulations. Instead, the numerics is automatically handled by the program. Thanks to this feature, the user can focus on the model itself and its practical relevance rather than the technical aspects of computer simulation. Furthermore, by providing a graphic view of the underlying reaction network Dynetica will facilitate the interactive manipulation and analysis of each model.

Dynetica's ability to perform both deterministic and stochastic simulations on the same model may facilitate comparative studies of these two approaches. Deterministic algorithms have been traditionally used to simulate the dynamics of a system of coupled reaction network. However, the small numbers of interacting components in some intracellular processes may become an issue. First, the continuity of these systems is no longer warranted. Second, fluctuations in the concentrations of the reacting components may significantly impact the system dynamics. Because of these issues, some researchers have questioned the use of deterministic algorithms in simulating the behaviors of

biological systems, and suggested using stochastic algorithms instead (Arkin et al 1998; Goss & Peccoud 1998; Kierzek 2002; Morton-Firth & Bray 1998; Srivastava et al 2002). They have shown that stochastic simulations often produce dynamics drastically different from what is predicted by deterministic simulations. Moreover, some have argued that a stochastic simulation more accurately and more completely accounts for the temporal evolution of a well-stirred chemical reaction network than does a deterministic algorithm (Gillespie 1977; McAdams & Arkin 1998). Nonetheless, since a stochastic algorithm only gives accurate solutions for a well-stirred system, it may not be applicable for intracellular processes. It is unclear whether it is more appropriate than a deterministic approach in modeling such processes. To this end, Dynetica may be employed to simulate a system using both deterministic and stochastic approaches and explore which approach is more appropriate in a particular situation.

With its present underlying software structure, Dynetica can easily be extended in its functionality and flexibility. It has a software module that automates the construction of a genetic network model based on the organization of genetic elements along the genome. In achieving this functionality, we made simplifying assumptions regarding the organization of the genome. For each gene, Dynetica will automatically generate a transcription reaction, a translation reaction, and degradation reactions for the resulting mRNA and protein. However, in reality, there are also genes for tRNA and rRNA that do not have protein products. Future modifications of the program will be needed to represent and distinguish different kinds of genes. New numerical algorithms can be implemented, so that the user will have the freedom in choosing the most appropriate

one for a given situation. Further, we are developing model templates for different types of biological systems, such as signaling pathways, viruses, and single cells. Like the document templates one may encounter in many word-processing programs such as Microsoft® Word, these templates will further facilitate the model-building process, particularly for new users. An emerging challenge for the modeling community lies in the interchange of models constructed using different software tools, as listed in the introduction section. Recently, there have been many efforts toward developing modeling standards for biology modeling, such as the SBML (Systems Biology Markup Language) project (http://www.cds.caltech.edu/erato) and the CellML (Cell Markup Language) project (http://www.cellml.org). To provide exchangeable mathematical models, we plan to implement software modules to import models constructed with other tools, or written in standard modeling languages. Finally, we plan to implement software modules to annotate models; we expect this functionality will further facilitate the communication of mathematical models as a representation of the underlying biological systems.

The evolution of biological network modeling can be compared to that of the molecular dynamics simulation, which uses physical principles to compute the structure and dynamics of biological molecules. Although the development and use of molecular dynamics simulation programs were initially much restricted to researchers with strong background in theoretic physics and mathematics, it is the development of powerful and user-friendly tools that has established this computational approach as a routine tool for structural studies of natural or synthetic biological molecules (Loew & Schaff 2001).

Similarly we envision that, Dynetica, together with other emerging modeling tools, will promote a broader application of mathematical models in cell biology by serving as a computational platform to create, analyze and exchange such models.

Acknowledgments

We thank Dr. William Loomis for clarifications on the aggregation stage network model, and Yi-Fan Chen for technical assistance. Funding of this work was provided by the National Science Foundation.

C h a p t e r 8

CONCLUDING REMARKS

"To see a world in a grain of sand
And a heaven in a wild flower,
Hold infinity in the palm of your hand
And eternity in an hour."

William Blake

8.1 Lessons from modeling phage T7

A genetically structured model of phage T7 intracellular growth in *E. coli* (Endy 1997; Endy et al 1997) has been recast in an object-oriented framework, extended by implementing a simple model representing the host physiology, and improved by incorporating a more mechanistic description of several steps of T7 infection. The development of the T7 model demonstrates the feasibility of building integrated models for complex biological systems based on data and mechanisms at the genetic and molecular levels.

The current phage T7 model, along with the previous T7 models, is distinguished from other kinetic models in biology in that it completely links the genotype and the phenotype of an organism – phage T7. The ability to account for the host environment further distinguishes the current T7 model from its precursors and other models. Many models, while detailed, have focused on the systems of interest themselves and treated

their environments as black boxes. By explicitly treating the host cell as bag of resources, whose levels are characterized by parameters obtained experimentally, we have been able to examine the development of phage T7 in the context of its host environment. This aspect has been exploited to investigate a number of biological questions of broad relevance. These applications have indeed demonstrated the important role of environments in various aspects of T7 biology. The environment clearly has a big impact on the development of the wild-type phage T7 as a whole (Chapter 3). Chapter 4 further suggests that the growth environment may play a role in the design of phage T7 in terms of its parametric configuration and its genomic structure. Specifically, the fact that every host cell will have limited resources appears to have forced phage T7 to adjust its design, through evolution, so that it can take full advantage of these resources for its development. Chapter 5 demonstrates how the environment may affect the nature and degree of epistasis among deleterious mutations at a population level.

In exploring biological questions, a major advantage of computer simulations is their high efficiency, which makes it feasible to test sophisticated hypotheses on a scale that is unattainable with real biological systems in a practical time frame. Another advantage is that by simulation we can accurately "measure" the growth of an organism, which is often a very difficult task in lab experiments. However, an immediate question regarding these simulations is to what extent the conclusions are applicable for real biological systems. Some conclusions of our work can be readily tested by experiment. The dependence of T7 growth on the host physiology has been validated by experiment (Chapter 3). The outstanding role of the translation machinery in determining this

dependence can also be tested by carefully designed experiment (Chapter 3, Discussion). Admittedly, other conclusions are based on more drastic extrapolations from the base model and may be more removed from lab experiments. To explore the design of phage T7 and the nature of epistasis, we created hundreds of thousands of mutants that have changed parameters or alternative genomes. In these mutants, parameter changes may be more easily realized experimentally. For example, a change in the activity of a promoter can be achieved by introducing point mutations to the promoter sequence, and the elongation rate of a polymerase (for synthesizing RNA or DNA) may be slowed down (Makarova et al 1995) and may even be increased (Sasaki et al 1998) by mutations in the polymerase gene. In fact, it has been argued that there has been little if any selection pressure to maximize the reaction rate constants for most metabolic enzymes (Kacser & Burns 1973). If this is the case, then it is probably feasible to experimentally change various kinetic parameters in either direction (Bray & Lay 1994). In comparison, changing the genomic structure experimentally will be more difficult to make but can still be done to a certain degree (Endy et al 2000).

Even when some *in silico* mutations, such as random genome permutations, are not readily realizable experimentally, conclusions from simulations are still useful. In fact, we can consider the T7 model as a digital virus that captures all the essential aspects of T7 infection: entry of DNA, expression of gene products, replication of DNA, assembly of viral procapsids, and the formation of progeny. Because of this direct mapping between the model and the reality, lessons from simulations may provide insights into the understanding of the real system.

8.2 Modeling beyond T7: challenges and opportunities

Although the T7 model is specific to T7 biology and may not be directly applied to a different system, the same kinetic principles and mathematical techniques can be used to describe other biological processes, as long as experimental data and mechanisms are available. As a generic modeling tool, Dynetica will facilitate modeling of biological systems, particularly for researchers unfamiliar with programming and numerical methods. However, Dynetica will still need the user to specify the details of the system, namely the components that are in the system and the interactions among these components. With the numerics and programming difficulties taken care of, the rate-limiting step for modeling shifts to the collection of experimental data and appropriate model formulation. Then, two important questions naturally come up: how much information we will need to build a sensible model? Given enough data and mechanisms, how much detail we should put into a model?

Answers to both questions will depend on the objective of modeling. For example, if we were to build a kinetic model that can accurately predict the entire life cycle of phage T7 infection, starting from the binding of T7 to the host cell (note that a simulation with the current model starts from the point when T7 DNA begins to enter the host cell) to the lysis of host cell by T7 progeny, we would not have enough experimental data. However, the current model is still useful in that it captures well the intracellular events of T7 infection. In this sense, we do have sufficient information. The same reasoning may apply to more complex systems. Although we may be quite far from a complete understanding of the growth of an entire cell or organism, we already

have sufficient information to build models for some isolated, well-studied networks that have identifiable network output, such as the *E. coli* chemotaxis signaling pathway (Barkai & Leibler 1997) and the segment network for *Drosophila* development pattern formation (von Dassow et al 2000). Similarly, the level of detail to be incorporated in the model should be consistent with the goal of the model. Again, take phage T7 model as an example. If our goal were to predict only the growth curve of the T7 infection, we would not need to put as much detail as we have done. In fact, a simpler model, such as the simplified T7 model presented in Chapter 7, may equally well predict the curve. However, the power of phage T7 model is beyond predicting a T7 growth curve only. The model also provides a means to explore how various host factors and viral components, including genes and their products, interact with one another and with their environment to determine the integrated behavior of the virus.

In light of these brief discussions, it seems appropriate to leave the primary responsibility of model formulation to the user of Dynetica (or other modeling tools). Yet, I should note that future enhancement of the program is likely to further facilitate the model-building process. In the long term, the ability of Dynetica to build models for generic reaction networks will present an opportunity to create a database of the integrated models of a wide spectrum of biological systems. I call such a database PhenoBank by analogy to GenBank. PhenoBank will differ from GenBank in that each entry in it will be an integrated model representing the knowledge base for a biological system, such as signaling pathways, viruses and single cells, and this model can be used to perform simulations that may reflect the behavior (in a sense, the phenotype) of the

corresponding system. Unlike the existing pathway databases, such as Science's STKE (Signal Transduction Knowledge Environment, http://stke.sciencemag.org/), PhenoBank will host models that can be used for dynamic simulations and analysis by a modeling tool, such as Dynetica.

Just as GenBank has greatly impacted biological research at large by facilitating the storage and access of genetic sequence data, and the comparative studies based on these data (a quick PubMed search with the keyword "GenBank" retrieved nearly 2,700 research articles as of 5/8/2002), I anticipate that PhenoBank will offer similar benefits by facilitating the process of storing and retrieving computer models, and by promoting comparative studies on the structures of related biological systems. Moreover, PhenoBank may impact the strategies of teaching biology in the future. It is probably not unreasonable to predict that, 10 years from now, some of the fundamental principles of gene regulation will be taught along with demonstrations using computer simulations, which may employ models stored in PhenoBank. With the increasing number of biological models documented in PhenoBank, we may begin to track the parameters that may appear frequently in different systems, for instance, the processing rates of RNA polymerases and ribosomes for various genetic systems. By doing so, we may determine a distribution of reasonable values for the parameters of interest. In situations where the value of such a parameter is unavailable for a system, we may select a value from the predetermined distribution and refine it when necessary. Further, we may also detect potentially erroneous parameter values by comparing the input values with the predetermined distribution.

I envision that Dynetica and the proposed PhenoBank may synergistically benefit broad biological research by providing a computational framework for creating, analyzing and sharing mathematical models and, at a higher level, the comparative study of related biological systems. In addition, it is potentially useful for the pharmaceutical industry, since it may facilitate the identification of pharmaceutical targets in disease-associated pathways, and the evaluation of pharmaceutical efficacy using computer simulations.

APPENDIX

Nomenclature and abbreviations

A0	The leftward EcRNAP promoter in T7 genome
AA	Amino acid
Ai (i= 1, 2, 3)	Major EcRNAP promoter i in T7 genome
bp	Base pairs
C	The length of the C period of the *E. coli* growth cycle (min)
CDA	Correlated deviation algorithm
C_{ij}	The correlation coefficient of the DDFs of proteins *i* and *j*
C_n	The T7 procapsid assembly nucleation level (*M*)
D	The length of the D period of the *E. coli* growth cycle (min)
DDF	Dynamic deviation factor
D_i	The DDF for protein *i*
$\overline{D_i}$	The time-averaged DDF for protein *i*
D_{ik}	The DDF from protein I evaluated at the *k*th time point on a discretized time course
DNAP	DNA polymerase
dNTP	deoxyribonucleoside triphosphate
EcRNAP	*E. coli* RNA polymerase
eTE	Efficiency of the terminator TE

$e_{T\phi}$	Efficiency of the terminator Tϕ
G_c	The total amount of DNA in an *E. coli* cell (number of *E. coli* genome equivalents)
K_1	The association constant between EcRNAP and gp0.7 (M^{-1})
K_2	The association constant between EcRNAP and gp2 (M^{-1})
K_3	The association constant between T7RNAP and gp3.5 (M^{-1})
k_a	The procapsid assembly rate constant ($M^{3.8}/s$)
k_{dp}	T7 protein degradation rate constant (1/s)
k_{dm}	T7 mRNA degradation rate constant (1/s)
k_E	Ribosomal elongation rate (bases/s)
k_{pack}	T7 DNA packaging rate constant (bp/s)
k_{PD}	Replisome elongation rate (bp/s)
k_{PE}	EcRNAP elongation rate (bases/s)
k_{PT7}	T7RNAP elongation rate (bases/s)
k_{Ti}	The translation rate coefficient for protein *i* ((protein molecules)/ [(mRNA molecules)·s])
L_{DNA}	The length of T7 DNA (base pairs)
L_{in}	The length of T7 DNA that has entered the host cell
L_{Ri}	The length of the internalized part of mRNA *i* (bases)
N_c	The number of T7 major capsid protein molecules required to form a procapsid
N_p	The initial number of active EcRNAPs in an *E. coli* cell.
N_R	The initial number of active ribosomes in an *E. coli* cell

N_r	The number of elongating replisomes
NTP	Nucleoside triphosphate
ODE	Ordinary differential equation
P	The total amount of protein in an *E. coli* cell (number of AA residues)
PCC	Protein correlation coefficient
PCM	Protein correlation matrix
PDE	Partial differential equation
P_i	The level of protein i (molecules/cell)
P_R	The level of the limiting species of procapsid and DNA during T7 DNA packaging
R	The total amount of RNA in the host cell (number of nucleotides)
R_i	The level of mRNA i (molecules/cell)
RNAP	RNA polymerase
S_{Ei}	The density of EcRNAP along the internalized region of mRNA i (molecules/base)
S_{T7i}	The density of T7RNAP along the internalized region of mRNA i (molecules/base)

T7RNAP	T7 RNA polymerase
TE	Transcription terminator for EcRNAP
Tϕ	Transcription terminator for T7RNAP
V_c	Volume of the host cell (*L*)
v_d	Depletion rate of a protein by assembly of T7 procapsids or the packaging of progeny particles (protein molecules/s)
v_{Di}	Depletion rate of protein *i* (protein molecules/s)
v_{Ti}	Production rate of protein *i* due to translation (protein molecules/s)
v_{T7}	Production rate of T7 progeny (T7 particles/s)
W(w)	Fitness (lower case represents normalized value)
ϕi	T7RNAP promoter *i* (*i* may be non-integer)

BIBLOGRAPHY

Abouhamad, W. N., Bray, D., Schuster, M., Boesch, K. C., Silversmith, R. E., Bourret, R. B. 1998. Computer-aided resolution of an experimental paradox in bacterial chemotaxis. *J Bacteriol* 180:3757-64.

Adams, M. H. 1959. *Bacteriophages* New York: Interscience

Akutsu, T., Miyano, S., Kuhara, S. 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*:17-28

Alon, U., Surette, M. G., Barkai, N., Leibler, S. 1999. Robustness in bacterial chemotaxis. *Nature* 397:168-71.

Arkin, A., Ross, J. 1995. Statistical Construction of Chemical Reaction Mechanism from Measured Time-Series. *J. Phys. Chem* 99:970-979

Arkin, A., Ross, J., McAdams, H. H. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* 149:1633-48.

Arkin, A., Shen, P., Ross, J. 1997. A Test Case of Correlation Metric Constructionof a Reaction Pathway from Measurements. *Science* 277:1275-1279

Arkin, A. P. 2001. Synthetic cell biology. *Curr Opin Biotechnol* 12:638-44.

Bailey, J. E. 2001. Complex biology with no parameters. *Nat Biotechnol* 19:503-4.

Barkai, N., Leibler, S. 1997. Robustness in simple biochemical networks. *Nature* 387:913-7.

Barkai, N., Leibler, S. 2000. Circadian clocks limited by noise. *Nature* 403:267-8.

Bartel, P. L., Roecklein, J. A., SenGupta, D., Fields, S. 1996. A protein linkage map of Escherichia coli bacteriophage T7. *Nat. Genet.* 12:72-77

Becskei, A., Seraphin, B., Serrano, L. 2001. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo J* 20:2528-35.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., Wheeler, D. L. 2002. GenBank. *Nucleic Acids Res* 30:17-20.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* 277:1453-74.

Bork, P., Koonin, E. V. 1998. Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet* 18:313-8.

Bray, D., Lay, S. 1994. Computer simulated evolution of a network of cell-signaling molecules. *Biophys. J.* 66:972-977

Bremer, H., Dennis, P. P. 1996. Modulation of chemical composition and other parameters of the cell by growth rate. In *Escherichia coli and Salmonella: Cellular and Molecular Biology II*, ed. F. C. Neidhardt, R. Curtiss III, C. G. J. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaecheter, H. E. Umbarger. pp. 1553-1569. Vol. II. Washington, DC: ASM Press

Butler, D. 2001. Publication of human genomes sparks fresh sequence debate. *Nature* 409:747-8.

Cairns, J., Stent, G. S., Watson, J. D., eds. 1992. *Phage and the origins of molecular biology*. Plainview, NY: Cold Spring Harbor Press

Calson, J. M., Doyle, J. 2000. Highly Optimized Tolerance: Robustness and Design in Complex Systems. *Physics Review Letters* 84:2529-2532

Carlton, R. M. 1999. Phage therapy: past history and future prospects. *Arch. Immunol. Ther. Exp.*

Charlesworth, B. 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res.* 55:199-221

Clarke, B. L. 1988. Stoichiometric network analysis. *Cell Biophys* 12:237-53.

Cohen, S. S. 1947. The synthesis of bacterial viruses in infected cells. *Cold Spring Harbor Symp. Quant. Biol.* 12:35-49

Cohen, S. S. 1949. Growth requirements of bacterial viruses. *Bact. Rev.* 13:1-24

Cohen, S. S. 1953. Studies on controlling mechanisms in the metabolism of virus-infected bacteria. *Cold Spring Harbor Symp. Quant. Biol.* 18:221-235

Crick, F. 1970. Central dogma of molecular biology. *Nature* 227:561-3.

Crow, J. F. 1970. Genetic loads and the cost of natural selection. In *Mathematical topics in population genetics*, ed. K. Kojima. pp. 128-177. Springer, Heidelberg, Germany:

Davies, E. K., Peters, A. D., Keightley, P. D. 1999. High frequency of cryptic deleterious mutations in Caenorhabditis elegans. *Science* 285:1748-51

Dayton, C. J., Prosen, D. E., Parker, K. L., Cech, C. L. 1984. Kinetic measurements of *Escherichia coli* RNA polymerase association with bacteriophage T7 early promoters. *J. Biol. Chem.* 259:1616-1621

de Visser, J. A., Hoekstra, R. F., van den Ende, H. 1997a. An experimental test for synergistic epistasis and its application in Chlamydomonas. *Genetics* 145:815-819.

de Visser, J. A. G. M., Hoekstra, R. F. 1998. Synergistic epistasis between loci affecting fitness: evidence in plants and fungi. *Genet. Res. Camb.* 71:39-49

de Visser, J. A. G. M., Hoekstra, R. F., van den Ende, H. 1996. The effect of sex and deleterious mutations on fitness in Chlamydomonas. *Proc. R. Soc. Lond. B* 263:193-200

de Visser, J. A. G. M., Hoekstra, R. F., van den Ende, H. 1997b. Test of interaction between genetic markers that affect fitness in Aspergillus niger. *Evolution* 51:1499-1505

Delbrück, M. 1946. Bacterial viruses or bacteriophages. *Biol. Revs. Cambridge Phil. Soc.* 21:30-40

Demerec, M., Fano, U. 1944. Bacteriophage-resistant mutants in E. coli. *Genetics* 30:119-136

DeRisi, J. L., Iyer, V. R., Brown, P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686

Donachie, W. D., Robinson, A. C. 1987. Cell division: parameter values and the process. In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular*

Biology, ed. J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, H. E. Umbarger. pp. 1578-1593. Vol. 2. Washington, DC: ASM Press

Dunn, J. J. 1976. RNase III cleavage of single-stranded RNA. Effect of ionic strength on the fideltiy of cleavage. *J Biol Chem* 251:3807-14

Dunn, J. J., Studier, F. W. 1983. Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of T7 Genetic Elements. *J. Mol. Biol.* 166:477-535

Eckerskorn, C., Strupat, K., Karas, M., Hillenkamp, F., Lottspeich, F. 1992. Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization. *Electrophoresis* 13:664-665

Edwards, J. S., Ibarra, R. U., Palsson, B. O. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19:125-30.

Edwards, J. S., Palsson, B. O. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem* 274:17410-6.

Egelman, H. H., Yu, X., Wild, R., Hingorani, M. M., Patel, S. S. 1995. Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proc. Natl. Acad. Sci. USA* 92:3869-3873

Eigen, M. 1971a. Molecular self-organization and the early stages of evolution. *Q Rev Biophys* 4:149-212.

Eigen, M. 1971b. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58:465-523.

Eigen, M., Biebricher, C. K., Gebinoga, M., Gardiner, W. C. 1991. The hypercycle. Coupling of RNA and protein biosynthesis in the infection cycle of an RNA bacteriophage. *Biochemistry* 30:11005-11018

Elena, S. F., Lenski, R. E. 1997. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390:395-398

Ellis, E. L., Delbrück, M. 1939. The growth of bacteriophage. *J. Gen. Physiol.* 22:365

Endy, D. 1997. Development and application of a genetically structured simulation for bacteriophage T7. Dartmouth College

Endy, D., Kong, D., Yin, J. 1997. Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7. *Biotech. Bioeng.* 55:375-389

Endy, D., Yin, J. 2000. Toward antiviral strategies that resist viral escape. *Antimicrob Agents Chemother* 44:1097-9.

Endy, D., You, L., Yin, J., Molineux, I. J. 2000. Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acad. Sci. U S A* 97:5375-5380

Fallon, E. M., Lauffenburger, D. A. 2000. Computational model for effects of ligand/receptor binding properties on interleukin-2 trafficking dynamics and T cell proliferation response. *Biotechnol Prog* 16:905-16.

Fell, D. A. 1992. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J* 286:313-30.

Fichmann, J., Westermeier, R. 1999. 2-D protein gel electrophoresis. An overview. *Methods Mol. Biol.* 112:1-7

Fields, S., Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340:245-6.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., et al. 2000. A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242:369-79.

Fry, J. D., Keightley, P. D., Heinsohn, S. L., Nuzhdin, S. V. 1999. New estimates of the rates and effects of mildly deleterious mutation in Drosophila melanogaster. *Proc Natl Acad Sci U S A* 96:574-9.

Garcia, L. R., Molineux, I. J. 1995a. Incomplete entry of bacteriophage T7 DNA into F plasmid-containing Escherichia coli. *J Bacteriol* 177:4077-83

Garcia, L. R., Molineux, I. J. 1995b. Rate of translocation of bacteriophage T7 DNA across the membranes of Escherichia coli. *J Bacteriol* 177:4066-76

Garcia, L. R., Molineux, I. J. 1996. Transcription-independent DNA translocation of bacteriophage T7 DNA into *Escherichia coli*. *J Bacteriol* 178:6921-6929

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141-7.

Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340-2361

Glass, L. 1975. Classification of biological networks by their qualitative dynamics. *J. Theor. Biol.* 54:85-107

Glass, L., Kauffman, S. A. 1973. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol* 39:103-29.

Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., et al. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science* 296:92-100

Goryanin, I., Hodgman, T. C., Selkov, E. 1999. Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics* 15:749-58.

Goss, P. J. E., Peccoud, J. 1998. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc. Natl. Acad. Sci. USA* 95:6750-6755

Green, P. 2002. Whole-genome disassembly. *Proc Natl Acad Sci U S A* 99:4143-4.

Hadas, H., Einav, M., Fishov, I., Zaritsky, A. 1997. Bacteriophage T4 development depends on the physiology of its host *Escherichia coli*. *Microbiology* 254:179-185

Haigh, J. 1978. The accumulation of deleterious genes in a population--Muller's Ratchet. *Theor. Popul. Biol.* 14:251-267.

Hartman, J. L. t., Garvik, B., Hartwell, L. 2001. Principles for the buffering of genetic variation. *Science* 291:1001-1004.

Hasty, J., McMillen, D., Isaacs, F., Collins, J. J. 2001. Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet* 2:268-79.

Hedén, C.-G. 1951. Studies of the infection of *E. coli* B with the bacteriophage T2. *Acta Pathologica et Microbiologica Scandinavica. Supplementum LXXXIX*

Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., Watanabe, C. 1993. Identifying proteins from two-dimensional gels by molecular mass searching of

peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 90:5011-5015

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415:180-3.

Hu, J. C. 2001. Model systems: Studying molecular recognition using bacterial n-hybrid systems. *Trends Microbiol* 9:219-22.

Huber, H. E., Tabor, S., Richardson, C. C. 1987. *Escherichia coli* thioredoxin stabilizes complexes of bacteriophage T7 DNA polymerase and primed templates. *J. Biol. Chem.* 262:16224-16232

Hui, A., de Boer, H. A. 1987. Specialized ribosome system: Preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc. Nat. Acad. Sci. USA* 84:4762-4766

Ikeda, R. A., Bailey, P. A. 1992. Inhibition of T7 RNA polymerase by T7 lysozyme in vitro. *J. Bio. Chem.* 267:20153--20158

Iranfar, N., Fuller, D., Sasik, R., Hwa, T., Laub, M., Loomis, W. F. 2001. Expression patterns of cell-type-specific genes in Dictyostelium. *Mol Biol Cell* 12:2590-600.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98:4569-74.

Kacser, H., Burns, J. A. 1973. The control of flux. *Symp Soc Exp Biol* 27:65-104

Karger, B. L., Chu, Y. H., Foret, F. 1995. Capillary electrophoresis of proteins and nucleic acids. *Annu. Rev. Biophys. Biomol. Struct.* 24:579-610

Keightley, P. D., Caballero, A., García-Dorado, A. 1998. Population genetics: Surviving under mutation pressure. *Current Biology* 8:R235-R237

Kierzek, A. M. 2002. STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm. *Bioinformatics* 18:470-81.

Kim, Y. T., Richardson, C. C. 1993. Bacteriophage T7 gene 2.5 protein: an essential protein for DNA replication. *Proc. Natl. Acad. Sci. USA* 90:10173-10177

Kimura, M., Maruyama, T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54:1337-1351

Kohn, K. W. 1999. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* 10:2703-34.

Kondrashov, A. S. 1993. Classification of hypotheses on the advantage of amphimixis. *J. Hered.* 84:372-387

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

Kumar, A., Patel, S. S. 1997. Inhibition of T7 RNA polymerase: transcription initiation and transition from initiation to elongation are inhibited by T7 lysozyme via a ternary complex with RNA polymerase and promoter DNA. *Biochemistry* 36:13954-13962

Kusakabe, T., Baradaran, K., Lee, J., Richardson, C. C. 1998. Roles of the helicase and primase domain of the gene 4 protein of bacteriophage T7 in accessing the primase recognition site. *EMBO J.* 17:1542-1552

Kusakabe, T., Richardson, C. C. 1997. Gene 4 DNA primase of bacteriophage T7 mediates the annealing and extension of ribo-oligonucleotides at primase recognition sites. *J. Biol. Chem.* 272:12446-12453

Kutter, E., Kellenberger, E., Carlson, K., Eddy, S., Neitel, J., et al. 1994. Effects of bacterial growth conditions and physiology on T4 infection. In *Molecular Biology of Bacteriophage T4*, ed. J. D. Karam. pp. 406-418. Washington, DC: ASM Press

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

Laub, M. T., Loomis, W. F. 1998. A molecular network that produces spontaneous oscillations in excitable cells of Dictyostelium. *Mol Biol Cell* 9:3521-32.

Lauffenburger, D. A. 2001. Bioengineering for the Science and Technology of Biological Systems. In *Frontiers of Engineering.* pp. 62-74. Washington, D. C.: National Academy Press

Lee, S. B., Bailey, J. E. 1984. Analysis of growth rate effects on productivity of recombinant *Escherichia coli* populations using molecular mechanism models. *Biotech. Bioeng.* 26:66-73

Lenski, R. E., Ofria, C., Collier, T. C., Adami, C. 1999. Genome complexity, robustness and genetic interactions in digital organisms. *Nature* 400:661-664

Lewin, B. 1997. *Genes*. VI ed. New York: Oxford University Press. 1260 pp.

Liang, S., Fuhrman, S., Somogyi, R. 1998. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pac. Symp. Biocomput.* pp. 18-29

Loew, L. M., Schaff, J. C. 2001. The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol* 19:401-6.

Loomis, W. F. 1998. Role of PKA in the timing of developmental events in Dictyostelium cells. *Microbiol Mol Biol Rev* 62:684-94.

Macdonald, L. E., Zhou, Y., McAllister, W. T. 1993. Termination and slippage by bacteriophage T7 RNA polymerase. *J. Mol. Biol.* 232:1030-1047

Makarova, O. V., Makarov, E. M., Sousa, R., Dreyfus, M. 1995. Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. *Proc Natl Acad Sci U S A* 92:12250-4.

Mann, M., Hendrickson, R. C., Pandey, A. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* 70:437-73

McAdams, H. H., Arkin, A. 1997. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* 94:814-9.

McAdams, H. H., Arkin, A. 1998. Simulation of prokaryotic genetic circuits. *Annu Rev Biophys Biomol Struct* 27:199-224

McAdams, H. H., Shapiro, L. 1995. Circuit simulation of genetic networks. *Science* 269:650-656

McAllister, W. T., Morris, C., Rosenberg, A. H., Studier, F. W. 1981. Utilization of bacteriophage T7 late promoters in recombinant plasmids during infection. *J. Mol. Biol.* 153:527-544

McCarron, R. J., McAllister, W. T. 1978. Effect of ribosomal loading on the structural stability of bacteriophage T7 early messenger RNAs. *Biochem. Biophys. Res. Comm.*

McCraith, S., Holtzman, T., Moss, B., Fields, S. 2000. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A* 97:4879-84.

Mendelman, L. V., Notarnicola, S. M., Richardson, C. C. 1992. Roles of bacteriophage T7 gene 4 proteins in providing primase and helicase functions in vivo. *Proc. Natl. Acad. Sci. USA* 89:10638-10642

Mendes, P. 1993. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci* 9:563-71.

Mendes, P. 1997. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22:361-363

Miller, J. H. 1992. *A Short Course in Bacterial Genetics. A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria* Cold Spring Harbor Laboratory Press

Molineux, I. J. 1999. Bacteriophage T7. In *Encyclopedia of Molecular Biology*, ed. T. E. Creighton. pp. 2495-2507. New York: Wiley

Molineux, I. J. 2001. No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. *Mol Microbiol* 40:1-8.

Morton-Firth, C. J., Bray, D. 1998. Predicting temporal fluctuations in an intracellular signalling pathway. *J Theor Biol* 192:117-28.

Mukai, T. 1969. The genetic structure of natural populations of Drosophila Melanoganster. VII. Synergistic interactions of spontaneous mutant polygenes controlling viability. *Genetics* 61:749-761

Mukai, T., Chigusa, S. I., Mettler, L. E., Crow, J. F. 1972. Mutation rate and dominance of genes affecting viability in Drosophila melanogaster. *Genetics* 72:335-55.

Muller, H. J. 1964. The relation of recombination to mutational advance. *Mutat. Res.* 1:2-9

Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D., Venter, J. C. 2002. On the sequencing and assembly of the human genome. *Proc Natl Acad Sci U S A* 99:4145-6.

Noble, D. 2002. Modeling the Heart--from Genes to Cells to the Whole Organ. *Science* 295:1678-82.

O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250:4007-4021

Oliver, S. 2000. Guilt-by-association goes global. *Nature* 403:601-3.

Peck, J. R., Waxman, D. 2000. Mutation and sex in a competitive world. *Nature* 406:399-404

Peters, A. D., Keightley, P. D. 2000. A test for epistasis among induced mutations in *Caenorhabditis elegans*. *Genetics* 156:1635-1647

Peters, A. D., Lively, C. M. 2000. Epistasis and the maintenance of sex. In *Epistasis and the evolutionary process*, ed. J. B. Wolf, E. D. Brodie, 3rd, M. J. Wade. pp. 99-112. New York: Oxford University Press

Pfennig-Yeh, M. L., Ponta, H., Hirsch-Kauffmann, M., Rahmsdorf, H. J., Herrlich, P., Schweiger, M. 1978. Early T7 gene expression. *Molec. Gen. Genet.* 166:127-140

Phillips, P. C., Otto, S. P., Whitlock, M. C. 2000. Beyond the average. In *Epistasis and the evolutionary process*, ed. J. B. Wolf, E. D. Brodie, 3rd, M. J. Wade. pp. 20-38. New York: Oxford University Press

Phizicky, E. M., Fields, S. 1995. Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* 59:94-123

Ponta, H., Reeve, J. N., Pfennig-Yeh, M., Hirsch-Kauffmann, M., Schweiger, M., Herlich, P. 1977. Productive T7 infection of Escherichia coli F+ cells and anucleate minicells. *Nature* 269:440-2.

Prevelige, P. E., Thomas, D., King, J. 1993. Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophys. J.* 64:824-835

Quick, D. J., Shuler, M. L. 1999. Use of in vitro data for construction of a physiologically based pharmacokinetic model for naphthalene in rats and mice to probe species differences. *Biotechnol Prog* 15:540-55.

Rabkin, S. D., Richardson, C. C. 1990. In vivo analysis of the initiation of bacteriophage T7. *Virology* 174:585-592

Rao, C. V., Arkin, A. P. 2001. Control Motifs for Intracellular Regulatory Networks. *Annu. Rev. Biomed. Eng.* 3:391-419

Reddy, B., Yin, J. 1999. Quantitative intracellular kinetics of HIV type 1. *AIDS Res. Hum. Retroviruses* 15:273-283

Reinitz, J., Vaisnys, J. R. 1990. Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of co-operativity. *J Theor Biol* 145:295-318.

Roeder, G. S., Sadowski, P. D. 1977. Bacteriophage T7 morphogenesis: phage-related particles in cells infected with wild-type and mutant T7 phage. *Virology* 76:263-285

Sadowski, P. D., Kerr, C. 1970. Degradation of *Escherichia coli* B deoxyribonucleic acid after infection with deoxyribonucleic acid-defective amber mutants of bacteriophage T7. *J. Virol.* 6:149-155

Sasaki, N., Izawa, M., Watahiki, M., Ozawa, K., Tanaka, T., et al. 1998. Transcriptional sequencing: A method for DNA sequencing using RNA polymerase. *Proc Natl Acad Sci U S A* 95:3455-60.

Sauro, H. M. 1993. SCAMP: a general-purpose simulator and metabolic control analysis program. *Comput. Appl. Biosci.* 9:441-450

Schaff, J., Fink, C. C., Slepchenko, B., Carson, J. H., Loew, L. M. 1997. A general computational framework for modeling cellular structure and function. *Biophys. J.* 73:1135-1146

Schaff, J., Loew, L. M. 1999. The virtual cell. *Pac Symp Biocomput*:228-39.

Schaff, J. C., Slepchenko, B. M., Loew, L. M. 2000. Physiological modeling with virtual cell framework. *Methods Enzymol* 321:1-23

Schena, M. 1996. Genome analysis with gene expression microarrays. *Bioessays* 18:427-431

Schena, M., Shalon, D., Davis, R. W., Brown, P. O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467-470

Schilling, C. H., Edwards, J. S., Palsson, B. O. 1999. Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol Prog* 15:288-95.

Schilling, C. H., Palsson, B. O. 1998. The underlying pathway structure of biochemical reaction networks. *Proc Natl Acad Sci U S A* 95:4193-8.

Schuster, S., Fell, D. A., Dandekar, T. 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* 18:326-32.

Shea, M. A., Ackers, G. K. 1985. The O_R Control System of Bacteriophage Lambda. A Physical-Chemical Model for Gene Regulation. *J. Mol. Biol.* 181:211-230

Shuler, M. L., Leung, S., Dick, C. C. 1979. A mathematical model for the growth of a single bacterial cell. *Ann. NY Acad. Sci.* 326:35-55

Smolen, P., Baxter, D. A., Byrne, J. H. 2001. Modeling circadian oscillations with interlocking positive and negative feedback loops. *J Neurosci* 21:6644-56.

Soderbom, F., Loomis, W. F. 1998. Cell-cell signaling during Dictyostelium development. *Trends Microbiol* 6:402-6.

Son, M., Watson, R. H., Serwer, P. 1993. The direction and rate of bacteriophage T7 DNA packaging in vitro. *Virology* 196:282-289

Spassky, B., Dobzhansky, T., Anderson, W. W. 1965. Genetics of natural populations. XXXVI. Epistatic interactions of the components of the genetic load in Drosophila pseudoobscura. *Genetics* 52:653-664

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.* 9:3273-3297

Spiro, P. A., Parkinson, J. S., Othmer, H. G. 1997. A model of excitation and adaptation in bacterial chemotaxis. *Proc. Natl. Acad. Sci. USA* 94:7263-7268

Srivastava, R., You, L., Summers, J., Yin, J. 2002. Stochastic versus deterministic modeling of intracellular viral kinetics. *J. Theor. Biol.*:In press

Stephanopoulos, G., Aristidou, A. A., Nielsen, J. 1998. *Metabolic Engineering. Principles and Methodologies* San Diego, CA, USA: Academic Press. 725 pp.

Steven, A. C., Trus, B. L. 1986. The structure of bacteriophage T7. In *Electron microscopy of proteins*, ed. J. R. Harris, R. W. Horne. pp. 1-35. Vol. 5. London: Academic Press

Struthers-Schlinke, J. S., Robins, W. P., Kemp, P., Molineux, I. J. 2000. The internal head protein Gp16 controls DNA ejection from the bacteriophage T7 virion. *J. Mol. Biol.* 301:35-45

Studier, F. W. 1969. The Genetics and Physics of Bacteriophage T7. *Virology* 39:562-574

Studier, F. W. 1972. Bacteriophage T7. *Science* 176:367-376

Studier, F. W., Dunn, J. J. 1983. Organization and expression of bacteriophage T7 DNA. *CSH Quant. Biol.* 47:999-1007

Summers, W. C. 1969. The process of infection with coliphage T7. I. Charaterization of T7 RNA by polyacrylamide gel electrophoretic analysis. *Virology* 39:175-182

Summers, W. C. 1970. The process of infection with coliphage T7 IV. Stability of RNA in bacteriophage-infected cells. *J. Mol. Biol.* 51:671-678

Tabor, S., Huber, H. E., Richardson, C. C. 1987. *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* 262:16212-16223

Thomas, R. 1973. Boolean formalization of genetic control circuits. *J Theor Biol* 42:563-85.

Tomita, M. 2001. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* 19:205-10.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., et al. 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15:72-84.

Uetz, P. 2002. Two-hybrid arrays. *Curr Opin Chem Biol* 6:57-62.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403:623-7.

Varner, J., Ramkrishna, D. 1999. Mathematical models of metabolic pathways. *Curr Opin Biotechnol* 10:146-50.

Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W. 1995. Serial analysis of gene expression. *Science* 270:484-487

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., et al. 2001. The sequence of the human genome. *Science* 291:1304-51.

Vidal, M., Legrain, P. 1999. Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res* 27:919-29.

von Dassow, G., Meir, E., Munro, E. M., Odell, G. M. 2000. The segment polarity network is a robust developmental module. *Nature* 406:188-92.

Waterston, R. H., Lander, E. S., Sulston, J. E. 2002. On the sequencing of the human genome. *Proc Natl Acad Sci U S A* 99:3712-6.

West, S. A., Peters, A. D., Barton, N. H. 1998. Testing for epistasis between deleterious mutations. *Genetics* 149:435-444

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., et al. 2002. The Protein Data Bank: unifying the archive. *Nucleic Acids Res* 30:245-8.

Whitlock, M. C., Bourguet, D. 2000. Factors affecting the genetic load in Drosophila: synergistic epistasis and correlations among fitness components. *Evolution* 54:1654-1660.

Whitlock, M. C., Phillips, P. C., Moore, F. B.-G., Tonsor, S. J. 1995. Multiple fitness peaks and epistasis. *Annu. Rev. Ecol. Syst.* 26:601-629

Wilke, C. O., Adami, C. 2001. Interaction between directional epistasis and average mutational effects. *Proc R Soc Lond B Biol Sci* 268:1469-74.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., Adami, C. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412:331-3.

Winslow, R. L., Scollan, D. F., Holmes, A., Yung, C. K., Zhang, J., Jafri, M. S. 2000. Electrophysiological modeling of cardiac ventricular function: from cell to organ. *Annu Rev Biomed Eng* 2:119-55

Wloch, D. M., Borts, R. H., Korona, R. 2001a. Epistatic interactions of spontaneous mutations in haploid strains of the yeast Sacchromyces cerevisiae. *J. Evol. Biol.* 14:310-316

Wloch, D. M., Szafraniec, K., Borts, R. H., Korona, R. 2001b. Direct Estimate of the Mutation Rate and the Distribution of Fitness Effects in the Yeast Saccharomyces cerevisiae. *Genetics* 159:441-52.

Wolf, J., Brodie, E. D., 3rd, Wade, M. J., eds. 2000. *Epistasis and the evolutionary process*. New York: Oxford University Press. 330 pp.

Yamada, Y., Nakada, D. 1976. Early to late switch in bacteriophage T7 development: no translational discrimination between T7 early messenger RNA and late messenger RNA. *J. Mol. Biol.* 100:35-45

Yamada, Y., Whitaker, P. A., Nakada, D. 1974. Early to late switch in bacteriophage T7 development: functional decay of T7 early messenger RNA. *J. Mol. Biol.* 89:293-303

Yi, T. M., Huang, Y., Simon, M. I., Doyle, J. 2000. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci U S A* 97:4649-53.

Yin, J., McCaskill, J. S. 1992. Replication of viruses in a growing plaque: a reaction-diffusion model. *Biophys. J.* 61:1540-1549

You, L., Suthers, P. F., Yin, J. 2002. Effects of Escherichia coli Physiology on Growth of Phage T7 In Vivo and In Silico. *J Bacteriol* 184:1888-94.

You, L., Yin, J. 1999. Amplification and spread of viruses in a growing plaque. *J Theor Biol* 200:365-73.

You, L., Yin, J. 2000. Patterns of regulation from mRNA and protein time series. *Metab Eng* 2:210-7.

You, L., Yin, J. 2001. Simulating the growth of viruses. *Pac Symp Biocomput*:532-43.

You, L., Yin, J. 2002. Dependence of Epistasis on Environment and Mutation Severity as Revealed by in Silico Mutagenesis of Phage T7. *Genetics* 160:1273-1281

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., et al. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science* 296:79-92

Zaritsky, A., Woldringh, C. L. 1973. Changes in cell size and shape associated with changes in the replication time of the chromosome of *Escherichia coli*. *J. Bacteriol.* 114:824-837

Zhang, X. 1995. T7 RNA polymerase and T7 lysozyme: Genetic, biochemical and structural analysis of their interaction and multiple roles in T7 infection. State University of New York at Stony Brook

Zhang, X., Studier, F. W. 1997. Mechanism of inhibition of bacteriophage T7 RNA polymerase by T7 lysozyme. *J Mol Biol* 269:10-27.