ABSTRACT

PONNALA, LALIT Analysis of Genetic Translation using Signal Processing. (Under the direction of Dr. D. L. Bitzer, Dr. M. A. Vouk and Dr. A. Stomp).

A series of free energy estimates can be calculated from the ribosome's progressive interaction with mRNA sequences during the process of translation elongation in eubacteria. A sinusoidal pattern of roughly constant phase has been detected in these free energy signals. Frameshifts of the +1 type occur when the ribosome skips an mRNA base in the 5'-3' direction, and can be associated with local phase-shifts in the free energy signal. We propose a mathematical model that captures the mechanism of frameshift based on the information content of the signal parameters and the relative abundance of tRNA in the bacterial cell. The model shows how translational speed can modulate translational accuracy to accomplish programmed +1 frameshifts and could have implications for the regulation of translational efficiency. Results are presented using experimentally verified frameshift genes across eubacteria.

Analysis of Genetic Translation using Signal Processing

by

Lalit Ponnala

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, NC

2007

Approved By:

Dr. Jeffrey L. Thorne	Dr. Tiffany M. Barnes
Dr. Anne-Marie Stomp Co-chair of Advisory Committee	Dr. Winser E. Alexander
Dr. Donald L. Bitzer Co-Chair of Advisory Committee	Dr. Mladen A. Vouk Co-Chair of Advisory Committee

To my parents and sister

Biography

Lalit Ponnala was born in the South Indian city of Hyderabad, where he spent 17 years of his early life. He then moved to the West coast of India in pursuit of a degree, and graduated from the National Institute of Technology, Surathkal with a bachelors in Electronics and Communication Engineering in 2001. Since then, he has been a graduate student in the Department of Electrical and Computer Engineering at North Carolina State University, Raleigh. While at NC State, he developed a keen interest in research and hopes to be able to pursue his interest in statistical signal processing for the rest of his life.

Acknowledgements

On a wintry November afternoon in 2002, I met an interesting person, Dr. Donald L Bitzer. Over the next four years, this person assumed many roles in my life - teacher, mentor, guide, philosopher, caretaker, guardian, friend and most importantly (lucky me!) my academic advisor. I would never be the person I am today if it weren't for Dr. Bitzer's constant support and confidence in my abilities.

Being cast into the web of inter-disciplinary research, about which I had no clue (and plenty of resentment), Dr. Bitzer's able guidance paved the way ahead and released me from mental roadblocks. I will be forever grateful to him for his inquisitive demeanor and infectious enthusiasm. As for Dr. Anne Stomp, her contributions to my work and building my personality can never be over-estimated. She has been a constant source of emotional support and I am greatly indebted to her for solving my administrative problems. This thesis would truly have never happened if it weren't for her.

To Dr. Alexander and Dr. Vouk, I owe special thanks for introducing me to this research area. I am indebted to Dr. Thorne and Dr. Barnes for their insightful discussions on my research topic.

NC State University has introduced me to some truly amazing teachers who instilled in me a deep sense of respect for imparting knowledge. In no particular order, I convey my gratitude to them:

- Dr. John Monahan, one of the most organized instructors, conveys material with crystal clarity and fields tricky questions with a sense of wit and humor. I thoroughly enjoyed taking Linear Models and Time Series Analysis with him.
- Dr. Samuel T. Alexander, the DSP guru, who's lecture on "DFT as a measure of correlation" remains etched in my mind to this day.
- Dr. Paul Franzon, for his sheer personality and charisma in class.
- Dr. Brian L. Hughes, whose class is worth attending just to hear him lecture.
- Dr. Roger Woodard, who laid solid foundations in elementary statistical analysis, paving the way for my minor in Statistics.
- Dr. Paul Fackler, whose firm grasp of both theory and applications leaves a lasting impact.

Serendipity plays an important role in anyone's life, and I have been lucky to have experienced some truly amazing interventions. In decreasing order of significance, I state them below:

- My internship: just at the point of my graduate life when I reached the limits of tolerance, I was lucky to land a summer internship in a wonderful company located just by the beach in California. The 3 months I spent there offered me much-needed respite from the rigors of academic work. Incidentally, I did get a lot of my writing done too!

- Exposure to another inter-disciplinary field (quantitative finance) and one of its foremost experts, Dr. Emanuel Derman. His writings brought out an appreciation and cultivation of good writing style, just at the time when I had almost given up hopes on my writing abilities. His explanation of model-building and how intuition meets practice is priceless.

- I have been really lucky to have access to Dr. Suvajit Samanta's mental abilities and guidance when the going got tough. I am grateful for his patient tolerance of my trivial questions at times. My graduate life at NCSU would have been utterly boring and insipid without him being around.

- Quentin Tarantino, master film-maker and story-teller par excellence, whose *Pulp Fiction* remains the one film that has made the most difference to my life.

To all these occurences and the higher powers that made them happen, I am immensely grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The complexity of living organisms makes them information-rich systems. As such, many processes are available for the application of signal processing analysis to reveal underlying mechanisms of information encoding and decoding. The mathematical methods of signal processing are well established and are used to extract encoded information from energetic patterns. These methods yield estimates of parameters that characterize the signal. Examples of the most basic parameters include frequency, phase and magnitude. Through study of system response to signal parameter change, the information content of signal parameters can be identified and the encoding and decoding rules can be defined. The application of signal processing analysis to a biological process requires the identification of a signal that could arise followed by characterization of signal parameters that correlate with process behavior.

It is well established that nucleic acid molecules, i.e. DNA and RNA, encode information in their nucleotide sequences that is essential to a number of cellular processes. Therefore, it is reasonable to use a signal processing approach to further our understanding of the rules and mechanisms of information encoding and decoding. The process of protein synthesis, or translation, is the most-studied biological process in which information encoded in the nucleotide sequence of mRNA is decoded into the correct sequence of amino acids in a polypeptide. Nucleic acids are long polymers of four nucleotide bases: adenine (A), guanine (G), cytosine (C) and thymidine (T, DNA) or uracil (U, mRNA). The chemical structure of the nucleotides provides for the formation of hydrogen bonds (hybridization) between pairs of nucleotide bases following specific rules. In Watson-Crick type hybridiza-

tion the rules are that adenine forms two hydrogen bonds with either thymidine or uracil and guanosine forms three hydrogen bonds with cytosine. If two single-stranded nucleic acid sequences can spatially align such that the hybridization can occur, they will form a stable, double helical structure and are said to be complementary. Hybridization of two nucleic acid molecules results in a change in free energy that is proportional to the number of hydrogen bonds formed between the two molecules. Watson-Crick hybridization can be thought of as a signal generating process in which the signal is the free energy change associated with nucleic acid alignment. Variation in the signal arises from the sequence variation which determines the degree to which the two sequences are complementary.

There are a number of biological processes in which nucleic acids participate that involve Watson-Crick hybridization including tRNA hybridization to mRNA during translation, recognition of the correct site for Okazaki fragment polymerization by primase during DNA replication [1], snRNA hybridization to pre-mRNA sequences during intron splicing [2], and siRNA hybridization to mRNAs during gene silencing [3]. In translation, the precision of hybridization between the anti-codon sequence of a tRNA molecule, carrying a specific amino acid, and the codon sequence of an mRNA molecule determines if that amino acid is polymerized into the polypeptide chain.

Two more examples of RNA-RNA hybridization encoding translation process information also exist. In 1974, Shine and Dalgarno [4] observed sequence complementarity between the 3'-terminal, single-stranded nucleotide sequence of the 16S rRNA (rRNA tail) and a window of mRNA sequence upstream of the start codon and hypothesized that the resulting hybridization could stabilize the mRNA/30S ribosome subunit complex. This observation was confirmed experimentally [5][6] and established 30S ribosome subunit recruitment as a role for the rRNA tail in translation initiation. More than a decade later, Weiss and co-workers [7][8] showed that hybridization between the rRNA tail and the mRNA was a critical component regulating a shift of reading frame during bacterial translation of the mRNA encoding the RF2 protein in *E. coli*. This was the first direct evidence of a role for hybridization of the rRNA tail with the mRNA during translation elongation. The requirements for exact sequence and exact spacing of sequence lead the investigators to conclude that the rRNA tail "...scans the mRNA during elongation ..." [8].

The idea of one nucleic acid molecule, the rRNA tail, "scanning" a second nucleic acid molecule, the mRNA, suggested to us the structure of a decoding algorithm from which a signal could arise. Each scanning alignment step would produce a free energy of hybridization value whose magnitude would be proportional to the degree of sequence complementarity. The linear series of these free energy values could constitute a signal indexed by nucleotide position on the mRNA molecule. The work of Weiss and co-workers [8] suggested to us that such a signal could encode information that the translation process utilizes for the maintenance of reading frame.

In considering this hypothesis, two expectations seemed critical. If information for the maintenance of reading frame exists in the rRNA tail signal, such an information signal would be expected to arise in the coding regions of a majority, if not all mRNA sequences. Additionally, if the signal did supply information for the maintenance of reading frame, it could exist across many species of bacteria if they employed the same mechanisms as *E. coli*. If the signal were found to exist across species, it would need to be maintained regardless of (G+C) content, known to vary across bacterial species. In Chapter 2, we establish that a free energy signal can be decoded from mRNA sequences utilizing an algorithm that models the mechanical movement of the mRNA through the ribosome during translation. Our study then characterizes this signal in terms of frequency, phase and magnitude. Our results indicate that coding regions of species tend to a mean species phase. Finally, we show that signal phase is a function of sequence (G+C) content, an indirect measure of codon bias. This last finding suggests the possibility that regulation of translational efficiency through codon usage could be mediated by signal phase.

To progress towards a model for control of reading frame, we applied electrical engineering concepts used for control system design. In electrical devices, input signals control device states. If the translating ribosome followed this design, its reading frame states, Frame 0, Frame +1 and Frame +2 (or -1), would be controlled by an input signal. In electrical devices, control system design takes the form of a mathematical model of a control system algorithm which decodes input signals to determine device state. The analytical tools of signal processing provide methods for detecting signals, extracting them from noise, characterizing signal parameters, and identifying the parameters and parameter behaviors that are predictive of device states. To use these tools requires a mathematical

model of the machine and an algorithm that simulates the machine process.

Our hypothesis is that the free energy signal arising from hybridization of the 16S rRNA tail with the mRNA is the input signal that controls reading frame. Modulation of reading frame could be accomplished through this signal if it supplied a force that adjusted the position of the mRNA relative to the ribosome. The first step towards validation of this hypothesis is the development of a mathematical model that defines ribosome position as a function of free energy signal parameters. The second step involves experimental testing of model predictions.

In Chapter 3 and Chapter 4, we develop the mathematical model describing control system design. In Chapter 5 we present an extensive set of results from applying our model to a variety of eubacterial species having known frameshifts. In Chapter 6 we summarize our findings and present avenues for further research.

Chapter 2

Free Energy Signal Characterization

2.1 Free energy calculation

A simple algorithm has been developed by Starmer and co-workers [9][10] and utilized for this study which generates a free energy signal as a function of nucleotide position (the decoding algorithm). Briefly, the algorithm requires a short nucleic acid sequence as the "decoder" that is successively aligned with a longer "message" sequence in which information is encoded (Figure 2.1). At each alignment, the algorithm calculates a free energy of nucleotide hybridization, ΔG° , for the optimal helical structure between the "decoder", for this study the 3'-terminal, single-stranded, nucleotides of the 16S rRNAs of bacterial species (16S rRNA tails), and the "message", the mRNA sequence that would be aligned with the 16S rRNA tail as the mRNA moves through the ribosomal complex as it is translated. The actual free energy calculation utilizes dynamic programming extended to allow for internal loops, to identify the minimal free energy conformation and the Individual Nearest Neighbor Hydrogen Bond model [11] to estimate the associated free energy value for that conformation. Adjustments to the free energy values for loop penalties [12] and for G/U mis-matches [13] are also incorporated. Bulges, more complex secondary structures involving only one of the two strands of RNA, are not considered in the calculation. This assumption was made based on structural models of the 70S ribosomal complex [14][15] in which the estimated space of the mRNA channel is thought to be insufficient for bulges

```
Position 0.    Free energy value = 0.0
rRNA: a u u c c u c c a c u a g
mRNA: G G U A A A A G A A U A A U G G C ...

Position 1.    Free energy value = 0.0
rRNA:       a u u c c u c c a c u a g
mRNA: ... G G U A A A A G A A U A A U G G C ...
⋮

Position 63.    Free energy value = -1.7
rRNA:          a u u c c u c c a c u a g
mRNA:  ... U C A C C G A G A U C C U G G U C ...

⋮

Position N-2.    Free energy value = 0.0
rRNA:            a u u c c u c c a c u a g
mRNA:  ... G C C G U C U G G U G A U G U A A

Position N-1.    Free energy value = -0.7
rRNA:              a u u c c u c c a c u a g
mRNA:  ... G C C G U C U G G U G A U G U A A
```

Figure 2.1: Alignment of the 16S rRNA tail with the mRNA sequence of gene *aceF* in *E. coli*. Free energy values of 0 indicate unfavorable binding. The length of the gene is N=1893 nucleotides.

and secondary structures to exist. The algorithm assigns the free energy value to a mRNA nucleotide. The alignment is then shifted one nucleotide downstream (in the 3' direction along the mRNA) and the free energy value of the new alignment is calculated and assigned. This approach generates a set of free energy values for an entire mRNA sequence indexed by nucleotide position. Our analysis assumes that the linear array of free energy values constitutes a discrete signal. This signal was examined using methods of time series analysis, with signal points indexed by nucleotide position, instead of time.

Sequence information and the genome databases used for this study are given in Table 2.1. Gene sequences for 12 eubacterial species, including *E. coli* K-12, were obtained from the NCBI GENBANK database (`http://www.ncbi.nlm.nih.gov/`). Using GENBANK annotation, the coding sequences were sorted into two categories: 1) verified

Table 2.1: List of eubacteria used in our study

Species Name	*GENBANK*	*16S tail*	*(G+C)%*
Buchnera aphidicola	NC_004545	auuccuccacuag	26
Borrelia burgdorferi	NC_001318	uuuccuccacuag	28
Bacillus licheniformis	NC_006322	uuuccuccacuag	46.2
Clostridium perfringens	NC_003366	uuuccuccacuag	27
Deinococcus radiodurans	NC_001263	uuuccuccacuag	66.6
Escherichia coli K-12	NC_000913	auuccuccacuag	50
Mycoplasma hyopneumoniae	NC_006360	uuuccuccacuag	28.6
Pseudomonas syringae	NC_005773	auuccuccacuag	55.6
Rhodobacter sphaeroides	NC_007493	uuuccuccacuag	68.8
Shigella boydii	NC_007613	auuccuccacuag	47.4
Salmonella enterica	NC_006511	auuccuccacuag	52.2
Thermus thermophilus	NC_005835	uuuccuccacuag	69.4

sequences, i.e. genes with a clearly annotated function and 2) hypothetical sequences, i.e. genes listed as hypothetical or putative. For *E. coli*, sequences encoding the 16S and 23S rRNAs were also used, designated as "non-coding" sequences to indicate that they do not encode amino acid sequence information. The 3'-terminal nucleotide sequences of the 16S rRNA (16S rRNA tails) for each species are also presented in Table 2.1. When calculating the free energy signals from a species population of mRNAs, the species' own 16S rRNA tail was used. These tails are the 3', single-stranded rRNA sequences that are potentially available for hybridization to the mRNA as it moves across the ribosome during translation.

A sample free energy signal, computed using the gene *aceF* sequence in *E. coli*, is shown in Figure 2.2. The estimated free energy for the alignment of the 5'-terminal nucleotide of the tail with the first base of the start codon is plotted at position 0 on the horizontal axis. The free energy estimates calculated for downstream alignments are plotted at positive indices while negative indices on the horizontal axis indicate free energy estimates for upstream alignments.

Two features of this variable free energy pattern are of note. There is a trough of negative free energy at nucleotide position -6. Earlier studies have identified the presence

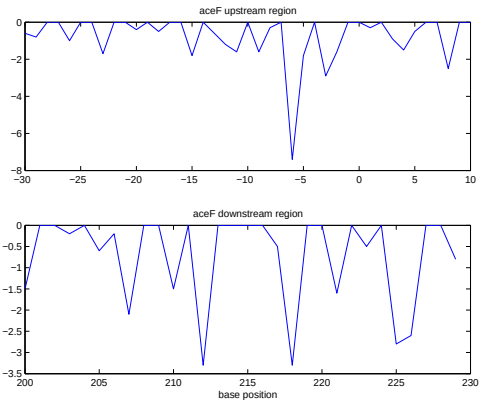


Figure 2.2: Free energy signal for gene *aceF* in *E. coli*

of an upstream free energy trough in genes of *E. coli* [16] and other bacteria [17]. This trough is interpreted as the signal feature for the Shine-Dalgarno region [18] [16] [19] [20] [17] [21] [22]. The other noteworthy feature is the pattern of negative free energy troughs that occur roughly every third nucleotide throughout the coding sequence. The suggestion of periodicity can be quantitatively confirmed using signal processing methodology.

2.2 Detection of periodicity

The set of free energy estimates are assumed to be a discrete signal, denoted as

$$\mathbf{y} = [y_0,\ y_1,\ \ldots,\ y_{N-1}] \tag{2.1}$$

The periodogram is defined as [23]

$$I_k = \frac{1}{N} |Y_k|^2, \quad k = 0 \ldots (N-1) \tag{2.2}$$

where

$$Y_k = \sum_{n=0}^{N-1} y_n e^{-j2\pi kn/N}, \quad k = 0 \ldots (N-1) \tag{2.3}$$

The periodogram of the free energy signal for a sample gene *aceF* reveals a dominant frequency of 1/3 cycles/base (Figure 2.3). The absence of other strong periodic components suggests that this signal can be modeled as the sum of a sine wave of frequency $f = 1/3$ and noise. A model for the signal can be written as

$$y_n = \mu + Asin(2\pi fn + \phi) + e_n \tag{2.4}$$

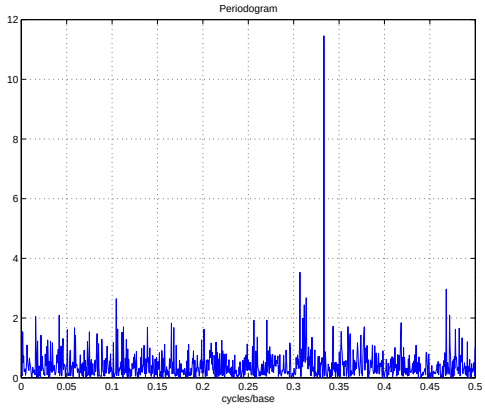


Figure 2.3: Periodogram for gene *aceF* in *E. coli*

where A is the amplitude, ϕ is the phase, $f = 1/3$ is the specified frequency and e_n is Gaussian white noise with variance σ^2. As per this model, if a periodic component of frequency $f = 1/3$ does not exist, the signal would be interpreted as white noise. To test the hypothesis that a free energy signal can be modeled from the variable free energy pattern arising from hybridization of the rRNA tail with the mRNA, the assumption is made that such a signal exists in the majority of coding regions. However, coding regions vary in length and signal length will affect the power of the statistical test. To ensure that the statistical test has sufficient power, the relationship between signal length, defined as nucleotide sequence length, and power was determined for a signal-to-noise ratio (SNR) of -18 dB, the mean SNR for *E. coli* K-12 coding regions (Table 2.3). As shown in Figure C.5, a power of 0.92 can be achieved using a signal length of greater than or equal to 900 nucleotides. Therefore, only coding regions of 900 nucleotides or greater were used to insure a robust statistical test.

The statistical test was performed with the null hypothesis that the free energy pattern contains only white noise, versus the alternate hypothesis that a signal does exist and it contains a dominant frequency component of $f = 1/3$ [24]. The signal model can be written in the equivalent form

$$y_n = \mu + C_1 \sin(2\pi f n) + C_2 \cos(2\pi f n) + e_n \tag{2.5}$$

where $C_1 = A\cos(\phi)$ and $C_2 = A\sin(\phi)$ are non-random constants.

The signal sum-of-squares, $|\mathbf{y}|^2$ can be partitioned by periodic components, allow-

ing the construction of a test of hypothesis [24]. Our null hypothesis is

$$H_0 : C_1 = C_2 = 0$$

and our alternate hypothesis is

$$H_1 : C_1 \textit{ and } C_2 \textit{ are both not zero}$$

From [24], we know that under H_0,

$$(2I_{N/3}) \quad \sim \quad \sigma^2\chi^2(2) \tag{2.6}$$

and $I_{N/3}$ is independent of

$$(\sum_{i=0}^{N-1} y_i^2 - I_0 - 2I_{N/3}) \quad \sim \quad \sigma^2\chi^2(N-3) \tag{2.7}$$

We may reject H_0 in favor of H_1 at level α if

$$\left[(N-3)I_{N/3}\right] / \left[\sum_{i=0}^{N-1} y_i^2 - I_0 - 2I_{N/3}\right] > F_{1-\alpha}(2, N-3) \tag{2.8}$$

The results of this test for the verified and hypothetical sequences greater than 900 nucleotides in various eubacteria are given in Table 2.2. The test is performed at level $\alpha = 0.05$. "Sample Size" indicates the number of sequences in each category. "Passed" indicates the number of sequences whose free energy signal shows only one periodic component of the assumed frequency for the hidden periodicity statistical test, i.e., $f = 1/3$. We observe that 95.9% of the selected verified sequences and 90.4% of the chosen hypothetical sequences in *E. coli* demonstrate strong periodicity at $f = 1/3$ in their free energy signals. For the other bacterial species in our study, whose genomic (G+C) contents ranged from 26% to 69.4% (Table 2.1), the majority of their verified and hypothetical sequences were also found to demonstrate strong periodicity at $f = 1/3$.

If the information encoded by the periodic signal is relevant to translation, we might expect that it would only be present in the coding sequences and not in the sequences that are not translated. To test this hypothesis would require applying our algorithm to non-coding sequences minimally 750 to 900 nucleotides in length, based on estimated relationship of statistical power and SNR, to have sufficient statistical power (Figure C.5).

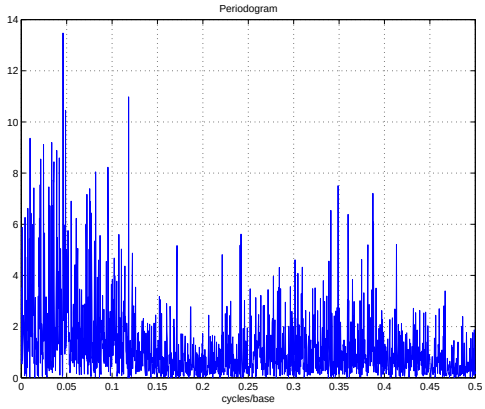


Figure 2.4: Periodogram calculated using the free energy signal for a 23S rRNA sequence in *E. coli*

In bacteria, the rRNA sequences are the only sequences that are sufficiently long to satisfy these considerations. Therefore, we used the 16S and 23S rRNA gene sequences, of which there are 7 each in *E. coli*, to test the hypothesis. The free energy patterns calculated using these sequences did not show periodicity at $f = 1/3$, consistent with the correlation between signal presence and periodicity and sequences that are translated. Figure 2.4 shows an example of the periodogram of a non-coding sequence, 23S rRNA.

2.3 Estimation of signal parameters

For those free energy signals for which our model (Equation (2.4)) is valid, we can evaluate the power of the 1/3 harmonic and estimate the noise variance using trigonometric regression [25][26]. The regression procedure performs a least-squares fit of the model described by Equation (2.5) to the free energy signal $\mathbf{y}$. Detailed mathematical derivations are shown in Appendix D.

The best-fit values of C_1 and C_2, denoted $\hat{C}_1$ and $\hat{C}_2$ respectively, can be used to estimate the magnitude and phase of the signal using Equations (2.9) and (2.10). It can be shown that the regression procedure is equivalent to maximum-likelihood estimation, under the assumption that the i.i.d. noise, e_n, follows a normal distribution [25].

$$\hat{A} = \sqrt{\hat{C}_1^2 + \hat{C}_2^2} \tag{2.9}$$

Table 2.2: Detection results

Species	*Sequence Type*	*Sample Size*	*Passed*
Buchnera aphidicola	Verified	206	197
	Hypothetical	34	32
Borrelia burgdorferi	Verified	265	242
	Hypothetical	140	99
Bacillus licheniformis	Verified	1318	1068
	Hypothetical	375	272
Clostridium perfringens	Verified	489	484
	Hypothetical	679	648
Deinococcus radiodurans	Verified	577	573
	Hypothetical	490	475
Escherichia coli	Verified	1193	1144
	Hypothetical	758	685
Mycoplasma hyopneumoniae	Verified	186	173
	Hypothetical	164	131
Pseudomonas syringae	Verified	1919	1888
	Hypothetical	472	440
Rhodobacter sphaeroides	Verified	977	972
	Hypothetical	359	357
Shigella boydii	Verified	875	838
	Hypothetical	715	653
Salmonella enterica	Verified	995	952
	Hypothetical	771	684
Thermus thermophilus	Verified	654	654
	Hypothetical	197	194

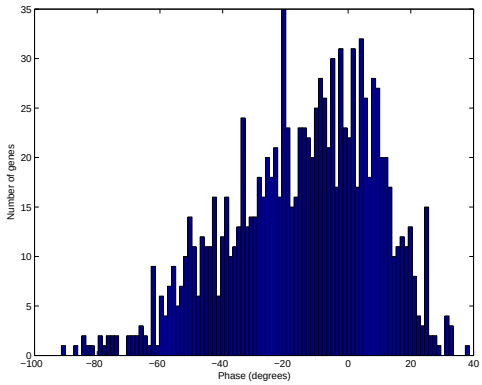


Figure 2.5: Histogram of Phase of verified sequences

Table 2.3: *E. coli* signal parameters

Parameter	*Mean*	*Std Dev*
Phase (degrees)	-14.53	23.26
SNR (dB)	-18.35	1.84

$$\hat{\phi} = \arctan(\hat{C}_2/\hat{C}_1) \tag{2.10}$$

The power of the sinusoidal component can be calculated using Equation (2.11). The mean-squared error (MSE) from regression yields an estimate of the noise variance $\hat{\sigma}^2$. The power of the noise and the signal-to-noise ratio (SNR) are calculated using Equations (2.12) and (2.13) respectively.

$$P_{signal} = 10\log_{10}\left((\hat{A}^2)/2\right) dB \tag{2.11}$$

$$P_{noise} = 10\log_{10}\left(\hat{\sigma}^2\right) dB \tag{2.12}$$

$$SNR = (P_{signal} - P_{noise})\, dB \tag{2.13}$$

Histograms for signal phase and SNR for verified genes in *E. coli* are shown in Figure 2.5 and Figure 2.6, respectively. The mean and standard deviation of the estimated parameter values are shown in Table 2.3. These values are calculated using verified genes in *E. coli* that pass our detection test (1144 in number).

The revelation of free energy periodicity embedded in coding regions provides the foundation for further studies to determine if the signal could provide information for the maintenance of reading frame. If this is its function, it would be reasonable to expect the

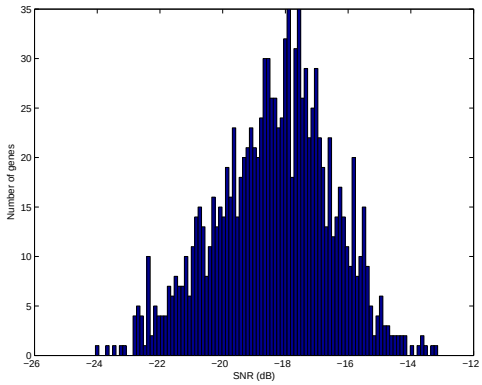


Figure 2.6: Histogram of SNR of verified sequences

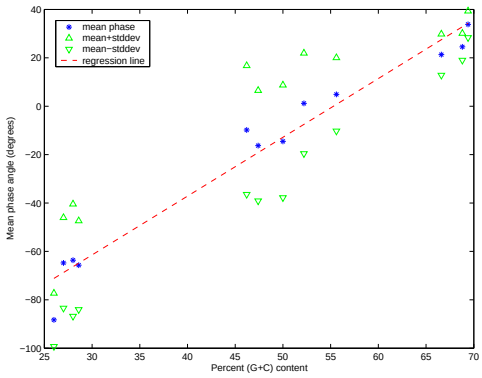


Figure 2.7: Phase as a function of (G+C) across eubacterial species

signal to be present in coding regions of eubacterial species in general. To determine if this is true, we selected 12 eubacteria of varying (G+C) content, listed in Table 2.1. The verified genes that passed the detection test for each species were used for analysis. The free energy signals for each species were calculated using its specific 16S tail, shown in Table 2.1. We found that a periodic signal is present in the coding regions of genes in all the species tested and that the mean phase of these signals is roughly proportional to the (G+C) content (Figure 2.7). An ANOVA test indicated a significant effect of (G+C) content on the signal phase.

2.4 Discussion

Our algorithm models the movement of the ribosome relative to the mRNA during translation. This model assumes that a continual series of mRNA sequence windows are accessible for hydrogen bond formation to occur between the 16S rRNA tail and the mRNA as they move by each other during the translation process. The free energy associated with each of these windows is a function of the degree of complementarity between the 16S rRNA tail and the mRNA sequence window. Using this model, it is clear that a periodic signal is encoded in the free energy variation. Standard signal processing and statistical analyses show that this signal has a dominant frequency 1/3 and that it is encoded in the majority of protein-encoding sequences of genes in a diverse group of eubacterial species, including *E. coli*. This periodic signal is not present in genomic sequences that encode rRNAs which do not participate in translation. Although this result is consistent with the signal being present only in sequences that are translated, the limited sample size (there are only 7, rRNA encoding genes in *E. coli*) prevents meaningful statistical confirmation of the hypothesis that the signal exists only in sequences encoding proteins. These results reveal a signal and provide a signal decoding mechanism, however they do not explain what parameters contribute to signal structure and what role it could play in translation.

In our model, the energetic variation of the signal arises from the variation in mRNA nucleotide sequence. That the signal has a frequency 1/3 implies that the mRNA nucleotide sequence has a frequency 1/3. Periodicity in the coding regions of genes has been observed prior to our results using statistical correlation analysis of coding regions. Lio and co-workers [27] have investigated prokaryotic and eukaryotic DNA sequences for the presence of sub-codes following a periodicity rule based on the ideas of several investigators [28] [29]. The analysis of individual gene sequences from both prokaryotes and eukaryotes revealed period-three recurrence of (G+C) bases in the codon third position, coherent with the reading frame for the gene ($(G+C)^{-3}$ periodicity). This period-three recurrence was found in some translated sequences in both prokaryotes and eukaryotes but was not found in introns, repetitive DNA or sequences encoding rRNAs or tRNAs [27]. These results are consistent with ours. The analysis of Lio and co-workers also identified translated sequences in which $(G+C)^{-3}$ periodicity could not be resolved however they did not exclude the possibility that a weaker period-three signal could be present. This result is consistent with a

relatively low SNR for their signal, impairing resolution of all but the strongest signals.

The new observation of a mean phase for *E. coli* genes suggested the subsequent study to determine if the presence of coding region periodicity with constant phase is a feature peculiar to *E. coli* or that is a more general feature of prokaryotic genomes. Our results indicate that each bacterial genome does have a distribution of signal phase, however, the mean phase for each species is different. Knowing that the (G+C) content of genomes varies, and that this variation is a reflection of the species preference for certain codons (generally referred to as Synonymous Codon Bias [30]), we hypothesized that signal phase is a function of (G+C) content. Our regression results indicate that phase is a function of (G+C) content and that there is a significant difference in the signal phase of species that are widely distributed across (G+C) content. The functional relationship between phase and (G+C) content means that signal phase can be manipulated through codon selection.

The role of Watson-Crick hybridization between 16S rRNA sequences, including the tail, and the mRNA during translation has long been the subject of investigation. Trifonov [31] suggested that this hybridization could play a role in maintenance of reading frame during translation. The elegant work of Weiss and co-workers [7][8] using mutant analysis of both the mRNA and the 16S rRNA clearly showed that hybridization between these two molecules was critical in the shift of reading frame that regulates the production of RF2 protein in *E. coli*. Our results suggested that parameters of the energetic signal, i.e. phase, could supply the translational process information for maintenance of reading frame.

Our findings are consistent with this hypothesis. To maintain the correct reading frame, the ribosome must translocate three nucleotides after each amino acid is incorporated into the polypeptide product of the translation process. Therefore, it would be expected that a signal encoding reading-frame information would have a dominant 1/3 frequency, as our signal does. In addition, using a robust statistical test, we found the signal to be present in genomic sequences that encode proteins, again an expected result. Our results also imply that specific manipulation of codon usage, which would modify (G+C) content, could locally adjust phase and potentially impact reading frame fidelity.

Chapter 3

Ribosomal Memory Model

Our free energy signal is noisy, resulting in a low signal-to-noise ratio (SNR). In order to extract information from the signal, some way of boosting the SNR is required. Since the signal pattern repeats itself every 3 bases, we could remove some of the noise by calculating base-wise averages of free energy triplets. Note that the ribosome reads the mRNA in steps of 3 bases. Therefore, we should be able to keep track of the ribosome's reading frame by tracking changes in the noise-reduced free energy signal pattern, specifically, its phase. Our hypothesis is that the ribosome needs to "see" a specific phase in order to stay in frame. This is in agreement with the "shifty sites" model of frameshifting [8].

3.1 Method of accumulation

Let us imagine a hypothetical memory for the ribosome system, consisting of a stack of 3 registers. Assuming that the interaction between the 16S rRNA tail and the mRNA sequence is indicative of reading frame, the memory system should maintain updates of the free energy released due to this interaction. As the energy flows into memory, information pertaining to the reading frame gets updated. Since frameshifting is a "localized" phenomenon triggered by short sequences [32][33], this memory model has potential utility. We will now present some details of our accumulation method.

We denote the register contents by the vector $\mathbf{R}^{(k)}, k = 1 \ldots \frac{L}{3}$, where $\frac{L}{3}$ is the number of codons. We store the first three energy values (corresponding to the first codon)

in consecutive registers i.e.

$$\mathbf{R}^{(1)} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

We then accumulate the free energies corresponding to the second codon, resulting in

$$\mathbf{R}^{(2)} = \begin{bmatrix} y_0 + y_3 \\ y_1 + y_4 \\ y_2 + y_5 \end{bmatrix}$$

After accumulating the signal for length of k codons, the register contents will be

$$\mathbf{R}^{(k)} = \begin{bmatrix} R_1^{(k)} \\ R_2^{(k)} \\ R_3^{(k)} \end{bmatrix} = \begin{bmatrix} \sum_{p=0}^{k-1} y_{3p} \\ \sum_{p=0}^{k-1} y_{3p+1} \\ \sum_{p=0}^{k-1} y_{3p+2} \end{bmatrix}$$

We repeat this procedure until we reach the last codon of the mRNA sequence, i.e. until $k = \frac{L}{3}$.

3.2 Cumulative magnitude and phase

The register contents $\mathbf{R}^{(k)}$ represent a snapshot of the free energy signal pattern. The three points have a sinusoidal nature due to the dominant periodicity of the energy pattern. This allows us to calculate the cumulative magnitude M_k and phase θ_k by interpolation. As a result, $\mathbf{R}^{(k)}$ can be represented as a complex phasor $\mathbf{V}_k = M_k e^{j\theta_k}$ [34]. We equate the contents of the registers, after subtracting their mean, to points on a sine-wave and solve Equations (3.1), (3.2) and (3.3) for M_k and θ_k.

$$r_1^{(k)} = R_1^{(k)} - \left(\frac{\sum_{n=1}^{3} R_n^{(k)}}{3} \right) = M_k sin\left(\theta_k\right) \tag{3.1}$$

$$r_2^{(k)} = R_2^{(k)} - \left(\frac{\sum_{n=1}^{3} R_n^{(k)}}{3} \right) = M_k sin\left(\theta_k + \frac{2\pi}{3}\right) \tag{3.2}$$

$$r_3^{(k)} = R_3^{(k)} - \left(\frac{\sum_{n=1}^{3} R_n^{(k)}}{3} \right) = M_k sin\left(\theta_k + \frac{4\pi}{3}\right) \tag{3.3}$$

3.3 Signal-to-Noise Ratio

Based on our free energy signal model (2.4), the register contents take the form

$$r_1^{(k)} = (kA)\, sin\,(\phi) + \left(\sum_{j=0}^{k-1} z_{3j} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \tag{3.4}$$

$$r_2^{(k)} = (kA)\, sin\left(\frac{2\pi}{3} + \phi\right) + \left(\sum_{j=0}^{k-1} z_{3j+1} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \tag{3.5}$$

$$r_3^{(k)} = (kA)\, sin\left(\frac{4\pi}{3} + \phi\right) + \left(\sum_{j=0}^{k-1} z_{3j+2} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \tag{3.6}$$

Therefore,

$$M_k = kA$$

and

$$\sigma_k^2 = \left(\frac{2k}{3}\right) \sigma^2$$

where σ_k^2 is the noise variance of the contents of the memory register $\mathbf{R}^{(k)}$. The SNR of the register contents is given by

$$\Gamma_k = \frac{M_k^2}{2\sigma_k^2} = \frac{3k}{2} \left(\frac{A^2}{2\sigma^2} \right)$$

Thus, the accumulation of points corresponding to the same sinusoidal pattern causes the SNR to grow linearly with the number of codons.

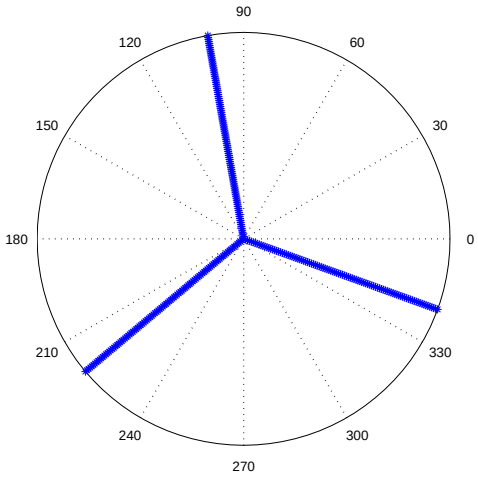


Figure 3.1: Thick lines indicate phase boundaries for each reading frame, relative to an initial signal phase of -20°

3.4 Visualization using polar plots

The magnitude M_k and phase θ_k of the register contents can be visualized on a polar plot, with the radial coordinate representing magnitude and the angular coordinate representing phase. Because the free energy signal frequency equals 1/3 cycles/nucleotide, each 120° sector of the polar plot represents one nucleotide (see Figure 3.1). For the free energy signal to play a role in reading frame determination, it would be expected that variation in M_k and/or θ_k would correlate with shifts in reading frame. To determine if such a correlation might exist, two genes were selected: *aceF*, a gene which does not encode a frameshift, and *prfB*, a well-studied gene whose mRNA sequence is known to encode a programmed frameshift at codon 26 [35].

Although the polar plot for *aceF* (Figure 3.2) shows some variation, the cumulative phase stays roughly constant at about -15°, within the sector of one nucleotide. Similar phase constancy was observed in all the 1673 verified genes in *E. coli* of length 200 codons or greater [36]. However, considerable variation in track within the nucleotide sector can occur (see Figure 3.3). By comparison, the polar plots of *prfB* (Figures 3.4 and 3.5) are quite different. The plot starts in the same nucleotide sector as that for *aceF*, but around codon 26 it swings through approximately 240°. When the phase change is complete, the plot re-establishes itself within a different nucleotide sector and remains there, with small

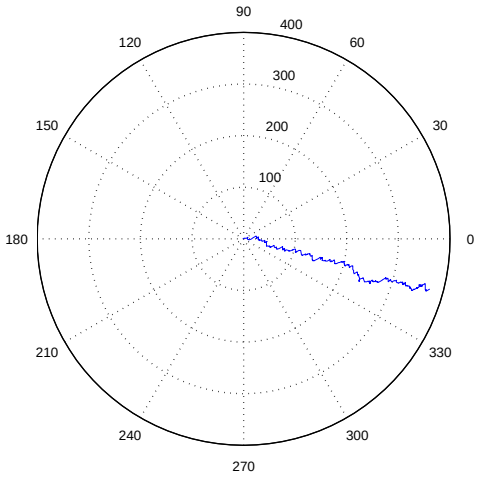


Figure 3.2: Polar plot for gene *aceF* in *E. coli*

variation, to the end of the gene. Although provocative and consistent with our hypothesis, analysis of other genes known to encode frameshifts would strengthen the correlation.

In a 1984 paper, Schoner et al [37] discussed their investigation of the conditions necessary for high-level expression of methionyl bovine growth hormone (Met-bGH) in *E.coli*. They found that by the introduction of additional codons 3' to the initiating AUG codon, expression levels of bGH get boosted to 30% of total cell protein. They attributed this high level of expression to variation in the efficiency of mRNA translation. We calculated free energy signals for bGH derivatives shown in the paper and examined them using our cumulative analysis (Figures 3.6 - 3.9). Note the difference between the plot for the sequence that yielded the highest amount of bGH, and the others that failed. All the sequences show a phase change in their polar plots. But in the case of the high-yielding sequence, it appears as though the ribosome is being drawn back into its original reading frame. In the others, the polar plot shows no sign of restoration to the species phase angle - the ribosome continues to pick the wrong codons and probably encounters a premature stop. We agree that the lengths of the sequences are small, leading to noisy estimates for cumulative phase. Nevertheless, the results support the utility of our methods, in a broad sense.

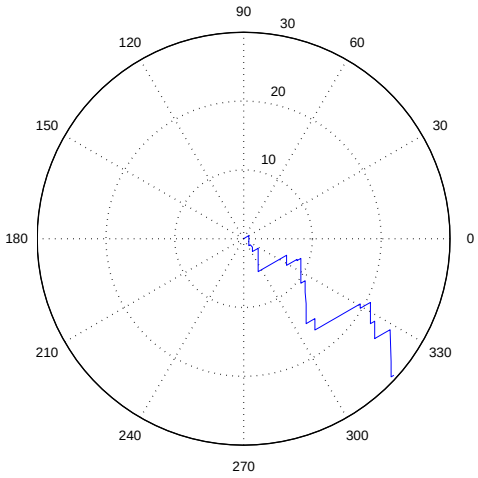


Figure 3.3: Polar plot for gene *tsf* in *E. coli*

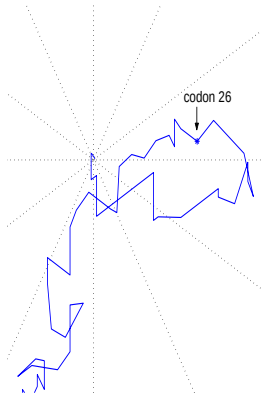


Figure 3.4: Partial polar plot for gene *prfB* in *E. coli*: arrow points to the location of frameshift, marked by a *

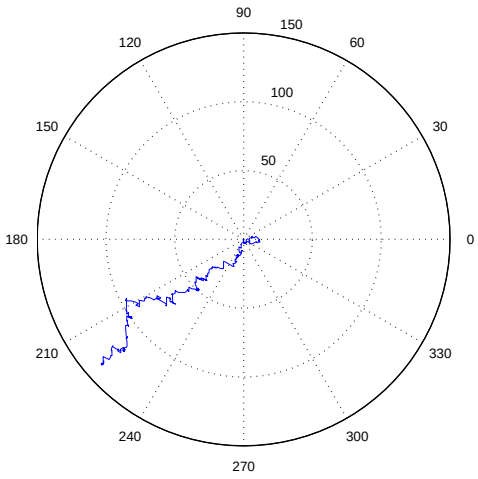


Figure 3.5: Polar plot for gene *prfB* in *E. coli*

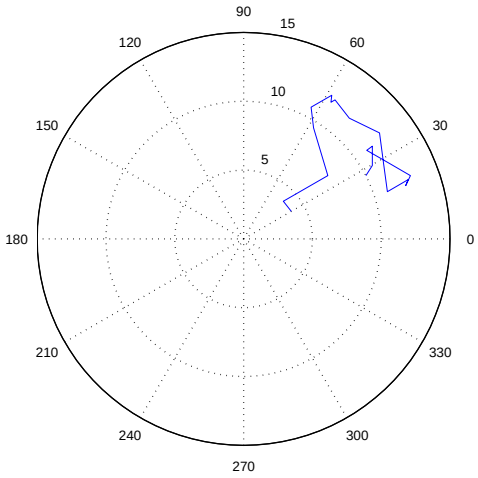


Figure 3.6: Sequence that yields 30% bGH

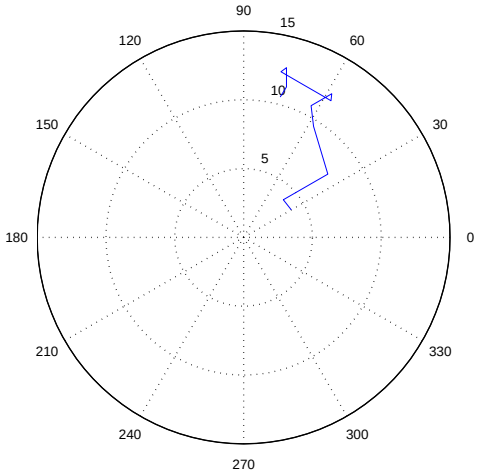


Figure 3.7: Sequence that yields 1.7% bGH

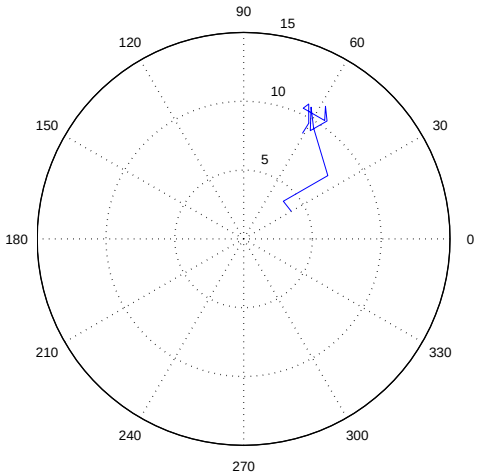


Figure 3.8: Sequence that yields 0.5% bGH

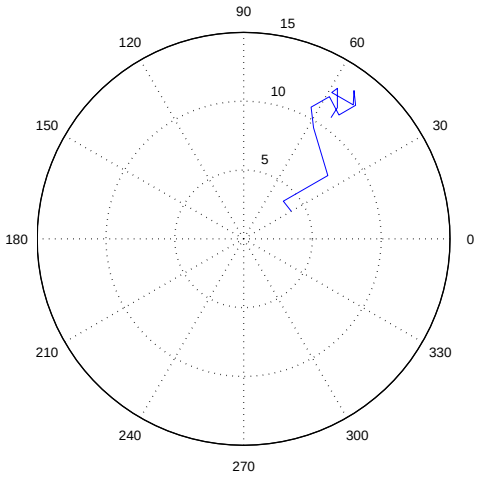


Figure 3.9: Sequence that yields 0.5% bGH

RECODE[1] is a database of non-canonical translational events such as frameshifts, ribosomal hops and codon redefinition [38][39]. Experimentally verified *prfB* gene sequences for twelve prokaryotes other than *E. coli* were obtained and their free energy signals were calculated using the corresponding species' 16S tail, and signal parameters were generated using the cumulative method. The *prfB* polar plots for all the examined species are shown in Chapter 4. A significant phase change is observed around the frameshift location in all these genes, consistent with the results obtained using the *prfB* gene in *E. coli*.

3.5 Drawbacks

Our cumulative model of signal phase, although useful for revealing frameshift sites encoded in gene sequences, has one significant drawback. For every additional codon, a greater perturbation of the free energy signal will be needed to shift the cumulative phase. This means that the model will have difficulty identifying frameshifts if they occur towards the end of a long gene sequence. Also, there is no experimental evidence that indicates that the entire gene sequence upstream of a frameshift site has a controlling influence on the frameshift. The sequence elements that result in a shift in reading frame during translation are small and can be localized in a short sequence within the coding region [35]. To accommodate these concerns we developed a new model that estimates instantaneous

[1] http://recode.genetics.utah.edu/

signal phase at each codon.

Chapter 4

Displacement Model

4.1 Conceptual framework

Let us first summarize our findings thus far. For a gene without a frameshift, the polar plot lengthens itself radially (due to growth in magnitude) but stays at a roughly constant phase angle ($\theta_k \approx \theta_{sp}$). When a +1 frameshift happens, the phase moves to a new nucleotide sector, +240° or -120° away. From the *prfB* polar plot, we see that the phase shifts about 60° before it gets to the frameshift location (from approximately -20° to approximately +40°), the equivalent of one-half of a nucleotide. Then it begins its track at the angle that re-establishes it in the new nucleotide sector, +240° from where it originated.

We designate $x = 0$ as the initial state, i.e., reading frame 0, as one of the two stable states of the ribosome-mRNA system. We assign unit increments in x for every 60° increment in phase, i.e. for every $\frac{1}{2}$ nucleotide-shift in the mRNA sequence. If the ribosome shifts a whole nucleotide, as it does in the +1 frameshift, we have $x = 2$. So a +1 frameshift can be modeled as a state transition from $x = 0$ to $x = 2$. The intermediate value $x = 1$ can be thought of as a *boundary* point, where there is equal likelihood of picking either the codon in Frame 0 or the codon in Frame +1. The value of displacement x reflects the proportion of the out-of-frame codon being viewed by the ribosome in its A-site. When $x = 1$, half of the first nucleotide of the +1 frame codon is exposed in the A-site (see Figure 4.1).

The angle θ_k estimates the true signal phase ϕ (from Equation (2.4)). As k gets

Reading frame 0, Perfect exposure, x = 0

```
-------------
|   |   |   |
| U | G | A |
|   |   |   |
-------------
```

Imperfect exposure, x = 1

```
  -------------
  |   |   |   |
| U | G | A | C
  |   |   |   |
  -------------
```

Reading frame +1, Perfect exposure, x = 2

```
-------------
|   |   |   |
| G | A | C |
|   |   |   |
-------------
```

Figure 4.1: Exposure of the codon in the A-site and its relationship to displacement x. Shown here is the 26th codon of the *prfB* gene where a shift into the +1 frame occurs.

larger, i.e. as we accumulate more of the signal, θ_k gives an increasingly accurate estimate of ϕ. From the polar plot for *prfB*, we observe that a +1 frameshift is accompanied by a 240° phase-shift. At first this seems counter intuitive. It might be expected that a +1 frameshift would result in a shift in phase of +120° (counter-clockwise) or one nucleotide sector, rather than a shift of +240° (counter-clockwise) or a shift of two nucleotide sectors. We will now lay the foundations of a model that captures this behavior.

4.2 The differential vector

As stated earlier, the cumulative energy signal, owing to its sinusoidal nature, can be represented as $\mathbf{V}_k = M_k e^{j\theta_k}$. We will refer to $\mathbf{V}_k$ as the *cumulative vector*. The contents of $\mathbf{V}_k$ contain a summation of the entire free energy signal up to codon k. The derivative of $\mathbf{V}_k$ with respect to codon position k gives the instantaneous energy available at codon k.

$$\mathbf{D}_k = \frac{d}{dk}\left(M_k e^{j\theta_k}\right) = M_k \frac{d}{dk}\left(e^{j\theta_k}\right) + e^{j\theta_k}\frac{dM_k}{dk} \tag{4.1}$$

The magnitude and phase of the differential vector $\mathbf{D}_k$, referred to as differential magnitude and differential phase, are given by Equation (4.2) and Equation (4.3) respectively.

$$|\mathbf{D}_k| = \sqrt{\left(\frac{dM_k}{dk}\right)^2 + \left(M_k \frac{d\theta_k}{dk}\right)^2} \tag{4.2}$$

$$\angle\mathbf{D}_k = \theta_k + \arctan\left(\frac{M_k \frac{d\theta_k}{dk}}{\frac{dM_k}{dk}}\right) \tag{4.3}$$

To calculate $|\mathbf{D}_k|$ and $\angle\mathbf{D}_k$, we will need the derivatives ($\frac{dM_k}{dk}$ and $\frac{d\theta_k}{dk}$), which can be evaluated using function approximation techniques [40]. A second order polynomial can be fitted to a window of points centered around M_k, to evaluate its derivative, $\frac{dM_k}{dk}$. An identical procedure is followed for computing $\frac{d\theta_k}{dk}$.

We observe that for a signal that stays roughly in phase, $\frac{d\theta_k}{dk} \approx 0$, and so, $|\mathbf{D}_k| \approx \frac{dM_k}{dk}$ and $\angle\mathbf{D}_k \approx \theta_k$. We know, from previous work that the free energy signals in a given eubacterium have a roughly constant phase [41]. For *E. coli*, that angle is $\theta_{sp} \approx -20°$. For a normal, non-frameshifting gene of length L nucleotides in *E. coli*, we see that $\theta_k \rightarrow \theta_{sp}$ as $k \rightarrow \frac{L}{3}$. Within the context of our hypothesis, the differential vector

$\mathbf{D}_k$ represents a force acting on the ribosome at codon k that adjusts the position of the ribosome relative to the mRNA, i.e., that modulates reading frame.

Another element believed to play an integral part in programmed frameshifts is ribosomal pausing [35]. Sipley and Goldman [42] provide experimental evidence that supports a frameshift model in which ribosomal pause time is a major determinant of frameshift probability, with pause time a function of tRNA availability. Therefore, we introduce the concept of *wait-time*, a measure of how long the ribosome waits for the tRNA to associate with the ribosome A-site, into our displacement model.

4.3 Estimating wait-time

The actual availability of tRNA, estimated using two-dimensional polyacrylamide gel electrophoresis, was found to be proportional to codon frequency for moderately expressed genes [43]. Using a set of mRNA sequences in *E. coli* that have N codons in all, the frequency of each codon (except the stop codons) can be calculated as

$$f_i = \frac{N_i}{N}, \quad i = 1 \ldots 61 \tag{4.4}$$

where N_i is the number of codons of type i. If a particular tRNA recognizes only one codon, then the codon frequency would be indicative of its availability. If there is more than one codon recognized by a tRNA isoacceptor, then the availability of that isoacceptor will be the sum of the individual codon frequencies. We estimate the availability of each tRNA isoacceptor using

$$\gamma_p = \sum_{i=1}^{n_p} f_i, \quad p = 1 \ldots 20 \tag{4.5}$$

where n_p is the number of codons that code for amino acid p.

Codons having abundant tRNAs would have short wait-times, and vice-versa. We assume a decreasing linear relationship between the wait-time τ and the tRNA availability γ, as shown in Equation (4.6). The wait-time gives an approximate number of cycles for which the ribosome can adjust itself while waiting for the appropriate tRNA. The number of wait cycles for a few sample codons are shown in Table 4.1.

$$\tau_p = \frac{\max(\gamma) - \gamma_p}{\min(\gamma)} \tag{4.6}$$

Table 4.1: Wait-times for a few sample codons in *E. coli*

Codon	*Amino-acid*	*Number of wait-cycles*
aac	Asn	7
ccu	Pro	16
acg	Thr	13
cuu	Leu	13
uuc	Phe	7
gca	Ala	2

4.4 The complete model

The vector $\mathbf{D}_k$ represents a force that could produce a linear movement of the ribsome one way or the other until the corresponding tRNA is found for the codon in the A-site. The displacement at each codon position is calculated incrementally (Δx), with the sign of Δx indicating the direction of movement (+ = downstream, - = upstream). The total displacement x_k is obtained by accumulating Δx for the corresponding number of wait cycles. When the ribosome is in reading frame 0, we define $x = 0$ and when it moves into the +1 frame, we define $x = 2$. We claim that the following equation captures the behavior in both reading frame states:

$$\Delta x_k = -C\,|\mathbf{D}_k| \sin\left(\angle\mathbf{D}_k + \frac{\pi x_k}{3} - \theta_{sp}\right) \tag{4.7}$$

The argument of the sine function contains the instantaneous measurement of phase:

$$\theta_{\Delta x} = \frac{\pi x_k}{3} - \theta_{sp} \tag{4.8}$$

Observe that when $x = 0$, the cumulative phase is at the species angle i.e., $\angle\mathbf{D}_k = \theta_{sp}$, leading to $\Delta x = 0$. When $x = 2$, we have $\angle\mathbf{D}_k = \theta_{sp} + \frac{4\pi}{3}$, again leading to $\Delta x = 0$. To calculate Δx, we introduce a constant of proportionality C, and calibrate it using the *prfB* signal. Mathematically, C measures the rate at which the ribosome adjusts itself to perturbations in x. For each unit of wait-time (also referred to as a *wait-cycle*), the incremental displacement Δx_k^j gets added onto the current position x_k^j. The total displacement is then assigned to the next codon $k+1$. Note that we are using the superscript j to index increments made during the wait-time of the ribosome. If the ribosome waits for τ cycles at codon k, the total initial displacement at codon $k+1$ would be assigned as

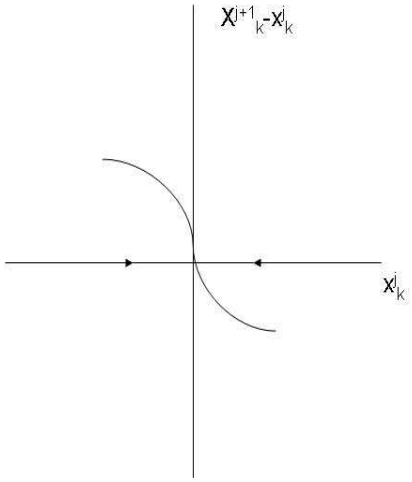


Figure 4.2: Vector field generated by Equation (4.7)

$$x_{k+1}^0 = \sum_{j=1}^{\tau} \Delta x_k^j \tag{4.9}$$

4.5 Stability

In practice, all the above equations hold approximately, so it is important to establish stability of the ribosome-mRNA system in a rigorous manner [44]. Equation (4.7) can be written as a recursive relation

$$x_k^{j+1} = x_k^j - C\,|\mathbf{D}_k| \sin\left(\angle\mathbf{D}_k + \frac{\pi x_k^j}{3} - \theta_{sp}\right) \tag{4.10}$$

Stability at $x^* = 0$:

When the ribosome is in reading frame 0, $x_k^j = 0$ and $\angle\mathbf{D}_k = \theta_{sp}$. Substituting $x_k^j = 0$ into Equation (4.10) leads to $x_k^{j+1} = x_k^j$, and hence, $x^* = 0$ is a fixed point. Let $\eta_j = x_k^j - x^*$ be a small perturbation away from x^*. To see whether the perturbation grows or decays, we substitute $x_k^j = \eta_j + x^*$ into Equation (4.10). The recursive relation can now be written as

$$x^* + \eta_{j+1} = x^* + \eta_j - C\,|\mathbf{D}_k| \sin\left(\angle\mathbf{D}_k + \frac{\pi(x^* + \eta_j)}{3} - \theta_{sp}\right)$$

Substituting $x^* = 0$, we get

$$\eta_{j+1} = \eta_j - C\,|\mathbf{D}_k| \sin\left(\frac{\pi\eta_j}{3}\right) \tag{4.11}$$

Since η_j is small, we have

$$\eta_{j+1} \approx \eta_j - C\,|\mathbf{D}_k|\,\frac{\pi\eta_j}{3} = \left(1 - C\frac{\pi\,|\mathbf{D}_k|}{3}\right)\eta_j$$

By making C fairly small, it can be ensured that $\left(C\frac{\pi|\mathbf{D}_k|}{3}\right) < 1\ \forall k$. This implies that η_j decays to zero as j gets large, since $\left(1 - \frac{\pi|\mathbf{D}_k|}{3}\right) < 1$. Thus, small perturbations cause the displacement to converge to the fixed point $x^* = 0$. The idea is illustrated in Figure 4.2.

Stability at $x^* = 2$:

When the ribosome is in reading frame +1, $x_k^j = 2$ and $\angle\mathbf{D}_k = \theta_{sp} + \frac{4\pi}{3}$. Substituting these into Equation (4.10) yields $x_k^{j+1} = x_k^j$, so $x^* = 2$ is a fixed point. For a nearby point $x_k^j = x^* + \eta_j$, the recursive relation takes the form

$$x^* + \eta_{j+1} = x^* + \eta_j - C\,|\mathbf{D}_k| \sin\left(\angle\mathbf{D}_k + \frac{\pi(x^* + \eta_j)}{3} - \theta_{sp}\right)$$

Substituting $x^* = 2$, we get an equation identical to Equation (4.11). Following identical steps, we may establish the stability of the fixed point $x^* = 2$.

The above arguments have established that the Equations (4.7) and (4.8) are structured so that the states $x = 0$ and $x = 2$ represent stable fixed points of the ribosome-mRNA system. Transition between the states is governed by the differential vector $\mathbf{D}_k$ and the time τ for which the ribsome waits at codon k.

Chapter 5

Results

5.1 Using *E. coli*

5.1.1 *prfB* gene

Two model parameters, the species phase angle, θ_{sp}, and the constant, C, must be specified to generate displacement values. The species phase angle θ_{sp} is the mean phase angle estimated from the set of verified genes as annotated in GENBANK, using the method described in [41]. For *E. coli*, the estimated value is $\theta_{sp} = -13°$. For gene *prfB* in *E. coli*, the value of $C = 0.005$ gave the highest resolution of a jump in displacement at codon 26. These values of θ_{sp} and C were used for subsequent analyses of other genes in *E. coli*. The values of these parameters for other bacteria are listed in Section 5.2. At the first codon of a gene sequence, the ribosome is locked into Frame 0, so we use $x_1 = 0$. The stop codons are assigned a large number of wait-cycles, typically 1000.

The displacement plots for the *aceF* and *prfB* genes of *E. coli* are given in Figures 5.1 and 5.2, respectively. Several features of these plots are of note. The displacement plot for *aceF* (Figure 5.1), a gene lacking a frameshift, shows that $x \approx 0$ for the entire length of the coding region. This behavior of x indicates that our method does not detect a frameshift in this gene, the expected result. In contrast, the displacement plot for the *prfB* gene (Figure 5.2) shows a sudden shift in x at codon 26, the absolute value of which is slightly greater than 2 and it is in the positive direction. Notice also that the displacement goes negative before shooting to the +2 position. This happens because of the way the

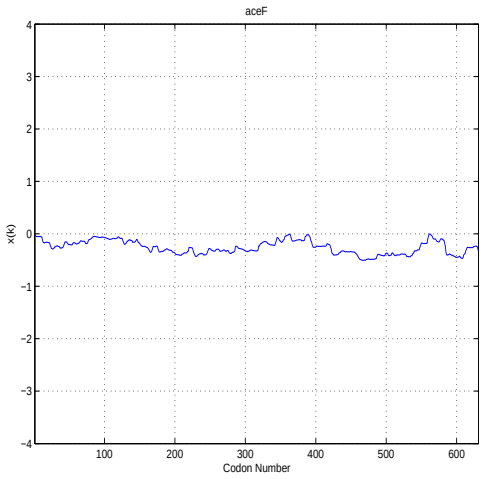


Figure 5.1: Displacement plot for gene *aceF* in *E. coli*

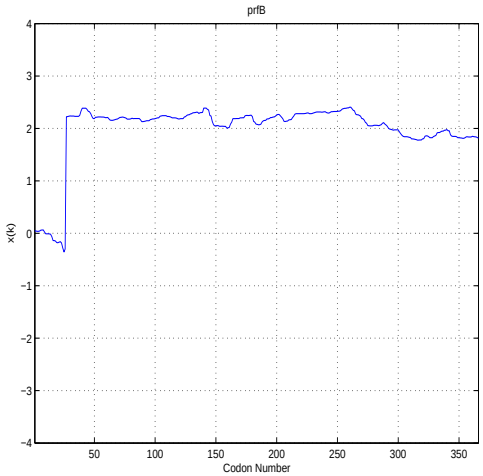


Figure 5.2: Displacement plot for gene *prfB* in *E. coli*

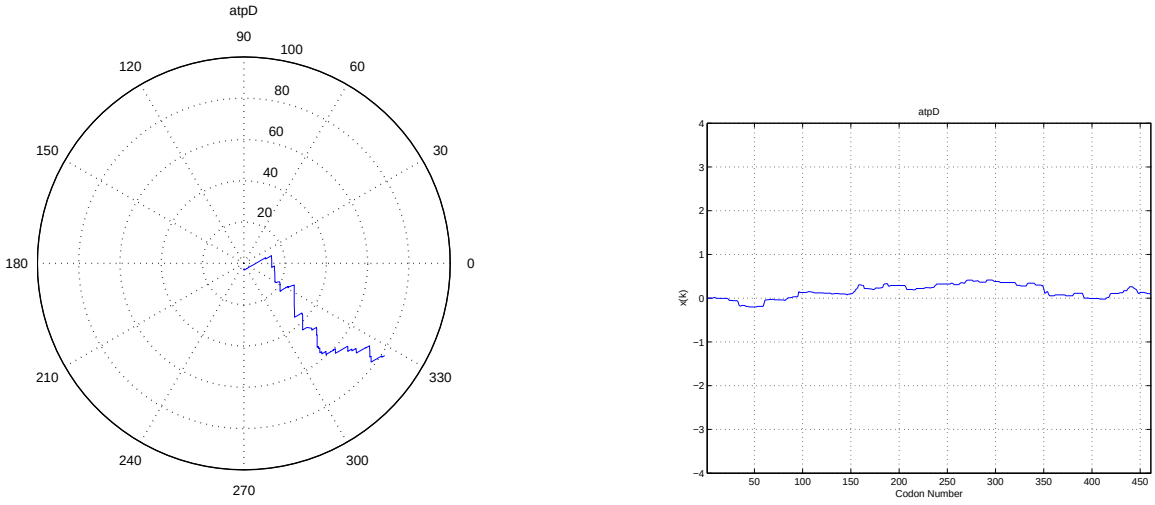
equation (4.7) is written. As ($\angle \mathbf{D}_k + \frac{\pi x}{3}$) gets larger than θ_{sp}, Δx goes negative. After a certain point, the argument of the sine function exceeds 180°, driving Δx positive.

Our algorithm is scaled such that a displacement value of $x = 2$ indicates a shift of one nucleotide, so in this case, the displacement indicates a +1 nucleotide shift in reading frame. This is also an expected result given that codon 26 is the location of a +1 frameshift in the *prfB* gene. For the remainder of the sequence, i.e., from codon 27 to the end of the gene, the value of x remains roughly at $x = 2$. This indicates that the gene stays in the new reading frame.

The *prfB* displacement plots for the remaining bacteria that we analyzed are given in Section 5.2.

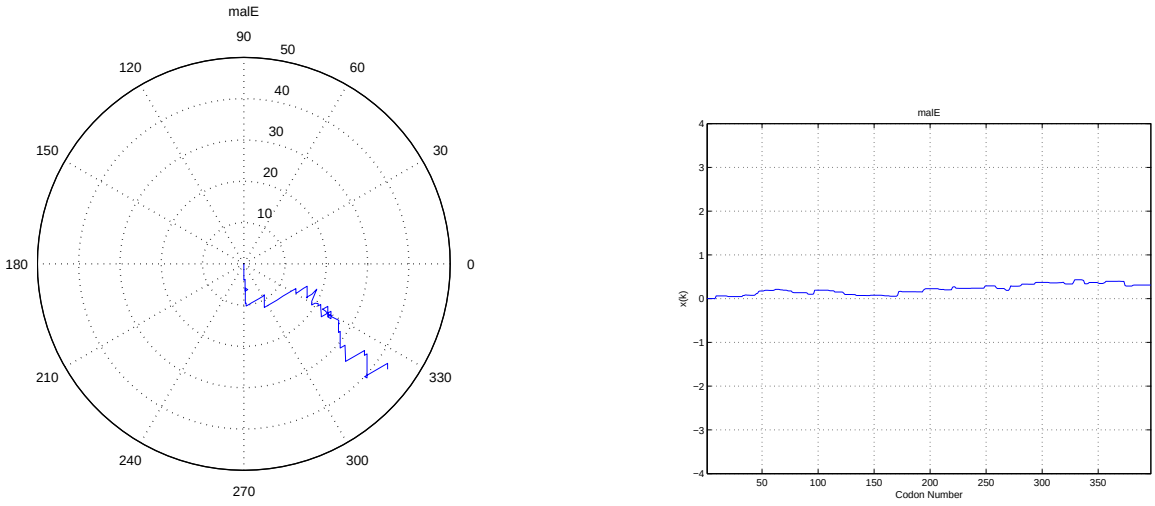
5.1.2 Link Genes

Link *et al.* [45] assessed the *in vivo* abundances of proteins in *E. coli* using electrophoresis, and ranked the genes in decreasing order of yield. We calculated the free energy signals for 87 such genes in *E. coli*, and analyzed them using our model. We found that for 86 of these genes, $-1 < x_k < 1$ for all values of k, indicating that the ribosome stays in frame for the entire length of each sequence. For the one remaining gene, we found slight deviation from the boundary value of $x_k = 1$ at $k = 70$, indicating a low probability of picking the in-frame codon at that location. The polar plots and displacement plots for 10 of these genes are included below.



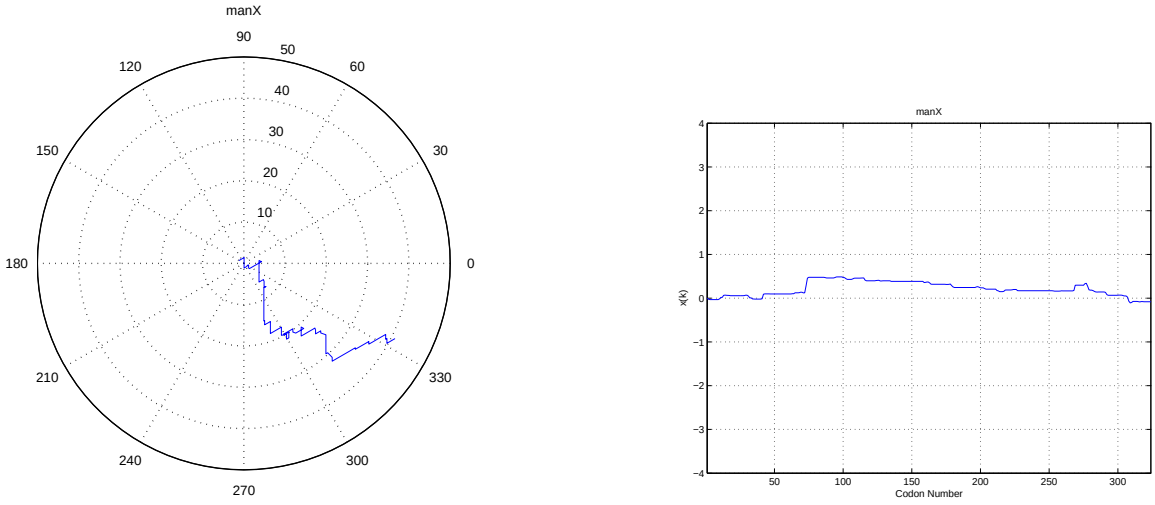
Polar plot **Displacement plot**

Figure 5.3: Analysis of gene *atpD* in *E. coli*



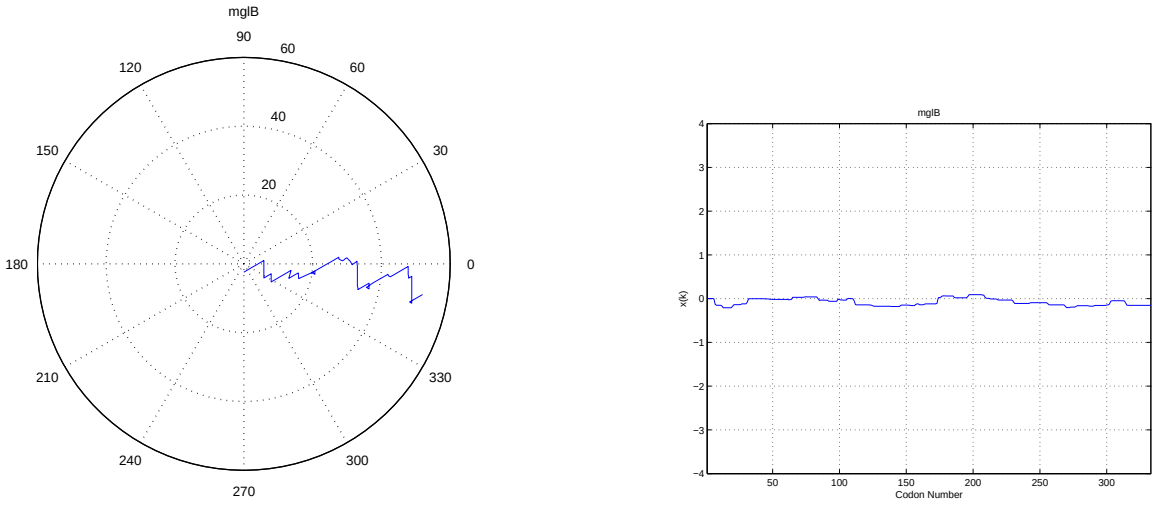
Polar plot **Displacement plot**

Figure 5.4: Analysis of gene *malE* in *E. coli*



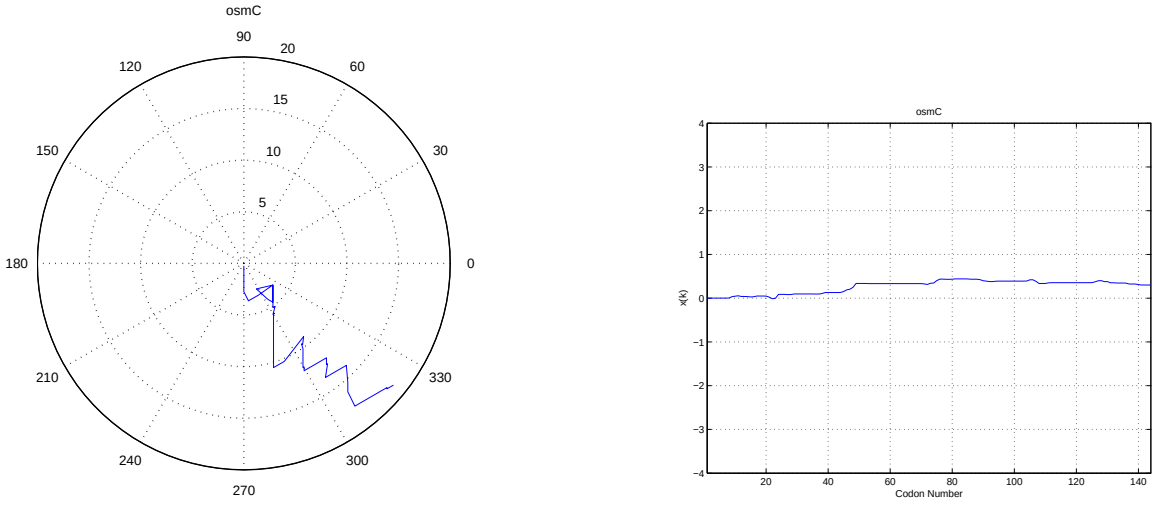
Polar plot **Displacement plot**

Figure 5.5: Analysis of gene *manX* in *E. coli*



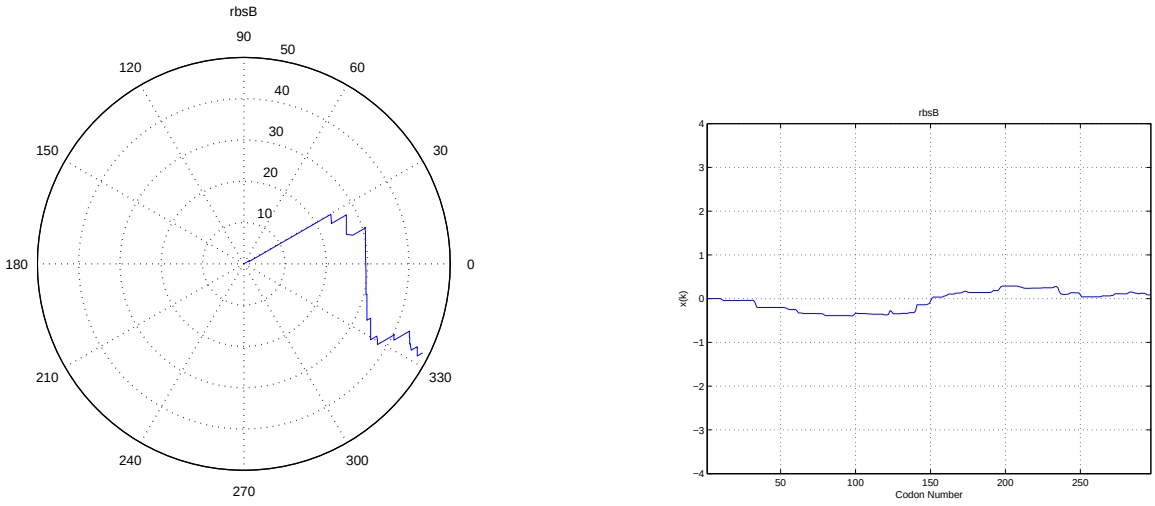
Polar plot **Displacement plot**

Figure 5.6: Analysis of gene *mglB* in *E. coli*



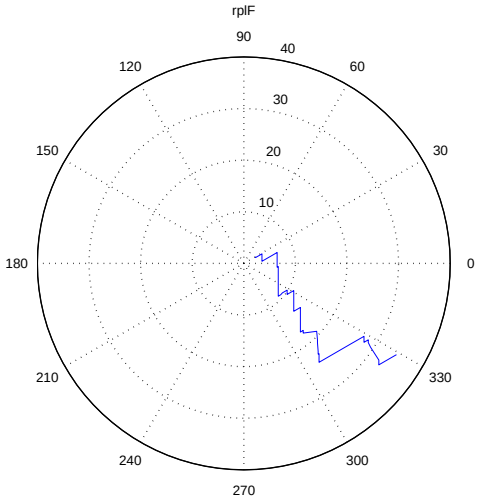
Polar plot **Displacement plot**

Figure 5.7: Analysis of gene *osmC* in *E. coli*

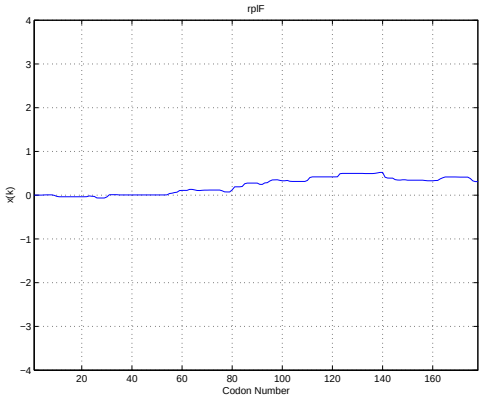


Polar plot **Displacement plot**

Figure 5.8: Analysis of gene *rbsB* in *E. coli*

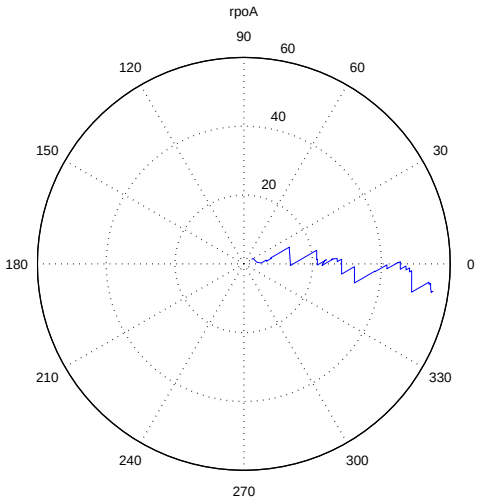


Polar plot

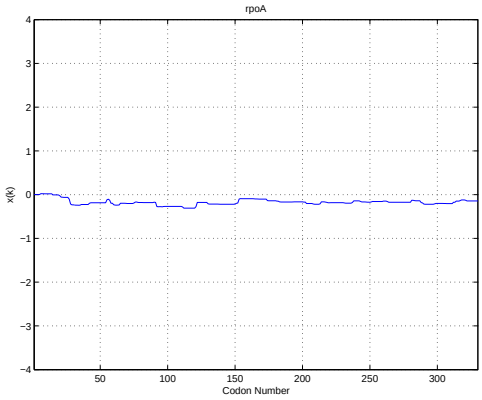


Displacement plot

Figure 5.9: Analysis of gene *rplF* in *E. coli*

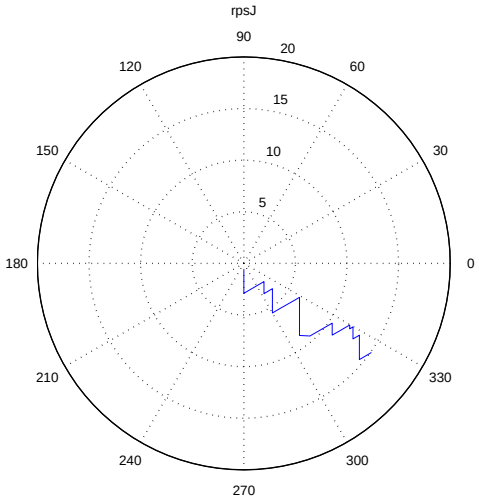


Polar plot

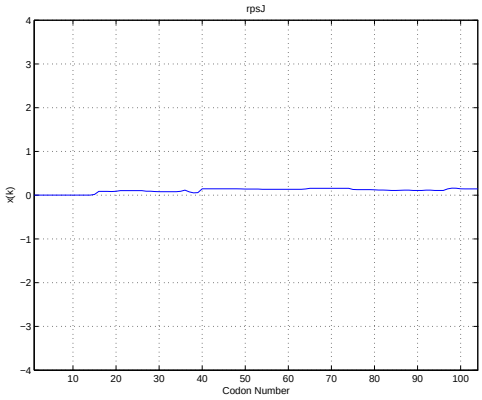


Displacement plot

Figure 5.10: Analysis of gene *rpoA* in *E. coli*

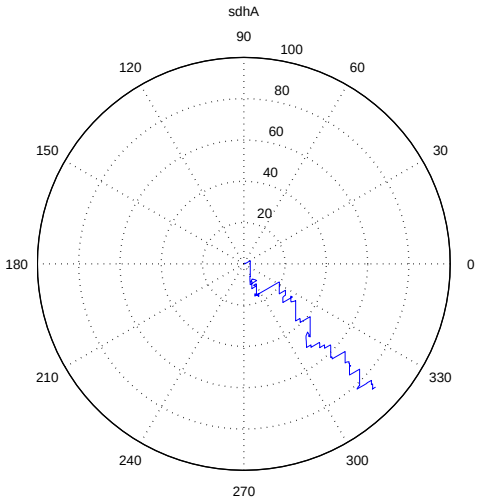


Polar plot

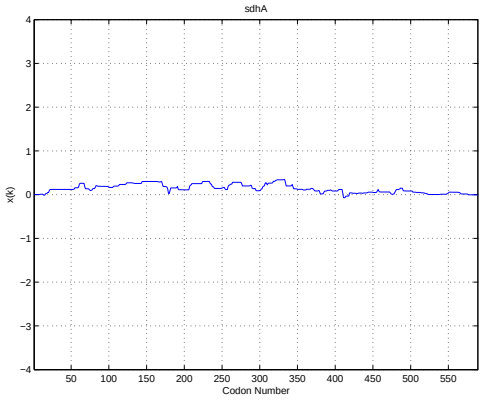


Displacement plot

Figure 5.11: Analysis of gene *rpsJ* in *E. coli*



Polar plot



Displacement plot

Figure 5.12: Analysis of gene *sdhA* in *E. coli*

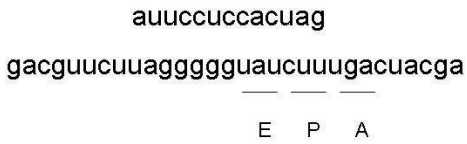


Figure 5.13: Alignment of 16S rRNA tail with the mRNA

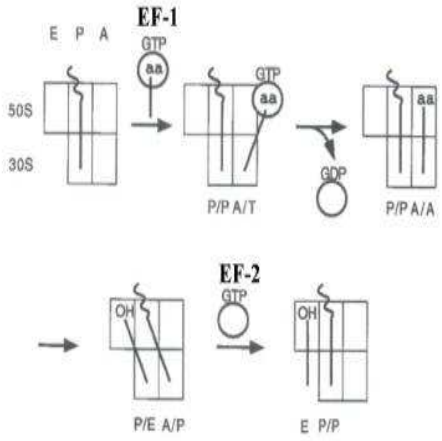


Figure 5.14: A, P, E sites of the ribosome

5.1.3 Issue of spacing

In our model, we have assumed the depicted alignment (Figure 5.13) of the 16S tail with the mRNA sequence at the frameshift codon. This alignment seems unrealistic since it leaves no room for the P-site (see Figure 5.14 showing the 3 sites of the ribosome).

In the absence of clear experimental evidence, this assumption is doubtful in accuracy [46][47][48][49]. In order to shed some light on this issue, we repeated our analysis by changing the spacing in the tail-mRNA alignment. By moving the tail backwards by a codon, we have a spacing of 1 codon, enough to accommodate the P-site (Figure 5.15). By moving it backwards by one more codon, we have a spacing of 2 codons, which could

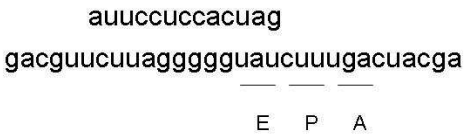


Figure 5.15: Alignment of 16S rRNA tail with the mRNA, 1 codon spacing

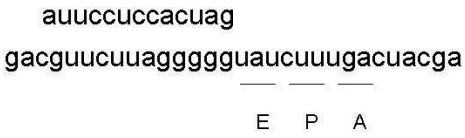


Figure 5.16: Alignment of 16S rRNA tail with the mRNA, 2 codon spacing

accommodate both the P-site and the E-site (Figure 5.16).

As mentioned previously, the following condition needs to be satisfied for the frameshift to be detected by our model:

$$\left(\angle \mathbf{D}_k + \frac{\pi x_k}{3} - \theta_{sp}\right) > \pi$$

For *E. coli*, the above condition is satisfied even when the spacing is increased to 1 codon, but fails for a spacing of 2 codons. As we increase the spacing, we are making an attempt to detect the frameshift by using a lesser portion of the free energy signal. The angle of the differential vector $\angle \mathbf{D}_k$ may not reach the desired value for the above condition to be satisfied. But, by shifting the value θ_{sp} a little, we may still be able to satisfy the condition, thereby making the model work. This may be all right, since there is a variance of about 25 ° associated with the estimate of θ_{sp}.

The estimated value of $\theta_{sp} = -15°$ works for *prfB* in *E. coli* for a spacing of 1 codon. For a spacing of 2 codons, the value must be decreased to at least $\theta_{sp} = -50°$. This is not intuitively satisfying, since we have changed the angle quite a bit.

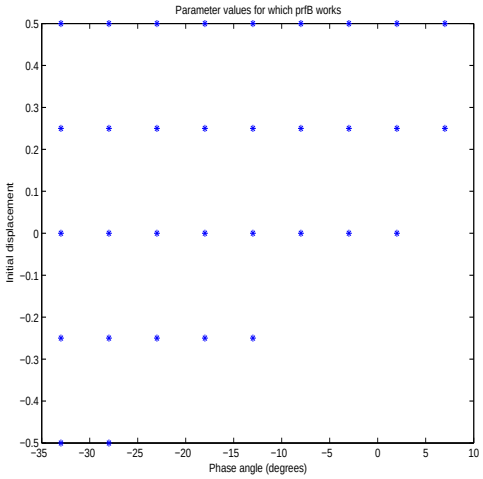


Figure 5.17: Parameter values for which *prfB* works in *E. coli*, spacing=0

5.1.4 Sensitivity Analysis

The two relevant parameters in our model are the species-specific phase angle θ_{sp} and the initial displacement x_0. These are unknown beforehand, and while θ_{sp} can be estimated from a sample of verified genes [41], there is no intuitive way to estimate x_0. To evaluate the sensitivity of our model to these parameters, we create a grid of values and proceed with the following two-step analysis:

1. First evaluate if *prfB* shows the expected frameshift at the known location. The criteria used to judge if the model works for *prfB* are:

   - Displacement changes from $x \approx 0$ to $x \approx 2$ at the right location
   - Displacement stays at $x \approx 2$ until the end of the gene sequence
   - The magnitude of the jump in displacement at the frameshift codon should be larger than the shift in displacement at other codons

   Using these conditions, we arrive at a set of points at which the model works for each value of spacing (see Figures 5.17, 5.18 and 5.19).

2. At each of these points, we evaluate how the model performs using a set of verified genes. We form a test dataset of $N = 100$ genes picked from Link *et al.*'s findings [45] and some long genes as per GENBANK annotation. As we noted earlier, if the

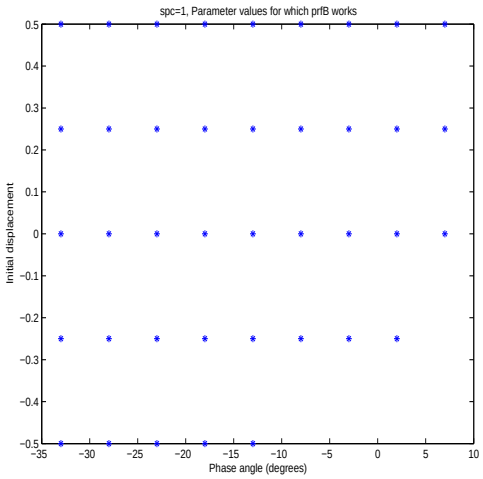


Figure 5.18: Parameter values for which *prfB* works in *E. coli*, spacing=1

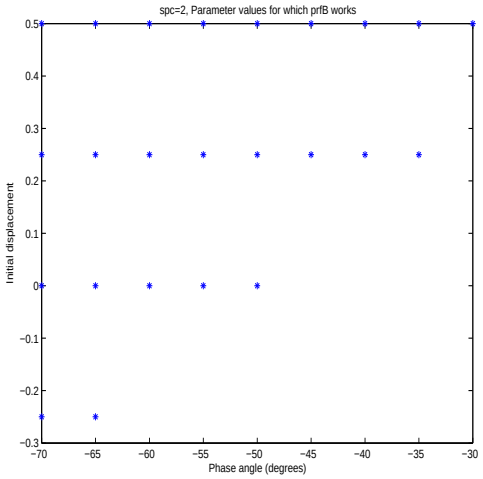


Figure 5.19: Parameter values for which *prfB* works in *E. coli*, spacing=2

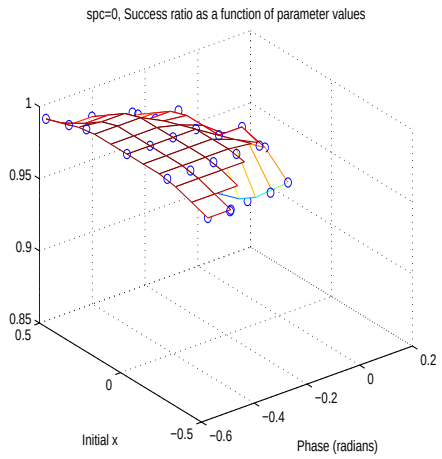


Figure 5.20: Success ratio as a function of parameter values in *E. coli*, spacing=0

model parameters are right, then for a normal (non-frameshift) gene, we must have $|x_k| < 1 \forall k$. Using each pair of parameter values identified in the previous step, we find the number of genes in our dataset that satisfy this condition (say, N_s). The success ratio $R = \frac{N_s}{N}$ is plotted as a function of parameter values. The analysis is repeated for all three values of spacing (see Figures 5.20, 5.21 and 5.22).

We observe that, using our original spacing (spc=0), we obtain a number of points at which *prfB* works in *E. coli*. The success ratio varies somewhat, but not too much over the parameter space. As seen from the above figures, the surface is roughly convex around the value of θ_{sp} and reaches a peak at $x_0 = 0$. This indicates that the model works best when the value of phase is close to the estimated value θ_{sp} and the initial displacement is close to zero. These results lend support to the assumption that the ribosome is locked in to Frame0 almost perfectly when elongation starts.

If the spacing were correct, then we would expect that the success ratio R is maximized around the estimated value of θ_{sp} and the surface is roughly concave about both θ_{sp} and $x_0 = 0$. Based on these considerations, the spacing of 1 codon as shown in Figure 5.15 seems to be optimal.

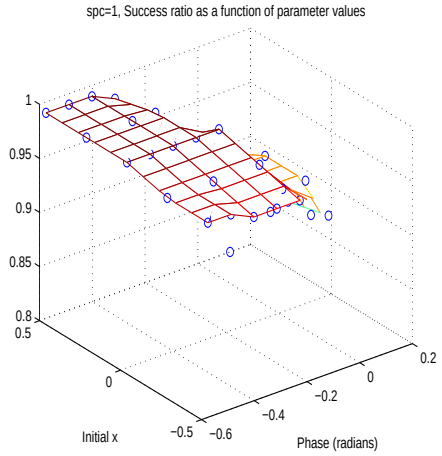


Figure 5.21: Success ratio as a function of parameter values in *E. coli*, spacing=1

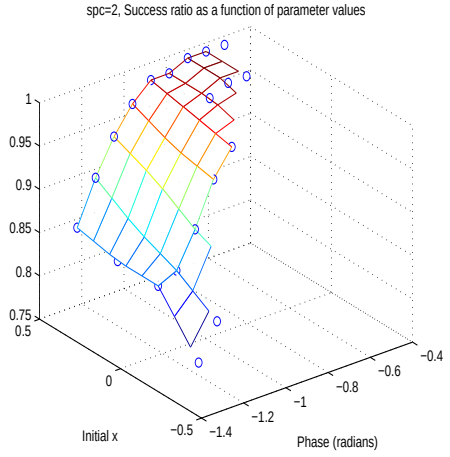


Figure 5.22: Success ratio as a function of parameter values in *E. coli*, spacing=2

5.2 Other eubacteria

A set of 11 eubacteria (apart from *E. coli*) have been selected for analysis, based on the following factors:

- Matching of accession number between RECODE (`http://recode.genetics.utah.edu/`) and GENBANK (`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`)
- Availability of a consensus sequence for the last 13 bases of the 16S rRNA, also referred to as the *16S tail*

For each species, Table 5.1 indicates

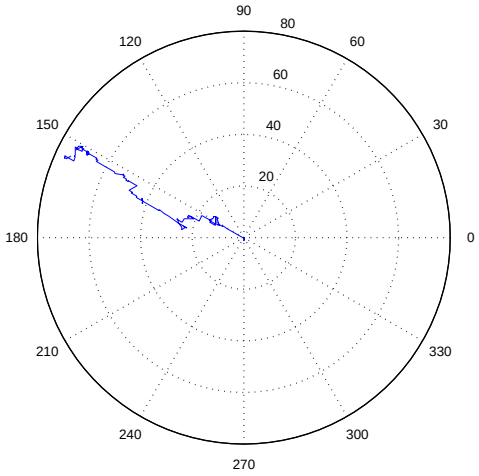
- its name
- its GENBANK accession number
- the 13 base-long *16S tail*
- the GC-content of the species, expressed as a percentage
- the mean species phase angle, θ_{sp}, in degrees
- the value of the parameter C, as defined in the model
- the number of the codon at which frameshift (FS) occurs, according to the RECODE database (following the convention that the first codon in the sequence, i.e. the start codon is numbered 1)

The analysis illustrated in Section 5.1 is performed for each of the species listed in Table 5.1.

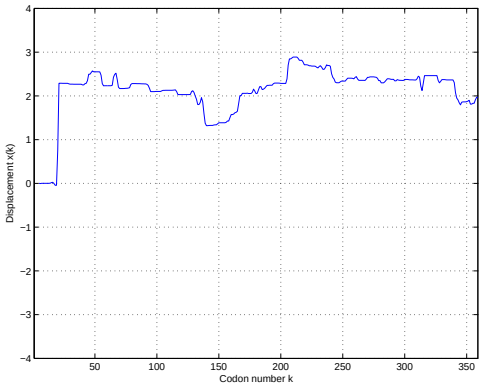
Table 5.1: Table of selected eubacteria

Name	*GENBANK*	*16S tail*	*(G+C)*	θ_{sp}	*C*	*FS codon*
B. burgdorferi	NC_001318	uuuccuccacuag	28.2	−63	0.005	20
B. halodurans	NC_002570	uuuccuccacuag	43.7	−23	0.005	25
B. subtilis	NC_000964	uuuccuccacuag	43.5	−24	0.01	25
C. pneumoniae	NC_000922	uuuccuccacuag	40.6	−54	0.005	24
C. trachomatis	NC_000117	uuuccuccacuag	41.3	−55	0.005	24
H. influenzae	NC_000907	auuccuccacuag	38.1	−58	0.005	26
P. multocida	NC_002663	auuccuccacuag	40.4	−48	0.01	26
S. mutans	NC_004350	uuuccuccacuag	36.8	−57	0.005	28
S. typhimurium	NC_003197	auuccuccacuag	52.2	3	0.005	26
T. pallidum	NC_000919	uuuccuccacuag	52.8	−8	0.005	25
X. fastidiosa	NC_002488	uuuccuccacuag	52.6	−15	0.005	26

5.2.1 Borrelia burgdorferi



Polar plot



Displacement plot

Figure 5.23: Analysis of gene *prfB* in *B. burgdorferi*

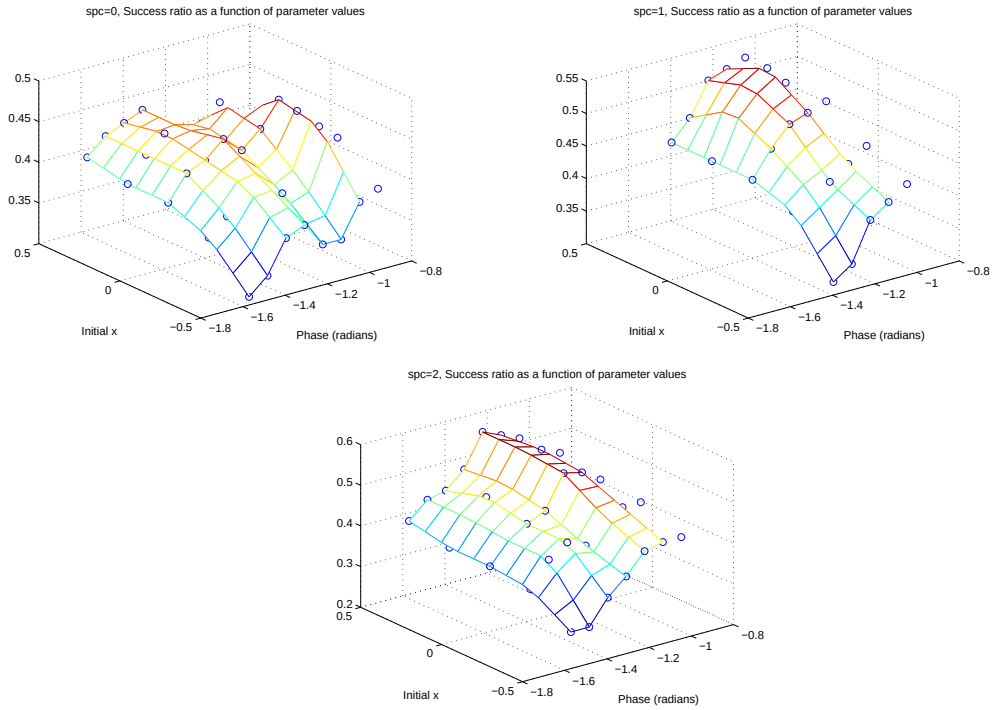
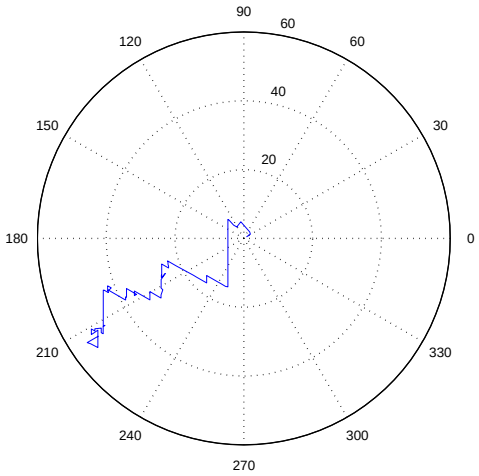
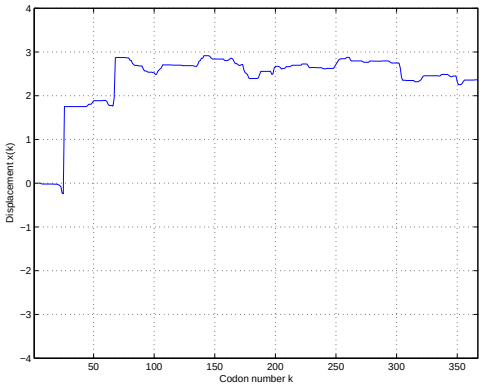


Figure 5.24: Sensitivity analysis using normal genes in *B. burgdorferi*

5.2.2 Bacillus halodurans



Polar plot



Displacement plot

Figure 5.25: Analysis of gene *prfB* in *B. halodurans*

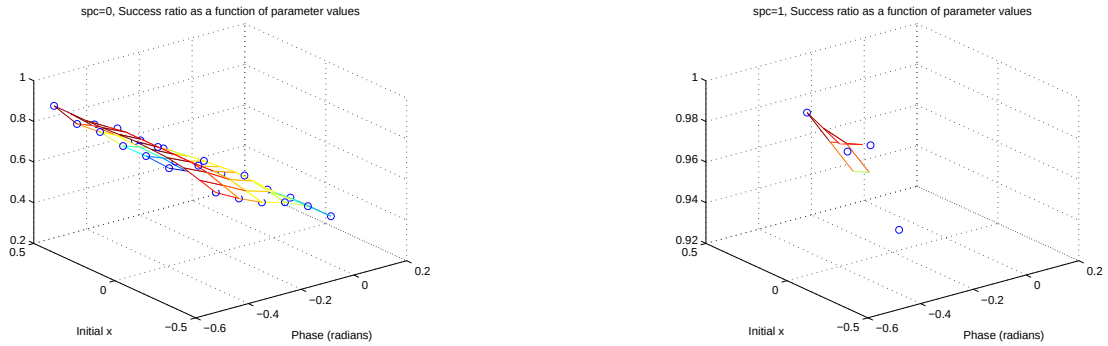
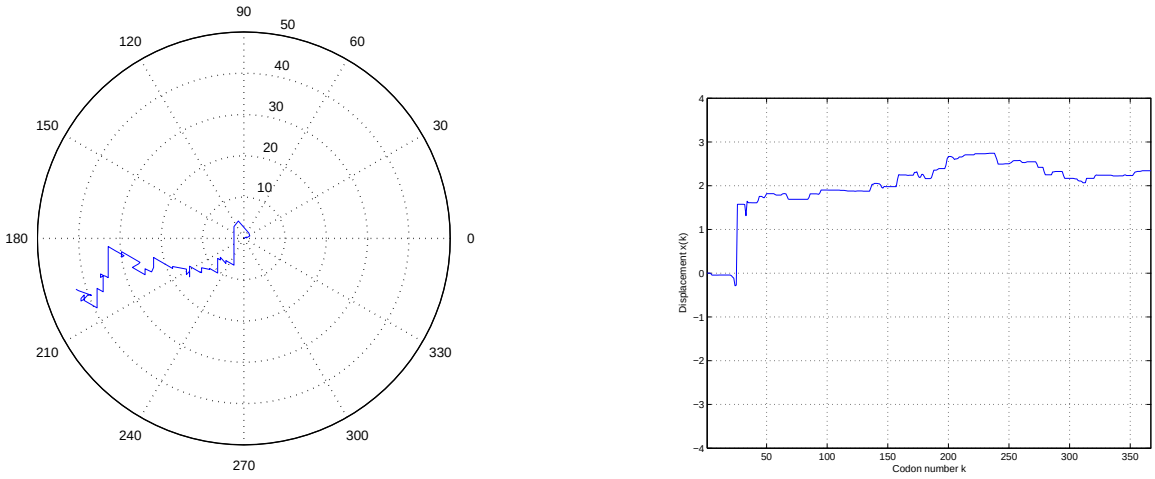


Figure 5.26: Sensitivity analysis using normal genes in *B. halodurans*

5.2.3 Bacillus subtilis



Polar plot **Displacement plot**

Figure 5.27: Analysis of gene *prfB* in *B. subtilis*

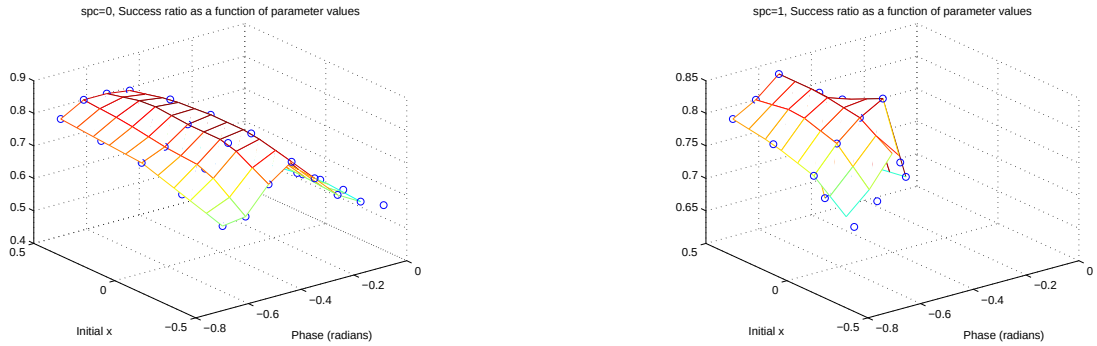
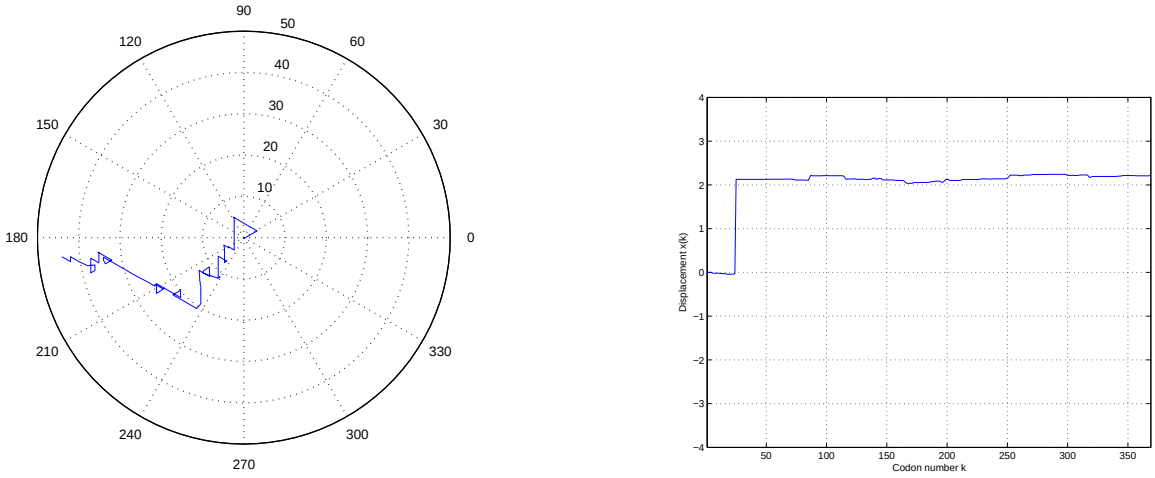


Figure 5.28: Sensitivity analysis using normal genes in *B. subtilis*

5.2.4 Chlamydophila pneumoniae



Polar plot **Displacement plot**

Figure 5.29: Analysis of gene *prfB* in *C. pneumoniae*

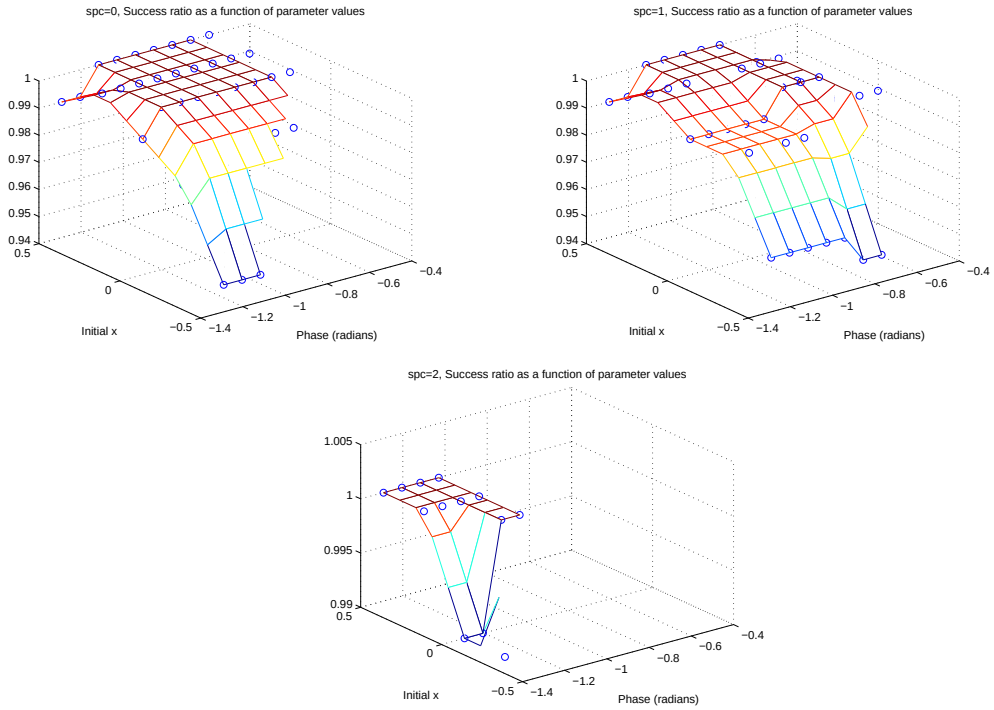
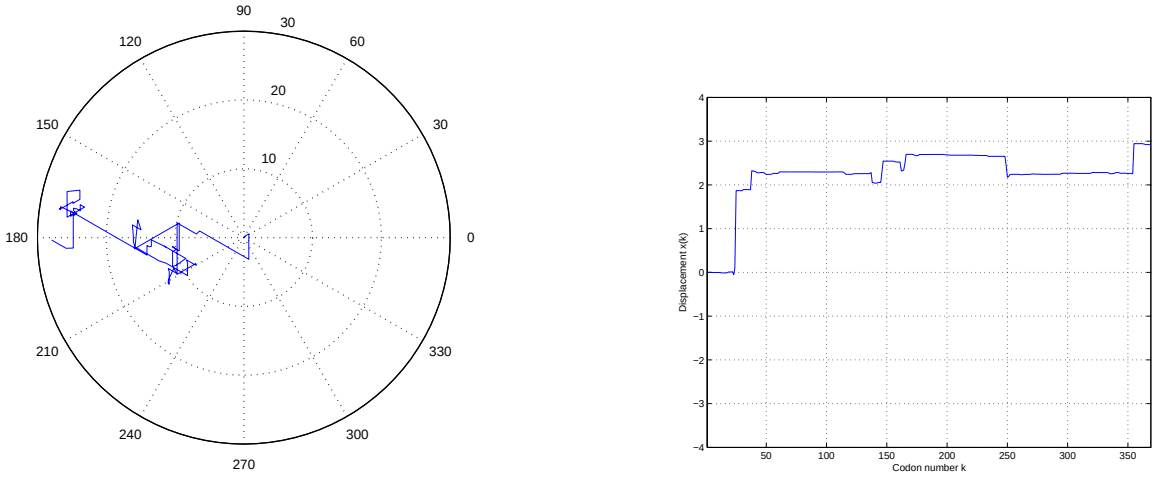


Figure 5.30: Sensitivity analysis using normal genes in *C. pneumoniae*

5.2.5 Chlamydia trachomatis



Polar plot **Displacement plot**

Figure 5.31: Analysis of gene *prfB* in *C. trachomatis*

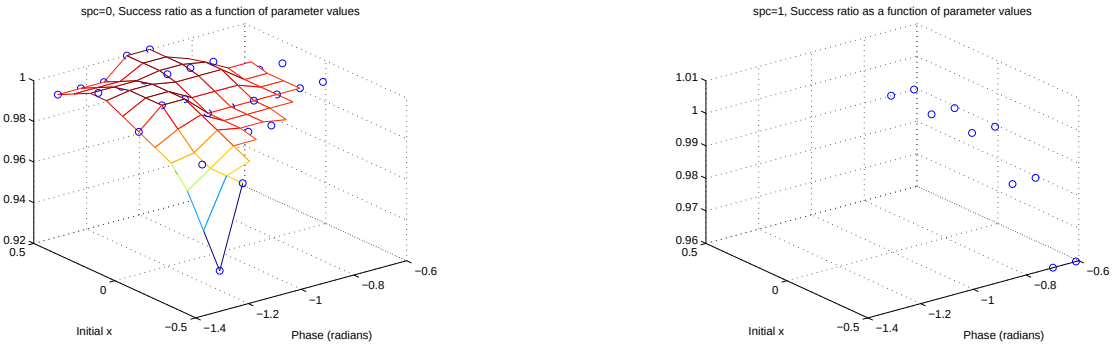
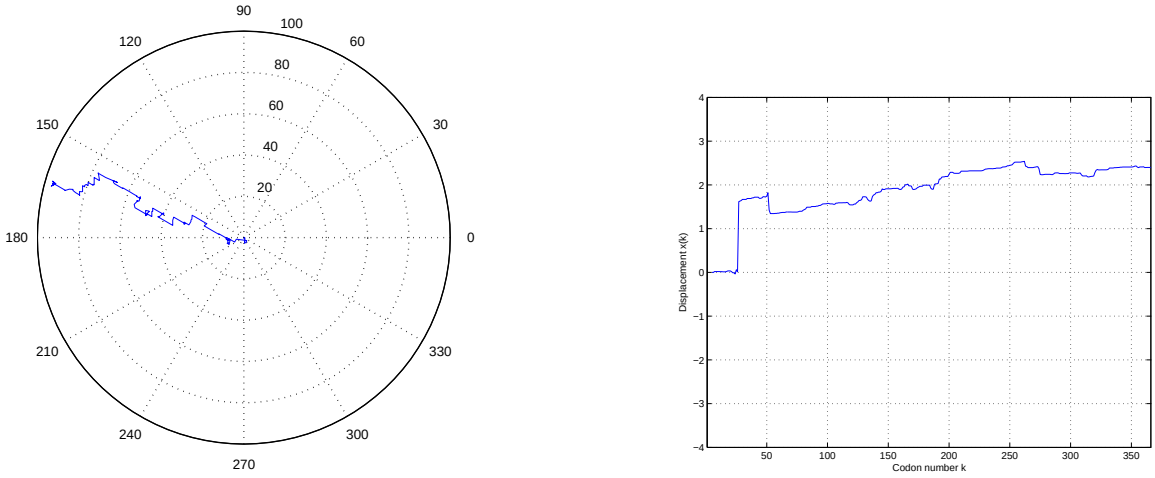


Figure 5.32: Sensitivity analysis using normal genes in *C. trachomatis*

5.2.6 Haemophilus influenzae



Polar plot **Displacement plot**

Figure 5.33: Analysis of gene *prfB* in *H. influenzae*

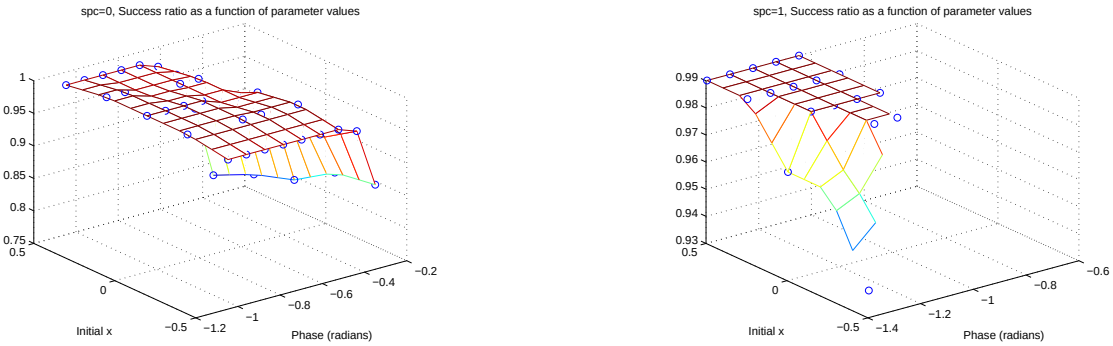
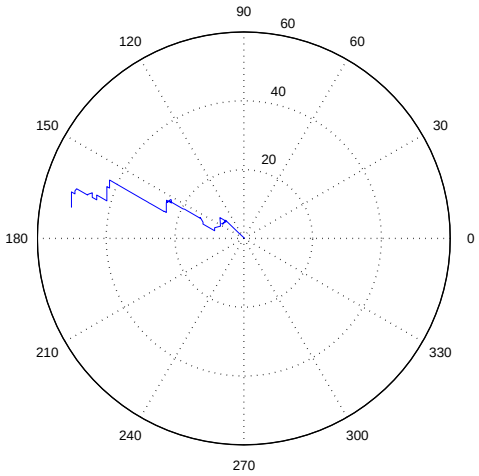
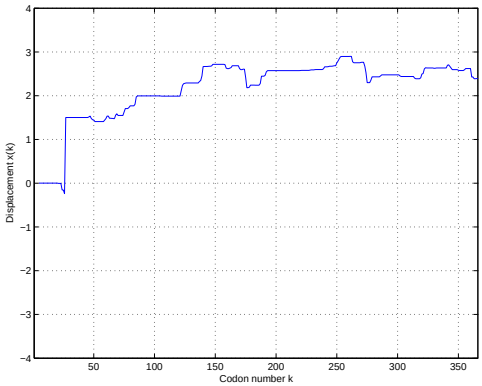


Figure 5.34: Sensitivity analysis using normal genes in *H. influenzae*

5.2.7 Pasteurella multocida



Polar plot



Displacement plot

Figure 5.35: Analysis of gene *prfB* in *P. multocida*

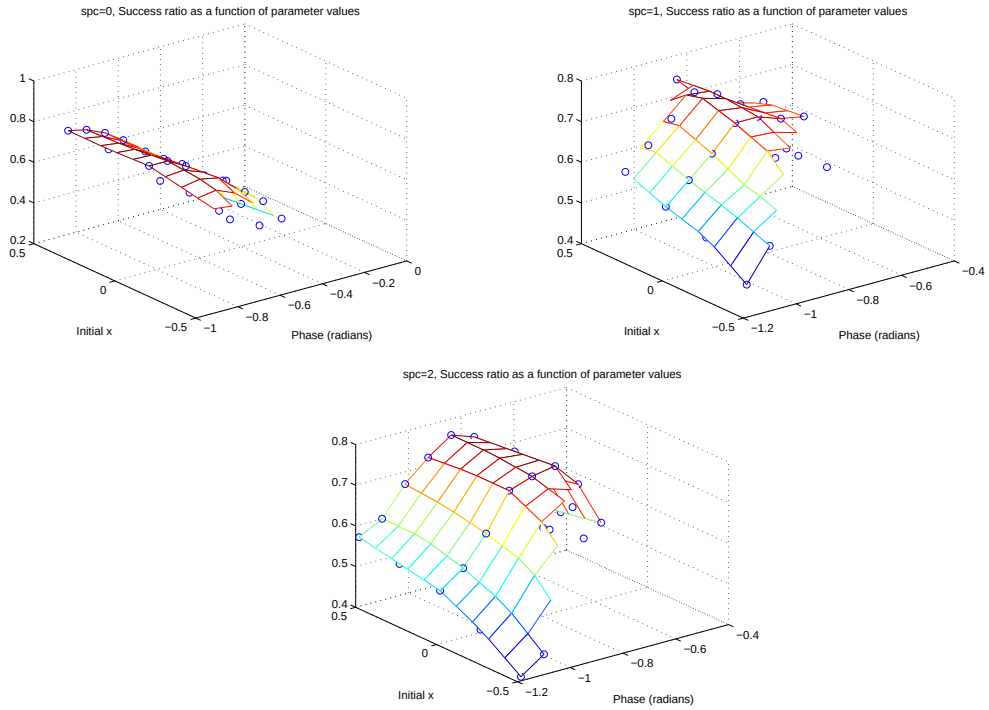
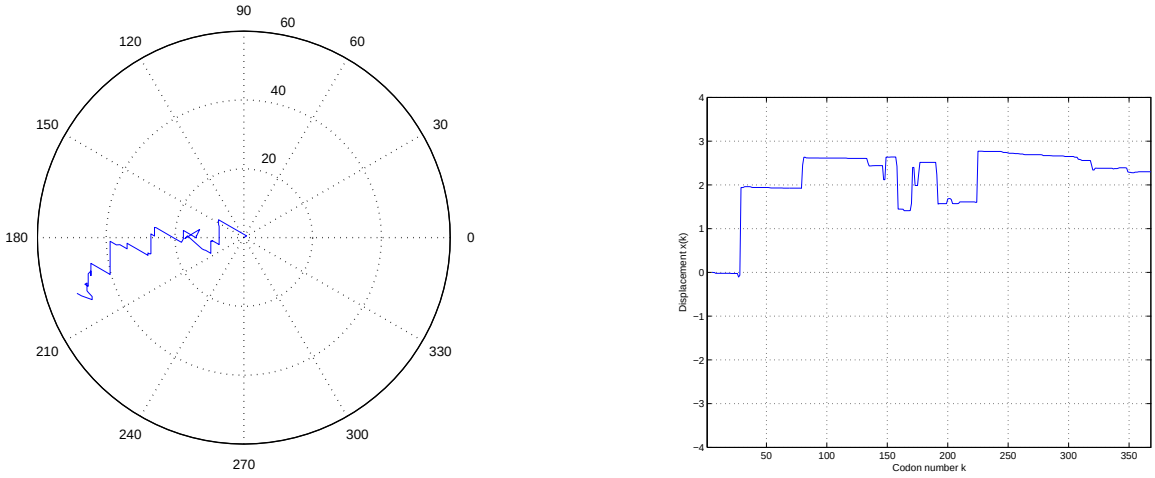


Figure 5.36: Sensitivity analysis using normal genes in *P. multocida*

5.2.8 Streptococcus mutans



Polar plot **Displacement plot**

Figure 5.37: Analysis of gene *prfB* in *S. mutans*

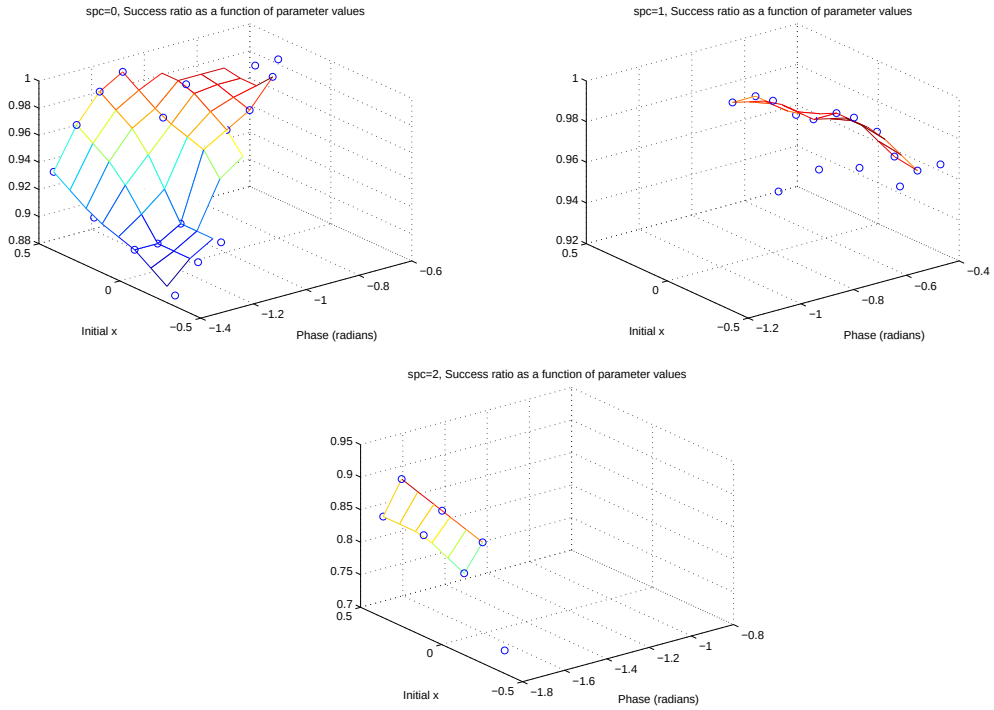
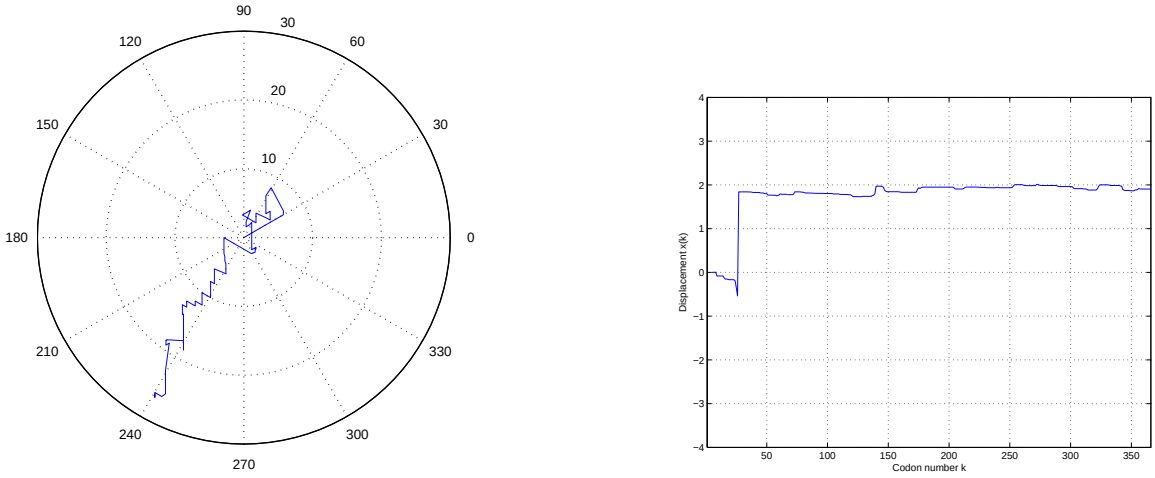


Figure 5.38: Sensitivity analysis using normal genes in *S. mutans*

5.2.9 Salmonella typhimurium



Polar plot **Displacement plot**

Figure 5.39: Analysis of gene *prfB* in *S. typhimurium*

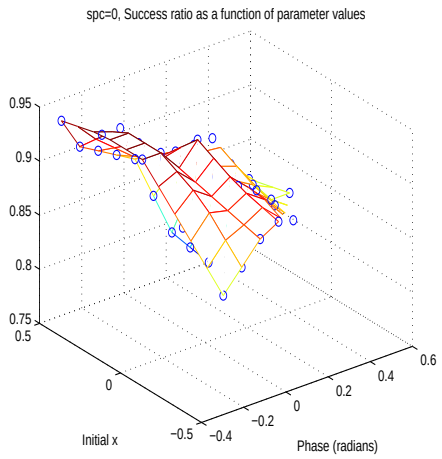
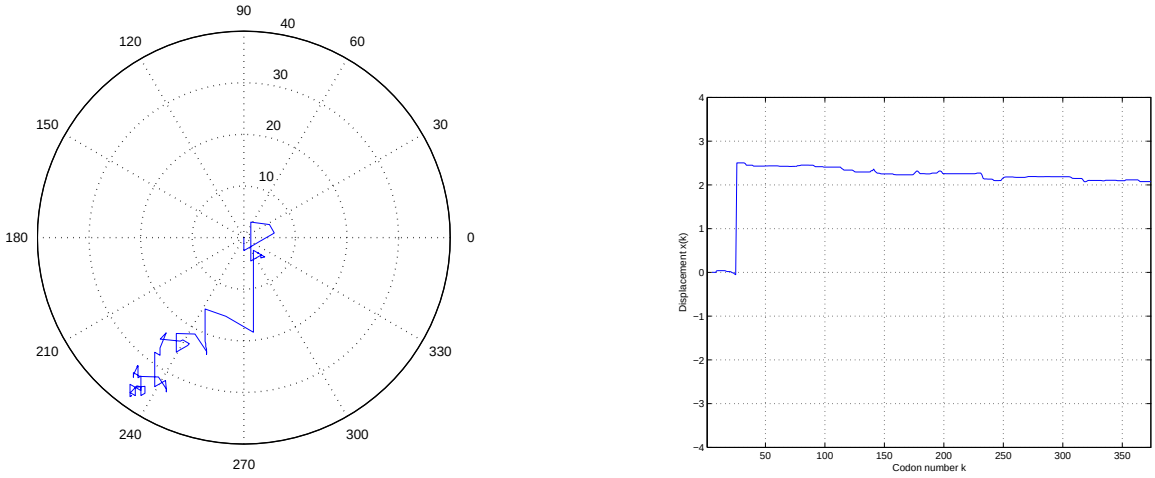


Figure 5.40: Sensitivity analysis using normal genes in *S. typhimurium*

5.2.10 Treponema pallidum



Polar plot **Displacement plot**

Figure 5.41: Analysis of gene *prfB* in *T. pallidum*

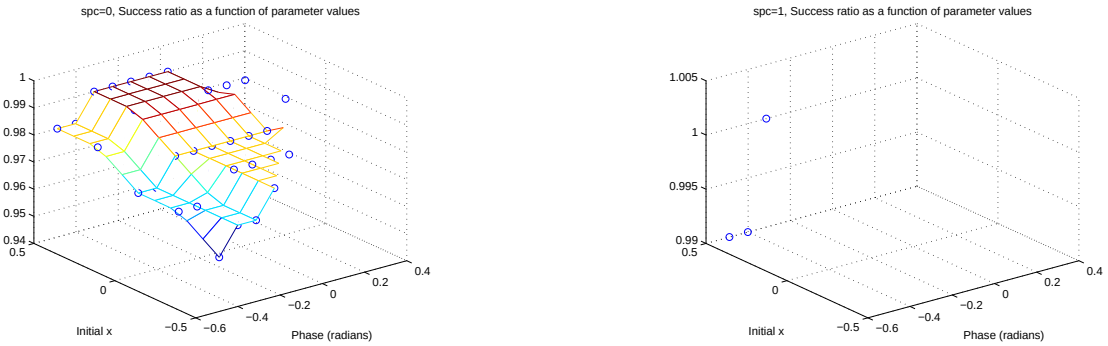
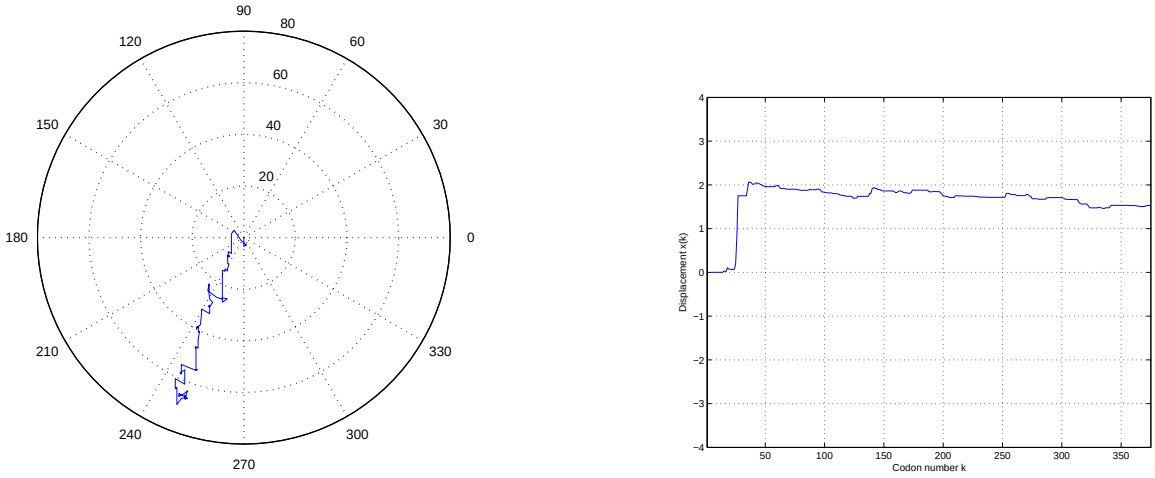


Figure 5.42: Sensitivity analysis using normal genes in *T. pallidum*

5.2.11 Xylella fastidiosa



Polar plot **Displacement plot**

Figure 5.43: Analysis of gene *prfB* in *X. fastidiosa*

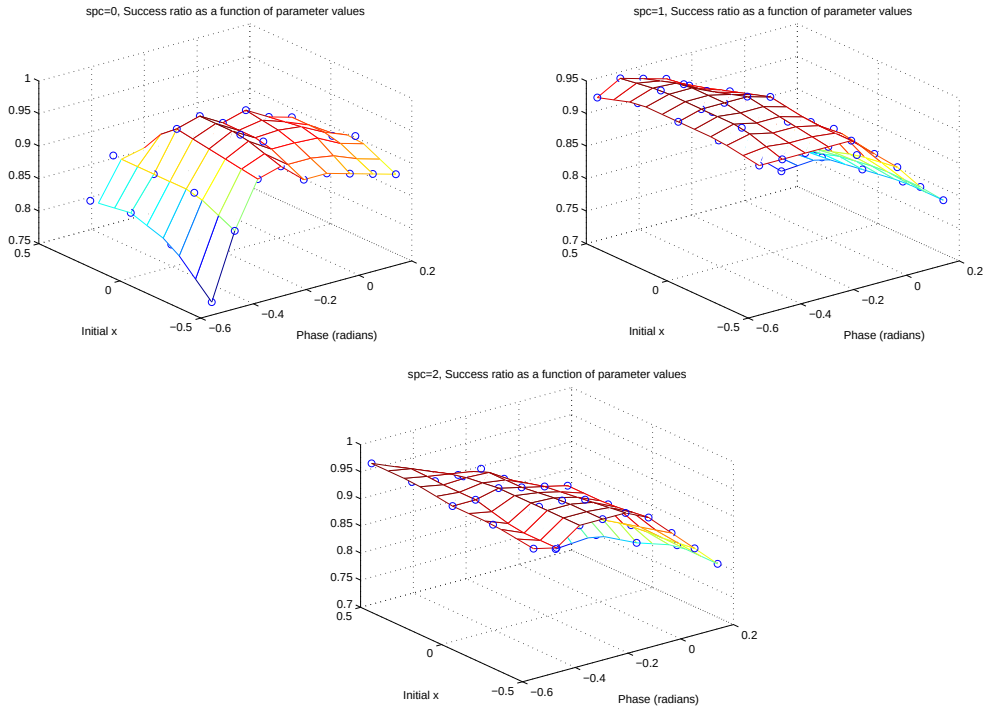


Figure 5.44: Sensitivity analysis using normal genes in *X. fastidiosa*

5.3 Discussion

From the results presented in the previous section, we observe that

- For 5 of the species, the model does not work when the spacing is changed to 2 codons. As mentioned previously, the angle-condition is not satisfied even when θ_{sp} is varied within reasonable bounds. For *S. typhimurium*, the model does not give the expected result even for a spacing of 1 codon.
- In some species (*B. halodurans*, *C. trachomatis*, *S. mutans*), the model is highly sensitive, i.e. only a few values of (θ_{sp}, x_0) would work.
- In some species, to make *prfB* work, the value of θ_{sp} needs to be decreased, while in some others, it needs to be increased (especially for spc=1,2)
- In general, the success ratio R peaks at the estimated value of θ_{sp}
- For 5 of the species, no clear concavity is observed about the value of θ_{sp}. Out of these, 3 species show a peak in success ratio at a phase θ_{sp}. For the remaining 8 species, the success ratio drops off with changing phase angle, showing the sensitivity of the model to θ_{sp}.
- The success ratio varies strongly with initial displacement in *C. pneumoniae*, *C. trachomatis*, *S. typhimurium* and *T. pallidum*. A reasonable amount of variation is seen in the remaining species, except for *B. halodurans* and *S. mutans*.

In summary, we have observed that our model is reasonably sensitive to its parameters, but also works over a wide range. Errors made in estimating the value of θ_{sp} (up to 20°) would not hamper the performance of our model greatly. The parameters work together, in the sense that small offsets in the value of one parameter may be compensated in some cases by offsets in the other. On the whole, our model is fairly robust and serves as a good indicator of frameshift tendency within bacterial genes.

Chapter 6

Conclusion

6.1 Summary

Our work defines an algorithm that simulates possible hybridization between the 3'-terminal nucleotides of the 16S rRNA and the mRNA. The algorithm revealed a periodic, free energy signal in the coding regions of the genes in a number of bacterial species [41]. Based on the ideas of Weiss *et al.* [8], Trifonov [50] and others, we hypothesized that this free energy signal could be supplying the information to modulate reading frame. Standard signal analysis methods were used to estimate the signal parameters, magnitude and phase. Signal phase was shown to be a function of coding region (G+C) content. We performed further analysis of the free energy signal, based on the hypothesis that the it could be supplying information used to maintain or shift translational reading frame.

This hypothesis is supported by three lines of prior evidence. Using statistical analysis of nucleotide frequency, Lio *et al.* [27] also detected periodicity in coding regions of both prokaryotic and eukaryotic genes. They suggested that this periodicity could function as a "reading frame correcting" signal in agreement with Grosjean *et al.* [51] and Trifonov [52]. Mutagenesis experiments of Weiss *et al.* [8] showed that a mutation which presumably decreased hybridization between the 16S rRNA tail and mRNA could alter frameshift frequency, and that a second sequence change restoring complementarity could largely restore frameshift frequency. The recent work of Yusupova *et al.* [48] has shown that the degree to which the 16S rRNA tail undergoes conformational change was correlated with the degree of complementarity between the tail and the mRNA sequence. The assumption in our

algorithm is that there is continual interaction of the 3'-terminal nucleotide tail of the 16S rRNA during translation that could produce a variable free energy pattern (signal). Signal variation would arise from the degree of complementarity between the fixed, 16S rRNA tail sequence and windows of mRNA sequence, made spatially accessible for hybridization due to the movement of the mRNA through the ribosomal complex during translation. Our approach assumes this free energy signal supplies the information needed to control reading frame: either to maintain reading frame through each elongation cycle or to change reading frame through a programmed frameshift.

Using the free energy signal we developed a mathematical model optimized to precisely predict the codon location of the frameshift site within the *prfB* coding sequence. The model is an adaptive algorithm that estimates the displacement of the ribosome from its original reading frame (Frame 0). This algorithm enables us to track the state of the ribosome-mRNA system. The physical interpretation of the differential vector, $\mathbf{D}_k$, in the model is that it represents the amount of force available at codon k to adjust the position of the mRNA. The amount of this adjustment potential that is actually realized is proportional to the time the ribosome waits for a tRNA to occupy the A-site. If the tRNA is relatively abundant, little of the adjustment is realized; if the tRNA is rare implying a long pause before the A-site is occupied, more adjustment of the mRNA relative to the ribosome occurs. The displacement x, captures the position adjustment. In a recursive form, the model starts with the previous position, derived from the energy signal for all the codons up to but not including the current codon, and uses the new displacement value to update the position, or state, of the mRNA relative to the ribosome.

In the course of developing our model, we have made several approximations and assumptions. One model assumption is that the presence of rare codons is the only factor modulating elongation rate. This assumption is consistent with Spirin [53] who asserts that the wait time due to the relative abundance of the tRNA can be assumed to be a dominating factor in inducing frameshifts. Although mRNA secondary structure is believed to result in ribosomal pausing, its absence from our model is based on the observation that a strong correlation has not been observed in all cases between mRNA secondary structure and frameshifting [54].

A second assumption concerns the proportionality between frequency of tRNA isoacceptor (calculated using Equation (4.5)) and actual tRNA availability. This proportionality is found to break down at low frequencies for genes encoding highly abundant proteins [43]. The codon bias in such genes is extreme, and this implies that the actual tRNA availability may be more than that estimated using our simple frequency calculation. This introduces a small error into the wait-time estimated using Equation (4.6). However, this small error would not significantly impact our overall results obtained by assuming that the wait-time is inversely proportional to our estimated tRNA availability. Another approximation involves the calculation of species mean phase angle θ_{sp}. We have used *all* the coding sequences annotated as "verified" in the GENBANK database, leading to a large variance in the estimate of θ_{sp}. A more confident estimate may be obtained by using genes whose authenticity has a greater degree of certainty, such as the genes studied by Link *et al.* [45]. The final approximation is the initial value of displacement. This value can be better estimated using the free energies computed for the "run-up" of the ribosome from the binding site to the start codon.

The utility of our model from the mechanistic perspective is that it suggests how both reading frame maintenance and reading frame shifts could be encoded in mRNA sequences using translational speed to modulate positional accuracy. The model captures the idea that the instantaneous component of hybridization energy, $\mathbf{D}_k$ (whose amount is a function of the mRNA sequence), is available to the ribosomal complex to adjust the position of the mRNA relative to the ribosomal decoding center by an amount that is proportional to the time required for a tRNA or release factor to fully occupy the A-site. The model implies that the codon bias of mRNAs could reflect the existence of a position-adjusting mechanism to maintain reading frame. Through codon selection, each mRNA sequence carries the information to fine-tune the position of each codon in the decoding center taking into consideration variable translational speed.

One consequence of our interpretation of the functional significance of codon bias is that it could give insight into the empirically demonstrated connection between native and recombinant protein yields and codon bias. Using the free energy signal parameters as indicators of elongation accuracy, one way to think about our model is that it yields a qualitative estimate of the frameshift tendency within a coding sequence. To the degree

that protein yield losses are determined by elongation errors, such as incorrect recruitment of tRNA, our model can show where such errors are most likely to occur in the coding sequence. Our model can also determine which possible sequence modifications would reduce the likelihood of such errors. By fitting a likelihood function to the displacement data x_k, we could quantify the "correctness" of a coding sequence for translation. These predictions would then need to be experimentally tested.

Our model also illustrates the value of applying engineering concepts to biological systems. The translation process operates with high reliability in potentially variable environments. As such, it can be considered a dynamic process in which the existence of a control system for reading frame maintenance is a reasonable engineering assumption. Mathematical modeling of control systems for dynamic processes has been the subject of considerable research [55]. Signal processing techniques have been used with considerable success to estimate the various states of a dynamic process using noisy measurements. The Kalman filter [56][57] is one of the most useful control system models. This filter uses recursive updating of the process state based on discrete sampling of input signal information. One example application is maintaining a ship's geographical position despite drift, a problem that bears some similarity to the problem faced by the ribosomal complex in maintaining reading frame.

6.2 Future Work

Our model has utility as both a tool that could be used for sequence annotation and for its implications as to the mechanism of reading frame maintenance and frameshifting. Sequence annotation is an early objective for genome sequencing projects. Frameshift sites are difficult to recognize [58] for current gene annotation programs such as GENMARK [59] and GLIMMER [60]. Our model implies that a free energy signal that is used to maintain reading frame is encoded in the coding regions of authentic genes. The existence of this signal can be visualized using either polar plots of signal phase and magnitude or in displacement plots. We are currently exploring this approach with the objective of developing an annotation program that can identify authentic coding regions and frameshift locations.

Each cycle of translation elongation requires the ribosomal complex to return to the

same "position", i.e., the positioning of the tRNA carrying the nascent polypeptide chain in the P-site. The precision of this position is critical as the P-site tRNA spatially defines the A-site boundary in the ribosomal complex [61]. The translational process must accomplish precise positioning of the P-site tRNA in the face of considerable process variation, including potentially changing environmental conditions of salt concentration, temperature, pH, and variable process components such as tRNAs and mRNA sequences. The requirement for the ribosomal complex to return to position in the face of environmental perturbations is analogous to the drift problem encountered in the ship example. In our model the equation for calculating instantaneous phase (Equation (4.8)) is analogous to the *measurement equation* of a Kalman filter, and the recursive relation (Equation (4.10)) is analogous to its *state update* equation. We have identified two states $x = 0$ and $x = 2$ corresponding to reading frames 0 and +1, respectively. The ribosome-mRNA system is shown to be stable in each of these two states, i.e., small perturbations to the state x_k arising from minor signal deviations will die out eventually. Our algorithm lays the ground work for using adaptive filtering techniques to detect frameshifts in coding sequences. The logical next step is to design an algorithm that describes the transition into the -1 frame, and thereby develop a generalized model of reading frame maintenance in bacteria.

Bibliography

[1] D. N. Frick and C. C. Richardson, "DNA primases," *Annu Rev Biochem*, vol. 70, pp. 39–80, 2001.

[2] B. C. Rymond and M. Rosbash, "Yeast pre-mRNA splicing," in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Vol. II. Gene Expression*, E. Jones, J. Pringle, and J. Broach, Eds. Cold Spring Harbor Laboratory Press, CSH, NY, 1992, pp. 143–192.

[3] J. G. Doench, C. P. Petersen, and P. A. Sharp, "siRNAs can function as miRNAs," *Genes Develop*, vol. 17, pp. 438–442, 2003.

[4] J. Shine and L. Dalgarno, "The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites," *Proc Natl Acad Sci USA*, vol. 71, no. 4, pp. 1342–1346, Apr 1974.

[5] J. A. Steitz and K. Jakes, "How ribosomes select initiator regions in mRNA: basepair formation between the 3' terminus of 16s rRNA and the mRNA during initiation of protein synthesis in Escherichia coli," *Proc Natl Acad Sci USA*, vol. 72, no. 12, pp. 4734–4738, Dec 1975.

[6] A. Hui and H. A. de Boer, "Specialized ribosome system: preferential translocation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli," *Proc Natl Acad Sci USA*, vol. 84, pp. 4762–4766, 1987.

[7] R. B. Weiss, D. M. Dunn, J. F. Atkins, and R. F. Gesteland, "Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting," in *Cold Spring Harb Symp Quant Biol*, vol. 52, 1987, pp. 687–693.

[8] R. B. Weiss, D. M. Dunn, A. E. Dahlberg, J. F. Atkins, and R. F. Gesteland, "Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli," *EMBO J*, vol. 7, no. 5, pp. 1503–1507, 1988.

[9] J. D. Starmer, "Free2Bind: Tools for computing minimum free energy binding between two separate RNA molecules," http://sourceforge.net/projects/free2bind/.

[10] J. Starmer, A. Stomp, M. Vouk, and D. Bitzer, "Predicting Shine-Dalgarno sequence locations exposes genome annotation errors," *PLoS Computat Biol*, vol. 2, no. 5, e57 DOI: 10.1371/journal.pcbi.0020057, 2006.

[11] T. Xia, J. SantaLucia Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs," *Biochemistry*, vol. 37, no. 42, pp. 14 719–14 735, Oct 20 1998.

[12] J. A. Jaeger, D. H. Turner, and M. Zuker, "Improved predictions of secondary structures for RNA," *Proc Natl Acad Sci USA*, vol. 86, no. 20, pp. 7706–7710, Oct 1989.

[13] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J Mol Biol*, vol. 288, no. 5, pp. 911–940, May 1999.

[14] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath, "Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution," *Cell*, vol. 102, no. 5, pp. 615–23, 2000.

[15] G. Z. Yusupova, M. M. Yusupov, J. H. Cate, and H. F. Noller, "The path of messenger RNA through the ribosome," *Cell*, vol. 106, no. 2, pp. 233–241, Jul 27 2001.

[16] T. Schurr, E. Nadir, and H. Margalit, "Identification and characterization of E.coli ribosomal binding sites by free energy computation," *Nucleic Acids Res*, vol. 21, no. 17, pp. 4019–4023, Aug 25 1993.

[17] Y. Osada, R. Saito, and M. Tomita, "Analysis of base-pairing potentials between 16s rRNA and 5' UTR for translation initiation in various prokaryotes," *Bioinformatics*, vol. 15, no. 7-8, pp. 578–581, Jul-Aug 1999.

[18] T. A. Thanaraj and M. W. Pandit, "An additional ribosome-binding site on mRNA of highly expressed genes and a bifunctional site on the colicin fragment of 16S rRNA from Escherichia coli: important determinants of the efficiency of translation-initiation," *Nucleic Acids Res*, vol. 17, no. 8, pp. 2973–2985, 1989.

[19] G. Lithwick and H. Margalit, "Hierarchy of sequence-dependent features associated with prokaryotic translation," *Genome Res*, vol. 13, no. 12, pp. 2665–2673, Dec 2003.

[20] K. Lee, C. Holland-Staley, and P. Cunningham, "Genetic analysis of Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations," *RNA*, vol. 2, no. 12, pp. 1270–1285, Dec 1996.

[21] A. Komarova, L. Tchufitsova, E. Supina, and I. Boni, "Extensive complementarity of the Shine-Dalgarno region and the 3'-end of 16S rRNA is inefficient for translation in vivo," *Russian Journal of Bioorganic Chemistry*, vol. 27, no. 4, pp. 248–255, 2001.

[22] J. Ma, A. Campbell, and S. Karlin, "Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures," *J Bacteriol*, vol. 184, pp. 5733–5745, 2002.

[23] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, 1st ed. Prentice-Hall, 1975.

[24] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. Springer-Verlag, New York, 1991.

[25] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. I: Estimation Theory.* Prentice-Hall, 1993.

[26] J. C. Brocklebank and D. A. Dickey, *SAS for Forecasting Time Series*, 2nd ed. Wiley-SAS, 2003.

[27] P. Lió, S. Ruffo, and M. Buiatti, "Third codon $G + C$ periodicity as a possible signal for an "internal" selective constraint," *J Theor Biol*, vol. 171, no. 2, pp. 215–223, Nov 21 1994.

[28] G. D'Onofrio and G. Bernardi, "A universal compositional correlation among codon positions," *Gene*, vol. 110, no. 1, pp. 81–88, Jan 2 1992.

[29] G. Cocho and J. L. Rius, "Structural constraints and gene dynamics," *Riv Biol*, vol. 82, no. 3-4, pp. 344–345, 416–417, 1989.

[30] M. Gouy and C. Gautier, "Codon usage in bacteria: correlation with gene expressivity," *Nucleic Acids Res*, vol. 10, no. 22, pp. 7055–7074, Nov 25 1982.

[31] E. N. Trifonov, "Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences," *J Mol Biol*, vol. 194, pp. 643–652, 1987.

[32] J. F. Atkins, R. B. Weiss, S. Thompson, and R. F. Gesteland, "Towards a genetic dissection of the basis of triplet decoding, and its natural subversion: programmed reading frame shifts and hops," *Annu Rev Genet*, vol. 25, pp. 201–228, 1991.

[33] J. F. Atkins, R. B. Weiss, and R. F. Gesteland, "Ribosome gymnastics - degree of difficulty 9.5, style 10.0," *Cell*, vol. 62, pp. 413–423, Aug 1990.

[34] D. C. Giancoli, *Physics for Scientists and Engineers.* Prentice Hall, 1989.

[35] P. J. Farabaugh, "Programmed translational frameshifting," *Microbiol Rev*, vol. 60, no. 1, pp. 103–134, Mar 1996.

[36] L. Ponnala, D. L. Bitzer, A. Stomp, and M. A. Vouk, "A computational model for reading frame maintenance," in *Proceedings of the 28th IEEE EMBS Annual International Conference.* IEEE, Aug 30-Sep 3, New York City, USA 2006, pp. 4540–4543, iSBN: 14244-0033-3.

[37] B. E. Schoner, H. M. Hsiung, R. M. Belagaje, N. G. Mayne, and R. G. Schoner, "Role of mRNA translational efficiency in bovine growth hormone expression in Escherichia coli," *Proc Natl Acad Sci USA*, vol. 81, pp. 5403–5407, Sep 1984.

[38] P. V. Baranov, O. L. Gurvich, O. Fayet, M. F. Prere, W. A. Miller, R. F. Gesteland, J. F. Atkins, and M. C. Giddings, "RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression," *Nucleic Acids Res*, vol. 29, no. 1, pp. 264–267, 2001.

[39] P. V. Baranov, O. L. Gurvich, A. W. Hammer, R. F. Gesteland, and J. F. Atkins, "RECODE 2003," *Nucleic Acids Res*, vol. 31, no. 1, pp. 87–89, 2003.

[40] W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*, 4th ed. Brooks/Cole Publishing Company, 1999.

[41] L. Ponnala, A.-M. Stomp, D. L. Bitzer, and M. A. Vouk, "Analysis of free energy signals arising from nucleotide hybridization between rrna and mrna sequences during translation in eubacteria," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2006, pp. Article ID 23 613, 9 pages, 2006, doi:10.1155/BSB/2006/23613.

[42] J. Sipley and E. Goldman, "Increased ribosomal accuracy increases a programmed translational frameshift in Escherichia coli." *Proc Natl Acad Sci USA*, vol. 90, no. 6, p. 23152319, March 15 1993.

[43] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Mol Biol Evol*, vol. 2, no. 1, pp. 13–34, Jan 1985.

[44] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering.* Perseus Books, Cambridge MA, 1994.

[45] A. J. Link, K. Robison, and G. Church, "Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli," *Electrophoresis*, vol. 18, pp. 1259–1313, 1997.

[46] H. Gao, J. Sengupta, M. Valle, A. Korostelev, N. Eswar, S. M. Stagg, P. Van Roey, R. K. Agrawal, S. C. Harvey, A. Sali, M. S. Chapman, and J. Frank, "Study of the structural dynamics of the E.coli 70S ribosome using real-space refinement," *Cell*, vol. 113, no. 6, pp. 789–801, Jun 13 2003.

[47] G. M. Culver, "Meanderings of the mRNA through the ribosome," *Structure (Camb)*, vol. 9, no. 9, pp. 751–758, Sep 2001.

[48] G. Yusupova, L. Jenner, B. Rees, D. Moras, and M. Yusupov, "Structural basis for messenger RNA movement on the ribosome," *Nature*, vol. 444, pp. 391–394, Nov 2006.

[49] J. M. Ogle, A. P. Carter, and V. Ramakrishnan, "Insights into the decoding mechanism from recent ribosome structures," *Trends Biochem Sci*, vol. 28, no. 5, pp. 259–266, May 2003.

[50] E. N. Trifonov, "Recognition of correct reading frame by the ribosome," *Biochimie*, vol. 74, no. 4, pp. 357–362, Apr 1992.

[51] H. J. Grosjean, S. De Henau, and D. M. Crothers, "On the physical basis for ambiguity in genetic coding interactions," *Proc Natl Acad Sci USA*, vol. 75, pp. 610–614, 1978.

[52] E. Trifonov, "The multiple codes of nucleotide sequences," *Bull Math Bio*, vol. 51, no. 4, pp. 417–432, 1989.

[53] A. S. Spirin, *Ribosomes*. Springer, 1999.

[54] H. Kontos, S. Napthine, and I. Brierley, "Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency," *Mol Cell Biol*, vol. 21, no. 24, pp. 8657–8670, Dec 2001.

[55] P. S. Maybeck, *Stochastic models, estimation, and control*, ser. Mathematics in Science and Engineering, 1979, vol. 141.

[56] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82 (Series D), pp. 35–45, March 1960.

[57] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, 2nd ed. John Wiley and Sons, Inc., 1992.

[58] S. Moon, Y. Byun, H.-J. Kim, S. Jeong, and K. Han, "Predicting genes expressed via -1 and +1 frameshifts," *Nucleic Acids Res*, vol. 32, no. 16, pp. 4884–4892, 2004.

[59] M. Borodovsky and J. McIninch, "GENMARK: Parallel gene recognition for both dna strands," *Computers Chem.*, vol. 17, no. 19, pp. 123–133, 1993.

[60] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Research*, vol. 27, no. 23, pp. 4636–4641, 1999.

[61] P. V. Baranov, R. F. Gesteland, and J. F. Atkins, "P-site tRNA is a crucial initiator of ribosomal frameshifting," *RNA*, vol. 10, pp. 221–230, 2004.

APPENDICES

Appendix A

Extracting sequences

A.1 GENBANK

This section describes the method of downloading sequences from GENBANK (illustrated using *E. coli* as an example). It is highly recommended to use Internet Explorer (IE7) for the following tasks:

- Searching for the GENBANK accession number (NC_000913 for *E. coli*) on the NCBI homepage (`http://www.ncbi.nlm.nih.gov/`) should take you to a page displaying database-wide search results. Click on the "Genome" link to go to the description page, and then click on the accession number to go to the summary page.
- Use the "RefSeq" link in the **Genome Info** column on the summary page to get the complete sequence. If sequence not displayed, uncheck the hide boxes and click the refresh button on the page.
- Download the GENBANK data (`http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi??db=nucleotide&val=NC_000913`) into a text file. Use the *File*→*SaveAs* (or *Page* → *SaveAs* in IE7) option to save as type: Text file (*.txt).
- Remove stuff from the top (before LOCUS) and bottom (after //) and save it as a .gbk file (Ecoli.gbk).
- Copy the sequence portion (after ORIGIN and before //) and save it as the sequence file (Ecoli_seqfile.txt)

- Go back to the summary page, and click on the "Protein coding" link in the **Features** column.
- Copy the entire protein coding table temporarily into an Excel file. If using IE7, just right-click on the top-left corner of the web-table and choose "Export to Microsoft Excel". If the right-most column of the table (the one containing diamonds) appears in the Excel file, just delete it. Make sure to remove any column-headings - the first row of the file should contain protein annotation, not stuff like "Protein Name", "Start", "End", etc. An easy way to remove the first row is to right-click and select *Delete→Entire row*. Save the resulting file in Excel95 format (Ecoli_95.xls) by selecting *Save as type* "Microsoft Excel 5.0/95 Workbook". If an annoying pop-up box appears asking if you really want to save in this format (and warning you that some features might be lost if you do), just click "Yes".
- Copy the structural RNA table into an Excel file. Follow the same instructions as for the protein coding table.

A.2 RECODE

In what follows,

- GENBANK refers to the list of complete microbial genomes (`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`)
- RECODE refers to the list of +1 frameshifters (`http://recode.genetics.utah.edu/type.cfm`)

There are 48 cases of +1 frameshift among prokaryota listed in RECODE, spread across 47 distinct species. *E. coli* is the only species that has two cases of +1 frameshift, *argI* and *prfB*). The remaining 46 species have *prfB* alone. Out of these 47 species:

- 25 have unique name-matches in GENBANK, out of which
  - 20 have matching accession numbers in RECODE and GENBANK
  - 3 do not have an accession number listed in RECODE
  - 2 have an accession number mismatch between RECODE and GENBANK

- 12 have multiple name-matches in GENBANK, out of which
  - 8 have matching accession numbers in RECODE and GENBANK
  - 1 has 3 matching accession numbers in RECODE and GENBANK
  - 1 links to a species with a different name
  - 2 have an accession number mismatch between RECODE and GENBANK
- 10 do not have name-matches in GENBANK

In our analysis, we use those species that have matching accession numbers in RECODE and GENBANK. For each of these species, the listed *prfB* sequence is copied manually into a text file (prfB_seq_RECODE.txt).

The RECODE database lists only the AE number for each species, not the NC_xxxxxx number. Click on the AE number to go to GENBANK, and make sure the name of the species matches completely with the species whose NC_xxxxxx data you are using.

Appendix B

Calculating signals using free2bind

The software to calculate free energy signals is written and maintained by J. D. Starmer [9][10], and can be downloaded from `http://sourceforge.net/projects/free2bind/`.

The *free_scan* program has certain output options that the user can specify. For analysis of several free energy signals, it is best to have one signal per line in a text file. This requirement can be specified on the command line as follows:

```
>> ./free_scan.pl -r -e -q auuccuccacuag EcoliSeq.fasta
                                    EcoliIndexData.txt >  EcoliSignals.txt
```

The output file *EcoliSignals.txt* would contain the signals corresponding to individual genes on separate rows. The detailed procedure is as follows:

- Copy the .gbk file (Ecoli.gbk) and the index file (EcoliIndexData) to the directory where *free_scan.pl* is stored
- Run *gbk2fasta.pl* to extract the sequence in FASTA format (command: `./gbk2fasta.pl Ecoli.gbk > EcoliSeq.fasta`)
- Run *free_scan.pl* to calculate signals in one-signal-per-row format. To calculate signals with a preset of 30, use the command:

  ```
  ./free_scan.pl -r -e -q -L 30 auuccuccacuag EcoliSeq.fasta
  EcoliIndexData > EcoliSignals.txt
  ```

Table B.1: Signal preset to be used for MATLAB-analysis, indicates number of points before position zero as per signal indexing scheme

Tail Length	*PERL command*	*Signal Preset*
8	$-L$ 30	23
11	$-L$ 30	20
13	$-L$ 30	18
n	$-L$ 30	$30-(n-1)$

To use FREIER energy values, give the command:

```
./free_scan.pl -r -e -q -L 30 -p FREIER auuccuccacuag EcoliSeq.fasta
EcoliIndexData > Ecoli_FREIER_Signals.txt
```

To know more about sequence presets and signal presets, see Table B.1.

- Remove the top line in the output file EcoliSignals.txt
- Delete: Ecoli.gbk, EcoliIndexData

Appendix C

Power of detection test

A free energy signal is obtained from the binding pattern of the ribosome tail with the mRNA sequence. The periodogram of such a signal shows activity at frequency $f = 1/3$. We wish to formally test for the presence of a sinusoid with specified frequency in such free energy signals. We use a statistical hypothesis test to detect the presence of periodicity and analyze the power of the test as a function of signal-to-noise ratio (SNR) and signal length (L).

We have verified that the signal is stationary, by doing the Augmented Dickey-Fuller Test [26]. This allows us to analyze the periodogram of the signal. Based on the statistical properties of the periodogram, we perform a test for the presence of a sinusoid of frequency $f = 1/3$ (see [24], pp 324-325 for test description, pp 323 for harmonic decomposition of the sum-of-squares).

We create a pure sinusoid of frequency $f = 1/3$ with length $L = 1023$. We add a fixed amount of noise to it, and produce a set of N signals, all having the same signal-to-noise ratio (SNR). We subject all of them to the periodicity test using a fixed value of α and determine the number of signals that pass the test i.e. show periodicity at $f = 1/3$, say N_p. The probability of correct detection P_D is calculated as

$$P_D = N_p/N$$

We remove the periodic component and repeat the procedure. Let N_0 be the number of

Table C.1: Detection results using $\alpha = 0.05$, $N = 1000$

SNR	P_D	P_F
-5	1.0	0.05
-10	1.0	0.05
-15	1.0	0.05
-18	0.96	0.05
-20	0.82	0.05

Table C.2: Detection results using $\alpha = 0.01$, $N = 1000$

SNR	P_D	P_F
-5	1.0	0.01
-10	1.0	0.01
-15	0.995	0.01
-18	0.88	0.01
-20	0.63	0.01

genes that show periodicity at $f = 1/3$. The probability of false alarm P_F is calculated as

$$P_F = N_0/N$$

We repeat the procedure using varying levels of SNR and α. The results are summarized in Table C.1 and Table C.2.

The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. In other words, the power of a hypothesis test is the probability of not committing a type II error. It is calculated by subtracting the probability of a type II error from 1, usually expressed as:

$$Power = 1 - P(typeIIerror) = (1 - \beta)$$

The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1.

We perform a set of simulations (Pass1) to calculate the power of our hypothesis test under varying values of SNR and L, at $\alpha = 0.05$. The typical SNR for the free-energy signals in E.coli K-12 is about $-18dB$. We only choose signals that are greater than 300 codons, i.e., 900 nucleotides for our experiments. Simulations show that at these parameter

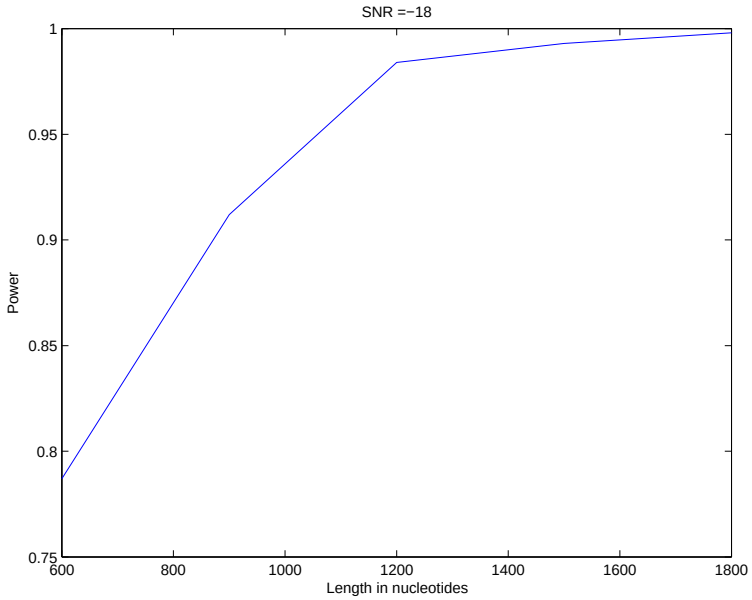


Figure C.1: (Pass1) Power Vs. Length at SNR = -18 dB

values, the power of the hypothesis test is 0.912. This is assumed to be good enough for our experiment.

Another run using more replications with closer-spaced lengths was performed (Pass2).

We find that our statistical test is a fairly powerful one for signals longer than 900 points, i.e. for genes longer than 300 codons.

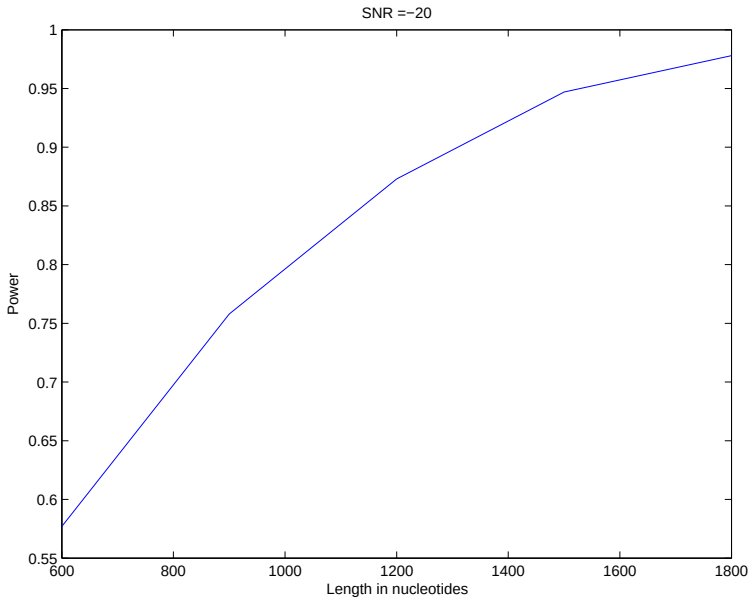


Figure C.2: (Pass1) Power Vs. Length at SNR = -20 dB

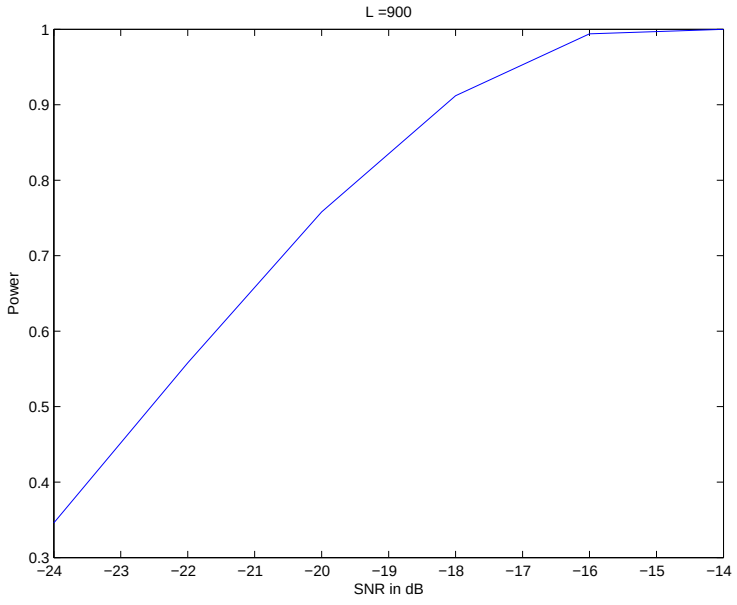


Figure C.3: (Pass1) Power Vs. SNR at Length = 900 points

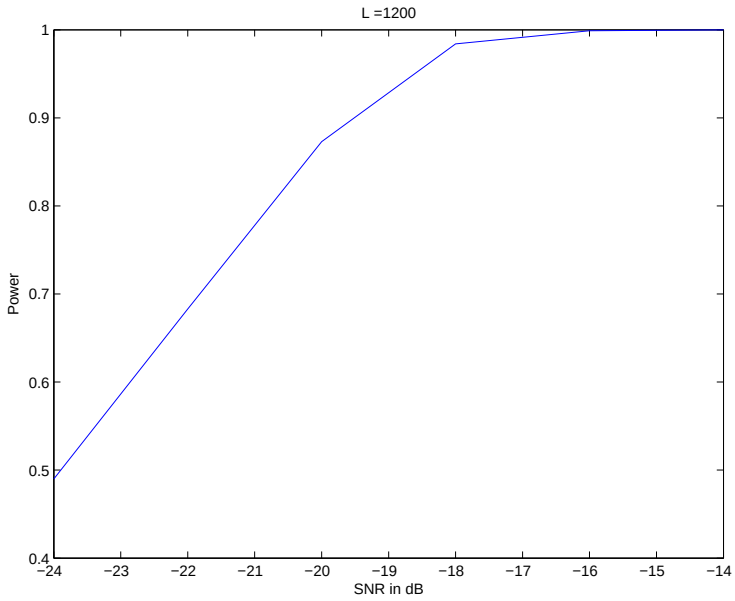


Figure C.4: (Pass1) Power Vs. SNR at Length = 1200 points

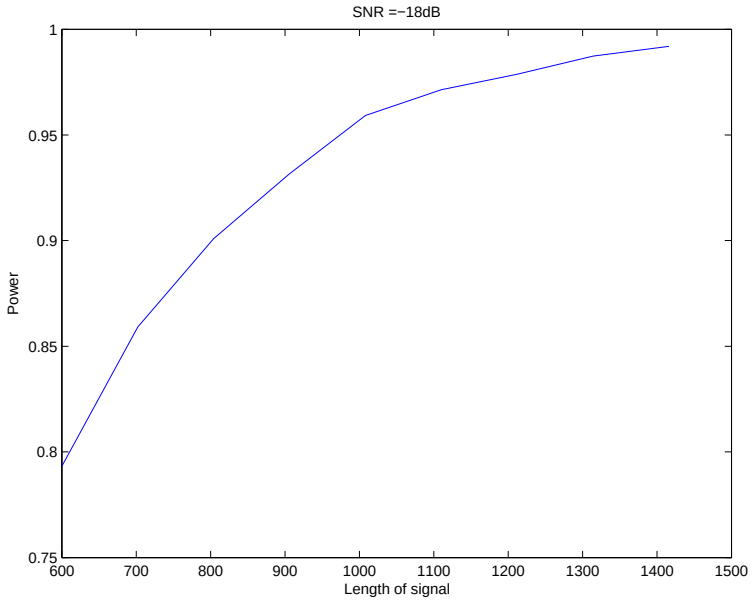


Figure C.5: (Pass2) Power Vs. Length at SNR = -18 dB

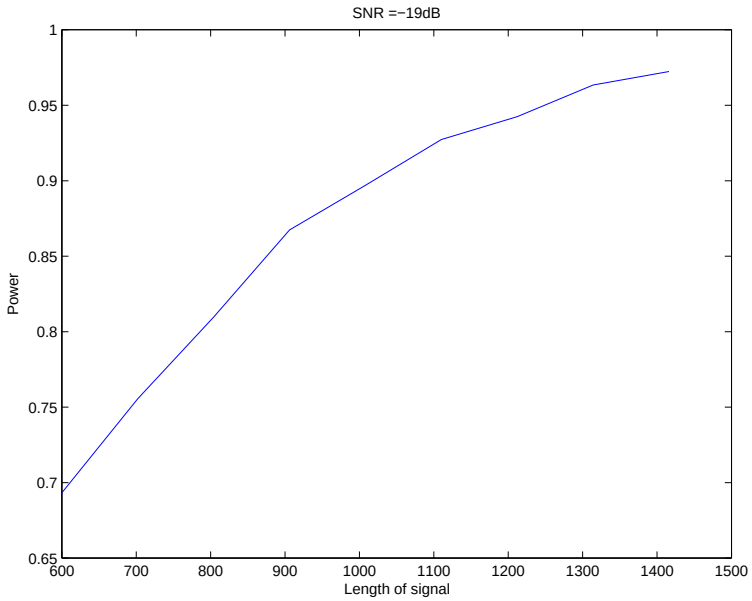


Figure C.6: (Pass2) Power Vs. Length at SNR = -19 dB

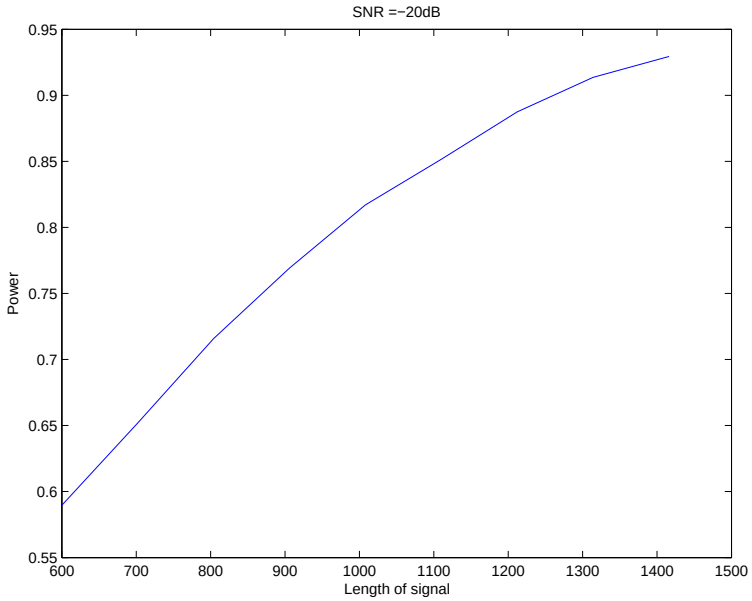


Figure C.7: (Pass2) Power Vs. Length at SNR = -20 dB

Appendix D

Signal Parameter Estimation

We are interested in estimating the phase and magnitude of a sinusoid corrupted by additive noise. We will describe statistical methods of estimation and give the bias and variance of the estimates as a function of the noise. We will rely heavily on material in Chapters 3, 7 and 8 of [25].

We will henceforth refer to the *pure* sinusoid as the "signal". All our experiments will be done using a signal of fixed frequency $f_0 = 1/3$. The following assumptions seem reasonable:

- The noise is IID, and has variance σ^2
- The noise is independent of the signal

So, we represent the model as

$$y[n] = \mu + Asin(2\pi(1/3)n + \phi) + w[n] \qquad n = 0, 1, ...(N-1) \tag{D.1}$$

where we are interested in estimating magnitude A and phase ϕ. The estimates of these quantities will be denoted by $\hat{A}$ and $\hat{\phi}$ respectively. We have explored the following methods:

1. Maximum likelihood
2. Least Squares

The concept of *identifiability* is important here. For our estimation techniques to work, we must have $A > 0$ and $0 < f_0 < 1/2$ (see pp 56 of [25] for explanation). The frequency condition is satisfied in our case since $f_0 = 1/3$, but the amplitude condition may not be if the model is incorrect, i.e. if we use *sin* instead of a *cos* or vice-versa.

D.1 Maximum Likelihood Estimation

Almost all practical estimators are based on the maximum likelihood principle. We now derive the MLE for amplitude and phase of a sinusoid. The derivation is based on Example 7.16 of [25], but modified for the *sin* model with fixed frequency f_0 and non-zero mean μ.

Assume that the noise $w[n]$ has a normal distribution, i.e. $\mathbf{w} \sim N(0, \mathbf{C})$ where $\mathbf{C} = \sigma^2\mathbf{I}$. The PDF of the data $\mathbf{y}$ is given by

$$p(\mathbf{y}; A, \phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(y[n] - \mu - Asin(2\pi(1/3)n + \phi))^2\right]$$

The MLE of amplitude A and phase ϕ is found by minimizing

$$J(A, \phi) = \sum_{n=0}^{N-1}(y[n] - \mu - Asin(2\pi(1/3)n + \phi))^2$$

We first expand the sine to yield

$$J(A, \phi) = \sum_{n=0}^{N-1}(y[n] - \mu - (Acos\phi)sin(2\pi(1/3)n) + (Asin\phi)cos(2\pi(1/3)n))^2$$

We may transform J to a quadratic function by letting

$$\alpha_1 = Acos\phi, \quad \alpha_2 = Asin\phi$$

which is a one-to-one transformation. The inverse transformation is given by

$$A = \sqrt{\alpha_1^2 + \alpha_2^2}, \quad \phi = \arctan(\frac{\alpha_2}{\alpha_1}) \tag{D.2}$$

Also, let

$$\mathbf{s} = [0 \; sin(2\pi(1/3)) \; \ldots \; sin(2\pi(1/3)(N-1))]^T$$

$$\mathbf{c} = [1 \; cos(2\pi(1/3)) \; \ldots \; cos(2\pi(1/3)(N-1))]^T$$

$$\mathbf{1} = [1 \; 1 \; 1 \; \ldots \; 1]^T$$

Then, we have

$$J(\alpha_1, \alpha_2) = (\mathbf{y} - \mu\mathbf{1} - \alpha_1\mathbf{s} - \alpha_2\mathbf{c})^T(\mathbf{y} - \mu\mathbf{1} - \alpha_1\mathbf{s} - \alpha_2\mathbf{c})$$

$$J(\alpha_1, \alpha_2) = (\mathbf{y} - \mathbf{H}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{H}\boldsymbol{\alpha})$$

where $\boldsymbol{\alpha} = [\mu \; \alpha_1 \; \alpha_2]^T$ and $\mathbf{H} = [\mathbf{1} \; \mathbf{s} \; \mathbf{c}]$. The minimizing solution is given by

$$\hat{\boldsymbol{\alpha}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{y}$$

Substituting $\mathbf{C} = \sigma^2\mathbf{I}$, we get

$$\hat{\boldsymbol{\alpha}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$$

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{s}^T \\ \mathbf{c}^T \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{s} & \mathbf{c} \end{bmatrix}$$

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{s} & \mathbf{1}^T\mathbf{c} \\ \mathbf{s}^T\mathbf{1} & \mathbf{s}^T\mathbf{s} & \mathbf{s}^T\mathbf{c} \\ \mathbf{c}^T\mathbf{1} & \mathbf{c}^T\mathbf{s} & \mathbf{c}^T\mathbf{c} \end{bmatrix}$$

An *approximate* MLE can be obtained in this case since f_0 is not near 0 or 1/2. We have

$$\frac{1}{N}\mathbf{1}^T\mathbf{s} \approx 0, \quad \frac{1}{N}\mathbf{1}^T\mathbf{c} \approx 0, \quad \frac{1}{N}\mathbf{c}^T\mathbf{s} \approx 0$$

$$\frac{1}{N}\mathbf{c}^T\mathbf{c} \approx 1/2, \quad \frac{1}{N}\mathbf{s}^T\mathbf{s} \approx 1/2$$

Note that the above approximations become exact when the length of the signal is an integral number of cycles, i.e. $N = 3p$ (in general, for $N = p/f_0$), where p is a positive integer. We now have

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} N & 0 & 0 \\ 0 & N/2 & 0 \\ 0 & 0 & N/2 \end{bmatrix}$$

$$(\mathbf{H}^T\mathbf{H})^{-1} = \begin{bmatrix} 1/N & 0 & 0 \\ 0 & 2/N & 0 \\ 0 & 0 & 2/N \end{bmatrix}$$

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \frac{1}{N}\sum_{n=0}^{N-1} y\,[n] \\ \frac{2}{N}\sum_{n=0}^{N-1} y\,[n]\, sin(2\pi(1/3)n) \\ \frac{2}{N}\sum_{n=0}^{N-1} y\,[n]\, cos(2\pi(1/3)n) \end{bmatrix}$$

and thus finally we have

$$\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} y\,[n]$$

$$\hat{\alpha}_1 = \frac{2}{N}\sum_{n=0}^{N-1} y\,[n]\, sin(2\pi(1/3)n), \quad \hat{\alpha}_2 = \frac{2}{N}\sum_{n=0}^{N-1} y\,[n]\, cos(2\pi(1/3)n)$$

$$\hat{A} = \sqrt{(\hat{\alpha}_1^2 + \hat{\alpha}_2^2)}, \quad \hat{\phi} = \arctan(\frac{\hat{\alpha}_2}{\hat{\alpha}_1})$$

Note that for the linear model, MLE is an efficient estimator in that it attains the CRLB and hence is the MVU estimator (see Theorem 7.5 on pp 186 of [25]). In the present case, the PDF of $\hat{\boldsymbol{\alpha}}$ is

$$\hat{\boldsymbol{\alpha}} \sim N(\boldsymbol{\alpha}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$$

This gives us

$$E(\hat{\mu}) = \mu, \quad var(\hat{\mu}) = \sigma^2/N$$

$$E(\hat{\alpha}_1) = \alpha_1, \quad var(\hat{\alpha}_1) = 2\sigma^2/N$$

$$E(\hat{\alpha}_2) = \alpha_2, \quad var(\hat{\alpha}_2) = 2\sigma^2/N$$

We will now derive the Cramer-Rao Lower Bound (CRLB) for the variance of the amplitude and phase estimators for a sinusoid of known frequency immersed in noise. The

derivation is based on Example 3.14 on pg 56 of [25]. Our model equation (D.1) may be written as

$$y[n] = s[n;\boldsymbol{\theta}] + w[n] \qquad n = 0, 1, ...(N-1)$$

where $\boldsymbol{\theta} = [\mu \;\; A \;\; \phi]^T$. Under the assumption that $w[n]$ is white Gaussian noise, the elements of the Fisher information matrix (see Example 3.9 leading to Equation 3.33 on pg 49 of [25]) are given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\frac{\partial s[n;\boldsymbol{\theta}]}{\partial\theta_i}\frac{\partial s[n;\boldsymbol{\theta}]}{\partial\theta_j}$$

We have

$$[\mathbf{I}(\boldsymbol{\theta})]_{11} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(1)(1) = \frac{N}{\sigma^2}$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{12} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(1)(sin(2\pi n/3+\phi)) \approx 0$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{13} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(1)(Acos(2\pi n/3+\phi)) \approx 0$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{22} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(sin(2\pi n/3+\phi))(sin(2\pi n/3+\phi)) \approx \frac{N}{2\sigma^2}$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{23} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(sin(2\pi n/3+\phi))(Acos(2\pi n/3+\phi)) \approx 0$$

$$[\mathbf{I}(\boldsymbol{\theta})]_{33} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(Acos(2\pi n/3+\phi))(Acos(2\pi n/3+\phi)) \approx \frac{NA^2}{2\sigma^2}$$

The Fisher information matrix now becomes

$$[\mathbf{I}(\boldsymbol{\theta})] = \begin{bmatrix} \frac{N}{\sigma^2} & 0 & 0 \\ 0 & \frac{N}{2\sigma^2} & 0 \\ 0 & 0 & \frac{NA^2}{2\sigma^2} \end{bmatrix}$$

We have upon inversion

$$var(\hat{\mu}) \geq \frac{\sigma^2}{N}$$

$$var(\hat{A}) \geq \frac{2\sigma^2}{N}$$

$$var(\hat{\phi}) \geq \frac{1}{\eta N}$$

where $\eta = \frac{A^2}{2\sigma^2}$ is the SNR. Note that the CRLB for each parameter decreases as $1/N$ and the bound for phase decreases as the SNR increases. Given that the MLE attains this bound, we can expect to have better estimates of amplitude and phase for longer signals with higher SNRs.

D.2 Least Squares Estimation

This method makes no probabilistic assumptions about the data, only a signal model is assumed. It is widely used in practice due to its ease of implementation, though its performance cannot be assessed without some assumptions about the structure of the data [25]. The LS procedure estimates model parameters $\boldsymbol{\theta}$ by minimizing

$$J = (\mathbf{y} - \mathbf{s}(\boldsymbol{\theta}))^T(\mathbf{y} - \mathbf{s}(\boldsymbol{\theta}))$$

where $\mathbf{s}(\boldsymbol{\theta})$ is the signal model for $\mathbf{y}$. Our model (D.1) can be simplified by transforming the parameters similar to the MLE method discussed above. If $(\mathbf{y} - \mathbf{s}(\boldsymbol{\theta})) \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, the LSE is also the MLE (see pp 254 of [25]).

D.3 Discussion

We need to be careful about the angle ϕ, since some methods require that it be within certain limits. If it is not, then the method will yield a spurious estimate. To avoid this problem, we do a series of tests (multiple confirmations) which will ensure that the phase angle we estimated is correct. One suggested approach is:

- Calculate phase angle using Least Squares, and write the full model equation
- Compare the above phase estimate with the MLE
- Compare the estimate with our cumulative calculation
- If the angle is coming out to be more than 90°, the model equation might need to be changed from *sin* to *cos* or vice-versa

Appendix E

Software Toolbox

The GSP toolbox contains software written in MATLAB for the analysis of free energy signals arising from nucleotide hybridization between rRNA and mRNA sequences during translation. It can be downloaded at: `http://www4.ncsu.edu/~lponnal/toolbox.htm`

The 'GSPtools' directory contains functions that perform specific tasks, such as detecting the $f = 1/3$ pattern, estimating signal parameters and calculating displacement. Type 'help function_name' at the MATLAB command prompt for more information on each function.

The 'GSPdemos' directory contains scripts that demonstrate some of the things one would want to do with free energy signals. Type 'help demo_name' at the MATLAB command prompt to see what each script demonstrates.

In order to examine a gene using the GSP tools, you will need three things:

- The sequence, stored in RNA or DNA format, either as a .fasta, .txt or .mat file (see Appendix A)
- The free energy signal for the above sequence, calculated using the appropriate 16S tail (see Appendix B)
- The availability of each tRNA in the species, expressed as a fraction (not a percentage).

At run-time, the sequence will need to be supplied as a character string in lower-case RNA format. Use either `getseq` or `readfasta` to load your sequence into a character string. A sequence (`seq`) is identified by its case (0 = small, 1 = capital) and type (0 = DNA, 1 = RNA). Use the trick

```
seq = num2char(char2num(seq,Case,Type),0,1)
```

to convert any sequence to lower-case RNA format.

The signal will need to be supplied as a row-vector, one free energy value for each consecutive alignment. See Appendix B for details.

Use the function `calctav` (see `demo3.m`) for calculating tRNA availability. You will need a set of verified sequences from the species, see Appendix A and B to learn how to get them. See the directory `GSP/GSPdemos` for sample files generated from the *E. coli* genome.

If you have everything ready, write a MATLAB script similar to `demo4.m` for analysis.

E.1 Storing sequence information

The GSP toolbox contains functions for storing sequence-related information pertaining to a species in easily-retrievable form (see `demo1.m`):

- `getanno` for extracting annotation from the protein coding table (Excel spreadsheet) and storing it as information matrices. An information matrix contains the following fields for each gene: function, strand (+/-), name, start, stop. Storing all this information in a matrix simplifies the task of finding a specific gene.
- `getseq` for extracting the sequence data into a .MAT file
- `writeseqs` (or `writeseqs2`) for writing verified and hypothetical sequences to separate text-files
- `checkwrseqs` for checking if the sequences have been written out correctly

- `calcgc` for calculating (G+C) content of the genes
- `writeindexfile` for writing index-files to be used with *free_scan*
- `gettails` for examining the 16S rRNA tails in the species

Functions that enable quick sequence retrieval are:

- `findgene`: search for a gene by its name
- `findgenes`: search for a set of genes by their names supplied in a text file

General-purpose functions included in the toolbox are:

- `char2num`: convert character sequence to numeric representation
- `num2char`: convert sequence from numeric to character form
- `codon2aacid`: name the amino acid corresponding to a codon
- `listcharbycodon`: list the codons in the input sequence and identify the corresponding amino acids
- `findaacid`: returns 3-letter abbreviation for a 1-letter amino acid
- `nseq2aaseq`: convert nucleotide sequence to amino-acid sequence
- `revcomp`: get the reverse-complement of an input sequence
- `seqdiff2`: locate differences between two character sequences
- `write2fasta`: write sequence to a file in FASTA format

E.2 Preliminary analysis of free energy signals

The purpose of preliminary signal analysis is to calculate the mean phase angle of the genes in the species and the variation in SNR across the genes. The following functions serve this purpose (see `demo2.m`):

- `checkwrsigs`: checks if the lengths of the signals written out to the text-file are correct

- `pickgg`: picks good genes based on signal criteria, calls the `detectf` function for testing the significance of the $f = 1/3$ component, and the `est_par` function for estimating signal parameters.

The procedure of model-fitting and parameter estimation has been described in [41]. Other useful functions are:

- `getseqsig1`: quickly extract the sequence and signal pertaining to a specific gene
- `est_par_res`: estimate parameters and return the residuals from model-fit

E.3 Polar plots and Displacement plots

The script `demo3.m` demonstrates the calculation of tRNA availability based on the relative abundance of each codon.

The following functions are used to examine each gene (see `demo4.m`):

- `cumm_mag_phase`: calculate the cumulative magnitude-phase profile of an input signal
- `calcmpx`: calculate the cumulative magnitude-phase profile, and the displacement profile for a gene
- `nloopcalc`: calculate the looping number (i.e., number of wait-cycles) based on the tRNA availability for each codon

Other useful functions are:

- `pplots`: produce polar plots for signals stored in a text-file
- `vecadd(d/r)`: add vectors with angles in (degrees/radians)
- `vecdiff(d/r)`: subtract vectors with angles in (degrees/radians)