UNIVERSITY OF CALIFORNIA, SAN DIEGO

Systems Analysis of Genomes: Towards a “Topobiology”

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Bioengineering

by

Timothy Eric Allen

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Shankar Subramaniam
Professor Gary A. Huber
Professor Milton H. Saier, Jr.
Professor Philip E. Bourne

2006

The dissertation of Timothy Eric Allen is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2006

To my parents

For setting me upon the right path.

“I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician; he is also a child placed before natural phenomena which impress him like a fairy tale.”

–Marie Curie

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank the National Science Foundation and the National Institutes of Health for the research and training grants that have funded my tenure as a graduate student at UCSD.

I am immensely thankful for my advisor, Bernhard Palsson, whose scientific guidance, professional mentorship, and boundless enthusiasm for progress in bioengineering research and education have provided a tremendously positive example and a source for motivation over the course of my progress throughout graduate school.

While I have benefited from the comraderie and knowledge of all of the graduate students and researchers with whom I have worked in the Palsson Lab, I feel I should especially acknowledge my long-time officemates, Jason Papin and Nathan Price. They have been not only invaluable resources of information and participants in many discussions and scientific debates but have also proven themselves to be great friends and are people of the highest character. I would also like to express my sincere gratitute to Markus Herrgård, whose patience and consistent willingness to share his knowledge and time have provided me with many of the skills and ideas presented in this dissertation.

I am also very grateful for all my friends outside of lab. Daniel Machemer and Ryan Drogo deserve special mention for their unwavering friendship and support and for helping me to maintain balance throughout my days in graduate school.

My scientific journey was set into motion over thirteen years ago when I had the privilege of being taught by Gayle Doran. She first introduced me to the marvels and intricacies of molecular biology, and to this day she remains the finest teacher I have had in any discipline.

Most of all, I thank my parents, Ann Lyn and Gene Allen, for their love and support, and for teaching me those things which are most important. Without them, none of this work would have been remotely possible. I also am thankful for

all of my extended family, and especially for my sister, Martha, whose outstanding work ethic and passion for education have left a lasting positive influence on me.

The text of Chapter Three, in full, is a reprint of the material as it appears in T.E. Allen and B.O. Palsson. 2003. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *Journal of Theoretical Biology*, 220:1-18. I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

The text of Chapter Four, in full, is a reprint of the material as it appears in T.E. Allen, M.J. Herrgård, L. Mingzhu, Y. Qiu, J.D. Glasner, F.R. Blattner, and B.O. Palsson. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *Journal of Bacteriology*, 185:6392-6399. I was the primary author of this publication and the co-authors participated and supervised the research which forms the basis for this chapter.

The text of Chapter Five, in part or in full, is a reprint of the material as it appears in T.E. Allen, N.D. Price, and B.O. Palsson. Sensitivity analysis of translational efficiency with respect to codon usage and tRNA abundance. In preparation. I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

The text of Chapters Seven and Eight, in part or in full, is a reprint of the material as it appears in T.E. Allen, N.D. Price, A.R. Joyce, and B.O. Palsson. 2005. Long-range patterns in prokaryotic genome sequences indicate significant multi-scale chromosomal organization. *PLoS Computational Biology*, (in press). I was the primary author of this publication and the co-authors participated and directed the research which forms the basis for this chapter.

The text of Chapter Eight, in part, is a reprint of the material as it appears in T.E. Allen, A.R. Joyce, and B.O. Palsson. Cross-correlation of spatial patterns in heterogeneous *Escherichia coli* data sets. In preparation. I was the primary author of this publication, and the co-authors participated and supervised the research which forms the basis for this chapter.

VITA

1999	B.S.E., Duke University
2001	M.S., University of California, San Diego
2006	Ph.D., University of California, San Diego

PUBLICATIONS

T.E. Allen and B.O. Palsson. 2003. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *Journal of Theoretical Biology*, 220: 1-18.

T.E. Allen, M.J. Herrgard, L. Mingzhu, Y. Qiu, J.D. Glasner, F.R. Blattner, and B.O. Palsson. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: a model-driven analysis of heterogeneous dataset. *Journal of Bacteriology*, 185: 6392-6399.

C.D. Herring, M. Raffaelle, T.E. Allen, E. Kanin, R. Landick, A.Z. Ansari, and B.O. Palsson. 2005. Immobilization of *Escherichia coli* RNA polymerase and location of binding sites using chromatin immunoprecipitation and microarrays. *Journal of Bacteriology*, 187: 6166-6174.

T.E. Allen, N.D. Price, A.R. Joyce, and B.O. Palsson. 2005. Long-range patterns in prokaryotic genome sequences indicate significant chromosomal organization. *PLoS Computational Biology.* In press.

A. Raghunathan, C. Honisch, C.D. Herring, T. Patel, M. Applebee, M. Mosko, B. Groff, T.E. Allen, E.M. Knight, C.R. Cantor, D. van den Boom, and B.O. Palsson. Genome-wide molecular changes in *Escherichia coli* during adaptive evolution on glycerol. In preparation.

T.E. Allen, N.D. Price, and B.O. Palsson. Sensitivity analysis of translational efficiency with respect to codon usage and tRNA abundance. In preparation.

T.E. Allen, A.R. Joyce, and B.O. Palsson. Cross-correlation of spatial patterns in heterogeneous *Escherichia coli* data sets. In preparation.

ABSTRACT OF THE DISSERTATION

Systems Analysis of Genomes: Towards a "Topobiology"

by

Timothy Eric Allen
Doctor of Philosophy in Bioengineering
University of California, San Diego, 2006

Professor Bernhard Ø. Palsson, Chair

Systems analysis of cell-scale reconstructions has become a staple of modern biological discovery in the era of high-throughput data generation. Reconstructions of bacteria currently assimilate the existing wealth of biochemical knowledge for a given organism, accounting for metabolites and the stoichiometry of their transport and transformation via enzymes. However, the bulk of the material and energy consumed during cellular growth is devoted to the synthesis of the macromolecules involved in information transfer: RNA, protein, and the genome itself. In this dissertation, a framework for incorporating the fundamental processes of bacterial transcription and translation is described. This framework is used to integrate mRNA expression and half-life data towards assessing the global transcription state of the *Escherichia coli* genome under numerous experimental conditions. To address how the translation network would be constrained within this framework, a model is used to determine the theoretical limits in translational efficiency achievable by altering the synonymous codon usage of each gene given measured tRNA abundances. A sensitivity analysis of translational efficiency, which can be varied by an average 6.5-fold, demonstrates that wild-type synonymous codon usage and measured tRNA abundances in *E. coli* are highly synchronized. However, the results from these studies also expose the limitations of network reconstructions that neglect three-dimensional spatial information. The transcription state of a genome can be highly nonrandom with respect to chromosomal position. Furthermore, tRNA diffusion limitations must be taken into account to accurately model translational efficiency. This dissertation thus advances a method for characterizing the spatial organization of chromosome position-dependent data. A comprehensive assessment of the periodic pattern content contained within 163 prokaryotic chromosomes is performed using wavelet analysis. The degree of patterning in sequence-derived properties correlates with genome-size, overall GC-content, and the occurrence of motility and chromosomal-binding proteins. Given additional functional data for *E. coli*, long-range patterns in multiple heterogeneous properties are shown to be highly correlated and are consistent with experimentally detected chromosomal macrodomains. Taken together, the findings reported in this dissertation demonstrate that the field of cell-scale modeling will ultimately enter a phase in which network connectivity is viewed within the context of topobiological, spatial constraints.

Chapter 1

Genome-scale Biology and the Need for Systems Analysis

> It seems inescapable that, at least at the level of molecules and cells, biology is moving from an era of data-collection to one of hypothesis-driven research. Progress in this new field will be driven by informed and increasingly quantitative theories—whatever name we choose to give it.
>
> Dennis Bray [29]

1.1 A brief history of genome-scale biology

The discovery of the structure of DNA in 1953 [328] and of the genetic code in the early 1960s [55, 207, 168] ushered in a golden age in molecular biology. Once researchers knew of the molecular basis for genetics and information transfer, a decades-long push was launched towards reductionism and the determination of all of the molecular components of a cell—including the genetic content itself. Before that final milestone was reached in 1995 [93], however, numerous important discoveries peppered the field of molecular biology in the 1970s and 1980s (particularly recombinant technology and cloning, the cornerstones of genetic engineering) [334]. The invention of DNA sequencers and whole-genome assembly methods then revolutionized the scale at which molecular biology was conducted in the mid-1990s. Since then, more than 200 organisms have been fully sequenced, ranging from *Mycoplasma genitalium* to *Homo sapiens*. This field of high-throughput gene content discovery became known as "genomics." Other "omics" data (some encompassing entire fields, such as proteomics) have since become widely available as a consequence of increasingly sophisticated technologies and widespread consortia and efforts to generate data [218]. These high-throughput data types include measurements of mRNA transcript levels ("transcriptomics," or simply gene expression data), protein levels ("proteomics"), small molecule concentrations ("metabolomics"), protein-protein interactions, transcription factor binding locations ("ChIP-chip" experiments), and phenotypic information ("phenomics"). With the exception of the attempts to measure large numbers of protein-protein interactions and ChIP-chip data, each of these "omics" data types represents the apotheosis of biological reductionism—the delineation of the cellular "parts list" down to the most detailed molecular level.

1.2 The need for systems analysis in biology

> We can call the human genome "the blueprint," the "Holy Grail," all sorts of things—it's a parts list. If I gave you the parts list for the Boeing 777, it's got 100,000 parts on it, but I don't think I could screw it together on the basis of that, and I certainly wouldn't understand why it flew.
>
> Eric Lander[1]

The recent proliferation of high-throughput biological data described in the last section—which includes sequence, gene expression, proteomics, metabolomic, and interaction data—has highlighted a need for systematic methods by which to integrate these data towards the goal of interpreting and predicting phenotypic behavior [218]. The ability to elucidate this genotype-phenotype (i.e. sequence-function) relationship will mark a fundamental step forward in biological under-

[1] 1999 speech given at the *Millennium Evening at the White House: Informatics Meets Genomics* event, `http://www.genome.gov/10001397`

standing at the cellular level. If fully realized, such an advance in understanding will have a profound impact on the diagnosis and treatment of human disease [343], and on the design of microbial organisms to synthesize desired drugs and chemicals [264, 187] and to assist in bioremediation and environmental clean-up [230]. With the ever-growing availability of high-throughput ("omics") data, however, the requirements for model development are changing. The gain of new knowledge from the current wealth of biological data requires the development of *in silico* models that meet the following five criteria:

1. Modern large-scale biological models must be data-driven;
2. New models will be based on large organism-specific databases;
3. They will need to integrate diverse data types (genomic, transcriptomic, proteomic, metabolomic, and phenomic data, to name the major types);
4. Modern models must be easily scalable to cell or genome-scale; and
5. They must account for inherent biological uncertainty.

It is important to note that these post-genomic era models are not expected to be able to compute cell functions with the same precision as we are used to in the disciplines of chemistry, physics, or engineering. These requirements necessitate a paradigm shift in the way large-scale in silico models are constructed [217, 218]. A framework that has been remarkably effective for elucidating the genotype-phenotype relationship in microbial organisms is network reconstruction, or "2-D annotation" [219, 248].

First, however, I will summarize the broad array of modeling approaches that have been applied to biological systems that has been employed over the past two and a half decades (and longer in some cases). Then I will present the basics of network reconstruction, as well as the biological lessons learned from analysis of cell-scale reconstructions of metabolism and transcriptional regulation.

1.3 A broad overview of existing modeling approaches in biology

Due to the wealth of available biochemical and physiological data, metabolism (and bacterial metabolism, in particular) has become the focus of many of the efforts currently underway to elucidate the genotype-phenotype relationship [311, 180, 262, 75, 263, 269, 76, 270, 77, 78, 335, 153, 246, 71]. Metabolic phenotypes can be defined in terms of flux distributions through a metabolic network. Interpreting and predicting metabolic flux distributions requires the application of mathematical modeling and computer simulation, for which a long history exists [15, 334]. Fundamentally different approaches include:

Topological analysis: Extreme pathway analysis (ExPA) [269, 222, 224], elementary flux modes [276], small-world/scale-free networks [329, 90, 154, 155, 17], SVD analysis [234, 86], and developmental modules and segment polarity in *Drosophila* genes [324].

Stoichiometry-based modeling: Stoichiometric network analysis [45]; chemical reaction network theory [87]; optimization-based techniques, including flux balance analysis (FBA) [315, 28], energy balance analysis [19, 233], mixed-integer linear programming [177, 34, 247], and bilevel optimization [35, 36]; and non-optimization-based techniques, including metabolic flux analysis [294].

Kinetics-based modeling: Primary kinetic analysis, including thermodynamic network theory [158], kinetic theory [250, 127], and osmotic and electric physico-chemical consideration; and secondary kinetic analysis, including metabolic control analysis (MCA) [89, 40], combined FBA/MCA [312], combined ExPA/dynamic modeling [142], biochemical systems theory & power law kinetics [265, 266], and cybernetic modeling [317].

Stochastic modeling: Stochastic simulation of the lambda phage [10] and of chemotaxis and signal processing [196], and stochastic kinetics of bacterial transcriptional regulation [123].

Model reduction and simplification techniques: Time-scale separation [249], modal analysis and temporal decomposition [214], kinetic assumptions [192], and modularity [122].

Some of these methods have been expanded to generate large-scale models. Whole-cell kinetic models have been developed, including E-CELL [304] and 3-D representation in V-cell [185], as well as a complete kinetic model of human red blood cell metabolism [152]. Constraint-based genome-scale metabolic models have been reconstructed for several microorganisms (reviewed in [248]). Large-scale models have also been constructed for signaling pathways [274, 223], mitosis [41], apoptosis [100], yeast core metabolism [299], the lambda and T4 phages [189], and for *E. coli* growth [286, 287]. Currently, several well developed mathematical approaches exist for the dynamic analysis of cellular metabolism and its regulation [287, 179, 215, 89, 18, 15, 208, 317, 318]. Most of these methods require detailed kinetic and concentration information about enzymes and various cofactors. Even though biological information is growing rapidly, we still do not have enough information to describe cellular metabolism in mathematical detail for a single cell [16]. The human red blood cell remains the lone notable exception [286, 287, 134, 275, 159, 238, 176, 197, 152].

1.4 Network reconstruction in a nutshell

The completion of the *H. influenzae* genome sequence in 1995 [93] marked a significant phase transition in the study of biology—and also in biological modeling, bringing into sharp relief the requirements for modern biological models enumerated in §1.2. The growing number of reliable high-throughput technologies in this nascent post-genome era has transformed biological research from that of a data-poor discipline into a (relatively) data-rich one. The fields of bioinformatics and theoretical biology are now moving to the forefront of biological discovery as scientists attempt to generate new knowledge from the large amount of information now readily available, through automated genome annotation, metabolic network reconstruction, protein structure determination and, more recently, regulatory network reconstruction from gene expression data. The reconstructions of metabolic networks that have appeared in recent years constitute rigorously-curated, chemically accurate databases of all biochemical transformations and transport reactions known to occur in a living cell [248]. When subjected to rigorous quantitative analysis methods, these reconstructions have proven to be powerful tools for driving biological discovery, as demonstrated by experimental proofs of principle, predictions of phenotype following adaptive evolution, and iterative network elucidation [77, 143, 54]. For a more in-depth review of metabolic network reconstruction, refer to [248].

1.5 Preview of the dissertation

The topic of this dissertation, broadly stated, covers the systems analysis of network reconstructions that have been fundamentally expanded to include the genome as a molecule and the information transfer processes as biochemical interactions and transformations. Prior to the work presented herein, no such analysis existed at the genome-scale. The overarching scientific questions are thus: How can we account for information transfer within network reconstructions in bacteria? How can this be used to integrate multiple heterogeneous data sets, and what are the results from such an integration?

The chapters in this dissertation thus deal with the following topics:

Chapter 1: This chapter describes the motivation for systems analysis in biology and introduces the concept of network reconstruction and analysis as a means to address the growing volumes of data in biology.

Chapter 2: Here I will provide a brief primer on the information transfer processes in bacterial cells, focusing on *Escherichia coli* as a model organism. I will describe the basic macromolecular synthesis reactions in *E. coli*, and I will show some order-of-magnitude analyses of these processes.

Chapter 3: The theoretical framework will be presented for expanding the scope of network reconstructions to include the reactions involved in the synthesis of RNA and protein in bacteria. The calculation of metabolic costs from the genome sequence will be discussed, as well.

Chapter 4: Given the framework laid out in the previous chapter, I will then describe the integration of heterogeneous data sets to analyze genome usage in *E. coli*. This chapter will present the first results which hint at the need for including spatial information in network reconstructions of macromolecular interactions.

Chapter 5: In this chapter I present a sensitivity analysis of translational efficiency in *E. coli* with respect to synonymous codon usage and experimentally measured tRNA abundances. This will wrap up the systems analyses of transcription and translation from a two-dimensional network perspective.

Chapter 6: This chapter acts as a bridge between the conceptual work presented in the previous chapters and the data analysis presented in the next two chapters. Here I first explicitly put forth the notion of a “3-D annotation” in biology, and I provide a broad review of the state of the art in our knowledge of the bacterial cell interior.

Chapter 7: Here I present an analysis of spatial patterns in most sequenced microbial genomes to-date. These results provide strong evidence for selection pressure for specific genome arrangements and orders, conclusively showing that a genome is much more than just a “parts list,” but a highly-structured map to the cell.

Chapter 8: This chapter extends the analysis of the previous chapter by delving more deeply into the pattern content of numerous genome position-dependent data sets in *E. coli.* The results presented here further argue on behalf of adding a spatial dimension to network reconstructions.

Chapter 9: In this concluding chapter, I summarize what was learned from the work presented in this dissertation, and I provide my thoughts on how this work fits in the larger context of the field and where future work is likely to lead us.

The material presented in this dissertation provides a fundamental conceptual advance in both the scope of network reconstructions and in the sorts of constraints that will eventually be necessary in any truly comprehensive cell-scale model. Let us now begin our discussion with a brief review of the processes of transcription and translation in bacteria.

Chapter 2

Information Transfer in Bacteria

A significant portion of this dissertation deals with the processes of transcription and translation—the information transfer processes—in bacteria, with particular emphasis upon their relation to cell-scale reconstructions. In order to best understand the next few chapters, however, a brief primer on information transfer in bacteria is needed. In this chapter, I will explain why the majority of the studies presented in this dissertation focus on the model organism, *Escherichia coli*, and I will describe the basic specifications and order-of-magnitude estimations of rates with respect to macromolecular synthesis processes in this well-studied organism.

2.1 *Escherichia coli* as a model organism

The enteric bacterium *E. coli* is one of the best-understood organisms in all of biology. Study of this organism's physiology, genetics, pathogenicity, and biochemistry has been ongoing for decades [201]. The chromosome of *E. coli* K-12 was sequenced in 1997 [26], and high-throughput gene expression data are currently being amassed [297, 331, 108, 4]. Due to the existence of a vast body of literature on *E. coli*, the current influx of phenotypic data for both wild-type *E. coli* K-12 and adaptively-evolved strains [77, 143, 94, 54, 95, 96], and the recently generated chromosome rearrangement and mutation data for these evolved strains [239, 133], this organism represents an ideal candidate for the systems analysis of a genome and its effect on protein synthesis *in silico*, towards the ultimate goal of elucidating the genotype to phenotype relationship [268].

In addition to the experimental data that have accumulated over the years for *E. coli*, this organism has a long history of meticulously generated high-quality genome-scale metabolic reconstructions [245]. Furthermore, a genome-scale set of Boolean regulatory rules [301] has been developed and experimentally verified for *E. coli* [54]. This detailed and comprehensive knowledge base for *E. coli* makes it emminently suitable for the explicit incorporation of the macromolecular synthesis reactions involved in information transfer.

2.2 The players in bacterial information transfer

The information transfer processes are essential to the bacterial life cycle, as they constitute the means by which the genetic material is copied and used in order to specify the protein machinery available to the cell. These processes constitute a large majority of the material and energy cost of cellular growth (∼75% [200]). The three processes that fall under the heading of information transfer are DNA replication[1], transcription, and translation. Table 2.1 lists the major players involved in information transfer in *E. coli*. (For more information on the spatial characteristics of many of these components, refer to Table 6.1 in Chapter 6.) The next two subsections briefly summarize the specifications involved in transcription and translation. Simplified schematic representations of both of these processes are provided in Chapter 3.

2.2.1 Transcription

Transcription is the process by which RNA is synthesized from a DNA template. RNA polymerase (RNAP) is the enzyme responsible for catalyzing the

[1] Replication falls outside the scope of this dissertation, so while I will not discuss it at length here, it is worth noting that DNA in *E. coli* is replicated at a rate of 800 nucleotides per second, with an error rate of only about one mistake every 10^{10} nucleotides [200].

Table 2.1: Components involved in information transfer in *E. coli.* Values reflect mid log-phase growth on rich media at 37°.

Component	Types	# per cell	Description	Reference
DNA	1	2.1	Genetic content of the organism	[200]
Replication proteins	7	1,615	Unwinding, replication, proofreading, supercoiling, nick-sealing, primer synthesis, etc.	[200]
RNA polymerase	1	2,500–3,000	Synthesizes RNA from DNA template	[150]
Sigma factors	7	700	Regulator involved in transcription initiation	[150]
mRNA	1,500–2,000	4,000	Template for protein synthesis	[200, 296]
NTP monomers	4	500,000–3,000,000	Building blocks for RNA and energy for protein synthesis (and dNTPs for DNA replication)	[296]
Amino acids	21	6,000,000	Building blocks for proteins	[296]
tRNA	46	200,000	Molecular "shuttles" involved in protein synthesis	[31, 68]
rRNA	3	18,700	Molecular machines for protein synthesis	[200]
Elongation factors	2	N/A	Involved in protein synthesis	[200, 193]

synthesis of RNA. There are three kinds of RNA: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). The mRNA transcripts are synthesized from genes known as coding sequences, since the information contained in the sequence is used to specify a particular primary sequence of amino acids according to the triplet base-pair genetic code [188]. Different mRNA transcripts are found in widely varying concentrations in *E. coli*, spanning four orders of magnitude [200, 4]. The other RNA molecules (tRNA and rRNA) are involved in the use of mRNA transcripts in synthesizing proteins, as described in the next subsection. A schematic of the basic lumped reactions taking place at each step in transcription is provided in Figure 2.1. In summary, RNAP and a sigma factor—an accessory protein usually involved in regulation and in recognizing specific promoters—binds to a promoter site on the DNA to form the initial ("closed") binary complex. If the "strength" of the given promoter is sufficiently high, the polymerase then opens up the double-stranded DNA to more easily access the strand containing the gene of interest. This complex is called the final ("open") binary complex. Typically, once the complex reaches this open form, the polymerization proceeds essentially irreversibly, although occasionally abortive transcripts result and the polymerase falls off the DNA. Soon after transcription elongation begins, the sigma factor dissociates from the complex, and the RNAP proceeds along the DNA, freeing the promoter site [244, 325] (Figure 2.1).

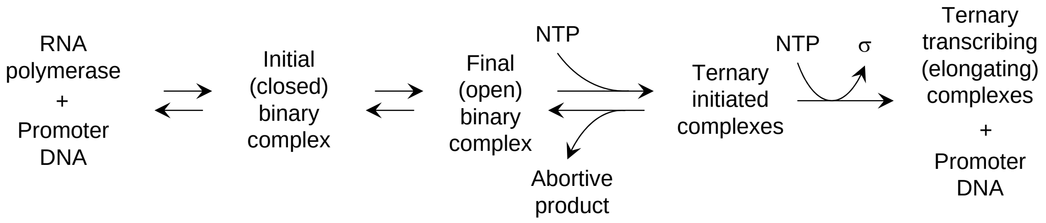

RNA polymerase + Promoter DNA
Initial (closed) binary complex
Final (open) binary complex
NTP
Abortive product
Ternary initiated complexes
NTP
σ
Ternary transcribing (elongating) complexes + Promoter DNA

Figure 2.1: Bacterial transcription reactions.

2.2.2 Translation

Translation is the process by which proteins are synthesized according to mRNA transcripts (“messages”). Many molecular machines and shuttles ($>$ 50 proteins, rRNA, and tRNA) as well as building blocks and energy molecules are essential for this process [188, 193]. The tRNA molecules act as shuttles that carry amino acids to the appropriate position in the ribosome, as specified by the mRNA transcript. This process requires two elongation factors (EF-Tu and EF-G) which are involved in binding of the aminoacyl-tRNA shuttles to the ribosome and translocation of the growing polypeptide chain to different ribosomal sites [200, 188]. Energy in the form of GTP is consumed during this process. The tRNA shuttles are then “recharged” with the appropriate amino acids (which requires ATP).

As summarized in Table 2.1, *E. coli* contains 46 tRNA species encoding 21 amino acids [170, 68] (also refer to Table 2.2). One of the tRNAs encodes the rare amino acid selenocysteine, encoded by the codon UGA which is usually a stop codon. There are two tRNA species that read the start codon, AUG, and incorporate *N*-formylmethionine as the first amino acid in almost every protein that is translated [200]. When AUG occurs in the middle of a transcript, a third cognate tRNA incorporates methionine. Overall, there are ∼200,000 tRNA molecules in a typical *E. coli* cell, ranging in number from ∼300–15,000 copies per cell, depending on the species [200, 68, 31]. The recognition of codon and tRNA anticodon is not necessarily unique, due to the Crick “wobble” hypothesis [56]. Thus, some codons can be recognized by as many as three tRNA species, and some tRNA molecules can bind to as many as three different codons [68]. What is remarkable is the fact that the nuclear region of *E. coli* is about 0.3 μm^3 and thus the spatial arrangement of all these processes is very intricate [48]. Also note that the process of translation is physically coupled with transcription in bacterial cells (whereas in eukaryotes, transcription occurs in the nucleus and translation in the cytoplasm). For more on the spatial organization of the cell interior, refer to Chapter 6.

Table 2.2: List of tRNA species, abundances, and cognate codons for *E. coli*. The data were compiled from [68].

	Anticodon	**Codon recognition**			**% of total**
tRNA	**(5′-3′)**	**1**	**2**	**3**	**tRNA**
Ala1B	UGC	GCU	GCA	GCG	5.06
Ala2	GGC	GCC			0.96
Arg2	ACG	CGU	CGC	CGA	7.39
Arg3	CCG	CGG			0.99
Arg4	UCU	AGA			1.35
Arg5	CCU	AGG			0.65
Asn	GUU	AAC	AAU		1.86
Asp1	GUC	GAC	GAU		3.73
Cys	GCA	UGC	UGU		2.47
Gln1	UUG	CAA			1.19
Gln2	CUG	CAG			1.37
Glu2	UUC	GAA	GAG		7.34
Gly1	CCC	GGG			1.33
Gly2	UCC	GGA	GGG		1.99
Gly3	GCC	GGC	GGU		6.78
His	GUG	CAC	CAU		0.99
Ile1	GAU	AUC	AUU		5.13
Ile2	CAU	AUA			0.27
Leu1	CAG	CUG			6.95
Leu2	GAG	CUC	CUU		1.47
Leu3	UAG	CUA	CUG		1.04
Leu4	CAA	UUG			2.98
Leu5	UAA	UUA	UUG		1.60
Lys	UUU	AAA	AAG		2.99
Met f1	CAU	AUG			1.88
Met f2	CAU	AUG			1.11
Met m	CAU	AUG			1.10
Phe	GAA	UUC	UUU		1.61
Pro1	CGG	CCG			1.40
Pro2	GGG	CCC	CCU		1.12
Pro3	UGG	CCA	CCU	CCG	0.90
Sel-Cys	UCA	UGA			0.34

Table 2.2, continued.

	Anticodon	**Codon recognition**			**% of total**
tRNA	**(5′-3′)**	**1**	**2**	**3**	**tRNA**
Ser1	UGA	UCA	UCU	UCG	2.02
Ser2	CGA	UCG			0.54
Ser3	GCU	AGC	AGU		2.19
Ser5	GGA	UCC	UCU		1.19
Thr1	GGU	ACC	ACU		0.16
Thr2	CGU	ACG			0.84
Thr3	GGU	ACC	ACU		1.70
Thr4	UGU	ACA	ACU	ACG	1.43
Trp	CCA	UGG			1.47
Tyr1	GUA	UAC	UAU		1.20
Tyr2	GUA	UAC	UAU		1.96
Val1	UAG	GUA	GUG	GUU	5.97
Val2A	GAC	GUC	GUU		0.98
Val2B	GAC	GUC	GUU		0.99

2.3 Macromolecular synthesis rates

The following specifications of *E. coli* are pertinent to any steady-state model of macromolecular synthesis. Typical RNA elongation rates during transcription range from 50–100 nucleotides per RNA polymerase (RNAP) per second, depending on the gene, the regulation present, and the growth rate [113, 323]. If we assume that there are approximately 2,500–3,000 RNAP molecules per cell [150], then a typical cell is capable of incorporating 125,000–300,000 nucleotides per second into RNA molecules. An average gene length of ∼1000 base pairs [26] implies that 125–300 mRNA molecules can be made by a single cell every second, assuming no rRNA and tRNA is being made. In reality, this value will be $< 5\%$ of this rate, such that ∼6–15 mRNA molecules will be transcribed per second. With respect to translation, each cell typically contains ∼18,000 ribosomes [200], each of which catalyzes peptide bond formation at the rate of 16 bonds per second [325]. Since each amino acid that is incorporated into a growing peptide chain is encoded by a triplet codon on the mRNA transcript, this rate corresponds to 48 nucleotides per second. (Interestingly, the rate of translation is comparable to the mRNA transcription rate, which occurs at ~50 nucleotides per second. The consistency in these measured rates is reassuring in light of the fact that the processes of transcription and translation are physically coupled in prokaryotes.) The average open reading frame (ORF) in *E. coli* encodes a protein that is 317 amino acids long [26]. If we assume the average protein in a given cell is 317 amino acids, a typical cell can produce a maximum of ~950 protein molecules per second [3]. These rates are summarized in Table 2.3.

Table 2.3: Macromolecular synthesis rates in *E. coli*. Values reflect mid log-phase growth on rich media at 37°.

Component	**Synthesis rate**	**Reference**
DNA	800 nt/sec	[200]
RNA	50–100 nt/sec/RNAP	[200, 31]
	$\hookrightarrow$ 125,000–300,000 nt/sec/cell	[3]
mRNA	6–15 mRNA/sec/cell	[3]
Protein	16 aa/sec/ribosome	[325]
	$\hookrightarrow$ 950 proteins/sec/cell	[3]

2.4 Control of information transfer

The information transfer processes are highly regulated and will vary considerably depending upon the environmental conditions in which the cell resides. Accordingly, the study of transcription regulation in bacteria comprises an extensive field [325], most of which falls beyond the scope of this dissertation. To summarize, there are numerous (> 300) transcription factors [228] which usually exhibit specific rules of operation that can be modeled in a Boolean fashion in bacteria [54]. These factors act as facilitator molecules which enhance the strength of specific promoters under particular sets of conditions [325]. For a much more in-depth review of transcription regulation in bacteria, consult one of the following texts: [201, 325, 237].

In exponentially-growing bacteria, the translational machinery (i.e. ribosomes and ribosomal proteins) increase in concentration exponentially with linear increases in growth rate [31]. There has been significant debate over the years as to precisely how the synthesis of rRNA and ribosomal proteins is regulated. Relatively recent studies have demonstrated that rRNA regulation is dependent upon the concentration of the initiating nucleotide triphosphate (either ATP or GTP) [101]. The synthesis of all the other translational machinery and ribosomal proteins are then ramped up or down to match the concentration of rRNA in the cell [325]. In this way, the cell has a means by which it can adjust its macromolecular synthesis—and hence its growth rate—according to the energy state of the cell. This energy state will depend upon the environment and, of course, upon the genes that are being actively transcribed. Thus, the regulation of the protein synthesis machinery constitutes somewhat of a “chicken-and-egg” problem that is still an active area of research [64].

2.5 Conclusions: Implications for network reconstruction

As mentioned in the previous chapter, genome-scale constraint-based models of metabolism have been reconstructed and used to predict metabolic phenotypic behaviors in a number of microbial organisms (see [236] and [248] for in-depth reviews of reconstructions and associated methods). However, prior to the work described in this dissertation, the chemical reactions whereby macromolecules—particularly RNA and proteins—are polymerized had not yet been explicitly defined in these models. With the addition of these protein synthesis-associated reactions, the stoichiometric constraints on genome-scale cellular networks will be more complete. Furthermore, macromolecular synthesis comprises the dominant energy and material cost in exponentially-growing bacterial cells [200, 31], and thus the integration of protein synthesis with existing constraint-based models constitutes a significant advance. These reactions—which include the information transfer processes of transcription and translation—are subject to mass balance, thermodynamic, and capacity constraints, as well as spatial diffusion limitations.

In accordance with meeting these needs, I have developed a framework by which to incorporate the information transfer reactions within constraint-based models of metabolism and regulation (Chapter 3; [3]). I will then describe the application of this framework towards the integration of heterogeneous high-throughput data types in *E. coli* (Chapter 4; [4]). In Chapter 5, I present an analysis of codon usage and tRNA abundances within the context of the stoichiometric reconstruction and flux-balance model of translation for *E. coli*.

Chapter 3

Expanding the Scope of Bacterial Reconstructions: Transcription and Translation

The large number of genome sequences completed in recent years has underscored the need to develop genome-scale models that can be used to elucidate phenotypic behavior from the genotype [73, 269]. The available annotated sequences, along with known organism-specific biochemical and physiological data, have been implemented in the reconstruction of genome-scale models of metabolism [163, 280, 209, 213, 49].

Kinetic models are very difficult to construct on a genome-scale due to the sheer number of parameters required [16]. A constraints-based approach can be used to successfully circumvent this problem under certain conditions. Such an approach relies upon the fact that metabolic networks are constrained by physicochemical laws which limit what phenotypes the cell is capable of attaining [217]. Thus, rather than calculating a unique phenotypic solution, one can determine the closed solution space within which the steady-state solution must lie, thereby defining the metabolic capabilities of the cellular network. Linear programming [44] can then be used to determine the solution within this space that optimizes a specified cellular objective. This approach, called flux balance analysis (FBA) [315, 28, 74, 110], has been successfully applied to genome-scale metabolic models of *Haemophilus influenzae* [75], *Escherichia coli* [76, 245], *Helicobacter pylori* [273], and *Saccharomyces cerevisiae* [97, 71].

Existing constraints-based genome-scale metabolic networks do not include sequence-based macromolecular polymerization reactions—namely, RNA and protein synthesis—except lumped as monomeric amino acid and nucleotide triphosphate demands for cellular growth [314]. These monomeric demands are determined from the cellular biomass constituents [200] and are thus independent of genome sequence. There is consequently a need to develop a constraints-based formalism for RNA and protein synthesis. Furthermore, this formalism needs to be readily scalable to the genome-scale.

General models of protein synthesis have included non-sequence dependent models within genome-scale metabolic networks [304] and mechanistically detailed kinetic, but not genome-scale, models [227, 70]. Detailed kinetic models have been developed for individual genes and operons and the proteins for which they encode, including the *lac* operon [339] and the *trp* operon [290, 261] in *E. coli*.

A sequence-based genome-scale model of protein synthesis, however, has not been developed, and currently no framework has been established for such a large-scale incorporation of protein synthesis to the current models. This paper describes a fundamental reaction scheme for protein synthesis that provides such a scalable framework. We analyze this basic network using flux balance and extreme pathway analyses, and we identify the parameters that govern both gene expression and protein synthesis.

3.1 Methods for including transcription and translation in steady-state models

3.1.1 Fundamental reaction scheme for protein synthesis

In order to develop a scalable framework within which to describe protein synthesis, it is necessary to identify the fundamental reactions that comprise an "idealized protein production scheme" (Figure 3.1). These reactions will comprise an "elemental system" for a particular gene and its protein, having conceptual systemic boundaries across which the building blocks and energy metabolites for polynucleotide and protein polymerization will be exchanged.

For a given gene, G, and the protein for which it encodes, we can write such a fundamental set of reactions (Table 3.1). This fundamental reaction set contains one gene encoding for one protein, and one type each of nucleotide, amino acid, and transfer RNA (tRNA), and is illustrated in Figure 3.1. The first six fluxes in Table 3.1 correspond to fluxes internal to the system, and the last nine correspond to exchange fluxes. A summary of all abbreviations and symbols used is provided at the end of the manuscript. The reaction set is as follows:

- **Transcription initiation:** The reaction corresponding to the flux, v_1, describes the binding of RNA polymerase (RNAP) to the promoter of G to form the open-promoter complex, G^*. This reaction is usually referred to as transcription initiation. We assume that the forward reaction from the closed RNAP-promoter complex to the open complex (G^*) is much faster than the reverse reaction [244]; thus, v_1 is essentially irreversible.

- **Transcription elongation:** The RNAP then proceeds along the gene during elongation, incorporating nucleotide triphosphates (NTPs) in a series of polymerization reactions represented by v_2. For every NTP added, a pyrophosphate ($\mathrm{PP_i}$) will be released. The liberated $\mathrm{PP_i}$ will immediately be hydrolyzed by pyrophosphatase into two inorganic phosphates ($\mathrm{P_i}$) to drive

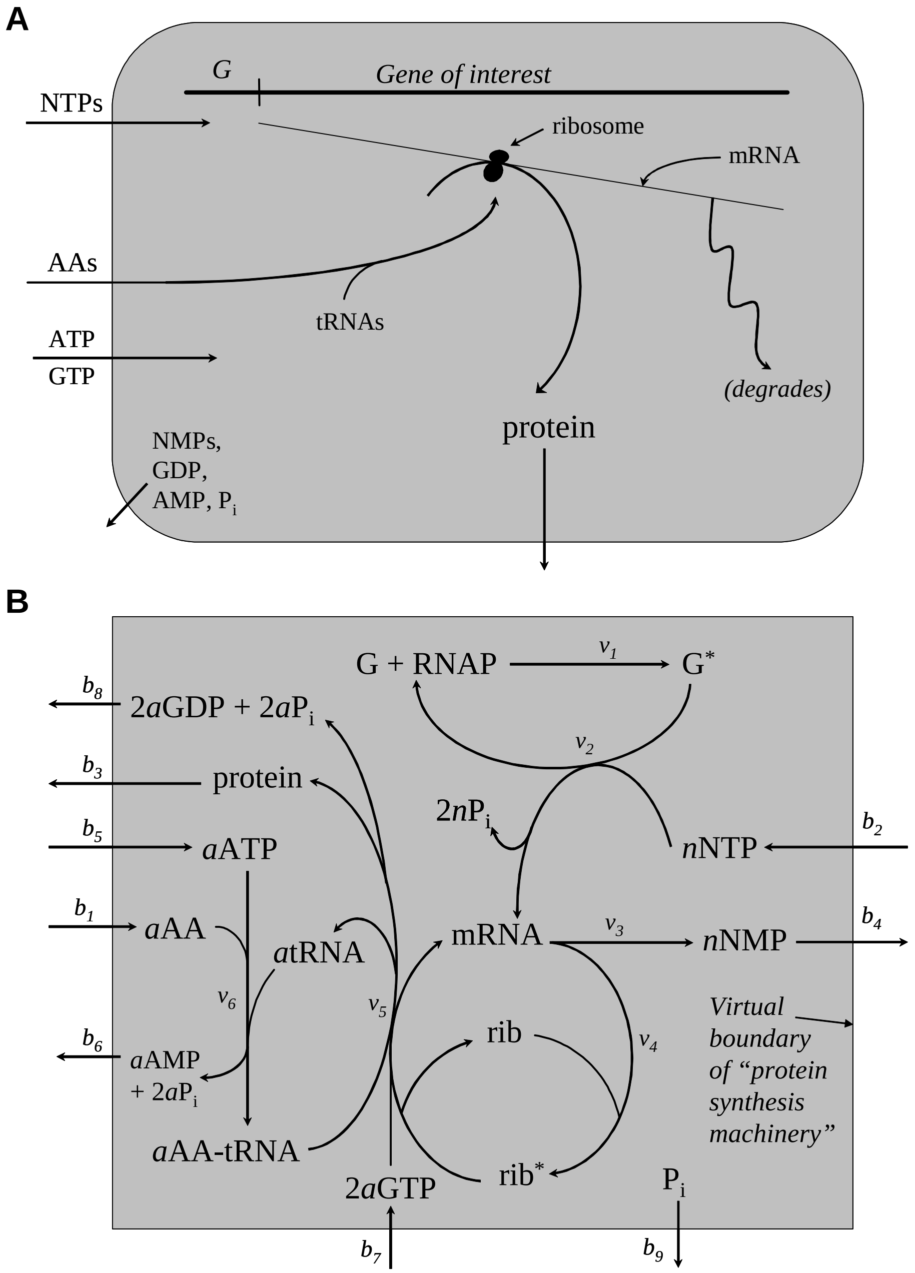


Figure 3.1: The fundamental reaction scheme for protein synthesis. Panel A provides a simplified schematic for the synthesis of a protein encoded by a generic gene. Panel B gives the complete fundamental reaction network based upon the individual reactions listed in Table 3.1 and discussed in the text.

Table 3.1: Simplified, fundamental reaction set for protein production. The first six reactions, which are discussed in the text, occur within the virtual systemic boundary within which the machinery for protein synthesis resides. The last nine reactions correspond to the exchange of building blocks (i.e., AAs and NTPs), protein, by-products (e.g., NMPs), and energy molecules (i.e., ATP and GTP) across the systemic boundary.

Transcription initiation:	$G + \mathrm{RNAP}$	$\xrightarrow{v_1}$	G^*
Transcription:	$G^* + n\mathrm{NTP}$	$\xrightarrow{v_2}$	$\mathrm{mRNA} + G + \mathrm{RNAP} + 2n\mathrm{P_i}$
mRNA decay:	mRNA	$\xrightarrow{v_3}$	$n\mathrm{NMP}$
Translation initiation:	$\mathrm{mRNA} + \mathrm{rib}$	$\xrightarrow{v_4}$	rib^*
Translation:	$\mathrm{rib}^* + a\mathrm{AAtRNA} + 2a\mathrm{GTP}$	$\xrightarrow{v_5}$	$a\mathrm{tRNA} + 2a\mathrm{GDP} + 2a\mathrm{P_i}$
			$+\mathrm{rib} + \mathrm{mRNA} + \mathrm{protein}$
tRNA charging:	$\mathrm{AA} + \mathrm{tRNA} + \mathrm{ATP}$	$\xrightarrow{v_6}$	$\mathrm{AMP} + 2\mathrm{P_i} + \mathrm{AAtRNA}$
Exchange fluxes:	$\mathrm{AA_{ext}}$	$\xrightarrow{b_1}$	AA
	$\mathrm{NTP_{ext}}$	$\xrightarrow{b_2}$	NTP
	$\mathrm{protein}$	$\xrightarrow{b_3}$	$\mathrm{protein_{ext}}$
	NMP	$\xrightarrow{b_4}$	$\mathrm{NMP_{ext}}$
	$\mathrm{ATP_{ext}}$	$\xrightarrow{b_5}$	ATP
	AMP	$\xrightarrow{b_6}$	$\mathrm{AMP_{ext}}$
	$\mathrm{GTP_{ext}}$	$\xrightarrow{b_7}$	GTP
	GDP	$\xrightarrow{b_8}$	$\mathrm{GDP_{ext}}$
	$\mathrm{P_i}$	$\xrightarrow{b_9}$	$\mathrm{P_{iext}}$

the reaction (v_2).

- **mRNA degradation:** The messenger RNA (mRNA) that is produced by v_2 will subsequently be degraded into its nucleotide monophosphate (NMP) constituents (v_3).

- **Translation initiation:** The mRNA will also bind to free ribosomes (translation initiation) to form the active ribosomal complex, rib* (v_4).

- **Translation elongation:** The polymerization reactions that incorporate amino acids (AAs) in the synthesis of the complete protein are lumped into v_5. Two GTPs are required per AA incorporated: one in the binding of the charged transfer RNA (AA-tRNA) to the ribosomal A-site, and the other in the translocation of the amino acid (with the rest of the nascent polypeptide) from the A- to the P-site [188].

- **tRNA charging:** In order to recharge the tRNAs, each AA binds ATP to form aminoacyl-AMP and $\mathrm{PP_i}$. The aminoacyl-AMP then reacts with a tRNA to produce the AA-tRNA and an AMP. These two reactions, driven by the hydrolysis of $\mathrm{PP_i}$ to $\mathrm{2P_i}$ by pyrophosphatase, are represented by v_6.

The AA and NTP inputs represent the building blocks (b_1, b_2), and the ATP and GTP inputs (b_5, b_7) represent the energy cost for the production of protein; the NMP, AMP, GDP, and $\mathrm{P_i}$ outputs represent the by-products (b_4, b_6, b_8, b_9); and the protein output (b_3) is simply the production rate of the protein.

This fundamental reaction set applies for any gene, G, regardless of the number of nucleotides or amino acids incorporated. For those prokaryotic genes which are present in operons, the entire operon is transcribed into a single mRNA. This polycistronic mRNA is then involved in separate reactions for each of the proteins for which it encodes. Thus, a set of reactions $v_{1,2,3}$ is written for every mRNA being produced, and reactions $v_{4,5}$ are written for every protein being synthesized.

3.1.2 The production of components of the protein synthesis “machinery”

The internal (to the synthesis system defined) production of ribosomes, transfer RNA, and RNA polymerase to the fundamental reaction scheme discussed above will add the reactions listed in Table 3.2 to the fundamental reaction set (Figure 3.2). For the untranslated RNA transcripts (i.e., tRNA and ribosomal RNA), production fluxes analogous to v_1 and v_2 are written. A degradation flux is not added for the untranslated transcripts since they are many times more stable than mRNA transcripts and thus unlikely to degrade within the time scale of cellular growth [203]. The synthesis of RNAP is analogous to that of the generic protein in the fundamental system, except that no exchange flux is included for RNAP since it remains internal to the system (Figure 3.2).

Table 3.2: Reactions added to the fundamental system when including the internal production of RNAP, tRNA, and rRNA. The subscript P denotes RNAP, the subscript t denotes tRNA, and the subscript r denotes rRNA. For example, G_t refers to the gene encoding for tRNA, and rib_P^* refers to the actively translating ribosomal complex for the synthesis of RNAP.

RNAP:	$G_P + \text{RNAP}$	$\xrightarrow{v_{1P}}$	G_P^*
	$G_P^* + n_P\text{NTP}$	$\xrightarrow{v_{2P}}$	$\text{mRNA}_P + G_P + \text{RNAP} + 2n_P\text{P}_\text{i}$
	mRNA_P	$\xrightarrow{v_{3P}}$	$n_P\text{NMP}$
	$\text{mRNA}_P + \text{rib}$	$\xrightarrow{v_{4P}}$	rib_P^*
	$\text{rib}_P^* + a_P\text{AAtRNA} + 2a_P\text{GTP}$	$\xrightarrow{v_{5P}}$	$a_P\text{tRNA} + 2a_P\text{GDP} + 2a_P\text{P}_\text{i}$
			$+\text{rib} + \text{mRNA}_P + \text{RNAP}$
tRNA:	$G_t + \text{RNAP}$	$\xrightarrow{v_{1t}}$	G_t^*
	$G_t^* + n_t\text{NTP}$	$\xrightarrow{v_{2t}}$	$\text{tRNA} + G_t + \text{RNAP} + 2n_t\text{P}_\text{i}$
rRNA:	$G_r + \text{RNAP}$	$\xrightarrow{v_{1r}}$	G_r^*
	$G_r^* + n_r\text{NTP}$	$\xrightarrow{v_{2r}}$	$\text{rib} + G_r + \text{RNAP} + 2n_r\text{P}_\text{i}$

3.1.3 Flux balance analysis (FBA)

FBA has been reviewed in detail previously [315, 28, 74, 110]. In short, a mass balance can be described for a system (e.g., the components of a cell’s protein

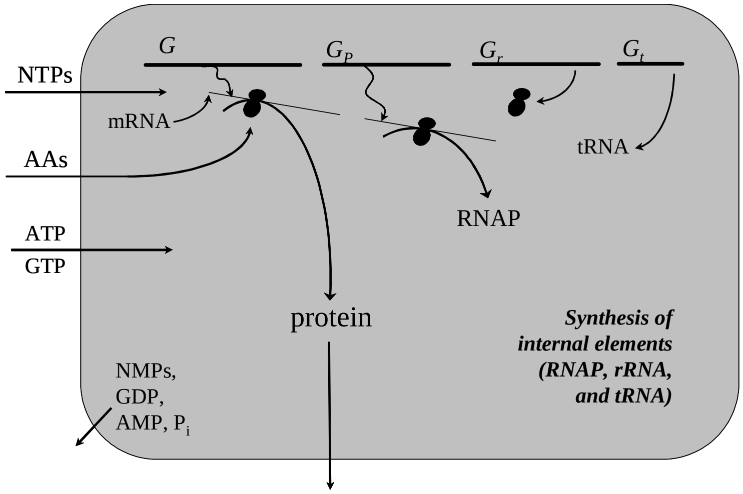


Figure 3.2: Addition of internal accessory elements to the fundamental protein synthesis scheme depicted in Figure 3.1. The basic reaction scheme is provided, in addition to the internal production of the protein synthesis machinery (RNAP, tRNA, and ribosomes).

production machinery), which in a steady state can be written as

$$\mathbf{Sv} = \mathbf{0}, \tag{3.1}$$

where $\mathbf{S}$ is the $m \times n$ stoichiometric matrix (having m metabolites and n reaction fluxes; e.g., see Table 3.3) and $\mathbf{v}$ is an $n \times 1$ vector containing the values of the fluxes through the reactions involved in the system. These fluxes will be subject to thermodynamic and capacity constraints (e.g., v_{max}'s of promoter bindings, maximum elongation rates, etc., as discussed below), described in general by inequality constraints of the form

$$\alpha_i \leq v_i \leq \beta_i, \tag{3.2}$$

where α_i and β_i represent the lower and upper bounds constraining each flux. Linear programming can be used to maximize protein production (i.e., b_3), given the stated constraints. Protein production is chosen as the objective in this study to determine, for a given set of environmental conditions and resources, how much the protein synthesis machinery within the cell can produce. Optimal flux distributions in this study were identified using a commercially available linear programming package (LINDO, Lindo Systems, Chicago), subject to the constraints given in

Equations 3.1 and 3.2.

3.1.4 Extreme pathway analysis

Since we have assumed that all of the reactions in Table 1 are essentially irreversible (i.e., $v_i \geq 0$), extreme pathway analysis may be used to generate a unique set of vectors spanning the nullspace of $\mathbf{S}$ [270]. A cone can be generated from this convex basis to circumscribe all allowable steady-state solutions to Equation 3.1:

$$C = \left\{ \mathbf{v} : \mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{p}_i, \alpha_i \geq 0, \forall i \right\}, \tag{3.3}$$

where $\mathbf{p}_i$ are the pathway vectors and α_i are positive weighting coefficients for each extreme pathway.

The pathway classification scheme developed previously [270] characterizes extreme pathways based on the activity of their exchange fluxes. Exchange fluxes can either be primary exchange fluxes (e.g., exchange of primary metabolites such as AAs, protein, etc.) or currency exchange fluxes (e.g., exchange of currency metabolites such as ATP, GTP, $\mathrm{P_i}$, etc.). An extreme pathway is classified as:

- Type I if it contains any non-zero primary exchange flux;
- Type II if the only active exchange fluxes are for currency metabolites; or
- Type III if there are no active exchange fluxes in the extreme pathway.

Here, we use a variant on this classification scheme in that we consider the NTPs used in RNA polymerization reactions to be primary metabolites rather than currency metabolites, since they are building blocks for the RNA and not simply an energy supply. Thus, in the fundamental system described in Figure 3.1, the primary exchange fluxes include b_1, b_2, b_3, and b_4.

Table 3.3: The stoichiometric matrix for the fundamental reaction scheme given in Table 3.2 and depicted in Figure 3.2.

$$\mathbf{S} = \begin{array}{ccccccccccccccc|ccccccccc l}
v_1 & v_{1P} & v_{1t} & v_{1r} & v_2 & v_{2P} & v_{2t} & v_{2r} & v_3 & v_{3P} & v_4 & v_{4P} & v_5 & v_{5P} & v_6 & b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 & b_8 & b_9 & \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G \\
0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G_{\mathrm{RNAP}} \\
0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G_{\mathrm{tRNA}} \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G_{\mathrm{rRNA}} \\
-1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{RNAP} \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G^* \\
0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G^*_{\mathrm{RNAP}} \\
0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G^*_{\mathrm{tRNA}} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & G^*_{\mathrm{rRNA}} \\
0 & 0 & 0 & 0 & -n & -n_P & -n_t & -n_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{NTP} \\
0 & 0 & 0 & 0 & 2n & 2n_P & 2n_t & 2n_r & 0 & 0 & 0 & 0 & 2a & 2a_P & 2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \mathrm{P}_i \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & n & n_P & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \mathrm{NMP} \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{mRNA} \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{mRNA}_P \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{rib} \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & a & a_P & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{tRNA} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{rib}^* \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{rib}^*_P \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2a & -2a_P & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & \mathrm{GTP} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -a & -a_P & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{AAtRNA} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{protein} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2a & 2a_P & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \mathrm{GDP} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & \mathrm{ATP} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{AA} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \mathrm{AMP}
\end{array}$$

3.1.5 Further constraints upon the system

Mass balance of mRNA production and degradation at steady-state implies $v_1 = v_2 = v_3$ and $v_4 = v_5$. The first three fluxes, $v_{1,2,3}$, cannot be directly coupled to the fluxes $v_{4,5}$ except in the following fashion.

If a particular protein is synthesized, then $v_{4,5} > 0$. Then, if $v_{4,5} > 0$, it must also hold that $v_{1,2,3} > 0$, since some amount of transcript must be present (and thus be maintained) for translation to occur. This constraint can be written as

$$v_{1,2,3} = \min\left(v_{\text{pr.str.}}, \text{limiting elongation rate}, b_2\right), \tag{3.4}$$

where $v_{\text{pr.str.}}$ is the transcription initiation flux which depends upon the promoter strength under the given set of conditions. The limiting elongation rate is the maximum speed at which the RNA polymerization can take place for a given transcript.

3.1.6 Promoter strengths

When constrained to nonzero values, the promoter-binding flux, v_1, can be set according to the promoter strength of the gene G under a specified set of conditions. Hence, if the influx of nucleotides is not limiting (i.e., $v_1 \leq nb_2$), the fluxes $v_{1,2,3}$ are set by the regulatory inputs to the system. If transcription regulation is to be taken into account, the regulatory “rules” under a specific set of conditions will determine whether or not v_1 is “on” [50], as well as its value under the specified conditions. For instance, if a gene is known to be down-regulated under a specific set of conditions, then the fluxes involved in synthesizing its mRNA transcript and the corresponding protein(s) will be set to zero. Bacterial promoter strengths have been studied extensively [161, 66, 344, 333].

3.1.7 Global maximum on transcription elongation

For particularly lengthy transcripts having strong promoters, it is possible that the elongation flux, v_2, will be limiting. Thus, an upper bound on the sum of all elongation (v_2) fluxes, $\beta_{2\text{-global max}}$, must be set based upon the sequence length and composition. For *E. coli*, typical elongation rates during transcription range from 50-100 nucleotides/RNAP/sec, depending on the gene, the regulation present, and the growth rate [113, 323]. If we assume that there are approximately 2500-3000 RNAP molecules per cell [150], then a typical cell is capable of incorporating 125,000-300,000 nucleotides/sec into RNA molecules. An average gene length of about ~1000 base pairs [26] implies that 125-300 mRNA molecules can be made by a single cell every second, assuming that no rRNA and tRNA is being made. In reality, however, this estimate will be significantly less, since 80% of the total RNA in *E. coli* is rRNA, and 15% is tRNA, with only 4% comprised of mRNA [200].

3.1.8 Calculation of mRNA concentrations

The mRNA degradation flux, v_3, is typically dependent upon the concentration of mRNA in a first-order, linear fashion:

$$v_3 = k[\text{mRNA}], \tag{3.5}$$

where k is the rate constant for the degradation of mRNA, which can be directly determined from the half-life of the particular mRNA [148, 171, 25]. Once $v_{1,2,3}$ is determined from the promoter strength for G, from a limiting nucleotide influx, or from the global maximum on the elongation fluxes, we can calculate the concentration of mRNA as follows:

$$[\text{mRNA}] = \begin{cases} v_{1,2,3}/k & \text{if} \quad v_{4,5} > 0 \\ 0 & \text{if} \quad v_{4,5} = 0 \end{cases}. \tag{3.6}$$

A Boolean logic representation has thus been assumed in that the gene is either “on” and being transcribed at a defined, condition-dependent rate, or it is “off,” and is not being transcribed at all [301]. In reality, there is a very low basal level of transcription at all times; accordingly, stochastic models have been used to take into account the "leakiness" of all promoters [10]. A small lower bound, α_1, on v_1 can thus be used. Hence, if the promoter strengths and mRNA half-lives for a given set of conditions are known (since both are subject to regulation), it is possible *a priori* to estimate genome-scale mRNA expression arrays.

3.1.9 Global maximum on translation initiation

Since there will be a finite number of free ribosomes available for protein synthesis at any given time, a global maximum must be set for the binding of each messenger RNA to a free ribosome [173, 91]. Thus, the following constraint must be applied to the v_4 fluxes for all of the proteins synthesized:

$$\sum_{i=1}^{N} v_{4,i} \leq \beta_{\text{4-global max}}, \tag{3.7}$$

where N is the total number of proteins being produced. Translation rates in *E. coli* are typically 16 amino acids/ribosome/sec (or 48 nucleotides/ribosome/sec) [325], which corresponds to 299,200 amino acids/cell/sec, assuming that there are 18,700 ribosomes per cell [200]. Assuming an average open reading frame (ORF) length of 317 amino acids [26], a typical *E. coli* cell can produce ∼950 protein molecules per second.

3.2 Results from analysis of fundamental protein synthesis network

The extreme pathway structure and key parameters governing protein production were identified for the following cases, ranging from highly simplified schema to a whole bacterial operon:

1. Fundamental system having 1 gene, 1 type of nucleotide, and 1 type of amino acid (Figure 3.1)

2. Fundamental system plus the production of the internal elements RNAP, tRNA, and rib (Figure 3.2)

3. Generalized N-gene operon (polycistronic mRNA)

4. Fundamental system having biologically meaningful values of 4 types of nucleotides and 20 types of amino acids

5. Production of malate dehydrogenase in *E. coli*

6. Production of the proteins encoded by the *lac* operon in *E. coli*

The extreme pathway results for the first four cases are summarized in Table 3.4.

3.2.1 Case 1—Fundamental system

An extreme pathway analysis of the network considered in the Case 1 divides the system into two functionally distinct categories (Figure 3.3). One extreme pathway (Panel A of Figure 3.3) corresponds to the maintenance of mRNA in the cell, and the other (Panal B of Figure 3.3) corresponds to the utilization of mRNA to manufacture protein. The maintenance flux for a particular mRNA is required whenever the encoded protein is being produced.

If the gene, G, is being transcribed, one of two parameters limits the protein production flux (b_3): either the amino acid influx (b_1) or the maximum flux allowed for translation initiation due to the finite ribosomal pool (β_4), whichever is smaller. The limitation of the protein production flux for the fundamental system may be mathematically desbribed as

$$b_3 \leq \min\left(\frac{b_1}{a}, \beta_4\right), \tag{3.8}$$

where a represents the number of amino acids in the protein. Note that, for this simplified system, there exists only one type of amino acid.

Table 3.4: Fundamamental extreme pathway structure for the first four cases of protein synthesis studied. In Case 2, n_X denotes the number of NTPs in mRNA and n_Y denotes the number of NTPs in mRNA$_\mathrm{P}$; in Case 3, the subscript A corresponds to the gene encoding protein A, and the subscript B corresponds to the gene encoding protein B; and in Case 4, the coefficients, $n_1, \ldots, n_4$ and $a_1, \ldots, a_{20}$ correspond to the number of each type of nucleotide/amino acid in the mRNA/protein of interest.

EP	**Net reaction equation**	**Primary function**
	Case 1:	
$\mathbf{p_1}$	$n\mathrm{NTP} \rightarrow n\mathrm{NMP} + 2n\mathrm{P_i}$	mRNA maintenance
$\mathbf{p_2}$	$a\mathrm{AA} + a\mathrm{ATP} + 2a\mathrm{GTP} \rightarrow \mathrm{protein} + a\mathrm{AMP} + 2a\mathrm{GDP} + 4a\mathrm{P_i}$	mRNA utilization
	Case 2:	
$\mathbf{p_1}$	$n_\mathrm{X}\mathrm{NTP} \rightarrow n_\mathrm{X}\mathrm{NMP} + 2n_\mathrm{X}\mathrm{P_i}$	prot. mRNA maint.
$\mathbf{p_2}$	$n_\mathrm{Y}\mathrm{NTP} \rightarrow n_\mathrm{Y}\mathrm{NMP} + 2n_\mathrm{Y}\mathrm{P_i}$	RNAP mRNA maint.
$\mathbf{p_3}$	$a\mathrm{AA} + a\mathrm{ATP} + 2a\mathrm{GTP} \rightarrow \mathrm{protein} + a\mathrm{AMP} + 2a\mathrm{GDP} + 4a\mathrm{P_i}$	prot. mRNA util.
	Case 3 (for a 2-gene operon):	
$\mathbf{p_1}$	$n\mathrm{NTP} \rightarrow n\mathrm{NMP} + 2n\mathrm{P_i}$	mRNA maintenance
$\mathbf{p_2}$	$a_\mathrm{A}\mathrm{AA} + a_\mathrm{A}\mathrm{ATP} + 2a_\mathrm{A}\mathrm{GTP} \rightarrow \mathrm{protein_A}$	mRNA utilization
	$+a_\mathrm{A}\mathrm{AMP} + 2a_\mathrm{A}\mathrm{GDP} + 4a_\mathrm{A}\mathrm{P_i}$	prod. of protein$_\mathrm{A}$
$\mathbf{p_2}$	$a_\mathrm{B}\mathrm{AA} + a_\mathrm{B}\mathrm{ATP} + 2a_\mathrm{B}\mathrm{GTP} \rightarrow \mathrm{protein_B}$	mRNA utilization
	$+a_\mathrm{B}\mathrm{AMP} + 2a_\mathrm{B}\mathrm{GDP} + 4a_\mathrm{B}\mathrm{P_i}$	prod. of protein$_\mathrm{B}$
	Case 4:	
$\mathbf{p_1}$	$n_1\mathrm{N_1TP} + n_2\mathrm{N_2TP} + n_3\mathrm{N_3TP} + n_4\mathrm{N_4TP} \rightarrow$	mRNA maintenance
	$n_1\mathrm{N_1MP} + n_2\mathrm{N_2MP} + n_3\mathrm{N_3MP} + n_4\mathrm{N_4MP} + 2\sum_{i=1}^{4} n_i\mathrm{P_i}$	
$\mathbf{p_2}$	$a_1\mathrm{AA_1} + a_2\mathrm{AA_2} + \cdots + a_{20}\mathrm{AA_{20}} + \sum_{i=1}^{20} a_i(\mathrm{ATP} + 2\mathrm{GTP}) \rightarrow$	mRNA utilization
	$\mathrm{protein} + \sum_{i=1}^{20} a_i(\mathrm{AMP} + 2\mathrm{GDP} + 4\mathrm{P_i})$	

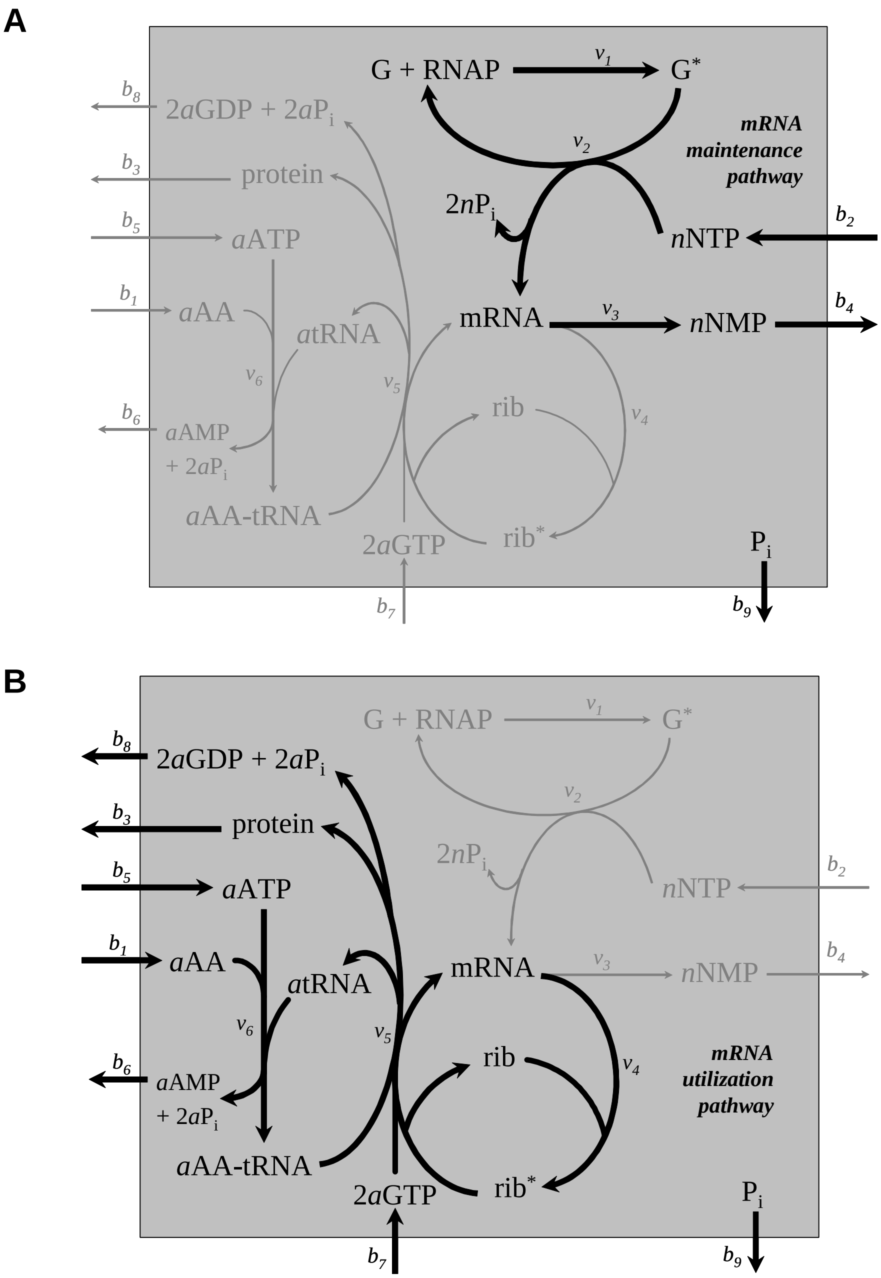


Figure 3.3: The extreme pathway structure of the fundamental protein synthesis network. There are two extreme pathways for the basic system: a) The extreme pathway for the maintenance of mRNA, and b) the extreme pathway for the synthesis of protein.

3.2.2 Case 2—Synthesis of internal components

Untranslated RNA transcripts included in the system (i.e., rRNAs and tRNAs) will not result in any additional extreme pathways, since we have assumed that the stable RNA is allowed neither to degrade nor to leave this idealized system. When the RNAP gene is included, one extreme pathway is added for the maintenance of the associated mRNA, but no "production" pathway is added since no exchange flux is written for RNAP (Table 3.4; Figure 3.4). The key parameters for protein production are thus unchanged from the system in Case 1. Thus, the internal elements are not included in any of the following cases.

3.2.3 Case 3—Generalized N-gene operon

In the case of polycistronic mRNA, there will be an extreme pathway corresponding to the maintenance of the mRNA and one extreme pathway for each protein encoded for on that particular mRNA. In general, an N-gene operon encodes for N proteins, and the resulting system thus has $N+1$ extreme pathways.

The maximum production flux of each of the N proteins is determined in the following manner:

$$b_3 \leq \min\left(\frac{b_1}{\sum_{i=1}^{N} a_i}, \frac{\beta_{4\text{-global max}}}{N}\right). \tag{3.9}$$

The ribosomal capacity is equally shared by the N transcripts since we assume that the half-lives are identical. (It is important to note that when the model is scaled to include more than one operon, the ribosomal capacity will be shared by all cellular transcripts and not just the transcripts of a particular operon.)

3.2.4 Case 4—Biological number of NTs and AAs

If we increase the types of nucleotides and amino acids to 4 and 20, respectively, as found in most living cells, the resulting system still contains one maintenance pathway per mRNA and one utilization pathway per protein produced. The extreme pathway analysis thus remains essentially unchanged from

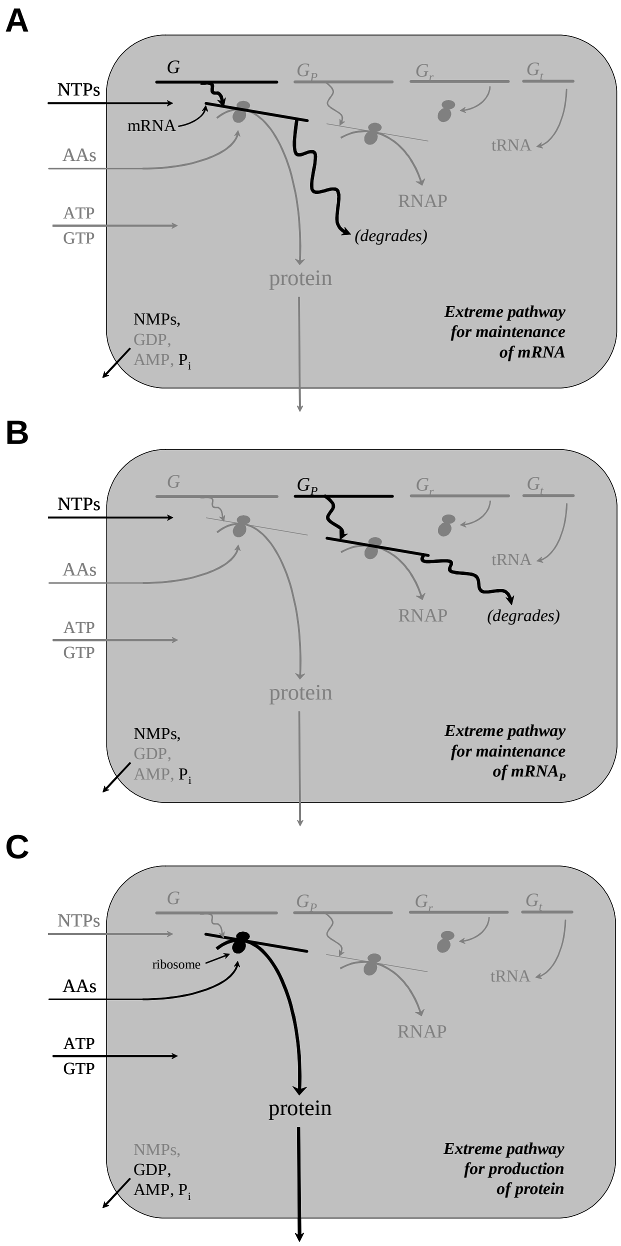


Figure 3.4: The extreme pathway structure when the synthesis of internal accessory components are included. There are three extreme pathways in this network: (A) The extreme pathway for the maintenance of mRNA, (B) the extreme pathway for the maintenance of the mRNA encoding for RNAP (mRNA_P), and (C) the extreme pathway for the synthesis of protein.

Case 3. We still have $N+1$, albeit more complicated, extreme pathways per N-gene operon (Table 3.4). Increasing the number of types of nucleotides or amino acids in the system has no effect on the extreme pathway structure, and the limiting amino acid influx to the system determines the protein production flux if the ribosomal pool is not limiting.

3.2.5 Case 5—Production of *E. coli* malate dehydrogenase

The nucleotide sequence for *mdh* and for the resulting amino acid sequence for malate dehydrogenase are given in Table 3.5 for *E. coli.* The mRNA encoded by this gene contains 219 adenine, 228 cytosine, 263 guanine, and 229 uracil residues, and the resulting malate dehydrogenase protein consists of the amino acid residues listed in Table 3.6. The inputs to the simplified system were increased to 4 types of nucleotide and 20 types of amino acid, and the *E. coli* gene, *mdh*, was selected as G [26]. This network exhibits two extreme pathways: a maintenance pathway for mRNA_{mdh} and a protein synthesis pathway for the production of the gene product, malate dehydrogenase. This result is analogous to that in Case 4 for $N=1$, except we now have numerical values that correspond to an actual gene in *E. coli.*

Figure 3.5 provides a schematic of the material costs (i.e., the NTP and AA inputs) and the energy costs (ATP and GTP required) for the maximal production of malate dehydrogenase. If all amino acid influxes are arbitrarily constrained to 10 (units of moles/cell/time) and if the ribosomal saturation is not limiting, then the maximal production flux of malate dehydrogenase is equal to 0.278 (moles/cell/time) (panel A of Figure 3.5). Note that this value is not intended to be a calculation of malate dehydrogenase actually produced in an *E. coli* cell (i.e., the units are arbitrary), but rather to highlight the limiting constraints on the network that synthesizes this enzyme. Glycine is therefore the limiting amino acid in this scenario since it is the most abundant amino acid in the malate dehydrogenase protein.

Next we considered the scenario depicted in panel B of Figure 3.5, in

Table 3.5: The nucleotide sequence of the gene *mdh* in *E. coli*, and the translated amino acid sequence of the corresponding malate dehydrogenase protein.

***mdh* nucleotide sequence**
`atgaaagtcgcagtcctcggcgctgctggcggtattggccaggcgcttgcactactgtta`
`aaaacccaactgccttcaggttcagaactctctctgtatgatatcgctccagtgactccc`
`ggtgtggctgtcgatctgagccatatccctactgctgtgaaaatcaaaggtttttctggt`
`gaagatgcgactccggcgctggaaggcgcagatgtcgttcttatctctgcaggcgtagcg`
`cgtaaaccgggtatggatcgttccgacctgtttaacgttaacgccggcatcgtgaaaaac`
`ctggtacagcaagttgcgaaaacctgcccgaaagcgtgcattggtattatcactaacccg`
`gttaacaccacagttgcaattgctgctgaagtgctgaaaaaagccggtgtttatgacaaa`
`aacaaactgttcggcgttaccacgctggatatcattcgttccaacacctttgttgcggaa`
`ctgaaaggcaaacagccaggcgaagttgaagtgccggttattggcggtcactctggtgtt`
`accattctgccgctgctgtcacaggttcctggcgttagttttaccgagcaggaagtggct`
`gatctgaccaaacgcatccagaacgcgggtactgaagtggttgaagcgaaggccggtggc`
`gggtctgcaaccctgtctatgggccaggcagctgcacgttttggtctgtctctggttcgt`
`gcactgcagggcgaacaaggcgttgtcgaatgtgcctacgttgaaggcgacggtcagtac`
`gcccgtttcttctctcaaccgctgctgctgggtaaaaacggcgtggaagagcgtaaatct`
`atcggtaccctgagcgcatttgaacagaacgcgctggaaggtatgctggatacgctgaag`
`aaagatatcgccctgggcgaagagttcgttaataagtaa`
Malate dehydrogenase amino acid sequence
`MKVAVLGAAGGIGQALALLLKTQLPSGSELSLYDIAPVTPGVAVDLSHIPTAVKIKGFSG`
`EDATPALEGADVVLISAGVARKPGMDRSDLFNVNAGIVKNLVQQVAKTCPKACIGIITNP`
`VNTTVAIAAEVLKKAGVYDKNKLFGVTTLDIIRSNTFVAELKGKQPGEVEVPVIGGHSGV`
`TILPLLSQVPGVSFTEQEVADLTKRIQNAGTEVVEAKAGGGSATLSMGQAAARFGLSLVR`
`ALQGEQGVVECAYVEGDGQYARFFSQPLLLGKNGVEERKSIGTLSAFEQNALEGMLDTLK`
`KDIALGEEFVNK`

Table 3.6: Amino acid composition of malate dehydrogenase, triosephosphate isomerase, and the proteins encoded by the *lac* operon, and the nucleotide composition of the corresponding genes.

	mdh	*lacZ*	*lacY*	*lacA*
A	219	678	240	189
C	228	842	284	125
G	263	888	297	137
T	229	667	433	161
Total	939	3075	1254	612
Ala	35	77	35	8
Arg	8	66	12	10
Asn	11	47	16	16
Asp	12	64	6	9
Cys	3	16	8	2
Gln	14	58	11	0
Glu	20	62	11	14
Gly	36	71	36	16
His	2	34	4	8
Ile	17	39	33	17
Leu	33	96	54	9
Lys	21	20	12	10
Met	4	24	14	7
Phe	10	38	56	8
Pro	13	62	12	12
Ser	17	60	29	12
Thr	18	56	19	12
Trp	0	39	6	2
Tyr	4	31	14	9
Val	34	64	29	22
Total	312	1024	417	203

which the ribosomal pool limits the overall protein production. In this example, 0.2 is the upper bound on $v_{4\text{-global}}$ (i.e., $\beta_{4\text{-global max}} = 0.2$ moles/cell/time). Thus, the maximum attainable protein production flux is also 0.2, and no amino acid is limiting.

3.2.6 Case 6—The *lac* Operon

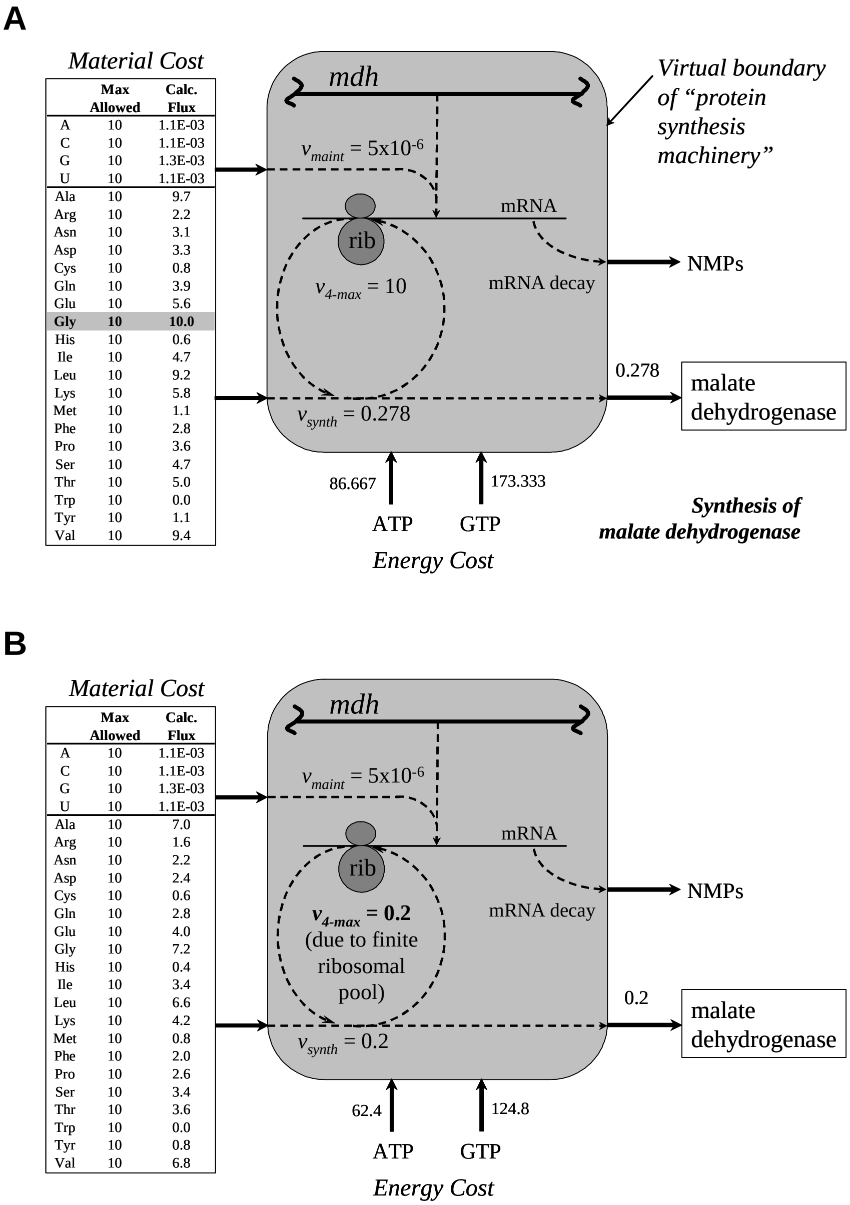
Expression of the *lac* operon involves 4 extreme pathways: one for the maintenance of the polycistronic mRNA, and the other three for the production of each of the three *lac* proteins. This case is analogous to Case 4 with $N = 3$. A schematic of the production of these three proteins is given in panel A of Figure 3.6, and the nucleotide and amino acid composition summaries are provided in Table 3.6.

If, as before, we constrain all amino acid influxes equally (and if the ribosomal saturation flux is left unconstrained), the limiting factor is the supply of leucine (panel B of Figure 3.6). This system can also be constrained by the ribosomal capacity or by any other amino acid if the amino acids are available in uneven supply.

3.3 Discussion

In this chapter I have demonstrated that it is possible to perform a stoichiometry-dependent structural analysis of gene expression and protein synthesis using extreme pathway and flux balance analyses. This framework provides a simple, clear, and detailed accounting of a cell's energy and material expenditure for protein synthesis. Furthermore, the fundamental model presented here is completely scalable, and can readily be expanded to genome-scale via direct use of genomic sequence data.

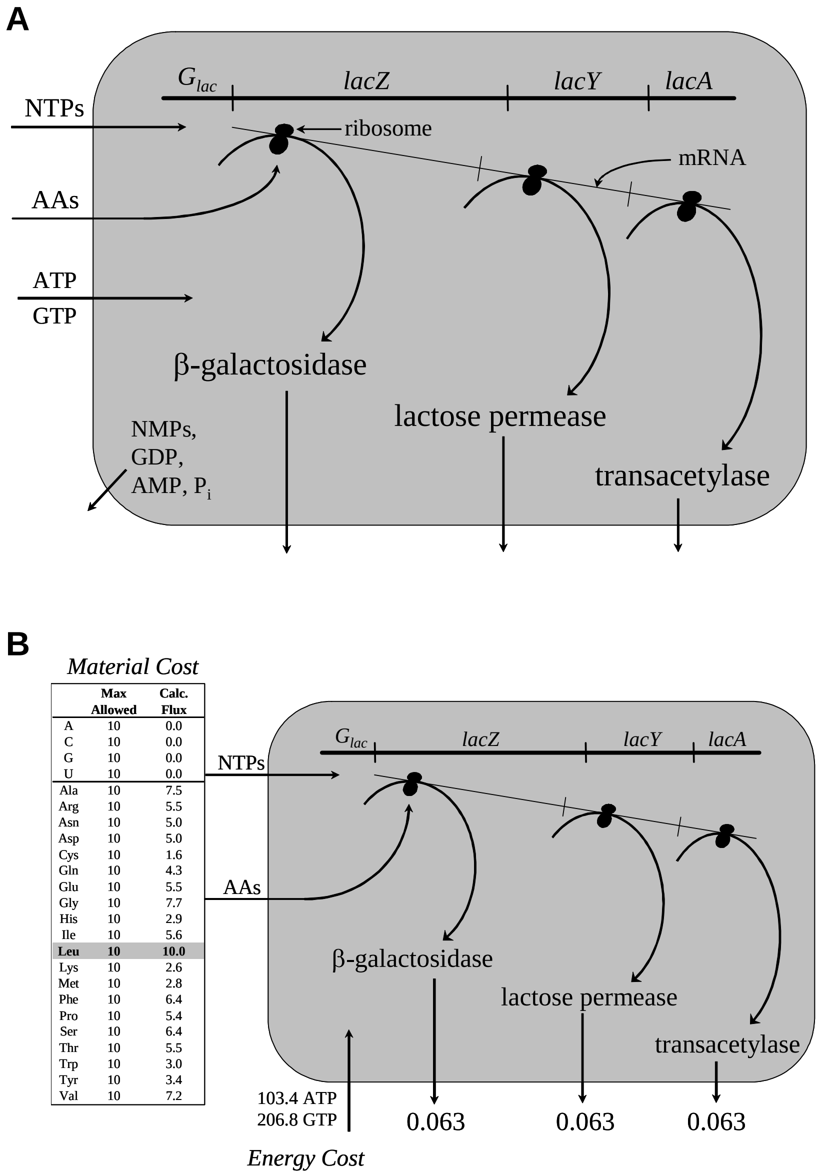
There are essentially two types of extreme pathways involved in protein production: those involved in the *maintenance* of messenger RNA and those in-



	Max Allowed	Calc. Flux
A	10	1.1E-03
C	10	1.1E-03
G	10	1.3E-03
U	10	1.1E-03
Ala	10	9.7
Arg	10	2.2
Asn	10	3.1
Asp	10	3.3
Cys	10	0.8
Gln	10	3.9
Glu	10	5.6
Gly	**10**	**10.0**
His	10	0.6
Ile	10	4.7
Leu	10	9.2
Lys	10	5.8
Met	10	1.1
Phe	10	2.8
Pro	10	3.6
Ser	10	4.7
Thr	10	5.0
Trp	10	0.0
Tyr	10	1.1
Val	10	9.4

	Max Allowed	Calc. Flux
A	10	1.1E-03
C	10	1.1E-03
G	10	1.3E-03
U	10	1.1E-03
Ala	10	7.0
Arg	10	1.6
Asn	10	2.2
Asp	10	2.4
Cys	10	0.6
Gln	10	2.8
Glu	10	4.0
Gly	10	7.2
His	10	0.4
Ile	10	3.4
Leu	10	6.6
Lys	10	4.2
Met	10	0.8
Phe	10	2.0
Pro	10	2.6
Ser	10	3.4
Thr	10	3.6
Trp	10	0.0
Tyr	10	0.8
Val	10	6.8

Figure 3.5: The synthesis of malate dehydrogenase in *E. coli*. The table on the left of each figure provides the constraints placed upon the nucleotide and amino acid influxes, as well as the calculated influxes upon optimization of protein production. The v_{maint} flux (equal to the $v_{1,2,3}$ flux in the text) was arbitrarily set to 5×10^{-6} (in units of concentration/time). The $v_{4-\text{max}}$ flux, which corresponds to the maximal ribomosal binding flux due to a finite ribosomal pool, is set as a constraint. The v_{synth} is the protein production flux which is being maximized. In panel (A), all amino acid influxes are constrained to 10, and the $v_{4-\text{max}}$ flux is also constrained to 10. In this case, the glycine influx is limiting. In panel (B), the amino acid constraints are unchanged, but the $v_{4-\text{max}}$ flux is constrained to 0.2. The ribosomal pool thus becomes limiting in this case.



	Max Allowed	Calc. Flux
A	10	0.0
C	10	0.0
G	10	0.0
U	10	0.0
Ala	10	7.5
Arg	10	5.5
Asn	10	5.0
Asp	10	5.0
Cys	10	1.6
Gln	10	4.3
Glu	10	5.5
Gly	10	7.7
His	10	2.9
Ile	10	5.6
Leu	**10**	**10.0**
Lys	10	2.6
Met	10	2.8
Phe	10	6.4
Pro	10	5.4
Ser	10	6.4
Thr	10	5.5
Trp	10	3.0
Tyr	10	3.4
Val	10	7.2

Figure 3.6: The expression of the *lac* operon. A simplified schematic for the synthesis of the proteins encoded by the *lac* operon in *E. coli* is provided in panel (A). The mRNA for this operon is polycistronic, encoding for three proteins. In panel (B), the production fluxes of three proteins encoded by the *lac* operon in *E. coli* are being maximized, and the ribosomal binding flux is not limiting. All amino acid influxes are constrained to 10, and the influx of leucine is limiting.

volved in the *utilization* of mRNA to synthesize protein. From a structural standpoint, these processes are decoupled, save that both pathway types are active if the corresponding gene is being expressed and its protein is being synthesized. Thus, their interdependence is strictly logical, whereas the actual flux values are determined by the following key parameters: the mRNA maintenance fluxes (for a particular gene or operon) are strictly dependent upon the promoter strength of the gene under a given set of environmental conditions, except in the event that nucleotide or RNAP availability is limited. The resulting mRNA concentration can then be calculated directly if the half-life of the mRNA is known [131, 39, 25]. For the cases studied, the fluxes involved in the utilization of expression information are dependent upon the total ribosomal pool or upon the availability of amino acids, whichever is limiting.

I have defined the properties of stoichiometric models for individual genes and operons. When one scales this framework to describe the protein production of an entire genome, one will need to deal with the interactions between these genes (and operons) and the machinery within the fundamental system considered herein. These interactions arise since all genes compete for a finite pool of available RNAP and ribosomes. Thus, the resulting mRNA maintenance fluxes are weighted according to the promoter strengths of the various genes. Similarly, the different mRNA transcripts must compete for a limited number of ribosomal binding sites. These translation initiation fluxes must therefore be weighted according to the relative abundances of each mRNA, which, in turn, can be calculated from the corresponding mRNA maintenance fluxes and half-lives (Equation 3.5). If large-scale promoter strength and mRNA half-life data are unavailable for a prokaryotic organism of interest, the weighting on the translation initiation (i.e., the relative mRNA abundances) fluxes may be estimated directly from gene expression profiles [297, 331].

The overall simplicity of the topology of the reactions involved in protein production is noteworthy in light of the complexity that has been found to exist in metabolic networks [220, 232]. Here, a lack of robustness is evident in that there are really no choices that can be made within the mRNA expression maintenance and mRNA expression utilization extreme pathways. The external environment provides a set of inputs that set the fluxes in a condition-dependent manner, and the corresponding proteins are produced from available amino acids and currency metabolites (i.e., ATP and GTP). Thus the protein synthesis network is more rigid than metabolism.

As genome sequences continue to become available and their gene products are elucidated (the "parts catalogue" of the cell, as it were), it is becoming increasingly evident that the interaction of simple components yields tremendous complexity in biology [216, 295, 7, 83]. We currently know most of the protein components that are encoded within a genome, and here we have described a fundamental network for the synthesis of each of these proteins. The task at hand is to scale these fundamental networks to include all protein components in a genome, and then to integrate these components with existing genome-scale metabolic networks, and corresponding regulatory networks when they become available [50].

Taken together, the results presented in this study show that the constraint-based approach of FBA can be used to describe protein synthesis. This approach is readily scaled-up to describe the activity of an entire bacterial genome, and can be integrated with metabolic FBA models.

The text of this chapter, in full, is a reprint of the material as it appears in T.E. Allen and B.O. Palsson. 2003. Sequence-Based Analysis of Metabolic Demands for Protein Synthesis in Prokaryotes. *Journal of Theoretical Biology*, 220:1-18. I was the primary author of this publication, and the co-author participated and supervised the research which forms the basis for this chapter.

Chapter 4

Data Integration: The Uses of the *Escherichia coli* Genome

As described in Chapter 1, the increasing availability of complete genome sequences has ushered in an era of genome-enabled science that allows for the construction of *in silico* models at the genome scale [49, 163, 209, 213, 280]. In addition to genome sequences, other high-throughput data types, including transcriptomic, proteomic, metabolomic, global mRNA decay data, and interaction data, are growing at an ever-increasing rate [115]. This wealth of genome-scale data highlights the need for scalable *in silico* methods by which to integrate and reconcile heterogeneous datasets [218].

In the last chapter, I described a sequence-based framework for calculating the metabolic costs of expressing a gene and synthesizing its gene product [3]. These costs are calculated directly from the DNA sequence, and estimations of ribosomal content can be used to scale the total protein producing capacity of the cell and the requisite costs. The established framework, when scaled to account for all the genes in the *Escherichia coli* K-12 MG1655 genome [26], would allow for the explicit calculation of the material and energy costs required for expressing the entire genome, in addition to the costs for synthesizing the resulting proteome. Fundamental values for cellular biomass requirements have been experimentally measured for *E. coli* [200], but these values have never been calculated directly from the merging of sequence data with high-throughput gene expression data. Previous sequence-based cost estimates for protein synthesis have been calculated from expression estimates based on codon usage [2], but have not integrated actual expression or mRNA half-life data. A method for integrating such heterogeneous datasets would provide fundamental material and energy cost values, estimated effective promoter strengths on a genome scale, and the genome-location distribution of gene expression in prokaryotes.

Expression profiling has been used to identify genes whose expression changes under shifting environmental conditions [9, 210, 252, 297, 347]. A variety of methods have been developed with which to analyze these data, including co-expression pattern analysis for operon prediction [258], dimensionality reduction techniques [132, 166], and several types of clustering methods [8]. A model-driven means by which to interpret and analyze expression data, however, has not been established. The availability of sequence data, expression data, and, most recently, global messenger RNA (mRNA) half-life data [25, 279] has created a need for such a structured analysis and integration of these disparate datasets. We have developed a method that accomplishes this goal, and have used it to study the overall cost of maintaining a particular expression state, the distribution of individual "effective" promoter strengths, and the corresponding genome-location dependent characteristics of gene expression.

4.1 Data integration methods

4.1.1 *In silico* analysis framework

The analysis framework established previously [3] describes a means of calculating the material and energy costs for maintaining a particular mRNA transcript and for synthesizing the resulting protein. For mRNA maintenance, the constituent nucleotide triphosphates are required to maintain the concentration of a transcript at a particular steady-state concentration [31]. If the transcription rate, v_{mRNA} (in units of numbers of transcripts per cell per time, typically transcripts/sec/cell), is known for a gene, the requisite nucleotide demands can be calculated directly from the gene sequence.

Similarly, if the abundance of a particular transcript (m_i) relative to the total mRNA content ($m_{\text{rel},i} = m_i/m_{\text{tot}}$, where $m_{\text{tot}} = \sum_k m_k$) and the ribosomal content of the cell are known, upper bounds on the amino acid requirements for synthesizing the encoded protein can be explicitly calculated. Thus, if the protein synthesis rate (i.e., number of protein molecules translated per cell per unit time) is known, the amino acid building blocks required to synthesize the encoded protein can be calculated directly from the sequence. In addition to the amino acid costs, 1 ATP and 2 GTP molecules will be required for each peptide bond that is formed [3, 188].

4.1.2 Calculation of transcription state

The transcription state is defined as the vector of all transcription rates in the genome, $v_{\text{mRNA},i}$ ($i = 1, \ldots, N$, where N represents the number of coding sequence ORFs in the genome). The transcription state of the *E. coli* genome can be explicitly calculated using sequence data if the following parameters are known: the effective promoter strengths (or ORF usages), the mRNA degradation rate of each transcript being synthesized, the mRNA abundances, and the free RNA polymerase (RNAP) concentration. At the genome scale, we can write for each transcript:

$$v_{\text{deg},i} = k_{\text{deg},i} m_i \tag{4.1}$$

where $v_{\text{deg},i}$, $k_{\text{deg},i}$, and m_i represent the degradation rate, the mRNA degradation rate constant, and the mRNA concentration, respectively, for the i^{th} gene.

The transcription initiation rates, $v_{\text{mRNA},i}$, can be approximated if the effective promoter strength for each gene (q_i, in units of $\text{M}^{-1}\text{s}^{-1}$) and the RNAP and promoter concentrations ($[\mathrm{P}]_i$) are known [244]:

$$v_{\mathrm{mRNA},i} = q_i[\mathrm{RNAP}][\mathrm{P}]_i. \tag{4.2}$$

It is assumed that transcription elongation is not limiting to protein synthesis, since once transcription initiation occurs, ribosomes may bind to the unfinished mRNA transcript and translation may commence at a rate comparable to the mRNA elongation rate [325].

In a steady-state, the transcription rate must balance the mRNA degradation rate:

$$v_{\mathrm{deg},i} = v_{\mathrm{mRNA},i} \tag{4.3}$$

It is therefore possible to reconcile data containing mRNA concentrations, effective promoter strengths, and mRNA degradation rates in the following manner:

$$k_{\mathrm{deg},i}m_i = q_i[\mathrm{RNAP}][\mathrm{P}]_i \tag{4.4}$$

The effective promoter strengths, q_i, that will depend on both the intracellular conditions and the regulation present, can thus be calculated globally given large-scale mRNA concentration data [278, 331] and mRNA half-life data [25, 279]. If log-phase growth is assumed, the copies per cell of each promoter can be estimated from each gene's position on the chromosome and the growth rate of the cell [31]. Since these q_i's are essentially normalized transcription rate constants, they will be subject to regulation. Thus, the variance of each q_i across many datasets becomes a useful quantity. The vector of all q_i's, $\mathbf{q} = (q_1 \ldots q_N)$, constitutes the promoter activation state of the genome, where N represents the number of coding sequences in the genome.

4.1.3 Metabolic cost of RNA synthesis

The synthesis rate of each mRNA transcript—which will determine the nucleotide triphosphates required—is set by the effective promoter strength, q_i, for each (i^{th}) gene. Neither the mRNA elongation rate nor the free RNAP concentration is assumed to be limiting to the synthesis rate of each transcript [31, 244]. In the absence of large-scale promoter strength data, however, the transcription rate for each transcript may be estimated from the relative mRNA abundances (estimated from expression data) and from available mRNA decay rates [25, 279] (Equation 4.1). One may normalize the nucleotides required for mRNA maintenance when the total mRNA concentration ($[\text{mRNA}]_{\text{tot}}$) at a given growth rate is known [31].

4.1.4 Metabolic cost of protein synthesis

The total protein synthesis rate (i.e., the overall capacity of the cell to synthesize protein) will be limited by the number of ribosomes available to the cell [31, 173]. Additionally, the relative abundance of each transcript ($m_{\text{rel},i}$) will determine the weighting of the synthesis rate for each protein since all mRNA transcripts will compete for the pool of available ribosomes. This disregard for the potential effect of transcript length on ribosomal occupancy is probably valid since the messages are not necessarily saturating. In fact, the number of ribosomes in a typical *E. coli* cell is about an order of magnitude greater than the total number of messages [200]. Thus, an upper bound on each protein synthesis rate can be set as follows:

$$v_{\text{prot},i} = \frac{\beta}{a_i} m_{\text{rel},i}, \tag{4.5}$$

where β is the maximal protein synthesis capacity of the cell (in units of number of peptide bonds formed per cell per time, about 340,000 peptide bonds per cell per second [31]) as limited by the number of ribosomes present, a_i is the number of amino acids in each protein, and $m_{\text{rel},i}$ is the relative abundance of each mRNA transcript. The corresponding amino acid costs for supporting these upper bounds on protein synthesis rates can be directly calculated from the known sequence. Additionally, the energy cost required for ribosomal binding, translocation along the ribosomes, and tRNA charging can be calculated for each protein synthesis rate.

Analyzing genome-location dependent patterns in gene expression The calculation of the transcription state of the genome calls for a means of analyzing potential patterns in expression along the chromosome. Wavelet transform techniques [21] can be used to analyze and visualize the genome-location dependent variability of gene expression. While standard Fourier transforms allow identifying periodic patterns in stationary signals, wavelet transforms allow identifying both periodic and non-periodic localized patterns and do not assume a stationary signal. In this work we used the continuous wavelet transform, which is better suited for visualizing patterns than its discrete counterpart [198]. The continuous wavelet transform of signal $x(t)$ (in our case, effective promoter strengths along the genome) is defined as

$$W(t,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g\left(\frac{t'-t}{a}\right) x(t')dt', \tag{4.6}$$

where $g((t'-t)/a)$ is the wavelet transform filter centered at t, and the width of the filter a is used to determine the scale at which patterns are analyzed. By choosing the filter function g we can extract different types of patterns from the data. Here we used the Morlet wavelet defined as $g(t) = \cos(5t)\exp(-t^2/2)$, which is particularly well suited for studying localized periodic patterns in data [21]. The wavelet transform can be visualized using a scalogram that displays the transform $W(t,a)$ as a contour plot with location along the genome, t, on one axis and the scale, a, on the other axis. We evaluated the significance of the spatial patterns extracted through wavelet analysis by randomizing the gene order in the *E. coli* genome and re-computing the transform for each randomized genome. A P-value for each individual $W(t,a)$ was then calculated based on 1000 randomized genomes by computing the number of times a specific $|W^*(t,a)|$ for a randomized genome is larger than the true $|W(t,a)|$.

4.1.5 Experimental methods and normalization

All mRNA expression data were generated from *E. coli* grown in batch culture and are available online [108].[1] Most experiments used the sequenced K-12 strain MG1655, 17 experiments involved strains derived from MG1655 with single ORF disruptions, and two experiments (single spotted array hybridizations) used strains DH5alpha and DH10B. The majority of experiments (39/49) used cells harvested at early exponential phase growth, and 10 experiments used cells from late exponential phase or stationary phase cultures. Forty-six of the experiments used cells grown in a MOPS-based minimal medium, while three used Luria-Bertani (LB) media. Glucose was used as the carbon source in most minimal medium experiments (43/49), and the others used acetate, glycerol, or proline as carbon sources. Data were collected by hybridization of fluorescently labeled cDNAs to either Affymetrix *E. coli* antisense oligonucleotide arrays (as described in [257]) or microarrays of spotted ORF-length PCR fragments (as described in [331]). The oligonucleotide arrays contained probes for both ORFs and intergenic regions, but only the data corresponding to ORFs were considered in this study. For each ORF on the Affymetrix array we calculated the average difference value using the Microarray Suite Software (Affymetrix, Inc., Santa Clara, CA). For spotted arrays the "signal" for each ORF was taken to be the average intensity of duplicate spots on the array. Fluorescently-labeled genomic DNA was used as a reference for the spotted arrays, thus providing an absolute measure of expression. To convert the signal values to estimates of transcript abundances, the simplifying assumption was made that for each experiment an average *E. coli* cell in the population contained 10,000 (gene-sized) mRNA transcripts [200]. The signal for each ORF on each array was scaled by a factor 10,000/sum of the signal intensities for each array. When replicate hybridizations were available, the scaled signal values were averaged across arrays. A small number of spots on each spotted microarray were disregarded when averaging across replicates because of poor quality PCR, spot-

[1] https://asap.ahabs.wisc.edu/annotation/php/logon.php

ting, or hybridization. For this reason the sum of the copies/cell estimates are slightly lower than 10,000 and vary across the spotted cDNA array experiments.

4.2 Results from integration of genome-scale data in *E. coli*

The *in silico* and experimental methods described above were used to address three questions: What are the metabolic resources required for expressing the entire *E. coli* genome under various conditions? What is the distribution of effective promoter strengths, and is this distribution gene-function dependent? Do these estimated promoter strength distributions reveal genome-location dependent patterns in gene expression?

4.2.1 Metabolic cost of genome expression

The cost of expressing the *E. coli* genome was calculated for a number of different steady-state mRNA concentration distributions. A number of random distributions were probed, as well as mRNA concentrations derived directly from the 49 gene expression datasets generated in this study. All of these cost calculations were normalized using parameters corresponding to a cell with a 40 minute doubling time (Table 4.1). Thus, for the mRNA maintenance cost, the mRNA concentrations were normalized to a specified total mRNA concentration ($[\mathrm{mRNA}]_{\mathrm{tot}} = \sum m_i = 4.188 \times 10^{-3}$ M). Similarly, the protein synthesis rates (and the corresponding costs) were normalized assuming 21,040 active ribosomes/cell [31], or $\beta = 3.37 \times 10^5$ peptide bonds/cell/second assuming a peptide elongation rate of 16 amino acids/ribosome/second [325]. Note that the amino acid costs provided are actually upper bounds on the costs, since possible tRNA abundance constraints have not been taken into account.

Table 4.1: Calculated amino acid and nucleotide demands for expressing the *E. coli* genome (mmol/g-DCW/hr). The average protein length and the resulting byproduct synthesis rates are included for each set of simulations. The calculations in the first column are derived from randomly generated datasets, while those in the second column are derived directly from the 49 gene expression datasets in this study. The third column gives the CVs for the data-based calculations across all 49 datasets. All results have been normalized using parameters corresponding to a doubling time of 40 minutes: total [mRNA] = 4.188×10^{-3} M, total ribosomal content = 21,040 active ribosomes, mass = 4.33×10^{-13} g-DCW/cell, and density = 382.72 g-DCW/L [31].

	Demands		
	Random	All Data	CVs (%)
AA's	316.93	276.10	6.4
ALA	0.66	0.66	1.7
ARG	0.38	0.39	1.6
ASN	0.27	0.28	1.4
ASP	0.36	0.37	2.2
CYS	0.08	0.07	8.6
GLN	0.31	0.30	3.0
GLU	0.40	0.43	4.7
GLY	0.51	0.53	1.8
HIS	0.16	0.15	4.8
ILE	0.42	0.42	1.3
LEU	0.74	0.69	3.4
LYS	0.31	0.35	7.3
MET	0.20	0.19	2.2
PHE	0.27	0.26	4.0
PRO	0.31	0.29	3.1
SER	0.40	0.39	2.8
THR	0.38	0.38	1.3
TRP	0.11	0.09	10.0
TYR	0.20	0.19	3.1
VAL	0.49	0.51	3.0
ATP	7.02	7.02	0.01
CTP	0.08	0.08	0.7
GTP	13.97	13.96	0.004
UTP	0.08	0.07	1.4

Simulated *in silico* expression profiles

The cost of expressing a particular distribution of mRNA transcripts and for synthesizing the encoded proteins was calculated for three random mRNA concentration distributions: uniform, normal, and exponential distributions. Since the calculations for any randomly generated expression profile, regardless of distribution, were nearly invariant, Table 4.1 provides the mean nucleotide and amino acid demands (as well as the resulting byproducts) for a typical simulation. The coefficients of variation (CVs) were found from calculating the costs given by 400 simulations, but are not shown in the table since they were all less than 1%.

Measured *in vivo* expression profiles

The material and energy costs were then calculated for mRNA concentration distributions derived from available experimentally determined gene expression data, and the resulting costs and CVs are provided in Table 4.1. Gene expression datasets from 49 separate experiments (corresponding to 91 hybridizations—41 Affymetrix and 50 spotted cDNA arrays) were generated as described, and transcript copies/cell estimations were made for most of the 4290 coding sequences in *E. coli* for each dataset. For the spotted arrays, the transcript copies/cell estimations were made from microarrays normalized using genomic DNA as described above. The experimental conditions from which these data were derived varied widely and include exponential and stationary-phase growth in glucose minimal medium, exponential growth in acetate and in glycerol minimal media, response to acid shock, response to cold shock, response to heat shock, growth in media containing an antibiotic, growth in LB broth, and various deletions grown on glucose minimal medium. In order to examine if the observed relative cost invariance held for datasets available elsewhere, additional datasets were obtained from the literature [303]. The results from these datasets (not shown) were comparable to those from our laboratory and did not alter the overall findings of this study.

Cost comparisons

The means and coefficients of variation from each computation of metabolic costs were compared. The variance in the results among the 400 random simulations was essentially negligible (all coefficients of variation $< 1\%$). The 49 simulations from expression data exhibited slightly higher variation (the average CV for the amino acid demands was 3.6%), but the coefficient of variation reached no higher than 10% (for the tryptophan cost). There was not a statistically significant difference in the costs for any of the amino acids or nucleotides resulting from randomly distributed mRNA concentrations or data-based simulations. The mean protein length was about 40 amino acids shorter for the data-based simulations than would be expected if the mRNA distribution were random. The highest CVs for the data-based cost calculations were for tryptophan (10.0%), cysteine (8.6%), and lysine (7.3%); and the lowest were for isoleucine (1.3%), threonine (1.3%), and asparagine (1.4%). The amino acid composition of a related strain of *E. coli* (B/r) has been experimentally determined [200], and the calculated costs for *E. coli* K-12 correlate relatively well with these biomass data (results not shown).

4.2.2 Distribution of estimated effective promoter strengths

Using global mRNA half-life data [25], we calculated the effective promoter strengths, q_i, for each of the 49 sets of mRNA concentrations estimated from expression data (which includes expression data from a variety of experimental conditions). The mean effective promoter strength and the corresponding coefficient of variation (CV) were plotted for each gene of the 3817 genes for which both expression data and half-life data were available (panel A of Figure 4.1). (Refer to the caption of Table 4.1 for the parameters used in the calculation of promoter strengths.) Here, the CV can be thought of as a measure of the extent to which a gene is subject to regulation under the experimental conditions tested. The highest expression levels generally corresponded to ribosomal protein components and associated protein synthesis enzymes, structural proteins, and membrane pore proteins (as classified according to [281]). Although the majority of CVs (60.9% of the 3817 mean effective promoter strengths) fall between 50% and 100%, 115 genes have standard deviations that are equal to or greater than double their average expression level. Over one-fifth of the genes (876, or 22.9%) had CVs of less than 50% (panel A of Figure 4.2).

If the genes known to take part in metabolism [76] are considered separately (panel B of Figure 4.2), their CVs (81.9%, on the average) are comparable to the average CV for the 3817 genes (78.2%). The average expression of the metabolic genes (891 $\mathrm{M}^{-1}\mathrm{s}^{-1}$), however, is significantly higher than that of the average gene (632 $\mathrm{M}^{-1}\mathrm{s}^{-1}$). The mean effective promoter strengths and CVs of genes implicated in regulation [260] are roughly equivalent to those of the overall genome (mean $q = 559\ \mathrm{M}^{-1}\mathrm{s}^{-1}$, mean $CV = 79.5\%$) (panel C of Figure 4.2).

4.2.3 Genome-location dependent patterns in gene expression

In order to elucidate potential genome-location dependent patterns in gene expression, wavelet transforms were applied to the effective promoter strength data as described above. Sliding averages of the calculated effective promoter strengths using Savitzky-Golay smoothing (panels B and C of Figure 4.1) indicate a non-random genome-location dependent variability along the *E. coli* chromosome. In particular, there appears to be a periodic large-scale pattern of regions with high average expression. This pattern is present in both the datasets generated from Affymetrix and from spotted array experiments, thus implying that the observed pattern is not likely to be an artifact of the experimental platform (refer to blue and red lines in Figure 4.1). In order to elucidate this pattern—in addition to other more subtle spatial patterns in the data—continuous wavelet and Fourier transforms were applied to the effective promoter strength data. The continuous wavelet transform of the average effective promoter strengths estimated from the 20 Affymetrix GeneChip experiments performed in this study (using the Morlet wavelet [21]) was represented in a scalogram (panel A of Figure 4.3). The

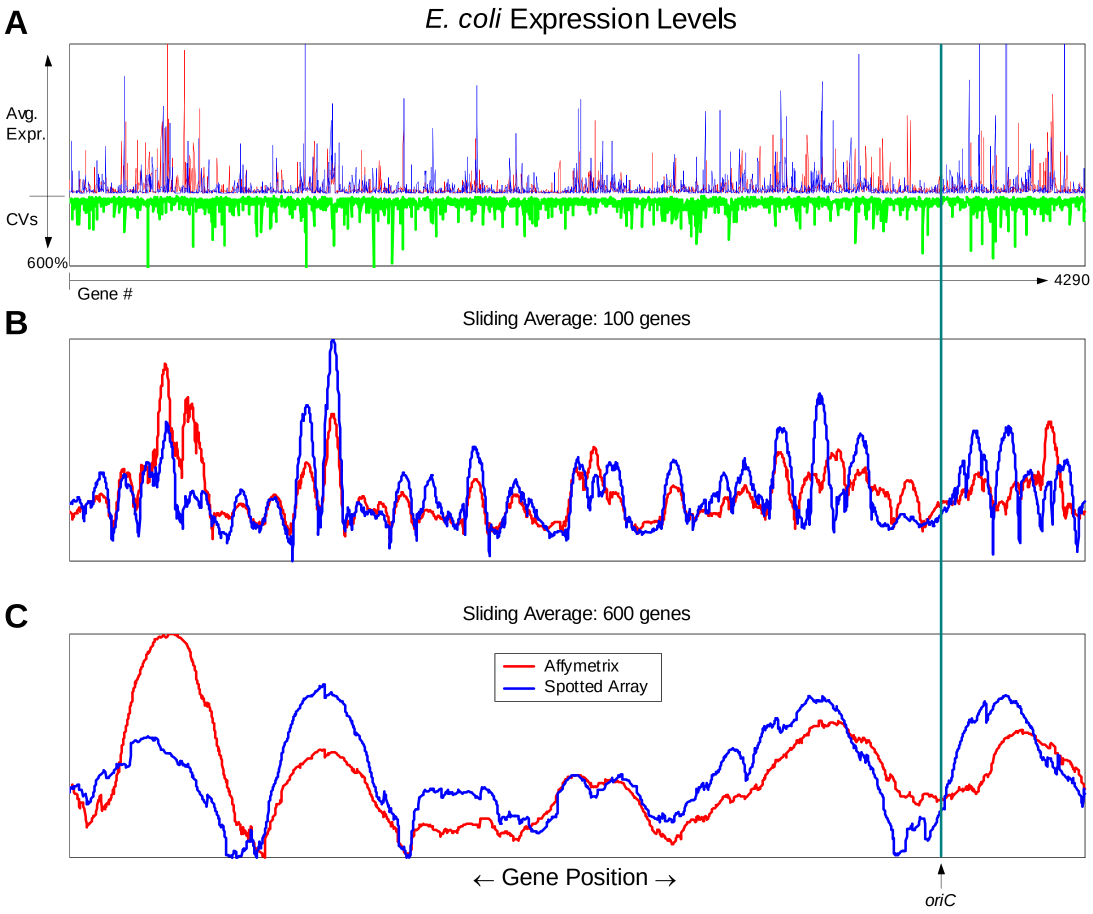


Figure 4.1: Calculated average effective promoter strengths at different sliding average scales. The units for the effective promoter strengths are in $M^{-1}s^{-1}$. The cellular parameters were chosen for a doubling time of 40 minutes (refer to the caption of Table 1), with an RNAP concentration of 1.456×10^{-6} M [31]. The concentration of each promoter, $[P]_i$, was chosen based on a C period of 45 minutes and a D period of 25 minutes [31]. The location of the origin of replication (*oriC*) is indicated for reference. (A) Plots of mean expression levels and coefficients of variation (CVs) for the 20 Affymetrix datasets and the 29 spotted array datasets. The blue bars represent the mean effective promoter strengths (q_i's) calculated from experiments performed using Affymetrix arrays, the red bars represent those from spotted array experiments, and the green bars represent the CVs spanning all 49 datasets used in the calculations. (B) Plots of mean expression levels over a sliding average (with 2nd-order Savitzky-Golay smoothing) of 100 genes for the Affymetrix (blue) and spotted array (red) datasets. (C) Same as panel B, but the sliding average is taken over a 600-gene window.

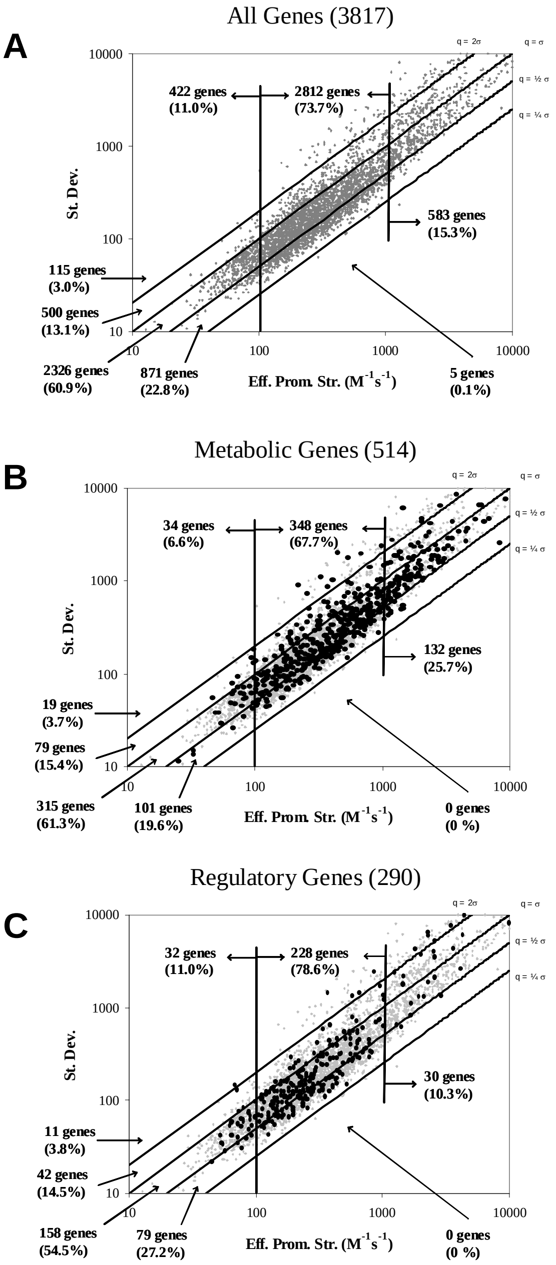


Figure 4.2: Log-log plots of the standard deviations vs. mean effective promoter strengths (q_i's) for individual ORFs in 49 expression datasets. The labels exterior to each plot indicate numbers of genes between each coefficient of variation demarcation, and the labels inside each plot denote numbers of genes whose promoter strengths are less than 100, between 100 and 1000, and greater than 1000 $M^{-1}s^{-1}$. (A) Plot of all 3817 genes for which effective promoter strengths were calculated. (B) Overlay of 514 metabolic genes [76]. (C) Overlay of 290 regulatory genes [260].

major feature of the transform was the clear periodic pattern at a scale of approximately 600 kb. This pattern was observed in the spotted array datasets and was also detected using other types of wavelet filters such as the Marr wavelet used in [198], indicating that the observed pattern is not an artifact due to either the experimental platform or the particular transform used (results not shown).

In the cross-section of the scalogram at the scale of 610 kb (panel B of Figure 4.3), the regular periodic pattern extending over almost the entire length of the genome was readily observed. The same periodic component identified through wavelet analysis can also be identified as a peak in the Fourier spectrum (panel C of Figure 4.3) at a period of approximately 600 bp. However, the periodic pattern does not extend in a regular fashion throughout the whole genome, making standard Fourier analysis somewhat less suitable for this study than wavelet analysis.

The observed periodic pattern appeared in all the individual effective promoter strength datasets computed using different expression profiles and hence does not seem to be specific to any particular experimental condition. No such pattern was observed in the raw mRNA half-life data. A periodic pattern was, however, detected in the raw gene expression data (not shown), but the pattern was somewhat less well defined than in the effective promoter strength data. Since the effective promoter strengths have been corrected for differential mRNA decay rates and distance from the replication origin, they would seem to be a more appropriate measure of the actual transcription rate than mRNA expression data alone.

Analysis of gene functional classes whose members are preferentially located in particular regions of high or low average expression within the periodic pattern (Fig. 3b) may elucidate the relationship between the observed periodicity and *E. coli* cellular function. Flagellar and other cell motility related genes and genes encoding ribosomal and other translation-related proteins are preferentially located in one or more of the high expression regions. On the other hand, genes involved in major metabolic functions such as energy metabolism, carbon utiliza-

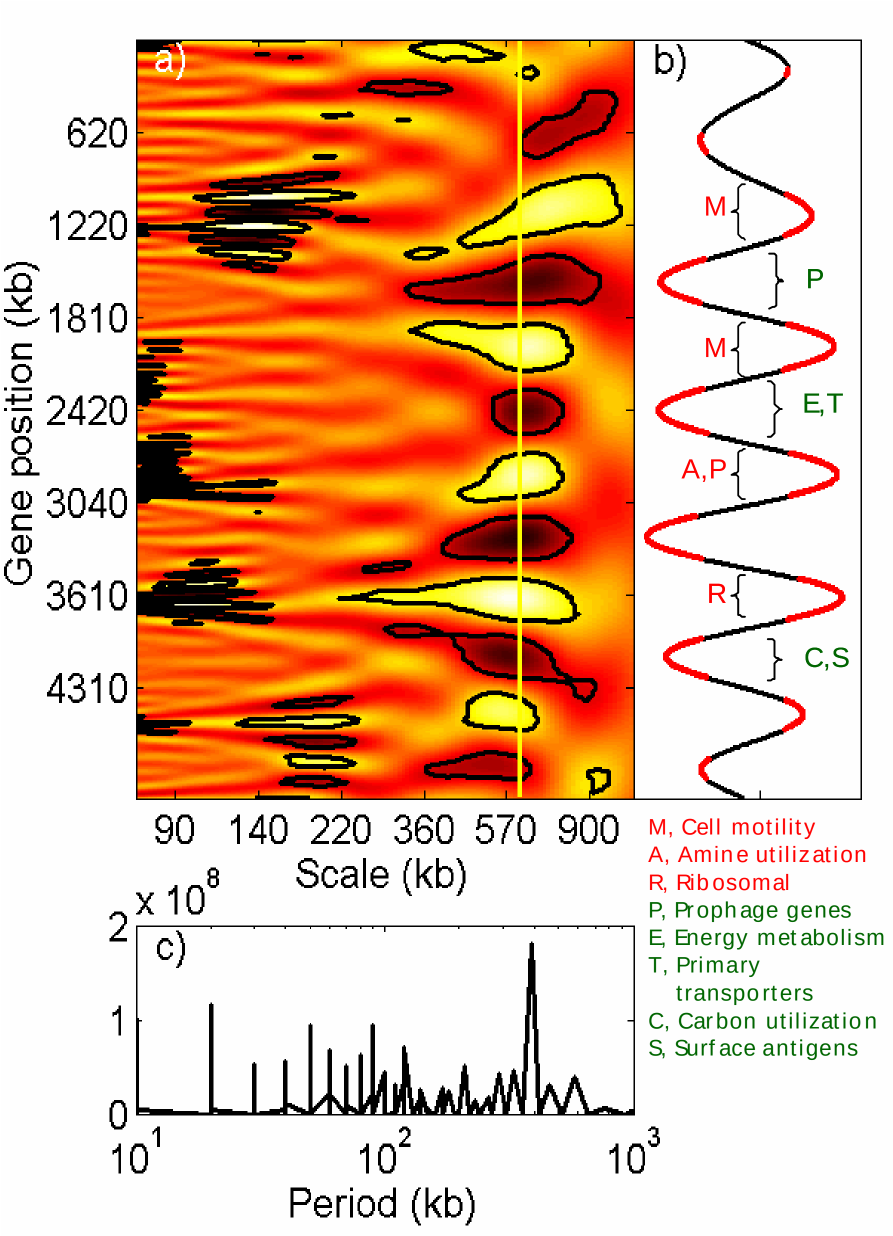


Figure 4.3: Spatial variability of gene expression along the *E. coli* genome studied using continuous wavelet and Fourier transforms of the effective promoter strength data. (A) Scalogram of the wavelet transform with gene position on the y-axis and transform scale on the x-axis. Lighter/darker regions correspond to higher/lower values of the coefficients. The regions encircled by black contour lines are deemed to be statistically significant patterns compared to spatially randomized effective promoter strengths ($P < 0.001$). (B) The cross-section of the wavelet scalogram in panel a) at the scale of 610 kb. The regions with significantly non-random wavelet coefficients are marked in red. Gene functional classes (classified according to GenProtEC [281]) preferentially located in particular high (red) or low (green) expression regions (hypergeometric $P < 0.001/(\text{number of functional classes})$) are also indicated. (C) Fourier transform of the effective promoter strength data. The only significant peak in the transform occurs at the period of approximately 600 kb.

tion, and transport tend to be located in the low expression regions. Furthermore, genes in certain functional classes are typically strongly enriched in only one or two of the high or low expression regions, indicating potentially distinct roles for each of these regions. Note that the only data generated were for protein-coding ORFs. Thus, the rRNA and tRNA transcription rates were not considered in the analysis of genome-location dependent patterns.

4.3 Discussion and implications

We have performed an integrated analysis of genome-scale gene expression in *E. coli*, based on simultaneous use of sequence data, gene expression data, and mRNA half-life data. The results from this integrative analysis are three-fold: 1) The relative material and energy costs used to express the *E. coli* genome are essentially independent of the distribution in mRNA concentrations; 2) The distribution of the effective promoter strengths was examined for 49 gene expression datasets, revealing that over 16% of the genes in *E. coli* vary in expression by more than 100% of the average promoter strengths under the conditions measured; and 3) A wavelet analysis of these distributions revealed a large-scale ($\sim$ 600 kb) periodic pattern in the expression of genes in *E. coli*. The methods used were computationally simple, and thus suitable for immediate integration into existing genome-scale metabolic models of *E. coli* [76, 245].

The apparent invariance of the costs for maintaining any expression state of the genome implies that the metabolic resources required to maintain a particular transcription and proteomic state are relatively constant and independent of external conditions. This invariance will not hold true, however, if a gene or small subset of genes with atypical amino acid composition is expressed at levels that are orders of magnitude higher than the rest of the genes (calculations not shown). Thus microbes genetically engineered to express a particular protein at a high level may experience significant phenotypic effects associated with the cost imposed by such atypical expression. It is also possible that the dynamic range of microarrays and genechips becomes limiting if a few transcripts are expressed at a very high level and therefore saturate the signal on the arrays [37, 251]. To test the significance of this effect, the cost simulations were performed in which the top 0.1% of genes with the highest expression levels were assigned copies/cell values that were 10% higher than the level reported by the arrays. The highest CV was raised to just over 20% (for tryptophan), while the average CV of the amino acid costs increased from 3.6% to 8.1%, thus suggesting that a limited dynamic range in the experimental technology could have some effect on the calculated costs. Finally, it is possible that the observed invariance may be due to a lack of probing the experimental conditions that would most alter the relative amino acid costs required for expression. However, the conditions chosen were quite varied in nature, and hence one would expect there to be differences in the overall metabolic costs between these conditions if such differences exist at all.

The variation in effective promoter strength was computed for the entire genome. In general, no clear patterns were found between gene category and variation in expression level. There was also no observed functional class bias in either the effective promoter strengths or in the variance across 49 different calculations. It is worth noting that these computations will be biased by the experimental conditions under which each expression profile was measured. To better ascertain genes that are subject to regulation, it will be necessary to test more varied growth conditions (e.g., growth on other carbon sources, anaerobic growth, growth during diauxic shifts, etc.). If M9 medium (which contains a relatively high amount of phosphate) were used instead of MOPS medium, for example, one might expect the genes involved in the phosphate regulon to exhibit altered effective promoter strengths (and, consequently, increased CVs in the subsequent analysis), thus revealing the extent to which those particular genes were differentially regulated under the changing media conditions [327]. As more datasets are included in this type of integrated analysis, a better gauge of the variability in gene expression will be obtained, thus more completely revealing the extent to which each gene is subject to regulation.

An approximately 600 kb periodic genome-location dependent pattern in gene expression in the *E. coli* genome was detected using wavelet analysis of the effective promoter strength data generated in this study. The origin and significance of this pattern, however, is not clear. One possible explanation for the observed pattern is the existence of topological domains with potentially different levels of supercoiling in the *E. coli* chromosome [289]. It has been estimated that there are 43(±10) of such domains so that the average domain size would be approximately 100 kb [289]. No significant 100 kb periodicity was detected in the wavelet analysis except for particular localized patterns (Fig. 3a), although an irregular periodicity at a sliding average of 100 genes ($\sim$ 100kb) was observed (Fig. 1b). As the 600 kb periodicity corresponds to a multiple of the 100 kb topological domain scale, it is possible that the potential differences in gene expression in different topological domains indeed explain the observed pattern. However, the nature and locations of the topological domain boundaries are not known in the *E. coli* genome [47, 226, 342], making comparisons of the topological domain structure to the observed periodicity in expression challenging. Even if the origin of the periodic expression pattern is somewhat obscure, there is a clear tendency of genes in certain functional classes to cluster in either the high or low expression regions within this pattern (Fig. 3b). If the periodic pattern and the corresponding functional class clusters continue to be observed as more datasets are generated, this tendency may suggest how a genome-location dependent constraint on gene expression could act to shape gene order in genomes.

As genome-scale data—including mRNA expression data, mRNA half-lives, and proteomic data—are becoming more widely available, the need for integrating these heterogeneous data types is becoming stronger [218]. As this study demonstrates, higher-order biological analysis can be performed based upon the integration of multiple data types that cannot be done based on the analysis of individual datasets. Such integrated data analysis is enabled by genome-scale *in silico* models. Different data types demand a model to explicitly relate their values, thus revealing emergent properties that would otherwise be inaccessible [125].

The proposed model integrates three types of genome-scale data: sequence, gene expression data, and mRNA half-life data. This structured framework constitutes a novel means by which to analyze expression data and interpret the expression state of a cell. The scalability of the methods used to generate these results should greatly facilitate the future integration of genomic expression state with existing genome-scale metabolic models. This method therefore constitutes an important step in our progress towards achieving truly genome-scale integrated models of cellular function.

The text of this chapter, in full, is a reprint of the material as it appears in T.E. Allen, M.J. Herrgård, L. Mingzhu, Y. Qiu, J.D. Glasner, F.R. Blattner, and B.O. Palsson. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *Journal of Bacteriology*, 185:6392-6399. I was the primary author of this publication, and the co-authors participated and supervised the research which forms the basis for this chapter.

Chapter 5

Sensitivity analysis of translational efficiency with respect to codon usage and tRNA availability

As described in Chapter 2, during the process of translation, most amino acids can be specified using multiple synonymous codons that bind particular tRNA species. The binding stoichiometry between the tRNA species and their cognate codons is not a trivial one-to-one relationship, as some tRNA species are able to bind to more than one cognate codon due to "wobble" in the third nucleotide position [56]. (Refer to Table 2.2 for a list tRNA species and their cognate codons in *Escherichia coli.* The possible complexities in tRNA-codon binding stoichiometry are illustrated schematically in Figure 5.1 for the amino acid threonine.) A measure known as codon adaptation index (CAI) was devised to quantify the extent to which synonymous codon usage in a given gene reflects that of a reference set of highly-expressed genes [282]. This quantity has often been used as a predictor for gene expression since there is typically a correlation between high CAI and high mRNA expression [69, 109] and protein expression [326]. The choice of synonymous codons is not random, and the existence of organism-specific codon preferences is widely accepted [1]. Studies of biased codon usage in multiple sequenced genomes suggest that genome-wide mutational processes act as the primary constraint on codon usage [42], although there is considerable variability among organisms [283].

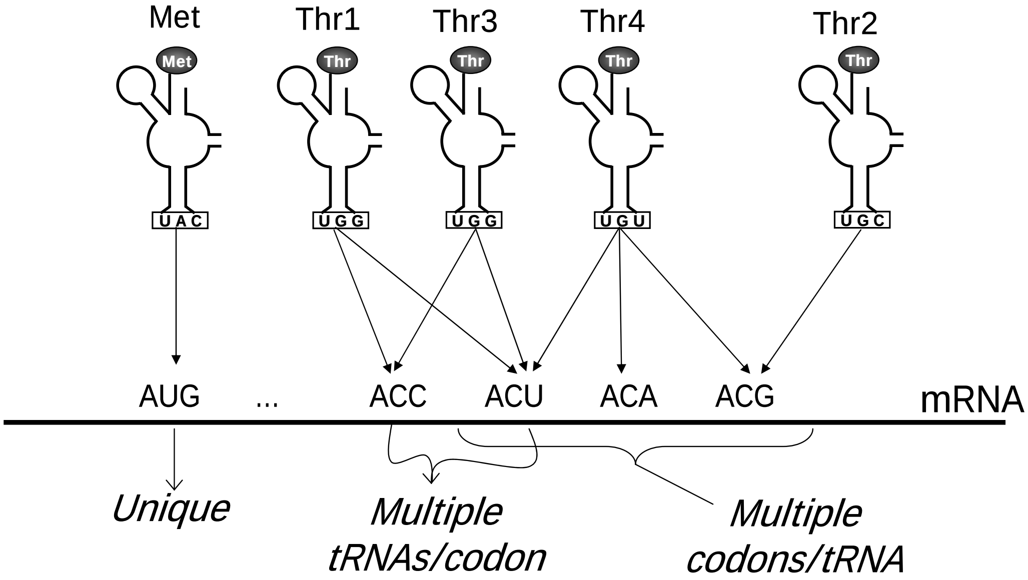


Figure 5.1: Stoichiometry for threonine tRNA-codon binding.

The synonymous codon usage of protein-coding genes in *E. coli* has been shown to correlate with measured *in vivo* abundances of the corresponding isoacceptor tRNA species [145, 68], suggesting the existence of an additional selection pressure on codon usage for the optimization of translational efficiency [23, 291]. Subsequent analysis of codon usage in multiple unicellular organisms has revealed an evolutionary constraint imposed by tRNA contents and translational efficiency on synonymous codon usage [162, 256]. Measurements of *in vivo* and *in vitro* translation rates have indicated that the availability of tRNA species is the rate-limiting step in elongation during protein synthesis and hence determines translational efficiency [313, 300, 292]. Furthermore, protein expression has been shown to be more closely related to codon bias than to the identity of the stop codon or the translation initiation strength associated with a particular mRNA Shine-Dalgarno sequence [183]. These findings suggest that the rate at which a given mRNA transcript can be translated (i.e. the translational efficiency) is limited by the supply-demand ratio of specific tRNA species to their cognate codons [291].

The correlation of synonymous codon choice with translational efficiency motivates an assessment of the range in efficiencies achievable simply by altering codon usage, as well as the optimality of the synonymous codon allocation for each of the wild-type genes in *E. coli* given measured tRNA abundances [145, 68]. Previous studies have been limited almost exclusively to genomic sequence data, including an analysis of bias in first codon positions in genes [119] and of relating the metabolic efficiency of organisms to their amino acid compositions by estimating proteomic profiles directly from codon usage [2]. Other studies have used measured tRNA abundances to predict the optimality of codon usage [23, 291] or of tRNA abundances [256] in a number of microbial organisms. In such analyses of codon-tRNA "optimality," however, much of the literature to date has adopted the notion of an "optimal codon" among each set of synonymous codons [99], due to the known optimality of binding efficiencies between specific codons and tRNAs [117] and the strong correlation between tRNA availability and codon usage [145, 162]. However, such an assumption neglects the need to optimize the supply-demand ratio of each tRNA species to its cognate codon(s).

In the present study, the translational efficiency for a given gene has been defined based upon the supply-demand ratios of the tRNA species to their cognate codons. I then computed the codon distributions which minimize or maximize translational efficiency in *E. coli* and examined the efficiency of the wild-type codon distribution for each gene in *E. coli*. Finally, under simplied assumptions I assessed how changes in the wild-type tRNA abundances will affect the optimality of the codon choice for each gene in *E. coli*.

5.1 Methods for modeling translational efficiency

5.1.1 Modeling framework and assumptions

Basic flux-balance framework In this study, the translational efficiency is defined as the steady-state rate at which each protein in a bacterium can be synthesized given ribosomal and tRNA abundance limitations. A mass balance can be written for every protein molecule in the cell:

$$\frac{d[\mathrm{Pr}]_i}{dt} = v_i - (\mu + k_i)[\mathrm{Pr}]_i, \tag{5.1}$$

where $[\mathrm{Pr}]_i$ is the concentration of the i^{th} protein, v_i is the corresponding translation rate for that protein, μ is the growth rate of the cell, and k_i is the degradation rate of the i^{th} protein. At steady state (i.e. $d[\mathrm{Pr}]_i/dt = 0$), the protein synthesis rate must balance the rate of dilution. Solving for the protein concentration yields:

$$[\mathrm{Pr}]_i = \frac{v_i}{\mu + k_i} \tag{5.2}$$

The protein degradation rates will not be considered for the remainder of this study since the vast majority of proteins are not subject to proteolysis in growing cells [194]. For the small number of actively degraded proteins in growing cells (e.g. some regulatory proteins), this degradation term may be accounted for [114].

Assumptions In addition to the steady-state flux balance assumption given above, I have made a number of assumptions regarding the process of translation. These assumptions have been based on experimental results and sequence-based analysis of codon usage in bacterial organisms.

Translation flux: The translation flux for each protein (v_i from the above mass-balance equation) us assumed to be equal to the number of corresponding mRNA molecules per cell times the translational yield (defined as the number of ribosomes allocated to each mRNA) times the maximum elongation rate per ribosome divided by the length of the protein. This expression for translation flux is presented, with key assumptions, in Figure 5.2.

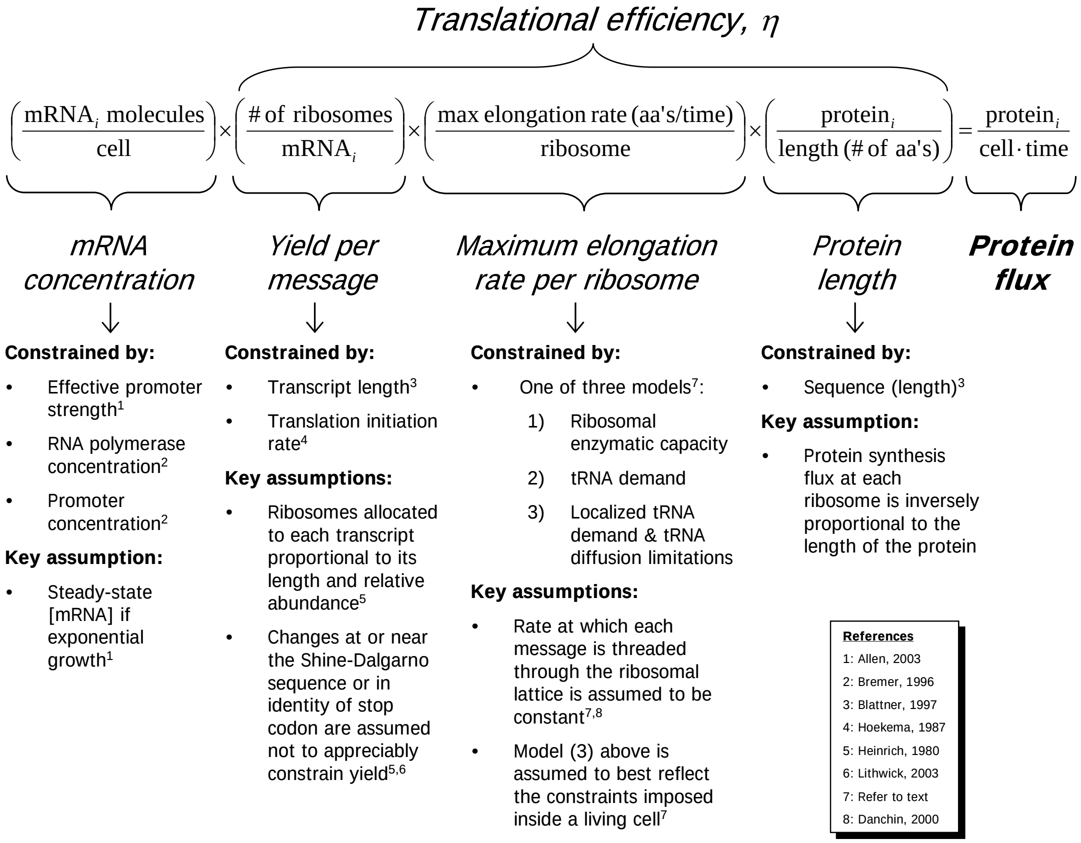

Translational efficiency, η
$\left(\frac{\text{mRNA}_i \text{ molecules}}{\text{cell}}\right) \times \left(\frac{\text{\# of ribosomes}}{\text{mRNA}_i}\right) \times \left(\frac{\text{max elongation rate (aa's/time)}}{\text{ribosome}}\right) \times \left(\frac{\text{protein}_i}{\text{length (\# of aa's)}}\right) = \frac{\text{protein}_i}{\text{cell} \cdot \text{time}}$
mRNA concentration
Constrained by:
• Effective promoter strength[1]
• RNA polymerase concentration[2]
• Promoter concentration[2]
Key assumption:
• Steady-state [mRNA] if exponential growth[1]
Yield per message
Constrained by:
• Transcript length[3]
• Translation initiation rate[4]
Key assumptions:
• Ribosomes allocated to each transcript proportional to its length and relative abundance[5]
• Changes at or near the Shine-Dalgarno sequence or in identity of stop codon are assumed not to appreciably constrain yield[5,6]
Maximum elongation rate per ribosome
Constrained by:
• One of three models[7]:
1) Ribosomal enzymatic capacity
2) tRNA demand
3) Localized tRNA demand & tRNA diffusion limitations
Key assumptions:
• Rate at which each message is threaded through the ribosomal lattice is assumed to be constant[7,8]
• Model (3) above is assumed to best reflect the constraints imposed inside a living cell[7]
Protein length
Constrained by:
• Sequence (length)[3]
Key assumption:
• Protein synthesis flux at each ribosome is inversely proportional to the length of the protein
Protein flux
References
1: Allen, 2003
2: Bremer, 1996
3: Blattner, 1997
4: Hoekema, 1987
5: Heinrich, 1980
6: Lithwick, 2003
7: Refer to text
8: Danchin, 2000

Figure 5.2: Calculation of the translation flux in bacteria. The product of the yield per message, maximum elongation rate, and inverse protein length is taken to be η, the translational efficiency.

Expression for translational efficiency: The translational efficiency is equal to the three terms on the right of the lefthand side of the equation in Figure 5.2.

Translational yield proportional to transcript length: The yield per mRNA molecule is assumed to be proportional to the length of the transcript, since longer transcripts will be occupied by more ribosomes [126, 148]. Variation at or near the Shine-Dalgarno sequence or in the identity of the stop codon are assumed not to significantly affect the yield [126, 183]. The translation initiation rate [130] is also assumed to be constant for each gene.

Parallelized model: Since the yield has been assumed to be proportional to transcript length, this term will essentially cancel out the length term. Thus, the translational efficiency will be equal to the translation elongation rate. I have thus assumed that translation is a parallel process in bacteria, rather than a serial process as it is often depicted (Figure 5.3).

Elongation rate depends on tRNA/codon supply-demand ratios: Several experimental studies have demonstrated that translation elongation depends upon synonymous codon usage [313, 225, 292, 293]. This elongation rate is assumed to be equal to the limiting ratio of tRNAs to the codons to which they are bound (see below).

Neglect differential binding affinities and selective tRNA charging: The present model does not take into account differential binding affinities [116, 58] for tRNA species with "wobble" (i.e. non Watson-Crick) base-pair matching because they have not been comprehensively measured. The key results shown in this study do not change if estimated binding parameters are incorporated (not shown). Selective tRNA charging [79] has also been neglected in this model since this phenomenon is specific to conditions in which particular amino acids are growth limited.

Neglect codon context and secondary structure effects: Potential effects of codon context [24] and secondary structural features of the mRNA transcripts [211] have been ignored. Available data do not indicate that there is any effect of mRNA secondary structure on translation rate [292].

The assumption regarding the dependence of elongation rate upon the local supply/demand ratios of tRNA species to their cognate codons is very important, for it deviates from the view of the cell as a well-mixed bag of dilute solutes ([80]; also see [111] for illustrations). The assumption that tRNA molecules are spatially confined (localized in the cell) has been discussed in the literature [59], and enables the first step towards the consideration of spatial constraints in genome-scale *in silico* models of microorganisms [235, 248]. More on such spatial aspects of the bacterial cell will be dealt with in Chapter 6 and the subsequent chapters.

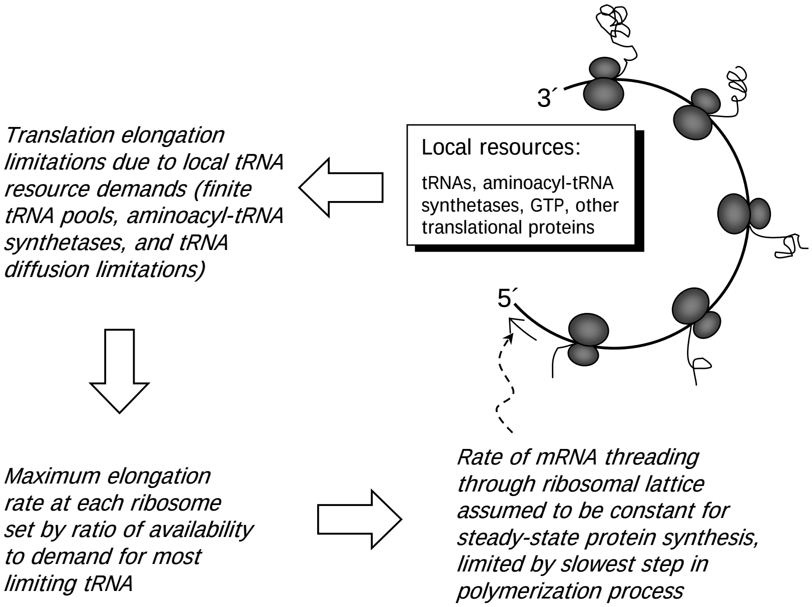

Translation elongation limitations due to local tRNA resource demands (finite tRNA pools, aminoacyl-tRNA synthetases, and tRNA diffusion limitations)
Local resources:
tRNAs, aminoacyl-tRNA synthetases, GTP, other translational proteins
3´
5´
Maximum elongation rate at each ribosome set by ratio of availability to demand for most limiting tRNA
Rate of mRNA threading through ribosomal lattice assumed to be constant for steady-state protein synthesis, limited by slowest step in polymerization process

Figure 5.3: Schematic depicting bacterial translation as a parallel process. In this model, a given mRNA transcript is threaded through a more or less fixed "factory" of ribosomes which draw upon local pools of tRNA molecules [59]. Since the number of ribosomes bound is assumed to be proportional to the transcript length, the overall protein synthesis flux will be independent of length [126].

Additional assumptions are that the initiating N-formylmethionine, the nonstandard amino acid selenocysteine, and the stop codons are not limiting in this process. It is also possible that the overall abundance of ribosomes may be limiting at very high growth rates [321, 173], but this is unlikely to be a constraint in wild-type *E. coli* in the absence of any high-copy number plasmids.

5.1.2 Calculating translation rates

The kinetics of translation initiation, elongation, and termination constitute a complex enzyme system [186]. Since the kinetic parameters are largely unknown for every binding event that occurs during translation, a lumped model that approximates these reactions yet still captures the overall behavior of protein synthesis is desirable (Chapter 3; see also [3]). In Chapter 4, I presented a simplistic model for determining the amino acid and energy demands for protein synthesis based solely upon the sequence, the availability of ribosomes, and the relative abundance of each mRNA transcript [4]. The translation rates in this simplified model were written as

$$v_i = \beta_{\mathrm{rib}} m_{\mathrm{rel},i} / a_i, \tag{5.3}$$

where $\beta_{\mathrm{rib}} \approx 300,000$ peptide bonds formed per cell per second (assuming typical *E. coli* parameters [200]), $m_{\mathrm{rel},i}$ is the relative abundance of each mRNA transcript ($m_{\mathrm{rel},i} = [\mathrm{mRNA}]_i / \sum_j [\mathrm{mRNA}]_j$), and a_i is the number of amino acids (i.e. one more than the number of peptide bonds required) in the finished protein. This model is based upon the assumption that each mRNA molecule competes equally for a finite pool of ribosomes, so that the number of ribosomes allocated to a particular mRNA transcript is proportional to the relative abundance of that transcript. Any competitive advantage due to transcript length (i.e. resulting in increased ribosomal transit time) has been ignored, since the variability in transcript abundance (about 3-fold on average, calculations not shown) typically outweighs the variability in transcript length (0.65-fold [26]).

Serial model of translation The simple model of translation presented above and proposed previously [4], however, neglects limitations imposed by finite tRNA abundances [313, 292]. If x_j is the fraction of tRNAs that binds codon j, the maximal rate at which each peptide bond can be formed will be equal to the fraction of tRNAs available for that particular step in elongation: $v_{\text{elongation step for codon } j} = \beta_{\text{rib}} m_{\text{rel},i} x_j$. If protein synthesis were a strictly serial process, the overall synthesis rate of each protein would be

$$v_i = \beta_{\text{rib}} m_{\text{rel},i} \left[\sum_{j=1}^{61} \frac{c_j}{x_j} \right]^{-1}, \tag{5.4}$$

where c_j denotes the number of times codon j (of a total of 61 sense, or non-stop, codons) appears in the transcript, mRNA$_i$. The overall rate is thus equal to the reciprocal of the sum of the reciprocals of the rate of the elongation steps corresponding to each type of codon, where the translation rate corresponding to each set of codons is x_j/c_j.

Parallel model of translation The serial model of translation elongation, however, is likely only valid at the near-zero growth rates which occur during starvation due to the dearth of ribosomes [31]. A more realistic model takes into account the partially parallel nature of protein synthesis due to multiple ribosomes binding simultaneously to each mRNA transcript. In a revision of the model given in Equation 5.4 above, the synthesis rate of each protein is more appropriately constrained by the most limiting set of tRNAs—i.e. the set of codons with the minimum ratio of the fractional bound quantities of the corresponding tRNAs to the number of times that the cognate codons appear in the transcript:

$$v_i = \beta_{\text{rib}} m_{\text{rel},i} \left(\frac{x_{\text{limiting codon},i}}{c_{\text{limiting codon},i}} \right), \quad \text{limiting codon} = \min \left(\frac{x_j}{c_j} \right) \tag{5.5}$$

In the remainder of this analysis, for each gene we can take β_{rib} and $m_{\text{rel},i}$ as constants. Thus, the translational efficiency, η, becomes the quantity of interest:

$$\eta_i = \min_j \left(\frac{x_j}{c_{j,i}} \right) = \frac{x_{\text{limiting codon},i}}{c_{\text{limiting codon},i}}, \tag{5.6}$$

where the limiting codon is the codon j that yields the minimum ratio, $x_j/c_{j,i}$ for protein i. In other words, the protein translation rate will be limited by the particular codon for which the ratio of the corresponding tRNA binding availability to the demand for that tRNA—as dictated by the number of times its cognate codon(s) appears in the mRNA transcript—is minimized. It is thus assumed that the protein synthesis rate is limited by the set of anticodon-codon binding events which are most constrained by the relative abundances of the tRNA pools localized at each (i^{th}) gene.

The matrix of tRNA-codon binding values for a given set of synonymous codons, $\mathbf{X}$, can be calculated from the vector of relative tRNA abundances for each amino acid and the number of each type of synonymous codon corresponding to each amino acid. Let us define the $\mathbf{B}$ matrix as the "binding" matrix describing the stoichiometry between tRNA species and their cognate codons. As an example, consider the case of threonine illustrated in Figure 5.1. This amino acid has four different tRNA species which read four different synonymous codons. For this example system the corresponding $\mathbf{B}$ matrix is as follows:

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix}, \qquad \mathbf{c} = [c_1 \quad c_2 \quad c_3 \quad c_4]$$

The rows in $\mathbf{B}$ correspond to the four tRNA species for threonine in *E. coli* (Thr1, Thr2, Thr3, and Thr4), and the columns correspond to the four threonine codons (ACA, ACC, ACG, and ACU). The allocation matrix of each tRNA species to each codon will depend upon the relative tRNA abundances and the number of each synonymous codon in the gene (i.e. the $\mathbf{t}$ and $\mathbf{c}$ vectors, respectively). First, let us define a distribution vector, d_i:

$$d_i = \frac{t_i}{\sum_j (B_{i,j} c_j)},$$

from which a distribution matrix (given $\mathbf{B}$) can be computed:

$$D_{i,j} = d_i B_{i,j}.$$

In order to compute the allocation matrix, **X**, from this distribution, two intermediate matrices are required:

$$DF_{i,j} = \frac{D_{i,j}}{\sum_k (D_{k,j})}, \qquad DG_{i,j} = \frac{DF_{i,j} c_j}{\sum_k (DF_{i,k} c_k)}.$$

Let $\mathbf{T}_{\text{diag}} = \mathbf{It}$, which is a diagonal matrix formed by multiplying the identity matrix, **I** with the tRNA abundance vector, **t**. The tRNA-codon allocation matrix, in which $X_{i,j}$ represents the amount of tRNA i bound to codon j, is then given by:

$$X_{i,j} = T_{\text{diag}_{i,j}} DG_{i,j}. \tag{5.7}$$

Given this tRNA-codon allocation matrix (**X**), the efficiency for this set of codons and amino acids (i.e. this particular amino acid, aa) is given by:

$$\eta_{\text{aa}} = \min \left(\sum_k X_{k,j} / c_j \right) \tag{5.8}$$

In order to compute the efficiency for a given gene, i, the efficiency for each set of synonymous codons must be computed when the number of amino acids of a particular type is greater than zero:

$$\eta_i = \min \left(\eta_{\text{aa}_1}, \eta_{\text{aa}_2}, \eta_{\text{aa}_3}, \ldots, \eta_{\text{aa}_{20}} \right), \qquad \forall\ N_{\text{aa}} > 0, \tag{5.9}$$

where N_{aa} is the number of amino acids of each type appearing in the given protein. The codon-tRNA binding stoichiometry and relative tRNA abundances for each amino acid in *E. coli* are provided in Table 5.1.

Although 32 rate constants for the association of each ternary complex to several of the cognate codon(s) have been estimated from a model assuming growth rate-optimized tRNA abundance and codon usage [23], only a limited number of these binding parameters have actually been experimentally measured [58]. Thus, the expression for translational efficiency presented in Equation 5.6 does not take differential binding affinities into account. In the event that these data become more completely available, each **B** matrix can be adjusted to reflect known binding affinities, rather than the binary values used in this study.

Table 5.1: Stoichiometry of binding between tRNA species and cognate codons for each amino acid in *E. coli.* The relative abundances for each tRNA are indicated in the column to the left of each binding matrix ($B_{i,j}$ from above). Note that these relative abundances do not sum exactly to unity since the two initiating tRNAs (for N-formylmethionine) and the tRNA for the rare amino acid selenocysteine have not been included.

Alanine	Abundance	GCA	GCC	GCG	GCU
Ala1B	0.0506	1	0	1	1
Ala2	0.0096	0	1	0	0

Arginine	Abundance	AGA	AGG	CGA	CGC	CGG	CGU
Arg2	0.0739	0	0	1	1	0	1
Arg3	0.0099	0	0	0	0	1	0
Arg4	0.0135	1	0	0	0	0	0
Arg5	0.0065	0	1	0	0	0	0

Asparagine	Abundance	AAC	AAU
Asn	0.0186	1	1

Aspartate	Abundance	GAC	GAU
Asp1	0.0373	1	1

Cysteine	Abundance	UGC	UGU
Cys	0.0247	1	1

Glutamine	Abundance	CAA	CAG
Gln1	0.0119	1	0
Gln2	0.0137	0	1

Glutamate	Abundance	GAA	GAG
Glu2	0.0734	1	1

Glycine	Abundance	GGA	GGC	GGG	GGU
Gly1	0.0133	0	0	1	0
Gly2	0.0199	1	0	1	0
Gly3	0.0678	0	1	0	1

Histidine	Abundance	CAC	CAU
His	0.0099	1	1

Isoleucine	Abundance	AUA	AUC	AUU
Ile1	0.0513	0	1	1
Ile2	0.0027	1	0	0

Leucine	Abundance	CUA	CUC	CUG	CUU	UUA	UUG
Leu1	0.0695	0	0	1	0	0	0
Leu2	0.0147	0	1	0	1	0	0
Leu3	0.0104	1	0	1	0	0	0
Leu4	0.0298	0	0	0	0	0	1
Leu5	0.0160	0	0	0	0	1	1

Lysine	Abundance	AAA	AAG
Lys	0.0299	1	1

Methionine	Abundance	AUG
Met	0.0110	1

Phenylalanine	Abundance	UUC	UUU
Phe	0.0161	1	1

Proline	Abundance	CCA	CCC	CCG	CCU
Pro1	0.0140	0	0	1	0
Pro2	0.0112	0	1	0	1
Pro3	0.0090	1	0	1	1

Serine	Abundance	AGC	AGU	UCA	UCC	UCG	UCU
Ser1	0.0202	0	0	1	0	1	1
Ser2	0.0054	0	0	0	0	1	0
Ser3	0.0219	1	1	0	0	0	0
Ser5	0.0119	0	0	0	1	0	1

Threonine	Abundance	ACA	ACC	ACG	ACU
Thr1	0.0016	0	1	0	1
Thr2	0.0084	0	0	1	0
Thr3	0.0170	0	1	0	1
Thr4	0.0143	1	0	1	1

Tryptophan	Abundance	UGG
Trp	0.0147	1

Tyrosine	Abundance	UAC	UAU
Tyr1	0.0120	1	1
Tyr2	0.0196	1	1

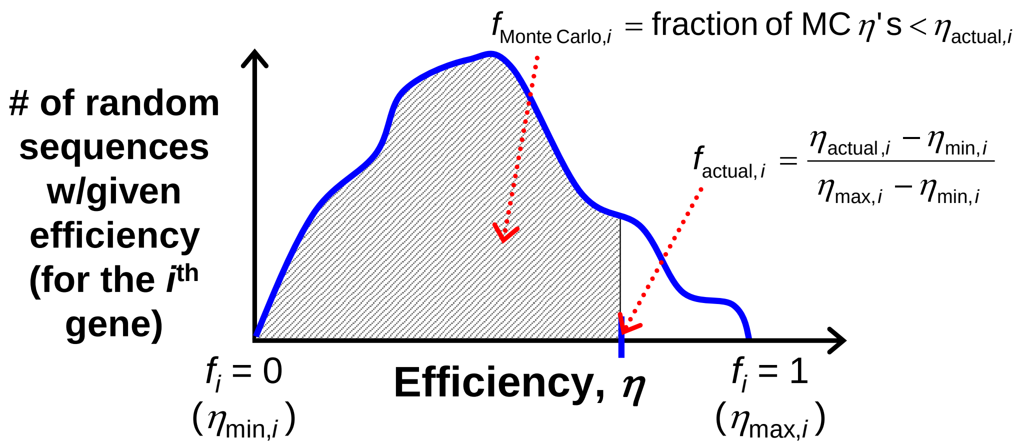
Valine	Abundance	GUA	GUC	GUG	GUU
Val1	0.0597	1	0	1	1
Val2A	0.0098	0	1	0	1
Val2B	0.0099	0	1	0	1

5.1.3 Finding optimal codon allocation schemes

Given a model for estimating the translational efficiency of a protein (Equations 5.6–5.9), and given the codon composition and the relative tRNA abundances, it is possible to probe the flexibility in protein synthesis (i.e. translational efficiency) that results from altering codon usage or by altering relative levels of tRNA species. The allowable range of synthesis rates for each protein can then be compared to those estimated for the actual sequenced organism.

Distribution of η_i's given random synonymous codon usage One means to probe the effect of synonymous codon usage on translational efficiency is to perform Monte Carlo simulations in which the synonymous codons for each gene are chosen at random, with uniform probability of choosing a given synonynous codons. If the codons are chosen in this manner (i.e. the only constraint on codon choice is to encode the required number of each type of amino acid in the protein), a histogram of efficiency values can be generated from the Monte Carlo simulations for each gene (refer to Figure 5.6 for examples). A histogram of these simulations for a hypothetical gene, i, is illustrated conceptually in Figure 5.4. Note that if one of these histograms only contains one efficiency value, the synthesis of that gene is completely inflexible with respect to altering its translational efficiency by changing codon usage (i.e. the range in possible efficiencies will be zero, as defined below).

Most genes (all but four: b0005, *hisL*, *pheL*, and *slyD*) have many possible translational efficiencies, especially genes that contain amino acids whose corresponding synonymous codons possess greater degeneracy. The size of this synonymous codon usage space is discussed in §5.2.1. The translational efficiencies of a large number (e.g. 1,000) of these Monte Carlo simulations will reveal bias in the synthesis spectrum for a particular protein. The results from the Monte Carlo simulations can be compared with the translational efficiency of the i^{th} wild-type *E. coli* protein ($\eta_{\text{actual},i}$) can be found by determining the fraction of simulated



$f_{\text{Monte Carlo},i}$ = fraction of MC η's < $\eta_{\text{actual},i}$
# of random sequences w/given efficiency (for the i^{th} gene)
$f_{\text{actual},i} = \frac{\eta_{\text{actual},i} - \eta_{\text{min},i}}{\eta_{\text{max},i} - \eta_{\text{min},i}}$
$f_i = 0$ $(\eta_{\text{min},i})$
Efficiency, η
$f_i = 1$ $(\eta_{\text{max},i})$

Figure 5.4: Conceptual illustration of probing translational efficiencies. For a hypothetical gene, i, the histogram shown represents the number of simulations exhibiting the translational efficiencies (η) indicated along the x-axis. The quantities η_{min}, η_{max}, $\eta_{\text{actual},i}$, $f_{\text{actual},i}$, and $f_{\text{Monte Carlo},i}$ are described in the text.

efficiencies that are less than the actual efficiency (Figure 5.4).

Minimizing translational efficiency The codon usage that yields the minimum translational efficiency for a given protein can be found by choosing the codon for each amino acid with the minimum $\sum_j (X_{i,j})$—i.e. the codon whose tRNA isoacceptors are least abundant. Choosing the codons in such a way will minimize the ratio $x_{\text{limiting codon},i}/c_{\text{limiting codon},i}$ for a given amino acid sequence, which will consequently minimize the efficiency with which that protein can be produced, $\eta_{\text{min},i}$ (notated at the origin of the sketched plot in Figure 5.4). In our example case of threonine, the minimum efficiency will always be to select the codon ACA to encode any threonine amino acids in a protein, since ACA corresponds to the smallest sum of relative tRNA abundances as measured in *E. coli* [68] (refer to threonine under Table 5.1).

Maximizing translational efficiency The synonymous codon usage yielding the maximum synthesis rate can be found by choosing, for a given number of a particular amino acid, the distribution of corresponding codons that maximizes the efficiency computed from Equations 5.6–5.9. Consider threonine, for example.

There are 4 codons that encode threonine [188] (Table 5.1), so there are 4 fractions of tRNA abundance which can bind to these codons ($t_{1\ldots4}$) with a binding stoichiometry given by the matrix $B_{i,j}$. If a protein contains N_{thr} threonine residues, the codon usage that yields the maximal translational efficiency if those residues become limiting can be found by solving the following “maximin” mixed-integer nonlinear programming (MINLP) problem [336]:

$$\text{Maximize } \min_j \left(\frac{\sum_i X_{i,j}}{c_j} \right) \text{ subject to the following constraints:}$$

$$\begin{aligned} X_{i,j} &= \text{nonlinear function of } t_i,\ c_j, \text{ and } B_{i,j} \\ c_1 + c_2 + c_3 + c_4 &= N_{\text{thr}} \\ c_1, \ldots, c_4 &\geq 0 \\ c_1, \ldots, c_4 &\in \text{integers} \end{aligned}$$

In this problem, c_j denotes the number of codons of type j that encode threonine, and $X_{i,j}$ is a matrix as defined in §5.1.2. Thus, the ratio $x_{\text{limiting codon},i}/c_{\text{limiting codon},i}$ will be maximized for a given amino acid sequence, which will consequently maximize the translational efficiency for that protein, $\eta_{\text{max},i}$. This maximum efficiency ($\eta_{\text{max},i}$) is indicated near the right of the x-axis in the hypothetical example sketched in Figure 5.4.

The solution to this maximization problem is nontrivial because: 1) the matrix $X_{i,j}$ is a nonlinear function of t_i, c_j, and $B_{i,j}$, 2) the codon allocation values are constrained to be nonnegative integer values, and 3) solutions in which any of the codons take values of zero result in divide-by-zero errors. This third issue can be dealt with by the introduction of logical variables into the formulation which can then be used to assign arbitrarily large values to the elements of $X_{i,k}$ corresponding to codons $c_k = 0$ in any possible solution so as to essentially remove those codons from consideration in the min portion of the objective. For amino acids having only one synonymous codon and/or one tRNA isoacceptor species, the solution is trivial (see righthand side of Table 5.2). For nontrivial amino acids with more complex binding stoichiometry (lefthand side of Table 5.2), I used an *ad* *hoc* solution-finding approach for each value of N_{aa} found in the entire collection of *E. coli* proteins. This approach involved using the Microsoft Excel Solver tool to identify candidate solutions in which the variables were c_j. I then used several Matlab scripts which stepped through alternate solutions at varying degrees of "distance" from these initial solutions (in terms of reallocation of codons from the starting point) in order to find the codon allocation scheme that yielded the maximum translational efficiency. These solutions are provided in Appendix A.

Table 5.2: Amino acids for which optimization of translational efficiency is trivial or nontrivial.

Nontrivial	**Trivial**
Ala	Asn
Arg	Asp
Gln	Cys
Gly	Glu
Ile	His
Pro	Lys
Leu	Met
Ser	Phe
Thr	Trp
Val	Tyr

5.2 Results

5.2.1 Possible codon allocation schemes

In general, there are numerous ways in which a particular protein sequence can be specified by differing sets of synonymous codons. For example, a peptide consisting of just two threonines can be specified in 10 different ways (since threonine has 4 cognate codons), and in as many as 16 different ways ($2^4 = 16$) if the order of the codons is taken into account. In this study, since potential codon context and mRNA secondary structure effects have been neglected, only the numbers of allocated synonymous codons have been considered (and not their order within the transcript).

The "codon allocation space" (i.e. the number of all possible synonymous codon allocation schemes for a given amino acid sequence) was computed for each protein expressed in *E. coli.* This was done by generating a lookup table of the number of possibilities given the number of amino acids of a given type in the protein (N_{aa}) and the number of synonymous codon choices available for each amino acid (e.g. for threonine, there are four). The lookup table covering the range of synonymous codon choices (which is never more than six, the number for arginine, leucine, and serine) is computed as follows, where $A_{i,j}$ corresponds to each value in Table 5.3.

Table 5.3: Lookup table to determine the number of possible synonymous codon configurations for a set of amino acids of one type.

	Choices (m)					
N_{aa}	1	2	3	4	5	6
0	1	1	1	1	1	1
1	1	2	3	4	5	6
2	1	3	6	10	15	21
3	1	4	10	20	35	56
4	1	5	15	35	70	126
5	1	6	21	56	126	252
$\vdots$	$\vdots$	$\vdots$			$\ddots$	$\vdots$
n	1	$A_{n,1} + A_{n-1,2}$	$\dots$	$\dots$	$\dots$	$A_{n,m-1} + A_{n-1,m}$

In order to compute the total number of synonymous codon allocation schemes for a given protein (i.e. amino acid sequence), the corresponding values from the lookup table (Table 5.3) must be multiplied for each amino acid. (If there are no amino acids of a particular type in a protein, the multiplier will simply be unity, as evidenced in the first row of the lookup table.) The results for *E. coli* are presented in Figure 5.5. The smallest gene still has 15,360 possible synonymous codon configurations, and the largest has 10^{75}. These values are typically highly correlated with ORF length.

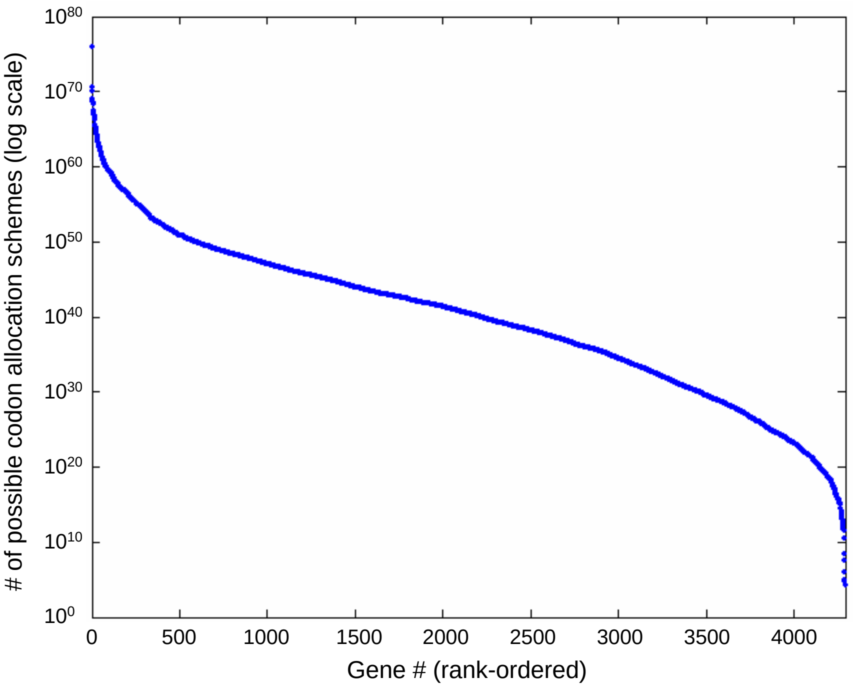


Figure 5.5: Number of possible synonymous codon allocation schemes for *E. coli* genes. The genes have been rank-ordered, and the number of possible synonymous codon schemes is plotted on a log scale.

5.2.2 Monte Carlo simulations of synonymous codon usage

Random synonymous codon configurations were assigned to each gene in *E. coli* for 1,000 Monte Carlo simulations (schematically illustrated in Figure 5.4). If an amino acid has four associated synonymous codons, for example, there will be a 25% probability that a specific one of those four codons will be selected for one of the amino acids of that type. The translational efficiency was computed for each of the randomized codon allocation schemes for every protein in *E. coli*. A histogram of efficiencies can then be plotted for each protein. Four example histograms are shown in Figure 5.6. These histograms are discontinuous because of the discrete nature of codon allocation schemes (i.e. the number of any given codon must be an integer value).

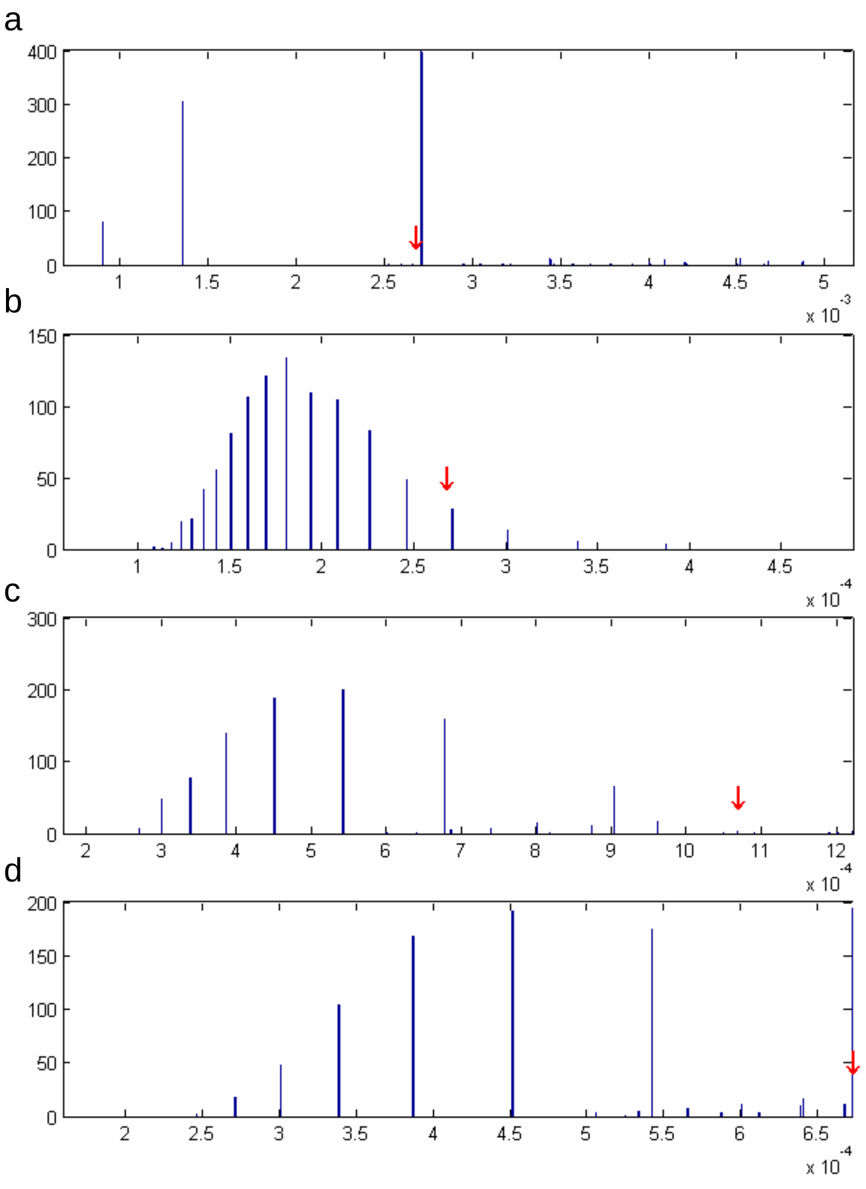


Figure 5.6: Histograms of translational efficiencies for four genes in *E. coli* given random synonymous codon choice. The histograms resulted from 1,000 Monte Carlo simulations given the model presented in Equations 5.6–5.9. Vertical red arrows indicate the efficiency of the wild-type sequence in *E. coli*. a. Histogram for *thrL* (b0001), which consists of only 21 amino acids. b. Histogram for *thrA* (b0002), which consists of 810 amino acids. c. Histogram for *thrB* (b0003), which consists of 310 amino acids. d. Histogram for *thrC* (b0004), which consists of 428 amino acids.

5.2.3 Range in protein synthesis

The calculated minimum and maximum translational efficiencies for a given protein that are possible by altering synonymous codon usage can be used to calculate the achievable range for each protein:

$$\xi_i = \frac{\eta_{\text{max},i} - \eta_{\text{min},i}}{\eta_{\text{min},i}} \tag{5.10}$$

This range represents the fold-change over which the protein levels can be adjusted, and the vector of ranges, ξ_i, constitutes what is essentially the flexibility of the proteome of an organism with respect to translational efficiency. This value is independent of the $[\text{mRNA}]_i$ or β_{rib} values chosen and is thus solely dependent upon a protein's amino acid sequence the the relative tRNA abundances. A histogram of these ranges for *E. coli* is provided in Figure 5.7. The mean range is 6.5, and the standard deviation is 2.6.

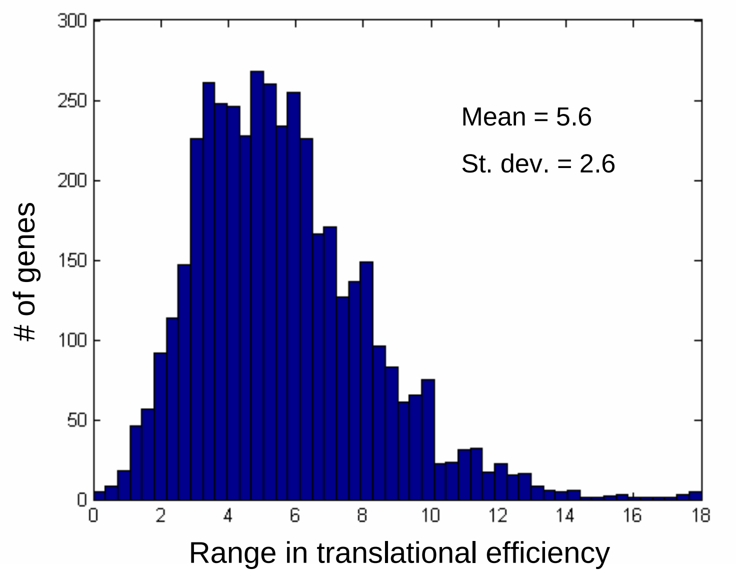


Figure 5.7: Histogram of achievable ranges in *E. coli* translational efficiency. These ranges represent the extent to which the translational efficiency for each gene in *E. coli* can be varied by altering synonymous codon usage (ξ_i in Equation 5.10). The mean range is 6.5, and the standard deviation is 2.6.

5.2.4 Optimality of codon usage in *E. coli*

Using the calculated $\eta_{\text{min},i}$ and $\eta_{\text{max},i}$ values, it is possible to define a measure of the efficiency of translation for a given organism (i.e. a given set of codon usages). If the η_i's are calculated for the codon usage found in the sequenced wild-type strain of an organism (denoted $\eta_{\text{actual},i}$), the optimality (i.e. translational efficiency) of each gene is given by:

$$f_{\text{actual},i} = \frac{\eta_{\text{actual},i} - \eta_{\text{min},i}}{\eta_{\text{max},i} - \eta_{\text{min},i}} \tag{5.11}$$

If the actual codon usage is optimal (i.e. best) in transcript mRNA_i, then the corresponding $\eta_{\text{actual},i} = 1$; and if the codon usage is worst, then $\eta_{\text{actual},i} = 0$. As was the case for the ranges, these normalized efficiency values, $f_{\text{actual},i}$, are independent of both $[\text{mRNA}]_i$ and β_{rib}, but will be dependent upon the codon usage and the relative tRNA abundances. This normalized efficiency measure is illustrated schematically in Figure 5.4.

Given these normalized efficiency values and the results from the Monte Carlo simulations, the efficiency of each gene in *E. coli* can be visualized in a 3-D histogram as shown in Figure 5.8a. The lefthand axis represents the $f_{\text{actual},i}$ values from Equation 5.11, which are indicative of how optimal the *E. coli* wild-type synonymous codon allocation is compared to that of the theoretical maximum and minimum ($\eta_{\text{min},i}$ and $\eta_{\text{max},i}$). The righthand axis represents the $f_{\text{Monte Carlo},i}$ metric presented in §5.1.3 and illustrated in Figure 5.4. This value is indicative of how optimal the wild-type *E. coli* translational efficiencies are relative to the efficiency results from the random sampling of the codon allocation space.

The results shown in the 3-D histograms in Figure 5.8a demonstrate that the codon usage in *E. coli* has been optimally tuned to the tRNA abundances present in the cell, or vice versa. The Wilcoxon rank-sum test [307] was used to ascertain whether or not genes belonging to particular functional classes [281] exhibited higher or lower efficiency than the rest of the genes. These results are presented in Table 5.4.

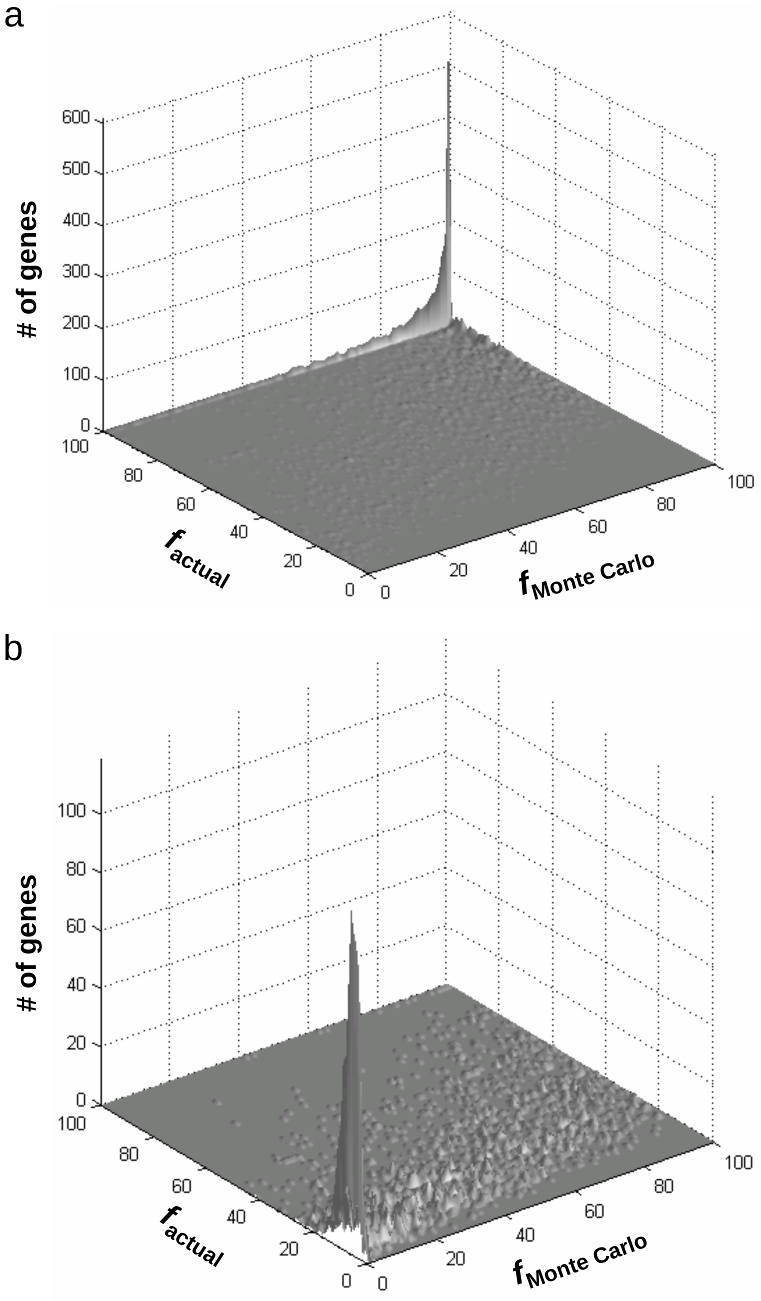


Figure 5.8: Optimality of codon usage in *E. coli*. The left axis on each 3-D histogram represents the f value given in Equation 5.11 expressed as a percentage, and the right axis represents the percentage of Monte Carlo simulations for each gene for which the calculated translational efficiency was less than the wild-type value. a. Efficiency of codon usage in *E. coli* given experimentally measured tRNA abundances [145, 68]. b. Efficiency of *E. coli* codon usage given randomized tRNA abundances.

Table 5.4: Functional class biases in translational efficiency in *E. coli.* Only functional classes for which significant differences in $f_{\text{actual},i}$ were detected have been included (Wilcoxon $p < 0.001$). The classes have been rank-ordered according to the mean $f_{\text{actual},i}$ for $i \in$ class

	Mean $f_{\mathbf{actual},i}$ for:	
Gene class	$i \in$ **class**	$i \notin$ **class**
Classes w/higher average efficiencies		
Ribosomal proteins	0.9356	0.8273
Potential-driven transporters	0.9234	0.8243
Energy production	0.9156	0.8270
Transport	0.8867	0.8183
Metabolism	0.8555	0.8133
Cell structure	0.8850	0.8103
Protein related	0.8833	0.8249
Primary active transporters	0.8829	0.8256
Classes w/lower average efficiencies		
Transposon related	0.6356	0.8320
Extrachromosomal proteins	0.6975	0.8403
Prophage genes	0.6982	0.8384
Unknown function	0.7923	0.8411

5.2.5 Sensitivity to tRNA abundance

The optimal solutions for codon allocation schemes presented in Appendix A are highly dependent upon the relative tRNA abundances, t_i. This motivates an assessment of how the translational efficiences vary with altered tRNA abundances. The tRNA abundances in *E. coli* were thus randomized relative to the actual measured values. Due to computational expediency, the nonlinear relationship between efficiency and codon usages presented earlier was linearized, thus forming an MILP problem [336]. The resulting efficiencies were very highly correlated to those computed using the nonlinear method in Equations 5.6–5.9 ($r = 0.96$). The $f_{\text{actual},i}$ values were also highly correlated with those determined from the linearized MILP method ($r = 0.78$). The 1,000 Monte Carlo simulations were also recomputed using this new linearized method. The resulting 3-D histogram is plotted in Figure 5.8b. The peak that was sharply located in the optimal range of the original histogram (Figure 5.8a) moved almost to the other corner of the efficiency spectrum when the tRNA abundances were randomized. This further demonstrates the highly-tuned nature of the tRNA species in *E. coli* with their cognate codons.

5.3 Discussion and Conclusions

The results presented in this chapter imply that the relative abundances of the components of a proteome (for a fixed transcriptome) can be varied ("wobbled") considerably without changing the amino acid sequence of each protein, assuming a parallelized model of translation and localized demands upon tRNA molecules. The average 6.5-fold theoretical range in translational efficiency in *E. coli* is surprisingly wide, given its potential impact on the composition of the proteome. Thus there is an alternative "regulatory mechanism" available to cells that is inherently built into the DNA sequence to modify relative protein abundances by changing codon usage through evolution. The existence of rare codons in bacterial organisms has previously been proposed as an evolutionarily determined means by which a cell regulates the protein expression of specialized genes [259].

It has been estimated that only 10-20 synonymous mutations per genome occurred over 20,000 generations of serially evolved *E. coli* [178]. This estimate was made from the random re-sequencing of 0.4% of the genome, thus potentially missing specific genes in which synonymous ("wobble") mutations might have been selected. *E. coli* has been subjected to adaptive evolution over 700 generations to optimize its growth rate on glycerol [143], and in a targeted re-sequencing of glycerol-related genes, the sequence of *glpR* contained the same synonymous mutation at nucleotide position 60 in two separately evolved strains [239]. Notwithstanding the low mutation rates predicted elsewhere [178], these findings raise an interesting question: To what extent can changes in synonymous codon choices affect the translational efficiency of a given protein in bacteria? This question has direct relevance to the optimization of protein synthesis for non-native genes that have been artificially inserted into bacteria such as *E. coli* towards genetically engineered goals [187] or for any other application in which heterologous protein expression is required [118]. Based on the results in this study and those published elsewhere [178, 239], it appears that *E. coli* may be utilizing the built-in "regulatory" mechanism (through changing synonymous codon usage) actively during adaptive evolution.

The results from this study demonstrate that the variability in translational efficiency achievable simply by altering synonymous codon usage is indeed wide enough to be subject to significant selection pressure [1]. These results also show that the codon selection in *E. coli* is far from random, with biases for both efficient and inefficient codon preferences and with respect to specific gene functional classifications [98]. The histograms in Figure 5.8 clearly show how highly-tuned the tRNA abundances in wild-type *E. coli* are to the synonymous codon choices. Furthermore, the Monte Carlo simulations revealed that the degree of taper on the tail of the histogram of efficiencies on any given gene (as seen in the examples in Figure 5.6) can provide a predictive clue as to what point along the range of efficiencies the wild-type value is likely to fall. The thinner the taper is on the tail of the distribution of efficiencies, the less likely the wild-type value will be optimal. In other words, perhaps there would be less chance for evolution to have "searched" the smaller space in the thinly-tapered tail.

There are a few avenues for enhancing and building upon the framework presented here to even more accurately determine translational efficiency in bacteria. As mentioned in §5.1.2, as more and more experimentally determined binding constants for tRNA isoacceptor/cognate codon interactions become available [58], the binary values in the **B** matrix can be converted into binding affinities. In this way, differences in codon-anticodon binding affinities can be accounted for [23, 291]. In general, base-pair matching which is not strictly Watson-Crick (i.e "wobble" pairing, [56]) will bind less readily than the usual G-C and A-U pairs. The differences in some of these binding affinities have been found to be as much as six-fold [58].

Additionally, the simulations in the present study do not take into account the possibility of hairpin formation in mRNA transcripts [211]. Such an event would certainly down-shift the translational efficiency for that protein (probably to nearly zero, but this would depend upon the probability that the hairpin actually exists at any given time). Another aspect of translation that may eventually be included in this framework is the use of codon usage to minimize the occurrence of errors during translation, which, along with translational efficiency, has been shown to be under selection pressure [199]. Clearly, there are multiple evolutionary forces at work with respect to influencing codon usage in microbial organisms, as codon bias alone has been deemed insufficient to explain differences in translational power among bacteria [65]. Additionally, mRNA abundance and stability are still very important determinants in protein expression in bacteria [341]. As a foundation upon which these additional characteristics can be built, the present study constitutes an important step towards systematically probing the process of translation in bacteria and towards establishing how genome-scale reconstructions of metabolism and information transfer could be incorporated within a flux-balance formalism [3, 248].

The text of this chapter, in part or in full, is a reprint of the material as it appears in T.E. Allen, N.D. Price, and B.O. Palsson. In preparation. Sensitivity analysis of translational efficiency with respect to codon usage and tRNA abundance. I was the primary author of this publication and the co-authors participated and supervised the research which forms the basis for this chapter.

Chapter 6

Spatial Considerations: "3-D Annotation" of Bacterial Genomes

Many cellular processes—particularly the processes of transcription and translation described and analyzed in the previous chapters—are subject to significant spatial constraints. Macromolecules such as ribosomes, tRNAs, and proteins, in addition to the DNA itself, are very tightly packed within the $\sim 1\ \mu m^3$ volume of the bacterial cell and must operate within an environment characterized by significant macromolecular crowding [111, 337, 80]. Furthermore, the processes of chromosomal replication and cell division constitute nontrivial spatial problems that many bacteria solve by rigorously controlling the three-dimensional arrangement of the segregating chromosomes [48, 338]. Aspects of cell cycle regulation, the physical division of the cell into two daughter cells, and even pathogenesis are closely linked to the localization of proteins to specific sites within the cell [107]. The flagella used to propel *E. coli* bacteria in their search for food are invariably located at just one of the poles of the rod-shaped organism. Furthermore, there is growing evidence that the location of the genes along the chromosome (and not simply the collection of genes specified in the genome) has a significant impact on the location of gene products and the overall structure of the cell [59]. This topological view of the genome is supported by several studies indicating that gene order and function in most bacteria is significantly nonrandom, as will be described in detail in Chapters 7 and 8. These location-specific examples, as well as many other processes within bacterial cells, demonstrate the importance of spatial considerations in any whole-cell reconstruction that explicitly includes macromolecules and the processes in which they participate.

The 1-D and 2-D annotations discussed in Chapter 1 have a relatively rich conceptual history and have been explicitly defined over a number of years [245, 236, 248, 219]. However, a rigorous definition of the "3-D annotation" of a genome remains to be established. Perhaps the closest to such a definition can be found in Gerald Edelman's visionary 1988 text on molecular embryology [72] in which he introduced the term *topobiology*, referring to the field dealing with the "critical issue of place-dependent molecular interactions at the cell surface," with particular emphasis upon the importance of spatial cues in the differentiation of cells in a developing embryo. More recently, Antoine Danchin has proposed that the chromosome itself not only serves as a "parts list" for the cell, but also as a "map" to the cell (i.e. the location of the genes on the chromosome are important for the localization of their products in the properly functioning bacterial cell) [59]. Along Danchin's line of reasoning, I have here extended Edelman's definition of topobiology to encompass the spatial organization of a cell interior, and in particular the arrangement of the genome and the impact of this arrangement (both in gene order and in chromosomal packing) on the cell phenotype.

In the next section, I will provide an overview of this newly emerging "dimension" of genome annotation, and the following sections will summarize the concepts and facts that will form the basis for this fast-moving and exciting field.

6.1 What is a “3-D annotation”?

As mentioned above, current reconstructions of metabolism take into account some rudimentary spatial characteristics via the assignment of cellular components to specific subcellular compartments. In gram-negative bacterial reconstructions to date, these compartments include the cytoplasm, the periplasmic space, and the extracellular space [245, 248]. Components in reconstructions of eukaryotic cells such as yeast are assigned to organelles such as the nucleus, mitochondria, lysosome, or peroxisome [71].

While these compartmental constraints are sufficient for reconstructing and modeling most of the metabolic and growth capabilities of cells, their predictive ability with respect to most processes falling under the umbrella of classical cell biology is severely limited. The growing body of data from the field of microbial cell biology summarized in the subsequent sections of this chapter calls for a conceptual framework by which the three-dimensional spatial arrangement of a cell's components may be reconstructed and modeled (not unlike current 2-D network reconstructions). In order to establish this conceptual framework, one must first clearly delineate what is meant by “3-D annotation”:

> **3-D annotation:** A description of the spatial location of each chromosomal locus and each gene product specified in a genome, including its movement throughout the life cycle of the cell.

This definition thus encompasses the spatial arrangement and localization of the chromosome within the cell (and the spatial coordinates of each gene locus), cytoskeletal-like proteins, motor proteins (e.g. flagella), cellular machinery (e.g. ribosomes), and signaling molecules.

Because the field of modern microbial cell biology is much younger than the fields of microbial biochemistry and genetics, the resolution at which we can annotate components in space is less concrete than our confidence in the stoichiometry of most biochemical reactions and small molecule transporters in 2-D network reconstructions. Moreover, unlike the “hard” constraints imposed by reaction stoichiometry and thermodynamics, subcellular localization and trafficking in bacteria is neither static nor 100% precise due to inherent stochasticity. As a result of these limitations, no single technique will be able to elucidate the exact spatio-temporal state of every component within a bacterial cell. Despite this limitation, however, a number of techniques have been developed which are useful towards achieving this 3-D annotation. A representative set of these methods is displayed in Figure 6.1, with the corresponding length scale and component resolution indicated on the axes.

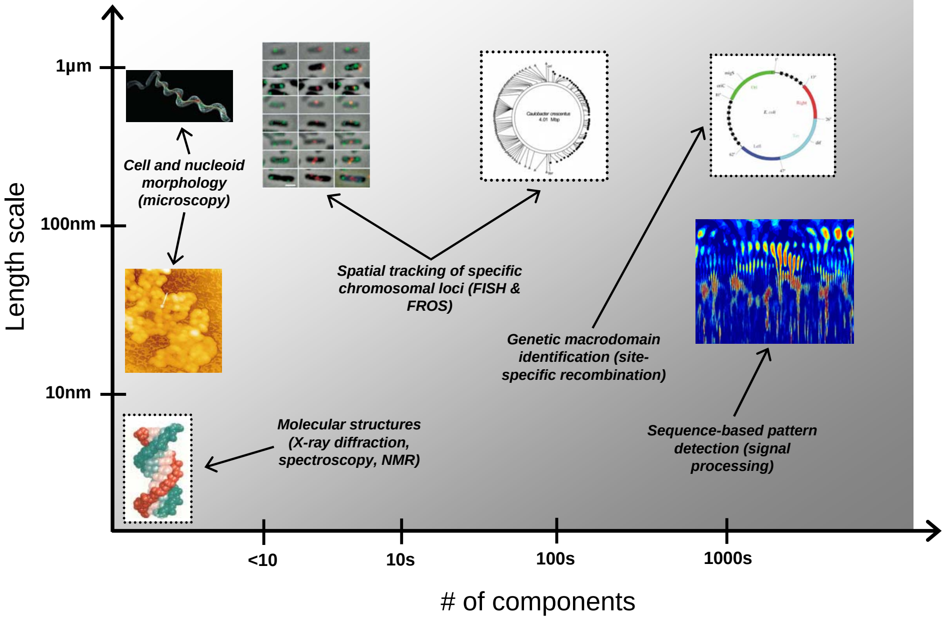

Length scale
1µm
100nm
10nm
Cell and nucleoid morphology (microscopy)
Spatial tracking of specific chromosomal loci (FISH & FROS)
Genetic macrodomain identification (site-specific recombination)
Sequence-based pattern detection (signal processing)
Molecular structures (X-ray diffraction, spectroscopy, NMR)
<10
10s
100s
1000s
# of components

Figure 6.1: Overview of 3-D annotation methods. The x-axis represents the number of cellular components being tracked in space, and the y-axis represents the length scales at which these components can be elucidated. In general, there is a dearth of data at finer resolutions of length at genome-scale.

Given the relatively nascent nature of the field of 3-D genome annotation, the aim of this chapter is to identify the spatial aspects of the cell interior and its components, highlight the methods used to elucidate these spatial characteristics, and discuss the current state of the art in this field.

6.2 The limitations of 2-D networks

The nodes in the cellular networks described in the preceding chapters are, in reality, physical molecules and structures that exist in three dimensions. Furthermore, these molecules not only occupy space but also exert non-trivial forces on their neighbors and the surrounding environment. The spatial arrangement and diffusion of small molecules (i.e. metabolites) can be considered negligible in most cases, and consequently their size does not factor in as a significant constraint in a cell-scale model. However, while water, metal ions, and small metabolites such as sugars and amino acids do not impose significant steric constraints on the cell interior, the aqueous environment is critical in setting the osmotic and electrostatic properties of the cytoplasm. For example, the presence of water favors the familiar B-form conformation of the DNA, and the absence of water favors the A-form. This effect of hydration is due to increased stabilization of the B-form due to the ability of its minor groove to accommodate a spine of water molecules [188].

A typical biochemical reaction can be written as follows, assuming that the participants in this reaction are not diffusion-limited and that they exist in what is essentially a well-mixed solution:

$$\mathsf{D\text{-}glucose} + \mathsf{ATP} \longrightarrow \mathsf{glucose\text{-}6\text{-}P} + \mathsf{ADP} + \mathsf{H}^+$$

The size of tiny molecules such as glucose and ATP is not likely to factor in to their participation in this reaction. Larger macromolecules such as the chromosome, however, are inadequately described by simple chemical equations like the one above. Such an oversimplified equation for the lumped polymerization of nucleotides in DNA replication might be written:

$$\mathsf{DNA} + \mathsf{DNA\ polymerase} + \mathsf{millions\ of\ dNTPs} \longrightarrow 2 \times \mathsf{DNA}$$

Clearly, this description of the DNA replication, while accurately accounting for the global metabolic costs that are required, does not account for the expression of individual genes, the interaction of these genes with regulatory and expression machinery, or the effects of DNA packing and gene order.

These spatial, *topobiological* aspects of the chromosome are inherently diffusion limited within the cramped confines of the cell interior [80]. Towards characterizing how the 3-D macromolecules are fit together within the cell, I will first describe the biophysical properties that characterize the intracellular millieu and the interaction of large molecules within the bacterial cell.

6.3 Intracellular biophysics

Before discussing some of the methods used to study the genome and its packaging within the bacterial cell at a more macroscopic level, it is important to gain a basic understanding of the biophysical principles of the DNA molecule and the intracellular milieu. Towards this aim, the current section will delineate the issues that arise inside the bacterial cell due to macromolecular crowding (§6.3.1) and the implications of this environment for macromolecular diffusion (§6.3.2).

6.3.1 Macromolecular crowding

Macromolecules generally constitute roughly 20-30% of the intracellular volume [200]. In fact, the total concentration of RNA and protein within *E. coli* has been estimated to be around 300-400 g/L, suggesting a specific volume for macromolecules of close to 1 mL/g [348]. Table 6.1 gives the volume and mass of the components found in a typical *E. coli* cell.

The importance of molecule size in volume exclusion is illustrated schematically in Figure 6.2. While the effect of crowding on small molecules is negligible (since a small molecule is free to diffuse to any available space; Figure 6.2a), macromolecules are excluded from all but a fraction of the available space due to their own size (Figure 6.2b). The concept of macromolecular crowding is illustrated beautifully in many of the drawings of David Goodsell (see, e.g., his excellent book *The Machinery of Life* [111]). Figure 6.2c provides a publicly available example of Goodsell's drawings of the bacterial interior, giving a sense of the steric con-

Table 6.1: Quantity and volume of components in an *E. coli* cell. The volume of molecules such as proteins and mRNA are estimated averages.

Molecule	% DCW[a]	Vol/molecule[b]	# of molecules	% total vol
Protein	55.0	65.7	2,360,000	17%
RNA:				
rRNA	16.7	2,674	18,700	5
tRNA	3.0	39	205,000	0.8
mRNA	0.8	1,449	1,380	0.2
DNA	3.1	4.76×10^6	2.1	1
Lipid	9.1	1.36	22,000,000	3
LPS[c]	3.4	8.3	1,200,000	1
Peptidoglycan	2.5	1×10^7	1	1
Glycogen	2.5	2,294	4,360	1
Soluble pool	3.9	0.09	138,000,000	1.3
Water	N/A	0.03	23,400,000,000	68.7

[a] DCW = "dry cell weight," or the % of the cell mass excluding water

[b] Volume = nm^3

[c] LPS = lipopolysaccharide

straints imposed by the chromosome, the ribosomes, and the many proteins that are present.

The considerable macromolecular crowding found in bacteria (and all cells, for that matter) has notable effects on many aspects of cell function:

Solute size and effective concentrations As already mentioned above and illustrated in Figure 6.2a-b, the effective concentration in crowded media is much higher than the actual concentration due to the available volume per macromolecule.

Reaction equilibria A consequence of the high effective concentrations in crowded media is that macromolecular associations are highly favored in such an environment. As a result, the activity coefficient[1] of enzymes is increased significantly in the cell interior, resulting in an increase in the equilibrium coefficient of as much as two to three orders of magnitude.

[1] The "activity coefficient" is defined as the ratio of effective concentration (due to macromolecular volume exclusion) to actual concentration.

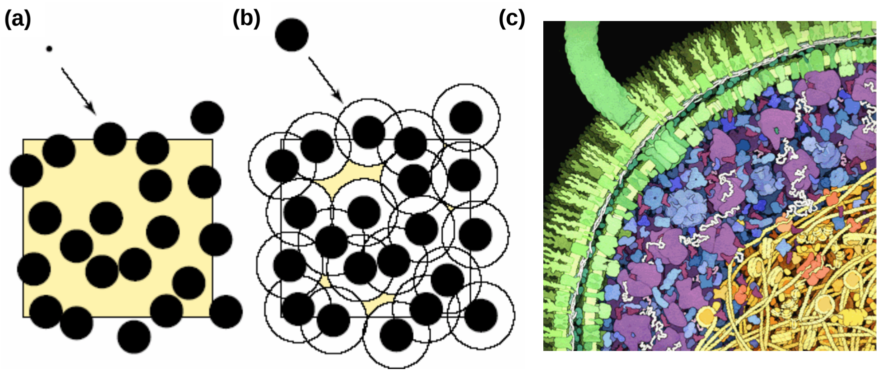


Figure 6.2: Macromolecular crowding. (a-b) The macromolecules shown in each of these squares occupy approximately 30% of the available space (adapted from [80]). (a) A small molecule is able to diffuse to the vast majority of the available 70% space (shaded in yellow). (b) A large macromolecule (of comparable size to the macromolecules already present), however, will be prevented from reaching most of the available volume because it is unable to approach the macromolecules closer than the open circles drawn around each one. (c) David Goodsell's representation of the approximate sizes, shapes, and density with which macromolecules are packed within *E. coli*.

Reaction rates A more intuitively obvious effect of macromolecular crowding is a reduction in molecular diffusion rates (especially for large molecules such as proteins and ribosomes). A commonly used analogy for this effect is the difference between the rate at which a person can walk across a room devoid of people (which may take a few seconds) and one that is full of people (which may take several minutes). We will discuss macromolecular diffusion in more detail in §6.3.2, but for now we will note that while macromolecular crowding increases equilibrium constants, it will limit the rate of any reaction that is subject to diffusion limitations.

In addition to these primary effects of macromolecular crowding *in vivo*, there are numerous examples of enhanced molecular chaperone activity as a result of crowding due to increased aggregation and increased functional activity (not unlike that observed for enzymes in a crowded environment) [310]. Consequently, there are those who argue that macromolecular crowding "play[s] a role in all biological processes that depend on noncovalent and/or conformational changes, such as protein and nucleic acid synthesis, intermediary metabolism and cell signalling, gene expression and the functioning of dynamic motile systems" (from [80]). Accordingly, this area continues to be an increasingly active realm of research.

6.3.2 Diffusion of macromolecules *in vivo*

As mentioned in the previous section, macromolecular crowding will reduce the rate of reactions that are diffusion-limited. From Fick's first equation,

$$-J = D\partial c/\partial x, \tag{6.1}$$

we know that the time it takes a molecule to diffuse a unit distance is inversely proportional to the diffusion constant, D. If a reaction is diffusion-controlled, it will be sensitive to increased macromolecular crowding since the crowded intracellular environment has been shown experimentally to decrease the diffusion constant D by as much as an order of magnitude.

However, recall from the previous section that an increase in macromolecular crowding also yields an increase in thermodynamic activity. Consider a simple condensation reaction of the form

$$\mathsf{A} + \mathsf{B} \longleftrightarrow \mathsf{AB}^* \longleftrightarrow \mathsf{AB},$$

where AB^* is the transition intermediate. If the first step in this process (the encounter of A and B) is rate-limiting, then macromolecular crowding will decrease the overall reaction rate due to the decreased diffusion constant. However, if the overall rate is limited by the activity coefficient of AB^*, then crowding will increase the overall rate. Thus, it should be clear that the effects of macromolecular crowding on intracellular reactions are nonlinear, particularly with respect to molecular size and concentration.

Measuring *in vivo* diffusion constants The most common method for measuring the translational diffusion constant of molecules *in vivo* involves fluorescence recovery after photobleaching (FRAP). In these experiments, a fluorescent molecule (e.g. green fluorescent protein, or GFP) is introduced into the cell, and an intense light pulse is used to irreversibly bleach these molecules within a specific area of the cell. The temporal change in the fluorescent intensity of this photobleached spot can then be used to compute the diffusion coefficient (D) of the fluorescent molecule within the living cell [320]. Using this method in *E. coli*, GFP was found to have an *in vivo* diffusion constant of $D = 7.7 \times 10^{-8}$ cm^2/s. A larger (72-kDa) fusion protein made up of GFP and a cytoplasmic maltose binding protein diffused more slowly ($D = 2.5 \times 10^{-8}$ cm^2/s) [82]. If the *E. coli* cell is assumed to be 2 μm long, these proteins can diffuse the length of the cell in 0.1-0.27 s (given Einstein's diffusion equation, $x = \sqrt{6Dt}$). The results of FRAP experiments should be interpreted cautiously, however, in light of phenomena such as incomplete fluorescence recovery or multi-component recovery [320].

An alternative method[2] for measuring solute and macromolecular diffusion *in vivo* is fluorescence correlation microscopy (FCM). This method takes advantage of the rate of fluctuations observed in a tiny volume fluoresced (not bleached) by a focused laser. The more rapid the observed fluctuations of the fluorescent molecule of interest in this spot, the greater the diffusion of that molecule. An autocorrelation function, G(τ), indicates the probability that a fluorescent particle found in the lighted area will also be found there at a later time point. The shape of the G(τ) curve can then be related to the diffusion coefficient, D.

In addition to measuring GFP diffusion in *E. coli*, the diffusion of single copies of mRNA has been experimentally determined. In order to accomplish this, a fusion protein composed of GFP and a bacteriophage coat protein may be used in concert with a reporter RNA containing repeated coat protein binding sites in order to fluorescently tag the mRNA reporter. The movement of these molecules can then be viewed by fluorescent microscopy.

[2] Both FRAP and FCM are useful for measuring translational diffusion. However, molecules can also rotate at a fixed point in space; this sort of mobility is called rotational diffusion. One method for measuring rotational diffusion *in vivo* is time-resolved anisotropy.

6.4 Supercoiling and DNA topology

The DNA molecule in *E. coli* would extend over a millimeter in length if cut from its circular form and stretched out completely. However, the entire *E. coli* cell is only ∼2 μm long, so the chromosome must be packed on the order of 1,000-fold in order to fit into the tiny cell interior [337]. Moreover, the entire chromosome must be reproduced in as few as 20 minutes, while simultaneously being accessed for transcription and translation throughout the cell cycle [31, 48]. Clearly, there must be significant chromosomal packing and organization in order for the cell to overcome this nontrivial spatial problem. I will now cover the major topics regarding DNA topology, and, in particular, supercoiling.

6.4.1 Supercoiling basics

Most bacterial DNA is circular in nature, rather than linear like the chromosomes in eukaryotic organisms. Because of this circular structure, the introduction of additional twist will affect the number of base pairs (bp) per turn in the DNA double helix. Consider the short DNA sequence shown schematically in Figure 6.3. If relaxed, double-stranded, linear DNA containing 105 bp and 10.5 bp/turn is bent into a circle and sealed (Figure 6.3b), it will contain 10 turns (twist, $T = 10$). However, if a negative turn is introduced into the linear DNA in Figure 6.3a before it is sealed, the resulting DNA will contain only 9 turns ($T = 9$) and 11.67 bp/turn (Figure 6.3c). Energetically, however, this DNA will not like having more or fewer than 10.5 bp/turn, and thus will not like the configuration shown in Figure 6.3c since it is considered to be underwound. To return the twist back to 10, it will supercoil by introducing a "writhe" (W) of -1 (Figure 6.3d). Alternatively, if we had introduced an extra turn to the right rather than to the left before sealing the DNA, it would have been overwound (with $T = 11$ and 9.54 bp/turn) and would have introduced a positive supercoil ($W = +1$) to return the twist to 10.

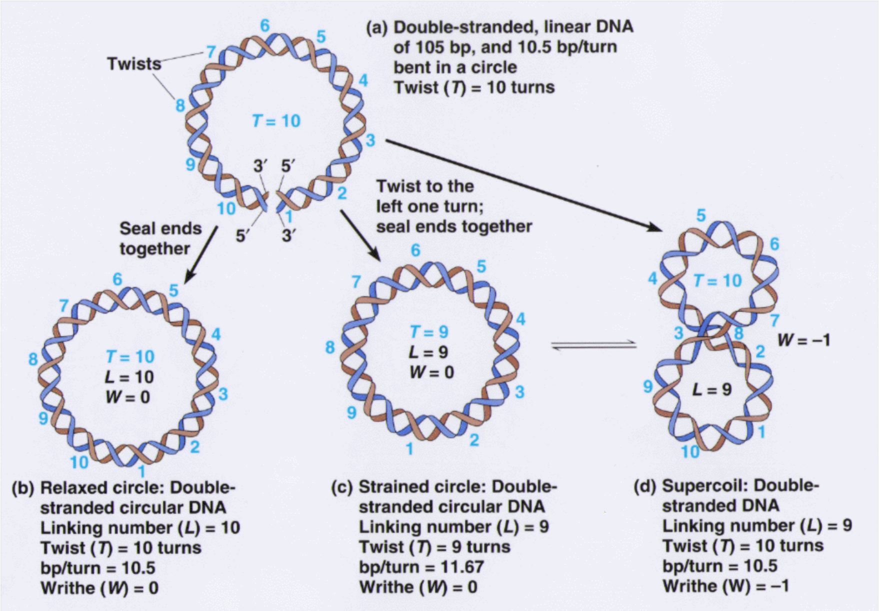

(a) Double-stranded, linear DNA of 105 bp, and 10.5 bp/turn bent in a circle Twist (T) = 10 turns
Twists
$T = 10$
3′ 5′
5′ 3′
Seal ends together
Twist to the left one turn; seal ends together
$T = 10$
$L = 10$
$W = 0$
$T = 9$
$L = 9$
$W = 0$
$T = 10$
$W = -1$
$L = 9$
1
2
3
4
5
6
7
8
9
10
(b) Relaxed circle: Double-stranded circular DNA Linking number (L) = 10 Twist (T) = 10 turns bp/turn = 10.5 Writhe (W) = 0
(c) Strained circle: Double-stranded circular DNA Linking number (L) = 9 Twist (T) = 9 turns bp/turn = 11.67 Writhe (W) = 0
(d) Supercoil: Double-stranded DNA Linking number (L) = 9 Twist (T) = 10 turns bp/turn = 10.5 Writhe (W) = −1

Figure 6.3: DNA supercoiling formation. The panels in this figure are described in the text. (Adapted from [188].)

In both of these scenarios, the sum of the twist and the writhe is a constant (9). This constant is called the "linking number" (L), which is related to the other quantitites by $L = T + W$. The linking number is more properly defined as the total number of times that the two circular DNA strands are interlinked. If this circular DNA is cut, twisted (or untwisted), and then re-sealed, the linking number will be changed. For sealed, circular DNA, however, L is a constant. Accordingly, when in Figure 6.3c-d we added an extra twist to the left, we changed the linking number by -1 ($\Delta L = -1$). Any strain that is introduced by extra or fewer turns (ΔL) must be distributed between a change in twist (ΔT) and a change in writhe (ΔW):

$$\Delta L = \Delta T + \Delta W. \tag{6.2}$$

A quantity describing the degree of superhelicity in a circular DNA molecule is the *superhelix density*, $\sigma = \Delta L/L_0$, where L_0 is the linking number of the DNA in its relaxed state. The circular DNA naturally occurring in *E. coli* and many other bacteria, for example, has a superhelix density of $\sigma = -0.06$. Since *E. coli* has 4.6 million bp—if it exists in the B-form and thus contains 10.0

bp/turn—will therefore contain 4,600,000 bp / (10.0 bp/turn) = 460,000 turns (i.e. $L_0 = 460,000$). If DNA gyrase were to twist the superhelix density to $\sigma = -0.06$, then $\Delta L = -0.06L_0 = -27600$. According to Eqn. 6.2, this change in linking number must be distributed in the twist and/or in the writhe. Since it is much easier for the DNA to simply bend (i.e. supercoil) rather than for *E. coli* to unwind the DNA, it will distribute the ΔL by setting $\Delta W = -27600$ and $\Delta T = 0$. Thus, the *E. coli* genome typically contains nearly 30,000 left-hand superhelical turns.

6.4.2 Topological domains

Biological processes such as replication require that DNA be unwound and/or cut. However, if there were nothing to stabilize the chromosome's supercoiling, such a cut would unwind all the negative supercoils in the cell. Unfortunately for bacteria such as *E. coli*, even a modest reduction in the degree of supercoiling is known to be lethal [346]. To avoid this significant problem, numerous DNA-binding proteins (and perhaps RNA molecules) serve as barriers that prevent the unwinding of the entire chromosome if an interruption is introduced (Figure 6.4). Because of these protein barriers—which can be thought to divide the chromosome into supercoiling *domains*—an interruption in the DNA only relieves the supercoiling in one of the domains, while leaving the necessary writhe intact in the remainder of the chromosome.

6.4.3 DNA binding proteins

Eukaryotes solve the significant problem of storing their genetic material inside the small space of a nucleus by packing their DNA around proteins called histones, which allow each chromosome to be densely packed into chromatin. Bacteria, however, do not contain histones, *per se*, but they do contain many similar DNA binding proteins which function as supercoiling domain barriers and allow much tighter packing than would be possible with naked DNA. Table 6.2 provides a list of the major proteins known to bind to the chromosome in *E. coli* (see also

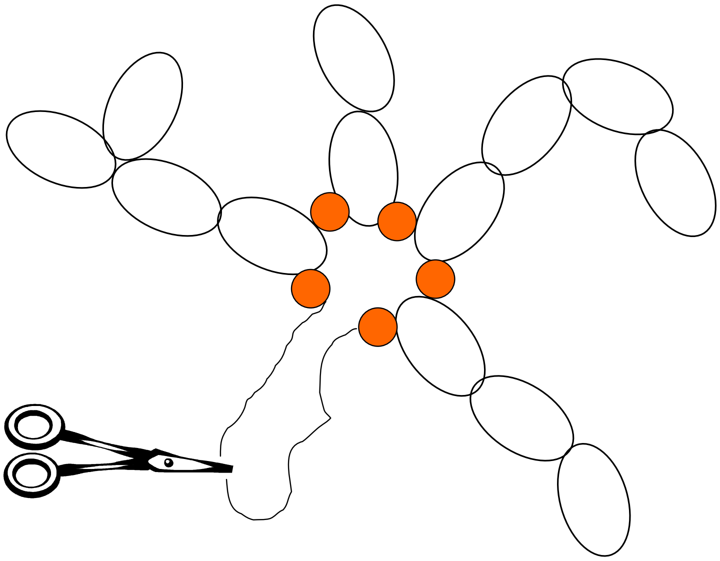


Figure 6.4: Topological organization of the bacterial chromosome. Bacterial DNA is tightly packed into supercoiled domains that are topologically disjoint from each other on account of supercoil diffusion barriers (orange). Therefore, certain cellular processes or DNA damage which introduce breaks into the DNA strand (indicated by scissors) will relax only one domain, leaving the topological state of the rest of the chromosome unchanged.

Table 7.3). Most of these (IHF, HU, H-NS) are histone-like proteins that often share strong sequence homology. All introduce varying degrees of bending (the protein called integration host factor, or IHF, introduces a nearly 180° bend in the DNA) and can bind to multiple sites on the DNA. There are 8 DNA binding proteins in addition to the 4 listed in Table 6.2, but of these only Dps (which replaces HU and FIS during stationary phase) is present in abundance. The other known DNA binding proteins are predominantly transcription factors and regulators such as Crp, CbpA, Lrp, and DnaA, but a thorough discussion of each of these is beyond the scope of this dissertation.

In addition to these proteins, the proteins Muk and SMC form complexes which serve to condense the DNA much like an accordion in an energy-dependent manner [204, 332, 267]. Further condensation is achieved by the introduction of negative supercoils by DNA gyrase, which serves to re-wind the chromosomal domains which are relaxed during DNA replication or due to the action of DNA topoisomerases [346].

Table 6.2: Major DNA binding proteins found in *E. coli*.

Protein	#/cell	Binding sites[a]	Comments
HU	60,000	N/A[b]	Constrains supercoiling
IHF	17,000-34,000	608	Causes U-turn in DNA
H-NS	20,000-60,000	N/A[b]	DNA condensation; regulation; constrains supercoiling
FIS	200-100,000[c]	∼6000	Regulation of rRNA & tRNA; other regulatory activities

[a] Binding sites predicted using hidden Markov models

[b] Non-specific binding

[c] Strongly dependent on growth phase

6.4.4 Elucidating supercoiling domains

In the late 1970s and early 1980s, researchers primarily used electron microscopy (EM) to observe the existence of supercoiling domains. In well-known experiment in 1981 [289], Sinden and Pettijohn introduced varying numbers of nicks into isolated *E. coli* DNA and used EM to measure the degree of chromosome relaxation. Using this method, they determined that the *E. coli* chromosome contained 42 ± 10 domains of supercoiling.

More recent work from Nicholas Cozzarelli and colleagues [231], however, used microarrays to measure the expression of more than 300 genes that were known to be sensitive to supercoiling. Nicks were then introduced into specific sites in the DNA using a restriction endonuclease, and the effects of these nicks (and the resulting reduction in supercoiling in the local domain) were measured using the expression arrays. This experiment, when coupled with additional EM visualization, revealed that supercoiling domains were smaller than had been previously reported and were highly variable in size, ranging from 2 to 66 kb, and averaging ∼10 kb. Furthermore, the locations of these ∼500, smaller supercoiling domains were not constant but were instead found to be fluid with respect to chromosomal position. There are numerous advantages for a bacterium such as *E. coli* to possess such small supercoiling domains, especially the fact that smaller domains mean less of the chromosome must be unwound when DNA interruptions occur. Additionally, small, fluid supercoiling domains permit the cell to undertake major processes such as replication, transcription, and translation simultaneously while still maintaining significant chromosomal structure [231].

6.5 Relation of structure to function

A key recurring theme in second half of this dissertation is that the bacterial chromosome not only contains information in its primary chemical sequence but also in its three-dimensional configuration. At the beginning of this chapter, Figure 6.1 shows the key methods and length scales by which the bacterial cell interior can be elucidated. In the last section, I reviewed the key aspects of DNA topology and supercoiling. The following question logically derives from these previously discussed topics: How does the 3-D arrangement of the chromosome and of the intracellular components affect cellular function and behavior? A summary of the latest findings that shed light upon this ultrastructure-function relationship in bacteria is the goal of this section. Much of the current knowledge of the structure and function of the bacterial nucleoid can also be found in a recent review from Lucy Shapiro's group [298].

First I will briefly discuss the impact of DNA structure and topology on gene expression and regulation. Then I will describe the proteins that are responsible for determining cell shape and structure to bacteria. Finally, I will review protein and chromosomal localization within the cell, with particular emphasis upon the effects of structure and localization on cell function.

6.5.1 Effect of DNA structure on gene regulation

The synthesis of cellular components such as proteins and RNA is generally controlled at the level of gene expression in bacterial cells. An oftentimes complex network of transcriptional control has evolved to allow bacteria to optimally respond to changing conditions and surroundings, thus providing them with a significant competitive advantage. The transcription of a specific gene is typically governed by the promoter region occurring upstream of the actual gene. Regulatory proteins called transcription factors bind to specific promoter regions with varying affinities, causing either an increase or a decrease in the transcription of the downstream gene, depending on the particular factor involved [325].

Some of the most common DNA binding proteins that were discussed in the previous section, however, do not bind to specific primary sequences or promoters. Instead, these proteins (which include HU and H-NS, among others) recognize particular 3-D conformations of the chromosome, such as the intrinsic curvature of a given stretch of DNA. (The intrinsic curvature inherent in any given stretch of DNA sequence is primarily a function of the GC content within that sequence. G-C nucleotide pairs are comprised of three hydrogen bonds, as opposed to only two hydrogen bonds for A-T pairs. As a result, GC-rich DNA will have increased stiffness as compared to AT-rich sequences and will thus possess a lower intrinsic curvature [226].) These so-called “architectural” proteins are often responsible for supporting a particular spatial configuration of a localized portion of DNA that either promotes or inhibits transcription in that region.

The level of DNA supercoiling can also affect transcriptional regulation by altering the binding affinities of transcription factors and/or RNA polymerase to a promoter region. Oftentimes this relationship between supercoiling and gene regulation leads to a localized effect in which the regulation of nearby genes are topologically coupled [325]. This type of coupling is created by the action of the RNA polymerase as it transcribes the DNA. Because the chromosome is essentially fixed such that the entire thing is unable to rotate as the polymerase progresses (due to not only its sheer size, but also to its anchoring to the translational machinery and sometimes the cell membrane), transcription naturally causes positive supercoiling in the locale just upstream of the gene and negative supercoiling downstream. Recall from earlier in this chapter that positively-supercoiled regions can be returned to normal by DNA gyrase, and negatively-supercoiled regions can be restored by topoisomerase I. If the transcription of a neighboring gene is sensitive to the level of supercoiling, this topological coupling may serve as a potent inhibitor or activator for such a gene. This phenomenon is called the twin-supercoiled domain model [325]. Additional detailed reviews of the role of supercoiling in transcriptional regulation can be found in Refs. [124, 306].

6.5.2 Cell shape

Eukaryotic cells have long been known to possess a well-characterized and structured cytoskeleton. This cytoskeleton is essential for maintaining cell shape and plays an integral role in many processes, including mitosis, motility, and protein trafficking. However, the presence of an analogous cytoskeletal structure within bacterial cells was long the subject of debate and has only recently been confirmed [38]. The basic cell shape cytoskeletal proteins in bacteria are shown schematically in Figure 6.5 and described below. Some cytoskeleton-like proteins known to occur in bacteria include:

FtsZ The protein FtsZ functions in dividing a bacterial cell near the middle of its long axis. FtsZ, a tubulin homolog GTPase present in virtually all bacteria and archea, aggregates near the cell division plane and constricts the cell membrane by an as yet unknown mechanism, recruiting additional cytokinetic factors [84]. More details on the self-organization properties of this system are provided in §6.5.3.

MreB, Mbl, & MreBH The protein MreB and its homologs play a key role in cell shape, polarity, and chromosome segregation in most non-spherical bacteria. This protein, an actin homolog, polymerizes *in vivo* into a spiral-shaped “skeleton” that gives the cell its elongated, semi-rigid shape [92, 106, 157].

ParM ParM, an actin homolog like MreB, plays a key role in plasmid separation

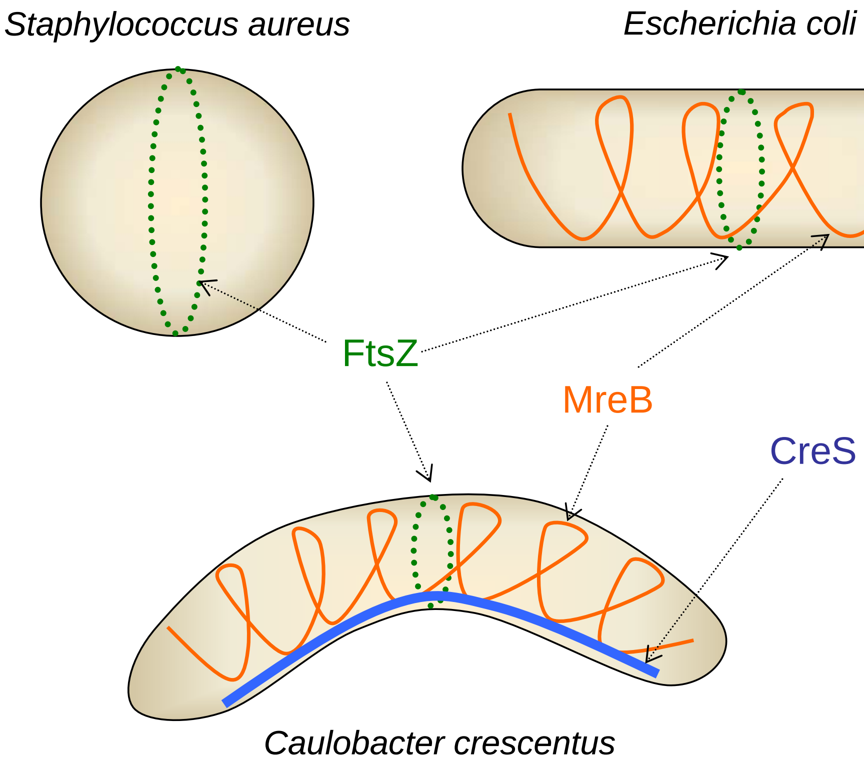


Figure 6.5: Cytoskeletal determinants of cell shape in bacteria. Three cytoskeletal proteins are common in bacteria: FtsZ, MreB, and CreS. All three of these appear in *Caulobacter crescentus*, FtsZ and MreB in the rod-shaped (but not crescent-shaped) *Escherichia coli*, and only the dividing ring FtsZ in the spherical coccus, *Staphylococcus aureus*. Since FtsZ imparts no shape in cells that are not dividing, an organism containing only this cytoskeletal protein will be spherical. MreB confers a rod shape, and CreS bends a rod-shaped bacterium into a crescent.

in bacteria. This protein is the most well-understood of the proteins involved in DNA segregation. Although it is an actin homolog, it exhibits dynamic instability (i.e. alternation between constant polymerization and rapid disassembly) reminiscent of eukaryotic tubulin [104, 195, 102].

CreS The protein CreS, a homolog of eukaryotic intermediate filaments, confers an inner curvature on crescent-shaped bacteria such as *Caulobacter crescentus* due to its coiled-coil shape [14].

6.5.3 Subcellular self-organization in bacteria

During cell division in *E. coli* and other bacteria, the structural protein FtsZ forms a ring at the center of the cell before it divides. However, the mechanisms of diffusion capture or targeted localization cannot, on their own, cause the medial placement of the FtsZ ring. This placement is governed by the Min proteins (Figure 6.6). The protein MinC inhibits the formation of the FtsZ ring and also binds to the MinD ATPase [137]. The two Min proteins (C & D) diffuse rapidly between the two poles [136, 242]. MinD polymerizes and binds to the membrane in its ATP-bound state, and then binds to a third Min protein, MinE. MinE, in turn, activates the hydrolysis of the MinD-ATP complex to MinD-ADP, causing the MinD to detach from the membrane. MinD is then recharged with ATP and aggregates at the point in the cell farthest from MinE (i.e. the opposite pole). This continuous oscillation of MinD (and MinE) drives the oscillation of MinC (since MinC binds to the MinD ATPase). The net result is that MinC concentrates at the poles and has the lowest concentration at the center of the cell [138, 172, 135]. Thus, FtsZ can polymerize most readily at the cell center. All of these steps are illustrated schematically in Figure 6.6.

6.5.4 Tracking individual loci during chromosomal segregation

Proteins are not the only molecules within the bacterial cell that are localized to specific regions during the cell cycle. In a pioneering set of experiments

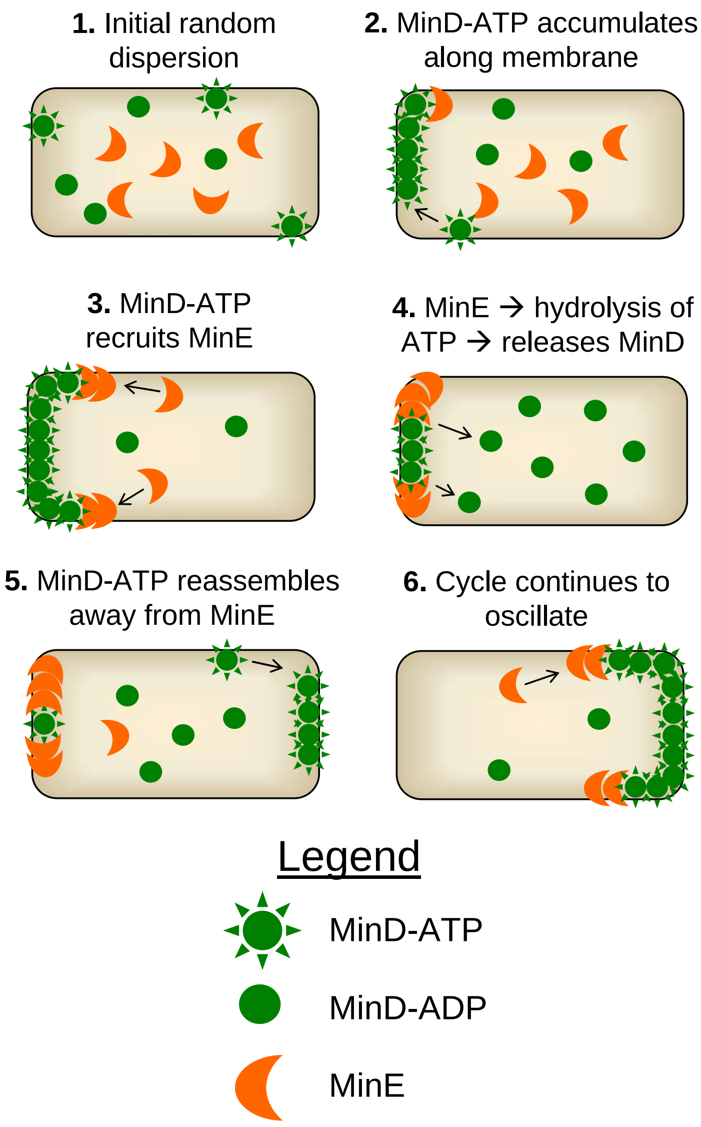

1. Initial random dispersion
2. MinD-ATP accumulates along membrane
3. MinD-ATP recruits MinE
4. MinE → hydrolysis of ATP → releases MinD
5. MinD-ATP reassembles away from MinE
6. Cycle continues to oscillate
Legend
MinD-ATP
MinD-ADP
MinE

Figure 6.6: Subcellular self-organization in bacteria: a system for finding the middle of the cellular axis. The polymerization of the FtsZ protein ring in *E. coli* occurs at the center of the cell's axis due to the rapid pole-to-pole oscillation of MinC, a protein which inhibits FtsZ polymerization and whose time-averaged smallest concentration occurs in the middle of the cell. Refer to the text for more details. (Adapted from [107]).

involving *Caulobacter crescentus*, Lucy Shapiro and colleagues used both fluorescence *in situ* hybridization (FISH; [205]) and a fluorescence repressor-operator system (FROS; [254, 112]) to determine the position of 112 specific chromosomal loci along the dividing *Caulobacter* cell [322]. They determined that the replicating chromosome in *Caulobacter* is organized in such a way as to preserve the linear order of the genes along the long axis of the cell. This highly ordered segregation occurs rapidly (on the order of minutes) and implies a high degree of chromosomal organization. The observation of the linear ordering of genes rules out several chromosomal arrangement configurations (e.g. random coils, rosette, or folded parallel to the long axis of the cell) and constitutes an early step towards elucidating the 3-D annotation of a bacterial chromosome [30]. Although such studies are more problematic in *E. coli* due to the difficulties inherent in synchronizing its cell cycle, some studies using FISH [205, 206] and FROS [81] have yielded similar results.

6.5.5 Other localization phenomena

Many bacteria use the localization of specific proteins within the cell for other processes in addition to chromosome segregation and cell division. Two of these processes—spore differentiation in *Bacillus subtilis* and pathogenesis in several organisms—are discussed below. These examples provide a glimpse of the breadth of bacterial localization phenomena that are only beginning to be elucidated.

Spore differentiation Certain types of bacteria have evolved very elaborate mechanisms for coping with extreme environmental stress by fundamentally altering their lifecycles to form spores under conditions of stress.[3] The best studied of the sporulating bacteria is *B. subtilis*.

As discussed in §6.5.3 and illustrated in Figure 6.6, proteins such as FtsZ can be localized at a specific point along the axis of the cell other than the poles.

[3] These spores can survive nearly indefinitely—in fact, a viable spore belonging to a previously undiscovered *Bacillus* species was recently unearthed from a 250-million-year-old salt crystal.

B. subtilis also contains the FtsZ protein, and under normal conditions, FtsZ forms a ring at the center of the cell just as it does in *E. coli* [27]. However, under conditions of severe stress to the cell, *B. subtilis* is able to shift the localization of FtsZ from the exact center of the cell towards one of the poles (typically about a quarter-cell distance from one of the poles). Forespore proteins will localize at the polar septal membrane generated by the off-center FtsZ ring. Meanwhile, the mother cell proteins are dispersed via diffusion throughout the larger part of the dividing cell (i.e. in the ∼75% of the cell on one side of the polar septum). Some of these proteins become enriched at the septum and enclose and surround the forespore proteins, eventually forming the nascent spore [27].

Pathogenesis Protein localization also plays a key role in the pathogenic activities of bacteria such as *Streptococcus pyogenes*, *Yersinia pestis*, *Listeria monocytogenes*, and *Shigella flexneri*. For example, in *Shigella*, the protein which causes pathogenesis (IcsA) is initially localized to one pole of the bacterium [253]. This localization is achieved by an IcsA-specific protease which selectively breaks down any IcsA that is not located at the selected pole. This localized protein then interacts with a host factor to generate actin "comet" tails which serve to propel the the bacterium around the host cell, thus decreasing the likelihood that the pathogen will be detected by the host's immune system [253].

6.6 Towards a genome-scale 3-D annotation

In this chapter, I have described the concept of a "3-D annotation" in biology, and I have reviewed the state of the art in our understanding of the interior of a bacterial cell. The information presented here largely derives from microscopy and X-ray diffraction, but the resolution and scale achieved with these methods is not currently sufficient to completely understand how the cellular components are arranged in space. An additional sort of analysis that was not presented here is analysis of the one-dimensional genome sequence. The sequence provides multiple parameters as a functional of chromosomal position, including gene locations, directionality of transcription, and estimates of intrinsic curvature per segment of DNA. However, there are currently few methods by which to extract large-scale spatial information and regularity from genome position-dependent data. In the next two chapters, I will describe my efforts to address this issue using signal processing techniques to elucidate the spatial organization of microbial genomes.

Chapter 7

Detecting Long-Range Patterns in Multiple Genomes

As described in the previous chapter, genomes in prokaryotic organisms typically are packed tightly into a nucleoid where they carry out multiple functions simultaneously [337, 349]. The condensed DNA within the bacterial nucleoid must not only be efficiently replicated and segregated during cell division [284], but it must also simultaneously participate in the information transfer processes of transcription and translation [48]. Recent studies have significantly advanced our understanding of the ultrastructural and multifunctional organization of prokaryotic chromosomes. DNA in *Escherichia coli* has been found to be packed into supercoiled domains ranging from 2-66 kb and averaging ∼10 kb [231]. At a slightly longer length scale, studies using fluorescence *in situ* hybridization (FISH) have revealed that the origin and terminus of replication in *E. coli* gravitate toward the poles of the cell throughout replication, but both migrate to the mid-cell region just prior to the initiation of chromosome replication [206]. Fluorescence experiments in synchronized cultures of the aquatic bacterium *Caulobacter crescentus* have revealed the cellular location of 112 individual chromosomal loci throughout replication and cell division [322] and of 124 loci near the origin in *E. coli* [81]. In addition to these imaging techniques, genetic dissection has been used to identify four macrodomains and two less-structured regions in the *E. coli* chromosome [309]. Two of these macrodomains were consistent with those found near the origin and terminus of replication using FISH [206]. However, many issues remain unresolved regarding the intricacies of this arrangement, and particularly the relationship between chromosomal ultrastructure and the processes of transcriptional regulation and protein synthesis [48, 298].

Several studies have revealed that genes in bacterial nucleoids tend to be arranged along the long axis of the cell (in the case of rod-shaped bacteria) so as to preserve the linear order of the genes along the chromosome [206, 322, 340, 30]. Given this linear arrangement, prokaryotic genome sequences inherently contain useful information relating to chromosomal ultrastructure since they provide numerous properties as a function of chromosome position [226]. However, the inference of 3-D genome-packing from direct examination of the raw sequence is somewhat challenging at the short length scales of the nucleotide, gene, or operon (1 bp-10 kb) due to the inherently one-dimensional nature of sequence data and hence the considerable sequence noise over shorter scales. Accordingly, various averaging and filtering methods have been used to identify long-range (i.e. > 10 kb) position-dependent patterns in genome-associated properties [226, 11, 12]. In order to detect such long-range periodic patterns in inherently noisy chromosome position-dependent data, wavelet analysis has been used in several studies [11, 182] (Figure 7.1). This method has previously been used to detect patterns in gene orientation [12], DNA bending profiles [13], and gene expression data [4, 156] in prokaryotes, as well as GC/AT skew oscillations in human chromosomes [202]. These studies have revealed that genome sequences are generally nonrandom with respect to chromosome position, and that long-range correlations in certain properties (e.g. gene orientation; [12]) exist across many length scales.

As more prokaryotic genome sequences become available, it should be increasingly possible to relate the quantitative degree of genome organization to global properties of each organism, including the presence of known nucleoid-

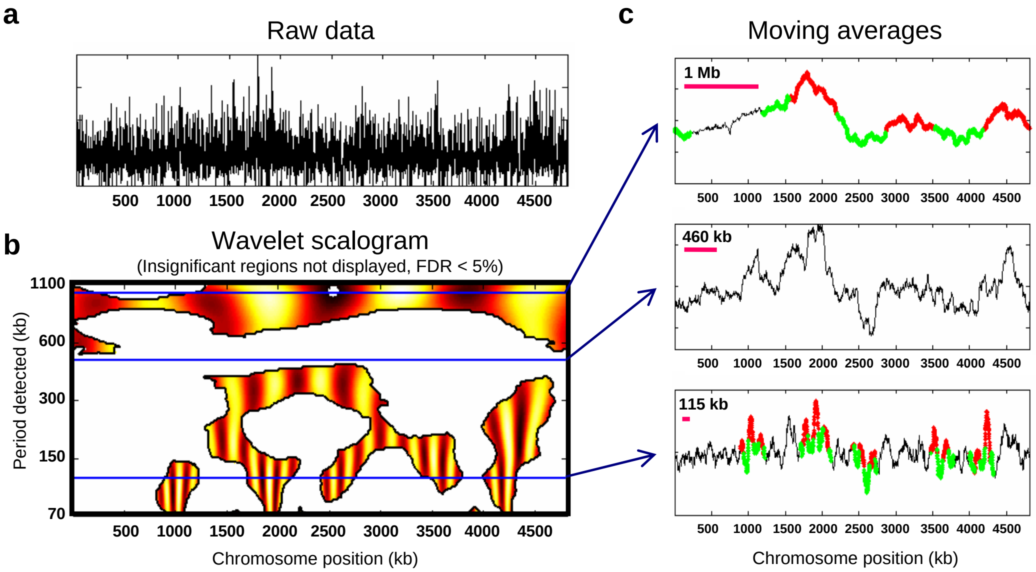


Figure 7.1: Approach for detecting genome position-dependent patterns. a. Raw sequence-derived data often contain patterns with respect to chromosome position that are not obvious from casual observance. (This example is for the fractional gene density per kilobase for *Salmonella enterica* serovar Typhi strain CT18.) b. Wavelet analysis was used to generate a scalogram showing significant chromosome position-dependent patterns in gene density over varying periodicities. The level of significance of the patterns was determined by randomizing the order of the raw sequence data 200 times and re-computing the real and imaginary portions of the Morlet wavelet transform values at each point in the scalogram for each randomization. Regions having a false discovery rate (FDR) greater than 5% are not displayed (white). The pattern strength for this data set is 33%. c. To facilitate the interpretation of the wavelet scalogram, three examples are shown for the moving averages of the raw data at three different length scales: 1 Mb, 460 kb, and 115 kb. Regions highlighted in red/green indicate significant regions of the scalogram at that scale that lie above/below the mean real transform value.

binding proteins [167], organism taxa, and genome size and composition. Observed correlations may indicate constraints that affect (or are affected by) genome organization. Therefore, the need exists to define an unbiased, quantitative measure of genome organization from sequence-derived data, compute this quantity for numerous sequenced prokaryotic genomes, and relate this quantity to global properties of each organism.

In this study, I have attempted to address these needs by employing wavelet analysis in concert with a bootstrap significance test (§7.1) to compute the pattern strengths of chromosome position-associated data sets derived from 163 sequenced prokaryotic chromosomes. This pattern strength provides a measure of the nonrandom nature of sequence-derived data that is independent of genome length. I then computed the pattern strength of genome position-dependent properties for nearly every sequenced prokaryotic genome, and we related this measure to taxonomic and physiological characteristics of each organism. The results presented in this chapter demonstrate that the degree of organization in bacterial genomes is highly variable and correlates with specific properties.

7.1 Pattern detection methodology and controls

7.1.1 Chromosome position-associated data sets

Data sets were analyzed from most prokaryotic genome sequences published to date (through January 2005) and were downloaded from the CBS Genome Atlas Database [121].[1] Four types of chromosome position-dependent data were analyzed for 151 prokaryotic organisms (corresponding to 163 chromosomes in 16 archaeal and 135 bacterial organisms): 1) GC/AT content averaged in kilobase bins, 2) gene orientation (i.e. strand), 3) fractional gene density (defined as the number of genes—or fractions of genes—per kilobase), and 4) codon adaptation index (CAI) [282] per gene. For the CAI, we used the global codon usage as the

[1] http://www.cbs.dtu.dk/services/GenomeAtlas/

reference set to maintain consistency, since the highly expressed genes for some of the organisms may not be predictable *a priori*. GC and AT content are by definition inversions of one another and are strictly anti-correlated, so any patterns present in either property will be identical. Thus, patterns in these properties are simply referred to patterns in “GC/AT content.”

7.1.2 Pattern detection by wavelet analysis and significance testing

Wavelet analysis [140], reviewed in detail elsewhere [305], is an approach whereby irregular patterns in biological data may be elucidated [12, 182, 4, 156, 202, 198]. In short, each genome-scale data set was ordered according to position along the chromosome. These ordered data, $f(x)$ (where x is defined as the nucleotide position along the chromosome), were then continuously integrated using a family of filter functions to obtain a transform value for numerous filter widths (i.e. scales, designated a) centered at each position x in the data set:

$$W(x,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g\left(\frac{x' - x}{a}\right) f(x') dx' \tag{7.1}$$

The filter function used in this study was the Morlet wavelet, defined as:

$$g(x) = e^{i5x} e^{-x^2/2}. \tag{7.2}$$

This particular wavelet was chosen because the length scale of the transform corresponds approximately to the period of any localized pattern [305]. The resulting transform values may be plotted in the form of a scalogram (Figure 7.1b), comprised of a contour plot in which the x-axis is the position along the genome (x), and the y-axis is the length scale (a) at which the transform is computed. Given that we employed the Morlet wavelet, this scalogram is useful for elucidating the strength of a range of periodicities localized at each point in time-series data (or, in this case, chromosome position-associated data). The particular voices (i.e. length scales) assessed in the transform for each genome were chosen such that the length scales presented on each scalogram correspond to periods between approximately 1.5 and 20% of the overall genome size.

Currently, no standard statistical methods of verifying patterns identified using continuous wavelet transforms are in common use. Thus, the significance of each transform value was ascertained by a bootstrap approach in which the order of the data points along the chromosome was randomized 200 times, and the real and imaginary portions of the Morlet wavelet transform were re-computed for each randomized data set (described previously for the real portion of the Morlet wavelet [4]). As described in the Supplementary Methods and Controls, the randomization of each genome position-associated data set was performed on either a gene-by-gene basis (for annotation-derived data) or on a kilobase-by-kilobase basis (for annotation-independent properties such as GC content). Thus, the null hypothesis against which each wavelet scalogram was tested consisted of the wavelet transform of a "scrambled" data set, where the unit of chromosome which was scrambled was either the gene or a kilobase segment. A *p*-value was then computed for each point in the scalogram based upon the number of times the magnitude of the transform value from each randomization exceeded that of the original transform. The *p*-value cutoff corresponding to a selected false discovery rate [20] (FDR $< 5\%$) was then determined from the distribution of *p*-values computed for each scalogram from the randomization tests.

7.1.3 Controls

Below are a set of positive and negative controls for the wavelet transform and bootstrap procedure described above. The negative controls showed that no significant patterns were detected in trivial or randomly-ordered data sets (for which no pattern would be expected *a priori*), thus effectively ruling out the possibility that the observed periodic patterns are simply artifacts inherent either in the wavelet filter used or spurious cyclic patterns caused by outliers in otherwise random data (called the Slutzky-Yule effect when observed in moving averages [307]).

Wavelet analysis was performed for a 1 Mb subset of the *Pseudomonas putida* GC/AT data set in order to rule out the possibility that the correlation shown in Figure 7.5a was due to an artifact of the wavelet voices chosen for the varying genome sizes. No significant decrease in fractional pattern strength was detected for the smaller subset.

Positive controls To conduct a positive control for the wavelet pattern detection methodology described above, I generated a test signal of length 4000 (comparable to the length of genome-scale data sets in *E. coli*) consisting of the sum of two sine waves, one having a period of 1/6 the overall length of the data set and the other having a period of 1/12 the overall data length (Figure 7.2a). The second test signal consisted of the first with noise added, where the magnitude of the noise varied linearly along the length of the dataset between 0 and 10 times the magnitude of the original signal, with the maximal 10× noise occurring at the midpoint of the data set, or $x = 2000$ (Figure 7.2b).

Wavelet analyses of these test signals was then performed, as described above. The scalogram corresponding to the first test signal clearly showed the two frequencies present, with only slight edge effects (Figure 7.2c). The scalogram corresponding to the noisy test signal reveals that the wavelet transform is still able to detect the lower-frequency signal across the length of the data set, even where noise was maximal (at coordinate 2000). Approximately three-fourths of the higher-frequency signal was elucidated and deemed significant (false discovery rate, or FDR, $< 5\%$, as described above). The noise in the localized region between points 1500 and 2000 obscured the original high-frequency signal over that span.

Negative controls Using FDR $< 5\%$ as the stringency cutoff, I then computed the wavelet scalogram and the significant regions of each scalogram for three trivial “data sets”: constant values (ones), uniformly random numbers between 0 and 1, and the *E. coli* gene expression data set for which the locus order has been randomized. These scalograms revealed no significant patterns at all (pattern

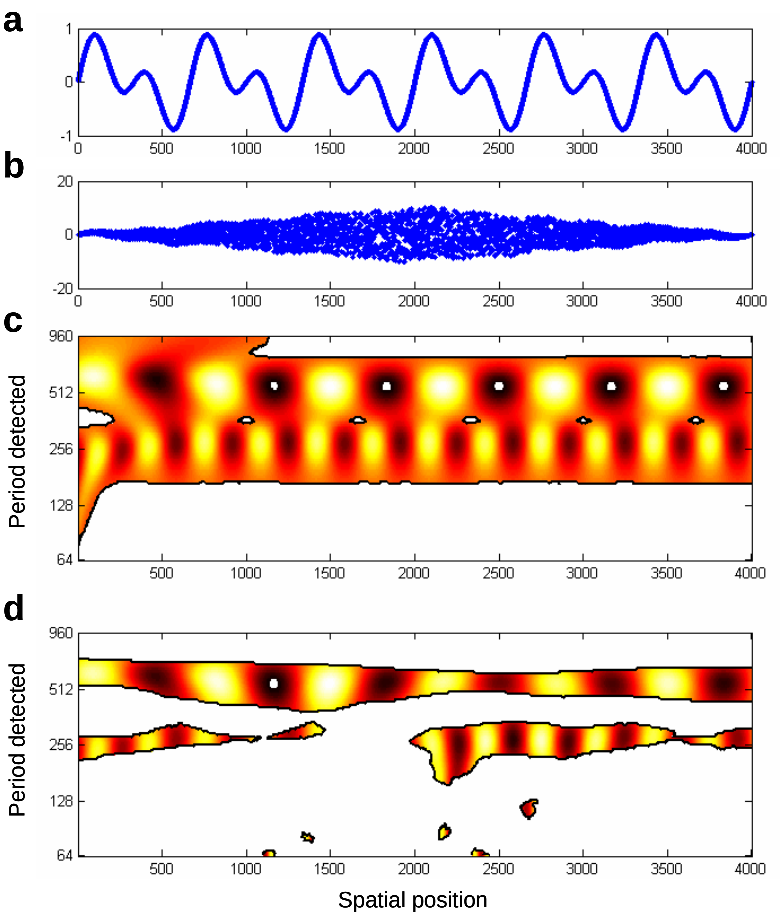


Figure 7.2: Wavelet analysis of test signals. a. Sum of two sine waves with wavelengths equal to 1/12 and 1/6 of the overall data length. b. Test signal with noise added (linearly uniform random noise between 0 and 10 times the magnitude of the original signal, with maximum noise at center of signal). c. Wavelet scalogram (real portion of Morlet wavelet) of original test signal. Insignificant portions of the scalogram from the real and imaginary portions of the Morlet wavelet transform (FDR < 5%) are shown in white. d. Wavelet scalogram of the noisy test signal shown in (b).

strength $< 1\%$).

A potential artifact that may arise when looking for any pattern or period in noisy data sets may be a spurious periodicity that is due to outliers in otherwise random data. (In moving averages of such data, this artifact is called the Slutzky-Yule effect [307].) For a data set whose dynamic range is fairly high, such as gene expression (for which the dynamic range is nearly 7000), a potential concern would be that the highly-expressed genes would introduce spurious periodicities in either moving average plots or in wavelet analysis scalograms of this data set. To test for this effect, I generated 100 data sets in which the gene expression data were randomized with respect to gene locus. For each of these "shuffled" expression data sets, I then performed the wavelet/randomization test to elucidate regions of significant pattern density (FDR $< 5\%$), and I summed binary matrices of these significant regions as determined from the real portion of the Morlet wavelet function to observe any overlap in patterns, as described above. As shown in Figure 7.3, however, the pattern overlap for the randomly-ordered gene expression data sets was minimal, reaching no higher than 18 of the 100 datasets for any point in the scalogram overlap plot. Essentially no significant patterns at all were observed, and certainly not any that even remotely resembled the patterns in any of the actual datasets considered in this study.

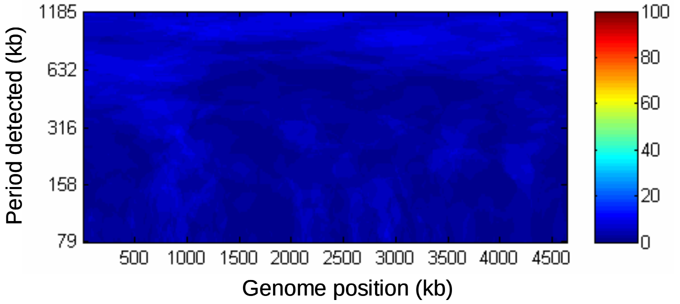


Figure 7.3: Overlap plot of significant regions of periodicity as determined from the real and imaginary portions of the Morlet wavelet scalogram (FDR $< 5\%$) for 100 randomly ordered gene expression data sets. The highest overlap at any given point was 11%, thus indicating that the observed patterns in this study are likely not an artifact resulting from the wavelet approach used.

I also tested whether the pattern strength might be an artifact due to data set length (i.e. accounting for the observation of increasing pattern strength with increasing genome length). Thus, I took short (∼500-1000 kb) segments of larger highly pattern genomes (e.g. *Pseudomonas putida*) and ran the wavelet analysis. The original higher-frequency patterns were still detected, thus ruling out a length-dependent artifact inherent in the wavelet method. Additionally, similar patterns to those observed in the data sets using the Morlet wavelet filter function in this study were also observed using another wavelet filter function designed to detect local and global periodic patterns (the Marr wavelet, also known as the "Mexican hat" wavelet). Thus, I am confident that the patterns observed in this study are indeed legitimate and are not simply an artifact resulting from the particular wavelet filter function used or the randomization procedure.

Data pre- and post-processing Since genes in prokaryotic organisms are not uniformly spaced along the genome and may even overlap slightly, the gene mid-points used as locus values for chromosomal position are not uniform. However, wavelet analysis technically only makes sense when performed on a set of uniform time points (or, in the case of this study, uniformly-spaced chromosomal loci). Thus, for all annotation-dependent datasets examined in this study (including gene expression, essentiality, ORF length, intergenic region length, etc.), it was necessary to linearly pre-interpolate the data before computing the wavelet transform values. This pre-interpolation did not introduce any spurious patterns (or alter any patterns observed in non-uniformly spaced data), probably due to the overall uniformity of ORFs in prokaryotes, as well as the absence of large intergenic regions. There was no need to pre-interpolate any of the annotation-independent data sets (e.g. G/C content, fractional gene density, etc.), as these were already uniform 1 kb segments.

Since some of the data sets (and thus the resulting scalograms) were of varying lengths, they had to be reduced in length to whichever data set contained the least number of elements in order to generate uniformly-sized binary matrices which could be summed for the overlap plots (see Figure 8.1a in the next chapter). Thus, after computing the wavelet transforms (and, for the figures, the moving averages) for each data set, I interpolated each row of the scalogram to the minimum data set length. Since these scalograms (and the resulting p-values) were smooth functions because of the nature of the wavelet filter function, I used a piecewise cubic Hermite interpolation to best preserve the observed patterns. Note that this interpolation was not performed on the primary data, but only on the wavelet scalograms and moving averages.

7.2 Patterns in multiple microbial genomes

7.2.1 Pattern strengths of sequenced prokaryotic organisms

Using the pattern detection method described in the previous section, I computed the pattern strengths for the GC/AT content, fractional gene density, and codon adaptation index (CAI) derived from 163 sequenced prokaryotic chromosomes (Figure 7.4). The average pattern strength for GC/AT content was 40% (standard deviation, or SD = 20%), for gene density was 19% (SD = 14%), and for CAI was 37% (SD = 22%). (The descriptive statistics for these distributions are summarized in Table 7.1.) The high standard deviations indicate that significant chromosome position-dependent patterns vary extensively for different organisms. The relative lack of patterning in gene density is a result of the low positional variability due to the short intergenic regions found in the generally gene-dense prokaryotic organisms. Rank-ordering the genomes by pattern strength revealed the variation in the degree of patterning in sequence-derived parameters in these chromosomes (right-hand panels of Figure 7.4). Table 7.2 lists the chromosomes containing the strongest and weakest patterns for each parameter, and the scalograms corresponding to the strongest patterns are indicated in the left-hand panels of Figure 7.4. The scalograms for *E. coli* are provided for reference (middle panels of Figure 7.4). Significant patterns were also detected in gene orientation (i.e. strand) for all but one of the chromosomes (not displayed; see Tables 7.4, 7.5, and 7.6 at the end of the chapter).

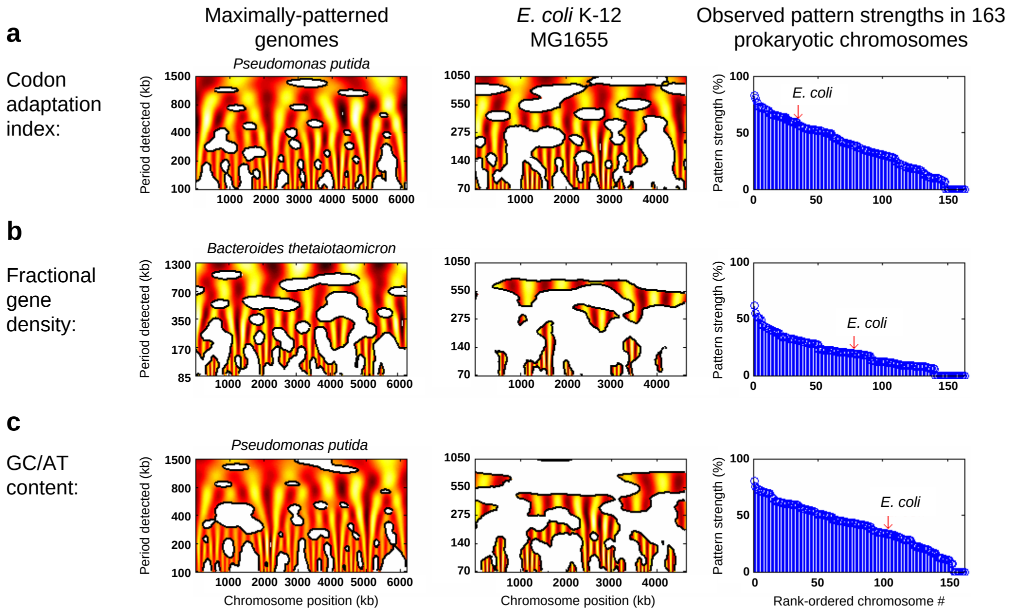


Figure 7.4: Generality of chromosome position-dependent patterns in sequence properties for 163 prokaryotic chromosomes. Continuous wavelet scalograms were computed for most prokaryotic chromosomes sequenced to date (through January 2005) to identify patterns in codon adaptation index per gene (a), fractional gene density per kilobase (b), and GC/AT content per kilobase (c). The colored portions of the scalogram indicate significant periodic patterns (FDR $< 5\%$). The degree of patterning for each prokaryotic sequence and each parameter (called the fractional pattern strength) was taken as the percentage of the area of the scalogram containing significant patterns. The first column shows the scalograms for the maximally-patterned chromosome found for each sequence property. For reference, the second column shows these scalograms for *E. coli* K-12 MG1655. The third column shows the rank-ordered fractional pattern strengths for the 163 sequenced prokaryotic chromosomes that were analyzed, with *E. coli* indicated relative to the other chromosomes on each plot.

7.2.2 Correlation of pattern strengths to organism-specific properties

Pattern strengths in the sequence-derived parameters for each chromosome were compared with global properties such as genome length, total AT composition, organism taxon, and the presence of specific nucleoid-binding proteins. Pattern strengths in CAI and GC/AT content were found to be weakly but significantly correlated with genome size ($r = 0.60$, $p = 2.4 \times 10^{-17}$; Figure 7.5a)

Table 7.1: Descriptive statistics for pattern strengths in GC/AT content, gene density, and CAI across 163 prokaryotic chromosomes. All values are percentages (%). (SD = standard deviation)

Property	**Mean**	**SD**	**Min**	**Max**
GC/AT content	39.6	19.7	0	80.0
Gene density	19.1	13.9	0	62.0
CAI	36.6	22.0	0	82.4

and anti-correlated with total AT composition ($r = -0.51$, $p = 2.0 \times 10^{-12}$; Figure 7.5b). These correlations are consistent with previously observed correlation between genome size and GC-content [22] and suggest an evolutionary requirement for greater genome organization in larger and more GC-rich organisms. However, a causal relationship among these three parameters is impossible to determine at this point. The potential evolutionary constraint regarding genome size may simply be the function of a requirement for a higher-level organization necessary to pack larger genomes into the bacterial cell. The tendency of GC-rich genomes to be more highly patterned is likely linked to physical constraints imposed by the more rigid DNA resulting from the triple hydrogen bond between guanine and cytosine.

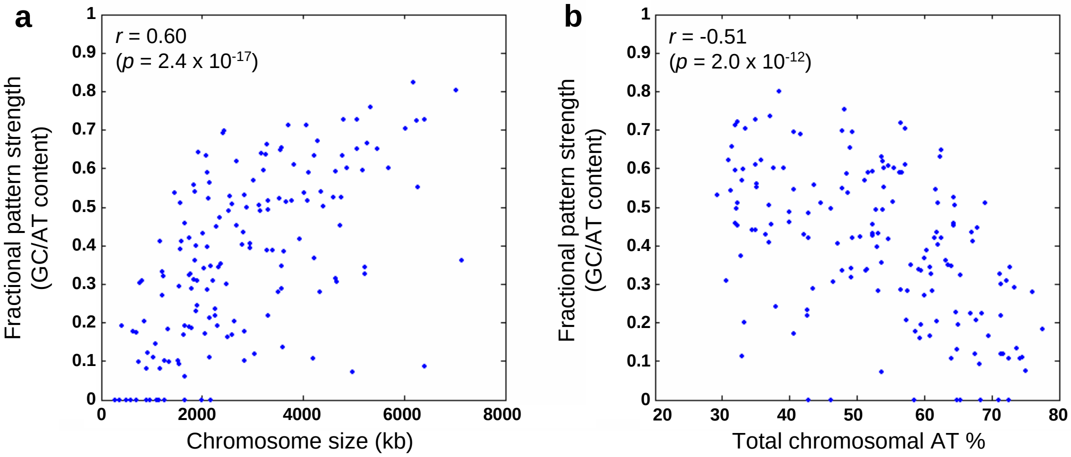


Figure 7.5: Correlations between pattern strength and organism-specific properties for 163 prokaryotic chromosomes. a. Correlation of fractional pattern strength in GC/AT content with chromosome length. b. Anti-correlation of fractional pattern strength in GC/AT content with total chromosomal AT%. The correlation coefficients and associated *p*-values are indicated on each graph.

Table 7.2: Organisms exhibiting either very high or very low chromosome position-dependent patterns in sequence-derived data (fractional pattern strength from wavelet scalograms indicated in parentheses, FDR $< 5\%$). See Supplementary Table 1 for complete listings, including strain names. All fractional pattern strengths are displayed as percentages.

Rank	**GC/AT content**	**CAI**	**Gene density**
1	*Pseudomonas putida* (80.0)	*Pseudomonas putida* (82.4)	*Bacteroides thetaiotaomicron* (62.0)
2	*Xylella fastidiosa* (75.4)	*Mesorhizobium loti* (80.4)	*Nocardia farcinica* (55.2)
3	*Mesorhizobium loti* (73.7)	*Bordetella bronchiseptica* (76.0)	*Synechocystis* sp. PCC6803 (55.2)
4	*Acinetobacter species* (72.6)	*Pseudomonas syringae* (72.7)	*Pseudomonas putida* (50.8)
5	*Burkholderia pseudomallei* chr. 1 (72.2)	*Erwinia carotovora* (72.6)	*Photorhabdus luminescens* (50.6)
6	*Bacillus subtilis* (71.8)	*Salmonella enterica* typhi (72.6)	*Bacteroides fragilis* (49.2)
7	*Bordetella bronchiseptica* (71.2)	*Bacteroides thetaiotaomicron* (72.3)	*Bifidobacterium longum* (44.7)
8	*Bacteroides thetaiotaomicron* (70.3)	*Burkholderia pseudomallei* chr. 1 (71.3)	*Bordetella bronchiseptica* (44.2)
9	*Pseudomonas aeruginosa* (70.2)	*Ralstonia solanacearum* chr. 1 (71.1)	*Burkholderia pseudomallei* chr. 2 (43.2)
10	*Geobacillus kaustophilus* (69.9)	*Nocardia farcinica* (70.4)	*Xanthomonas axonopodis* (42.6)
. . .	. . .	. . .	. . .
154	*Tropheryma whippelii* (7.3)	*Rickettsia typhi* (0)	*Leptospira interrogans* chr. 1 (0)
155	*Nanoarchaeum equitans* (0)	*Rickettsia prowazekii* (0)	*Helicobacter hepaticus* (0)
156	*Haloarcula marismortui* chr. 2 (0)	*Mycoplasma pulmonis* (0)	*Deinococcus radiodurans* chr. 1 (0)
157	*Wolbachia pipientis* (0)	*Mycoplasma genitalium* (0)	*Coxiella burnetii* (0)
158	*Thermosynechococcus elongatus* (0)	*Leptospira interrogans* chr. 2 (0)	*Corynebacterium efficiens* (0)
159	*Rickettsia typhi* (0)	*Helicobacter pylori* (0)	*Chlorobium tepidum* (0)
160	*Rickettsia prowazekii* (0)	*Fusobacterium nucleatum* (0)	*Campylobacter jejuni* (0)
161	*Parachlamydia species* (0)	*Coxiella burnetii* (0)	*Brucella Suis* chr. 1 (0)
162	*Ehrlichia ruminantium* (0)	*Borrelia garinii* (0)	*Brucella melitensis* chr. 2 (0)
163	*Anabaena nostoc* (0)	*Borrelia burgdorferi* (0)	*Brucella melitensis* chr. 1 (0)

The correlation between GC/AT pattern strength and genome length was assessed for high GC content genomes (> 50%, > 55%, > 60%, and > 65%, Figure 7.6a) and low GC content genomes (< 50%, < 45%, < 40%, and < 35%, Figure 7.6b). This analysis showed that the strong positive correlation identified and presented in Figure 7.5 holds for all cases except for the low GC content (< 40% and < 35% GC), which are presumably the least rigid genomes.

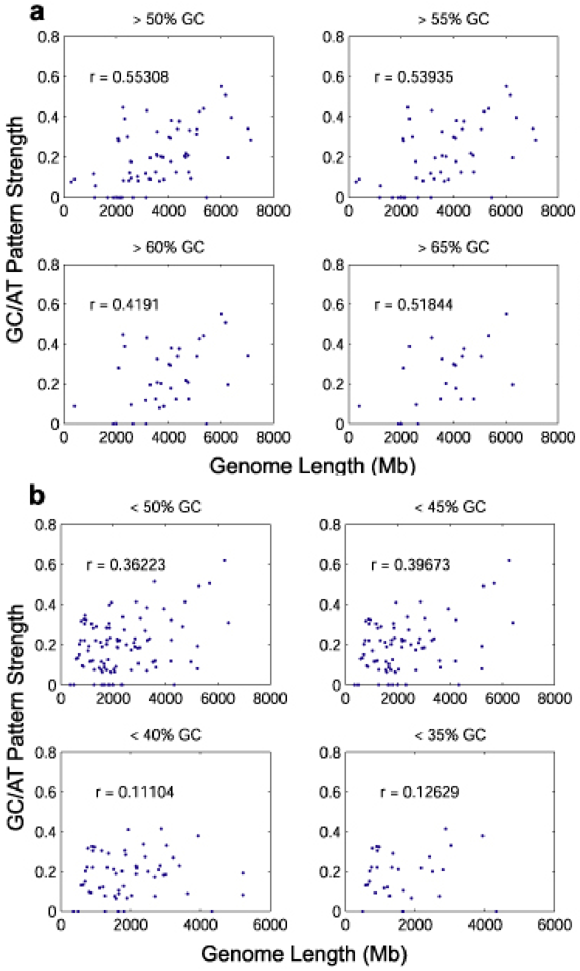


Figure 7.6: Correlation of GC/AT pattern strength and genome length for genomes of varying % GC content. Subplot titles specify GC% cutoff values (i.e. the > 50% GC plot includes genomes with greater than 50% GC content). Pearson correlation coefficient values are depicted directly on each subplot. a. High GC content genome correlations between GC/AT pattern strength and genome length. b. Low GC content genome correlations between GC/AT pattern strength and genome length.

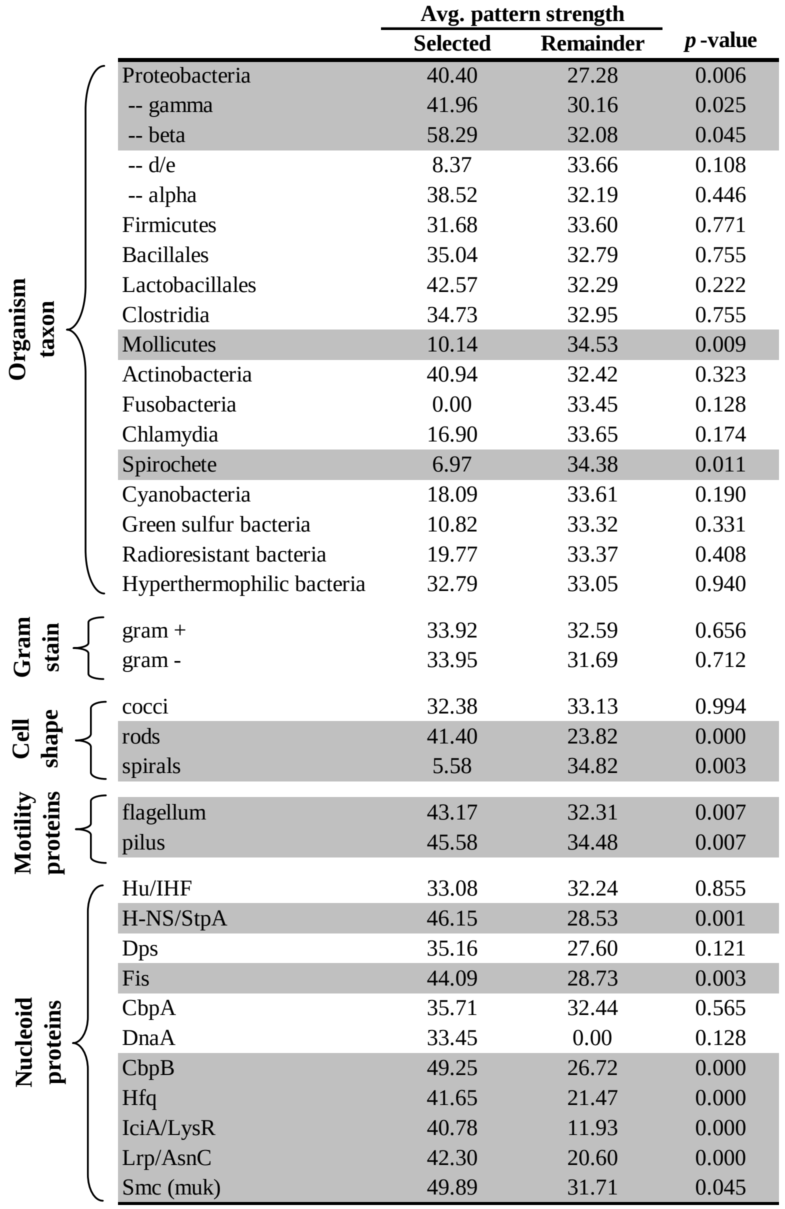
I then examined correlation of pattern strength with particular organism-specific characteristics relating to taxon, gram stain, cell shape, and the presence of particular classes of proteins in each organism (summarized in Table 7.3). The Wilcoxon rank-sum test ($p < 0.05$) was used to assess significance. With respect to organism taxa, patterns in CAI were found to be stronger among the proteobacteria and weaker among the mollicutes and spirochetes. Cell shape biases in pattern strength included a preference for stronger patterns in rod-shaped bacteria and weaker patterns in spiral-shaped bacteria. No other correlations relating to organism taxa, staining characteristics, or cell shape were observed. However, this analysis is inherently biased by the particular genomes that have been sequenced to date and are thus somewhat skewed towards enteric bacteria and pathogens. As the physiological and morphological diversity of sequenced prokaryotes increases, more definitive conclusions can be drawn regarding possible correlation between genome patterning and such properties as organism lifestyle and cell shape.

Genomes exhibiting the strongest patterns in CAI and GC/AT content had a higher likelihood (Wilcoxon rank-sum $p < 0.05$) of containing genes for flagella and pili than would be expected if the existence of these structures were uncorrelated with pattern strength. As shown in Table 7.3, the presence of genes encoding the specific nucleoid binding proteins H-NS, Fis, CbpB, Hfq, IciA, Lrp, and Muk was also found to be correlated with overall patterning in CAI. Comparisons of pattern strengths for each sequence-derived parameter revealed no significant correlations, with the exception of GC/AT content versus CAI ($r = 0.74$, $p = 5.6 \times 10^{-29}$). This correlation reflects the fact that CAI and GC/AT content are not actually independent properties, since GC-rich stretches of DNA will favor synonymous codons containing G and C.

7.3 Implications and Conclusions

As demonstrated in the analyses described above, genome sequences and sequence-derived properties are significantly patterned (i.e. non-randomly distrib-

Table 7.3: Correlation between pattern strength in codon adaptation index (CAI) and organism taxon, gram staining, cell shape, and the presence of known motility and nucleoid proteins. The "selected" column indicates the average pattern strength in the organisms meeting each criterion in the leftmost column, and the "remainder" column shows the average pattern strength of all the remaining organisms. The p-values were computed from the Wilcoxon rank-sum test, and the shaded rows met a cutoff of $p < 0.05$.



		Avg. pattern strength		
		Selected	**Remainder**	***p*-value**
Organism taxon	Proteobacteria	40.40	27.28	0.006
	-- gamma	41.96	30.16	0.025
	-- beta	58.29	32.08	0.045
	-- d/e	8.37	33.66	0.108
	-- alpha	38.52	32.19	0.446
	Firmicutes	31.68	33.60	0.771
	Bacillales	35.04	32.79	0.755
	Lactobacillales	42.57	32.29	0.222
	Clostridia	34.73	32.95	0.755
	Mollicutes	10.14	34.53	0.009
	Actinobacteria	40.94	32.42	0.323
	Fusobacteria	0.00	33.45	0.128
	Chlamydia	16.90	33.65	0.174
	Spirochete	6.97	34.38	0.011
	Cyanobacteria	18.09	33.61	0.190
	Green sulfur bacteria	10.82	33.32	0.331
	Radioresistant bacteria	19.77	33.37	0.408
	Hyperthermophilic bacteria	32.79	33.05	0.940
Gram stain	gram +	33.92	32.59	0.656
	gram -	33.95	31.69	0.712
Cell shape	cocci	32.38	33.13	0.994
	rods	41.40	23.82	0.000
	spirals	5.58	34.82	0.003
Motility proteins	flagellum	43.17	32.31	0.007
	pilus	45.58	34.48	0.007
Nucleoid proteins	Hu/IHF	33.08	32.24	0.855
	H-NS/StpA	46.15	28.53	0.001
	Dps	35.16	27.60	0.121
	Fis	44.09	28.73	0.003
	CbpA	35.71	32.44	0.565
	DnaA	33.45	0.00	0.128
	CbpB	49.25	26.72	0.000
	Hfq	41.65	21.47	0.000
	IciA/LysR	40.78	11.93	0.000
	Lrp/AsnC	42.30	20.60	0.000
	Smc (muk)	49.89	31.71	0.045

uted) with respect to chromosome position in most of the prokaryotic genomes sequenced to date (Figure 2). The degree of patterning in a bacterial organism is positively correlated with genome size, overall GC-content, the presence of several known nucleoid-binding proteins, and the presence of flagellar proteins (Figure 7.5; Table 7.3). These results strongly suggest the existence of structural constraints imposed by organism-specific features on the evolution of genome organization and base-pair composition in each organism.

In the next chapter, I will delve in greater detail into patterns in multiple heterogeneous data sets for *E. coli.* In Chapter 4, a ∼650 kb pattern was detected in gene expression data sets using wavelet analysis. In addition to gene expression, there are numerous available data for *E. coli* which include gene essentiality [105], evolutionary conservation [105] of each gene, biophysical properties as a function of nucleotide [226], and gene functional classes [281], as well as the locations of chromosome macrodomains using genetic dissection experiments [309]. The results presented in this chapter and in the next constitute early steps in the evolution of systems biology from analyses of component (1-D) and systemic (2-D) annotations [219] towards the systems analysis of three-dimensional genome organization.

The text of this chapter, in full, is a reprint of the material as it appears in T.E. Allen, N.D. Price, A.R. Joyce, and B.O. Palsson. 2005. Long-range Periodic Patterns in Microbial Genomes Indicate Significant Multi-scale Chromosomal Organization. *PLoS Computational Biology*, (in press). I was the primary author of this publication, and the co-authors participated and supervised the research which forms the basis for this chapter.

Table 7.4: Patterns in sequenced prokaryotic genomes — Part 1/3.

						Pattern strengths (%)			
				Gene	Genome	GC/AT	Gene		
Organism/strain/chromosome	Taxon ID	Taxon	AT%	count	length	content	density	CAI	Strand
Aspecies_ADP1_Main	BProt GP	Proteobacteria	59.6	3325	3599	33.4	12.2	13.7	42.0
Atumefaciens_C58_1	BProt AR	Proteobacteria	40.6	2722	2842	17.2	8.2	10.1	33.3
Atumefaciens_C58_2	BProt AR	Proteobacteria	40.7	1834	2075	54.4	29.1	63.4	38.3
Anostoc_PCC7120_Main	BCyano NN	Cyanobacteria	58.6	5366	6414	0.0	30.9	8.7	52.4
Aaeolicus_VF5_Main	BAqui AA	Aquificae	56.5	1522	1552	28.4	11.7	29.4	17.1
Aspecies_EbN1_Main	BProt BR	Proteobacteria	34.9	4133	4297	72.6	12.4	67.0	63.3
Banthracis_Ames_Main	BFirm BB	Firmicutes	64.6	11091	5228	50.5	19.2	34.2	52.0
Bcereus_ATCC10987_Main	BFirm BB	Firmicutes	64.4	5603	5225	52.4	8.2	32.6	49.2
Bhalodurans_C125_Main	BFirm BB	Firmicutes	56.3	4066	4203	58.9	11.9	10.8	53.8
Blicheniformis_ATCC14580_Main	BFirm BB	Firmicutes	53.8	4161	4223	61.9	22.3	63.3	52.1
Bsubtilis_168_Main	BFirm BB	Firmicutes	56.5	4106	4215	71.8	32.3	36.8	56.0
Bfragilis_YCH46_Main	BBBB	Bacteroidetes/Chlorobi	56.7	4578	5278	59.0	49.2	66.5	72.2
Bthetaiotaomicron_VPI5482_Main	BBBB	Bacteroidetes/Chlorobi	57.2	4778	6261	70.3	62.0	72.3	69.0
Bhenselae_Houston-1_Main	BProt AR	Proteobacteria	61.8	1612	1932	43.5	41.1	64.1	42.5
Bquintana_Toulouse_Main	BProt AR	Proteobacteria	61.2	1308	1582	28.3	19.0	38.9	48.8
Bbacteriovorus_HD100_Main	BProt DB	Proteobacteria	49.3	3583	3783	42.0	19.9	51.5	43.6
Blongum_NCC2705_Main	BActin AB	Actinobacteria	39.9	1727	2257	46.1	44.7	21.8	61.1
Bfloridanus_Strain_Main	BProt GE	Proteobacteria	72.6	589	706	10.7	20.2	17.5	32.0
Bbronchiseptica_RB50_Main	BProt BB	Proteobacteria	31.9	5006	5340	71.2	44.2	76.0	38.4
Bparapertussis_12822_Main	BProt BB	Proteobacteria	31.9	4402	4774	59.4	12.5	63.2	35.1
Bpertussis_Tohamal_Main	BProt BB	Proteobacteria	32.3	3806	4087	51.1	17.8	51.5	30.6
Bburgdorferi_B31_Main	BSpiro SS	Spirochetes	71.4	850	911	11.8	30.5	0.0	46.1
Bgarinii_PBi_Main	BSpiro SS	Spirochetes	71.7	832	905	11.8	22.1	0.0	45.6
Bmelitensis_16M_1	BProt AR	Proteobacteria	42.8	2059	2118	48.4	0.0	52.3	43.3
Bmelitensis_16M_2	BProt AR	Proteobacteria	42.7	1139	1178	23.2	0.0	41.1	29.8
BSuis_1330_1	BProt AR	Proteobacteria	42.8	2116	2108	41.8	0.0	28.4	32.2
BSuis_1330_2	BProt AR	Proteobacteria	42.7	1148	1208	21.7	5.7	33.1	20.3
Baphidicola_APS_Main	BProt GB	Proteobacteria	73.7	564	641	13.3	13.4	17.8	40.2
Bmallei_ATCC23344_1	BProt BB	Proteobacteria	31.9	2996	3511	45.8	12.5	28.0	44.2
Bmallei_ATCC23344_2	BProt BB	Proteobacteria	31	1768	2326	62.0	39.0	34.4	43.1
Bpseudomallei_K96243_1	BProt BB	Proteobacteria	32.3	3460	4075	72.2	29.4	71.3	51.7
Bpseudomallei_K96243_2	BProt BB	Proteobacteria	31.5	2395	3174	65.6	43.2	64.0	45.1
Cjejuni_NCTC11168_Main	BProt EC	Proteobacteria	69.5	1654	1642	16.7	0.0	16.7	40.6
Ccresentus_CB15_Main	BProt AC	Proteobacteria	32.8	3737	4017	37.4	29.9	53.6	40.2
Cpneumoniae_CWL029_Main	BChlam CC	Chlamydiae/Verrucomicrobia	59.4	1052	1231	16.0	28.4	31.9	29.3
Ctrachomatis_DUW_3CX_Main	BChlam CC	Chlamydiae/Verrucomicrobia	58.7	894	1043	17.7	16.9	10.8	43.2
Ccaviae_GPIC_Main	BChlam CC	Chlamydiae/Verrucomicrobia	60.8	998	1174	16.5	30.5	7.9	40.4
Ctepidum_TLS_Main	BCCC	Bacteroidetes/Chlorobi	43.5	2252	2155	28.9	0.0	10.8	40.7
Cviolaceum_ATCC12472_Main	BProt BN	Proteobacteria	35.2	4407	4752	55.2	20.8	52.5	48.6
Cacetobutylicum_ATCC824_Main	BFirm CC	Firmicutes	69.1	3672	3941	51.1	37.8	41.8	62.6
Cperfringens_13_Main	BFirm CC	Firmicutes	71.4	2660	3032	30.0	33.1	11.9	41.9
Ctetani_E88_Main	BFirm CC	Firmicutes	71.2	2373	2800	32.6	20.9	40.1	43.4
Cdiphtheriae_NCTC13129_Main	BActin AA	Actinobacteria	46.5	2320	2489	30.5	8.3	30.1	45.6
Cefficiens_YS314_Main	BActin AA	Actinobacteria	36.9	2942	3148	50.4	0.0	48.8	32.9
Cglutamicum_ATCC13032_Main	BActin AA	Actinobacteria	46.2	2993	3310	49.6	8.9	49.2	31.9
Cburnetii_RSA493_Main	BProt GL	Proteobacteria	57.3	2009	1996	20.5	0.0	0.0	45.0
Dethenogenes_195_Main	BCDDhalo	Chloroflexi	51.1	1580	1470	56.7	8.6	53.7	32.4
Dradiodurans_R1_1	BDDD	Deinococcus-Thermus	33	2579	2649	11.3	0.0	20.3	19.2
Dradiodurans_R1_2	BDDD	Deinococcus-Thermus	33.3	357	413	20.1	8.8	19.2	23.7
Dpsychrophila_LSv54_Main	BProt DD	Proteobacteria	53.2	3118	3524	43.2	11.6	52.3	49.7
Dvulgaris_Hildenborough_Main	BProt DD	Proteobacteria	36.9	3379	3571	40.8	20.7	34.6	34.8
Eruminantium_Welgevonden_Main	BProt AR	Proteobacteria	72.5	920	1517	0.0	21.1	10.0	29.0
Efaecalis_V583_Main	BFirm LE	Firmicutes	62.5	3113	3219	64.6	27.0	59.4	35.8
Ecarotovora_SCRI1043_Main	BProt GE	Proteobacteria	49	4492	5065	65.4	31.3	72.6	62.7
Ecoli_K-12_MG1655_Main	BProt GE	Proteobacteria	49.2	4289	4640	34.1	19.9	59.1	29.2
Ftularensis_SCHUS4_Main	BProt GT	Proteobacteria	67.7	1804	1893	20.7	6.6	22.9	48.5
Fnucleatum_ATCC25586_Main	BFuso FF	Fusobacterium	72.8	2067	2175	34.2	21.8	0.0	64.2
Gkaustophilus_HTA426_Main	BFirm BB	Firmicutes	47.9	3498	3545	69.9	21.3	64.8	45.7
Gsulfurreducens_PCA_Main	BProt DD	Proteobacteria	39.1	3445	3815	60.1	8.8	61.0	67.2
Gviolaceus_PCC7421_Main	BCyano CG	Cyanobacteria	38	4430	4660	24.0	21.8	30.5	16.8
Hinfluenzae_Rd_Main	BProt GP	Proteobacteria	61.8	1709	1831	20.5	30.6	31.2	39.7
Hhepaticus_ATCC51449_Main	BProt EC	Proteobacteria	64.1	1875	1800	10.7	0.0	18.7	51.4
Hpylori_26695_Main	BProt EC	Proteobacteria	61.1	1566	1668	32.6	8.8	0.0	55.5
Iloihiensis_L2TR_Main	BProt GA	Proteobacteria	53	2628	2840	39.8	19.0	53.0	39.8
Ljohnsonii_NCC533_Main	BFirm LL	Firmicutes	65.4	1821	1331	32.2	22.1	18.3	7.7
Lplantarum_WCFS1_Main	BFirm LL	Firmicutes	55.5	3051	3309	60.0	22.1	51.5	45.9

Table 7.5: Patterns in sequenced prokaryotic genomes — Part 2/3.

Llactis_lactis_Main	BFirm LS	Firmicutes	64.7	2266	2366	22.6	33.9	35.3	22.3
Lpneumophila_Philadelphia1_Main	BProt GL	Proteobacteria	61.7	2942	3398	54.5	22.9	38.7	23.7
Lxyli_CTCB07_Main	BActin AA	Actinobacteria	32.3	2030	2585	45.1	9.6	50.8	48.1
Linterrogans_56601_1	BSpiro SL	Spirochetes	65	4358	4333	19.4	0.0	27.9	43.2
Linterrogans_56601_2	BSpiro SL	Spirochetes	64.9	367	359	13.1	0.0	0.0	18.2
Linnocua_Clip11262_Main	BFirm BL	Firmicutes	62.6	2981	3012	42.1	18.5	56.7	49.3
Lmonocytogenes_EGD_Main	BFirm BL	Firmicutes	62	2855	2945	51.0	18.3	39.4	52.1
Msucciniciproducens_MBEL55E_Main	BProt GP	Proteobacteria	57.5	2384	2315	28.2	0.0	19.3	43.1
Mloti_MAFF303099_Main	BProt AR	Proteobacteria	37.2	6752	7037	73.7	34.0	80.4	38.0
Mcapsulatus_Bath_Main	BProt GM	Proteobacteria	36.4	2959	3305	42.9	19.6	21.9	63.3
Mbovis_AF212297_Main	BActin AA	Actinobacteria	34.4	3953	4346	43.9	33.9	53.8	48.9
Mleprae_TN_Main	BActin AA	Actinobacteria	42.2	2720	3269	43.0	9.2	63.6	54.1
Mtuberculosis_H37Rv_Main	BActin AA	Actinobacteria	34.4	3999	4412	43.9	37.8	50.0	52.1
Mgenitalium_G37_Main	BFirm MM	Firmicutes	68.3	480	581	9.3	13.1	0.0	7.0
Mhyopneumoniae_232_Main	BFirm MM	Firmicutes	71.4	691	893	21.9	32.7	8.0	39.6
Mmobile_163K_Main	BFirm MM	Firmicutes	75	633	778	7.4	31.7	30.1	18.6
Mmycoides_SC_Main	BFirm MM	Firmicutes	76	1016	1212	28.0	12.2	27.0	46.6
Mpenetrans_HF2_Main	BFirm MM	Firmicutes	74.3	1037	1359	10.7	29.1	9.9	32.5
Mpneumoniae_M129_Main	BFirm MM	Firmicutes	60	688	817	27.2	9.7	30.9	18.4
Mpulmonis_UAB_CTIP_Main	BFirm MM	Firmicutes	73.4	782	964	29.2	32.4	0.0	49.2
Nmeningitidis_B_MC58_Main	BProt BN	Proteobacteria	48.5	2025	2273	58.7	8.3	44.9	42.5
Neuropaea_Schmidt_Main	BProt BN	Proteobacteria	49.3	2574	2813	69.5	10.2	43.4	34.6
Nfarcinica_IFM10152_Main	BActin AA	Actinobacteria	29.2	11495	6022	53.1	55.2	70.4	43.9
Oiheyensis_HTE831_Main	BFirm BB	Firmicutes	64.3	3496	3631	45.2	8.8	38.4	53.2
Pspecies_UWE25_Main	BChlam CP	Chlamydiae/Verrucomicrobia	65.3	2031	2415	0.0	27.5	69.2	43.9
Pmultocida_Pm70_Main	BProt GP	Proteobacteria	59.6	2014	2258	19.5	22.1	23.6	31.3
Pluminescens_laumondiiTTO1_Main	BProt GE	Proteobacteria	57.2	4905	5689	61.1	50.6	60.1	73.5
Pasteris_OY_Main	BFirm MA	Firmicutes	72.3	754	861	30.9	9.2	20.3	40.2
Pgingivalis_W83_Main	BBBB	Bacteroidetes/Chlorobi	51.7	1909	2344	58.9	0.0	47.2	42.4
Pmarinus_SS120_Main	BCyano PP	Cyanobacteria	63.6	1882	1752	35.0	20.1	41.9	19.0
Pacnes_KPA171202_Main	BActin AA	Actinobacteria	40	2297	2561	48.6	17.8	52.7	43.8
Paeruginosa_PAO1_Main	BProt GP	Proteobacteria	33.4	5566	6265	70.2	19.7	55.0	40.3
Pputida_KT2440_Main	BProt GP	Proteobacteria	38.5	5350	6182	80.0	50.8	82.4	29.3
Psyringae_DC3000_Main	BProt GP	Proteobacteria	41.6	5471	6398	69.0	39.5	72.7	39.6
Rsolanacearum_GMI1000_1	BProt BR	Proteobacteria	33	3442	3717	56.8	20.2	71.1	38.9
Rsolanacearum_GMI1000_2	BProt BR	Proteobacteria	33.1	1678	2095	59.9	28.0	58.8	53.1
Rbaltica_strain1_Main	BPPP	Planctomycetes	44.6	7325	7146	50.9	28.3	36.1	37.4
Rpalustris_CGA009_Main	BProt AR	Proteobacteria	35	4831	5460	43.9	0.0	65.1	49.4
Rconorii_Malish7_Main	BProt AR	Proteobacteria	67.6	1374	1269	12.0	7.7	10.0	44.1
Rprowazekii_Madrid-E_Main	BProt AR	Proteobacteria	71	834	1112	0.0	11.7	0.0	34.6
Rtyphi_Wilmington_Main	BProt AR	Proteobacteria	71.1	838	1112	0.0	11.8	0.0	34.4
Senterica_typhiCT18_Main	BProt GE	Proteobacteria	47.9	4600	4810	54.9	33.2	72.6	50.8
Styphimurium_LT2_Main	BProt GE	Proteobacteria	47.8	4452	4858	33.5	9.4	60.0	34.0
Soneidensis_MR1_Main	BProt GA	Proteobacteria	54	4630	4970	55.0	11.6	7.0	42.6
Sflexneri_2a301_Main	BProt GE	Proteobacteria	49.1	4434	4608	31.7	20.5	52.6	31.8
Spomeroyi_DSS3_Main	BProt AR	Proteobacteria	35.8	3810	4110	62.1	38.1	59.0	30.8
Smeliloti_Rm1021_Main	BProt AR	Proteobacteria	37.3	3341	3655	45.4	8.0	51.4	21.9
Saureus_Mu50_Main	BFirm BS	Firmicutes	67.1	2714	2879	43.4	41.5	49.9	50.2
Sepidermidis_ATCC12228_Main	BFirm BS	Firmicutes	67.9	2419	2500	44.6	20.0	16.3	34.6
Sagalactiae_V2603_Main	BFirm LS	Firmicutes	64.3	2124	2161	45.8	18.8	56.2	34.5
Smutans_UAB159_Main	BFirm LS	Firmicutes	63.2	1960	2031	36.0	7.8	34.1	36.3
Spneumoniae_TIGR4_Main	BFirm LS	Firmicutes	60.3	2094	2161	38.7	22.7	21.3	36.7
Spyogenes_SF370_Main	BFirm LS	Firmicutes	61.5	1696	1853	42.0	28.4	53.8	31.3
Sthermophilus_CNRZ1066_Main	BFirm LS	Firmicutes	60.9	1915	1797	34.3	12.7	28.7	31.0
Sthermophilum_Strain_Main	BActin S	Actinobacteria	31.3	3337	3567	54.3	32.6	65.5	54.7
Ssp_WH8102_Main	BCyano CS	Cyanobacteria	40.6	2526	2435	69.6	30.1	69.6	41.2
SPCC6803_Strain_Main	BCyano CS	Cyanobacteria	52.3	3169	3574	43.2	51.5	28.8	21.3
Ttengcongensis_MB4T_Main	BFirm CT	Firmicutes	62.4	2588	2690	63.0	17.3	45.2	45.9
Telongatus_BP1_Main	BCyano CT	Cyanobacteria	46.1	2475	2594	0.0	17.6	16.8	24.4
Tmaritima_MSB8_Main	BThermt TT	Thermotogales	53.8	1846	1861	49.3	0.0	36.2	55.1
Tthermophilus_HB27_Main	BDDT	Deinococcus-Thermus	30.6	1982	1895	31.0	0.0	24.5	30.6
Tdenticola_ATCC35405_Main	BSpiro SS	Spirochetes	62.1	2767	2844	40.2	22.9	17.7	51.8
Tpallidum_Nichols_Main	BSpiro SS	Spirochetes	47.2	1031	1139	40.5	11.6	0.0	38.4
Twhippelii_TW0827_Main	BActin AA	Actinobacteria	53.7	784	926	7.3	34.6	12.2	18.6
Uurealyticum_serovar3_Main	BFirm MM	Firmicutes	74.5	611	752	11.0	22.2	9.9	29.9
Vcholerae_N16961_1	BProt GV	Proteobacteria	52.3	2736	2962	59.3	10.8	40.5	12.8
Vcholerae_N16961_2	BProt GV	Proteobacteria	53.1	1092	1073	28.2	19.2	14.4	28.1
Vparahaemolyticus_RIMD2210633_1	BProt GV	Proteobacteria	54.6	3080	3289	60.5	38.4	38.6	38.2
Vparahaemolyticus_RIMD2210633_2	BProt GV	Proteobacteria	54.7	1752	1878	41.8	23.3	40.0	20.8

Table 7.6: Patterns in sequenced prokaryotic genomes — Part 3/3.

Vvulnificus_CMCP6_1	BProt GV	Proteobacteria	53.6	2972	3282	62.9	7.2	66.2	41.0
Vvulnificus_CMCP6_2	BProt GV	Proteobacteria	52.9	1565	1845	49.2	39.8	55.7	21.4
Wglossinidia_Strain_Main	BProt GE	Proteobacteria	77.5	654	698	18.4	15.1	0.0	31.5
Wpipientis_wMel_Main	BProt AR	Proteobacteria	64.8	1195	1268	0.0	0.0	0.0	34.1
Wsuccinogenes_DSMZ1740_Main	BProt EC	Proteobacteria	51.5	2044	2111	33.8	9.2	39.7	30.3
Xaxonopodis_citri306_Main	BProt GX	Proteobacteria	35.2	4312	5176	56.0	42.6	59.6	55.0
Xcampestris_ATCC33913_Main	BProt GX	Proteobacteria	34.9	4181	5077	60.9	33.8	65.0	59.9
Xfastidiosa_Temecula1_Main	BProt GX	Proteobacteria	48.2	2034	2520	75.4	12.2	48.9	43.0
Ypestis_CO-92BiovarOrientalis_Main	BProt GE	Proteobacteria	52.4	4008	4654	45.4	29.0	31.4	59.3
Ypseudotuberculosis_IP32953_Main	BProt GE	Proteobacteria	52.4	3974	4745	42.5	41.4	45.3	45.8
Zmobilis_ZM4_Main	BProt AS	Proteobacteria	53.7	1998	2057	35.6	6.5	17.2	23.9
Archaea:									
Apernix_K1_Main	ACTD	Crenarchaeota	43.7	1841	1670	55.8	0.0	45.9	29.8
Afulgidus_DSM4304_Main	AEAAA	Euryarchaeota	51.4	2407	2179	33.5	18.5	34.7	42.4
Hmarismortui_ATCC43049_1	AEHH	Euryarchaeota	37.6	6382	3132	60.1	11.8	50.4	21.1
Hmarismortui_ATCC43049_2	AEHH	Euryarchaeota	42.8	579	289	0.0	7.7	0.0	0.0
Hspecies_NRC1_Main	AEHH	Euryarchaeota	32.1	2058	2006	49.6	0.0	43.1	11.6
Mthermoautotrophicum_delta-H_Main	AEMbMM	Euryarchaeota	50.5	1869	1752	42.2	28.4	32.3	39.1
Mjannaschii_DSM2661_Main	AEMcM	Euryarchaeota	68.6	1715	1665	22.3	0.0	19.1	31.6
Mmaripaludis_S2_Main	AEMMc	Euryarchaeota	66.9	1722	1662	22.5	10.5	6.1	31.5
Nequitans_Kin4M_Main	ANN	Nanoarchaeota	68.4	563	491	0.0	0.0	0.0	6.4
Ptorridus_DSM9790_Main	AETT	Euryarchaeota	64	1535	1546	34.8	7.6	9.3	16.9
Paerophilum_IM2_Main	ACTTTP	Crenarchaeota	48.6	5402	2223	53.8	0.0	30.8	20.8
Pabyssi_GE5_Main	AETTP	Euryarchaeota	55.3	1765	1766	51.2	22.2	32.7	38.8
Pfuriosus_DSM3638_Main	AETTTP	Euryarchaeota	59.2	2065	1909	33.7	6.2	30.7	41.5
Phorikoshii_OT3_Main	AETT	Euryarchaeota	58.1	1956	1739	35.1	7.6	18.9	32.7
Stokodaii_7_Main	ACSSS	Crenarchaeota	67.2	2826	2695	41.1	7.5	62.0	22.1
Tacidophilum_DSM1728_Main	AETTT	Euryarchaeota	54	1478	1565	60.1	0.0	50.9	17.9
Tvolcanium_GSS1_Main	AETTT	Euryarchaeota	60.1	1499	1585	36.7	6.4	40.9	27.8

Data summary:	
Bacterial chromosomes:	146
Archaeal chromosomes:	17
	163
Bacterial species:	135
Archaeal species:	16
	151

Chapter 8

Patterns in *Escherichia coli* Datasets

As described in the previous chapter, bacterial genomes contain a wide range of spatial patterns and regularity in several sequence-derived properties as a function of chromosome position. This motivates an assessment of pattern content in genome location-dependent data in a well-studied model organism. *Escherichia coli* is an ideal organism in which to conduct such an analysis given that gene expression data [4], gene essentiality data [105], the evolutionary conservation of each gene [105], sequence-derived biophysical and bending parameters [226], and well-curated gene classifications [281] are all available, in addition to "raw" sequence-derived properties [26] as discussed in the previous chapter.

Furthermore, a study of genome position-dependent patterns in heterogeneous data types in a well-studied model organism such as *E. coli* (e.g. gene expression vs. specific codon preferences) may reveal properties that are spatially linked. Thus, there is a need to determine the spatial coupling of multiple heterogeneous properties for a well-studied model organism. Accordingly, in this chapter I describe the methodology and results for the examination of disparate genome position-dependent data available for *E. coli* to determine properties that are spatially correlated over multiple length scales. This analysis of patterns in multiple

E. coli data sets supports the notion that the overall organization of the bacterial chromosome results from the simultaneous optimization of functional and structural constraints.

In addition to analyzing pattern overlaps and correlations, I will present analyses of cross-correlations of pattern content in 20 fundamentally distinct *E. coli* data sets, totaling 200 different data types (which includes differing counts of codon content per gene and amino acid demands as a function of chromosome position). These cross-correlations were examined globally over the entire length of the chromosome, as well as locally in 12 sub-chromosomal segments (demarcated according to the regions of high and low gene expression identified in Figure 4.3.

The nonrandom nature of gene expression (and numerous other properties) along the *E. coli* chromosome motivates an experimental assessment of the impact of large chromosomal inversions on gene expression patterns and on fitness. I have thus developed a prospective experimental design procedure used to identify inversions that are predicted to most disrupt the strong 650 kb pattern observed in gene expression. These results point toward experiments that may shed light on the biological importance of large-scale positional biases in genome location-dependent properties in a model organism such as *E. coli*.

8.1 Pattern correlation in *E. coli* data sets

8.1.1 Computing pattern overlay plots

As described in the previous chapter, the p-value cutoff corresponding to a selected false discovery rate [20] (FDR $<$ 5%) was determined from the distribution of p-values computed for each scalogram from 200 randomization tests per data set. Given this cutoff, one can generate a binary matrix (the same size as the scalogram) containing unity for each point in the scalogram for which FDR $<$ 0.05, and zeros elsewhere. The ratio of the sum of the non-zero elements in this binary matrix to the total matrix size is taken to be the pattern strength of a given data set (colored areas in Figure 7.1b). For matrices of the same size (as for *E. coli* gene expression, essentiality, and evolutionary retention), the sum of the binary significance matrices yields the degree of pattern overlap, as illustrated schematically in Figure 8.1a. The analysis of data sets from *E. coli* K-12 MG1655 included sequence-derived biophysical parameters averaged across 1 kb segments [226], gene classifications and product locations [281], gene expression [4], gene essentiality [105], and evolutionary retention indices computed based upon homology with 32 representative bacterial sequences [105].

8.1.2 Overlap of patterns in heterogeneous data sets in *E. coli*

Since a 600-650 kb periodic pattern has previously been detected in *E. coli* gene expression [4, 156, 184], the above results motivated an assessment of chromosome position-dependent patterns in functional properties specifically in *E. coli* (in addition to the patterns in GC/AT content, CAI, gene density, and gene orientation discussed above). Correlation of similar patterns in these heterogeneous data sets allows for an evaluation of the structural and functional organization of the *E. coli* genome. Binary matrices of significant pattern density regions were generated for a p-value cutoff corresponding to a specified false discovery rate [20] (FDR $<$ 5% for our analysis). Unity was assigned to regions in the scalogram deemed to have statistically significant patterning and zeros assigned elsewhere (see Figure 8.1a). For any given collection of data sets, the corresponding binary pattern-significance matrices can then be collated and visualized as a contour plot to reveal the extent of overlap in regions of the wavelet scalograms sharing significant p-values (Figure 8.1a).

Analysis of pattern overlap among functional genome position-dependent data sets in *E. coli* revealed that gene expression [4], gene essentiality [105], and the evolutionary retention index [105] contain significant periodic patterns overlapping at the 650 kb length scale (Figure 8.1b) and are strongly (positively) correlated (Figure 8.1c). Significant patterns in gene expression at the 600-650 kb period were

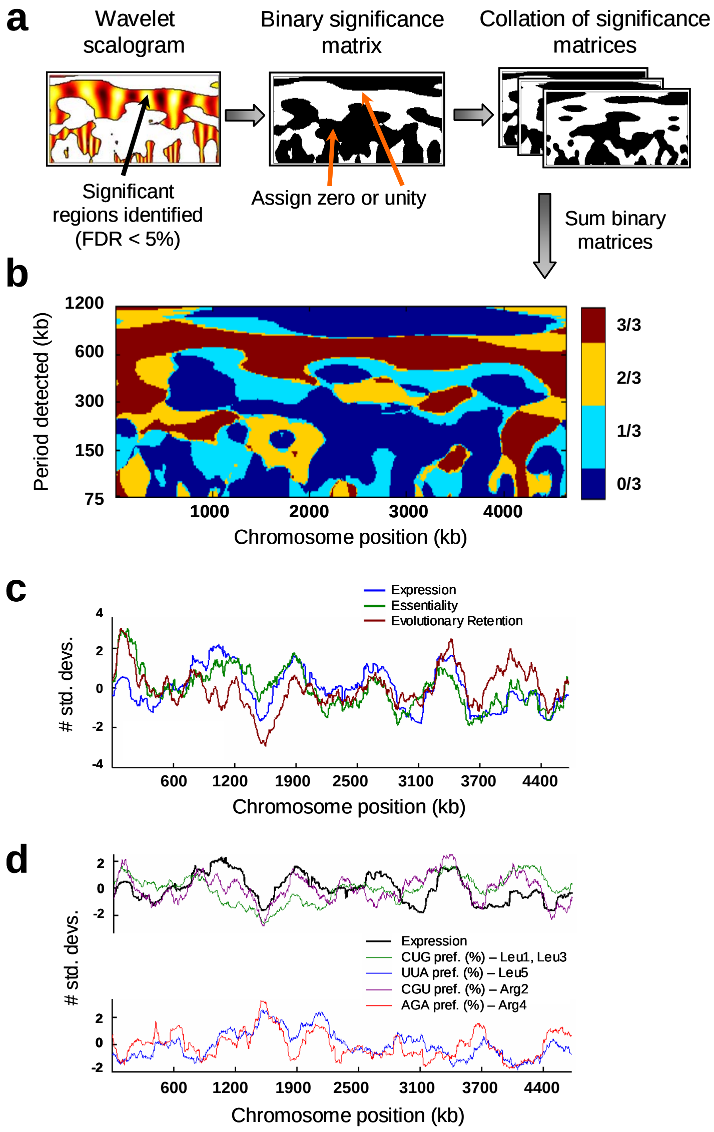


Figure 8.1: Correlation of specific chromosome position-dependent patterns in *E. coli* functional properties. a. Wavelet scalograms calculated for gene expression, gene essentiality, and evolutionary retention index were converted to a binary significance matrix by setting each significant point in a scalogram (FDR $<$ 5%) to unity and each non-significant point to zero. b. These binary matrices were summed across the three properties listed above to determine chromosome position-dependent patterns that were consistent across the different properties, and the resulting map was color-coded according to how many of the properties shared significant patterns. The red-colored segments indicate the periods and chromosome positions at which all three properties exhibited significant patterns. The averaged data have been normalized such that the mean is zero and the tick marks indicate standard deviations from the mean value. c. Correlation of gene expression, essentiality, and evolutionary retention averaged at a window of 325 kb (650 kb period). d. Correlation of intragenic codon preferences for two of the major codons encoding leucine (CUG) and arginine (CGU) and of expression, and anti-correlation of these with preferences for minor codons (UUA and AGA, respectively), at a moving average of 325 kb. The labels are as described above in (c).

also found to overlap with patterns in fractional gene density and CAI over most of the genome (Figure 8.2a). This observation is consistent with the known coupling of transcription and translation in prokaryotes [200], since shared positional biases in CAI and expression imply that codon usage (which affects translation) is spatially coupled to gene expression (transcription). Additionally, large-scale periodic patterns (most at the ∼650 kb length scale) in the intragenic preference of specific synonymous codons were detected in *E. coli*, implying consequent positional biases in the corresponding tRNA species. Thus, certain tRNA species will be preferentially demanded over specific regions of the chromosome; e.g. different tRNAs for arginine and lysine will be demanded at regions of either high or low gene expression at the 600-650 kb length scale (Figure 8.1d). The observed chromosome-position biases in gene expression and specific codon preferences in *E. coli*, along with the codon adaptation patterns observed in most of the 163 prokaryotic chromosomes analyzed in this study, suggest the existence of spatial gradients in the functional state of specific domains within each folded nucleoid [255]. These gradients may lead to spatial gradients in tRNA concentration that result from differential local demands for specific tRNA species [59].

The analysis of all 163 chromosomes presented in the previous chapter revealed that long-range patterns in synonymous codon usage (CAI) are not strictly independent from those in GC/AT composition. However, patterns in sequence-derived DNA bending parameters for *E. coli* (e.g. intrinsic curvature, propeller twist, stacking energy, etc.) almost completely overlap with patterns in GC/AT content (Figure 8.2b). As described previously, the GC/AT content reflects the average bendability of the chromosome over multiple length scales [226]. Thus, the observed correlation of pattern strengths in CAI and GC/AT content implies a general coupling of information storage with chromosomal bending. The strongest overlap in nucleotide sequence content and sequence-derived bending parameters in *E. coli* consists of a 600-650 kb periodic pattern near the origin of replication between the 3800-250 kb nucleotide coordinates ($82'$ to $5'$). This region closely

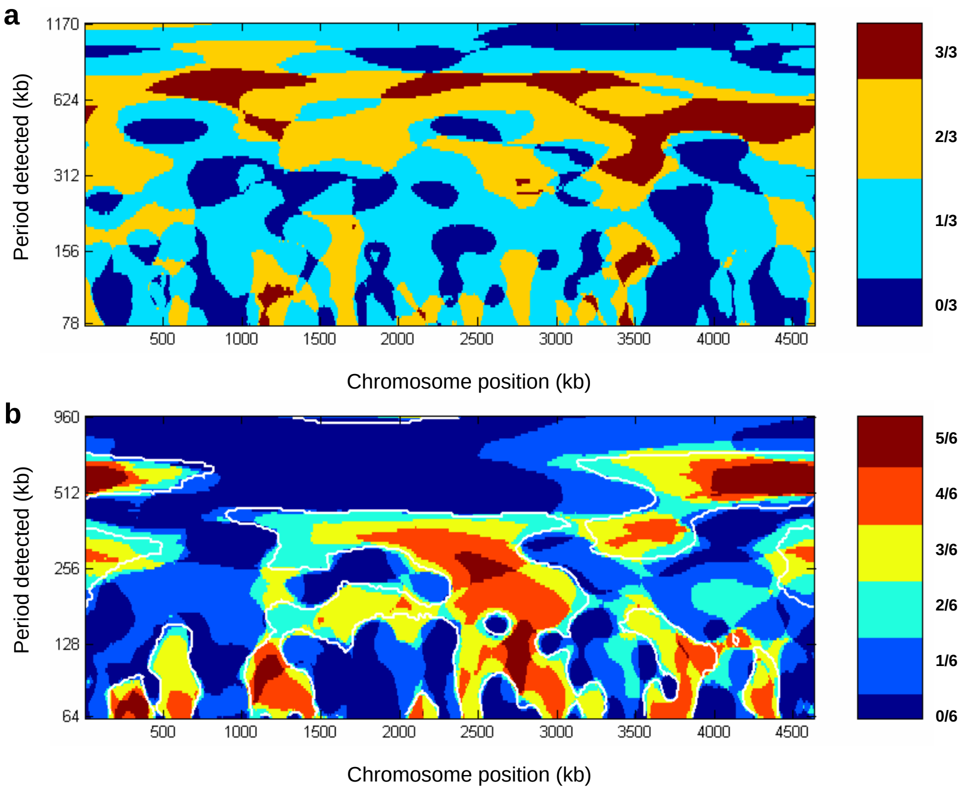


Figure 8.2: Overlay plots of significant regions of wavelet scalograms for various *E. coli* parameters. a. Degree of significant pattern overlap in expression, gene density, and codon adaptation in *E. coli.* Binary matrices corresponding to significant regions of wavelet scalograms (FDR $<$ 5%) for gene expression, codon adaptation index (CAI), and fractional gene density in *E. coli* were summed as described above. A periodic pattern of 600-650 kb can be seen across nearly three-quarters of the chromosome. b. Degree of significant pattern overlap sequence-derived DNA-bending parameters in *E. coli.* Binary matrices corresponding to significant regions of wavelet scalograms (FDR $<$ 5%) for intrinsic curvature, DNAseI sensitivity, protein-induced deformability, propeller twist, stacking energy, and nucleosome position preference in *E. coli* [226] were summed as described in the text. The white contour lines outline the significant regions of the wavelet scalogram for GC/AT content, thus demonstrating that these parameters are not independent.

coincides with the *E. coli* origin macrodomain detected in previous studies [309], and the structural regularity at the 600-kb length scale may facilitate localization of the origin to one of the cell poles during replication [206]. These DNA-bending associated data sets also contain localized periodic patterns at length scales on the order of 80-100 kb that occur in specific regions of the chromosome. The maximum pattern density in GC/AT content in this range occurred at the 74 kb period, containing eight localized patterns. Six of these eight pattern-rich segments were found to be significantly enriched (hypergeometric $p < 0.001$) with genes belonging to particular functional classes [281] which included prophage-related genes and genes encoding membrane-associated proteins (flagellar, energy production and transport, and cell surface antigens). The enrichment of patterned regions with genes of extrachromosomal origin implies a preferred regularity in chromosome structure and nucleotide content that facilitates foreign DNA incorporation. In the case of the regions enriched in membrane-associated proteins (flagellar, cell surface, etc.), the translocation of these proteins [338] may be enhanced by regular structure at the 80-100 kb length scale.

Genome topology has recently been shown to be a selection target in the long-term evolution of *E. coli* [57]. Our results demonstrate that prokaryotic genomes generally possess significant organization that increases with genome size, overall GC composition, and the presence of several known nucleoid-binding proteins. Thus, genome composition and size may impose additional constraints on the evolution of gene order and chromosomal arrangement in prokaryotes. Given that the spatial organization of chromosomal loci within a replicating *E. coli* cell is linearly ordered along the cellular axis [206, 30], the analysis presented here would imply the existence of 6 "sub-chromosomal" functional domains in the *E. coli* genome [184]. This notion of highly expressed topological domains has been suggested before [308] and is consistent with the macrodomains elucidated by genetic dissection of *E. coli* [309]. The boundaries of those four domains and two less-structured regions [309] align with the boundaries of the regions of high and low gene expression, gene essentiality, and evolutionary retention in *E. coli* at the 600-650 kb length scale (Figure 8.3). The observed patterns reveal that information transfer and chromosomal organization within the *E. coli* nucleoid are spatially interlinked.

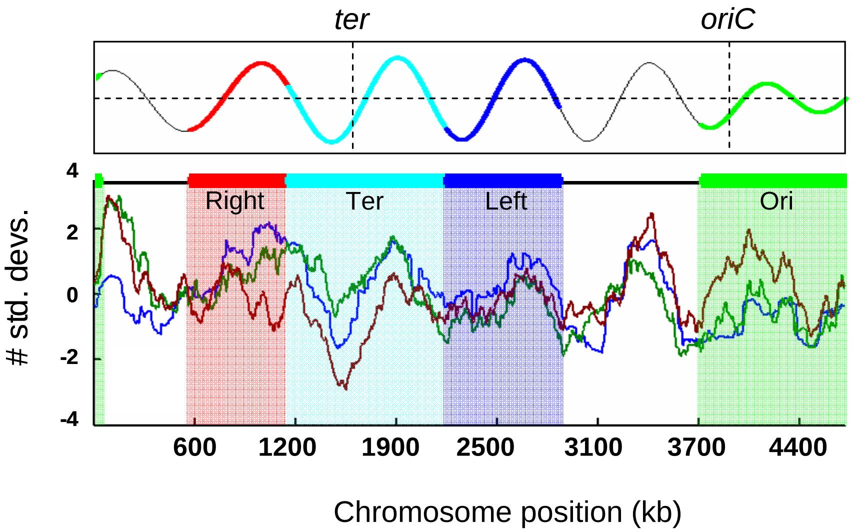


Figure 8.3: Comparison of *E. coli* gene expression, essentiality, and evolutionary retention at 600-650 kb length scale with experimentally-identified chromosome macrodomains [309]. The four shaded regions correspond to four macrodomains identified previously based upon the frequency of recombination events following genetic dissection of the *E. coli* chromosome. The two unshaded regions correspond to less-structured macrodomains. The traces in the lower panel are exactly as described in Figure 8.1c. The upper panel is a section of the wavelet scalogram for *E. coli* gene expression at a 650 kb period. Segments of this wavelet transform trace have been colored to correspond to the experimentally-identified chromosome macrodomains.

8.2 Cross-correlation of patterns

The method presented above was shown to be useful for revealing general patterns and pattern correlations in *E. coli* at similar length scales using an overlay visualization plot (Figure 8.1). However, this method does not currently distinguish between pattern correlation or anti-correlation since that information is lost when both the real and imaginary portions of the Morlet wavelet function are considered. An additional sort of pattern overlap not elucidated by the above general method involves the localization of higher-frequency periodic patterns (at the 80 kb length scale) preferentially within regions of low gene expression at the 650 kb length scale. Thus, an automated method is needed by which to identify specific pattern correlations, anti-correlations, and cross-correlations in genome position-associated data.

8.2.1 Method for determining pattern cross-correlations

In order to systematize the elucidation of this sort of pattern preference (in addition to correlations at similar scales), I developed the following method: For two genome position-associated datasets (e.g. gene expression and evolutionary retention index; Figure 8.4a), the degree of overlap in the genome loci corresponding to regions of significant localized periodicity is computed at each pair-wise length scale. This overlap has been termed the pattern density cross-correlation (PDC) "strength," defined as the percentage of loci in both data sets sharing significant Morlet wavelet transform values (Figure 8.4b). Once computed for each pair of scales between two data sets, a PDC "landscape" is obtained. Using image processing tools, one can compute the scales corresponding to the local maxima in each PDC landscape (Figure 8.4c).

The interpretation of a PDC landscape plot is given schematically in Figure 8.5a. The strong correlations along the diagonal of a PDC landscape contour plot reveal correlation at similar scales, similar to those revealed by visual inspection detected by the method diagrammed in Figure 8.1. The off-diagonals in a PDC landscape plot reveal preferential regularities in which high-frequency periodic patterns are clustered within particular high or low regions of another parameter at a lower frequency (e.g. the preferential clustering of high-frequency patterns in biophysical parameters [226] within regions of low expression [4]; Figure 8.5b-c).

8.2.2 Cross-correlations in *E. coli* patterns

Scoring scheme for correlations The results in Figure 8.5 demonstrate that low-expression regions of the *E. coli* chromosome contain regular (high-frequency)

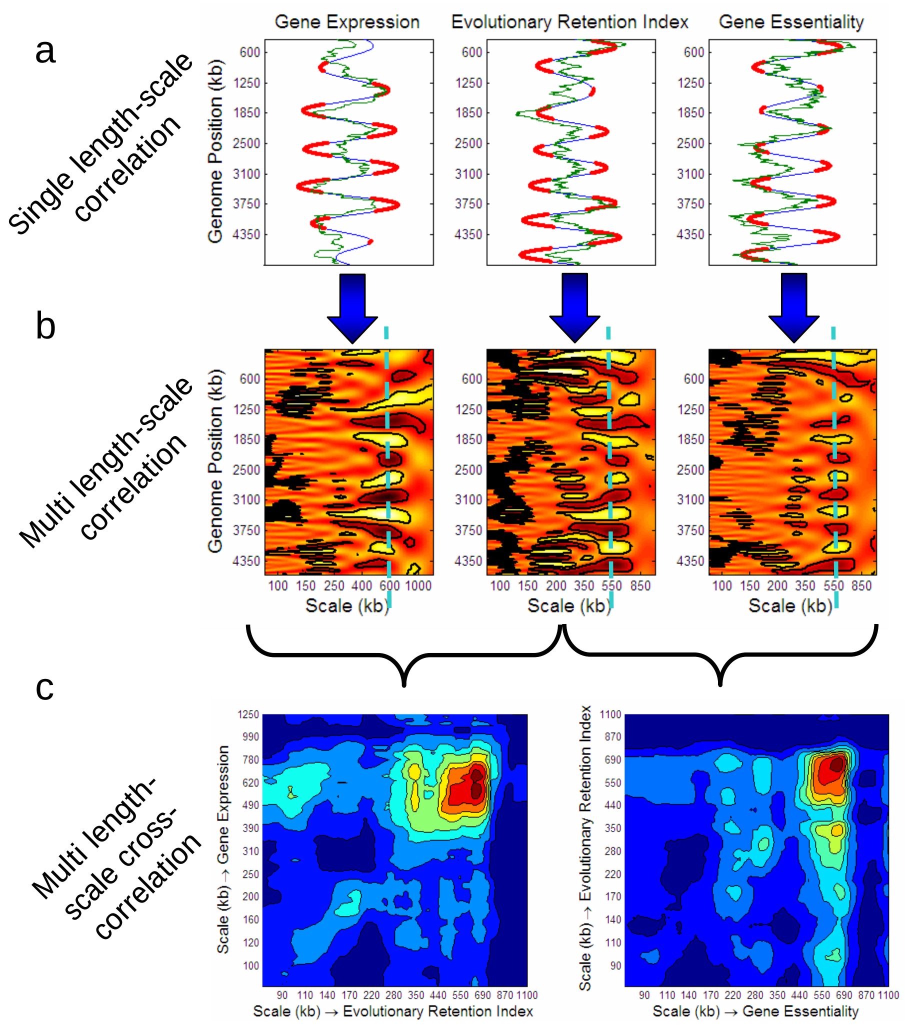


Figure 8.4: Identification of cross-correlations in multiple genome location-dependent data sets in *E. coli.* This schematic illustrates the progression in methodology from single length scale correlations to multi-length scale correlations to multi-length scale cross-correlations. The contour plots at the lower portion of the figure are pattern density cross-correlation (PDC) "landscape" plots, as described in the text.

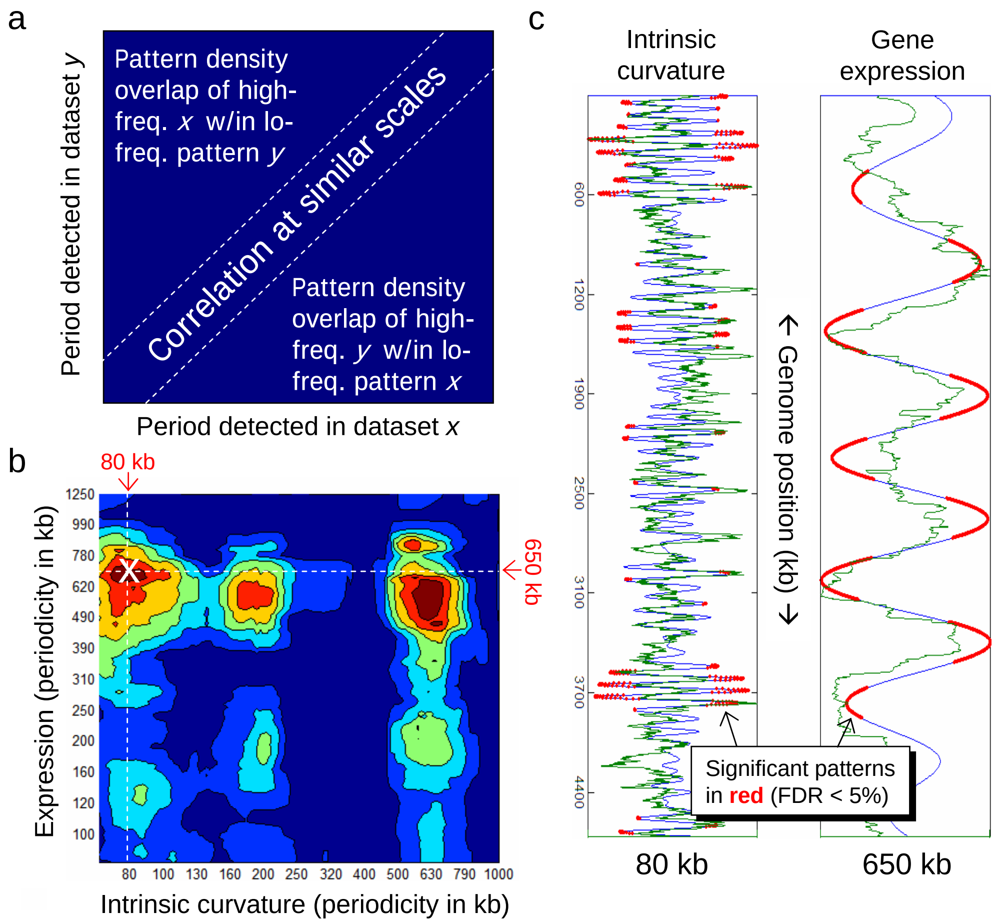


Figure 8.5: Pattern cross-correlation between two *E. coli* data sets. a. Schematic representation of the interpretation of PDC landscape contour plots (described in the text). b. PDC landscape for gene expression [4] and intrinsic curvature [226] in *E. coli*. Peaks were identified at the 80 kb, 200 kb, and 650 kb periods in curvature, all at the 650 kb period in gene expression. The strongest cross-correlation (at the 650 kb period in expression and the 80 kb period in curvature) is marked by a white ×. c. Trace of the wavelet transform (blue) and moving average (green) for the two data sets at the periodicities marked by the × in (b). The regions of significant high or low transform values (FDR < 5% for the real portion of the Morlet wavelet) are highlighted in red.

periodicities in intrinsic curvature (computed from the sequence [226]). In addition, the pattern correlations revealed by the overlay method in the previous section were confirmed for gene expression and evolutionary conservation using the PDC landscape methodology (Figure 8.4). A quantitative scoring scheme will be needed for two different scenarios:

1. The degree of pattern correlation or anti-correlation for the PDC peaks that fall along the diagonal, $r_{\mathrm{pos/neg}}$; and
2. The degree to which high-frequency patterns are preferentially localized within regions comprising either high or low transform values of a lower-frequency pattern (the off-diagonals), $r_{\mathrm{hi/lo}}$.

To compute these measures of correlation/anti-correlation and biased high-frequency pattern enrichment, we must first determine the real Morlet wavelet transform "slice" at a particular scale for each data set. (The length scales for each pair-wise comparison were chosen by the method described above.) For a pair of periods, a "density" vector, d_i (for each i nucleotide coordinate) can be defined:

$$d_i = \begin{cases} 1, & \text{FDR values for } \textit{both} \text{ sets of } p\text{-values} < 5\% \\ 0, & \text{otherwise} \end{cases} \tag{8.1}$$

An element in this density vector equals unity if the transform values of both data sets are deemed significant for the corresponding genome position coordinates, and zero otherwise. In other words, this binary vector represents the extent of overlap between the significant real Morlet wavelet scalograms at selected length scales. If the wavelet transform values at each length scale in a pair are defined as $\mathrm{cw}_{i,1}$ and $\mathrm{cw}_{i,2}$, the degree of correlation at high significant transform values (between two data sets) can be differentiated from correlation at low significant transform values:

$$\begin{aligned} \mathrm{hi}_{i,j} &= d_i \times (\mathrm{cw}_{i,j} > \mathrm{mean}(\mathrm{cw}_{i,j})) \\ \mathrm{lo}_{i,j} &= d_i \times (\mathrm{cw}_{i,j} < \mathrm{mean}(\mathrm{cw}_{i,j})) \end{aligned} \tag{8.2}$$

The index j refers to the data set in the pairwise comparison (either 1 or 2). The correlation at both high and low transform values (i.e. the positive or negative correlation), normalized to the density sum such that a perfectly correlated pair of transform slices yields a value of 1 and a perfectly anti-correlated pair yields -1, is given by the expression:

$$r_{\text{pos/neg}} = \frac{2 \times \sum_i (\text{hi}_{i,1}\text{hi}_{i,2} + \text{lo}_{i,1}\text{lo}_{i,2})}{\sum_i d_i} - 1 \tag{8.3}$$

The second kind of quantity that is needed is a measure of the preferential enrichment of high frequency signals of one data type within chromosomal segments having either high or low wavelet transform values in a low-frequency signal of a different data type (identified in the off-diagonals of Figure 8.5). This enrichment measure, $r_{\text{hi/lo}}$ will be equal to 1 if the high-frequency signal is exclusively localized within the high portions of the low-frequency signal, and -1 if located exclusively within the low portions:

$$r_{\text{hi/lo}} = \frac{2 \times \sum_i \text{hi}_{i,j}}{\sum_i d_i} - 1, \tag{8.4}$$

where j corresponds to the lower-frequency signal in the pair. A value near zero indicates that there is little to no preferential enrichment of high-frequency signals within either high or low regions of a low-frequency signal.

Results for *E. coli* The real component of the Morlet wavelet transform was computed for each of 134 different *E. coli* parameters, along with a bootstrap significance test as described in the previous chapter. These parameters are listed in Table 8.1, and the include sequence-derived parameters (GC content, sequence-derived biophysical parameters, etc.), annotation-based parameters (e.g. ORF length and codon preferences), gene functional classes, and functional data types such as gene expression and essentiality measures. Note that these parameters are not all independent from one another (e.g. positional preferences for the codon GGG will correlate with spatial patterns in GC content).

The pairwise cross-correlations at every length scale (i.e. period) were computed for each pair of data types listed in Table 8.1. This analysis thus exam-

Table 8.1: List of 134 parameters analyzed for pattern cross-correlations in *E. coli*.

Class	**No. of parameters**	**Reference**
Gene expression	1	[4]
Gene essentiality	1	[105]
Evolutionary retention	1	[105]
GCAT content/skews	8	[26]
Biophysical parameters	6	[226]
Gene classifications	51	[281]
Codon preferences	64	[26]
ORF lengths	1	[26]
Intergenic region lengths	1	[26]

ined $(134 \times 133)/2 = 8{,}911$ pairs of data sets, and $120 \times 120 = 14{,}400$ pairs of scales for each data set pair. All told, $8911 \times 14400 = 128{,}318{,}400$ correlations were examined and scored based on the p-values computed from the bootstrap significance tests. From these, 58,862 significant peaks in PDC landscapes were identified, and approximately 10,000 correlations were filtered out based on the strength of the pattern overlaps. For each of these correlations, the above-described correlation measures were computed ($r_{\mathrm{pos/neg}}$ and $r_{\mathrm{hi/lo}}$). A summary of the results for the absolute value of the correlation quantity is provided in the scatterplot shown in Figure 8.6.

Perfect correlations and anti-correlations tended to cluster at particular length scales, as shown in the histogram in the right-hand side of Figure 8.6. These periods, which include 200 kb, 350 kb, 650 kb, and 1 Mb, are likely biased somewhat by similar patterns in non-independent data sets (such as the different sets of codon preferences). However, this method was useful in its ability to systematically detect positive and negative correlations of the sort identified by visual inspection earlier in this chapter (Figure 8.1).

8.2.3 Sub-regions of the *E. coli* chromosome

Many of the higher-frequency patterns ($<$ 300 kb) that have been observed in *E. coli* have been localized to particular regions of the chromosome rather

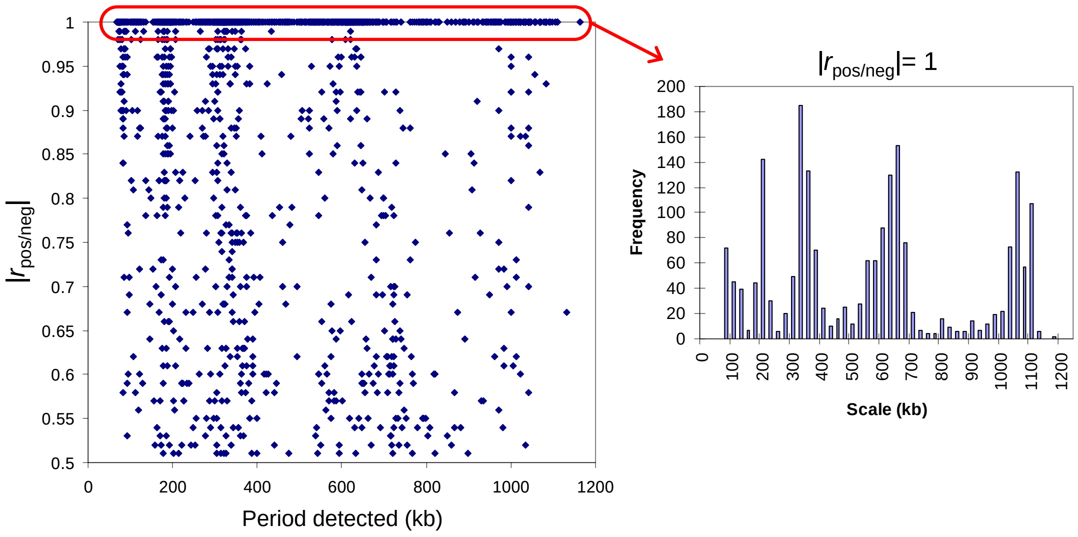


Figure 8.6: Positive and negative spatial correlations between multiple *E. coli* data sets. The quantity $r_{\mathrm{pos/neg}}$ is described in the text. The histogram at the right reveals a preference for spatial correlations to occur at length scales of 200 kb, 350 kb, 650 kb, and 1 Mb.

than occurring globally across the genome. This result motivates an assessment of pattern cross-correlation in "sub-regions" of the chromosome. Thus, the same cross-correlation analysis described above was performed not only on a data set by data set basis, but also for each of 12 sub-regions of of the *E. coli* chromosome. The boundaries of these regions were determined from the regions of high and low expression identified previously using wavelet analysis [4] (upper-left panel of Figure 8.7).

The results for regio-specific cross-correlations between *E. coli* gene expression and intrinsic curvature are shown in Figure 8.7. Each ∼600 kb segment of the chromosome exhibited differing degrees of correlation, as shown in the bottom portion of the figure. This result highlights the need for a more extensive analysis of localized patterns in all of the data sets examined thus far.

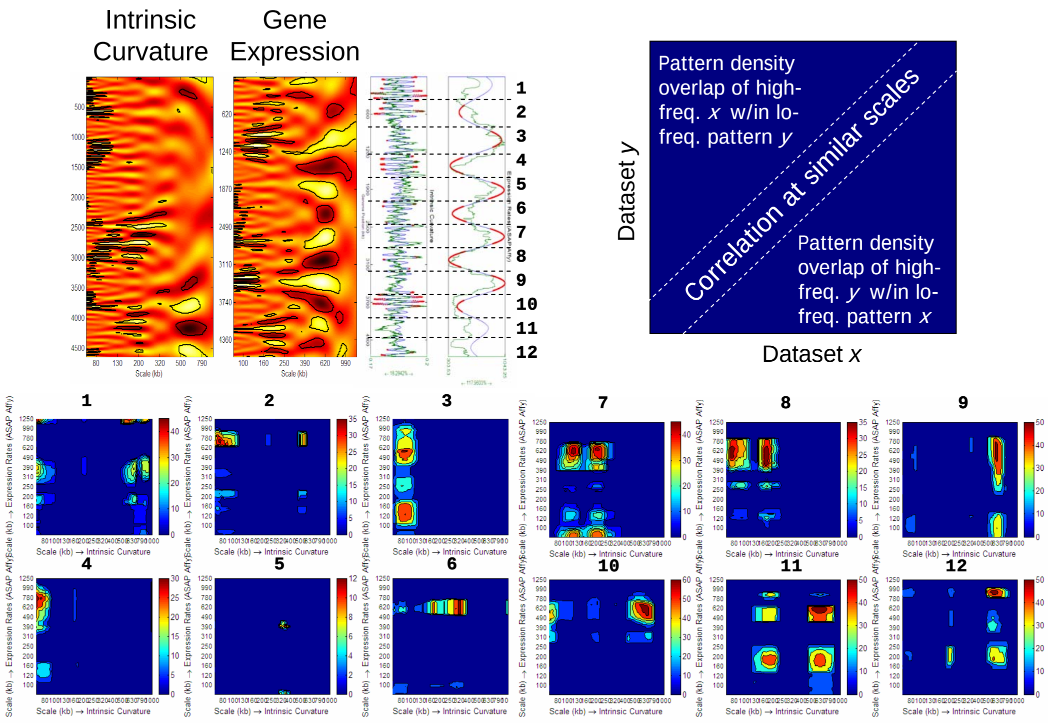


Figure 8.7: Cross-correlations identified between gene expression and intrinsic curvature in the *E. coli* genome, divided into 12 sub-regions. The upper-left panel shows the demarcation of the *E. coli* genome into subunits, and the 12 PDC landscapes show the significant correlations and cross-correlations found within each data set.

8.3 Prospective experimental design to investigate genome organization

The results presented thus far reveal a highly nonrandom spatial organization and pattern interlinking in many chromosomal position-dependent attributes in *E. coli*, and the results in Chapter 7 provide strong evidence that gene order and chromosomal organization in most bacteria are subject to selection pressure. These results thus raise the question of the robustness of the observed long-range patterns to specific alterations in the large-scale genome organization. In other words, are there ways to manually rearrange the genome in a viable cell for which long-range patterns would be disrupted or even disappear? If so, what would be the effect of such rearrangements on phenotype (e.g. fitness and gene expression)? Initial analysis of patterns in gene expression as measured by Affymetrix chips for *E. coli* strains that have been adaptively evolved on alternative carbon sources indicate that such patterns are preserved when large-scale (∼1 Mb) duplications and inversions occur [240].

Towards answering the above question, I have applied the wavelet pattern detection methodology described in 7 to conduct *in silico* rearrangements of the *E. coli* genome and, more specifically, to use this method to identify specific chromosomal rearrangements that will most strongly perturb the long-range patterns that have been observed previously (in gene expression and in other parameters). This involved an algorithmic approach illustrated schematically in Figure 8.8. In this procedure, the realm of possible rearrangements (in this case, inversions, since those can be specifically generated experimentally) is sampled in an iterative fashion until the rearrangement that produces the maximum perturbation of the wavelet scalogram is encountered. This procedure can then be repeated for segments of varying sizes in order to perturb patterns at multiple length scales. Since the combinatorial size of possible inversions and duplications is very large (over 4000 points along the genome and spanning sizes from ∼70 kb to 1 Mb), an exhaustive effort of this sort is computationally infeasible. However, numerous chromosomal arrangements can be ruled out *a priori* since they have been demonstrated experimentally to be unviable [277]. Additionally, the size of possible duplication-inversions can be limited to lengths which would be expected to perturb observed periods (rather than all possible lengths). Once an appropriate duplication size was selected, it was then inserted *in silico* at intervals spaced 50 kb apart. The wavelet analysis was then re-run for each of these simulated rearrangements, and the pattern strength (defined in the previous chapter) was used as a scoring metric and applied it to rank-order each set of rearrangements.

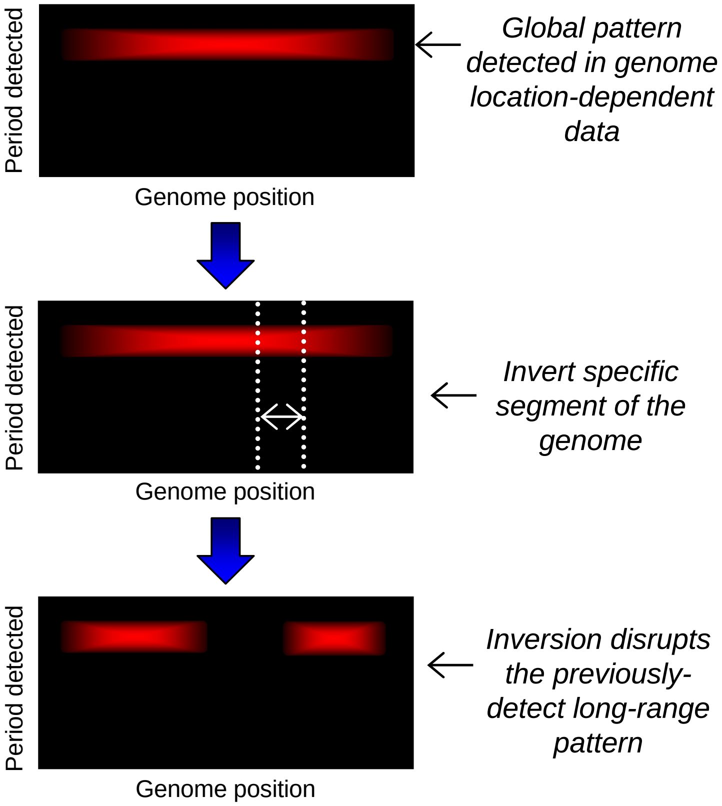

Period detected
Genome position
Global pattern detected in genome location-dependent data
Period detected
Genome position
Invert specific segment of the genome
Period detected
Genome position
Inversion disrupts the previously-detect long-range pattern

Figure 8.8: Identification of specific rearrangements that disrupt long-range patterns in genome position-dependent data.

Since the major global periodicity that has been observed in gene expression in *E. coli* is ∼600 kb, I have initially tested *in silico* rearrangements (inversions) that are 600 kb in length. I then examined the consequences of an inversion of this size occurring at numerous start sites in the *E. coli* genome, regularly spaced every 50 genes (Figure 8.9b).

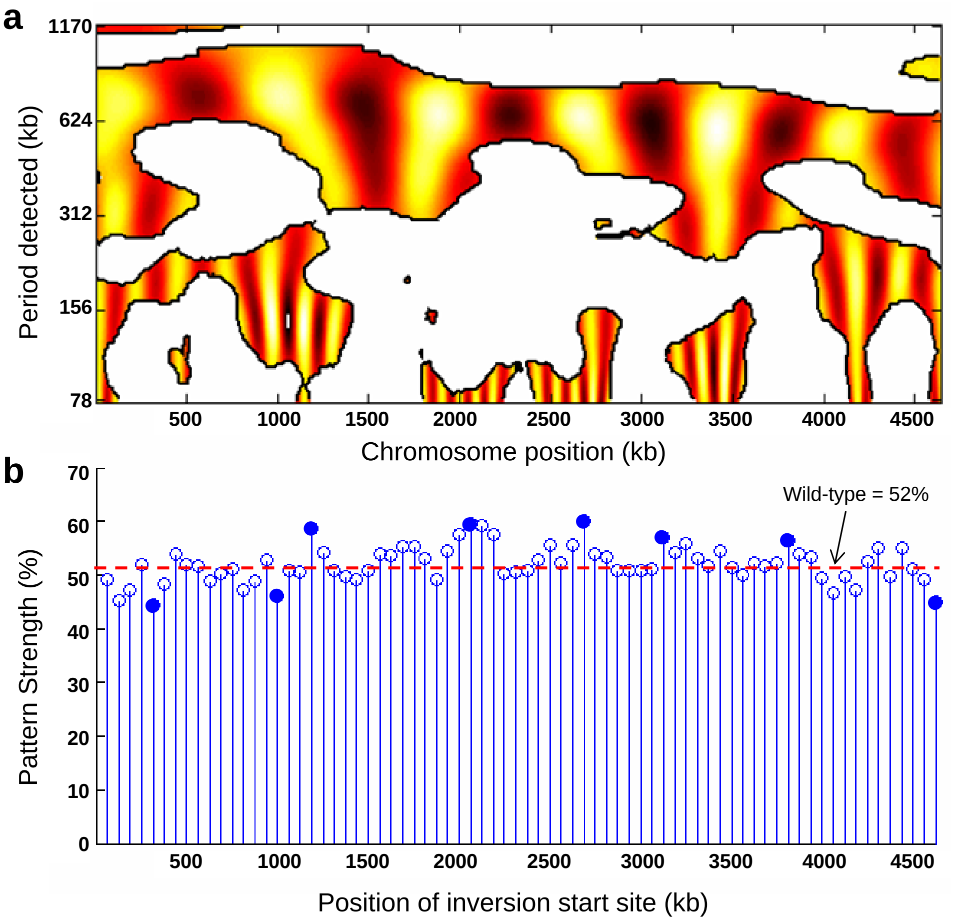


Figure 8.9: Selection of candidates for targeted genome inversions in *E. coli*. a. Continuous wavelet scalogram of gene expression data for *E. coli*. The colored portions of the scalogram indicate significant periodic patterns (FDR < 5%). b. The fractional pattern strengths are shown for the scalogram associated with each *in silico* rearranged genome. Each rearrangement consisted of a 600 kb inversion beginning at the locus indicated in the x-axis. The pattern strength for the wild-type *E. coli* genome is shown as a dashed red line for reference. Potential candidates for genome inversions are highlighted as solid blue circles.

As shown in Figure 8.9b, over half of the *in silico* inversions actually increased the overall pattern strength in gene expression, relative to the wild-type *E. coli* data. Additionally, several inversions result in significantly lower pattern strengths. Highlighted in blue on the figure are candidate inversions for experimental testing based on this analysis. However, the inversion sites near the origin and terminus can be ruled out since those have been shown to be unviable rearrangements in the literature, as mentioned above [277].

8.4 Conclusions

In *E. coli*, a detailed analysis of available data demonstrates that patterns in multiple disparate properties are interlinked (Figures 8.1, 8.2, and 8.3). The consistency of the 650 kb chromosome macrodomains identified using wavelet analysis of expression data [4] with those identified from genetics experiments [309] indicates that large-scale genome packing is indeed linked to transcription, as has been previously hypothesized [48] (Figure 8.3). An exhaustive analysis of millions of cross-correlations between 134 data sets in *E. coli* revealed significant positive and negative correlations both globally across the entire genome and localized to specific sub-regions. This work has additional implications for *de novo* genome design [288], in that gene order and composition—and the resulting chromosomal ultrastructure—are significant design variables that will likely need to be taken into account. Given the non-random distribution of these parameters in nearly all sequenced prokaryotes, as well as the linked nature of disparate parameters in *E. coli*, it is clear that any genome design endeavor will involve a multi-variable, multi-dimensional optimization problem.

The text of this chapter, in part, is a reprint of the material as it appears in T.E. Allen, N.D. Price, A.R. Joyce, and B.O. Palsson. 2005. Long-range Periodic Patterns in Microbial Genomes Indicate Significant Multi-scale Chromosomal Organization. *PLoS Computational Biology*, (in press). I was the primary author of this publication, and the co-authors participated and supervised the research which forms the basis for this chapter.

The text of this chapter, in part, is a reprint of the material as it appears in T.E. Allen, A.R. Joyce, and B.O. Palsson. Cross-correlation of spatial patterns in heterogeneous *Escherichia coli* data sets. In preparation. I was the primary author of this publication, and the co-authors participated and supervised the research which forms the basis for this chapter.

Chapter 9

Conclusion: Towards a "Topobiology"

> Determining how to formalize the problem of emergent features and multiscale description is one of the goals of the science of complex systems. Biology, whose object of study extends within one same entity across many scales, from molecule to animated organism, should thankfully embrace the efforts and watch their progress carefully.
>
> Sui Huang [139]

> Is the map of the cell in the chromosome?
>
> Antoine Danchin [60]

The preceding chapters described the conceptual advance associated with a systems analysis of bacterial genomes—and of the information transfer processes in bacteria—within the context of network reconstruction. Several key scientific conclusions were presented in this dissertation:

- Macromolecular synthesis reactions are chemical transformations and thus are not fundamentally different from the small-molecule reactions in metabolism.
- Network reconstructions can represent information transfer, explicitly accounting for the material and energy cost of RNA and protein synthesis.

- This method can be scaled to include the synthesis of every protein specified in a genome because the pathways involving macromolecular synthesis are linear and lack the complexity found in metabolic networks.
- Genome-scale reconstructions of multiple network types can be integrated to analyze heterogeneous data types and compute the material and energy costs for genome usage and maintenance.
- Translational efficiency can be varied by altering how the DNA chooses to encode each protein, and as such, the synonymous codon usage and measured tRNA abundances found in *E. coli* are highly synchronized.
- Genome position-dependent data from multiple microbial organisms contain highly varying degrees of patterns, the strengths of which correlate with several organism-specific characteristics.
- The spatial correlation of disparate data types in *E. coli* revealed that any genome design endeavor will involve a multi-parameter, multi-length scale optimization problem.

These conclusions can be distilled into two overarching discoveries:

1. Information synthesis can be accounted for within cell-scale network reconstructions to yield novel results and biological discovery; and
2. The physical locations of genes and their components—in addition to simply the "parts list" of the components and their network interactions—are highly nonrandom and are intricately intertwined with the function of a bacterial cell.

Curiously, however, the full implications of this second point are seemingly contradictory with the first point. What, then, are these implications?

A strictly 2-D network reconstruction formalism is insufficient to fully describe the information transfer processes and the interaction of macromolecules with the genome. An additional constraint must be taken into account in order to reconstruct these processes: spatial constraints, e.g. the location of genes along the chromosome, or the localization of certain proteins at specific regions within the cell (as described for several cases in Chapter 6). This sort of spatial constraint goes beyond component interactions and connectivity within a 2-D map, but would also take into account the intracellular locations of genes and gene products.

The two overarching discoveries delineated above do not necessarily contradict one another, however. The reason for this is that the methods used to reconstruct the processes of transcription and translation implicitly assign gene locations to each gene and each transcription event by virtue of the fact that a specific gene (or its promoter) indicates a specific chromosomal locus (as well as a molecule that participates in some reaction). Additionally, the sensitivity analysis presented in Chapter 5 included an implicit tRNA localization constraint. To compute the translational efficiencies in that study, each gene was considered as a separate, closed system. In other words, if a set of synonymous codons in Gene A demanded $tRNA_1$, but there were many codons in an adjacent Gene B which also demanded $tRNA_1$, it is conceivable that the translational efficiency of both genes would be reduced by having to share the same localized tRNA pool. Thus, the assumption presented in Chapter 5 constituted a highly idealized case, but nevertheless a case that more closely approximated reality than the alternative idealized case. Let us consider this alternative in the next section.

9.1 Revisiting the tRNA localization assumption: What if the cell is a "mixed bag"?

In Chapter 5, I described a sensitivity analysis of translational efficiency in *E. coli* with respect to the synonymous codon usage of each gene if pools of tRNA molecules are assumed to be localized near each transcript. In this analysis, I have assumed that the demands placed upon the pools of tRNA species are also localized near each translation site. Additionally, the results presented in Chapter 8 clearly showed that long-range genome location-dependent patterns exist in both gene expression [4, 5] and in codon usage in *E. coli* [5].

For the sake of argument, however, let us consider a hypothetical cell in which tRNA molecules are not subject to diffusion limitations are are free to bind to any codon in the cell instantaneously. When the ranges in protein synthesis were calculated under this idealized assumption (that the tRNA molecules are not diffusion-limited and are free to bind to any codon in the genome at any location), the histogram of ranges in translational efficiency becomes strikingly different (Figure 9.1) from that calculated for the localized tRNA case back in Chapter 5 (Figure 5.7).

If the assumption that tRNAs can freely diffuse to any available codon in the cell were indeed true, then there would be essentially zero flexibility in protein synthesis achievable simply by altering synonymous codon usage. Thus, there would be no selective pressure imposed by translational efficiency to generate codon biases. However, codon biases certainly do exist [1], implying that tRNA diffusion limitations lend credence to the notion of tRNA localization with respect to position along the chromosome [59]. Furthermore, if the translational step time is roughly 0.05 sec (assuming an elongation rate of 16 peptide bonds/sec [325], Table 2.3), the diffusion time for a particular tRNA molecule is on the order of 20-fold longer if D_{tRNA} (the diffusion coefficient for a tRNA molecule) is approximately 10^{-8} cm^2/sec and the diffusion length is 1 micron [320]. This implies that tRNA species are diffusion-limited at the cellular length scale and would thus be localized *in vivo*. Detection of periodicity in expression also supports the idea of tRNA localization and genome packing to minimize the diffusion distances of rare tRNAs [255].

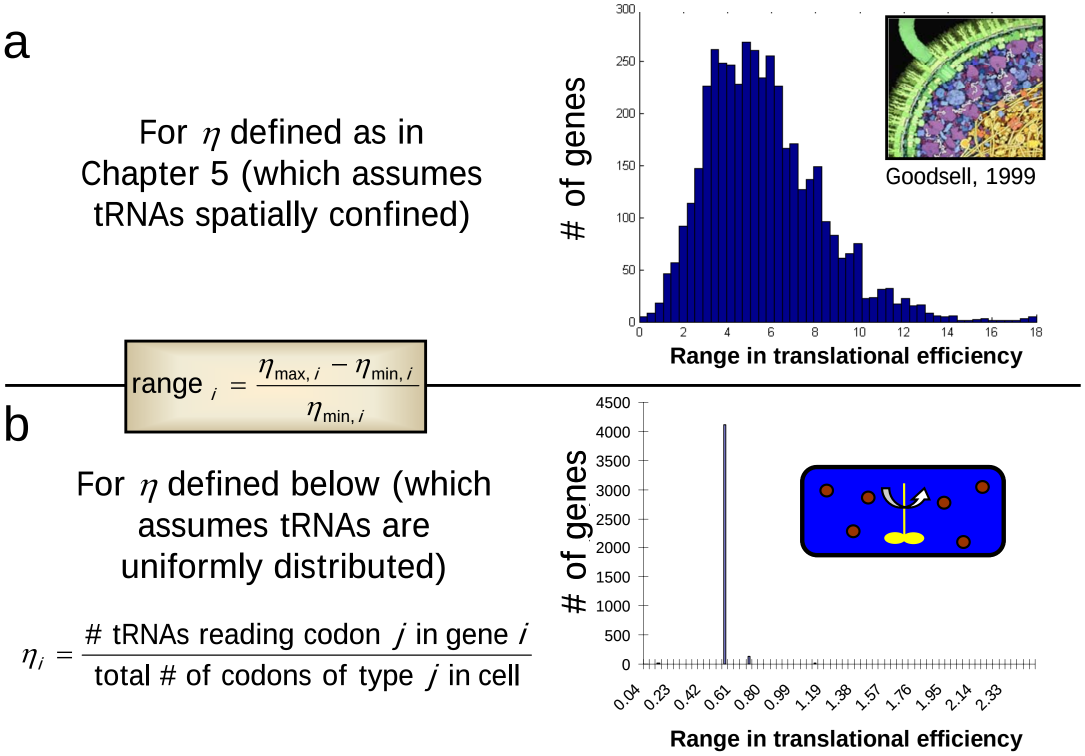


Figure 9.1: Revisiting the tRNA localization assumption. These panels show a comparison of histograms of allowable ranges in protein translational efficiency of all *E. coli* proteins given alterations in synonymous codon usage for two fundamentally different assumptions. a. tRNA molecules are assumed to be strictly localized in the vicinity of each transcript (as per drawing in inset and as described in Chapter 5). Thus, changes in codon usage can have a significant effect on protein translational efficiency (on average 6.5-fold). b. tRNA molecules are free to diffuse and bind to any available codon within the cell, no matter its location along the chromosome (see inset). Given this assumption, the extent to which protein synthesis can be altered by means of changing synonymous codon usage is negligible.

9.2 Towards spatial constraints for genome-scale reconstructions

Genome location-dependent patterns in multiple sequence-based parameters in numerous microbial organisms (Chapter 7) and in *E. coli* gene expression [4, 156, 5], codon usage [302, 5], and essential gene loci [105, 5] (Chapter 8) that have been detected imply that the ultrastructure of the chromosome is nonrandom [255]. Thus, it is likely that both gene order on the chromosome and the way in which the chromosome is packed within the bacterial nucleoid have evolved to optimize the expression of the genes whose products are required under a particular set of conditions [48]. Accordingly, the analysis presented in Chapter 5 was an *in silico* sensitivity analysis of protein synthesis in *E. coli* with respect to synonymous codon usage on a gene-by-gene basis. If diffusion limitations (Chapter 6) are ignored, however, the results presented in Figure 9.1 imply that the fields of biochemical network reconstruction and constraint-based analysis (the only network analysis currently practiced at genome-scale [235, 217, 248]) may enter a phase in which network stoichiometry must be viewed within the context of topobiological, spatial constraints—particularly when accounting for macromolecular synthesis reactions [3] (Chapters 3-5). Thus the network-based viewpoints of complex cellular functions will ultimately require reconciliation with the physical realities of the cell's three-dimensional interior (Figure 9.2).

9.3 Concluding thoughts

The work presented in this dissertation describes a conceptual advance in the scope of genome-scale reconstructions with the analysis of information transfer in *E. coli.* In Chapters 2–5, I described the fundamental methods by which macromolecules can be incorporated into existing cell-scale reconstructions of metabolism and regulation. However, in Chapters 6–8 I proceeded to reveal the limitations of network reconstructions (so-called "2-D annotations" [219, 248]) with respect

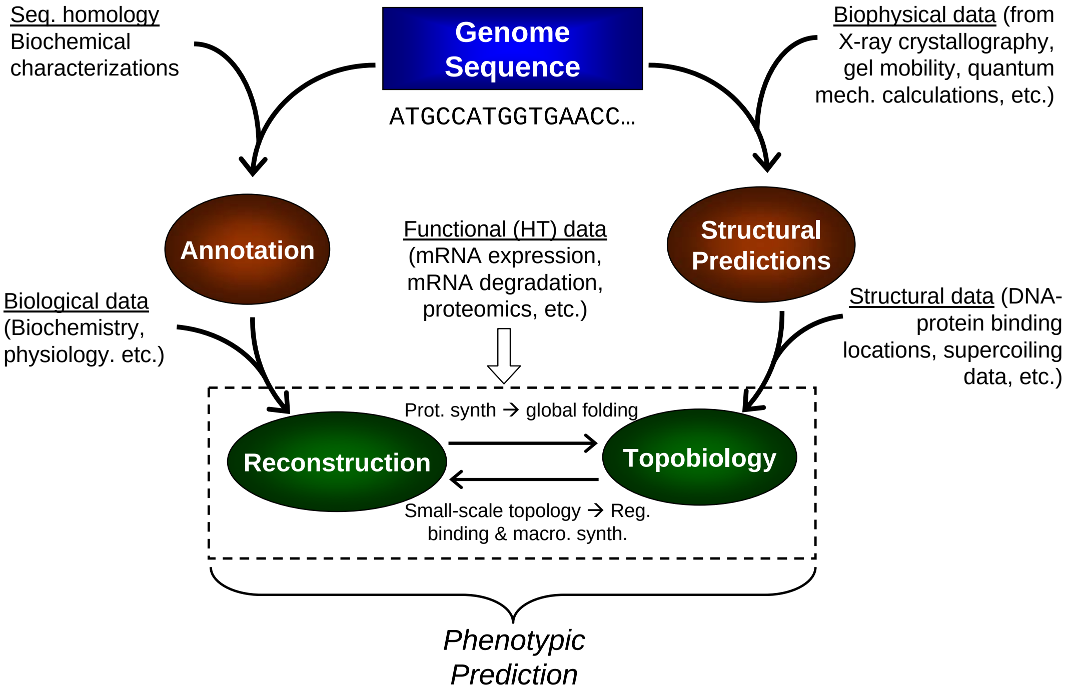

Seq. homology
Biochemical characterizations
Genome Sequence
ATGCCATGGTGAACC…
Biophysical data (from X-ray crystallography, gel mobility, quantum mech. calculations, etc.)
Annotation
Functional (HT) data (mRNA expression, mRNA degradation, proteomics, etc.)
Structural Predictions
Biological data (Biochemistry, physiology. etc.)
Structural data (DNA-protein binding locations, supercoiling data, etc.)
Prot. synth → global folding
Reconstruction
Topobiology
Small-scale topology → Reg. binding & macro. synth.
Phenotypic Prediction

Figure 9.2: Towards a 3-D reconstruction. The figure presents a conceptual outline of how various data types may eventually be integrated towards building a 3-D reconstruction of an organism. The left side of the schematic shows the established method of 2-D network reconstruction. Once macromolecules are explicitly included in reconstructions, however, multiple other data types can be incorporated, including functional data (e.g. gene expression datasets), sequence-based biophysical parameters (e.g. intrinsic curvature), and protein targeting and gene localization. Once the topobiological aspects of cell organization are well understood, these will eventually be integrated with the network reconstruction to achieve a 3-D, whole-cell reconstruction of a bacterial organism.

to spatial constraints and three-dimensional packing within the cell. This work involved both 1) assessing, *in silico*, the constraints that act on the macromolecular synthesis reactions in bacteria, and 2) analyzing genome-location dependent patterns in numerous datasets. This topobiological aspect of complex cellular processes has received limited attention in the past, but now with the availability of multiple genome-scale data sets we are in a position to address this fundamental issue in biology.

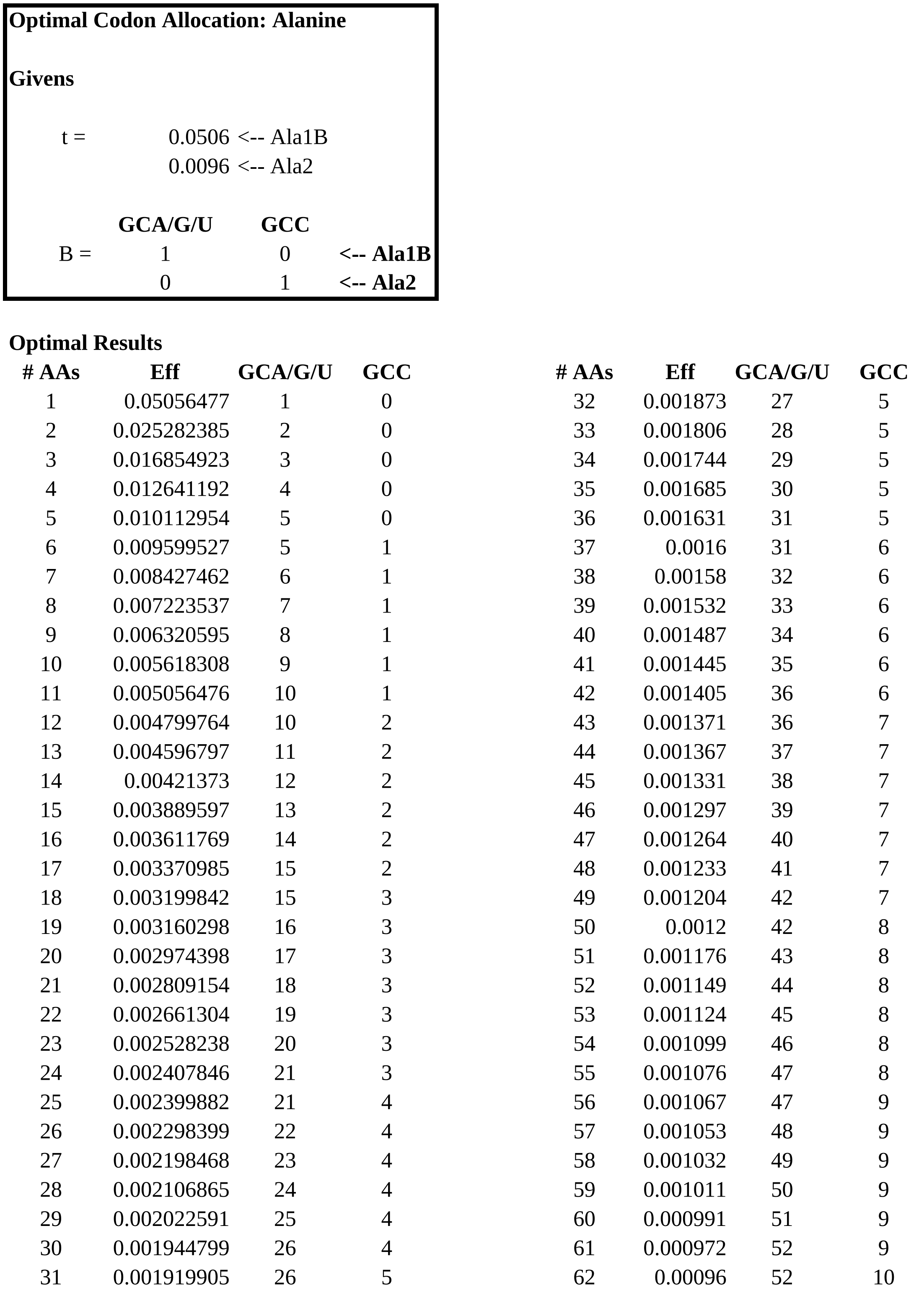
Future work in this field must also address the development of methodology and concepts needed to apply chromosomal position-dependent constraints to genome-scale *in silico* models. Given the results presented in this dissertation, I strongly believe that the field of whole-cell modeling is entering a phase in which network connectivity will increasingly be viewed within the context of topobiological, spatial constraints. Accordingly, high-throughput experimental methods need to be developed by which the spatial arrangement of a cell interior is probed. Such data can then be incorporated into a systems framework that would allow for the comprehensive assessment of location-specific interactions in cells, ultimately significantly improving the interpretive and predictive capabilities of cell-scale models. Further gains in our ability to use high-throughput data to elucidate chromosomal structure and predict its impact on phenotype—in not only model organisms such as *E. coli*, but also mammalian systems—would lead to an *in silico* framework by which topobiological constraints are explicitly accounted for. The methodological foundation that would result from such a framework may ultimately expedite the treatment of diseases such as cancer for which there is emerging evidence that alterations in chromosomal structure are an integral part of the disease etiology [241]. Once systems biology advances towards—and embraces—such a topobiology, we may finally witness the achievement of a true Kuhnian advance in biological understanding.

Appendix A

Optimal Codon Allocation Tables

This appendix contains the optimal synonymous codon allocation tables generated by the *ad hoc* methods described in Chapter 5. Ten of the amino acids contained nontrivial solutions as a result of the binding stoichiometry ($B_{i,j}$ matrix) between the tRNA species and their cognate codons for these amino acids (refer to lefthand side of Table 5.2): alanine, arginine, glutamine, glycine, isoleucine, leucine, proline, serine, threonine, and valine. These tables provide the number of amino acids of each of these ten types to be specified (in other words, the total number of synonymous codons to be allocated for each amino acid), the calculated maximal translational effiency ($\eta_{\max}$), and the synonymous codon allocation which yielded that maximum. Note that there may be multiple optimal solutions that yield the same maximum efficiency.

Table A.1: Optimal synonymous codon allocation schemes for alanine.



Optimal Codon Allocation: Alanine

Givens

t =	0.0506	<-- Ala1B	
	0.0096	<-- Ala2	

	GCA/G/U	**GCC**	
B =	1	0	**<-- Ala1B**
	0	1	**<-- Ala2**

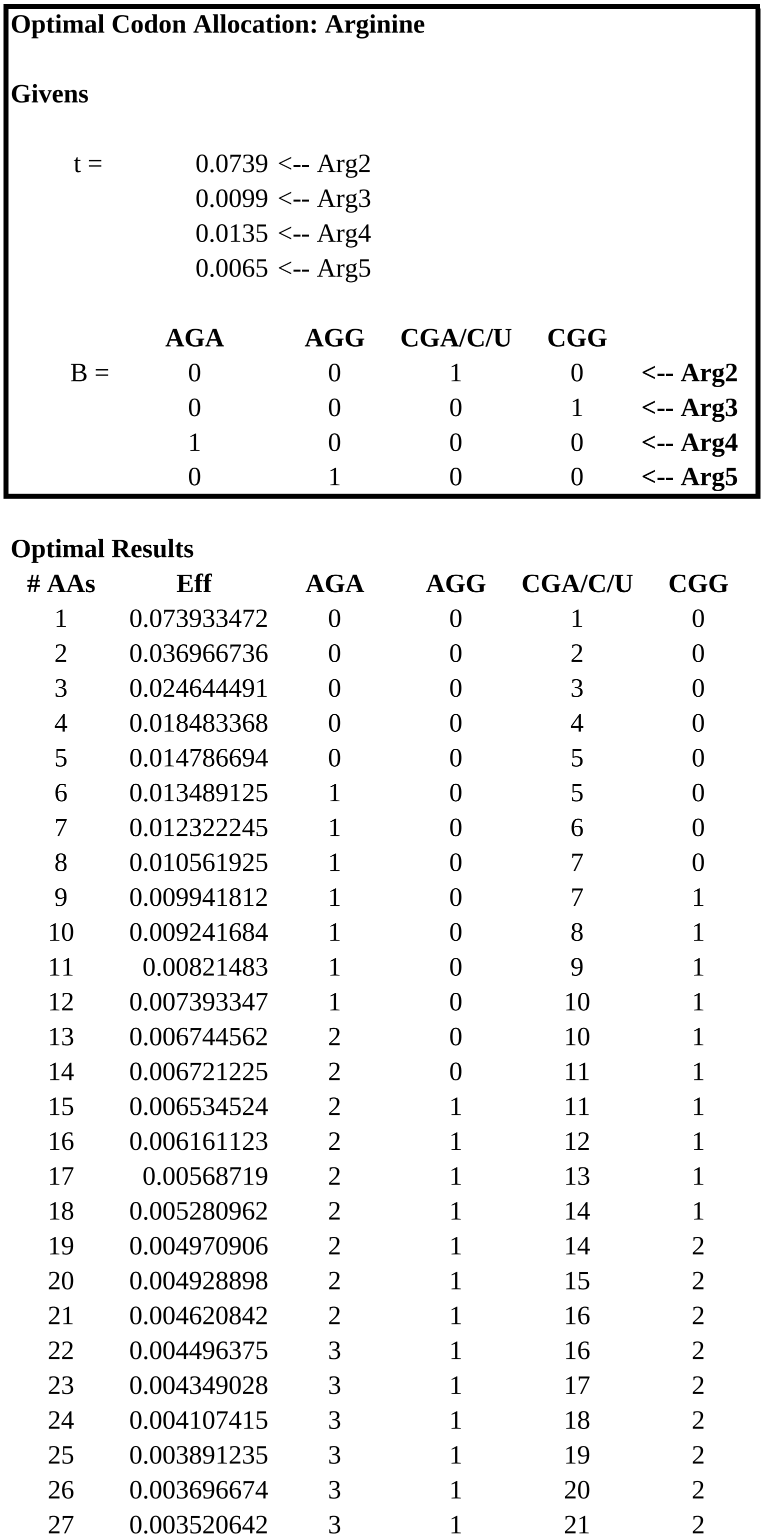
Optimal Results

# AAs	**Eff**	**GCA/G/U**	**GCC**	**# AAs**	**Eff**	**GCA/G/U**	**GCC**
1	0.05056477	1	0	32	0.001873	27	5
2	0.025282385	2	0	33	0.001806	28	5
3	0.016854923	3	0	34	0.001744	29	5
4	0.012641192	4	0	35	0.001685	30	5
5	0.010112954	5	0	36	0.001631	31	5
6	0.009599527	5	1	37	0.0016	31	6
7	0.008427462	6	1	38	0.00158	32	6
8	0.007223537	7	1	39	0.001532	33	6
9	0.006320595	8	1	40	0.001487	34	6
10	0.005618308	9	1	41	0.001445	35	6
11	0.005056476	10	1	42	0.001405	36	6
12	0.004799764	10	2	43	0.001371	36	7
13	0.004596797	11	2	44	0.001367	37	7
14	0.00421373	12	2	45	0.001331	38	7
15	0.003889597	13	2	46	0.001297	39	7
16	0.003611769	14	2	47	0.001264	40	7
17	0.003370985	15	2	48	0.001233	41	7
18	0.003199842	15	3	49	0.001204	42	7
19	0.003160298	16	3	50	0.0012	42	8
20	0.002974398	17	3	51	0.001176	43	8
21	0.002809154	18	3	52	0.001149	44	8
22	0.002661304	19	3	53	0.001124	45	8
23	0.002528238	20	3	54	0.001099	46	8
24	0.002407846	21	3	55	0.001076	47	8
25	0.002399882	21	4	56	0.001067	47	9
26	0.002298399	22	4	57	0.001053	48	9
27	0.002198468	23	4	58	0.001032	49	9
28	0.002106865	24	4	59	0.001011	50	9
29	0.002022591	25	4	60	0.000991	51	9
30	0.001944799	26	4	61	0.000972	52	9
31	0.001919905	26	5	62	0.00096	52	10

Table A.1, continued.

# AAs	Eff	GCA/G/U	GCC
63	0.000954052	53	10
64	0.000936385	54	10
65	0.000919359	55	10
66	0.000902942	56	10
67	0.000887101	57	10
68	0.000872684	57	11
69	0.000871806	58	11
70	0.00085703	59	11
71	0.000842746	60	11
72	0.000828931	61	11
73	0.000815561	62	11
74	0.000802615	63	11
75	0.000799961	63	12
76	0.000790075	64	12
77	0.00077792	65	12
78	0.000766133	66	12
79	0.000754698	67	12
80	0.0007436	68	12
81	0.000738425	68	13
82	0.000732823	69	13
83	0.000722354	70	13
84	0.00071218	71	13
85	0.000702288	72	13
86	0.000692668	73	13
87	0.000685681	73	14
88	0.000683308	74	14
89	0.000674197	75	14
90	0.000665326	76	14
91	0.000656685	77	14
92	0.000648266	78	14
93	0.00064006	79	14
94	0.000639968	79	15
95	0.00063206	80	15
96	0.000624256	81	15
97	0.000616644	82	15
99	0.000601962	84	15
100	0.00059997	84	16
101	0.00059488	85	16
102	0.000587962	86	16
103	0.000581204	87	16
104	0.0005746	88	16
105	0.000568143	89	16
106	0.000564678	89	17
107	0.000562	90	17
108	0.000556	91	17
109	0.00055	92	17
110	0.000544	93	17
112	0.000533	94	18
115	0.000521	97	18
117	0.000511	99	18
118	0.000506	100	18
119	0.000505	100	19
120	0.000501	101	19
121	0.000496	102	19
124	0.000482	105	19
125	0.00048	105	20
126	0.000477	106	20
129	0.000464	109	20
130	0.00046	110	20
133	0.000451	112	21
136	0.00044	115	21
143	0.000418	121	22
145	0.000414	122	23
146	0.000411	123	23
148	0.000405	125	23
149	0.000401	126	23
157	0.000383	132	25
160	0.000375	135	25
166	0.000361	140	26
222	0.00027	187	35
244	0.000246	205	39

Table A.2: Optimal synonymous codon allocation schemes for arginine.



Optimal Codon Allocation: Arginine

Givens

t =	0.0739	<-- Arg2
	0.0099	<-- Arg3
	0.0135	<-- Arg4
	0.0065	<-- Arg5

	AGA	**AGG**	**CGA/C/U**	**CGG**	
B =	0	0	1	0	**<-- Arg2**
	0	0	0	1	**<-- Arg3**
	1	0	0	0	**<-- Arg4**
	0	1	0	0	**<-- Arg5**

Optimal Results

# AAs	**Eff**	**AGA**	**AGG**	**CGA/C/U**	**CGG**
1	0.073933472	0	0	1	0
2	0.036966736	0	0	2	0
3	0.024644491	0	0	3	0
4	0.018483368	0	0	4	0
5	0.014786694	0	0	5	0
6	0.013489125	1	0	5	0
7	0.012322245	1	0	6	0
8	0.010561925	1	0	7	0
9	0.009941812	1	0	7	1
10	0.009241684	1	0	8	1
11	0.00821483	1	0	9	1
12	0.007393347	1	0	10	1
13	0.006744562	2	0	10	1
14	0.006721225	2	0	11	1
15	0.006534524	2	1	11	1
16	0.006161123	2	1	12	1
17	0.00568719	2	1	13	1
18	0.005280962	2	1	14	1
19	0.004970906	2	1	14	2
20	0.004928898	2	1	15	2
21	0.004620842	2	1	16	2
22	0.004496375	3	1	16	2
23	0.004349028	3	1	17	2
24	0.004107415	3	1	18	2
25	0.003891235	3	1	19	2
26	0.003696674	3	1	20	2
27	0.003520642	3	1	21	2

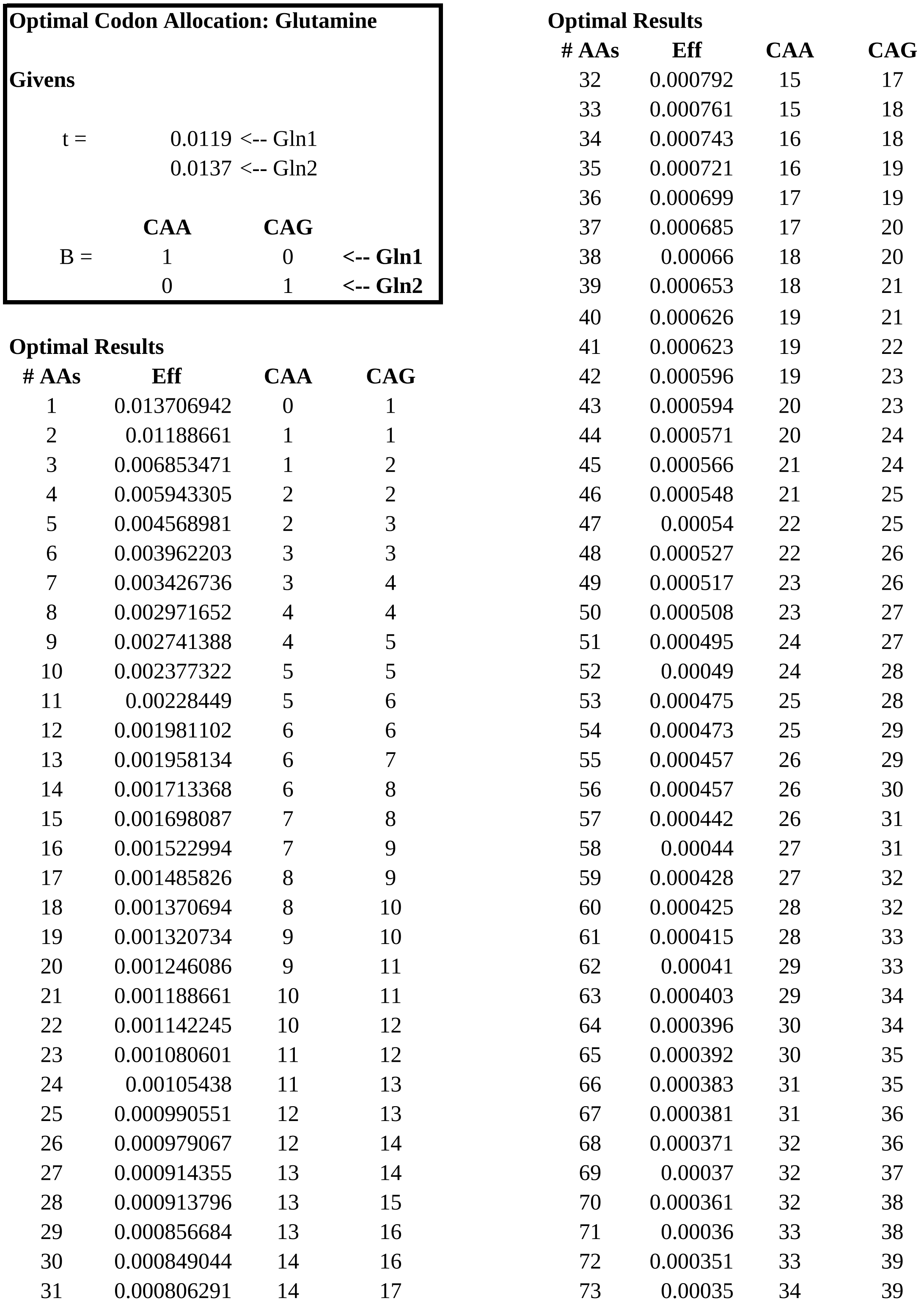
Table A.2, continued.

# AAs	Eff	AGA	AGG	CGA/C/U	CGG
28	0.003372281	4	1	21	2
29	0.003360612	4	1	22	2
30	0.003313937	4	1	22	3
31	0.003267262	4	2	22	3
32	0.003214499	4	2	23	3
33	0.003080561	4	2	24	3
34	0.002957339	4	2	25	3
35	0.002843595	4	2	26	3
36	0.002738277	4	2	27	3
37	0.002697825	5	2	27	3
38	0.002640481	5	2	28	3
39	0.00254943	5	2	29	3
40	0.002485453	5	2	29	4
41	0.002464449	5	2	30	4
42	0.002384951	5	2	31	4
43	0.002310421	5	2	32	4
44	0.002248187	6	2	32	4
45	0.002240408	6	2	33	4
46	0.002178175	6	3	33	4
47	0.002174514	6	3	34	4
48	0.002112385	6	3	35	4
49	0.002053708	6	3	36	4
50	0.001998202	6	3	37	4
51	0.001988362	6	3	37	5
52	0.001945618	6	3	38	5
53	0.001927018	7	3	38	5
54	0.00189573	7	3	39	5
55	0.001848337	7	3	40	5
56	0.001803255	7	3	41	5
57	0.001760321	7	3	42	5
58	0.001719383	7	3	43	5
59	0.001686141	8	3	43	5
60	0.001680306	8	3	44	5
61	0.001656969	8	3	44	6
62	0.001642966	8	3	45	6
63	0.001633631	8	4	45	6
64	0.001607249	8	4	46	6
65	0.001573053	8	4	47	6
66	0.001540281	8	4	48	6
67	0.001508846	8	4	49	6
68	0.001498792	9	4	49	6
69	0.001478669	9	4	50	6
70	0.001449676	9	4	51	6

Table A.2, continued.

# AAs	Eff	AGA	AGG	CGA/C/U	CGG
71	0.001421798	9	4	52	6
72	0.001420259	9	4	52	7
74	0.001369138	9	4	54	7
75	0.001348912	10	4	54	7
76	0.001344245	10	4	55	7
77	0.001320241	10	4	56	7
78	0.001306905	10	5	56	7
79	0.001297078	10	5	57	7
81	0.00125311	10	5	59	7
82	0.001242726	10	5	59	8
84	0.001226284	11	5	60	8
86	0.001192475	11	5	62	8
87	0.001173547	11	5	63	8
90	0.001124094	12	5	65	8
92	0.001104646	12	5	66	9
93	0.001103485	12	5	67	9
99	0.001037625	13	6	71	9
102	0.000999101	13	6	74	9
107	0.000960175	14	6	77	10
125	0.000821483	16	7	90	12
126	0.000816816	16	8	90	12
127	0.000812456	16	8	91	12
131	0.000786526	17	8	94	12
145	0.000710129	18	9	104	14
147	0.000704128	19	9	105	14

Table A.3: Optimal synonymous codon allocation schemes for glutamine.



Optimal Codon Allocation: Glutamine

Givens

t = 0.0119 <-- Gln1
0.0137 <-- Gln2

	CAA	**CAG**	
B =	1	0	**<-- Gln1**
	0	1	**<-- Gln2**

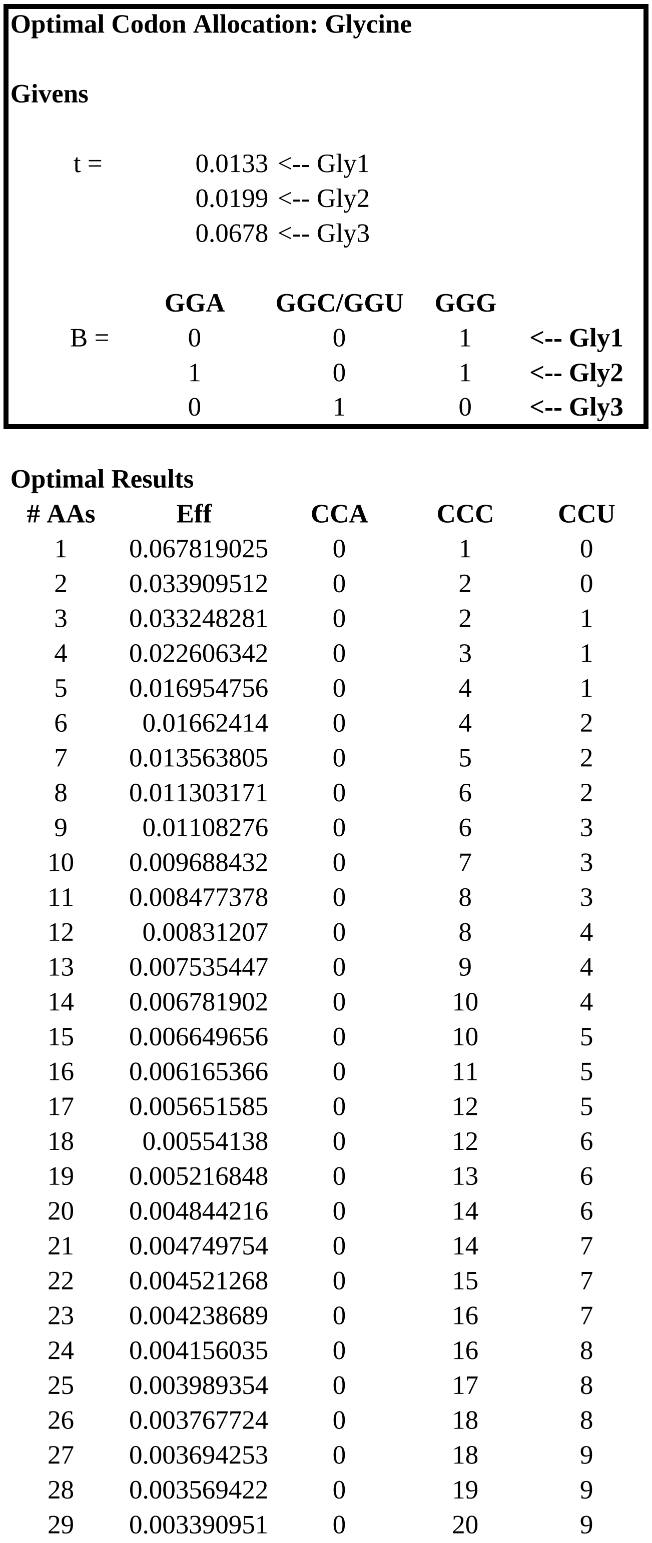
Optimal Results

# AAs	**Eff**	**CAA**	**CAG**
1	0.013706942	0	1
2	0.01188661	1	1
3	0.006853471	1	2
4	0.005943305	2	2
5	0.004568981	2	3
6	0.003962203	3	3
7	0.003426736	3	4
8	0.002971652	4	4
9	0.002741388	4	5
10	0.002377322	5	5
11	0.00228449	5	6
12	0.001981102	6	6
13	0.001958134	6	7
14	0.001713368	6	8
15	0.001698087	7	8
16	0.001522994	7	9
17	0.001485826	8	9
18	0.001370694	8	10
19	0.001320734	9	10
20	0.001246086	9	11
21	0.001188661	10	11
22	0.001142245	10	12
23	0.001080601	11	12
24	0.00105438	11	13
25	0.000990551	12	13
26	0.000979067	12	14
27	0.000914355	13	14
28	0.000913796	13	15
29	0.000856684	13	16
30	0.000849044	14	16
31	0.000806291	14	17
32	0.000792	15	17
33	0.000761	15	18
34	0.000743	16	18
35	0.000721	16	19
36	0.000699	17	19
37	0.000685	17	20
38	0.00066	18	20
39	0.000653	18	21
40	0.000626	19	21
41	0.000623	19	22
42	0.000596	19	23
43	0.000594	20	23
44	0.000571	20	24
45	0.000566	21	24
46	0.000548	21	25
47	0.00054	22	25
48	0.000527	22	26
49	0.000517	23	26
50	0.000508	23	27
51	0.000495	24	27
52	0.00049	24	28
53	0.000475	25	28
54	0.000473	25	29
55	0.000457	26	29
56	0.000457	26	30
57	0.000442	26	31
58	0.00044	27	31
59	0.000428	27	32
60	0.000425	28	32
61	0.000415	28	33
62	0.00041	29	33
63	0.000403	29	34
64	0.000396	30	34
65	0.000392	30	35
66	0.000383	31	35
67	0.000381	31	36
68	0.000371	32	36
69	0.00037	32	37
70	0.000361	32	38
71	0.00036	33	38
72	0.000351	33	39
73	0.00035	34	39

Table A.3, continued.

# AAs	Eff	CAA	CAG
74	0.000342674	34	40
75	0.000339617	35	40
76	0.000334316	35	41
77	0.000330184	36	41
78	0.000326356	36	42
79	0.00032126	37	42
80	0.000318766	37	43
81	0.000312806	38	43
82	0.000311521	38	44
83	0.000304785	39	44
84	0.000304599	39	45
85	0.000297977	39	46
86	0.000297165	40	46
87	0.000291637	40	47
88	0.000289917	41	47
89	0.000285561	41	48
90	0.000283015	42	48
91	0.000279734	42	49
92	0.000276433	43	49
93	0.000274139	43	50
94	0.00027015	44	50
95	0.000268764	44	51
96	0.000264147	45	51
97	0.000263595	45	52
98	0.000258622	45	53
99	0.000258405	46	53
100	0.000253832	46	54
101	0.000252907	47	54
102	0.000249217	47	55
103	0.000247638	48	55
104	0.000244767	48	56
105	0.000242584	49	56
106	0.000240473	49	57
107	0.000237732	50	57
108	0.000236327	50	58
109	0.000233071	51	58
110	0.000232321	51	59
111	0.000228589	52	59
112	0.000228449	52	60
113	0.000224704	52	61
114	0.000224276	53	61
115	0.00022108	53	62
116	0.000220122	54	62
117	0.000218	54	63
118	0.000216	55	63
119	0.000214	55	64
120	0.000212	56	64
121	0.000211	56	65
122	0.000209	57	65
123	0.000208	57	66
124	0.000205	58	66
125	0.000205	58	67
126	0.000202	58	68
127	0.000201	59	68
128	0.000199	59	69
129	0.000198	60	69
130	0.000196	60	70
131	0.000195	61	70
132	0.000193	61	71
133	0.000192	62	71
134	0.00019	62	72
135	0.000189	63	72
136	0.000188	63	73
137	0.000186	64	73
138	0.000185	64	74
139	0.000183	65	74
140	0.000183	65	75
141	0.00018	65	76
142	0.00018	66	76
143	0.000178	66	77
144	0.000177	67	77
145	0.000176	67	78
146	0.000175	68	78
147	0.000174	68	79
148	0.000172	69	79
149	0.000171	69	80
150	0.00017	70	80
151	0.000169	70	81
152	0.000167	71	81
153	0.000167	71	82
154	0.000165	71	83
155	0.000165	72	83
156	0.000163	72	84
157	0.000163	73	84
158	0.000161	73	85

Table A.4: Optimal synonymous codon allocation schemes for glycine.



Optimal Codon Allocation: Glycine

Givens

t =	0.0133	<-- Gly1
	0.0199	<-- Gly2
	0.0678	<-- Gly3

	GGA	**GGC/GGU**	**GGG**	
B =	0	0	1	**<-- Gly1**
	1	0	1	**<-- Gly2**
	0	1	0	**<-- Gly3**

Optimal Results

# AAs	**Eff**	**CCA**	**CCC**	**CCU**
1	0.067819025	0	1	0
2	0.033909512	0	2	0
3	0.033248281	0	2	1
4	0.022606342	0	3	1
5	0.016954756	0	4	1
6	0.01662414	0	4	2
7	0.013563805	0	5	2
8	0.011303171	0	6	2
9	0.01108276	0	6	3
10	0.009688432	0	7	3
11	0.008477378	0	8	3
12	0.00831207	0	8	4
13	0.007535447	0	9	4
14	0.006781902	0	10	4
15	0.006649656	0	10	5
16	0.006165366	0	11	5
17	0.005651585	0	12	5
18	0.00554138	0	12	6
19	0.005216848	0	13	6
20	0.004844216	0	14	6
21	0.004749754	0	14	7
22	0.004521268	0	15	7
23	0.004238689	0	16	7
24	0.004156035	0	16	8
25	0.003989354	0	17	8
26	0.003767724	0	18	8
27	0.003694253	0	18	9
28	0.003569422	0	19	9
29	0.003390951	0	20	9

Table A.4, continued.

# AAs	Eff	CCA	CCC	CCU
30	0.003324828	0	20	10
31	0.003229477	0	21	10
32	0.003082683	0	22	10
33	0.003022571	0	22	11
34	0.002948653	0	23	11
35	0.002825793	0	24	11
36	0.00277069	0	24	12
37	0.002712761	0	25	12
38	0.002608424	0	26	12
39	0.00255756	0	26	13
40	0.002511816	0	27	13
41	0.002422108	0	28	13
42	0.002374877	0	28	14
43	0.002338587	0	29	14
44	0.002260634	0	30	14
45	0.002216552	0	30	15
46	0.00218771	0	31	15
47	0.002119345	0	32	15
48	0.002078018	0	32	16
49	0.002055122	0	33	16
50	0.001994677	0	34	16
51	0.001955781	0	34	17
52	0.001937686	0	35	17
53	0.001883862	0	36	17
54	0.001847127	0	36	18
55	0.001832947	0	37	18
56	0.001784711	0	38	18
57	0.00174991	0	38	19
58	0.001738949	0	39	19
59	0.001695476	0	40	19
60	0.001662414	0	40	20
61	0.001654123	0	41	20
62	0.001614739	0	42	20
63	0.001583251	0	42	21
64	0.001577187	0	43	21
65	0.001541341	0	44	21
66	0.001511285	0	44	22
67	0.001507089	0	45	22
68	0.001474327	0	46	22
69	0.001445577	0	46	23
70	0.001442958	0	47	23
71	0.001412896	0	48	23
72	0.001385345	0	48	24

Table A.4, continued.

# AAs	Eff	CCA	CCC	CCU
73	0.001384062	0	49	24
74	0.00135638	0	50	24
75	0.001329931	0	50	25
76	0.001329785	0	51	25
77	0.001304212	0	52	25
78	0.001279604	0	53	25
79	0.00127878	0	53	26
80	0.001255908	0	54	26
81	0.001233073	0	55	26
83	0.001211054	0	56	27
84	0.001189807	0	57	27
85	0.001187439	0	57	28
86	0.001169294	0	58	28
87	0.001149475	0	59	28
89	0.001130317	0	60	29
90	0.001111787	0	61	29
92	0.001093855	0	62	30
94	0.001072525	0	63	31
97	0.001039009	0	65	32
102	0.000982884	0	69	33
104	0.000968843	0	70	34
106	0.000949951	0	71	35
113	0.000892356	0	76	37
114	0.000880767	0	77	37
115	0.000874955	0	77	38
117	0.000858469	0	79	38
118	0.00085252	0	79	39
124	0.000810934	0	83	41
125	0.000807369	0	84	41
131	0.000770671	0	88	43
134	0.000753545	0	90	44
135	0.000745264	0	91	44
137	0.000737163	0	92	45
140	0.000721479	0	94	46
155	0.000651927	0	104	51
159	0.000633823	0	107	52
185	0.000545054	0	124	61
206	0.000488945	0	138	68
213	0.000474259	0	143	70

Table A.5: Optimal synonymous codon allocation schemes for isoleucine.

Optimal Codon Allocation: Isoleucine

Givens

t =	0.0513	<-- Ile1	
	0.0027	<-- Ile2	
	AUA	**AUC/AUU**	
B =	0	1	**<-- Ile1**
	1	0	**<-- Ile2**

Optimal Results

# AAs	Eff	AUA	AUC/AUU
1	0.051342689	0	1
2	0.025671345	0	2
3	0.01711423	0	3
4	0.012835672	0	4
5	0.010268538	0	5
6	0.008557115	0	6
7	0.00733467	0	7
8	0.006417835	0	8
9	0.005704743	0	9
10	0.005134269	0	10
11	0.004667517	0	11
12	0.004278557	0	12
13	0.003949438	0	13
14	0.003667335	0	14
15	0.003422846	0	15
16	0.003208918	0	16
17	0.003020158	0	17
18	0.002852372	0	18
19	0.00270716	1	18
20	0.002702247	1	19
21	0.002567134	1	20
22	0.00244489	1	21
23	0.002333759	1	22
24	0.002232291	1	23
25	0.002139279	1	24
26	0.002053708	1	25
27	0.001974719	1	26
28	0.001901581	1	27
29	0.001833667	1	28
30	0.001770438	1	29
31	0.001711423	1	30

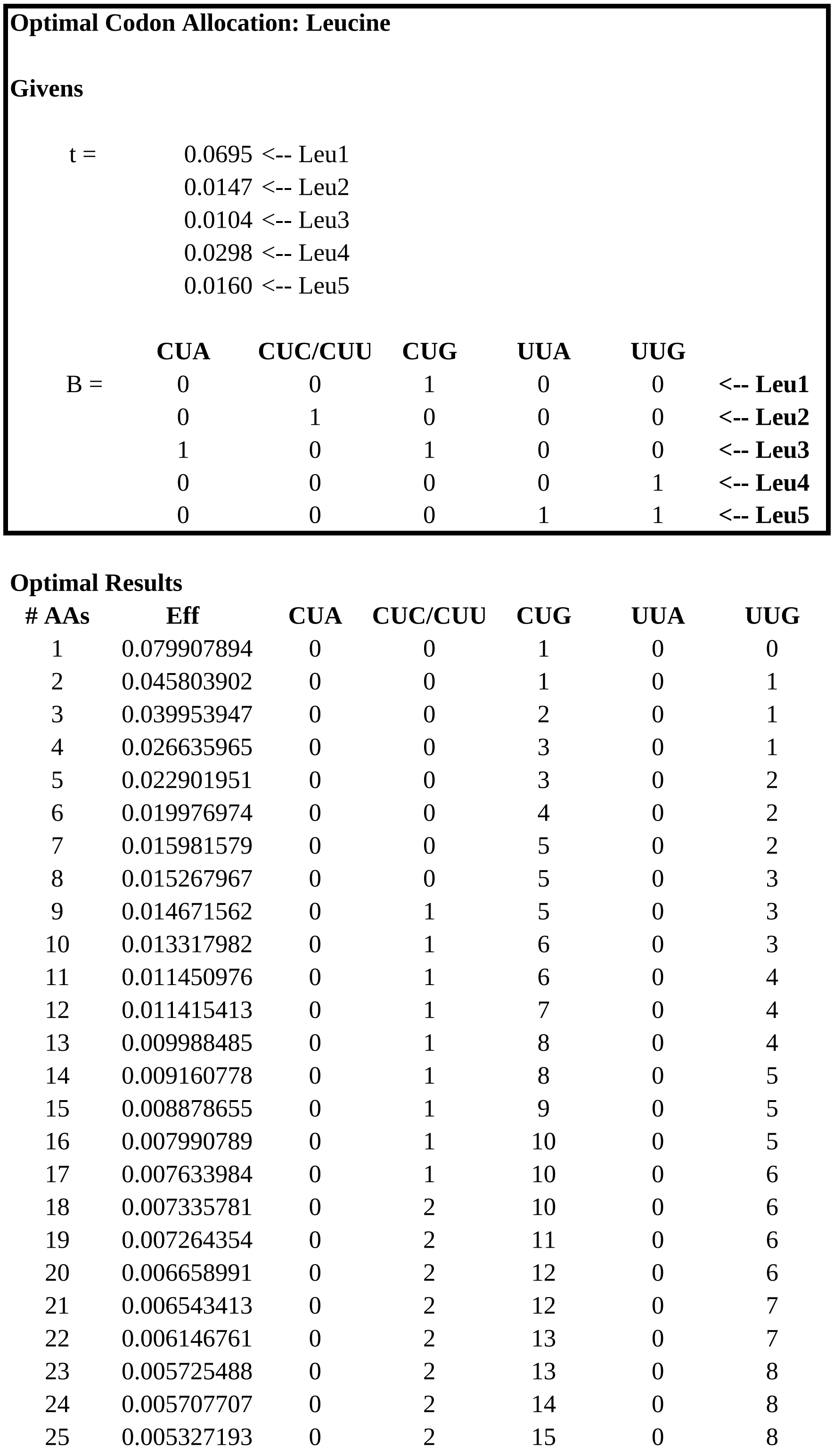
Optimal Results

# AAs	Eff	AUA	AUC/AUU
32	0.001656	1	31
33	0.001604	1	32
34	0.001556	1	33
35	0.00151	1	34
36	0.001467	1	35
37	0.001426	1	36
38	0.001388	1	37
39	0.001354	2	37
40	0.001351	2	38
41	0.001316	2	39
42	0.001284	2	40
43	0.001252	2	41
44	0.001222	2	42
45	0.001194	2	43
46	0.001167	2	44
47	0.001141	2	45
48	0.001116	2	46
49	0.001092	2	47
50	0.00107	2	48
51	0.001048	2	49
52	0.001027	2	50
53	0.001007	2	51
54	0.000987	2	52
55	0.000969	2	53
56	0.000951	2	54
57	0.000934	2	55
58	0.000917	2	56
59	0.000902	3	56
60	0.000901	3	57
61	0.000885	3	58
62	0.00087	3	59

Table A.5, continued.

# AAs	Eff	AUA	AUC/AUU
63	0.000855711	3	60
64	0.000841683	3	61
65	0.000828108	3	62
66	0.000814963	3	63
67	0.00080223	3	64
68	0.000789888	3	65
69	0.00077792	3	66
70	0.000766309	3	67
71	0.00075504	3	68
72	0.000744097	3	69
73	0.000733467	3	70
74	0.000723136	3	71
75	0.000713093	3	72
76	0.000703324	3	73
77	0.00069382	3	74
78	0.000684569	3	75
79	0.00067679	4	75
80	0.000675562	4	76
81	0.000666788	4	77
82	0.00065824	4	78
83	0.000649907	4	79
84	0.000641784	4	80
85	0.00063386	4	81
86	0.00062613	4	82
87	0.000618587	4	83
88	0.000611222	4	84
89	0.000604032	4	85
90	0.000597008	4	86
91	0.000590146	4	87
92	0.00058344	4	88
93	0.000576884	4	89
94	0.000570474	4	90
95	0.000564205	4	91
96	0.000558073	4	92

Table A.6: Optimal synonymous codon allocation schemes for leucine.



Optimal Codon Allocation: Leucine

Givens

t =	0.0695	<-- Leu1
	0.0147	<-- Leu2
	0.0104	<-- Leu3
	0.0298	<-- Leu4
	0.0160	<-- Leu5

	CUA	**CUC/CUU**	**CUG**	**UUA**	**UUG**	
B =	0	0	1	0	0	**<-- Leu1**
	0	1	0	0	0	**<-- Leu2**
	1	0	1	0	0	**<-- Leu3**
	0	0	0	0	1	**<-- Leu4**
	0	0	0	1	1	**<-- Leu5**

Optimal Results

# AAs	**Eff**	**CUA**	**CUC/CUU**	**CUG**	**UUA**	**UUG**
1	0.079907894	0	0	1	0	0
2	0.045803902	0	0	1	0	1
3	0.039953947	0	0	2	0	1
4	0.026635965	0	0	3	0	1
5	0.022901951	0	0	3	0	2
6	0.019976974	0	0	4	0	2
7	0.015981579	0	0	5	0	2
8	0.015267967	0	0	5	0	3
9	0.014671562	0	1	5	0	3
10	0.013317982	0	1	6	0	3
11	0.011450976	0	1	6	0	4
12	0.011415413	0	1	7	0	4
13	0.009988485	0	1	8	0	4
14	0.009160778	0	1	8	0	5
15	0.008878655	0	1	9	0	5
16	0.007990789	0	1	10	0	5
17	0.007633984	0	1	10	0	6
18	0.007335781	0	2	10	0	6
19	0.007264354	0	2	11	0	6
20	0.006658991	0	2	12	0	6
21	0.006543413	0	2	12	0	7
22	0.006146761	0	2	13	0	7
23	0.005725488	0	2	13	0	8
24	0.005707707	0	2	14	0	8
25	0.005327193	0	2	15	0	8

Table A.6, continued.

# AAs	Eff	CUA	CUC/CUU	CUG	UUA	UUG
26	0.005089322	0	2	15	0	9
27	0.004994243	0	2	16	0	9
28	0.004890521	0	3	16	0	9
29	0.004700464	0	3	17	0	9
30	0.00458039	0	3	17	0	10
31	0.004439327	0	3	18	0	10
32	0.004205679	0	3	19	0	10
33	0.004163991	0	3	19	0	11
34	0.003995395	0	3	20	0	11
35	0.003816992	0	3	20	0	12
36	0.003805138	0	3	21	0	12
37	0.003667891	0	4	21	0	12
38	0.003632177	0	4	22	0	12
39	0.003523377	0	4	22	0	13
40	0.003474256	0	4	23	0	13
41	0.003329496	0	4	24	0	13
42	0.003271707	0	4	24	0	14
43	0.003196316	0	4	25	0	14
44	0.003073381	0	4	26	0	14
45	0.003053593	0	4	26	0	15
46	0.002959552	0	4	27	0	15
47	0.002934312	0	5	27	0	15
48	0.002862744	0	5	27	0	16
49	0.002853853	0	5	28	0	16
50	0.002755445	0	5	29	0	16
51	0.002694347	0	5	29	0	17
52	0.002663596	0	5	30	0	17
53	0.002577674	0	5	31	0	17
54	0.002544661	0	5	31	0	18
55	0.002497122	0	5	32	0	18
56	0.00244526	0	6	32	0	18
57	0.002421451	0	6	33	0	18
58	0.002410732	0	6	33	0	19
59	0.002350232	0	6	34	0	19
60	0.002290195	0	6	34	0	20
61	0.002283083	0	6	35	0	20
62	0.002219664	0	6	36	0	20
63	0.002181138	0	6	36	0	21
64	0.002159673	0	6	37	0	21
65	0.002102839	0	6	38	0	21
66	0.002095937	0	7	38	0	21
67	0.002081996	0	7	38	0	22
68	0.00204892	0	7	39	0	22

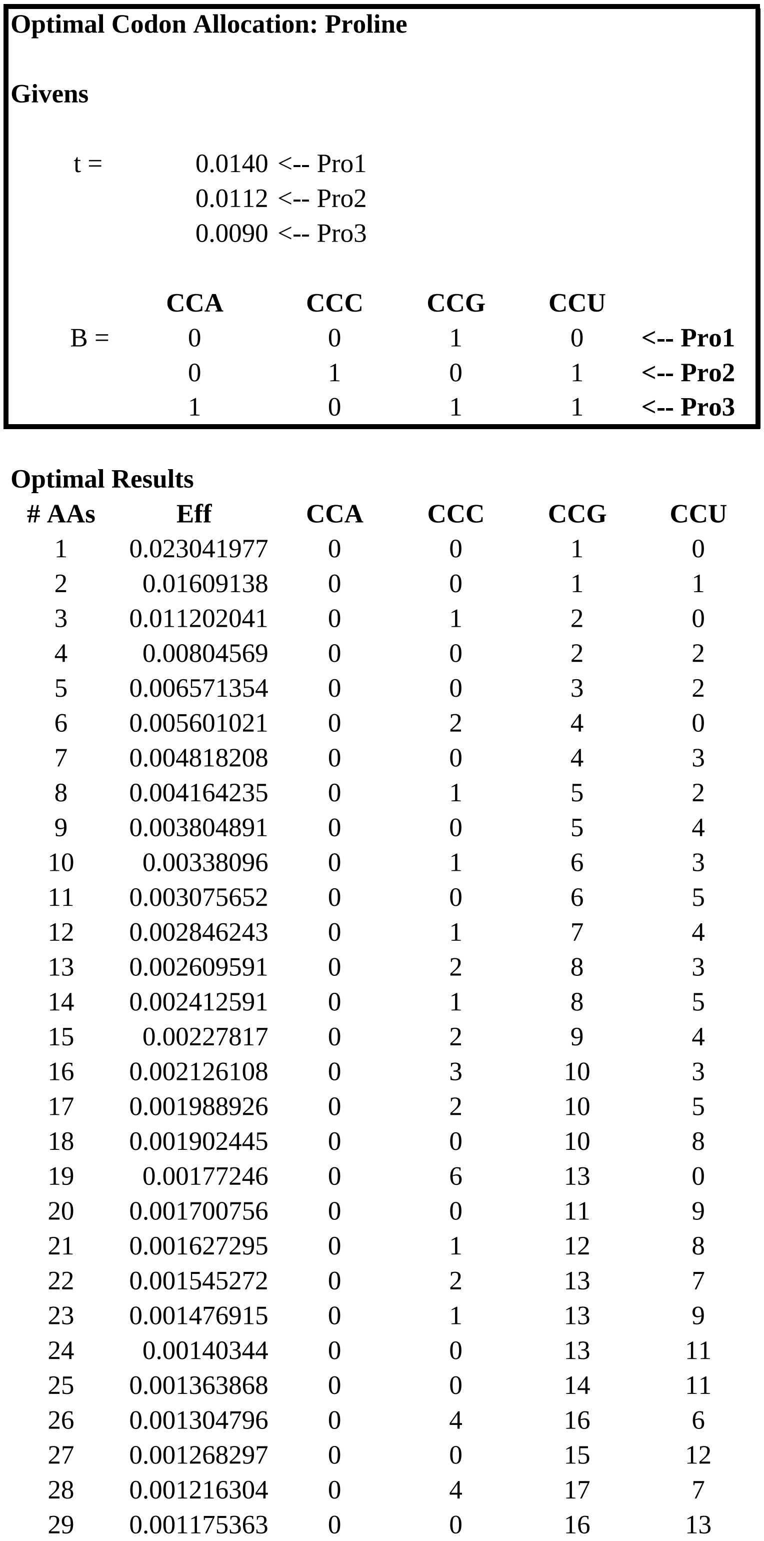
Table A.6, continued.

# AAs	Eff	CUA	CUC/CUU	CUG	UUA	UUG
69	0.001997697	0	7	40	0	22
70	0.001991474	0	7	40	0	23
71	0.001948973	0	7	41	0	23
72	0.001908496	0	7	41	0	24
73	0.001902569	0	7	42	0	24
74	0.001858323	0	7	43	0	24
75	0.001833945	0	8	43	0	24
76	0.001832156	0	8	43	0	25
77	0.001816089	0	8	44	0	25
78	0.001775731	0	8	45	0	25
79	0.001761689	0	8	45	0	26
80	0.001737128	0	8	46	0	26
81	0.001700168	0	8	47	0	26
82	0.001696441	0	8	47	0	27
83	0.001664748	0	8	48	0	27
84	0.001635854	0	8	48	0	28
85	0.001630773	0	8	49	0	28
86	0.001630174	0	9	49	0	28
87	0.001598158	0	9	50	0	28
88	0.001579445	0	9	50	0	29
89	0.001566821	0	9	51	0	29
90	0.00153669	0	9	52	0	29
91	0.001526797	0	9	52	0	30
92	0.001507696	0	9	53	0	30
93	0.001479776	0	9	54	0	30
94	0.001477545	0	9	54	0	31
95	0.001467156	0	10	54	0	31
96	0.001452871	0	10	55	0	31
97	0.001431372	0	10	55	0	32
98	0.001426927	0	10	56	0	32
99	0.001401893	0	10	57	0	32
100	0.001387997	0	10	57	0	33
102	0.001354371	0	10	59	0	33
103	0.001347174	0	10	59	0	34
104	0.001333778	0	11	59	0	34
105	0.001331798	0	11	60	0	34
106	0.001309965	0	11	61	0	34
107	0.001308683	0	11	61	0	35
108	0.001288837	0	11	62	0	35
109	0.001272331	0	11	62	0	36
110	0.001268379	0	11	63	0	36
111	0.001248561	0	11	64	0	36
112	0.001237943	0	11	64	0	37

Table A.6, continued.

# AAs	Eff	CUA	CUC/CUU	CUG	UUA	UUG
113	0.001229352	0	11	65	0	37
114	0.00122263	0	12	65	0	37
115	0.001210726	0	12	66	0	37
116	0.001205366	0	12	66	0	38
117	0.001192655	0	12	67	0	38
118	0.001175116	0	12	68	0	38
119	0.001174459	0	12	68	0	39
120	0.001158085	0	12	69	0	39
121	0.001145098	0	12	69	0	40
122	0.001141541	0	12	70	0	40
123	0.001128582	0	13	70	0	40
124	0.001125463	0	13	71	0	40
125	0.001117168	0	13	71	0	41
126	0.001109832	0	13	72	0	41
127	0.001094629	0	13	73	0	41
128	0.001090569	0	13	73	0	42
129	0.001079836	0	13	74	0	42
130	0.001065439	0	13	75	0	42
131	0.001065207	0	13	75	0	43
132	0.00105142	0	13	76	0	43
139	0.000998849	0	14	80	0	45
140	0.000995737	0	14	80	0	46
142	0.000978104	0	15	81	0	46
144	0.000974487	0	15	82	0	47
145	0.000962746	0	15	83	0	47
150	0.000929162	0	15	86	0	49
153	0.000916078	0	16	87	0	50
154	0.000908044	0	16	88	0	50
156	0.000897842	0	16	89	0	51
159	0.000878109	0	16	91	0	52
160	0.000868564	0	16	92	0	52
161	0.000864225	0	16	92	0	53
165	0.00084822	0	17	94	0	54
170	0.000817927	0	17	97	0	56
177	0.000789722	0	18	101	0	58
185	0.000753848	0	19	106	0	60
200	0.000698646	0	21	114	0	65

Table A.7: Optimal synonymous codon allocation schemes for proline.



Optimal Codon Allocation: Proline

Givens

t =	0.0140	<-- Pro1
	0.0112	<-- Pro2
	0.0090	<-- Pro3

	CCA	**CCC**	**CCG**	**CCU**	
B =	0	0	1	0	**<-- Pro1**
	0	1	0	1	**<-- Pro2**
	1	0	1	1	**<-- Pro3**

Optimal Results

# AAs	**Eff**	**CCA**	**CCC**	**CCG**	**CCU**
1	0.023041977	0	0	1	0
2	0.01609138	0	0	1	1
3	0.011202041	0	1	2	0
4	0.00804569	0	0	2	2
5	0.006571354	0	0	3	2
6	0.005601021	0	2	4	0
7	0.004818208	0	0	4	3
8	0.004164235	0	1	5	2
9	0.003804891	0	0	5	4
10	0.00338096	0	1	6	3
11	0.003075652	0	0	6	5
12	0.002846243	0	1	7	4
13	0.002609591	0	2	8	3
14	0.002412591	0	1	8	5
15	0.00227817	0	2	9	4
16	0.002126108	0	3	10	3
17	0.001988926	0	2	10	5
18	0.001902445	0	0	10	8
19	0.00177246	0	6	13	0
20	0.001700756	0	0	11	9
21	0.001627295	0	1	12	8
22	0.001545272	0	2	13	7
23	0.001476915	0	1	13	9
24	0.00140344	0	0	13	11
25	0.001363868	0	0	14	11
26	0.001304796	0	4	16	6
27	0.001268297	0	0	15	12
28	0.001216304	0	4	17	7
29	0.001175363	0	0	16	13

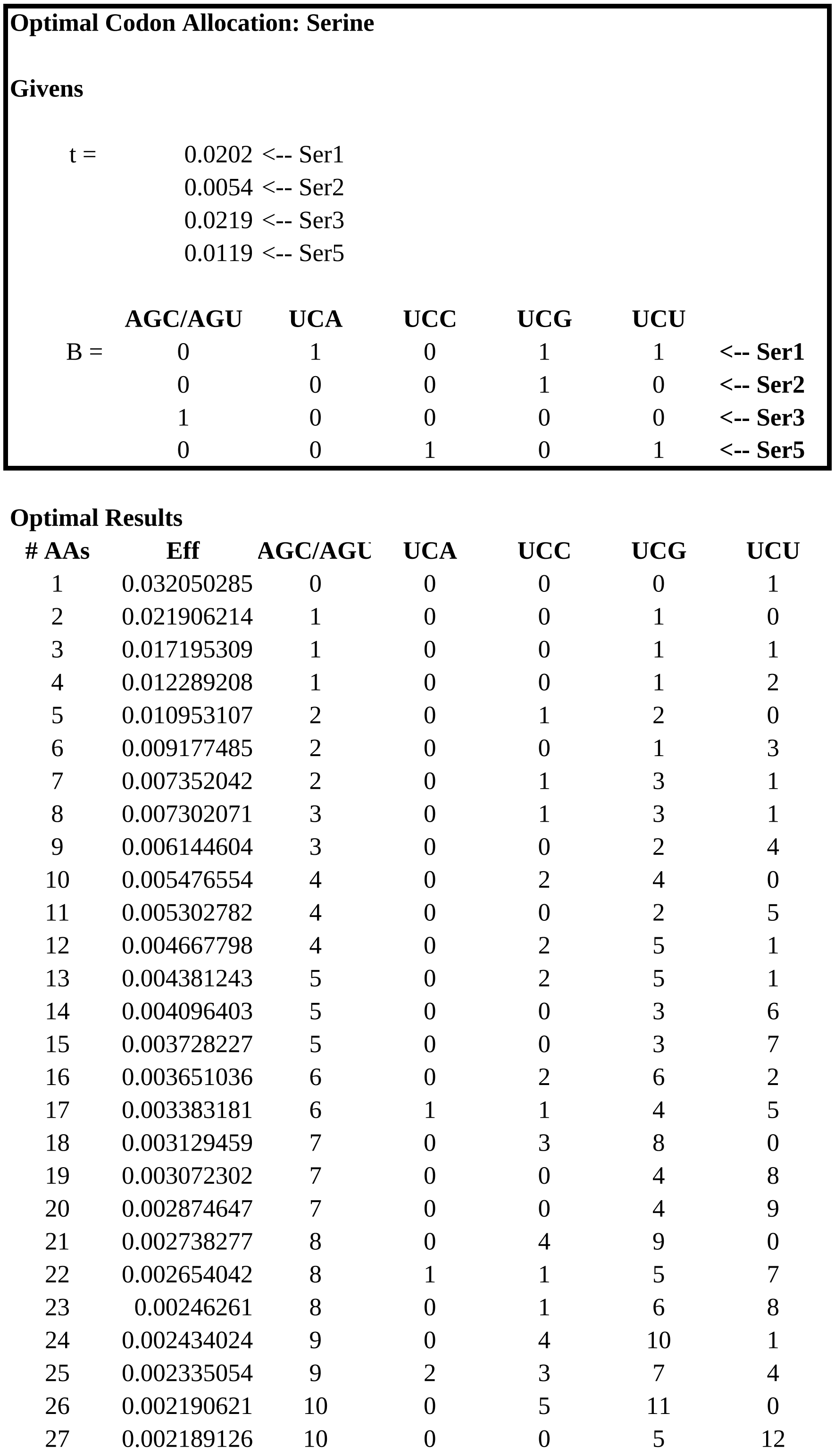
Table A.7, continued.

# AAs	Eff	CCA	CCC	CCG	CCU
30	0.001139085	0	4	18	8
31	0.001099624	0	5	19	7
32	0.001064043	0	1	18	13
33	0.001035342	0	5	20	8
34	0.001003985	0	0	19	15
35	0.000970885	0	5	21	9
36	0.000951223	0	0	20	16
37	0.000921679	0	12	25	0
38	0.000897969	0	0	21	17
39	0.000870836	0	0	22	17
40	0.000853407	0	13	27	0
41	0.000831546	0	1	23	17
42	0.000807591	0	0	23	19
43	0.000794551	0	14	29	0
44	0.000771663	0	8	27	9
45	0.000760978	0	0	25	20
46	0.00074329	0	15	31	0
47	0.000726508	0	0	26	21
48	0.000711967	0	7	29	12
49	0.000698242	0	16	33	0
50	0.000681934	0	0	28	22
51	0.000669769	0	8	31	12
52	0.000658342	0	17	35	0
53	0.000642181	0	8	32	13
54	0.000634148	0	0	30	24
55	0.000622336	0	18	37	0
56	0.000610028	0	0	31	25
57	0.000598718	0	10	35	12
59	0.000578289	0	0	33	26
60	0.000568533	1	11	36	12
61	0.000560383	0	0	34	27
62	0.000550985	0	13	39	10
63	0.000543556	0	0	35	28
65	0.000525738	0	0	36	29
67	0.000510364	1	11	40	15
71	0.000482043	1	16	44	10
72	0.000475611	0	0	40	32
73	0.000468771	1	13	44	15
75	0.000456169	1	18	47	9
76	0.000450066	1	14	46	15
78	0.000438934	1	19	49	9
79	0.000432875	0	0	44	35
81	0.000422766	0	0	45	36

Table A.7, continued.

# AAs	Eff	CCA	CCC	CCG	CCU
83	0.000412427	1	17	51	14
86	0.000397986	1	18	53	14
87	0.000393208	1	19	54	13
92	0.000371847	1	17	56	18
93	0.000368013	1	21	58	13
132	0.000259237	1	36	85	10

Table A.8: Optimal synonymous codon allocation schemes for serine.



Optimal Codon Allocation: Serine

Givens

t =	0.0202	<-- Ser1
	0.0054	<-- Ser2
	0.0219	<-- Ser3
	0.0119	<-- Ser5

	AGC/AGU	**UCA**	**UCC**	**UCG**	**UCU**	
B =	0	1	0	1	1	**<-- Ser1**
	0	0	0	1	0	**<-- Ser2**
	1	0	0	0	0	**<-- Ser3**
	0	0	1	0	1	**<-- Ser5**

Optimal Results

# AAs	**Eff**	**AGC/AGU**	**UCA**	**UCC**	**UCG**	**UCU**
1	0.032050285	0	0	0	0	1
2	0.021906214	1	0	0	1	0
3	0.017195309	1	0	0	1	1
4	0.012289208	1	0	0	1	2
5	0.010953107	2	0	1	2	0
6	0.009177485	2	0	0	1	3
7	0.007352042	2	0	1	3	1
8	0.007302071	3	0	1	3	1
9	0.006144604	3	0	0	2	4
10	0.005476554	4	0	2	4	0
11	0.005302782	4	0	0	2	5
12	0.004667798	4	0	2	5	1
13	0.004381243	5	0	2	5	1
14	0.004096403	5	0	0	3	6
15	0.003728227	5	0	0	3	7
16	0.003651036	6	0	2	6	2
17	0.003383181	6	1	1	4	5
18	0.003129459	7	0	3	8	0
19	0.003072302	7	0	0	4	8
20	0.002874647	7	0	0	4	9
21	0.002738277	8	0	4	9	0
22	0.002654042	8	1	1	5	7
23	0.00246261	8	0	1	6	8
24	0.002434024	9	0	4	10	1
25	0.002335054	9	2	3	7	4
26	0.002190621	10	0	5	11	0
27	0.002189126	10	0	0	5	12

Table A.8, continued.

# AAs	Eff	AGC/AGU	UCA	UCC	UCG	UCU
28	0.00206104	10	1	1	6	10
29	0.001991474	11	0	5	12	1
30	0.001962751	11	0	6	13	0
31	0.001864114	11	0	0	6	14
32	0.001825518	12	0	5	13	2
33	0.001774932	12	3	4	9	5
34	0.001698087	12	0	7	15	0
35	0.001685093	13	0	7	15	0
36	0.001623132	13	0	0	7	16
37	0.00156473	14	0	7	16	0
38	0.001556204	14	2	6	13	3
39	0.001487353	14	2	2	9	12
40	0.001460414	15	0	8	17	0
41	0.001437324	15	0	0	8	18
42	0.001383139	15	1	8	17	1
43	0.001369138	16	0	8	18	1
44	0.001328548	16	2	3	11	12
45	0.001289687	16	0	0	9	20
46	0.001288601	17	0	0	9	20
47	0.001244259	17	0	9	20	1
48	0.001217012	18	0	8	20	2
49	0.001201052	18	3	3	11	14
50	0.001167527	18	4	6	14	8
51	0.001152959	19	0	10	22	0
52	0.001130816	19	0	0	10	23
53	0.001096113	19	3	3	12	16
54	0.001095311	20	3	3	12	16
55	0.001067479	20	2	10	21	2
56	0.001043153	21	0	11	24	0
57	0.001037386	21	0	0	11	25
58	0.001009544	21	5	8	17	7
59	0.000995737	22	0	11	25	1
60	0.000982294	22	0	11	25	2
61	0.000958216	22	0	0	12	27
62	0.000952444	23	4	9	20	6
63	0.000934197	23	4	10	21	5
64	0.000912759	24	0	13	27	0
67	0.000876249	25	0	13	29	0
68	0.000869069	25	3	12	25	3
69	0.000849154	25	5	5	16	18
70	0.000842547	26	0	14	30	0
71	0.0008308	26	0	0	14	31
74	0.000795005	27	5	5	17	20

Table A.8, continued.

# AAs	Eff	AGC/AGU	UCA	UCC	UCG	UCU
75	0.000782365	28	0	15	32	0
76	0.000778622	28	5	12	25	6
79	0.000747447	29	5	5	18	22
80	0.000732844	29	4	14	29	4
82	0.000718999	30	6	6	19	21
84	0.000704888	31	5	5	19	24
85	0.000691569	31	2	16	34	2
86	0.000684569	32	0	17	37	0
87	0.000679799	32	6	6	20	23
90	0.000655989	33	2	17	36	2
92	0.0006443	34	0	0	18	40
93	0.000633537	34	5	16	33	5
94	0.000625892	35	0	18	40	1
96	0.000612733	35	0	0	19	42
98	0.000602527	36	4	18	37	3
100	0.00059206	37	0	20	43	0
105	0.000561698	39	0	21	45	0
107	0.000549756	39	0	0	21	47
108	0.000547655	40	0	21	46	1
111	0.000533854	41	5	20	41	4
131	0.000450136	48	0	22	53	8
156	0.000377693	58	0	31	67	0
227	0.000260788	84	0	45	97	1

Table A.9: Optimal synonymous codon allocation schemes for threonine.

Optimal Codon Allocation: Threonine

Givens

t =	0.0016	<-- Thr1
	0.0084	<-- Thr2
	0.0170	<-- Thr3
	0.0143	<-- Thr4

	ACA	**ACC**	**ACG**	**ACU**	
B =	0	1	0	1	**<-- Thr1**
	0	0	1	0	**<-- Thr2**
	0	1	0	1	**<-- Thr3**
	1	0	1	1	**<-- Thr4**

Optimal Results

# AAs	**Eff**	**ACA**	**ACC**	**ACG**	**ACU**
1	0.032905996	0	0	0	1
2	0.01865451	0	1	1	0
3	0.011905338	0	1	1	1
4	0.009614212	0	1	2	1
5	0.008179338	0	1	2	2
6	0.006709022	0	2	3	1
7	0.005755204	0	2	3	2
8	0.005152234	0	3	4	1
9	0.004533715	0	4	5	0
10	0.004089669	0	2	4	4
11	0.003730902	0	5	6	0
12	0.003386128	0	5	6	1
13	0.003160384	0	4	6	3
14	0.002917109	0	5	7	2
15	0.002726446	0	3	6	6
16	0.002576117	0	6	8	2
17	0.002415403	0	7	9	1
18	0.002289102	1	5	7	5
19	0.002165262	0	8	10	1
20	0.00206078	0	9	11	0
21	0.001963717	1	7	9	4
22	0.001868279	1	8	10	3
23	0.001794003	1	4	8	10
24	0.001717411	0	9	12	3
25	0.001645773	0	10	13	2
26	0.001585776	1	8	11	6
27	0.001524807	0	11	14	2

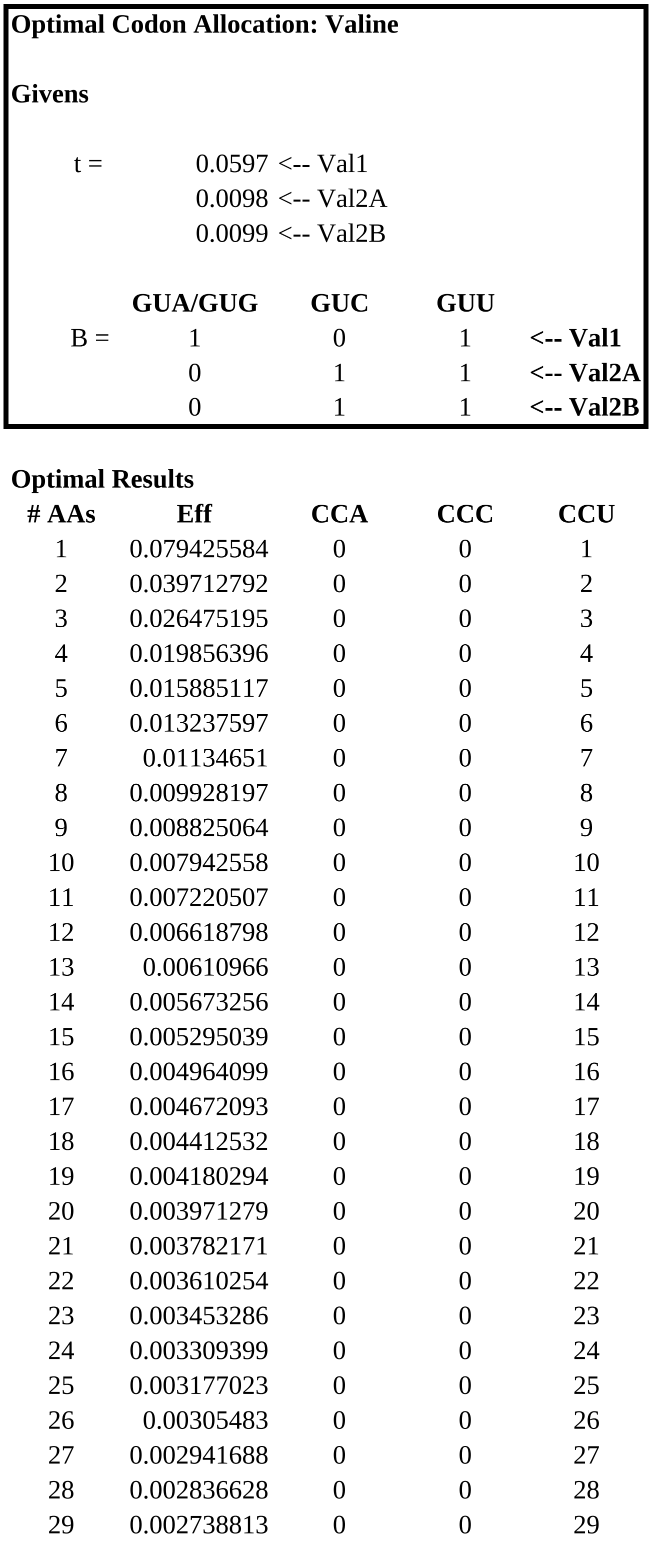
Table A.9, continued.

# AAs	Eff	ACA	ACC	ACG	ACU
28	0.001473277	0	12	15	1
29	0.001421958	1	10	13	5
30	0.001371044	1	11	14	4
31	0.001332465	0	14	17	0
32	0.001288058	0	12	16	4
33	0.001248891	1	13	16	3
34	0.001212846	2	11	14	7
35	0.001176672	0	14	18	3
36	0.001146908	1	15	18	2
37	0.00111445	1	13	17	6
38	0.001083375	2	14	17	5
39	0.001058991	0	17	21	1
40	0.001030447	0	15	20	5
41	0.001005512	1	16	20	4
42	0.000981859	2	14	18	8
43	0.000958395	2	15	19	7
44	0.000938386	1	18	22	3
45	0.000916283	0	16	22	7
46	0.000896666	2	17	21	6
47	0.000878576	0	20	25	2
48	0.000859122	0	21	26	1
49	0.000842622	2	19	23	5
50	0.000824894	0	22	27	1
51	0.000809592	0	23	28	0
52	0.000794035	1	21	26	4
53	0.000777945	0	19	26	8
54	0.000763723	2	20	25	7
55	0.000750654	0	23	29	3
56	0.000737067	3	19	24	10
57	0.000724236	2	22	27	6
58	0.00071117	0	25	31	2
59	0.000699943	0	26	32	1
60	0.000688481	2	24	29	5
61	0.000676047	3	20	26	12
62	0.000666249	3	23	28	8
63	0.000655234	0	26	33	4
64	0.000644772	1	27	33	3
65	0.000635129	3	25	30	7
66	0.00062498	0	28	35	3
67	0.000616502	1	29	35	2
68	0.000607334	2	27	33	6
69	0.000597868	4	23	29	13
70	0.000590058	0	31	38	1

Table A.9, continued.

# AAs	Eff	ACA	ACC	ACG	ACU
71	0.00058133	0	29	37	5
72	0.000573454	2	30	36	4
73	0.000565883	3	28	34	8
74	0.000557542	3	29	35	7
75	0.000550754	1	32	39	3
76	0.000543298	1	30	38	7
77	0.000536243	4	26	33	14
78	0.000529496	0	34	42	2
81	0.000509881	3	31	38	9
83	0.000497833	2	35	42	4
84	0.000491476	0	33	43	8
99	0.000417273	2	41	50	6
101	0.00040901	4	40	48	9
106	0.000389766	1	46	56	3
110	0.000375395	0	49	60	1
113	0.000365622	0	51	62	0
114	0.000362472	2	49	59	4
118	0.000349982	1	52	63	2
122	0.000338665	2	52	63	5
139	0.000297088	3	59	71	6
169	0.000244446	8	62	76	23
273	0.000151354	1	122	148	2

Table A.10: Optimal synonymous codon allocation schemes for valine.



Optimal Codon Allocation: Valine

Givens

t =	0.0597	<-- Val1
	0.0098	<-- Val2A
	0.0099	<-- Val2B

	GUA/GUG	**GUC**	**GUU**	
B =	1	0	1	**<-- Val1**
	0	1	1	**<-- Val2A**
	0	1	1	**<-- Val2B**

Optimal Results

# AAs	**Eff**	**CCA**	**CCC**	**CCU**
1	0.079425584	0	0	1
2	0.039712792	0	0	2
3	0.026475195	0	0	3
4	0.019856396	0	0	4
5	0.015885117	0	0	5
6	0.013237597	0	0	6
7	0.01134651	0	0	7
8	0.009928197	0	0	8
9	0.008825064	0	0	9
10	0.007942558	0	0	10
11	0.007220507	0	0	11
12	0.006618798	0	0	12
13	0.00610966	0	0	13
14	0.005673256	0	0	14
15	0.005295039	0	0	15
16	0.004964099	0	0	16
17	0.004672093	0	0	17
18	0.004412532	0	0	18
19	0.004180294	0	0	19
20	0.003971279	0	0	20
21	0.003782171	0	0	21
22	0.003610254	0	0	22
23	0.003453286	0	0	23
24	0.003309399	0	0	24
25	0.003177023	0	0	25
26	0.00305483	0	0	26
27	0.002941688	0	0	27
28	0.002836628	0	0	28
29	0.002738813	0	0	29

Table A.10, continued.

# AAs	Eff	CCA	CCC	CCU
30	0.002647519	0	0	30
31	0.002562116	0	0	31
32	0.002482049	0	0	32
33	0.002406836	0	0	33
34	0.002336047	0	0	34
35	0.002269302	0	0	35
36	0.002206266	0	0	36
37	0.002146637	0	0	37
38	0.002090147	0	0	38
39	0.002036553	0	0	39
40	0.00198564	0	0	40
41	0.001937209	0	0	41
42	0.001891085	0	0	42
43	0.001847107	0	0	43
44	0.001805127	0	0	44
45	0.001765013	0	0	45
46	0.001726643	0	0	46
47	0.001689906	0	0	47
48	0.0016547	0	0	48
49	0.00162093	0	0	49
50	0.001588512	0	0	50
51	0.001557364	0	0	51
52	0.001527415	0	0	52
53	0.001498596	0	0	53
54	0.001470844	0	0	54
55	0.001444102	0	0	55
56	0.001418314	0	0	56
57	0.001393431	0	0	57
58	0.001369407	0	0	58
59	0.001346196	0	0	59
60	0.00132376	0	0	60
61	0.001302059	0	0	61
62	0.001281058	0	0	62
63	0.001260724	0	0	63
64	0.001241025	0	0	64
65	0.001221932	0	0	65
66	0.001203418	0	0	66
67	0.001185456	0	0	67
68	0.001168023	0	0	68
69	0.001151095	0	0	69
70	0.001134651	0	0	70
71	0.00111867	0	0	71
72	0.001103133	0	0	72

Table A.10, continued.

# AAs	Eff	CCA	CCC	CCU
73	0.001088022	0	0	73
74	0.001073319	0	0	74
75	0.001059008	0	0	75
77	0.001031501	0	0	77
78	0.001018277	0	0	78
79	0.001005387	0	0	79
80	0.00099282	0	0	80
82	0.000968605	0	0	82
83	0.000956935	0	0	83
85	0.000934419	0	0	85
87	0.000912938	0	0	87
89	0.000892422	0	0	89
90	0.000882506	0	0	90
91	0.000872809	0	0	91
92	0.000863322	0	0	92
94	0.000844953	0	0	94
96	0.00082735	0	0	96
97	0.00081882	0	0	97
99	0.000802279	0	0	99
101	0.000786392	0	0	101
102	0.000778682	0	0	102
103	0.000771122	0	0	103
107	0.000740564	3	0	104
108	0.000735422	0	0	108
109	0.000728675	0	0	109
110	0.000722051	0	0	110
116	0.000684703	0	0	116
117	0.000678851	0	0	117
209	0.000379833	62	9	138

Bibliography

[1] Akashi, H. 2001. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.*, 11:660-666.

[2] Akashi, H. and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA*, 99:3695-3700.

[3] Allen, T.E. and B.O. Palsson. 2003. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J. Theor. Biol.*, 220:1-18.

[4] Allen, T.E., M.J. Herrgård, M. Liu, Y. Qiu, J.D. Glasner, F.R. Blattner, B.O. Palsson. 2003. Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.*, 185:6392-6399.

[5] Allen, T.E., N.D. Price, A.R. Joyce, and B.O. Palsson. 2005. Long-range patterns in prokaryotic genome sequences indicate significant multi-scale chromosomal organization. *PLoS Comput. Biol.*, in press. DOI: 10.1371/journal.pcbi.0020002.eor

[6] Almaas, E., B. Kovacs, T. Vicsek, Z.N. Oltvai, and A.L. Barabasi. 2004. Global orgaization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427:839-843.

[7] Alon, U., M.G. Surette, N. Barkai, and S. Leibler. 1999. Robustness in bacterial chemotaxis. *Nature*, 397:168-171.

[8] Altman, R.B. and S. Raychaudhuri. 2001. Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, 11:340-347.

[9] Arfin, S.M., A.D. Long, E.T. Ito, L. Tolleri, M.M. Riehle, E.S. Paegle, and G.W. Hatfield. 2000. Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. *J. Biol. Chem.*, 275:29672-29684.

[10] Arkin, A., J. Ross, and H.H. McAdams. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, 149:1633-1648.

[11] Arneodo, A., E. Bacry, P.V. Graves, and J.F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.*, 74:3293-3296.

[12] Audit, B. and C.A. Ouzounis. 2003. From genes to genomes: universal scale-invariant properties of microbial chromosome organisation. *J. Mol. Biol.*, 332:617-633.

[13] Audit, B., C. Vaillant, A. Arneodo, Y. D'Aubenton-Carafa, and C. Thermes. 2004. Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences. *J. Biol. Phys.*, 30:33-81.

[14] Ausmees, N., J.R. Kuhn, and C. Jacobs-Wagner. 2003. The bacterial cytoskeleton: an intermediate filament-like function in cell shape. *Cell*, 115:705-713.

[15] Bailey, J.E. 1998. Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnol. Prog.*, 14:8-20.

[16] Bailey, J.E. 2001. Complex biology with no parameters. *Nat. Biotechnol.*, 19:503-504.

[17] Barabasi, A.L. 2002. *Linked: the new science of networks*. Perseus Pub. Cambridge, Mass.

[18] Barkai, N. and S. Leibler. 1997. Robustness in simple biochemical networks. *Nature*, 387:913-917.

[19] Beard, D.A., S.D. Liang, and H. Qian. 2002. Energy balance for analysis of complex metabolic networks. *Biophys. J.*, 83:79-86.

[20] Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57:289-300.

[21] Bentley, P.M. and J.T.E. McDonnell. 1994. Wavelet transforms: an introduction. *IEE Electron. Commun. Eng. J.*, 6:175-186.

[22] Bentley, S.D. and J. Parkhill. .2004. Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* 38:771-791.

[23] Berg, O.G. and C.G. Kurland. 1997. Growth rate-optimised tRNA abundance and codon usage. *J. Mol. Biol.*, 270:544-550.

[24] Berg, O.G. and P.J.N. Silva. 1997. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res.*, 25:1397-1404.

[25] Bernstein, J.A., A.B. Khodursky, P.-H. Lin, S. Lin-Chao, and S.N. Cohen. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. USA*, 99:9697-9702.

[26] Blattner, F.R., *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453-1474.

[27] Blaylock, B., X. Jiang, A. Rubio, C.P. Moran, Jr., and K. Pogliano. 2004. Zipper-like interaction between proteins in adjacent daughter cells mediates protein localization. *Genes Dev.*, 18:2916-2928.

[28] Bonarius, H.P.J., G. Schmid, and J. Tramper. 1997. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. *Trends Biotechnol.*, 15:308-314.

[29] Bray, D. 2001. Reasoning for results. *Nature*, 412:863.

[30] Breier, A.M. and N.R. Cozzarelli. 2004. Linear ordering and dynamic segregation of the bacterial chromosome. *Proc. Natl. Acad. Sci. USA*, 101:9175-9176.

[31] Bremer, H. and P.P. Dennis. 1996. Modulation of chemical composition and other parameters of the cell by growth rate. In: Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F.C., *et al.*, eds). ASM Press, Washington.

[32] Brown, T.A. 1999. *Genomes.* Wiley-Liss, New York.

[33] Bulmer, M. 1988. Codon usage and intragenic position. *J. Theor. Biol.*, 133:67-71.

[34] Burgard, A.P. and C.D. Maranas. 2001. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.*, 74:364-75.

[35] Burgard, A.P. and C.D. Maranas. 2003. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.*, 82:670-677.

[36] Burgard, A.P., E.V. Nikolaev, C.H. Schilling, and C.D. Maranas. 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.*, 14:301-312.

[37] Bustin, S.A. and S. Dorudi. 2002. The value of microarray techniques for quantitative gene profiling in molecular diagnostics. *Trends Mol. Med.*, 8:269-272.

[38] Cabeen, M.T. and C. Jacobs-Wagner. 2005. Bacterial cell shape. *Nat. Rev. Microbiol.*, 3:601-610.

[39] Cao, D. and R. Parker. 2001. Computational modeling of eukaryotic mRNA turnover. *RNA*, 7:1192-1212.

[40] Cascante, M., L.G. Boros, B. Comin-Anduix, P. de Atauri, J.J. Centelles, and P.W. Lee. 2002. Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.*, 20:243-249.

[41] Chen, K.C., A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J.J. Tyson. 2000. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell*, 11:369-391.

[42] Chen, S.L., W. Lee, A.J. Hottes, L. Shapiro, and H.H. McAdams. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA*, 101:3480-3485.

[43] Chudin, E., R. Walker, A. Kosaka, S.X. Wu, D. Rabert, T.K. Chang, and D.E. Kreder. 2001. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.*, 3:research0005.

[44] Chvátal, V. 1983. *Linear Programming.* W.H. Freeman, New York.

[45] Clarke, B.L. 1980. Stability of complex reaction networks. *Adv. Chem. Phys.*, 43:1-215.

[46] Clarke, B.L. 1988. Stoichiometric network analysis. *Cell Biophys.*, 12:237-253.

[47] Condemine, G. and C.L. Smith. 1990. Transcription regulates oxolinic acid-induced DNA gyrase cleavage at specific sites on the *E. coli* chromosome. *Nucleic Acids Res.*, 18:7389-7396.

[48] Cook, P.R. 2002. Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.*, 32:347-352.

[49] Covert, M.W., C.H. Schilling, I. Famili, J.S. Edwards, I.I. Goryanin, E. Selkov, and B.O. Palsson. 2001. Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.*, 26:179-186.

[50] Covert, M.W., C.H. Schilling, and B.O. Palsson. 2001. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, 213:73-88.

[51] Covert, M.W. and B.O. Palsson. 2002. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.*, 277:28058-28064.

[52] Covert, M.W. and B.O. Palsson. 2003. Constraints-based models: Regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol.*, 221:309-325.

[53] Covert, M.W., I. Famili, and B.O. Palsson. 2003. Identifying constraints that govern cell behavior: a key to converting conceptual computaional models in biology? *Biotechnol. Bioeng.*, 84:763-772.

[54] Covert, M.W., E.M. Knight, J.L. Reed, M.J. Herrgård, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429:92-96.

[55] Crick, F.H.C., L. Barnett, S. Brenner, and R.J. Watts-Tobin. 1961. General nature of the genetic code for proteins. *Nature*, 192:1227-1232.

[56] Crick, F.H.C. 1966. Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, 19:548-555.

[57] Crozat, E., N. Philippe, R.E. Lenski, J. Geiselmann, and D. Schneider. 2005. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics*, 169:523-532.

[58] Curran, J.F. and M. Yarus. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons *in vivo*. *J. Mol. Biol.*, 209:65-77.

[59] Danchin, A., P. Guerdoux-Jamet, I. Moszer, and P. Nitschké. 2000. Mapping the bacterial cell architecture into the chromosome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 355:179-190.

[60] Danchin, A. 2002. *The Delphic Boat: What Genomes Tell Us.* Harvard, Cambridge, MA.

[61] Dandekar, T., S. Schuster, B. Snel, M. Huynen, and P. Bork. 1999. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, 343:115-124.

[62] Davidson, E.H., *el al.* 2002. A genomic regulatory network for development. *Science*, 295:1669-1678.

[63] de Jong, H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9:67-103.

[64] Dennis, P.P., M. Ehrenberg, and H. Bremer. 2004. Control of rRNA synthesis in *Escherichia coli*: a systems biology approach. *Microbiol. Mol. Biol. Rev.*, 68:639-668.

[65] Dethlefsen, L. and T.M. Schmidt. 2005. Differences in codon bias cannot explain differences in translational power among microbes. *BMC Bioinformatics*, 6:3.

[66] Deuschle, U., W. Kammerer, R. Gentz, and H. Bujard. 1986. Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. *EMBO J.*, 5:2987-2994.

[67] Dittmar, K.A., E.M. Mobley, A.J. Radek, and T. Pan. 2004. Exploring the regulation of tRNA distribution on the genomic scale. *J. Mol. Biol.*, 337:31-47.

[68] Dong, H., L. Nilsson, and C.G. Kurland. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.*, 260:649-663.

[69] dos Reis, M., L. Wernisch, and R. Savva. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.*, 31:6976-6985.

[70] Drew, D.A. 2001. A mathematical model for prokaryotic protein synthesis. *Bull. Math. Biol.*, 63:329-351.

[71] Duarte, N.C., M.J. Herrgård, and B.O. Palsson. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.*, 14:1298-1309.

[72] Edelman, G.M. 1988. *Topobiology: An Introduction to Molecular Embryology.* BasicBooks, New York.

[73] Edwards, J.S. and B.O. Palsson. 1998. How will bioinformatics influence metabolic engineering? *Biotechnol. Bioeng.*, 58:162-169.

[74] Edwards, J.S., R. Ramarishna, C.H. Schilling, and B.O. Palsson. 1999. Metabolic flux balance analysis, in: *Metabolic Engineering*, eds. S.Y. Lee and E.T. Papoutsakis. Marcel Deker, New York.

[75] Edwards, J.S. and B.O. Palsson. 1999. Systems properties of the *Haemophilus infleunzae* Rd metabolic genotype. *J. Biol. Chem.*, 274:17410-17416.

[76] Edwards, J.S. and B.O. Palsson. 2000. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, 97:5528-5533.

[77] Edwards, J.S., R.U. Ibarra, and B.O. Palsson. 2001. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, 19:125:130.

[78] Edwards, J.S., R. Ramakrishna, and B.O. Palsson. 2002. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol. Bioeng.*, 77:27-36.

[79] Elf, J., D. Nilsson, T. Tenson, and M. Ehrenberg. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, 300:1718-1722.

[80] Ellis, R.J. 2001. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26:597-604.

[81] Elmore, S., M. Müller, N. Vischer, T. Odijk, and C.L. Woldringh. 2005. Single-particle tracking of *oriC*-GFP fluorescent spots during chromosome segregation in *Escherichia coli. J. Struct. Biol.*, 151:275-287.

[82] Elowitz, M.B., M.G. Surette, P.E. Wolf, J.B. Stock, and S. Leibler. 1999. Protein mobility in the cytoplasm of *Escherichia coli. J. Bacteriol.*, 181:197-203.

[83] Eisenberg, D., E.M. Marcotte, I. Xenartos, and T.O. Yeates. 2000. Protein function in the post-genomic era. *Nature*, 405:823-826.

[84] Errington, J., R.A. Daniel, and D.J. Scheffers. 2003. Cytokinesis in bacteria. *Microbiol. Mol. Biol. Rev.*, 67:52-65.

[85] Famili, I. and B.O. Palsson. 2003. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys J.*, 85:16-26.

[86] Famili, I. and B.O. Palsson. 2003. Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices. *J. Theor. Biol.*, 224:87-96.

[87] Feinberg, M. 1980. Chemical oscillation, multiple equilibria, and reaction network structure. In: *Dynamics and Modelling of Reactive Systems* (Stewart, W., W.H. Ray, and C. Conley, eds). Academic Press, New York.

[88] Fell, D.A. 1992. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J.*, 286:313-330.

[89] Fell, D.A. 1996. *Understanding the Control of Metabolism.* Portland Press, London.

[90] Fell, D.A. and A. Wagner. 2000. The small world of metabolism. *Nat. Biotechnol.*, 18:1121-1122.

[91] Fell, D.A. 2001. Beyond genomics. *Trends Genet.*, 17:680-682.

[92] Figge, R.M., A.V. Divakaruni, and J.W. Gober. 2004. MreB, the cell shape-determining bacterial actin homologue, co-ordinates cell wall morphogenesis in *Caulobacter crescentus. Mol. Microbiol.*, 51:1321-1332.

[93] Fleischmann, R.D., *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269:496-512.

[94] Fong, S.S., J.Y. Marciniak, and B.O. Palsson. 2003. Description and interpretation of adaptive evolution of *Escherichia coli* K12 MG1655 using a genome-scale *in silico* metabolic model. *J. Bacteriol.*, 185:6400-6408.

[95] Fong, S.S. and B.O. Palsson. 2004. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.*, 36:1056-1058.

[96] Fong, S.S., A.R. Joyce, and B.O. Palsson. 2005. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different expression states. *Genome Res.*, 15:1365-1372.

[97] Förster, J., I. Famili, P. Fu, B.O. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.*, 13:244-253.

[98] Fuglsang, A. 2003. Strong associations between gene function and codon usage. *APMIS*, 111:843-847.

[99] Fuglsang, A. 2003. The effective number of codons for individual amino acids: some codons are more optimal than others. *Gene*, 320:185-190.

[100] Fussenegger, M., J.E. Bailey, and J. Varner. 2000. A mathematical model of caspase function in apoptosis. *Nat. Biotechnol.*, 18:768-774.

[101] Gaal, T., M.S. Bartlett, W. Ross, C.L. Turnbough, Jr., and R.L. Gourse. 1997. Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science*, 278:2092-2097.

[102] Garner, E.C., C.S. Campbell, and R.D. Mullins. 2004. Dynamic instability in a DNA-segregating prokaryotic actin homolog. *Science*, 306:1021-1025.

[103] Ge, H., A.J. Walhout, and M. Vidal. 2003. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, 19:551-560.

[104] Gerdes, K., J. Moller-Jensen, G. Ebersbach, T. Kruse, and K. Nordstrom. 2004. Bacterial mitotic machineries. *Cell*, 116:359-366.

[105] Gerdes, S.Y., *et al.* 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, 185:5673-5684.

[106] Gitai, Z., N. Dye, and L. Shapiro. 2004. An actin-like gene can determine cell polarity in bacteria. *Proc. Natl. Acad. Sci. USA*, 101:8643-8648.

[107] Gitai, Z. 2005. The new bacterial cell biology: moving parts and subcellular architecture. *Cell*, 120:577-586.

[108] Glasner, J.D., P. Liss, G. Plunkett 3rd, A. Darling, R. Prasad, M. Rusch, A. Byrnes, M. Gilson, B. Biehl, F.R. Blattner, and N.T. Perna. 2003. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, 31:147-151.

[109] Goetz, R.M. and A. Fuglsang. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, 327:4-7.

[110] Gombert, A.K. and J. Nielsen. 2000. Mathematical modeling of metabolism. *Curr. Opin. Biotechnol.*, 11:180-186.

[111] Goodsell, D.S. 1993. *The Machinery of Life*. Springer-Verlag, New York.

[112] Gordon, G.S., D. Sitnikov, C.D. Webb, A. Teleman, A. Straight, R. Losick, A.W. Murray, and A. Wright. 1997. Chromosome and low copy plasmid segregation in *E. coli*: visual evidence for distinct mechanisms. *Cell*, 90:1113-1121.

[113] Gotta, S.L., O.L. Miller Jr., and S.L. French. 1991. Ribosomal RNA transcription rate in *Escherichia coli*. *J. Bacteriol.*, 173:6647-6649.

[114] Gottesman, S. and M. R. Maurizi. 1992. Regulation by proteolysis: energy-dependent proteases and their targets. *Microbiol. Rev.*, 56:592-621.

[115] Greenbaum, D., N.M. Luscombe, R. Jansen, J. Qian, and M. Gerstein. 2001. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.*, 11:1463-1468.

[116] Grosjean, H.J., S. de Henau, and D.M. Crothers. 1978. On the physical basis for ambiguity in genetic coding interactions. *Proc. Natl. Acad. Sci. USA*, 75:610-614.

[117] Grosjean, H. and W. Fiers. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18:199-209.

[118] Gustafsson, C., S. Govindarajan, and J. Minshull. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.*, 22:346-353.

[119] Gutiérrez, G., L. Márquez, and A. Marín. 1996. Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucleic Acids Res.*, 24:2525-2527.

[120] Hall, D.A., H. Zhu, X. Zhu, T. Royce, M. Gerstein, and M. Snyder. 2004. Regulation of gene expression by a metabolic enzyme. *Science*, 306:482-484.

[121] Hallin, P.F. and D.W. Ussery. 2004. CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics*, 20:3682-3686.

[122] Hartwell, L.H., J.J. Hopfield, S. Leibler, and A.W. Murray. 1999. From molecular to modular cell biology. *Nature*, 402:C47-C52.

[123] Hasty, J., J. Pradines, M. Dolnik, and J.J. Collins. 2000. Noise-based switches and amplifiers for gene expression. *Proc. Natl. Acad. Sci. USA*, 97:2075-2080.

[124] Hatfield, G.W., and C.J. Benham. 2002. DNA topology-mediated control of global gene expression in Escherichia coli. *Annu. Rev. Genet.*, 36:175-203.

[125] Hatzimanikatis, V. and K.H. Lee. 1999. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metab. Eng.*, 1:275-281.

[126] Heinrich, R. and T.A. Rapoport. 1980. Mathematical modelling of translation of mRNA in eucaryotes; steady state, time-dependent processes and application to reticulocytes. *J. Theor. Biol.*, 86:279-313.

[127] Heinrich, R. and S. Schuster. 1996. *The Regulation of Cellular Systems.* Chapman and Hall, New York.

[128] Herrgård, M.J., M.W. Covert, and B.O. Palsson. 2003. Reconciling gene expression data with known gnome-scale regulatory network structures. *Genome Res.*, 13:2423-2434.

[129] Herring, C.D., M. Raffaelle, T.E. Allen, E. Kanin, R. Landick, A.Z. Ansari, and B.O. Palsson. 2005. Immobilization of *Escherichia coli* RNA polymerase and location of binding sites using chromatin immunoprecipitation and microarrays. *J. Bacteriol.*, 187:6166-6174.

[130] Hoekema, A., R.A. Kastelein, M. Vasser, and H.A. de Boer. 1987. Codon replacement in the PGK1 gene of *Saccharomyces cerevisiae*: experimental approach to study the role of biased codon usage in gene expression. *Mol. Cell Biol.*, 7:2914-2924.

[131] Holstege, F.C.P., E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717-728.

[132] Holter, N.S., M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar, and N.V. Fedoroff. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA*, 97:8409-8414.

[133] Honisch, C., A. Raghunathan, C.R. Cantor, B.O. Palsson, and D. van den Boom. 2004. High-throughput mutation detection underlying adaptive evolution of *Escherichia coli*-K12. *Genome Res.*, 14:2495-2502.

[134] Holzhütter, H.G., G. Jacobasch, and A. Bisdorff. 1985. Mathematical modelling of metabolic pathways affected by an enzyme deficiency. A mathematical model of glycolysis in normal and pyruvate-kinase-deficient red blood cells. *Eur. J. Biochem.*, 149:101-111.

[135] Howard, M., A.D. Rutenberg, and S. de Vet. 2001. Dynamic compartmentalization of bacteria: accurate division in *E. coli. Phys. Rev. Lett.*, 87:278102.

[136] Hu, Z., and J. Lutkenhaus. 1999. Topological regulation of cell division in *Escherichia coli* involves rapid pole to pole oscillation of the division inhibitor MinC under the control of MinD and MinE. *Mol. Microbiol.*, 34:82-90.

[137] Hu, Z., and J. Lutkenhaus. 2000. Analysis of MinC reveals two independent domains involved in interaction with MinD and FtsZ. *J. Bacteriol.*, 182:3965-3971.

[138] Hu, Z., E.P. Gogol, and J. Lutkenhaus. 2002. Dynamic assembly of MinD on phospholipid vesicles regulated by ATP and MinE. *Proc. Natl. Acad. Sci. USA*, 99:6761-6766.

[139] Huang, S. 2000. The practical problems of post-genomic biology. *Nat. Biotechnol.*, 18:471-472.

[140] Hubbard, B.B. 1998. *The World According to Wavelets: The Story of a Mathematical Technique in the Making.* A K Peters, Natick, MA.

[141] Huh, W.K., J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, and E.K. O'Shea. 2003. Global analysis of protein localization in budding yeast. *Nature*, 425:686-691.

[142] Hynne, F., S. Dano, and P.G. Sorensen. 2001. Full-scale model of glycolysis in *Saccharomyces cerevisiae. Biophys. Chem.*, 94:121-163.

[143] Ibarra, R.U., J.S. Edwards, and B.O. Palsson. 2002. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*, 420:186-189.

[144] Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292:929-934.

[145] Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.*, 146:1-21.

[146] Ingber, D.E. 2003. Tensegrity I. Cell structure and hierarchical systems biology. *J. Cell Sci.*, 116:1157-1173.

[147] Ingber, D.E. 2003. Tensegrity II. How structural networks influence cellular information processing networks. *J. Cell Sci.*, 116:1397-1408.

[148] Iost, I. and M. Dreyfus. 1995. The stability of *Escherichia coli lacZ* mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J.*, 14:3252-3261.

[149] Isaacs, F.J., J. Hasty, C.R. Cantor, and J.J. Collins. 2003. Prediction and measurement of an autoregulatory genetic module. *Proc. Natl. Acad. Sci. USA*, 100:7714-7719.

[150] Iwakura, Y., K. Ito, and A. Ishihama. 1974. Biosynthesis of RNA polymerase in *Escherichia coli*. I. Control of RNA polymerase content at various growth rates. *Mol. Gen. Genet.*, 133:1-23.

[151] Jacob, F. and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins *J. Mol. Biol.*, 3:318-356.

[152] Jamshidi, N., J.S. Edwards, T. Fahland, G.M. Church, and B.O. Palsson. 2001. Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics*, 17:286-287.

[153] Jamshidi, N., S.J. Wiback, S.J., and B.O. Palsson. 2002. *In silico* model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism. *Genome Res.*, 12:1687-1692.

[154] Jeong, H., B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi. 2000. The large-scale organization of metabolic networks. *Nature*, 407:651-654,

[155] Jeong, H., S.P. Mason, A.L. Barabasi, and Z.N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature*, 411:41-42.

[156] Jeong, K.S., J. Ahn, and A.B. Khodursky. 2004. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.*, 5:R86.

[157] Jones, L.J., R. Carballido-Lopez, and J. Errington. 2001. Control of cell shape in bacteria: helical, actin-like filaments in *Bacillus subtilis*. *Cell*, 104:913-922.

[158] Jones, M.N. 1988. *Biochemical Thermodynamics (Studies in Modern Thermodynamics, No 8)*. Elsevier, New York.

[159] Joshi, A. and B.O. Palsson. 1989. Metabolic dynamics in the human red cell. Part I - A comprehensive kinetic model. *J. Theor. Biol.*, 141:515-528.

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

[161] Kajitani, M. and A. Ishihama. 1983. Determination of the promoter strength in the mixed transcription system. II. Promoters of ribosomal RNA, ribosomal protein S1 and recA protein operons from *Escherichia coli. Nucleic Acids Res.*, 11:3873-3888.

[162] Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238:143-155.

[163] Karp, P.D., M. Riley, S.M. Paley, and A. Pelligrini-Toole. 1996. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, 24:32-39.

[164] Karp, P.D. 2001. Pathway databases: a case study in computational symbolic theories. *Science*, 293:2040-2044.

[165] Kauffman, K.J., P. Prakash. and J.S. Edwards. 2003. Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, 14:491-496.

[166] Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson. 2001. A gene expression map for *Caenorhabditis elegans. Science*, 293:2087-2092.

[167] Kim, J., S.H. Yoshimura, K. Hizume, R.L. Ohniwa, A. Ishihama, and K. Takeyasu. 2004. Fundamental structural units of the *Escherichia coli* nucleoid revealed by atomic force microscopy. *Nucleic Acids Res.* 32:1982-1992.

[168] Khorana, H.G. 1965. Polynucleotide synthesis and the genetic code. *Fed. Proc.*, 24:1473-1487.

[169] Klamt, S., J. Stelling, M. Ginkel, E.D. Gilles. 2003. FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, 19:261-269.

[170] Komine, Y., T. Adachi, H. Inokuchi, and H. Ozeki. 1990. Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. *J. Mol. Biol.*, 212:579-598.

[171] Kushner, S.R. 1996. mRNA decay. In: Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F.C., *et al.*, eds). ASM Press, Washington.

[172] Lackner, L.L., D.M. Raskin, and P.A. de Boer. 2003. ATP-dependent interactions between *Escherichia coli* Min proteins and the phospholipid membrane in vitro. *J. Bacteriol.*, 185:735-749.

[173] Laffend, L. and M.L. Shuler. 1994. Ribosomal protein limitations in *Escherichia coli* under conditions of high translational activity. *Biotechnol. Bioeng.*, 43:388-398.

[174] Lander, E.S., *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature*, 409:860-921.

[175] Lazebnik, Y. 2002. Can a biologist fix a radio? - Or, what I learned while studying apoptosis. *Cancer Cell*, 2:179-182.

[176] Lee, I.-D. and B.O. Palsson. 1991. A Comprehensive model of human erythrocyte metabolism: extensions to include pH effects. *Biomed. Biochim. Acta*, 49:771-789.

[177] Lee, S., C. Phalakornkule, M.M. Domach, and I.E. Grossmann. 2000. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng.*, 24:711-716.

[178] Lenski, R.E., C.L. Winkworth, and M.A. Riley. 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. *J. Mol. Evol.*, 56:498-508.

[179] Liao, J.C. 1993. Modelling and analysis of metabolic pathways. *Curr. Opin. Biotechnol.*, 4:211-216.

[180] Liao, J.C., S.Y. Hou, and Y.P. Chao. 1996. Pathway analysis, engineering and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.*, 52:129-140.

[181] Liljenstr om, H. and G. von Heijne. 1987. Translation rate modification by preferential codon usage: intragenic position effects. *J. Theor. Biol.*, 124:43-55.

[182] Liò, P. 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19:2-9.

[183] Lithwick, G. and H. Margalit. 2003. Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res.*, 13:2665-2673.

[184] Løbner-Olesen, A., M.G. Marinus, and F.G. Hansen. 2003. Role of SeqA and Dam in *Escherichia coli* gene expression: A global/microarray analysis. *Proc. Natl. Acad. Sci. USA*, 100:4672-4677.

[185] Loew, L.M. and J.C. Schaff. 2001. The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol.*, 19:401-406.

[186] MacDonald, C.T., J.H. Gibbs, and A.C. Pipkin. 1968. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers*, 6:1-5.

[187] Martin, V.J., D.J. Pitera, S.T. Withers, J.D. Newman, and J.D. Keasling. 2003. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.*, 21:796-802.

[188] Mathews, C.K. and K.E. van Holde. 1996. *Biochemistry.* Benjamin/Cummings, Menlo Park, CA.

[189] McAdams, H.H. and L. Shapiro. 1995. Circuit simulation of genetic networks. *Science*, 269:651-656.

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

[191] McAdams, H.H. and L. Shapiro. 2003. A bacterial cell-cycle regulatory network operating in time and space. *Science*, 301:1874-1877.

[192] Michaelis, L. and M. Menten. 1913. Die Kinetik der Invertinwirkung. *Biochim. Z.*, 49:333-369.

[193] Michal, G. 1999. *Biochemical pathways: an atlas of biochemistry and molecular biology.* Wiley, New York.

[194] Miller, C.G. 1996. Protein degradation and proteolytic modification. In: Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F.C., *et al.*, eds). ASM Press, Washington.

[195] Moller-Jensen, J., J. Borch, M. Dam, R.B. Jensen, P. Roepstorff, and K. Gerdes. 2003. Bacterial mitosis: ParM of plasmid R1 moves plasmid DNA by an actin-like insertional polymerization mechanism. *Mol. Cell*, 12:1477-1487.

[196] Morton-Firth, C.J. and D. Bray. 1998. Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.*, 192:117-128.

[197] Mulquiney, P.J. and P.W. Kuchel. 1999. Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: computer simulation and metabolic control analysis. *Biochem. J.*, 342:597-604.

[198] Murray, K.B., D. Gorse, and J.M. Thornton. 2002. Wavelet transform for the characterization and detection of repeating motifs. *J. Mol. Biol.*, 316:341-363.

[199] Najafabadi, H.S., H. Goodarzi, and N. Torabi. 2005. Optimality of codon usage in *Escherichia coli* due to load minimization. *J. Theor. Biol.*, 237:203-209.

[200] Neidhardt, F.C., J.L. Ingraham, and M. Schaechter. 1990. *Physiology of the Bacterial Cell.* Sinauer, Sunderland, MA.

[201] Neidhardt, F.C., *et al.*, eds. 1996. Escherichia coli *and* Salmonella: *Cellular and Molecular Biology.* ASM Press, Washington.

[202] Nicolay, S., F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo. 2004. Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? *Phys. Rev. Lett.*, 93:108101.

[203] Nierlich, D.P. 1978. Regulation of bacterial growth, RNA, and protein synthesis. *Annu. Rev. Microbiol.*, 32:393-432.

[204] Niki, H., A. Jaffe, R. Imamura, T. Ogura, and S. Hiraga. The new gene *mukB* codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of *E. coli. EMBO J.*, 10:183-193.

[205] Niki, H. and S. Hiraga. 1997. Subcellular distribution of actively partitioning F plasmid during the cell division cycle in *E. coli. Cell*, 90:951-957.

[206] Niki, H., Y. Yamaichi, and S. Hiraga. 2000. Dynamic organization of chromosomal DNA in *Escherichia coli. Genes Dev.*, 14:212-223.

[207] Nirenberg, M.W. and J.H. Matthaei. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA*, 47:1588-1602.

[208] Novak, B., A. Toth, A. Csikasz-Nagy, B. Gyorffy, J.J. Tyson, and K. Nasmyth. 1999. Finishing the cell cycle. *J. Theor. Biol.*, 199:223-233.

[209] Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 27:29-34.

[210] Oh, M.K., L. Rohlin, K.C. Kao, and J.C. Liao. 2002. Global expression profiling of acetate-grown *Escherichia coli. J. Biol. Chem.*, 277:13175-13183.

[211] Orešič, M. and D. Shalloway. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.*, 281:31-48.

[212] Ouzounis, C.A. and P.D. Karp. 2000. Global properties of the metabolic map of *Escherichia coli. Genome Res.*, 10:568-576.

[213] Overbeek, R., N. Larsen, G.D. Pusch, M. D'Souza, E. Selkov Jr., N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, 28:123-125.

[214] Palsson, B.O., A. Joshi, and S.S. Ozturk, 1987. Reducing complexity in metabolic networks: making metabolic meshes manageable. *Fed. Proc.*, 46:2485-2489.

[215] Palsson, B.O. and I.D. Lee. 1993. Model complexity has a significant effect on the numerical value and interpretation of metabolic sensitivity coefficients. *J. Theor. Biol.*, 161:299-315.

[216] Palsson, B.O. 1997. What lies beyond bioinformatics? *Nat. Biotechnol.*, 15:3-4.

[217] Palsson, B.O. 2000. The challenges of *in silico* biology. *Nat. Biotechnol.*, 18:1147-1150.

[218] Palsson, B.O. 2002. *In silico* biology through "omics". *Nat. Biotechnol.*, 20:649-650.

[219] Palsson, B.O. 2004. Two-dimensional annotation of genomes. *Nat. Biotechnol.*, 22:1218-1219.

[220] Papin, J.A., N.D. Price, J.S. Edwards, and B.O. Palsson. 2002. The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J. Theor. Biol.*, 215:67-82.

[221] Papin, J.A., N.D. Price, and B.O. Palsson. 2002. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res.*, 12:1889-1900.

[222] Papin, J.A., N.D. Price, S.J. Wiback, D.A. Fell, and B.O. Palsson. 2003. Metabolic pathways in the post-genome era. *Trends Biochem. Sci.*, 28:250-258.

[223] Papin, J.A. and B.O. Palsson. 2004. The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.*, 87:37-46.

[224] Papin, J.A., J. Stelling, N.D. Price, S. Klamt, S. Schuster, and B.O. Palsson. 2004. Comparison of network-based pathway analysis methods. *Trends Biotechnol.*, 22:400-405.

[225] Pedersen, S. 1984. *Escherichia coli* ribosomes translate *in vivo* with variable rate. *EMBO J.*, 3:2895-2898.

[226] Pedersen, A.G., L.J. Jensen, S. Brunak, H.-H. Stæfeldt, and D.W. Ussery. 2000. A DNA structural atlas for *Escherichia coli. J. Mol. Biol.*, 299:907-930.

[227] Peretti, S.W. and J.E. Bailey. 1986. Mechanistically detailed model of cellular metabolism for glucose-limited growth of *Escherichia coli* B/r-A. *Biotechnol. Bioeng.*, 28:1672-1689.

[228] Perez-Rueda, E. and J. Collado-Vides. 2000. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.*, 28:1838-1847.

[229] Peter, B.J., J. Arsuaga, A.M. Breier, A.B. Khodursky, P.O. Brown, and N.R. Cozzarelli. 2004. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli. Genome Biol.*, 5:R87.

[230] Pieper, D.H. and W. Reineke. 2000. Engineering bacteria for bioremediation. *Curr. Opin. Biotechnol.*, 11:262-270.

[231] Postow, L., C.D. Hardy, J. Arsuaga, and N.R. Cozzarelli. 2004. Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev.*, 18:1766-1779.

[232] Price, N.D., J.A. Papin, and B.O. Palsson. 2002. Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.*, 12:760-769.

[233] Price, N.D., I. Famili, D.A. Beard, and B.O. Palsson. 2002. Extreme pathways and Kirchhoff's second law. *Biophys. J.*, 83:2879-2882.

[234] Price, N.D., J.L. Reed, J.A. Papin, I. Famili, and B.O. Palsson. 2003. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys. J.*, 84:794-804.

[235] Price, N.D., J.A. Papin, C.H. Schilling, and B.O. Palsson. 2003. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol.*, 21:162-169.

[236] Price, N.D., J.L. Reed, and B.O. Palsson. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, 2:886-897.

[237] Ptashne, M. and A. Gann. 2002. *Genes & Signals.* Cold Spring Harbor Laboratory Press, New York.

[238] Rae, C., S.J. Berners-Price, B.T. Bulliman, and P.W. Kuchel. 1990. Kinetic analysis of the human erythrocyte glyoxalase system using 1H NMR and a computer model. *Eur. J. Biochem.*, 193:83-90.

[239] Raghunathan, A. and B.O. Palsson. 2003. Scalable method to determine mutations that occur during adaptive evolution of *Escherichia coli. Biotechnol. Lett.*, 25:435-441.

[240] Raghunathan, A., B.-S. Lee, S. Toyama, T.E. Allen, S.S. Fong, A.R. Joyce, and B.O. Palsson. Characteristic chromosomal rearrangements during adaptive evolution of *Escherichia coli* on different carbon sources. In preparation.

[241] Rajagopalan, H., M.A. Nowak, B. Vogelstein, and C. Lengauer. 2003. The significance of unstable chromosomes in colorectal cancer. *Nat. Rev. Cancer*, 3:695-701.

[242] Raskin, D.M., and P.A. de Boer. 1999. Rapid pole-to-pole oscillation of a protein required for directing division to the middle of *Escherichia coli. Proc. Natl. Acad. Sci. USA*, 96:4971-4976.

[243] Ravasz, E., A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabasi AL. 2002. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551-1555.

[244] Record Jr., M.T., W.S. Reznikoff, M.L. Craig, K.L. McQuade, and P.J. Schlax. 1996. *Escherichia coli* RNA polymerase (Eσ^{70}), promoters, and the kinetics of the steps of transcription initiation. In: Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F.C., *et al.*, eds). ASM Press, Washington.

[245] Reed, J.L. and B.O. Palsson. 2003. Thirteen years of building constraint-based *in silico* models of *Escherichia coli. J. Bacteriol.*, 185:2692-2699.

[246] Reed, J.L., T.D. Vo, C.H. Schilling, and B.O. Palsson. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, 4:R54.

[247] Reed, J.L. and B.O. Palsson. 2004. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.*, 14:1797-1805.

[248] Reed, J.L., I. Famili, I. Thiele, and B.O. Palsson. 2005. Towards multidimensional genome annotation. *Nat. Rev. Genet.*, in press.

[249] Reich, J.G. and E.E. Sel'kov. 1975. Time hierarchy, equilibrium and non-equilibrium in metabolic systems. *Biosystems*, 7:39-50.

[250] Reich, J.G. and E. Selkov. 1981. *Energy Metabolism of the Cell: A Theoretical Treatise.* Academic Press, London.

[251] Relógio, A., C. Schwager, A. Richter, W. Ansorge, and J. Valcárel. 2002. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Aicds Res.*, 30:e51.

[252] Richmond, C.S., J.D. Glasner, R. Mau, H. Jin, and F.R. Blattner. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.*, 27:3821-3835.

[253] Robbins, J.R., D. Monack, S.J. McCallum, A. Vegas, E. Pham, M.B. Goldberg, and J.A. Theriot. 2001. The making of a gradient: IcsA (VirG) polarity in *Shigella flexneri. Mol. Microbiol.*, 41:861-872.

[254] Robinett, C.C., A. Straight, G. Li, C. Willhelm, G. Sudlow, A. Murray, and A.S. Belmont. 1996. In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition. *J. Cell Biol.*, 135:1685-1700.

[255] Rocha, E.P.C., P. Guerdoux-Jamet, I. Moszer, A. Viari, and A. Danchin. 2000. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotechnol.*, 78:209-219.

[256] Rocha, E.P.C. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, 14:2279-2286.

[257] Rosenow, C., R.M. Saxena, M. Durst, and T.R. Gingeras. 2001. Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic Acids Res.*, 29:e112.

[258] Sabatti, C., L. Rohlin, M.K. Oh, and J.C. Liao. 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, 30:2886-2893.

[259] Saier Jr., M.H. 1995. Differential codon usage: a safeguard against inappropriate expression of specialized genes? *FEBS Lett.*, 362:1-4.

[260] Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millán-Zárate, E. Díaz-Peredo, F. Sánchez-Solano, E. Pérez-Rueda, C. Bonavides-Martínez, and J. Collado-Vides. 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, 29:72-74.

[261] Santillán, M. and M.C. Mackey. 2001. Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data. *Proc. Natl. Acad. Sci. USA*, 98:1364-1369.

[262] Sauer, U., D.C. Cameron, and J.E. Bailey. 1998. Metabolic capacity of *Bacillus subtilis* for the production of purine nucleosides, riboflavin, and folic acid. *Biotechnol. Bioeng.*, 59:227-238.

[263] Sauer, U. and J.E. Bailey. 1999. Estimation of P-to-O ratio in *Bacillus subtilis* and its influence on maximum riboflavin yield. *Biotechnol. Bioeng.*, 64:750-754.

[264] Sauer, U. 2001. Evolutionary engineering of industrially important microbial phenotypes. *Adv. Biochem. Eng. Biotechnol.*, 73:129-169.

[265] Savageau, M.A., E.O. Voit, and D.H. Irvine. 1987. Biochemical systems theory and metabolic control theory: I. Fundamental similarities and differences. *Mathematical Biosciences*, 86:127-145.

[266] Savageau, M.A., E.O. Voit, and D.H. Irvine. 1987. Biochemical systems theory and metabolic control theory: II. The role of summation and connectivity relationships. *Mathematical Biosciences*, 86:147-169.

[267] Sawitzke, J.A. and S. Austin. 2000. Suppression of chromosome segregation defects of *Escherichia coli muk* mutants by mutations in topoisomerase I. *Proc. Natl. Acad. Sci. USA*, 97:1671-1676.

[268] Schilling, C.H., J.S. Edwards, and B.O. Palsson. 1999. Towards metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol. Prog.*, 15:288-295.

[269] Schilling, C.H., S. Schuster, B.O. Palsson, and R. Heinrich. 1999. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 15:296-303.

[270] Schilling, C.H., D. Letscher, and B.O. Palsson. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203:229-248.

[271] Schilling, C.H. and B.O. Palsson. 2000. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, 203:249-283.

[272] Schilling, C.H., J.S. Edwards, D. Letscher, and B.O. Palsson. 2000. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.*, 71:286-306.

[273] Schilling, C.H., M.W. Covert, I. Famili, G.M. Church, J.S. Edwards, and B.O. Palsson. 2002. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.*, 184:4582-4593.

[274] Schoeberl, B., C. Eichler-Jonsson, E.D. Gilles, and G. Muller. 2002. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.*, 20:370-375.

[275] Schuster, R., H.G. Holzhütter, and G. Jacobasch. 1988. Interrelations between glycolysis and the hexose monophosphate shunt in erythrocytes as studied on the basis of a mathematical model. *Biosystems*, 22:19-36.

[276] Schuster, S. and C. Hilgetag. 1994. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2:165-182.

[277] Segall, A., M.J. Mahan, and J.R. Roth. 1988. Rearrangement of the bacterial chromosome: forbidden inversions. *Science*, 241:1314-1318.

[278] Selinger, D.W., K.J. Cheung, R. Mei, E.M. Johansson, C.S. Richmond, F.R. Blattner, D.J. Lockhart, and G.M. Church. 2000. RNA expression analysis

using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.*, 18:1262-1268.

[279] Selinger, D.W., R.M. Saxena, K.J. Cheung, G.M. Church, and C. Rosenow. 2003. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.*, 13:216-223.

[280] Selkov, E., Jr., Y. Grechkin, N. Mikhailova, and E. Selkov. 1998. MPW: the Metabolic Pathways Database. *Nucleic Acids Res.*, 26:43-45.

[281] Serres, M.H. and M. Riley. 2000. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, 5:205-222.

[282] Sharp, P.M. and W.-H. Li. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15:1281-1295.

[283] Sharp, P.M., E. Bailes, R.J. Grocock, J.F. Peden, and R.E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, 33:1141-1153.

[284] Sherratt, D.J. 2003. Bacterial chromosome dynamics. *Science*, 301:780-785.

[285] Shi, Y. and Y. Shi. 2004. Metabolic enzymes and coenzymes in transcription– a direct link between metabolism and transcription? *Trends Genet.*, 20:445-452.

[286] Shuler, M.L., S. Leung, and C.C. Dick. 1979. A mathematical model for the growth of a single bacterial cell. *Ann. N.Y. Acad. Sci.*, 326:35.

[287] Shuler, M.L. and M.M. Domach. 1983. Mathematical models of the growth of individual cells. In: *Foundations of Biochemical Engineering* (Blanch, H.W., E.T. Papoutsakis, and G. Stephanopoulos, eds). American Chemical Society, Washington.

[288] Smith, H.O., C.A. Hutchison III, C. Pfannkoch, and J.C. Venter. 2003. Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. USA*, 100:15440-15445.

[289] Sinden, R.R. and D.E. Pettijohn. 1981. Chromosomes in living *Escherichia coli* growing on minimal and rich media. *Proc. Natl. Acad. Sci. USA*, 78:224-228.

[290] Sinha, S. 1988. Theoretical study of the tryptophan operon: application in microbial technology. *Biotechnol. Bioeng.*, 31:117-124.

[291] Solomovici, J., T. Lesnik, and C. Reiss. 1997. Does *Escherichia coli* optimize the economics of the translation process? *J. Theor. Biol.*, 185:511-521.

[292] Sørensen, M.A., C.G. Kurland, and S. Pedersen. 1989. Codon usage determines translation rate in *Escherichia coli. J. Mol. Biol.*, 207:365-377.

[293] Sørensen, M.A. and S. Pedersen. 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.*, 222:265-280.

[294] Stephanopoulos, G.N., A.A. Aristidou, and J. Nielsen. 1998. *Metabolic Engineering: Principles and Methodologies*. Academic Press, San Diego.

[295] Strothman, R.C. 1997. The coming Kuhnian Revolution in biology. *Nat. Biotechnol.*, 15:194-199.

[296] Sundararaj, S., A. Guo, B. Habibi-Nazhad, M. Rouani, P. Stothard, M. Ellison, and D.S. Wishart. 2004. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli. Nucleic Acids Res.*, 32:D293-295.

[297] Tao, H., C. Bausch, C. Richmond, F.R. Blattner, and T. Conway. 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.*, 181:6425-6440.

[298] Thanbichler, M., P.H. Viollier, and L. Shapiro. 2005. The structure and function of the bacterial chromosome. *Curr. Opin. Genet. Dev.*, 15:153-162.

[299] Theobald, U., W. Mailinger, M. Baltes, M. Rizzi, and M. Reuss. 1997. In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: I. Experimental observations. *Biotechnol. Bioeng.*, 55:305-316.

[300] Thomas, L.K., D.B. Dix, and R.C. Thompson. 1988. Codon choice and gene expression: synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes *in vitro*. *Proc. Natl. Acad. Sci. USA*, 85:4242-4246.

[301] Thomas, R. 1991. Regulatory networks seen as asynchronous automata: a logical description. *J. Theor. Biol.*, 153:1-23.

[302] Tillier, E.R. and R.A. Collins. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, 50:249-257.

[303] Tjaden, B., R.M. Saxena, S. Stolyar, D.R. Haynor, E. Kolker, and C. Rosenow. 2002. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res.*, 30:3732-3738.

[304] Tomita, M., K. Hashimoto, K. Takahashi, T.S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tamida, K. Yugi, J.C. Venter, and C.A. Hutchison III. 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15:72-84.

[305] Torrence, C. and G.P. Compo. 1998. A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.* 79:61-78.

[306] Travers, A. and G. Muskhelishvili. 2005. DNA supercoiling – a global transcriptional regulator for enterobacterial growth? *Nat. Rev. Microbiol.*, 3:157-169.

[307] Upton, G. and I. Cook. 2002. *A Dictionary of Statistics*. Oxford, New York.

[308] Ussery. D., T.S. Larsen, K.T. Wilkes, C. Friis, P. Worning, A. Krogh, and S. Brunak. 2001. Genome organisation and chromatin structure in *Escherichia coli. Biochimie*, 83:201-212.

[309] Valens, M., S. Penaud, M. Rossignol, F. Cornet, and F. Boccard. 2004. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.*, 23:4330-4341.

[310] van den Berg, B., R.J. Ellis, and C.M. Dobson. 1999. Effects of macromolecular crowding on protein folding and aggregation. *EMBO J.*, 18:6927-6933.

[311] van Gulik, W.M. and J.J. Heijnen. 1995. Metabolic network stoichiometry analysis of microbial growth and product formation. *Biotechnol. Bioeng.*, 48:681-698.

[312] van Riel, N.A., M.L. Giuseppin, and C.T. Verrips. 2000. Dynamic optimal control of homeostasis: an integrative system approach for modeling of the central nitrogen metabolism in *Saccharomyces cerevisiae. Metab. Eng.*, 2:49-68.

[313] Varenne, S., J. Buc, R. Lloures, and C. Ladzunski. 1984. Translation is a non-uniform process: Effect of tRNA availability on the rate of elongation of the nascent polypeptide chains. *J. Mol. Biol.*, 180:549-576.

[314] Varma, A. and B.O. Palsson. 1993. Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *J. Theor. Biol.*, 165:503-522.

[315] Varma, A. and B.O. Palsson. 1994. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology*, 12:994-998.

[316] Varma, A. and B.O. Palsson. 1994. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.*, 60:3724-3731.

[317] Varner, J. and D. Ramkrishna. 1999. Metabolic engineering from a cybernetic perspective. 1. Theoretical preliminaries. *Biotechnol. Prog.*, 15:407-425.

[318] Vaseghi, S., A. Baumeister, M. Rizzi, and M. Reuss. 1999. *In vivo* dynamics of the pentose phosphate pathway in *Saccharomyces cerevisiae. Metab. Eng.*, 1:128-140.

[319] Venter, JC, *et al.* 2001. The sequence of the human genome. *Science*, 291:1304-1351.

[320] Verkman, A.S. 2002. Solute and macromolecule diffusion in cellular aqueous compartments. *Trends Biochem. Sci.*, 27:27-33.

[321] Vind, J., M.A. Sørensen, M.D. Rasmussen, and S. Pedersen. 1993. Synthesis of proteins in *Escherichia coli* is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. *J. Mol. Biol.*, 231:678-688.

[322] Viollier, P.H., M. Thanbichler, P.T. McGrath, L. West, M. Meewan, H.H. McAdams, and L. Shapiro. 2004. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc. Natl. Acad. Sci. USA*, 101:9257-9262.

[323] Vogel, U. and K.F. Jensen. 1994. The RNA chain elongation rate in *Escherichia coli* depends on growth rate. *J. Bacteriol.*, 176:2807-2813.

[324] von Dassow, G., E. Meir, E.M. Munro, and G.M. Odell. 2000. The segment polarity network is a robust developmental module. *Nature*, 406:188-192.

[325] Wagner, R. 2000. *Transcription Regulation in Prokaryotes.* Oxford, New York.

[326] Wang, R., J.T. Prince, and E.M. Marcotte. 2005. Mass spectrometry of the *M. smegmatis* proteome: Protein expression levels correlate with function, operons, and codon bias. *Genome Res.*, 15:1118-1126.

[327] Wanner, B.L. 1996. Phosphorous assimilation and control of the phosphate regulon. In: Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F.C., *et al.*, eds). ASM Press, Washington.

[328] Watson, J.D. and F.H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737-738.

[329] Watts, D.J. 1999. *Small worlds: the dynamics of networks between order and randomness.* Princeton University Press, Princeton, N.J.

[330] Webb, C.D., A. Teleman, S. Gordon, A. Straight, A. Belmont, D. C.-H. Lin, A.D. Grossman, A. Wright, and R. Losick. 1997. Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of *B. subtilis. Cell*, 88:667-674.

[331] Wei, Y., J.-M. Lee, C. Richmond, F.R. Blattner, J.A. Rafalski, and R.A. La Rossa. 2001. High-density microarray-mediated gene expression profiling of *Escherichia coli*. *J. Bacteriol.*, 183:545-556.

[332] Weitao, T., K. Nordstrom, and S. Dasgupta. 1999. Mutual suppression of *mukB* and *seqA* phenotypes might arise from their opposing influences on the *Escherichia coli* nucleoid structure. *Mol. Microbiol.*, 34:157-168.

[333] Weller, K. and R.D. Recknagel. 1994. Promoter strength prediction based on occurrence frequencies of consensus patterns. *J. Theor. Biol.*, 171:355-359.

[334] Westerhoff, H.V. and B.O. Palsson. 2004. The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, 22:1249-1252.

[335] Wiback, S.J. and B.O. Palsson. 2002. Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.*, 83:808-818.

[336] Williams, H.P. 1999. *Model Building in Mathematical Programming*, 4th ed. Wiley, New York.

[337] Woldringh, C.L. and T. Odijk. 1999. Structure of DNA within the bacterial cell: physics and physiology. In: *Organization of the Prokaryotic Genome* (Charlebois, R.L., ed). ASM Press, Washington.

[338] Woldringh, C.L. 2002. The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol. Microbiol.*, 45:17-29.

[339] Wong, P., S. Gladney, and J.D. Keasling. 1997. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on flucose and lactose. *Biotechnol. Prog.*, 13:132-143.

[340] Wu, L.J. and J. Errington. 1998. Use of asymmetric cell division and spoI-IIE mutants to probe chromosome orientation and organization in *Bacillus subtilis*. *Mol. Microbiol.*, 27:777-786.

[341] Wu, X., H. Jörnvall, K.D. Berndt, and U. Oppermann. 2004. Codon optimization of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance. *Biochem. Biophys. Res. Commun.*, 313:89-96.

[342] Yang, Y. and G.F. Ames. 1988. DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. *Proc. Natl. Acad. Sci. USA*, 85:8850-8854.

[343] Yarmush, M. and F. Berthiaume. 1997. Metabolic engineering and human disease. *Nat. Biotechnol.*, 15:525-528.

[344] Zacharias, M., G. Theissen, C. Bradaczek, and R. Wagner. 1991. Analysis of sequence elements important for the synthesis and control of ribosomal RNA in *E. coli*. *Biochimie*, 73:699-712.

[345] Zaslaver, A., A.E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M.G. Surette, and U. Alon. 2004. Just-in-time transcription program in metabolic pathways. *Nat. Genet.*, 36:486-491.

[346] Zechiedrich, E.L., A.B. Khodursky, and N.R. Cozzarelli. 1997. Topoisomerase IV, not gyrase, decatenates products of site-specific recombination in *Escherichia coli*. *Genes Dev.*, 11:2580-2592.

[347] Zheng, M., X. Wang, L.J. Templeton, D.R. Smulski, R.A. LaRossa, and G. Storz. 2001. DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. *J. Bacteriol.*, 183:4562-4570.

[348] Zimmerman, S.B. and S.O. Trach. 1991. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*. *J. Mol. Biol.*, 222:599-620.

[349] Zimmerman, S.B. 2003. Underlying regularity in the shapes of nucleoids of *Escherichia coli*: implications for nucleoid organization and partition. *J. Struct. Biol.*, 142:256-265.