ABSTRACT

MADHUP MISHRA. The Role of Free Energy Synchronization Signal in Translation of Prokaryotes. (Under the direction of Dr. Donald L. Bitzer and Dr. Mladen A. Vouk).

Sequences upstream of the coding region in prokaryotes show a consensus sequence called the Shine Dalgarno sequence. This sequence is the Watson-Crick complement to the 3' tail of 16S ribosomal RNA. Rosnick analyzed the ensemble free energy scores between the 3' tail-end of the RNA and the underlying mRNA. He found that the affinity between the tail-end and the mRNA is not just restricted to the upstream Shine Dalgarno region (SD region), but also extends downstream throughout the length of the gene. He confirmed the SD region as the lock signal and found an ensemble periodic free energy signal called the synchronization signal in the downstream region with a harmonic that peaks every third nucleotide with respect to the start codon. The periodic signal is hypothesized to either have a role in keeping the ribosome in frame with the mRNA being translated, or being a good predictive indicator of that state. The current work:

- Studies the hypothesis that the lock and the periodic signal seen in the ensemble of the species coding regions extends beyond just E.coli. Specifically the work is concerned with analysis of sample of species across bacteria and archae kingdoms of the prokaryotes. The analysis shows that the periodic signal is present in the coding regions but not in the non-coding regions and that in some cases a lock signal is not present. This work proposes an *Exponential Binding Index Locking Model* to account for the genes with no upstream lock signal.

- Proposes a novel methodology for analysis of the synchronization signal over individual genes. The approach leverages the ensemble periodicity information through sinusoidal wave interpolation with frequency of $1/3^{rd}$ to approximate the synchronization signal and study its magnitude and phase characteristics. The synchronization signal is seen as a good indicator of the process of translation as suggested by our investigations. The starting phase of the signal is dependent on the frame in which the Shine Dalgarno lock, where present happens upstream of start. An application of individual synchronization signal is identification of frameshifts in general in genes.The +1 programmed frameshifting gene prfB of E.Coli K12 was used as a case study to demonstrate this application.

The Role of Free Energy Synchronization Signal in Translation of Prokaryotes

by

Madhup Mishra

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial satisfaction of the
requirements for the Degree of
Master of Science

Department of Computer Science

Raleigh

2004

Approved By:

Dr. Steffen Heber

Dr. Donald L. Bitzer
Chair of Advisory Committee

Dr. Mladen A. Vouk
Co-Chair of Advisory Committee

I dedicate this work to my family, my teachers, my friends and my dog Leo ...

Biography

Madhup Mishra was born in a small town of Lucknow,India in the foothills of Himalayas on the 29th of October in 1978. He lived in Lcuknow for 18 years and did his entire schooling from the Saint Francis College there. Madhup made his first move out of Lucknow to Bangalore, India for his Bachelor of Engineering degree in Computer Science from the old Bangalore University. On successful completion of his Bachelor degree, he got admitted to the North Carolina State University for an MS degree in Computer Science Department.

Acknowledgements

Thanks to Dr Ann Stomp in Department of Forestry for her help in the research. Thanks to Scott K Vu for his help in doing the Cross Species Experiment. Thanks to all the members of the research group including Lalit Ponnala, Joshua Starmer, David Jerole, Tiffany Barnes and Chinhua Xing. Lastly thanks to anyone who helped me in anyway with this research or otherwise.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Central Dogma

DNA stores the genetic information for an organism needed to specify the proteins in coded form. The central dogma of molecular biology revolves around the paradigm that genes are perpetuated as sequences of nucleic acid, but function by being expressed in the form of proteins[25]. Three types of processes are responsible for the inheritance of genetic information and its conversion from RNA form to the proteins. Copies of DNA are produced by the process of replication. In organism like bacteria which have no nucleus, the DNA replicates in the cytoplasm of their cells. This can be viewed as perpetuation of information. The information is expressed by a 2 stage process of Transcription and Translation. Transcription generates a single-stranded messenger RNA(mRNA) identical in sequence with one of the strands of the duplex DNA(with Uracil replacing the DNA counterpart Thymine). The reverse process namely reverse transcription, though rare, is known to occur in some viruses, called retroviruses [25]. HIV is a notable example of a retrovirus. Through the process of translation, the mRNA present is converted into the sequence of amino acids comprising a protein. The series of 3 nucleotides, or codon, is responsible for coding one amino acid. The amino acids are carried to the site of translation by transfer RNA (tRNA). Long chains of these 20 different amino acids form proteins.

1.2 Cellular Machinery of Protein Synthesis

Three types of RNA, designated by their roles, all of which relate to the process of protein synthesis, are described in detail in this section[24][25].

1.2.1 Messenger RNA Composition

Messenger RNA(mRNA) and other cellular RNA species are single stranded polynucleotide unlike the DNA. It consists of a linear combination of 4 nucleotide residues - adenine(A), guanine (G), cytosine (C) and uracil (U) - sequentially connected by phosphodiester bonds between the 3' - end position of the ribose of one nucleotide and the 5' - position of the adjacent one as seen the figure 1.1. The terminal nucleotide, the 5' - position of which does not participate in forming the internucleotide bond, is referred to as the 5' - end of RNA. The terminal nucleotide with free 3' - hydroxyl is referred to as the 3' end. The mRNA is generally read from the 5' to the 3' end. The terminal 5'- position in natural mRNAs is always substituted. In prokaryotes, this end is either simply phosphorylated or carried the triphosphate group.

The length of mRNA is always greater than the length of the coding sequence. Each amino acid is represented in the mRNA by a sequence of three nucleotides or codons. The codons are arranged in contiguous non overlapping manner. The region of the message that contains the codons is called the reading frame. This reading frame is invariably flanked on both sides by variable length untranslated/non-coding regions. The 5' flanking region is referred to as the leader(upstream) sequence and the 3' flanking region as the trailer region. Identification of the factors that determine the starting point of the coding nucleotide sequence within the mRNA chain is an important problem. Each polypeptide is known to begin with an N - terminal methionine residue, and therefore the first codon in the coding sequence should be that of methionine. In most prokaryotic genes AUG and less frequently GUG or UUG form the initiation codon. It should however be noted that by no means does every AUG, GUG or UUG triplet become an initiation codon. Hence the choice of a given codon as an initiation codon depends not only on the codon structure, that is, its nucleotide composition and sequence, but also on the position of the codon in the mRNA. It has been shown that the nucleotide sequence preceding the initiation codon, as well as the particular secondary and tertiary structures of this mRNA region, are vital

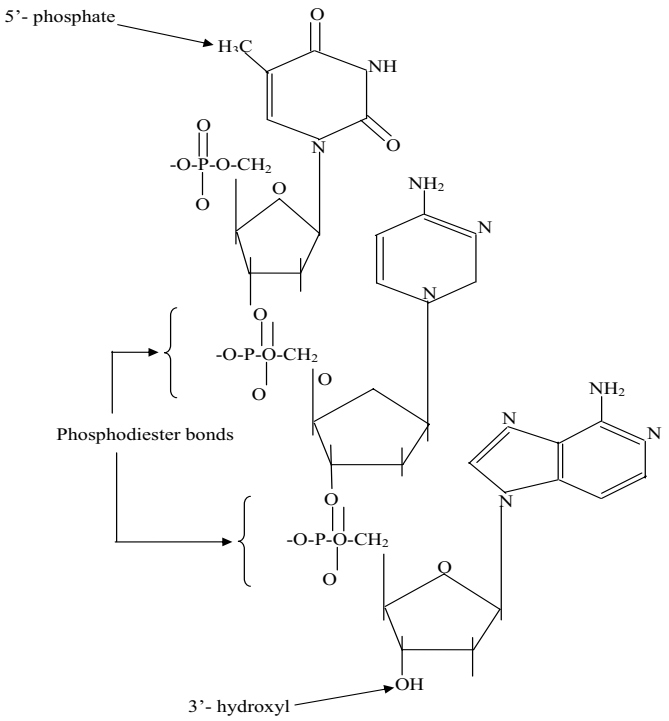

Figure 1.1: Nucleic acids (RNA and DNA) are formed by the condensation of nucleotides, catalyzed by polymerases. The bond that is formed is called a phosphodiester bond. Notice that the bonds in this case are formed between the 3' carbon of one sugar and the 5' carbon of the next sugar. If we follow the phosphodiester backbone of this trinucleotide from top to bottom, we see that there is a direction. That is, we begin at the phosphate on the 5' carbon and go 5' - 3' along the backbone. Picture taken from www.blc.arizona.edu website

for the corresponding triplet to be exposed as an initiation codon. Like the specific start codons, the two universal stop codons are UAA and UAG.

A given mRNA chain does not necessarily contain just one coding sequence. In prokaryotic organism mRNAs it is common for one polynucleotide chain to have coding regions for several proteins. Such mRNAs are usually termed as polycistronic mRNA. Different coding regions or cistrons within a given chain are usually separated by internal non-coding sequences which begin from the termination codons onwards of the preceding cistrons.

1.2.2 Transfer RNA Composition

Transfer RNA(tRNA) is one of the 3 types of cellular RNAs. It is believed that the 3-dimensional structures of most tRNAs are very similar except for the variable loop even though tRNAs show considerable sequence variability. These molecules fold back on themselves to form a clover-leaf with either four or five double-stranded base-paired stems and either three or four single-stranded loops. They fold in such a way as to bring the 5' and the 3' termini together in what is called the acceptor stem. The amino acid attaches to the 3' end of the acceptor stem. The other unpaired regions or loops of tRNA are named according to their unique structural features. The stem-loop structures of the tRNA are called its arms.

The function of transfer RNA is to transport free amino acids to the ribosomes where they become linked to form polypeptide chains. A tRNA has 2 crucial properties:

- It is able to represent only one amino acid to which it is covalently linked.
- It contains a trinucleotide sequence, the anticodon, which is a Watson-Crick compliment to the codon representing its amino acid. The anticodon enables the tRNA to recognize the codon in the mRNA sequence via complementary base pairing.

When a tRNA is charged with the amino acid corresponding to its anticodon, it becomes aminoacyl-tRNA. The amino acid is linked by an ester bond from its carboxyl group to the 2' or 3' hydroxyl group of the ribose of the 3' terminal base of the tRNA which is always adenine. There is at least one tRNA for each amino acid.

1.2.3 Ribosome and Its Sub-units

Amino acid are assembled into proteins by the ribosome, a compact ribonucleoprotein particle consisting of two subunits. The two subunits work together as part of the complete ribosome, but each undertakes distinct reactions in protein synthesis. Each ribosome subunit consists of several proteins associated with a long RNA molecule; these RNAs are known as ribosomal RNA(rRNA), the third type of RNA. The prokaryotic ribosome is named "70S" after its sedimentation coefficient. More accurately "S" here refers to the Svedberg coefficient, a measure of the rate of movement of a molecule or structure in a centrifugal field[24]. A characteristic feature of one of the visible ribosomal projections is

a groove dividing the ribosome into two unequal separable subunits. Under certain conditions of sufficiently low Mg,Na^+,Li^+ or urea concentration in the medium, the ribosome dissociates into the two subunits with a mass ratio of 2:1. High concentrations of about 0.5 M of each "physiological" monovalent cautions as K^+ and $NH4^+$ have similar effect on the ribosome. The 70S ribosome dissociates into subunits with a sedimentation coefficients 50S(molecular mass 1.65×10^6 daltons) and 30S (molecular mass 0.85×10^6 daltons)[24].

$$70S \leftrightarrow 50S + 30S \tag{1.1}$$

The re-association is promoted by Ca^{2+}, diamines, polyamines and alcohols.

The prokaryotic 30S subunit may be itself subdivided into lobes which are referred to as the "head" (H), "body" (B), and "side bulge" or "platform" (SB). It should be noted that the small ribosomal 30S subunit of archae-bacteria has a morphology which is intermediate between that of the eubacterial 30S subunit and eukaryotic 40S subunit. The small ribosomal subunit contains one molecule of high molecular weight rRNA that is designated 16S rRNA.

The 50S larger subunit is more isometric than the small one. Three peripheral protuberances can be distinguished: the central one (CP) is termed the head, the lateral finger-like protuberance is called the L7/L12 stalk and the other lateral protuberance, located on the other side of the central protuberance is referred to as the side lobe or the L1 ridge. The large ribosomal subunit contains a high molecular mass called 23S rRNA. The bacterial 23S is polynucleotide like the 16S. In addition to one molecule of 23S rRNA, the ribosomal subunits of cytoplasmic ribosomes of all prokaryotes contain one 5S rRNA molecule. It forms a separate domain of the large subunit.

In an intact ribosome the 2 subunits are joined very specifically. The flattened side of the 50S is involved in the contact between the subunits; if the subunit is viewed from this surface, the head of the subunit is up, and the stalk is on the right. The subunits are associated in the head to head and side lobe to side lobe manner.

The ribosome provides an environment that controls the recognition of a codon of mRNA by the anticodon of tRNA. To accomplish the sequential synthesis of a protein, the ribosome moves along the mRNA, one codon at a time. A ribosome attaches to mRNA at or near the 5' end of a coding region; moving along the RNA towards the 3' end, it translates each triplet codon into an amino acid en route. The effective contact length between the ribosome and the template at any one point of time is about 40 to 50 nucleotides. As

the ribosome proceeds, the appropriate aminoacyl-tRNAs associate with it, donating their amino acids to the polypeptide chain. At any given moment, the ribosome can accommodate the 2 aminoacyl-tRNAs corresponding to successive codons at its P and A sites, making it possible for a peptide bond to form between the two corresponding amino acids. At each step, the growing polypeptide chain becomes longer by one amino acid.

A number of ribosomes perform a readout of the same information on the template and synthesize identical polypeptide chains. Of course at any given point in time each of these chains are in different phases of completion. This complex between one ribosome attached to the template is called a polyribosome. The existence of polyribosomes as translating ribosomes in a cell indicate that the ribosomes are abundant in the cell while the quantity of mRNA is low. Indeed the rRNA comprises of around 80% of the cellular RNA while mRNA is not more than around 5% [24].

1.3 Stages of Translation

The process of translation in prokaryotes can be fundamentally divided into three parts, namely, initiation, elongation and termination. They are discussed in detail in subsequent subsections [24][25].

1.3.1 Initiation

A ribosome begins to read off information from the beginning of mRNa's coding region. This starting point(start codon) is at a certain, sometimes significant distance from the 5'-end of the polynucleotide chain. The series of events in which the ribosome identifies the origin of readout and binds to it is called as initiation of translation. Initiation requires a special initiation codon, initiator tRNA, and proteins, which are referred to as initiation factors. The initiation factors (IF1,IF2 and IF3) along with other translational factors and their functions are covered in the table 1.1. The dissociation of the non-translating 70S ribosome into 30S and 50S precedes the initiation. IF1 may accelerate this dissociation reaction while IF3 binds to 30S subunit and removes them from the 70S equilibrium state. The precise placement of the 30S subunit at the start codon is achieved by a base pairing between the 3' tail-end of the 16S rRNA of the 30S subunit and sequences found just

upstream from the initial start codon. This base-pairing interaction was discovered by John Shine and Lynn Dalgarno and the sequence in the mRNA is therefore called the Shine-Dalgarno sequence (SD sequence) or ribosome binding sequence (RBS)[30]. The SD-sequence suggests how prokaryotic mRNA can be polycistronic. Since the 30S subunit finds its way to each start codon, several different starts can be made on the same mRNA, if they have the SD-sequence upstream. The interaction is as shown in the equation 1.2[24].

$$30S : (IF3 : IF1 : IF2) + mRNA \rightarrow mRNA : 30S : (IF3 : IF1 : IF2) \tag{1.2}$$

Once the 30S subunit complex with the IF1,IF2 and IF3, is located correctly, the initiator tRNA($tRNA_f^{Met}$) is positioned over the start codon forming the initiation complex as shown in the equation 1.3 [24]. The binding is GTP dependent and mediated by IF2.

$$\begin{aligned} mRNA : 30S : (IF3 : IF1 : IF2) + GTP + tRNA_f^{Met} \rightarrow \\ mRNA : 30S : IF1 : IF2 : GTP : tRNA_f^{Met} : IF3 \end{aligned} \tag{1.3}$$

However if free IF2 with GTP encounters the $tRNA_f^{Met}$ in solution, they first form the ternary complex $tRNA_f^{Met}$:IF2:GTP and then this complex binds to ther mRNA:30S complex[24].

After the 30S subunit and appropriate tRNA have recognized the mRNA through the SD-sequence, the 50S subunit joins the complex, to form the 70S initiation complex. IF1 seems to be released concurrently with the subunit association. The factor-binding site of the 50S interacts with the IF2 and induces the GTPase activity of the factor. GTP is hydrolyzed, resulting in the loss of affinity of IF2 for $tRNA_f^{Met}$ and the ribosome. Thus 70S ribosome is formed precisely at the initiation codon with initiator $tRNA_f^{Met}$ in the P site[24].

$$\begin{aligned} mRNA : 30S : IF1 : IF2 : GTP : tRNA_f^{Met} + 50S \rightarrow \\ (mRNA : 30S : IF2 : GTP : tRNA_f^{Met} : 50S) + IF1 \rightarrow \\ mRNA : 70S : tRNA_f^{Met} + IF1 + IF2 + GDP \end{aligned} \tag{1.4}$$

Among the amino acids found in proteins, methionine plays a unique role, as it is found in the beginning of all polypeptide chains. It is also found internally like other amino acids. Hence there are two tRNAs, one for recognizing the start methionine ($tRNA_f^{Met}$) and the other for recognizing the internal methionine ($tRNA_m^{Met}$). The formylated methionine

Table 1.1: Translation Factors in Prokaryotes

Factor Name	Factor Type	Function
fheightIF1	Initiation	Stimulates IF2 and IF3
IF2	Initiation	Binds fMet-tRNAf; GTPase
IF3	Initiation	Prevents Rb association; monitors correct fMet-tRNAf-initiation codon interaction
EF1A(EF-Tu)	Elongation	Forms ternary complex with aa-tRNA and GTP; binds aa-tRNA to Rb A-site; GTPase
EF1B(EF-Ts)	Elongation	Promotes guanine nucleotide exchange on EF1A
EF2(EF-G)	Elongation	Promotes translocation reaction, GTPase
RF1	Termination	Promotes termination at UAA, UAG
RF2	Termination	Promotes termination at UAA, UGA
RF3	Termination	Promotes action of RFI and RF2; GTPase
RF4	Termination	Dissociates mRNA and tRNAs from Rb

is what is required at the start codon. The following reaction 1.5 is what makes them different [24].

$$\begin{aligned} tRNA_f^{Met} + Catalyst(Synthase) &\rightarrow methionyl - tRNA_f^{Met} \\ methionyl - tRNA_f^{Met} + Catalyst(Transpormylase) &\rightarrow fMet - tRNA_f^{Met} \end{aligned} \tag{1.5}$$

$$\begin{aligned} tRNA_m^{Met} + Catalyst(Synthase) &\rightarrow methionyl - tRNA_m^{Met} \\ methionyl - tRNA_m^{Met} + Catalyst(Transpormylase) &\rightarrow NoReaction \end{aligned} \tag{1.6}$$

1.3.2 Elongation

After translation initiation at the start codon, the next amino acid to be added to the protein chain is brought by its tRNA to the A site of the ribosome. This is carried out by EF-Tu. This requires that Tu be bound to a GTP. The placement of the aminoacyl-tRNA into the A site is at the expense of energy, and the GTP is hydrolyzed to GDP which leaves the protein. At this point the EF-Tu must be recycled by the addition of another GTP. This is accomplished by EF-Ts. The tRNA that was already present in the ribosome (say, the initiator tRNA) is in the P site of the ribosome. A bond is now formed between the COOH terminus of the amino acid or peptide chain on the tRNA in the P-site (actually

where it's linked to the tRNA) and the NH2 terminus of the amino acid in the A-site. The reaction is catalyzed by a part of the large subunit of the ribosome. The actual catalytic activity is probably a function of the large rRNA and is therefore a ribozyme reaction. At the end of this reaction, the growing peptide chain is carried by the tRNA in the A site of the ribosome. The ribosome is now going to move relative to the messenger RNA, down the mRNA in the 5' to 3' direction. The net result of this movement will be to transfer the tRNA carrying the polypeptide chain into the P-site, and the empty tRNA into the E-site. This leaves the A-site vacant, positioned over the next codon, and ready for the next aminoacyl-tRNA. This reaction requires another elongation factor, EF-G. Once the empty tRNA is in the E-site, it leaves the ribosome and the entire machine is ready for the next elongation step. This means bringing the next aminoacyl-tRNA into the A-site, again using EF-Tu/GTP. In this way, the ribosome consecutively reads out mRNA codons in the direction of its 3'-end which implies a synthesis of the polypeptide chain coded by the mRNA by sequential addition of amino acid residues to the nascent polypeptide chain. This is the polypeptide elongation.

1.3.3 Termination

The synthesis stops when the ribosome reaches the termination codon. In the presence of the termination codon, the ribosome does not bind any aminoacyl-tRNA. Instead, specialized proteins called termination factors induce the release of the polypeptide chain. In prokaryotes there are three release factors (RF-1, RF-2, and RF-3) whose functions mentioned in the table 1.1. Termination results in the release of the ribosome from the mRNA and the removal of the peptide from its link to the tRNA. The stage is rightly called termination of translation.

After termination, the ribosome generally jumps oftf the mRNA by breaking into its subunits. It may however continue to slip along the mRNA without translating and re-initiates when it comes across a new initiation codon. This is proposed to explain one of the reasons for the absence of the consensus signal of identification of the start sites in some Escherichia coli genes(Shine Dalgarno sequence) in section 5, table A.1.

1.4 Codon Bias

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC		UCC		UAC		UGC	
UUA	Leu	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Gln	CGA	
CUG		CCG		CAG		CGG	
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lys	AGA	Arg
AUG	Met	ACG		AAG		AGG	
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	Glu	GGA	
GUG		GCG		GAG		GGG	

Figure 1.2: Coding Dictionary with many to one mapping from codons to amino-acids. Picture taken from http://mmcalear.web.wesleyan.edu lecture notes.

All amino acids except Methionine and Tryptophan are coded for by two to six codons as shown in 1.2. DNA-sequence data from diverse organisms clearly show that synonymous codons for any amino acid are not used with equal frequency even though choices among the codons should be equivalent in terms of protein structures. Grantham et al have found that synonymous codons are used differently by different kinds of organisms and that each type of genome has a particular coding strategy-that is, the choices among synonymous codons are consistently similar for all genes within each type of genome [28][14]. This finding has been designated the genome hypothesis.

Ikemura et al proposed that organism-specific codon choice is related to organism-specific populations of isoaccepting tRNAs[1],specifically in the cases of Escherichia coli and yeast Saccharomyces cerevisiae [23]. For most amino acids, choices among synonymous codons are biased, and clear similarities of choice exist among the genes of each organism,

[1]Isoaccepting tRNAs are tRNAs that are charged with the same amino acid but usually respond to different codons for that amino acid.

in spite of the wide variety of gene functions. This organism-specific codon choice was called the *dialect* of the organism. The extent of the bias in codon usage found for individual genes in either organism was found to be closely related to the level of protein production for each gene [14][23]. Ikemura et al also found that highly expressed genes exclusively use one or a few synonymous codons, to the nearly complete exclusion of others. They conjectured that for moderately and poorly expressed genes however, the same type of dialect exists, but the extent of the bias is more moderate-that is, the codons that are exclusively used in the highly expressed genes are usually preferred, but other synonyms are also used at significant levels. They believed availability of tRNA molecules has been found to be a major factor in producing the codon dialects of E. coli and yeast. Codon choices observed in the foreign-type genes such as phage, transposon,and plasmid genes- are somewhat similar to those embodied in the host dialect, but the level of similarity is clearly lower than that found among host genes [23]. The codon-choice pattern of yeast mitochondrial genes differs totally from the pattern of its nuclear genes [17].

1.5 Frame Shifting

Alterations to the the reading frame occur extremely rarely during translation, yet some genes have evolved sequences that efficiently induce frameshifting. These sequences are termed as programmed frameshift sites. Frameshifting can be broadly classified into +1 and -1 shifts of frame. The -1 shift of frame is more common than its +1 counterpart. The rarest form of frameshifting however are the translational hop sites which program the ribosome to bypass a region of several dozen nucleotides. Each of these types of events are stimulated by distinct mechanisms. All of the events share a common phenomenology in which the programmed frameshift site causes the ribosome to pause during elongation so that the kinetically unfavorable alternative decoding event can occur. During this pause most frameshifts occur because one or more ribosome-bound tRNAs slip between the slippage and next codons. However, even this generalization is not entirely consistent, since some frameshifts occur without slippage. Because of their similarity to rarer translational errors, programmed frameshift sites provide a tool with which to probe the mechanism of frame maintenance.

+1 Programmed frameshift sites[1] occur less commonly than do -1 frameshift sites; however, they are as widely dispersed evolutionarily since they occur in bacteria , yeast, and mammalian cells[22]. The prfB gene exemplifies the major features of +1 frameshifting. The shift occurs when a peptidyl-tRNA is bound to the ribosomal P site and the A site is empty. It then requires a translational pause with the ribosome positioned over the frameshift site (25th codon). In prfB, the A site codon is a UGA terminator recognized by peptide release factor 2 (RF2), the protein product of prfB. When RF2 is limiting, recognition of the UGA is slow, and consequently frameshifting occurs; when RF2 is abundant, termination occurs instead of frameshifting. This produces an autogenous regulatory loop controlling levels of RF2 by regulating frameshifting [2][22]. The shift in frames appears to require that during the translational pause caused by slow recognition of the UGA, peptidyl-$tRNA^{GAG}_{Leu}$ slips from CUU onto a +1 overlapping UUU codon. Curran used mutant variants of the prfB site to demonstrate that the efficiency of frameshifting varies directly with the stability of the peptidyl-tRNA in the shifted frame, with the CUU codon giving the highest efficiency of those tested[2]. As with the bacterial-1 frameshifts, an upstream Shine-Dalgarno interaction stimulates frameshifting on prfB. The site is only 3 nt upstream of the CUU slip codon, much closer than in -1 frameshift sites. Mutagenesis of the site and the complementary sequence in 16S rRNA re-confirms the role of the 16S and that the interaction stimulates frameshifting[22]. Frameshifting occurs when the CUU-bound peptidyl-$tRNA^{UAG}_{Leu}$ slips from CUU to the overlapping UUA codon during the slow decoding of the AGG codon. The event appears to be entirely stochastic, and it requires no other stimulatory sequences.

1.6 Goals

The thesis reported in this work is concerned with two things.

- An investigation to find if the Free-Energy based periodic synchronization signal (discussed in detail in chapter 2 and further on) found by Rosnick et al. [29], is present in prokaryotes in general. It uses the existing ensemble synchronization signal averaging approach for the above investigation.Although the analysis over the ensemble tells us about the behavior of the species in general, an approach which could help investigate the behavior of the synchronization signal and the translation process in individual

[1] shift reading in the rightward direction by one base

genes is needed.

- The present work proposes a new signal vector based analysis approach as a tool for analysis of the signal in individual genes. This Vector based approach is used to shed some light on the properties of this signal in general. Also investigated is the possibility of the synchronization signal being a more precise indicator of the translation process than simple codon bias. The current work tries to investigate the possible relation of the synchronization signal with the upstream Shine Dalgarno lock using the Vector based approach. Finally the vector based approach is used to demonstrate how synchronization signal can be used to identify some properties of the coding region like frameshifting.

Chapter 2 talks about the previous research done in this area. It talks about Rosnick's work on which this current work is based. It also tries to capture the motivation behind the current work. Chapter 3 talks about the free energy score calculations over the ensemble of genes as well as individual genes. Chapter 4 lists the proof of concept of the ensemble synchronization signal over a set of species across bacteria and archae kingdoms. The 5 chapter proposes an alternate *Exponential Binding Index Locking Model* for the upstream lock. Chapter 6 discusses how to extract synchronization signal over an individual genes. It tries to a)find out how good an indicator of the translation process, individual Synchronization signal is, and b) Demonstrate one of the applications of Individual Synchronization Signal, namely Frameshifting. Chapter 7 discusses signal relationship with the upstream SD lock. Analysis of the work is covered in chapter 8. Finally chapter **??** concludes the current work and talks about future scope of the research.

Chapter 2

Related Work

The genomic revolution that started in mid 90s with the sequencing of the Haemophilus influenza genome has led to the sequencing of hundreds of genomes. The complete list of these sequences is available from the public database[1][2] . This growth has almost been exponentially increasing in genomic sequence data[26]. There is more than a theoretical interest in determining the factors separating coding sequences from non-coding open reading frames. The ability to predict translatable coding sequences is critical to the construction of new genes and the transgenic protein synthesis[12][8].

A number of approaches to distinguish the coding regions from non coding ones have been studied which we intend to summarize in this section. Special attention is payed to the free energy based approach proposed by Rosnick [29] on which the current work is based.

2.1 Comparative Gene Finding Approaches

These methods use similarity search procedures using putative proteins as queries to known databases to find homology information between these putative proteins and known proteins from the same or the other organism. A direct comparison of a genomic sequence with databases of expressed sequence tags (ESTs), using programs such as BLASTN

[1] http://www.ncbi.nlm.nih.gov/Genbank/
[2] http://www.tigr.org/tdb/

2.0 and AAT, can identify regions of a contig that correspond to processed mRNA[1][16]. This is a popular approach of recognizing putative genes similarities to cDNAs from the same or a closely related organism. An approach similar in principle is the comparison of a genomic sequence that is translated in all six reading frames with protein sequence databases, using a program such as BLASTX 2.0, can identify probable coding regions[1].

2.2 Markov Approaches

Homology information alone does not solve the annotation problem completely because of around 20-40% have no significant similarity with other known sequences or display only partial similarity with other known genes[11][15]. Hence there are intrinsic methods which train on DNA sequence only and determine gene locations using statistical patterns of nucleotides inside and outside coding regions along with the patterns at gene boundaries. Some intrinsic computer methods for gene finding employ a local Baysean approach such as GeneMark and Glimmer[11][18][26] and organism specific search program ECOPARSE[7] . These methods try to capture the compositional differences among coding regions, shadow coding regions (coding on the opposite DNA strand), and non-coding DNA. Such approaches still have difficulties in predicting the precise position of the start of translation[1]. Recently a system called GENMARK Genesis has been developed that automatically clusters ORFs from an uncharacterized bacterial genome and derives separate Markov models for each cluster obtained[1].

Several different types of sequence generating models can be combined into a unified scheme by using a hidden Markov model (HMM) framework. In this approach, transitions between sub models corresponding to particular gene components are modeled as unobserved(hidden) Markov processes, which determine the probability of generating particular (observable) nucleotides. The HMM architecture is in fact quite general and has been applied successfully to many problems in computational biology and in other fields. One such approach based on the hidden Markov model is the PROD-HMM[11]. The PROD-HMM is a composite of two such HMM models that simultaneously model the statistics of pairs of orthologous DNA sequences through species-specific transition and emission probabilities.

An important distinction between Markov model architectures used in gene finding

is that some programs (e.g. GENSCAN and GENMARK) use explicitly double-stranded models that allow for the occurrence of multiple genes on either or both DNA strands, whereas most others analyze only one strand at a time and assume that the input sequence contains a single complete gene[1].

2.3 Information Theory based Approach

Schneider et al proposed an informational theory based approach of computational characterization of start sites of genes. It is based on the physics underlying molecular binding interactions which was created by Shannon. The information contained in a set of binding sites can be computed by summing the information content across the base positions of the binding sites[32]. As one would admit, the information theory study only shows the average sequence conservation and inferring the conservation for individual sequences is difficult. To overcome this drawback, Schneider introduced the individual information technique. The method begins by generating a weight matrix from the frequencies of each nucleotide at each position of aligned sequences. This matrix is then applied to each of the sequences to determine the sequence conservation of individual sequences. It should be noted that it is not possible to determine the information content from a single sequence alone because the actual contact could be anywhere in the single gene. Also the information becomes really biased on a single gene. The individual information method depends on an aligned set of sequences. This works under the assumption that we know the starts and stops of the sequence that we need to do an information theory study on. The alignment of the sequences is important because the information content shows up on a consensus of genes and not just one or couple of them. This method is really helpful if we have an idea where to find the consensus. While multiple alignment is a difficult problem in general, for most binding sites, gaps are not required to make good alignments because protein binding sites are generally small objects with little flexibility observed along the sequence. A general theory for individual information with gaps is not available, although the uncertainty introduced by gaps has been considered [32] and hidden Markov models(2.2) may provide the basis for a solution.

2.4 Coding Theory based Approach

Coding theory approach is helpful in overcoming the shortcoming of the Information Theory approach discussed in the previous subsection. May et al hypothesized if the genetic information in the DNA sequence is encoded in a manner equivalent to block encoding, the mRNA, should conform to the block coding model[6]. Using basic block coding principles they developed a method and a decoding model based in chemical and biological characteristics of the ribosome and the ribosome binding site, located in the leader region of the mRNA. Their hypothesis functionally considers the mRNA as a noisy, convolutionally encoded signal and the ribosome as a table-driven convolutional decoder[6]. The idea behind the approach is that various leader sequences having the Shine Dalgarno are encoded signals with noise in them which can be decoded to the exactly same error-free information(lock or not lock in this case) by the ribosome. The block code method claims to show distinction between translated sequence groups and non-translated sequence groups.

2.5 Signal Processing and Free Energy based Approaches

In 1985, Veljkovic' et al examined various functionally similar DNA sequences and determined the existence of each corresponding consensus spectrum[20]. The multiple cross-spectrum of sequences of unrelated functions however failed to produce that. Signal Processing based approaches are still actively being researched on.

In addition to investigating new roles for the RNA, recent work has expanded thought on how to measure interactions of rRNA and the mRNA from a thermodynamic perspective. The change in free energy due to binding between the two strands is a logical start. The Gibbs free energy is used to measure the spontaneity of any biochemical reaction (Δ G < 0) with the idea being more negative free energy implies more spontaneous reaction. Thanaraj et al examined the 16S pairing with the mRNA upstream regions and suggested a strong binding leading to strong stability for ribosome formation[33]. Schurr et al employed free energy in lieu of base-pair counts as a measure of complementarity [31].

David Rosnick in his PhD work calculated the free energy from 50 positions upstream of the start codon to the stop codon for 2085 coding sequences [29][4][3]. He overlayed the tail-end (13 base pairs) of the 16S rRNA of the E.coli on the mRNA and calculated the free energy in kcal/mol by adding the energy expended and released to form this binding

[29][5]. The tail-end was then moved over downstream by one base pair and the new free energy was calculated. Finally, he aligned the genes he considered on their start codons, and then took position specific averages of free energy across the ensemble as covered in the section 3.4. He observed that for E. coli, on an average over the ensemble there is an upstream lock (strong binding) of around -2 to -2.5 kcal/mol centered around 10 upstream bases and consistent with the SD location which confirmed all the previous findings. This is called the Shine Dalgarno (SD) lock signal. This SD lock has been shown to relate to translation rates of the genes to which they precede. This SD lock however is not the sole indicator for translation [27][31].

While the SD lock is important for the translational initiation, the affinity between the 16S tail-end and the mRNA sequence continues throughout the gene-length (ie. elongation period)[24]. Experimental research in downstream homology before Rosnick's work indicates that the SD sequence impacts the efficiency of a shifty stop [13]. Other works of the affinity between the 16S tail-end and the mRNA have mostly been restricted to the upstream regions or simpler base preferences and relatively with a small sample size or mixing data from different organisms[31][34]. Rosnick used the same free energy calculations that he used for the upstream lock throughout the length of the gene. By examining the ensemble average free energy values and subsequent Fourier analysis, a periodic "synchronization" signal repeating every 3 base positions was detected in the coding region[29]. Discovery of this synchronization signal went beyond the previous work because the observed periodicity is tied directly to the 16S, rather than the broad statistics of codon preference or any Markov model. Also the computed signal has a physical basis directly related to the experimental evidence of binding between the 16S and mRNA. Knowing that there is a signal of any strength in the ensemble sum implies a preference of that signal in each of the individual coding sequences considered in the ensemble. As the nature seems to favor this signal and maintains it over the coding region he hypothesized that the synchronization signal is advantageous in some way. Rosnick et al conjectured that the signal is an essential property of the coding region and that it keeps the ribosome in place and helps prevent frame shifting. If the phase of the signal deviates sufficiently, the likelihood of a frameshift was hypothesized. Hence this model was consistent with the shift stop model of frameshifting.

2.6 Motivation for Current Work

The principal assumption of this work is that the binding of the exposed part of the 3' end of the 16S rRNA to the underlying mRNA is a good indicator of the overall state of the translation process regarding its initiation as well as its translational efficiency [4]. This work is based on the biochemical model for Escherichia Coli suggested by David Rosnick as a part of his PhD work as discussed in chapter 2. The motivation behind the current work was get a handle on the process of translation in prokaryotes so that we can, with sufficient ease and efficiently, produce useful human proteins using Escherichia Coli (or other prokaryotes). Another motivation behind the current work was to come up with a better identification tool for the identification of coding region in general. The various investigations that we performed in the current work for the better understanding of the translation process are itemized as follows:

- Investigate the possibility of cross-species protein production based on the synchronization signal by ensemble synchronization signal analysis.
- Use the ensemble average synchronization signal to study the relationship of the synchronization signal to the Shine Dalgarno lock.
- Extraction of the synchronization signal on an individual gene basis with good signal to noise ratio.
- Use the extracted signal to study if synchronization signal in individual genes is a good indicator of the translation process.
- Use the extracted signal to study the characteristics of the process of translation and coding region like frameshifting.

Chapter 3

Methodology and Calculations

This chapter deals with the free energy based scoring that we have used throughout our work. It talks about the methodology of ensemble synchronization signal for a set of genes. It further discusses another approach based on the individual genes which is helpful in studying the characteristics of coding regions of individual genes.

3.1 Source of Data

The species investigated in this study were chosen to explore and demonstrate how taxonomically far from the E coli does the ensemble synchronization signal phenomenon observed by Rosnick et al. extend [21]. The species from a different class, order, phylum, and kingdom relative to E.coli were chosen for the experiment. The common condition was that the selected species had a fully sequenced genome in the GenBank [1]. The list of species and their GenBank accession numbers are shown in Table 4.1. We also used the Gutell Laboratory Database [2] to determine where the hairpin starts on the exposed tail of the 16S small subunit for all the species as shown in table 4.2 . In the cases where there was a discrepancy between the Gutell Laboratory and GenBank, data from GenBank was used. The hairpin at the 3' end of the 16S was assumed to start after the first sequence "tag" reading from the 3' end. This was chosen because E.coli K-12 exposed tail ends with

[1] http://www.ncbi.nlm.nih.gov/Genbank/
[2] http://www.rna.icmb.utexas.edu/

"tag". For most of the species we investigated, the 3' end exposed tails looked very similar to E.coli K-12's exposed tail with a little difference in length. The sequence "3 - ctccactag - 5" of the exposed tail exists in all species.

3.2 Software Used

The software languages used in this work were primarily C and Perl. Perl is a powerful parsing tool which can be used to come up with specialized state-of-the-art text mining tools that can be used to parse out sequence of genes and their properties like if the gene under the study is verified, ok or just putative or hypothetical. The were C programs written for doing the dynamic programming based free energy calculations that we will discuss in a later subsection. C programs were written to do the ensemble averages on the free energy from various genes of each species. Vector analysis on the synchronization signal was performed by other C modules.

MATLAB was the primary tool used for analysis throughout the work. The C modules mentioned generate code in Matlab for plots relating to the free energy and the polar plots in the vector analysis that were used. Microsoft's Excel sheets were used in a limited way to do some basic statistical calculations during the binding index locking and the codon-bias calculations.

This document was prepared using the LATEX documentation tools. Acrobat was used to create the illustrations.

3.3 Free Energy Score Calculations

The free energy for the formation of binding between two strands of RNA or an RNA strand to form a secondary structure is a thermodynamic constant that gives the amount of energy required for or released by a reaction. It is measured in kcal/mol as the parameter ΔG. Reactions that require energy have a positive value. Reactions that release free energy have negative value. Energy must be released overall to form a stable RNA-RNA or RNA secondary structure formation [25]. The stability of such a structure is determined by the amount of energy released. For example a binding with ΔG=-21 kcal/mol is less likely than a binding with ΔG=-35 kcal/mol. Specifically, a particular pairing conformation

between the mRNA and 16S rRNA is more stable if the formation of the hydrogen bonds via Watson-Crick pairing results in a more negative free energy. The conformation resulting in the most negative free energy is considered optimal [29]. Various sub-optimal conformations may also be considered by restricting the manner of the binding. For example, a binding site along an mRNA strand doesn't necessarily reflect the optimal conformation between the mRNA and the bound 16S, but rather a sub-optimal conformation restricted to binding in a particular region.

Throughout this work free energy is used as an exclusive quantitative measure for the likelihood of the stability of the rRNA-mRNA interaction. Inface we shall use the term binding energy, free energy score, free energy and binding strength interchangeably throughout the text. These free energy calculations are based on the same principles that are used for folding a single RNA molecule and are executed by similar algorithms [25][19][31]. Dynamic programming is used for finding the best match between the RNA strands based on free energies released by matching base-pairs. There is a penalty mechanism in place that penalizes mismatches. For every series of mispaired bases, called a loop, a +0.8kcal/mol is applied as penalty[25].

The free-energy changes for binding initiation and propagation were obtained by multiple regression to observed thermodynamic parameters upon the nearest-neighbor model [19][10][25]. These thermodynamic parameters of binding formation were obtained by two methods.

- Individual melting curves were fit to a two-state model with sloping base lines and the enthalpy and entropy changes derived from the fits were averaged.
- Reciprocal melting temperature t_m^{-1} vs log(C_t) was plotted to yield enthalpy and entropy changes.

On the basis of reproducibility, the estimated error limits are ± 5 % for the enthalpy and entropy changes, ΔH° and ΔS°, and $\pm 2\%$ for the free energy change, ΔG° at the melting temperature t_m [19]. According to the nearest-neighbor model, the standard-state free energy ΔG° of formation of hydrogen bonds among two strands of the RNAs is the sum of 3 terms:

- A free-energy change for helix initiation associated associated with forming the first base pair of the binding.

- A sum of propagation free energies for forming each subsequent base pair.
- A symmetry correction if the sequence is self complementary.

$$\Delta G^\circ = \Delta H^\circ - t \times \Delta S^\circ = \Delta H^\circ - 310.15 \times \Delta S^\circ \tag{3.1}$$

Here the ΔH° is the enthalpy change and ΔS° is the entropy change in the same binding reaction. The enthalpy and free energy changes for the two-state transitions were used with the nearest neighbor model to obtain parameters for initiation and propagation of binding as shown in table 3.1

Based on the above points the overall free energy of 2 strands of RNA binding against each other can be given by the general equation 3.2

$$\Delta G^\circ_{total} = \Delta G^\circ_i + \sum \Delta G^\circ_x + \sum \Delta G^\circ_u \tag{3.2}$$

The individual terms in the equation 3.2 are calculated as [25]:

- ΔG°_i is the free energy for initiation of the binding. It takes a positive value of +3.4 kcal/mol, representing the energy required to form the first base pair. This applies to intermolecular duplex formation and doesn't apply to intramolecular duplexes (such as hair-pins).
- $\sum \Delta G^\circ_x$ is the sum of the individual reactions involved in propagating the binding as each base pair is added. The formation of each base pair releases energy. Hence this sum is negative.
- $\sum \Delta G^\circ_u$ is the sum of individual instances encountered in which the opposing bases are not complementary. It represents the energy required to hold these bases in an unpaired state. Hence this sum is positive.

The specific pairing up of the strands of the rRNA and mRNA depends on factors like temperature, pressure, ion concentration etc which have not been considered in our computations yet. Bulges and internal loops are not considered in the calculations. In fact these are seen as a future scope which is likely to refine the results by a substantial amounts.

In the present work the free energy scores for the interaction between the 13 bases of the 3' tail-end of the 16S rRNA and the mRNA of the individual genes are calculated as described in [25][31] using free energy values using from [19]. Let the score estimate for

Table 3.1: Thermodynamic parameters for RNA-RNA binding initiation and propagation in 1M NaCl

Propagation sequence $\longrightarrow$ $\longleftarrow$	**ΔH° kCal/mol**	**ΔS° eu**	**ΔG°_{37} kCal/mol**
$\frac{AA}{UU}$	-6.6	-18.4	-0.9
$\frac{AU}{UA}$	-5.7	-15.5	-0.9
$\frac{UA}{AU}$	-8.1	-22.6	-1.1
$\frac{CA}{GU}$	-10.5	-27.8	-1.8
$\frac{CU}{GA}$	-7.6	-19.2	-1.7
$\frac{GA}{CU}$	-13.3	-35.5	-2.3
$\frac{GU}{CA}$	-10.2	-26.2	-2.1
$\frac{CG}{GC}$	-8.0	-19.4	-2.0
$\frac{GC}{CG}$	-14.2	-34.9	-3.4
$\frac{GG}{CC}$	-12.2	-29.7	-2.9
Initiation	(0)	-10.8	3.4
Symmetry Correction (self-complementary)	0	-1.4	0.4
Symmetry Correction (non-self-complementary)	0	0	0

the i^{th} gene for x^{th} base position relative to the first base of the start codon be e^i_x. Let the position of the first base of the start codon be denoted by x=0. In addition, let codons be numbered k=0,1,2,3.... beginning with the start codon.

Consider the following illustration. Let the first sequence below be the tail-end of the 16S rRNA. Let the second sequence below be the region upstream of the start in mRNA. Let x mark be the position to which the calculated score will be assigned.

3′A$\underline{U}$UC$\underline{CUC}$C$\underline{AC}$UAG-5′(16S rRNA tail-end).

5′C$\underline{A}$UA$\underline{GAG}$U$\underline{UG}$GCA_x-3′(i^{th}) mRNA sequence).

Then

‘ $e^i_x = \Delta G^\circ_{total-at-x} = \Delta G^\circ_i + \sum \Delta G^\circ_x + \sum \Delta G^\circ_u$

Then in our example

$e^i_x = \Delta G^\circ_{total-at-x}$ = +3.4 -2.3 -1.7 +0.8 -2.1 = -1.9kCal/mol. based on the table 3.1. This calculated os repeated for all the n_i bases of the i^{th} gene giving its free energy score vector $\{....e^i_{-2}, e^i_{-1}, e^i_0, e^i_1, e^i_2, ..., e^i_{n_i-1}, e^i_{n_i}\}$.

3.3.1 Indexing Scheme

As seen from the position of x in the illustration before, the indexing scheme is clear. The 3' side of the mRNA sequence which matches with the 13^{th} base from the 5' end of the 16S stream is the position which gets the value of this total free energy score allocated. So when the 13 bases of the 16S are aligned over the mRNA stream such that the G at the 5' side of the 16S above lands on the first base of the start codon of the gene, it is called the 0^th position with respect to the start codon. To move on to -1 position, the ribosome is moved one left on the mRNA strand. Similarly the ribosome moved one towards the right will take it to the +1 position.

3.4 Ensemble Synchronization Signal Analysis

To see the synchronization signal over an ensemble of genes for a particular species, all genes in the ensemble, were lined up on their start codons and the 3$\acute{t}$ail-end of the 16S rRNA of the same species (different species 16S for cross-species experiment) was layed over the mRNA from 50 bases upstream till the stop codon for each of the genes. The free energy released due to the base pairing between the Watson-Crick pairs between the mRNA stream and the 16S tail-end was calculated at any one position on the mRNA using dynamic programming approach of longest subsequence match as explained in the section 3.3. The indexing scheme used to allocate this energy value to a specific position in the mRNA is done using the scheme explained in the subsection 3.3.1. This would give us free energy released in kCal/mole for that interaction of the two. Now the 16S tail-end is slid over by one base pair on the mRNA stream and the new set of free energy interactions are computed like the previous case. The 16S tail-end is hence moved over one complete gene. The free energy signal in kCal/mole was calculated in this way for all the genes of the ensemble. The free energy vector thus obtained for i^{th} gene is $\{....e^i_{-2}, e^i_{-1}, e^i_0, e^i_1, e^i_2, ..., e^i_{n_i-1}, e^i_{n_i}\}$. Free energy score vectors were similarly obtained for all the genes from 1 thru N where N is total number of genes in the ensemble. As the genes were aligned on their start codon, we get a matrix of free energies $M \times N$ where M is the length of the free energy vector for the longest gene and N is the number of genes in the ensemble. The matrix is as follows:

$$\begin{bmatrix} .... & e^1_{-2} & e^1_{-1} & e^1_0 & e^1_1 & e^1_2 & .... \\ .... & e^2_{-2} & e^2_{-1} & e^2_0 & e^2_1 & e^2_2 & .... \\ .... & .... & .... & .... & .... & .... & .... \\ .... & e^i_{-2} & e^i_{-1} & e^i_0 & e^i_1 & e^i_2 & .... \\ .... & .... & .... & .... & .... & .... & .... \\ .... & e^N_{-2} & e^N_{-1} & e^N_0 & e^N_1 & e^N_2 & .... \end{bmatrix}$$

Now as the genes were lined up on their start codons, we take position specific averages(averaging vertically over columns) of these free energy values across the ensemble. It should be noted that the since the genes were of different length that the average at any one position would not be always be calculated over total number of genes (N). Lets say that the number of genes at any one particular column x is M(x). Then the position specific averaged signal vector would be of the form:
$\{..., \frac{\sum_{z=0}^{M(-2)} e^z_{-2}}{M(-2)}, \frac{\sum_{z=0}^{M(-1)} e^z_{-1}}{M(-2)}, \frac{\sum_{z=0}^{M(0)} e^z_0}{M(0)}, \frac{\sum_{z=0}^{M(1)} e^z_1}{M(1)}, ...\}$.

This way of locating the synchronization signal over an ensemble has the advantage of filtering out the individual variations in energy at gene-level thereby filtering out most of the noise. Hence it makes it easier to identify the average signal as opposed to the one present in the individual genes.

3.5 Individual Synchronization Signal Analysis

Ensemble averaging shown in the section 3.4 is a good indicator of synchronization signal behavior over a group of genes. However to discern the signal in an individual gene becomes difficult because the noise-level goes up substantially. To look at the individual genes more carefully we need some way to extract this signal in a single gene keeping good signal to noise ratio. From previous work of Rosnick et al., we know that the dominant frequency of the synchronization signal is $1/3^{rd}$ [4]. Based on that, we approximate it using a sinusoidal wave anchored in the three bases of a codon. We represent this sinusoidal waveform as a vector with definite magnitude and phase. In order to get the sinusoidal wave anchored in three bases which has a low noise level, we accumulate free-energy scores for each of the three positions in the reading frame relative to the start codon over a certain

number of codons. Let the three accumulated scores over the first k codons be:

$$A_k = \sum_{x=0,3,6}^{3k} e_x$$
$$B_k = \sum_{x=1,4,7}^{3k+1} e_x$$
$$C_k = \sum_{x=2,5,8}^{3k+2} e_x \tag{3.3}$$

In order to remove and DC term from the signal we take the average of A_k,B_k and C_k (3.4) and subtract from each of them to get a_k,b_k and c_k respectively(3.5).

$$DC = \frac{A_k + B_k + C_k}{3} \tag{3.4}$$

$$a_k = A_k - DC$$
$$b_k = B_k - DC$$
$$c_k = C_k - DC \tag{3.5}$$

Now we use the 3 points a_k, b_k and c_k in equation 3.5 to be three points on the sinusoid.

$$a_k = M_k sin(\phi_k) \tag{3.6}$$

$$b_k = M_k sin(\phi_k + \frac{2\pi}{3}) \tag{3.7}$$

$$c_k = M_k sin(\phi_k + \frac{4\pi}{3}) \tag{3.8}$$

From equations 3.7 and 3.8 we get

$$M_k cos(\phi_k) = \frac{b_k - c_k}{\sqrt{3}} \tag{3.9}$$

Using 3.6 and 3.9, M (magnitude) and ϕ (phase) can be computed as shown below

$$\phi_k = tan^{-1}(\frac{a_k\sqrt{3}}{b_k - c_k})$$
$$M_k = \sqrt{(a_k)^2 + \frac{(b_k - c_k)^2}{3}} \tag{3.10}$$

Plotting this magnitude and phase over the entire length of a gene in polar coordinates as shown in figure 3.1.

The cumulative nature of this approach helps keep the noise-level under control. It is interesting to note how we compensate the loss of data over an ensemble by cumulatively adding up energies across a single gene. During the course of this report we shall call this cumulative vector summation as being horizontal as opposed to the ensemble averaging which can be seen as vertical.

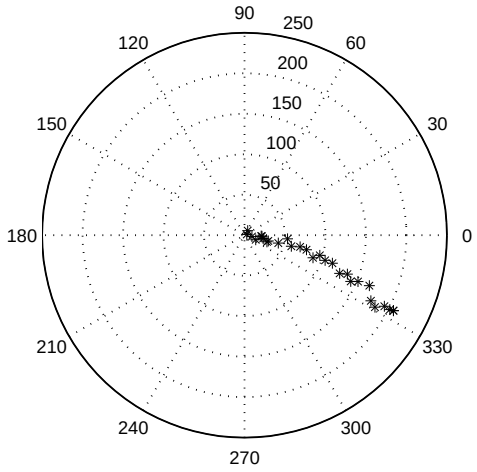

Figure 3.1: The Synchronization signal shown for a single E.coli K12 gene of length 1451 base-pairs in the polar plot with the angle of cumulative vector is the phase of synchronization signal and the distance from the center is its magnitude

Chapter 4

Ensemble Synchronization Signal Experiments and Results

Ensemble Synchronization signal is important in the sense it tells the general information of the coding region of a set of genes by suppressing the variations of individual gene synchronization signal. In the subsection 4.1 we use it to see the extension of the model proposed by Rosnick.

4.1 Cross Species Analysis

The Cross Species Analysis was done to find out if the ensemble energy patterns observed for E.coli existed in other prokaryotes, and to assess the investigated ribosomes with respect to their behavior in potential cross species translation. Ensemble free energy analysis was done on a variety of species at taxonomically different levels from E.coli to identify the lock signal and the synchronization signal in them and their roles in translation. Eight species were chosen from different genus, order, class, phylum and kingdom from the E.coli. The Salmonella is in a different genus from E.coli in bacteria kingdom. Pseudomonas is in a different order than that of E.coli. Rickettsia lies in a different class. Aquifex and Lactobacillus are in a different phylum. Thermoplasma, Sulfolobus and Halobacterium were

chosen from the archae kingdom which is different from the bacteria kingdom of E.coli. The details of the species chosen and their Genbank accession numbers are mentioned in the table 4.1.

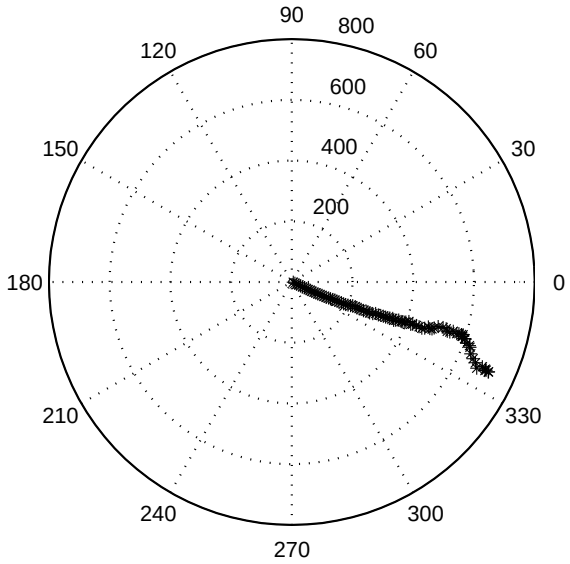

Figure 4.1: The Synchronization signal found over an ensemble of 200 genes from E.coli K12 length 4604 base-pairs in the polar plot with the angle of cumulative vector is the phase of synchronization signal and the distance from the center is its magnitude. This plot shows the species consensus over the phase of the synchronization signal

200 non-putative non-hypothetical genes were chosen from each of the species. The details on how the genes and the 16S tail ends were chosen have been covered in the previous section 3.1. The genes were chosen to be of different lengths keeping the gene-length distributions of the sampled genes across all of the species somewhat alike. This was to prevent any kind of bias in the lengths of the gene sample affecting the results in any way.

As explained in the section 3.4 the free energy of the average signal was calculated over 200 E.coli genes with its own 16S tail-end of rRNA. This ensemble free energy signal obtained was analyzed for the lock signal and the synchronization signal that Rosnick et al had talked about. For the lock signal the ensemble free energy of the region upstream of the start codon was analyzed. Also for the synchronization signal, the region for the length of the entire average signal was analyzed. This was then repeated for 200 genes of each of

Table 4.1: Prokaryotic species selection table

Prokaryotic Species	Abbreviation	Accession #	Taxonomical difference from Escherichia coli K-12
Escherichia coli K-12	E.coli	U00096	None
Salmonella typhimurium	Samon	NC_003197	Genus
Pseudomonas aeruginosa	Pseudo	NC_002516	Order
Rickettsia conorii	Rick	NC_003103	Class
Aquifex aeolicus	Aquifex	NC_000918	Phylum
Lactobacillus plantarum WCFS1	Lacto	NC_004567	Phylum
Thermoplasma volcanium	Thermo	NC_002689	Kingdom
Sulfolobus solfataricus	Sulfo	NC_002754	Kingdom
Halobacterium sp. NRC-1	Halo	NC_002607	Kingdom

Table 4.2: Species 16S 3' Exposed Tail

Prokaryotic Species	3' exposed tail of 16S
E.coli	3 - attcctccactag - 5
Samon	3- attcctccactag - 5
Pseudo	3 - attcctccactag - 5
Rick	3 - attcctccattag - 5
Aquifex	3- atttcctccactag - 5
Lacto	3 - tctttcctccactag - 5
Thermo	3- cctccactag -5
Sulfo	3 - ctccactag - 5
Halo	3- tcctccactag - 5

the other species with their respective 16Ss. Their average signals were analyzed for both the signals. The aim of this was to explore how taxonomically far from the E.coli can these signals occur.

There was another experiment performed with the same set of genes from the 9 species. This time for each of the gene of a particular species say A, the 16S tail-end from another species say B was used to find the free energy. The ensemble free-energy signal for 200 genes of A with the 16S tail end from B was calculated again as shown in section 3.4. This free-energy signal was for the lock and the synchronization signal. This was with the aim of exploring the possibility of cross-species protein production with free energy as the parameter. It should be noted that is was done under the assumption that the 16S tail end

is a good indicator for the process of translation not taking into consideration other factors like temperature, pressure, salt concentration etc. This experiment was repeated for all the genes from each of the 9 species with the 16S tail-ends from the rest of the species.

Figures A.1 thru A.18 show the results for the various species with their own 16Ss as well as the other species 16S tail-ends. The ensemble of the each species mentioned in the table 4.1 is plotted at a)The upstream region in search of the lock signal for various 16Ss and b)The region downstream(19 to 180 base-positions) in search of the downstream periodic synchronization signal for various 16S tail-ends.

The Tables 4.3-4.4 summarize the locking and signal analyses for both self a nd cross species interactions between the 16S tail-end of rRNA and mRNA. In the table, L is the estimated locking free energy in kcal/mol; M is the average synchronization signal amplitude in kcal/mol per base; ϕ is the average synchronization signal phase; DC is the average free energy of the synchronization signal; Codon # is the number of codons for a particular species used for the experiment and min,max are the smallest,largest gene length in # of codons for the particular species. An amplitude of about 0.2 kcal/mol per base around the average is considered acceptable. 'None' means that the locking signal was not distinguishable from the noise.

Based on the lock signal strength and the synchronization characteristics as shown in tables 4.3-4.4, we organized the results in the following four categories.

- *Good upstream lock and strong synchronization signal*: The mRNA sequences from E.coli(Row 1, A.1, A.2), Salmonella(Row 2, A.3, A.4), Pseudomonas(Row 3, A.5, A.6), Aquifex(Row 5, A.9, A.10) and Lactobacillus(Row 6, A.11, A.12) show a strong upstream lock (L) and a decently strong periodic synchronization signal (M and ϕ), both with their respective tail-ends and the other tail-ends in this small subgroup.

- *Moderate upstream lock and moderate synchronization signal*: Thermoplasma(Row 7, A.13, A.14) and Sulfolobus(Row 8, A.15, A.16) fall into this category.

- *Weak upstream lock and weak synchronization signal*: Rickettsia(Row 4, A.7, A.8), for reasons not entirely clear, has a weak upstream locking and a weak synchronization signal. Validity of data is under investigation.

- *No upstream lock but strong synchronization signal*: Halobacterium(Row 9, A.17, A.18) doesn't have exhibit a noticeable upstream locking but shows a really strong syn-

chronization signal. We hypothesize that in spite of lack of an upstream locking, Halobacterium derives enough energy from the periodic synchronization signal to lock the ribosome into frame and initiate translation.

Table 4.3: Cross species lock energy and magnitude and phase of synchronous signal

16S tail-end→ mRNA↓ Codon# (min,max)↓	Ecoli	Rick	Aquifex	Lacto	Thermo	Sulfo
Ecoli:62222 (46,1518)	L: -2.05 M: 0.47 ϕ: 335 DC:-0.77	L:-1.93 M: 0.33 ϕ: 345 DC:-0.61	L: -2.08 M: 0.47 ϕ: 334 DC:-0.78	L: -2.15 M: 0.38 ϕ: 328 DC:-0.88	L: -1.66 M: 0.40 ϕ: 338 DC:-0.66	L: -1.28 M: 0.38 ϕ: 329 DC:-0.58
Samon:70539 (56,962)	L: -2.41 M: 0.40 ϕ: 346 DC:-0.75	L: -2.30 M: 0.30 ϕ: 2 DC:-0.60	L: -2.42 M: 0.40 ϕ: 345 DC:-0.76	L: -2.49 M: 0.30 ϕ: 340 DC:-0.87	L: -1.59 M: 0.34 ϕ: 348 DC:-0.65	L: -1.30 M: 0.30 ϕ: 337 DC:-0.55
Pseudo:76610 (45,1101)	L: -2.04 M: 0.67 ϕ: 17 DC:-0.87	L: -1.90 M: 0.53 ϕ: 25 DC:-0.71	L: -2.15 M: 0.66 ϕ: 17 DC:-0.88	L: -2.43 M: 0.57 ϕ: 18 DC:-0.97	L: -1.59 M: 0.58 ϕ: 20 DC:-0.73	L: -1.40 M: 0.52 ϕ: 16 DC:-0.64
Rick:46943 (51,907)	L: None M: 0.33 ϕ: 289 DC:-0.49	L: None M: 0.23 ϕ: 300 DC:-0.60	L: None M: 0.35 ϕ: 288 DC:-0.61	L: None M: 0.28 ϕ: 274 DC:-0.71	L: None M: 0.29 ϕ: 289 DC:-0.49	L: None M: 0.29 ϕ: 281 DC:-0.42
Aquifex:64360 (50,1157)	L: -2.24 M: 0.33 ϕ: 353 DC:-0.96	L: -2.23 M: 0.29 ϕ: 13 DC:-0.80	L: -2.31 M: 0.33 ϕ: 351 DC:-1.00	L: -2.40 M: 0.21 ϕ: 350 DC:-1.20	L: -1.78 M: 0.26 ϕ: 352 DC:-0.77	L: -1.28 M: 0.21 ϕ: 345 DC:-0.62
Lacto:62778 (46,1250)	L: -3.29 M: 0.35 ϕ: 330 DC:-0.78	L: -3.11 M: 0.35 ϕ: 344 DC:-0.62	L: -3.41 M: 0.35 ϕ: 328 DC:-0.79	L: -3.74 M: 0.28 ϕ: 320 DC:-0.90	L: -2.49 M: 0.31 ϕ: 328 DC:-0.67	L: -1.74 M: 0.31 ϕ: 320 DC:-0.57
Thermo:57072 (52,1345)	L: -1.44 M: 0.18 ϕ: 318 DC:-0.81	L: -1.06 M: 0.12 ϕ: 350 DC:-0.67	L: -1.45 M: 0.19 ϕ: 319 DC:-0.83	L: -1.58 M: 0.14 ϕ: 304 DC:-0.96	L: -1.27 M: 0.16 ϕ: 318 DC:-0.67	L: -1.16 M: 0.14 ϕ: 310 DC:-0.55
Sulfo:58398 (47,1167)	L: -1.72 M: 0.19 ϕ: 326 DC:-0.74	L: -1.31 M: 0.16 ϕ: 352 DC:-0.61	L: -1.75 M: 0.21 ϕ: 322 DC:-0.76	L: -1.81 M: 0.14 ϕ: 309 DC:-0.90	L: -1.56 M: 0.17 ϕ: 318 DC:-0.60	L: -1.42 M: 0.16 ϕ: 305 DC:-0.49
Halo:79584 (58,1071)	L: -None M: 0.67 ϕ: 17 DC:-1.01	L: None M: 0.58 ϕ: 30 DC:-0.81	L: None M: 0.67 ϕ: 18 DC:-1.01	L: None M: 0.55 ϕ: 20 DC:-1.15	L: None M: 0.58 ϕ:18 DC:-0.86	L: None M: 0.48 ϕ: 13 DC:-0.72

Table 4.4: Cross species lock energy and magnitude and phase of synchronous signal with Halobacterium 16S

16S tail-end→ mRNA↓	Halobacterium
Ecoli:62222	L: -1.93 M: 0.45 ϕ:337 DC:-0.75
Samon:70539	L: -2.16 M: 0.39 ϕ: 169 DC:-0.72
Pseudo:76610	L: -1.86 M: 0.65 ϕ: 18 DC:-0.84
Rick:46943	L: None M: 0.32 ϕ: 290 DC:-0.55
Aquifex:64360	L: -2.06 M: 0.30 ϕ: 353 DC:-0.88
Lacto:62778	L: -2.97 M: 0.33 ϕ: 330 DC:-0.75
Thermo:57072	L: -1.37 M: 0.18 ϕ: 315 DC:-0.76
Sulfo:58398	L: -1.65 M: 0.19 ϕ: 321 DC:-0.67
Halo:79584	L: None M: 0.65 ϕ: 18 DC:-0.98

Chapter 5

Binding Index Locking Model

As we have already talked about in previous chapters, the Shine Dalgarno sequence in the region upstream to the coding region is seen as a consensus sequence over a lot of genes. We have observed and further confirmed by the literature that in E.coli that even though there exists an average binding between the 3' end of 16S and region upstream of mRNA coding region, the individual bindings might not be strong enough for the ribosome to hold on to one position on the mRNA[24][29]. The literature claims this as an absence of Shine Dalgarno sequence in those genes[24]. We have also seen Halobacterium which doesn't have a SD lock up-front of its start codon. These results suggest one of two things. Either SD lock may not be important when the synchronization signal starts in the correct phase. Or the current model for SD locking is not good enough and we need an alternate model for locking which is consistent with our generic model. Hence in this part we try to explain the ribosomal binding at these places based on an alternate *Exponential Binding-Index Locking Model*. The relative time that the small subunit of the ribosome spends at any one particular position of the upstream region might not be long enough for the construction of the entire ribosome. We tried to investigate on the possibility that even though free-energy score at one position may not be enough for the SD lock to happen, the lock could be a cumulative effect of free-energies observed in 5 positions upstream of the start codon. The intuitive approach was to linearly add the free energies for those 5 positions to see if that accumulates enough energy score for the SD lock to happen. This approach had a severe drawback that a lot of background noise add up to indicate false starts. Hence the

summation had to be non-linear in nature, preferably exponential. The exponential nature of the summation was in lieu with constructing a binding energy based exponential filter. This filter would suppress the lower free energy segments which happen at random (noise in our case) while boosting the genuine ones to add up to a large binding index which when compared to a threshold will indicate the SD lock. The biological hypothesis that ties this model to our existing generic model is that the small ribosomal subunit might lock on at 5 consecutive upstream consecutive positions immediately upstream of the start codon for relative periods of time proportional to its free energy binding at those places. The length of the cumulative time that it spends over these sites determines the success of the larger subunit to land on it and the whole ribosome to get constructed.

860 genes from Escherichia coli K-12 genome from the GenBank were used which the Ecogene Database[1] classified as verified protein coding genes[9]. The set of genes was randomly divided into training and test datasets. We took 466 out of 860 genes to be in the training set and the rest of the genes were in the testing set to test the accuracy of the model. The exponential binding index locking function for a position i would be of the form as shown below.

$$F_i = e^{-k \times E_i} - 1 \tag{5.1}$$

In equation 5.1, F_i is the exponential binding index at i^{th} position upstream of the start codon, E_i is the binding energy of 16S tail end with the mRNA stream at i^{th} position (as per the indexing scheme 3.3.1) and k is the binding index constant for the function. The other factors apart from binding energy like temperature, pressure, ion-concentration etc, which could play an important role in the determining of the binding index, are assumed to be constant throughout the analysis. The binding index at 5 positions upstream of the start codon were calculated and added up to find the cumulative binding index for each of the genes as shown in equation 5.2.

$$F = \sum_{i=-1}^{-5} e^{-k \times E_i} - 1 \tag{5.2}$$

If this cumulative F was above a threshold cumulative F_t it qualified for an appropriate upstream lock. The values of the k and the threshold F_t needed to be fixed. While defining this index we fixed the value of $F_t = 1.0$ (define the function threshold) for E=-4.0

[1] http://bmb.med.miami.edu/EcoGene/EcoWeb/

Table 5.1: Comparison of # of genes with an *energy peak as lock* and *binding index based lock*

	# of genes	Energy Lock	Binding Index lock
Parameters		Energy Th=-4.0kCal/mol	k=0.18&F_{th}=0.9
Clockwise	183	136	164
Percentage Passed		74.31%	89.61%
Counter-clockwise	211	160	196
Percentage Passed		75.83%	92.89%
Total	394	75.13%	91.37%

kCal/mole which was a sufficient amount of energy for an upstream lock and calculated k = 0.173286. The k was then varied from 0.18 to 0.15 and for various values of F_t we found the number of genes out of 466 verified ones in our training dataset which would pass for an upstream lock (P_t). We did the same for the whole Escherichia coli K-12 genome by computing cumulative binding indices for every 5 consecutive positions for the entire genome considering each set of 5 places as a prospect for putative binding. Let P_t be the number of genes that pass for a lock for a particular value of k and F_t is 466-P_t. Also let P_{wg} be the number of 5 consecutive sites in the whole genome that would pass for a lock and F_{wg} is total number of 5 consecutive sites P_{wg}. The ratio (P_t/F_t) to (P_{wg}/F_{wg}) was then optimized to get the best value of F_t for each of the k's. The whole idea of this optimization was to increase the true starts and supress the false starts. The plot is shown in figure 5.1. We looked at peaks of the above ratio to fix the values of F_t and k. The values of F_t for which the ratio peaks in the range of 0 to 2.0 or 3.0 units were considered. Based on this analysis the value of k was fixed to 0.19 and F_t at 0.9. Putting this back in equation 5.1 and solving for E we get E = 3.5 kCal/mole which is enough energy for a free energy lock in mRNA secondary structure bindings[7]. Using these values for k and F_t we considered all the 394 verified protein in the test datasets starts to see how many of them would pass for an upstream locking. We then compared these results to the ones we get from visually locating the SD locks as peaks right before the start codons.

. The results are tabulated in 5.1. The results of the genes for the test dataset of 394 genes for the binding index based model were compared with those by visual peak inspection. The Binding Index based model has an accuracy of 91.37% as opposed to 75.83 % in the visual peak inspection. However 38 clockwise and 51 counter-clockwise exceptional

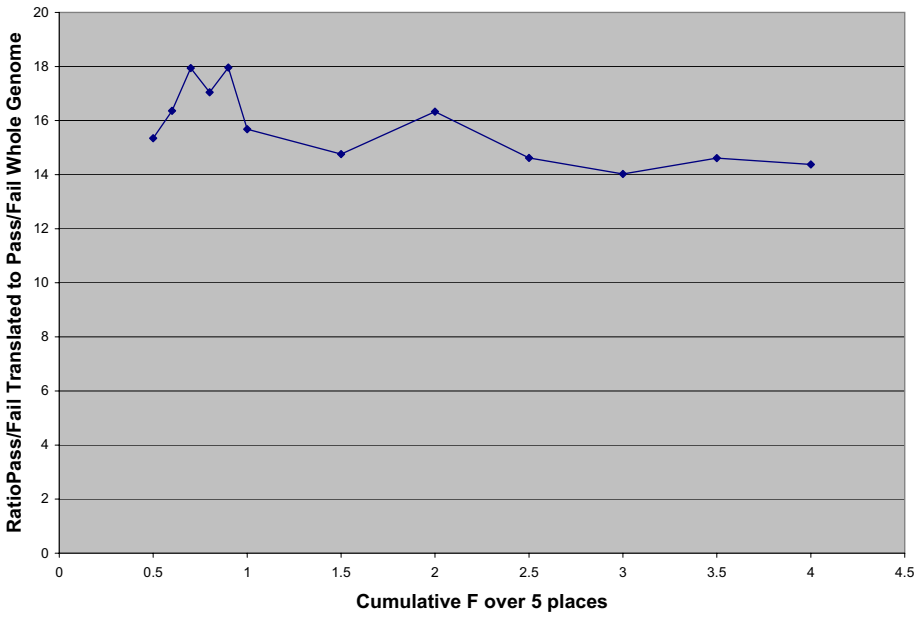

Figure 5.1: Setting up the threshold Binding Index F_t=0.9 for k=0.18 for which the signal to noise ratio is maximum at 17.96

genes out of the total set of 860 verified genes needed explanation. The table A.1 lists the clockwise and the table A.2 lists the counter-clockwise exceptional genes. We then tried to use a filtering process to identify the cause of the exceptions to the model. Based on the plausible reason for their failures the exceptional genes were divided into one of the following categories.

- Some genes are within 5-7 bases ahead of an already existing gene stop might not require the Lock signal.
- Some genes might start at another start codon in frame of the current one downstream.
- Some Genes might have a possible wobble at one base-pair somewhere in the upstream region to give enough energy for locking.
- Finally genes which dont fit in any of the explanations.

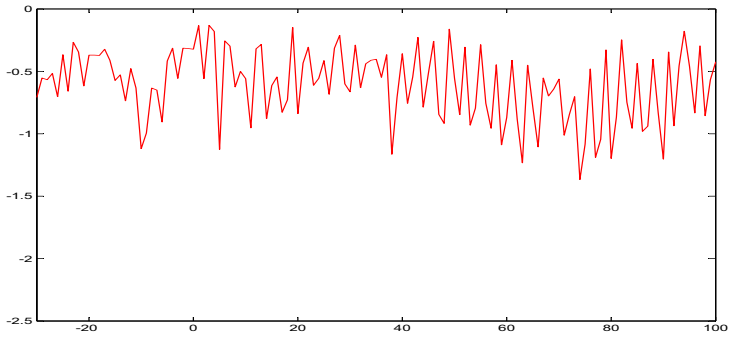

Figure 5.2: The Synchronization signal seen over the ensemble from -30 to 100 base-positions of 36 verified non-locking genes from the E.coli K12. It should be noted that there is no SD locking as shown by the ensemble.

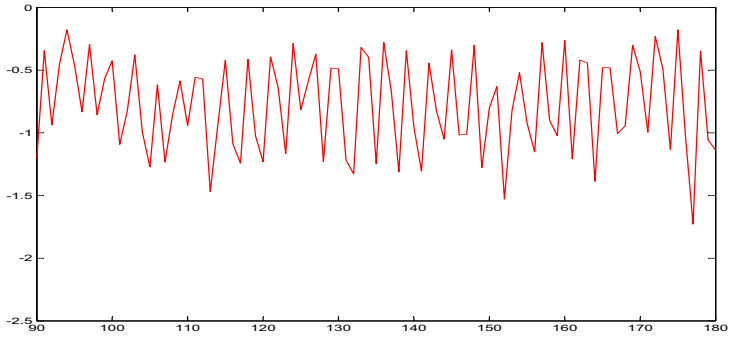

Figure 5.3: The Synchronization signal over the ensemble of 36 verified non-locking genes from the E.coli K12 from 90 to 180 base-positions

Chapter 6

Individual Synchronization Signal Experiments and Results

With the method of extraction of the synchronization signal on a single gene basis, we made an attempt to perfom a set of experiments to better understand the characteristics of the coding region. We started off by investigating the possibility of the synchronization signal being a good indicator of the translation process. Subsequently we tried to apply this individual synchronization signal approach (3.5) to investigate frameshifting.

6.1 Synchronization Signal and Codon Bias

In the subsequent sections we performed a set of experiments in an attempt to see if the synchronization signal, though a product of the codon bias, is a more precise indicator of the coding region.

6.1.1 Synchronization Signal Begin Middle and End

To analyse of the synchronization signal strength over the length of any gene, we performed a position specific vector analysis over an ensemble of 390, more than a

1000 base-pair long verified genes. We aligned all the genes at the start, middle and end maintaining the frame of translation. The genes were aligned on the start codons and the synchronization signal vector magnitude was found for each of them over a 100 base-pairs. The mean and variance around the mean was calculated for the starting vector magnitude. We did a similar analysis for the middle of the gene aligning all the genes on their respective half-length maintaining frame and the distribution was analyzed. Finally the procedure was repeated for the last 100 base-pairs of all the genes aligning them over the stop codons. The calculations were similar to those discussed in 3.5 by taking a 100 base-pairs worth of energy to compute the vector each time. The results are as shown in 6.1. It is clear that the means of the beginning synchronization signal vector distribution is off by the means of the mid and end synchronization signal vector distributions. This suggests that the synchronization signal is specific to its position along the length of the gene.

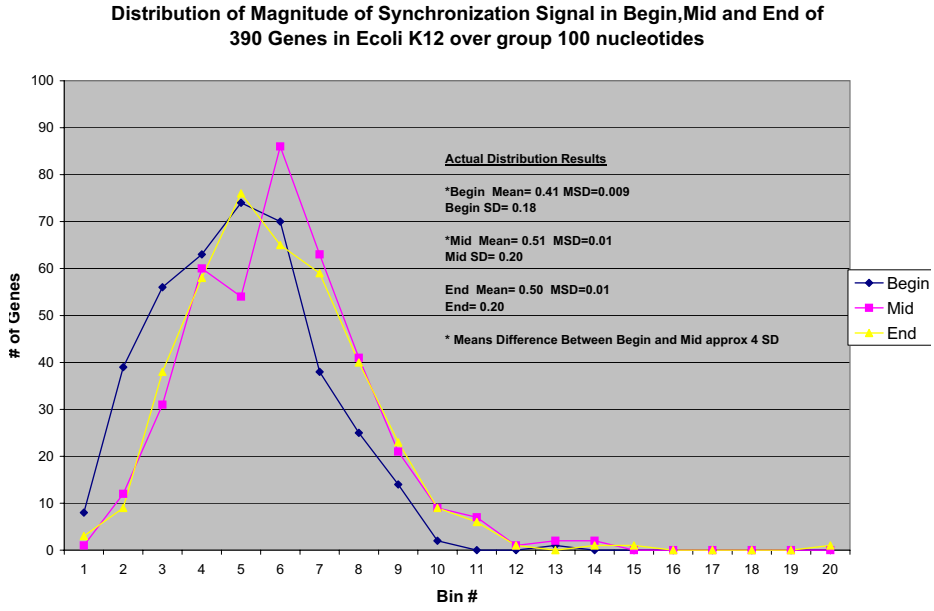

Figure 6.1: Synchronization Signal over beginning 100, mid 100 and end 100 base-pairs over 390 verified long(>1000 base-pairs) genes

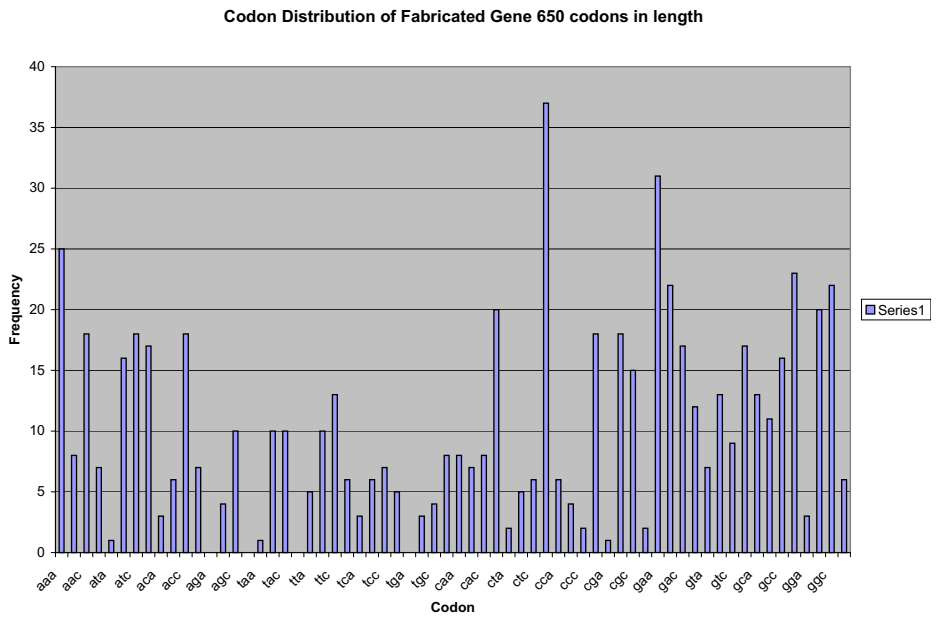

Figure 6.2: Codon Distribution of the fabricated gene(650 codons) which copies the "codon dialect" of E.coli species

6.1.2 Codon Shuffling Experiment and Results

The codon bias as discussed in section 1.4 depends on the availability of tRNAs. The frequency of $1/3^{rd}$ may lead us to think that the syncrhonization signal is just an outcome of the codon bias in the genes. We started to explore if synchronization signal was an outcome of the codon bias or something more fundamental in nature. The idea behind the experiment was to see if the synchronization signal was a good indicator for the process of translation , or it was just another measure of the codon bias.

To start with the experiment 25 verified genes, more than a 1000 bases in length were taken from the E.coli and each one of them were independantly shuffled on per codon-basis not in any specific order. This means that we took the first codon of a gene and swapped it with a different codon at a randomly chosen position in the same gene. This was done repeatedly till we were able to shuffle the whole gene. This process was repeated for all the 25 genes. These shuffled genes were treated as putative genes and free energy analysis was performed on them by moving over the tail-end of the 16S rRNA individually on each one of them and calculating free-energy for each of those positions. Now the

ensemble position specific averages were taken as already discussed in section 3.4. There was an average synchronization signal present over the ensemble. This is because the effect of the shuffling of the genes is lost over the ensemble of genes. Hence it was required to do the analysis on the individual genes and not over the ensemble. Vector analysis as already discussed in the section 3.5, was used for this signal analysis for the individual genes. The idea of evaluating the genes individually was to see changes in the synchronization signal keeping exactly the same codon bias as the actual genes. This was done by shuffling the codons in each of the 25 genes, in many different ways and treating all the shuffled permutations as putative genes. The free energy was calculated for each of those putative genes. We did that for the first 51 bases and then repeated for increments of 51 cumulatively calculating the magnitude and phase of the vector.

We further investigated the average codon bias over 390 verified and long E.coli genes 6.2. We used this distribution to select 650 codons representing the codon bias of the species. We used these 650 codons arranged in different orders to generate hypothetical genes. We found the cumulative synchronization signal vector for each of them. Figures 6.3 thru 6.5 show the cumulative synchronization signal vectors for three such genes. This was to see if the neighboring codons had a second order effect on the cumulative synchronization signal. As seen in the figures the 650 long fabricated genes produced cumulative synchronization signal vectors with different magnitude and phase reflecting this second order effect.

By selectively choosing sequences of codons we can achieve a different magnitude and a different phase for the synchronization signal for the same codon distribution as the average bias across the species. Figure 6.3 is the 650 codon-long fabricated E.coli gene with the general phase direction at roughly 0° which is quite different from the general direction of vector for the E.coli genes. The fabricated gene was then shuffled to get two sets of different vectors starting in different phase from itself with different final magnitude as shown in figures 6.4 and 6.5. So all the 3 fabricated genes have exactly the same codons but different magnitude and phase of the synchronization signal over their lengths.

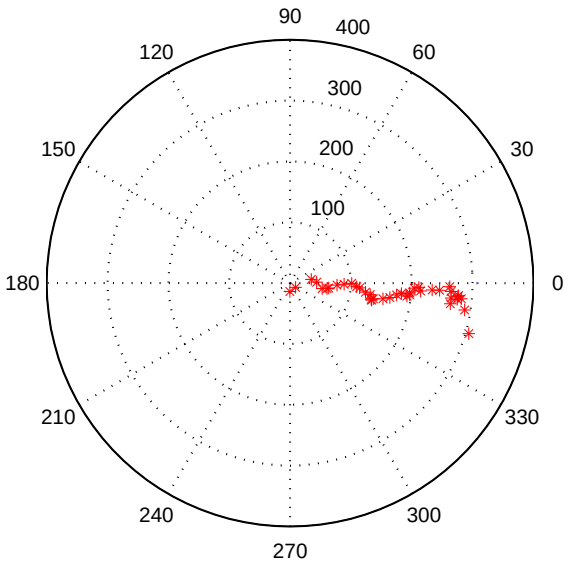

Figure 6.3: Fabricated Gene I with 650 codons with same codon bias as E.coli K12

6.2 Frameshifting Experiment and Results

In order to illustrate the capabilities of our individual vector based methodology, we focus on two E. Coli genes. One nonframe shifting, and one that exhibits frameshifting. Both are verified genes. The non- frameshfiting gene is thrA. The frameshifting gene is prfB. Programmed +1 frameshifting in E.Coli prfB gene exemplifies the major features of +1 frameshifting identified in [22]. It has a purine rich (Shine-Dalgarno type) region 3 nucleotides upstream of the slippage codon CUU (25th codon, k=24), a stop codon (UAG, k=25) next to the slippage codon, and the relatively good stability of the peptidyl-tRNA-GAG with the +1 overlapping UUU codon as compared to slippage codon CUU [22][2]. Our hypothesis is that there would be a change in the behavior of the prfB synchronization signal in this region. To investigate this we first performed signal analysis over the length of the non-frameshifting thrA. The results are shown in figures 6.6 and 6.7. We then compare this set of results with the behavior of the signal for the frameshifting prfB gene. The latter is shown in figures 6.8 and 6.9.

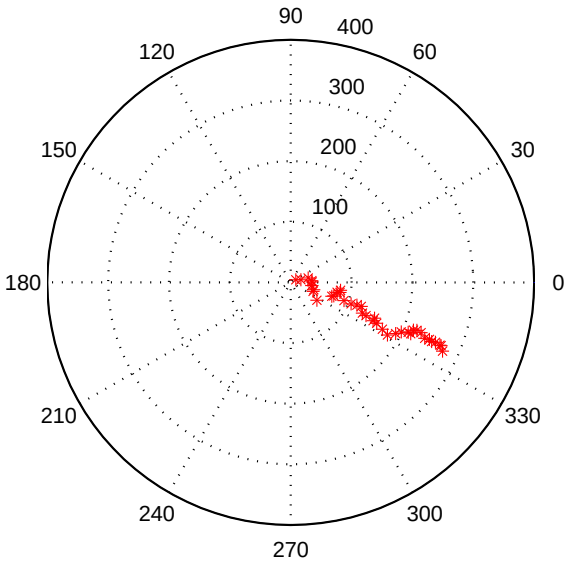

Figure 6.4: Fabricated Gene II with 650 codons with same codon bias as E.coli K12 at a different synchronization signal than Fabricated Gene I

6.2.1 Role of Synchronization Signal in identification of slippage site in frameshifting gene prfB

Figures 6.6 and 6.7 show the magnitude (in kCal/mol) and phase (in radians) of the signal as it accumulates over the first 100 bases of the thrA gene. There are two things to note. One is that the phase has a transient region which extends about 30 to 40 codons downstream from the start codon. It then stabilizes and attains an asymptotic value close to zero. This indicates that the ribosome tail remains in phase with, or in the same frame as, the start codon. The other item to note is that the value of Mk grows in an almost monotonic fashion. This indicates that the signal vector contributions from the individual codons tend in the same direction. Note that the average signal magnitude is M_k/k,where k is the number codons. Similar behavior of the signal magnitude and phase was noted for other non-frameshifting E. Coli genes we examined.

Figures 6.8 and 6.9 show the magnitude (in kCal/mol) and phase (in radians) of the signal for the prfB frameshifting gene accumulated over its first 100 codons. There are two important differences with respect to the behavior of the signal in non-frameshifting genes. First, while the phase is close to zero4 up to about 30 to 40 codons downstream

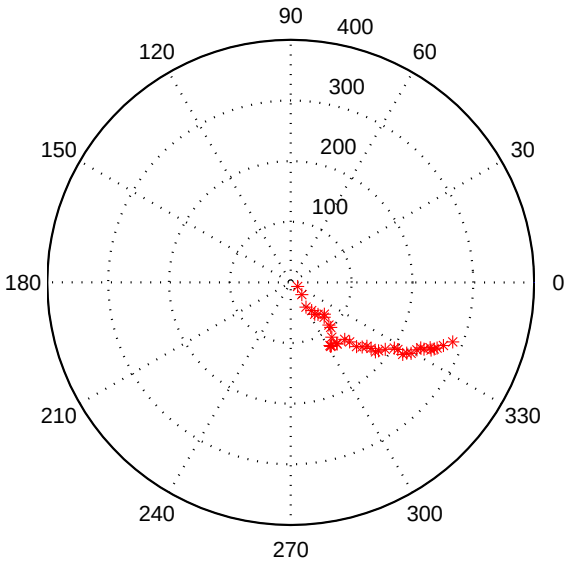

Figure 6.5: Fabricated Gene III with 650 codons with same codon bias as E.coli K12 at a different synchronization signal than Fabricated Gene I

from the start codon, it then changes to achieve an asymptotic value around 4 radians or about 230 degrees. This implies +1 change in the reading frame, and is consistent with the observed change in prfB reading frame that occurs around the 26th codon. The second thing to note (Figure 6.8) is that there is a peak in the magnitude of the signal vector around codon 25 followed by a nearly monotonic growth in the vector magnitude of the accumulated signal around codon 40. The accepted slippage codon is at k=24 (25th codon) [22][2]. This is an indication that the differential signal vector changed direction at the frameshift point. The terminal phase, as seen in Figure 6.9, is heading to an asymptotic value approximately 120 degrees from that observed for the thrA gene.

Another way of directly checking for a change in the phase angle of the signal is to estimate the differential vector phase by producing a sliding window summation over some number of codons, say 10 codons. Figure 6.10 shows the phase of the differential signal. We can see that some 16 to 30 codons downstream , the phase starts changing from a value around zero (and in phase with the start codon) to a phase around 4 radians (or about 230 degrees). Again, this is consistent with the observed frameshift at codon 26.

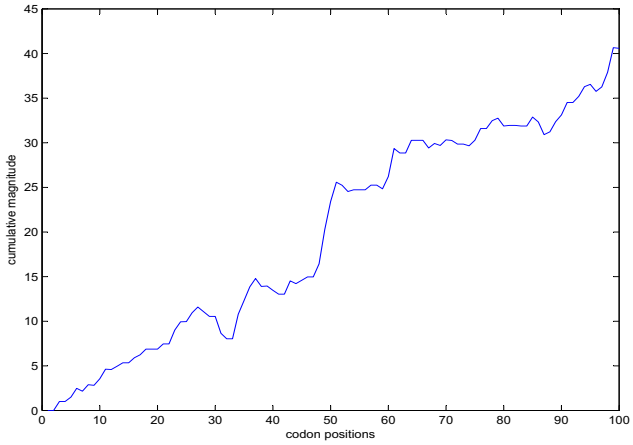

Figure 6.6: Magnitude of Cumulative Synchronization Signal vector for the first 100 codons in non-frameshifting thrA gene of E.coli

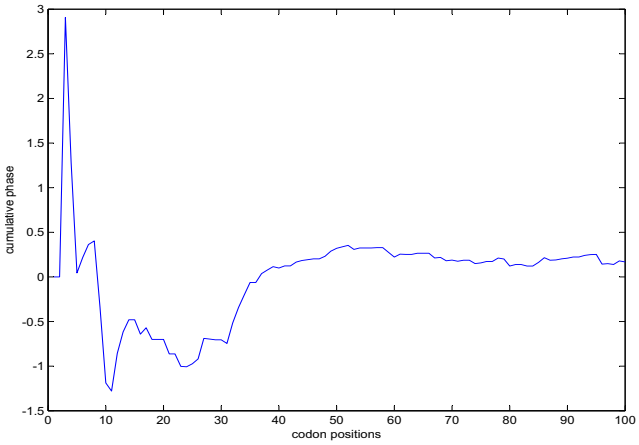

Figure 6.7: Phase of Cumulative Synchronization Signal vector for the first 100 codons in non-frameshifting thrA gene of E.coli

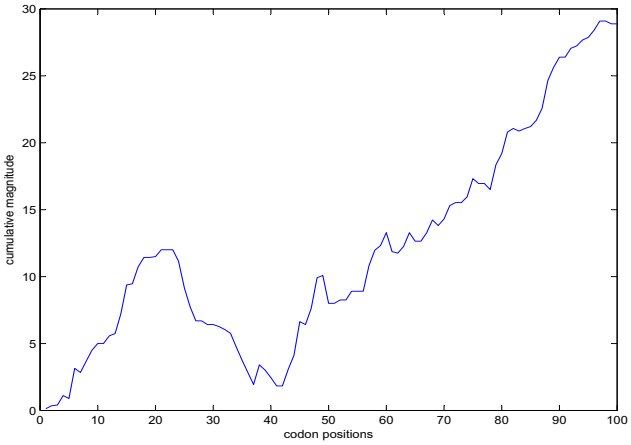

Figure 6.8: Magnitude of Cumulative Synchronization Signal vector for the first 100 codons in frameshifting prfB gene of E.coli

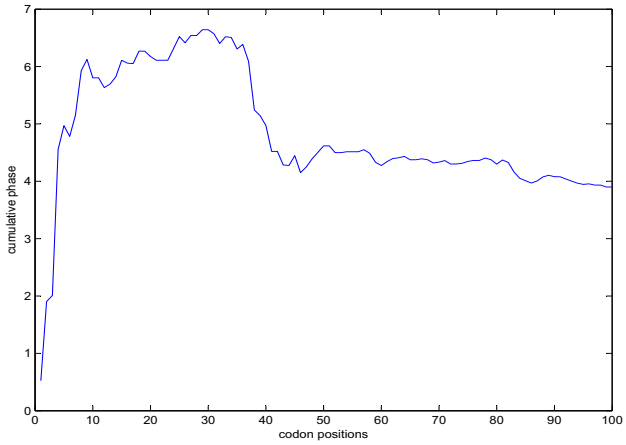

Figure 6.9: Phase of Cumulative Synchronization Signal vector for the first 100 codons in frameshifting prfB gene of E.coli

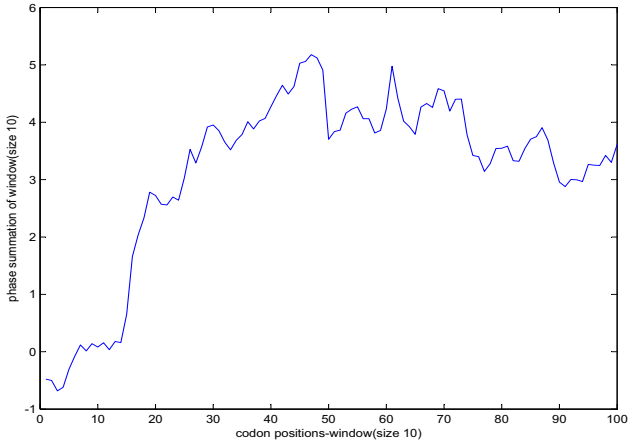

Figure 6.10: Phase of Differential Synchronization Signal vector for the first 100 codons(moving window 10 codons wide) in frameshifting prfB gene of E.coli

Chapter 7

Signal Relationship with Shine Dalgarno Locking

During the previous chapters we have seen that there is a set of genes which are verified but don't show a SD lock in the region upstream of the start codon. We also identified a synchronization signal over their ensemble not just downstream as shown in figures 5.2 and 5.3. In an attempt to throw some more light on the ribosome's behavior in the translation initiation process, we attempted to explore the possibility of a synchronization signal before the start codon. This was done for all the 860 genes in general and then specifically for the 36 genes which didn't have a SD lock in any one of them. Vector analysis was done over them in the region around the start codon. The magnitude and phase for the synchronization signal accumulated every base triplet (third harmonic of the signal) was calculated from -54th base position (-18^{th} base triplet or codon) to 180 base positions downstream of start codon over the ensemble average. The results for the 36 genes seen from the figure 7.1 (also see figure 7.2) show that there is a synchronization signal in the upstream region just before the start codon. However the behavior was missing over the ensemble of 860 genes (figure 7.3). This goes with the fact that the synchronization signal is not present outside the coding region.

To analyze the possible relation of the absence of synchronization signal in the

upstream region to the SD lock frame with respect to the start of coding region we repeat the analysis over 3 sets of genes. 100 genes from the 860 genes which had the maximum free-energy for the SD lock in frame position 1, 2 and 3 were chosen to be in each of the 3 sets. Frame position 3 corresponds to the -1^{st} position of our free energy indexing scheme. Similarly the frame positon 2 and 1 correspond to the -2^{nd} and -3^{rd} position respectively. The cumulative magnitude and phase for frame position 1 are shown in figures **??** and A.20. The cumulative magnitude and phase for frame position 2 are as shown in figures A.23 and A.24. Finally figures A.27 and A.28 show the same for the ensemble in frame 3. The figures referred above show the mean behavior of the cumulative magnitude and phase in each of the 3 frames for the 3 ensembles from -54^{th} to 180^{th} base position. They also show the standard deviations around the means at each point as a seperate plot in each of the figures. The details of the calculations of mean cumulative magnitude and phase and standard deviations around these means are shown in subsection 7.1.

The cumulative vector at position i shows the effect of the synchronization signal over a group of positions. Differential analysis of these cumulative vectors gives the instantaneous value of the synchronization signal at any codon positon. The synchronization signal doesn't always move the ribosome by exactly 3 base pairs. This differential vector can be viewed as a correction signal to the synchronization signal that readjusts the normal 3 base-pair shift at a time. The differential magnitude and phase calculations are covered in subsection 7.2. The results for the differential magnitude and phase for frame 1 are shown in figures A.21 and A.22. Differential magnitude and phase for frame 2 are shown in figures A.25 and A.26. Finally the figures A.29 and A.30 illustrate the differential magnitude and phase for the genes which lock in frame 3.

From the vector plots A.20,A.24 and A.28, we can see that the starting phase for the set of genes in each of the 3 starting positions is statistically different. The phase of the synchronization signal tries to achieve the final equilibrium phase for the gene.

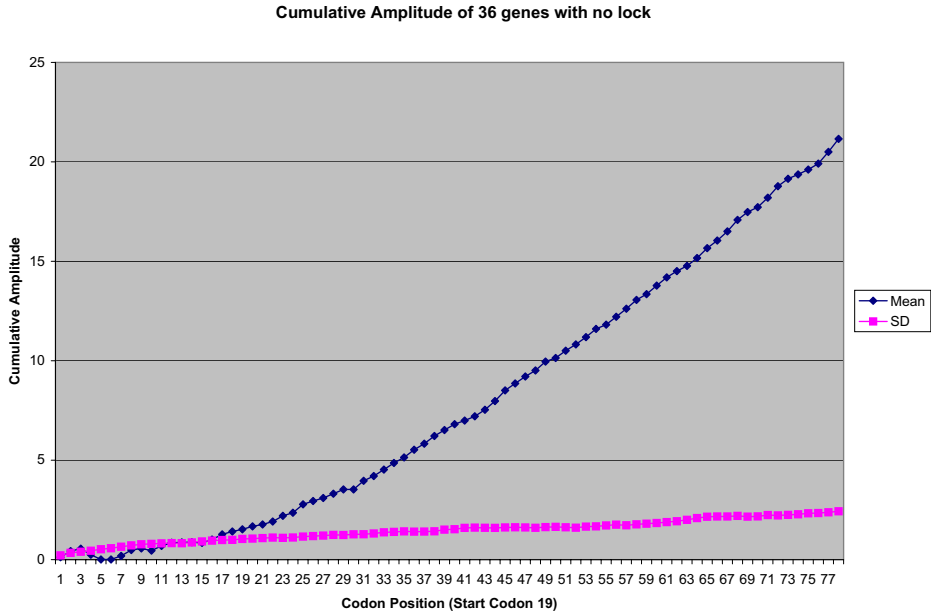

Figure 7.1: Magnitude accumulated over from -18 codons to +60 codon position for an average of 36 verified genes with no SD lock

7.1 Mean and Standard Deviations of the Cumulative Ensemble Vector

Consider for any particular codon position c_i there are vectors for each of the 100 genes with magnitude M_i^k and phase ϕ_i^k for the k_{th} gene in the ensemble of 100. The vector calculations are already covered in the part 3.5.

We can find from calculations from before that for the vector $Me^{j\phi}$, the x and y components are given by equation 7.1.

$$\begin{aligned} M_i^k sin(\phi) &= a_i^k \\ M_i^k cos(\phi) &= \frac{b_i^k - c_i^k}{\sqrt{3}} \end{aligned} \tag{7.1}$$

Now the means of the 2 components of the vectors are given by equation 7.2.

$$\overline{M_i sin(\phi)} = E_y = \frac{1}{p} \sum_{i=1}^{p} a_i^k$$

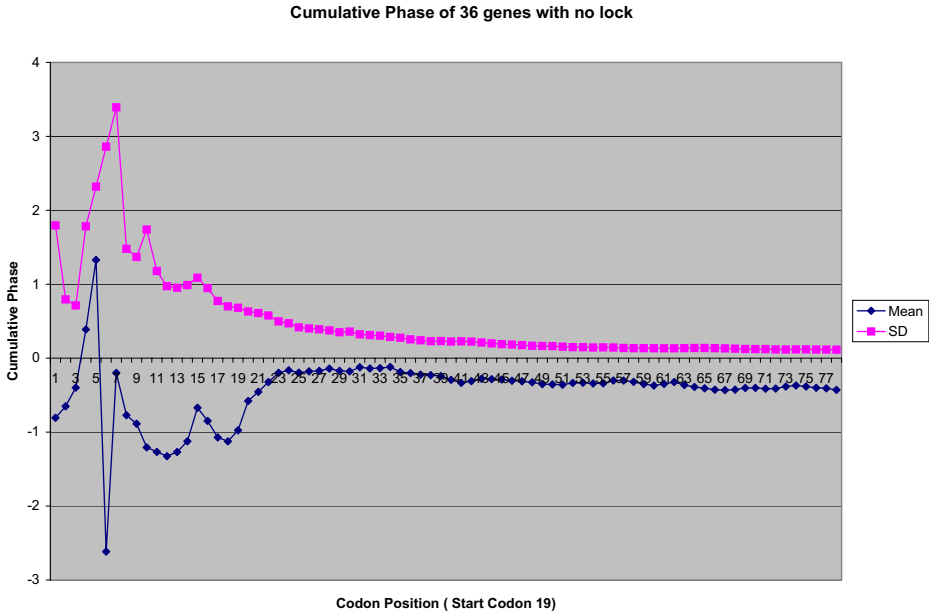

Figure 7.2: Phase accumulated over from -18 codons to +60 codon position for an average of 36 verified genes with no SD lock

$$\overline{M_i cos(\phi)} = E_x = \frac{1}{p}\sum_{i=1}^{p}\frac{b_i^k - c_i^k}{\sqrt{3}} \tag{7.2}$$

The standard deviations for the set of vectors are given by equation 7.3.

$$\begin{aligned}\sigma_{M_i sin(\phi)} &= \frac{1}{\sqrt{p-1}}\sqrt{\sum_{i=1}^{p}(a_i^k - E_y)^2} \\ \sigma_{M_i cos(\phi)} &= \frac{1}{\sqrt{p-1}}\sqrt{\sum_{i=1}^{p}(\frac{b_i^k - c_i^k}{\sqrt{3}} - E_x)^2}\end{aligned} \tag{7.3}$$

The Standard deviations around the means are given by equation 7.4.

$$\begin{aligned}\sigma_{E_y} &= \frac{1}{\sqrt{p}}\sigma_{M_i sin(\phi)} \\ \sigma_{E_x} &= \frac{1}{\sqrt{p}}\sigma_{M_i cos(\phi)}\end{aligned} \tag{7.4}$$

We approximate the standard deviation around the mean magnitude using the equation 7.5.

$$\sigma_{M_i} = \sqrt{{\sigma_{E_y}}^2 + {\sigma_{E_x}}^2} \tag{7.5}$$

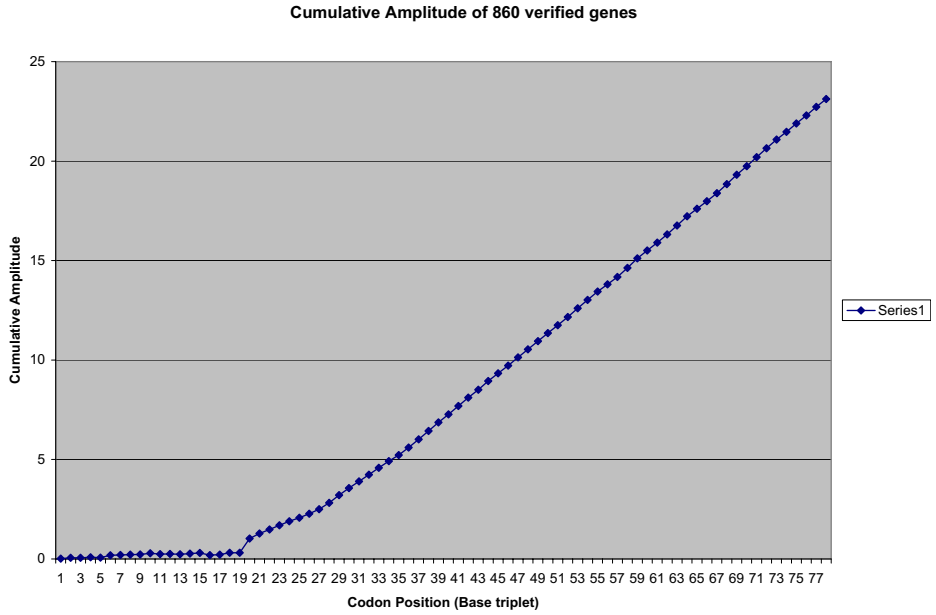

Figure 7.3: Magnitude accumulated over from -18 codons to +60 codon position for an average of 860 verified genes

Also to get the standard deviation around the mean phase we use 7.6.

$$\sigma_{\phi_i} = \frac{\sigma_{M_i}}{M_i} \tag{7.6}$$

7.2 Differential Magnitude and Phase Calculations

In order to see the differential behavior of the corrective vector of synchronization signal, we found the vector differential of the cumulative magnitude and cumulative phase using principles of vector calculus. The cumulative vector at any particular codon position i can be represented as,

$$F_c = Me^{j\phi} \tag{7.7}$$

Next we find the change of F_c with respect to the codon position(C)

$$\begin{aligned} \frac{\delta F_c}{\delta C} &= \frac{\delta F_c}{\delta M}\frac{\delta M}{\delta C} + \frac{\delta F_c}{\delta \phi}\frac{\delta \phi}{\delta C} \\ &= e_{j\phi}[\frac{\delta M}{\delta C} + jM\frac{\delta \phi}{\delta C}] \end{aligned} \tag{7.8}$$

The $\frac{\delta M}{\delta C}$ and $\frac{\delta \phi}{\delta C}$ in equation 7.8 are the slopes of the cumulative magnitude and phase plots. We use a mathematical procedure for finding the best fitting straight line to a given set of points by minimizing the sum of the squares of the offsets ("the residuals") of the points from the curve. Using this method of Least Square Fitting, we get the slopes of the two cumulative plots. The line is fitted over 3 sets of points (1 thru 3 say).

$$\frac{\delta M_2}{\delta C} = \frac{\sum_{i=1}^{3} 3C_i A_i - \sum_{i=1}^{3} C_i \sum_{i=1}^{3} A_i}{\sum_{i=1}^{3} 3C_i^2 - (\sum_{i=1}^{3} C_i)^2} \tag{7.9}$$

$$\frac{\delta \phi_2}{\delta C} = \frac{\sum_{i=1}^{3} 3C_i \phi_i - \sum_{i=1}^{3} C_i \sum_{i=1}^{3} \phi_i}{\sum_{i=1}^{3} 3C_i^2 - (\sum_{i=1}^{3} C_i)^2} \tag{7.10}$$

The differential of F_c with respect to C 7.8 is a vector in itself with a differential magnitude(M_d) and a differential phase (M_ϕ).

$$M_d = \sqrt{(\frac{\delta M}{\delta C})^2 + (\frac{\delta \phi}{\delta C})^2} \tag{7.11}$$

$$\phi_d = \phi + tan^{-1} \frac{M \frac{\delta \phi}{\delta C}}{\frac{\delta M}{\delta C}} \tag{7.12}$$

Chapter 8

Discussion and Conclusion

The current work was under the basic assumption that the free-energy interaction of the 16S tail-end with the underlying mRNA sequence is an indicator of the state of the process of translation. We investigated the concept of at least 2 signals occurring in the translation process, namely, the lock signal and the periodic synchronization signal. Synchronization signal, although related to and generated by the codon bias appears to have additional significance in the translational process.

The possibility of the lock and the synchronization signals was explored across an array of species from the bacteria and archae kingdoms. As discussed in section 4.1, it was found that all the 9 species spread over the bacteria and the archae kingdoms displayed the synchronization signal for the set of genes we considered. The strength(Magnitude M) of the synchronization signal varied from one species to the other. However for any one species the magnitude doesn't vary a lot when matched against the 16S tail-ends other than its own. Similarly the average phase (ϕ) of each of the species for tail-ends different than its own, lied in the close vicinity of the phase obtained with its own 16S tail-end. Similar behavior over the different 16S tail-ends for any particular species implies that the 16S is a good filter to read off information from the underlying mRNA sequences. The presence of synchronization signal suggests the possibility of cross-species protein production considering all the other factors like temperature etc are not taken into account.

The results of the cross-species experiment were logical in the sense that the species organized themselves based on the biological taxonomy. The lock and synchronization signal

of bacterial and archaeal kingdoms showed different characteristics for the lock and the synchronization signal except for some exceptional species.

All the species in the bacteria kingdom like E.coli, Salmonella, Lactobacills, Pseudomonas, and Aquifex had a decently strong lock signal and a strong synchronization signal. This goes with them being in the same bacterial kingdom. Rickettsia, was the exception to this observation, lacking the SD lock signal and a weak synchronization signal. Rickettsia needs a further investigation to investigate the presence of the lock and the synchronization signals in its ancestors. In the archae kingdom, Thermoplasma and Sulfolobus showed a medium lock and medium synchronization signal. Halobacterium was the outlying in the archaeal subset. It doesn't have show a noticeable SD lock signal but shows a strong synchronization signal. This led us to investigate into the locking signal in more detail including alternate models of locking. We were also interested in lock's relationship with the synchronization signal to see if the latter compensates for the missing lock. The search for SD locking took us back to the E.coli to see if all the verified protein-coding genes had a SD lock signal up-front. The experiment is covered under the chapter 5. This was of primary importance for the identification of a ribosomal-binding-site by a ribosome. We investigated the 860 verified protein-coding genes to see an upstream lock signal. It should be pointed out that a spike in the free energy spectrum at position -1,-2 or -3 noticeably more than the background noise was considered a SD lock. Based on our calculations, we couldn't find it in a lot of the verified genes. This wasn't too surprising as the literature talks about a reasonable number of genes lacking the SD sequence in the upstream region.

We suggested that it may be that the free-energy needed for a ribosome to construct in the upstream region is not just derived from one point. We defined the upstream lock to be an exponential integral of energies accumulated over 5 upstream positions. This new term that we were measuring at 5 upstream positions was called the Binding Index. If the total binding index accumulated over 5 upstream positions was more than a threshold, it was considered a lock. This threshold was fixed to maximize the true start sites while suppressing the false ones.

One possible physical interpretation of the exponential binding index is that it measures the relative time that the ribosome spends in an upstream position. On similar lines, the total binding index can be interpreted as the relative time that the ribosome needs to construct itself and start translating.

Having seen the universal presence of the synchronization signal over the coding-

regions of an array of species with a frequency of $1/3^{rd}$, we wanted to explore if the synchronization signal was an effect of the abundance of the tRNAs (in other words the codon bias). The question was whether the neighboring codons have a second order effect on the generation of the synchronization signal. We can see how this makes the synchronization signal different from the codon bias which just measures the frequencies of each type of codon in a particular region. If this was the case, the synchronization signal was a better indicator of the state of translation than the codon bias. We conducted a set of experiments to see which one of the synchronization signal and the codon bias is a more comprehensive indicator of the coding region.

We started off by analyzing the distribution of the synchronization signal in the beginning of genes, in their middle regions, and towards their ends. As suggested by the results in section 6.1.1, the means of the magnitude distributions of the synchronization signal in beginning of genes compared to those in the middle of the genes are significantly different. The standard deviations around the means being really small suggest that the two are significantly different. Numerically the synchronization signal is statistically weaker in the beginning than in the middle. The middle and the ends of the genes show statistically similar mean magnitudes of the synchronization signal.The non-uniformity of the synchronization signal along the length of the genes hints on the synchronization signal being a more comprehensive feature of coding regions than the codon bias.

In section 6.1.2 we shuffled the codons of around 25 genes and took ensemble average of energies as shown in section 3.4 and examined the average free energy plot. The ensemble of the 25 shuffled genes had a synchronization signal in them. This was effectively because the synchronization signal maintains its magnitude and phase in the generally correct direction in each of the individual genes. Hence when we see the average effect of those individual vectors over the ensemble the synchronization signal is observed. The synchronization signals however when analyzed individually for each of those genes were varying around the mean direction of the ensemble synchronization signal. With an aim to remove the variance of the codon distribution, we fabricated a gene with 650 codons (the distribution for the 650 also made to be the general E.coli codon dialect for the species in general) and shuffled it many different ways to see if we get any difference in the magnitude and phase of the synchronization signal keeping the exact same codons. In section 6.1.2 we have given 3 examples of such shuffling which used exactly the same codons and yet a different magnitude and phase of the synchronization signal. This strongly suggests that

the neighboring codons have a second order effect on the synchronization signal which may not be captured by the codon bias.

A conceivable drawback of using codon bias in a single gene is is hard to compute when the number of codons is small (<= 64). However, we can derive a synchronization signal from just one gene as suggested by the section 3.5. This makes the synchronization signal to more precisely indicate the process. The identification of the +1 frameshifting site in the prfB gene of E.Coli demonstrates the use of the synchronization signal in individual genes.

Finally in chapter 7 we tried to explain why there was some synchronization signal upstream of the start codon for the 36 non-locking genes,(7.1) and not over the species in general(??) by investigating the relationship of the lock and the synchronization signals. The synchronization signal vectors before the start of the 824 genes were in different direction with comparable amplitudes. Hence their vector addition over the ensemble had almost zero amplitude before the start codon. The amplitude grew noticeable after the start codon, implying the lining up of the synchronization signal vectors in a common direction on the average. Hence we are able to see the synchronization signal in the coding regions and not in the non-coding regions. However for the 36 non-locking genes, the synchronization signal vector wasn't randomly distributed which got some amplitude even before the start codon over the ensemble. This was a really interesting observation as it hinted on the comparability of the upstream synchronization signal and the SD lock. In section 7 we took 3 sets of genes locking in frames 1,2 and 3 respectively but aligned on the the same reading frame. The cumulative phase over the ensemble over each of the 3 sets turned out to be statistically different. This means that the upstream ensemble synchronization signal vectors for each of the 3 sets has different initial phases. This better explained the difference in behavior of the cumulative amplitude plots over the ensemble being different in frame 2 as compared to the frame 1 and frame 2. Considering the 824 genes to be locked in either of the 3 frames, taking the average over the whole ensemble is averaging out vectors in different phases with comparable magnitudes. Hence this shows up as almost a zero vector for the upstream region over the average of 824 genes.

The current work helped us better understand the synchronization signal, and tried to show how good an indicator of the translational process it is. The frameshifting site identification of prfB gene in E.Coli demonstrates the ability of the process to identify characteristics of coding region in general, particularly, we believe, any kind of frameshifting.

This work needs to be extended in eukaryotes to see if the synchronization signal is preserved in differently spliced exons. One can also evaluate the synchronization signal as a tool for identifying intron-exon sites and the way the exons can be spliced together.

Furthermore work needs to be done to set up some wet-lab experiments to test the model and improve it more.

Bibliography

[1] C.B. Burge and S. Karlin. Finding the genes in genomic dna. *Curr. Opin. Struct. Biol.*, 8:346–354, 1998.

[2] J.F. Curran. Analysis of effects of trnas; message stability on frameshift frequency at the escherichia coli rf2 programmed frameshifting site. *Nucleic Acid Res.*, 21:1837–1843, 1993.

[3] M.A. Vouk D.I. Rosnick, D.L. Bitzer and E.E. May. Free energy periodicity in e.coli. *The First Joint BMES/EMBS conference*, 2:1216, 1999.

[4] M.A. Vouk D.I. Rosnick, D.L. Bitzer and E.E. May. Free energy periodicity in e.coli coding. In *Proceedings of* 22^nd *Annual EMBS International Conference*, pages 2470–2473, July 2000.

[5] M.A. Vouk D.I. Rosnick, D.L. Bitzer and E.E. May. Escherichia coli protein coding sequence detection by analysis of free energy periodicity. *GENSIPS*, 2002.

[6] D.L. Bitzer E.E. May, M.A. Vouk and D.I. Rosnick. Ribosome as a table-driven convolutional decoder for escherichia coli k-12 translation initiation system. In *Proceedings of* 22^nd *Annual EMBS International Conference*, pages 24766–2469, July 2000.

[7] A. Krogh et al. A hidden markov model that finds genes in e.coli dna. *Nucleic Acids Res.*, 22(22):4768–4778, 1994.

[8] B.E. Schoner et al. Role of mrna translation efficiency in bovine growth hormone expression in escherichia coli. *Nat. Acad. of Sci.*, 81(17):5403–5407, 1984.

[9] D.A. Benson et. al. Genbank. *Nucleic Acid Res.*, 26(1):1–7, 1988.

[10] J. A. Jaeger et al. Improved predictions of secondary structures for rna. *Proc. Natl. Acad. Sci.*, 86:7706–7710, 1989.

[11] M. Walker et al. A comparative genomic method for computational identification of prokaryotic translation initiation sites. *Nucleic Acids Res.*, 30(14):3181–3191, 2002.

[12] R. Crea et al. Chemical synthesis of genes for human insulin. *Proc.. Natl. Acad. Sci.*, 75(12):5765–5769, 1978.

[13] R.B.Weiss et al. Reading frame switch caused by base-pair formation between the 3' end of the 16s rrna and the mrna during elongation of protein synthesis. *EMBO J.*, 7:1503–1507, 1998.

[14] R.Grantham et al. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, 9:43–74, 1981.

[15] S. Altschul et al. Gapped blast and psi-blast: a new generation of protein database search algorithm. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[16] S.F. Altschul et al. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[17] S.G. Bonitz et al. Codon recognition rules in yeast mitochondria. *Proc.. Natl. Acad. Sci.*, 77:3167–3170, 1980.

[18] S.L. Salzberg et al. Microbial gene identification using interpolated markov models. *Nucleic Acid Res.*, 26:544–548, 1998.

[19] S.M. Freier et al. Improved free-energy parameters for predictions of rna duplex stability. *Proc. Natl. Acad. Sci.*, 83:9373–9377, 1986.

[20] V. Veljković et al. Is it possible to analyze dna and protein sequences by methods of digital signal processing? *IEEE Transactions on Biomedical Engineering*, 32(5):337–341, 1985.

[21] George Garrity et.al. *The Bergey's Manual of Systematic Bacteriology.Second Ed. Vol One"*. Springer-Verlag, 2001.

[22] P.J. Farabaugh. Programmed translational frameshifting. *Annual Rev. Genet.*, 30:507–528, 1996.

[23] T. Ikemura. Codon usage and trna content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2(1):13–34, 1985.

[24] A. S. Lewin. *Ribsomes.* Kluwer Academic Plenum Publishers, 1999.

[25] B. Lewin. *Genes VI.* Oxford University Press, 1994.

[26] M.Borodovsky and J. McIninch. Genemark: Parallel gene recognition for both dna strands. *Computers Chem.*, 17(2):123–133, 1993.

[27] M.A. Vouk P. Bermel, E.E. May and D.L. Bitzer. On the import of the shine dalgarno seriesto the expression of mrna sequences. Department of Computer Science, North Carolina State University.

[28] R.Grantham. Working of the genetic code. *Trends Biochem. Sci.*, 5:327–331, 1980.

[29] D.I. Rosnick. *Free Energy Periodicity and Memory Model for Genetic Coding.* PhD thesis, North Carolina State University, 2001.

[30] J. Shine and L. Dalgarno. The 3'-terminal sequence of e.coli 16s ribosomal rna:complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci.*, 71:1342–1346, 1974.

[31] E. Nadir T. Schurr and H. Margalit. Identification and characterization of e.coli ribosomal binding sites by free energy calculation. *Biochimie*, 74:357–362, 1992.

[32] T.D.Schneider. Informational content of individual genetic sequences. *J. theor.Biol.*, 189:427,441, 1997.

[33] T.A. Thanaraj and M.W. Pandit. An additional ribosome-binding site on mrna of highly expressed genes and a bifuntional site on the colicin fragment of 16s rrna from escherichia coli: important determinants of the efficiency of translation-initiation. *Nucleic Acid Res.*, 17:2973–2985, 1989.

[34] E.N. Trifonov. Recognition of the correct reading frame by the ribosome. *Nucleic Acids Res.*, 21:4019–4023, 1993.

Appendix A

Appendix

This appendix consists the plots for the *ensemble synchronization signal cross species experiment*. It also has the plots for the *signal relationship with Shine Dalgarno lock*. Finally it consists a set of tables specifying the genes related to *binding index based locking model*.

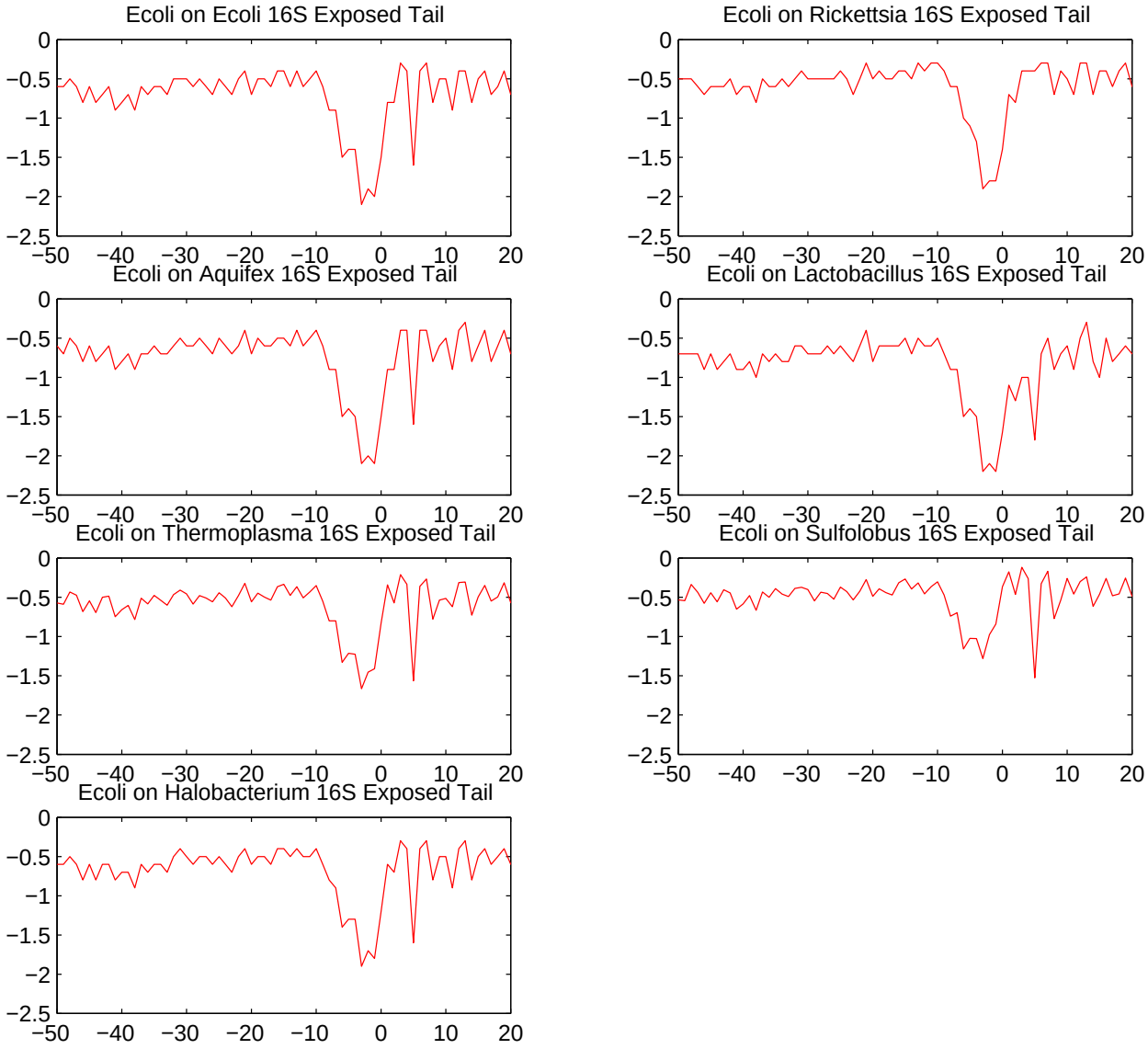

Figure A.1: Average Free Energy over 200 genes for E.coli mRNA with all 16S Tails: Lock Signal

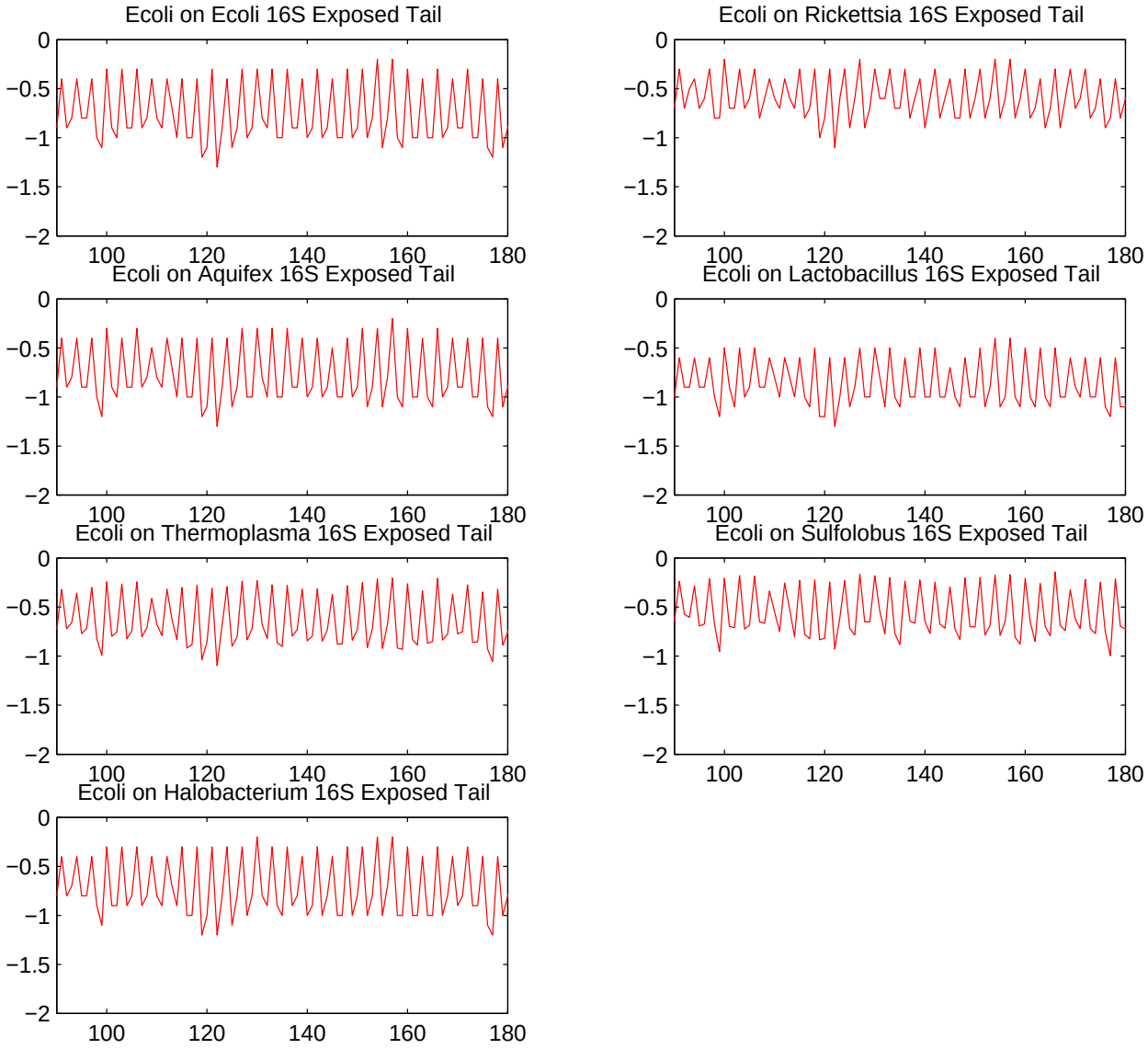

Figure A.2: Average Free Energy over 200 genes for E.coli mRNA with all 16S Tails: Synchronization Signal

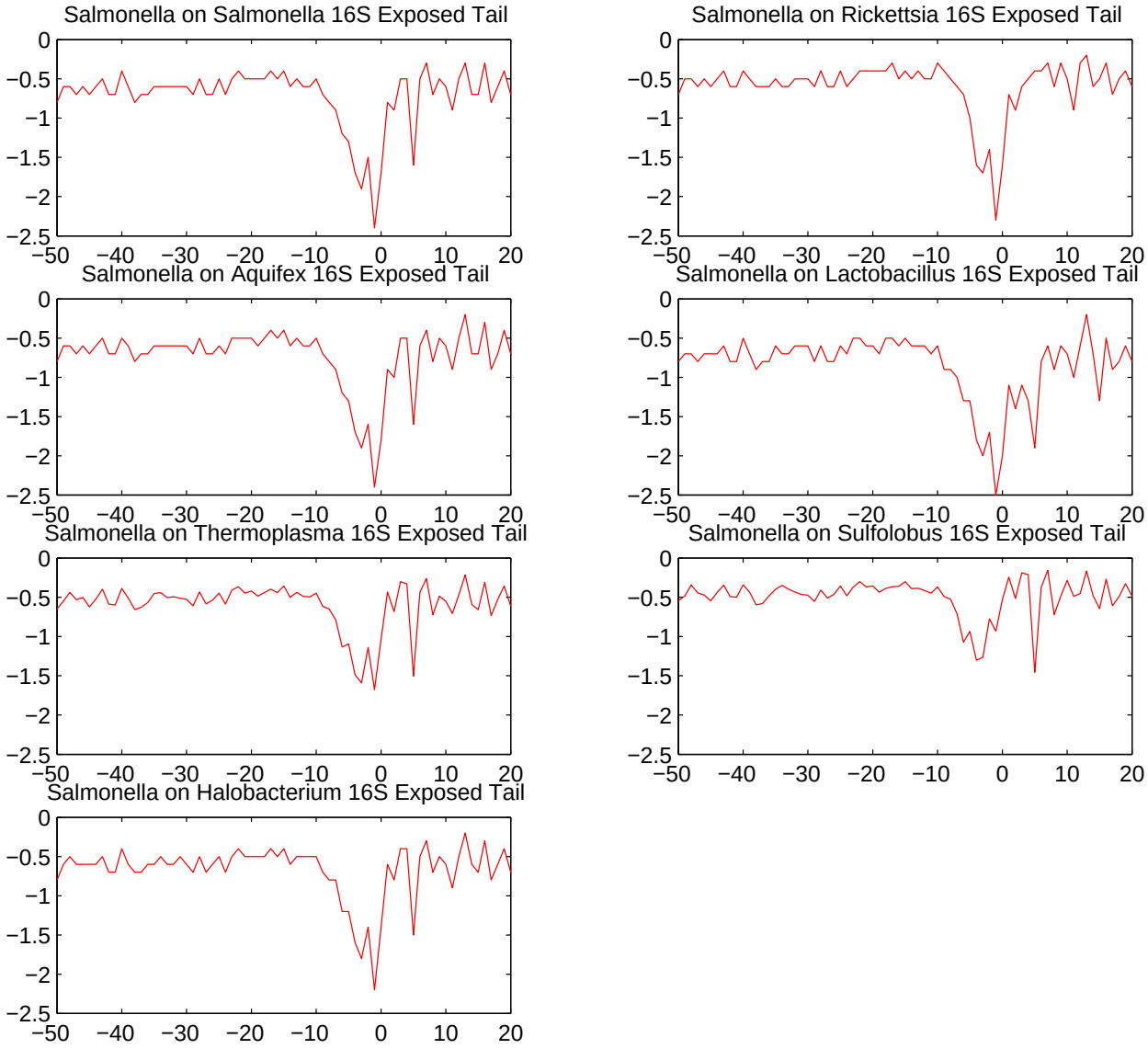

Figure A.3: Average Free Energy over 200 genes for Salmonella mRNA with all 16S Tails: Lock Signal

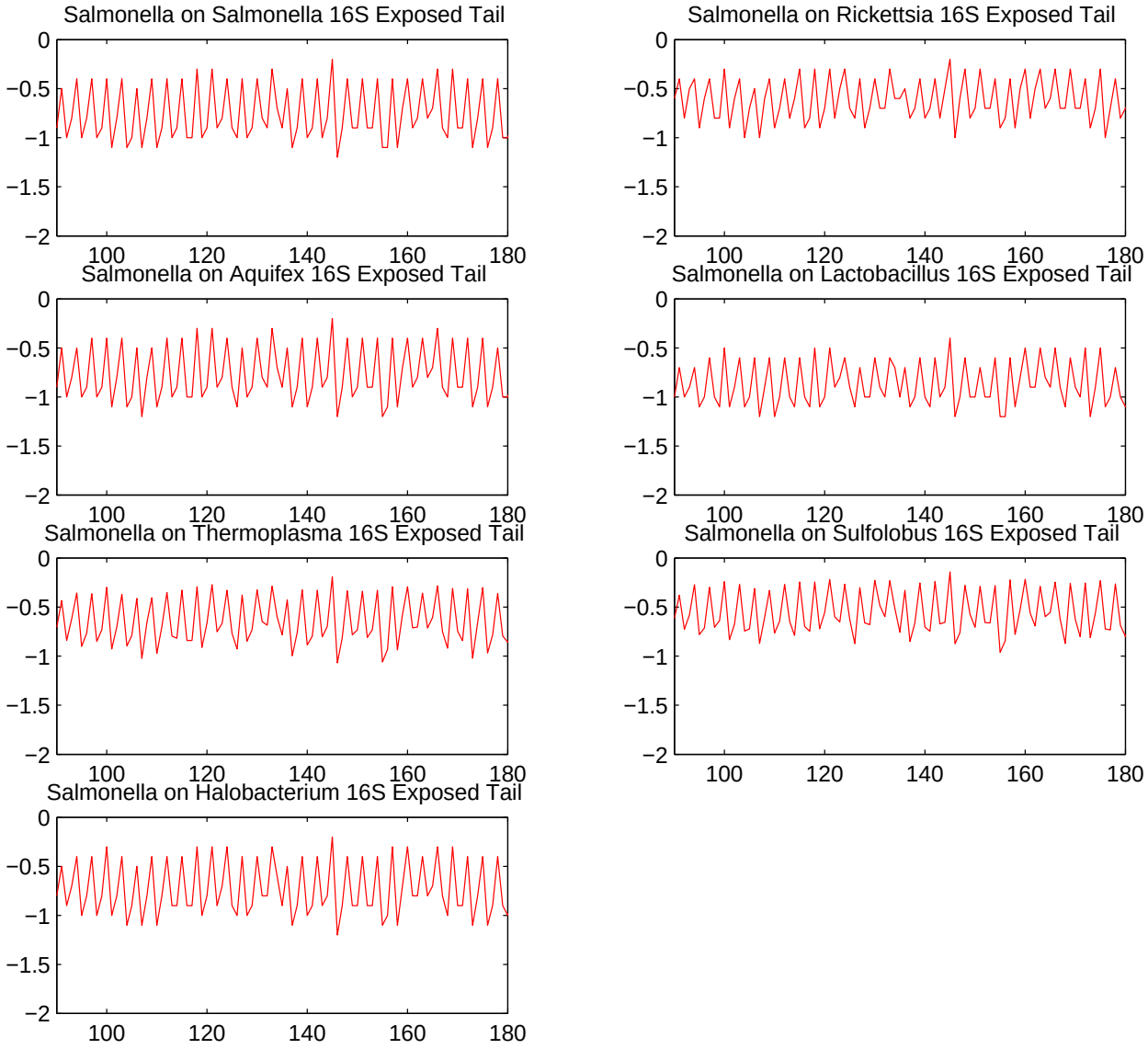

Figure A.4: Average Free Energy over 200 genes for Salmonella mRNA with all 16S Tails: Synchronization Signal

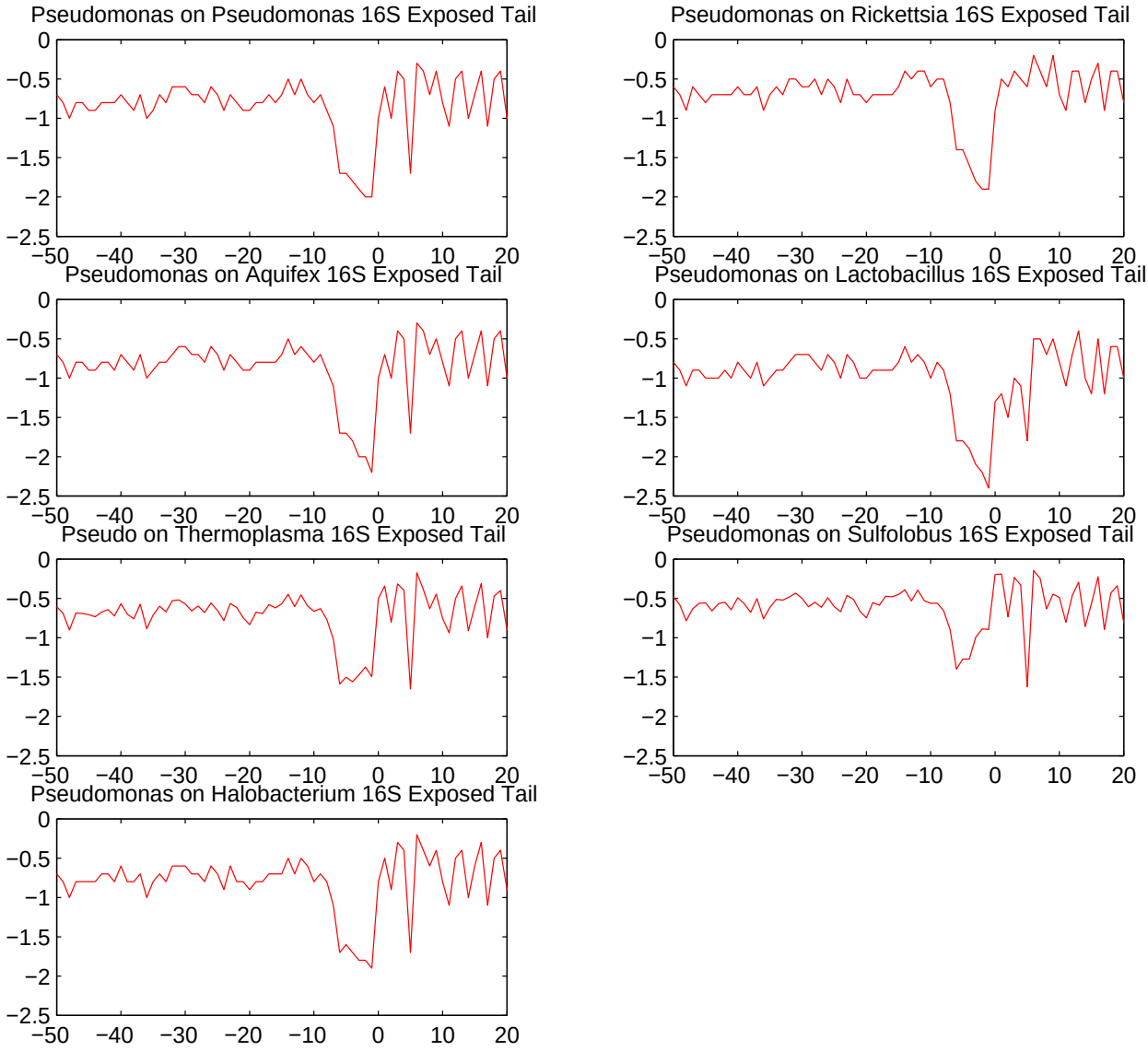

Figure A.5: Average Free Energy over 200 genes for Pseudomonas mRNA with all 16S Tails: Lock Signal

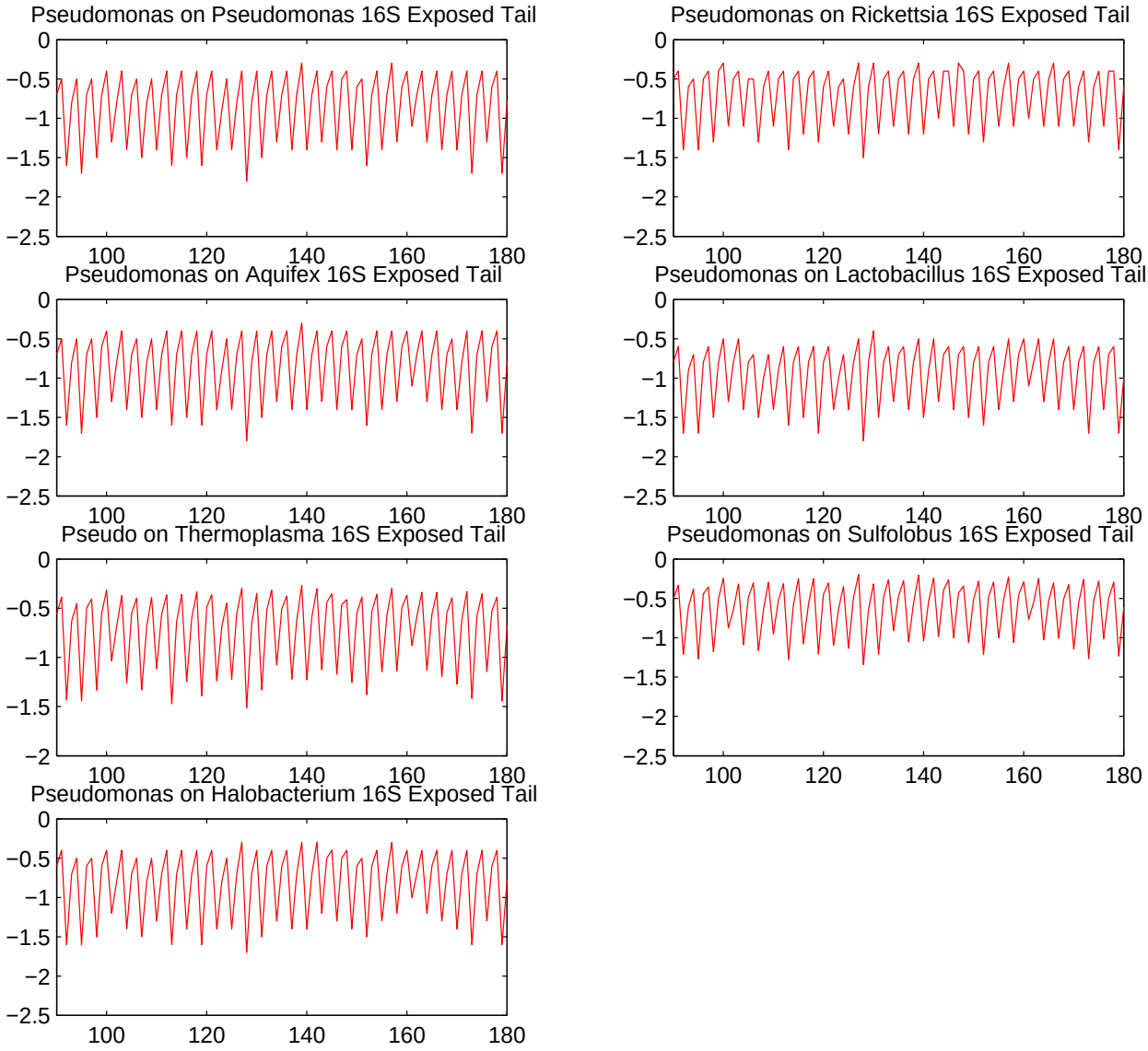

Figure A.6: Average Free Energy over 200 genes for Pseudomonas mRNA with all 16S Tails: Synchronization Signal

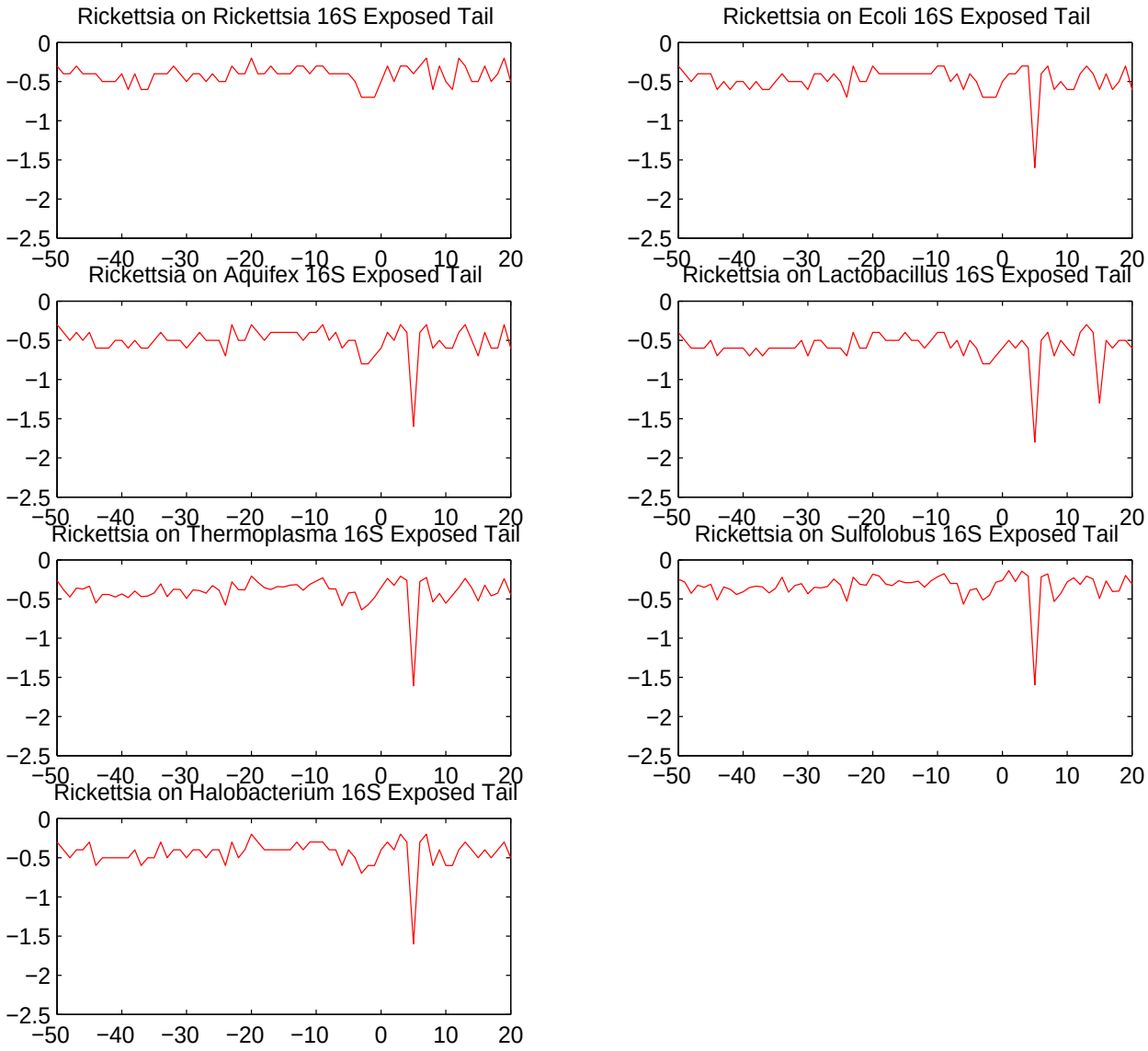

Figure A.7: Average Free Energy over 200 genes for Rickettsia mRNA with all 16S Tails: Lock Signal

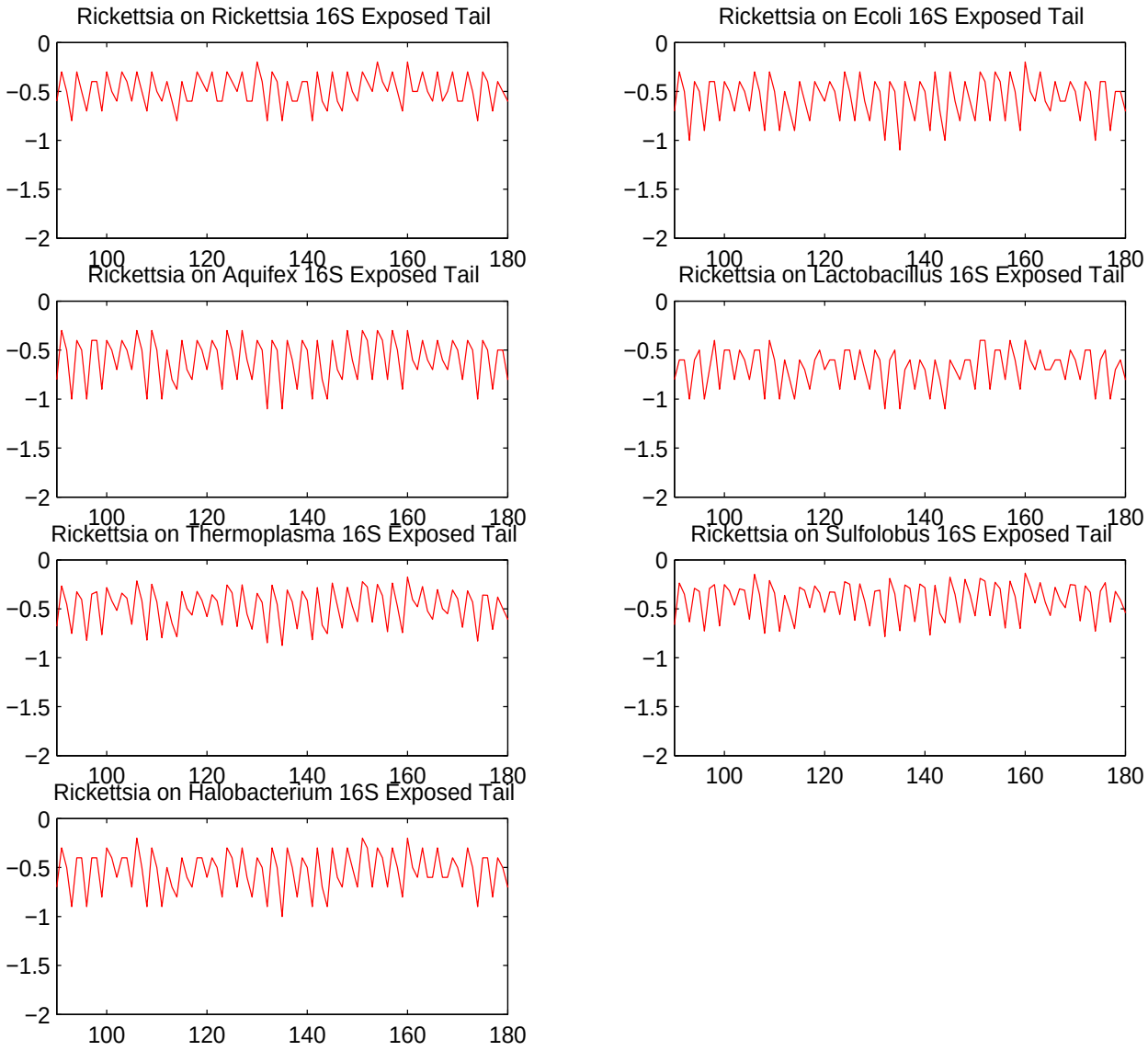

Figure A.8: Average Free Energy over 200 genes for Rickettsia mRNA with all 16S Tails: Synchronization Signal

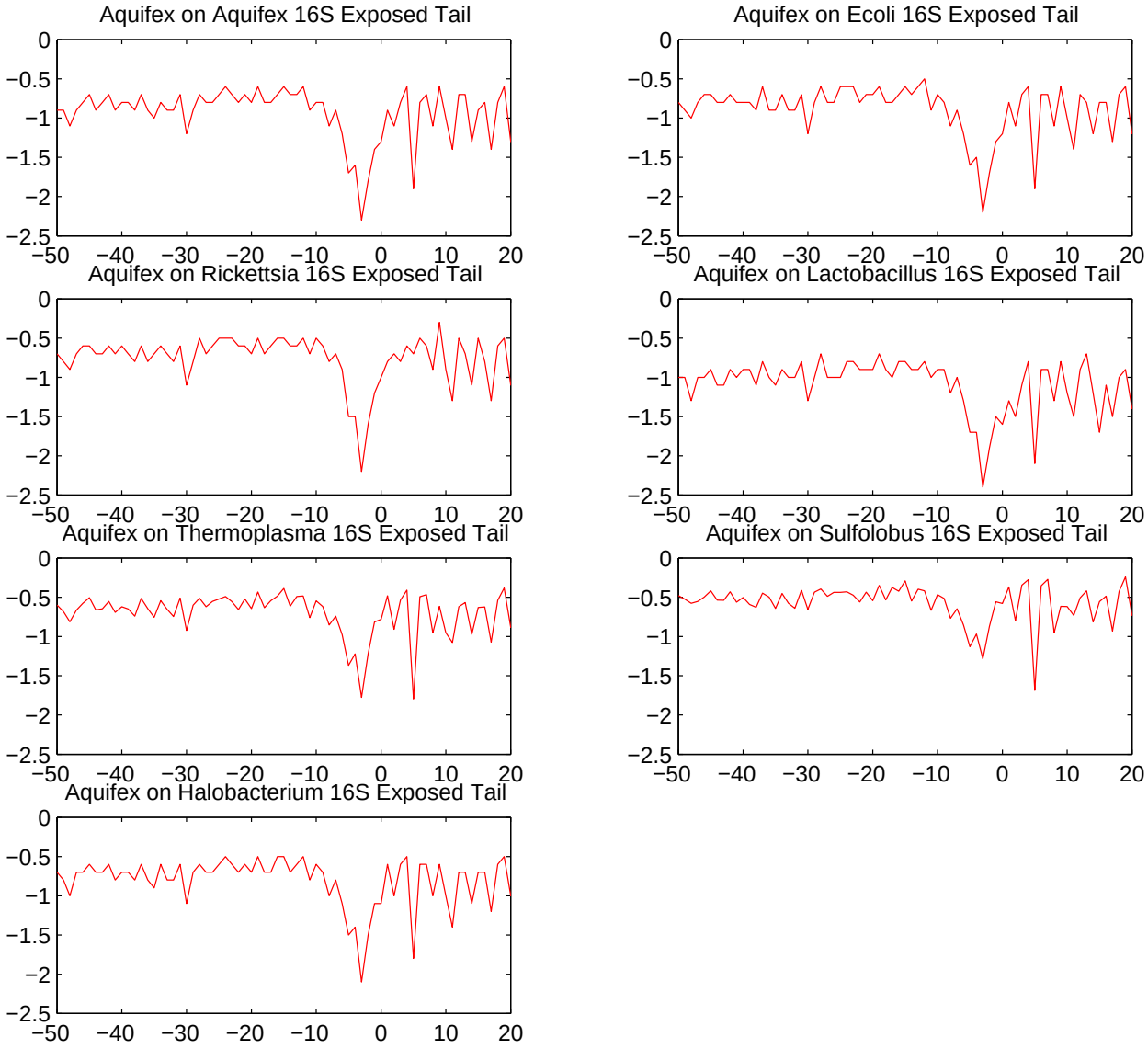

Figure A.9: Average Free Energy over 200 genes for Aquifex mRNA with all 16S Tails: Lock Signal

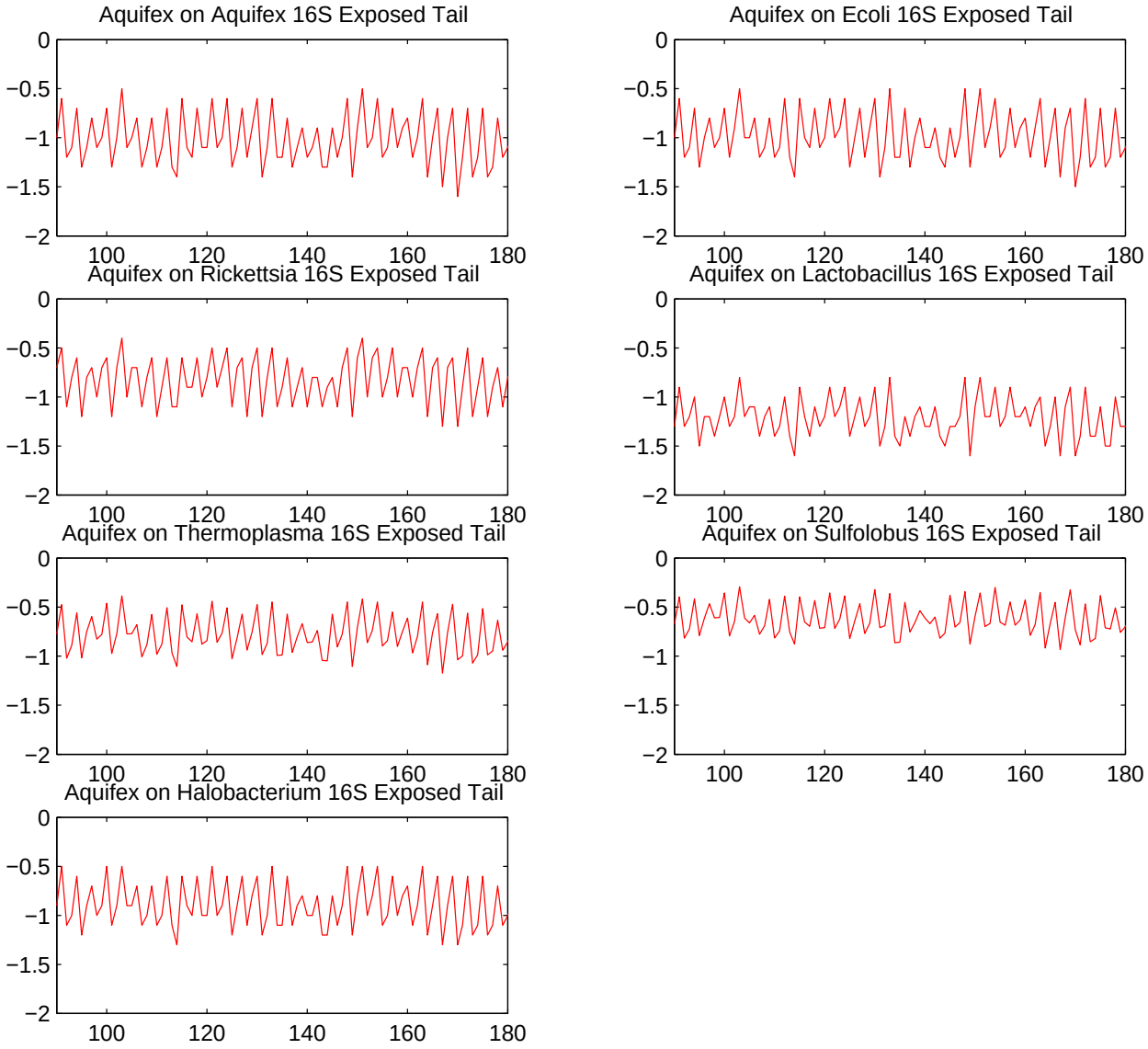

Figure A.10: Average Free Energy over 200 genes for Aquifex mRNA with all 16S Tails: Synchronization Signal

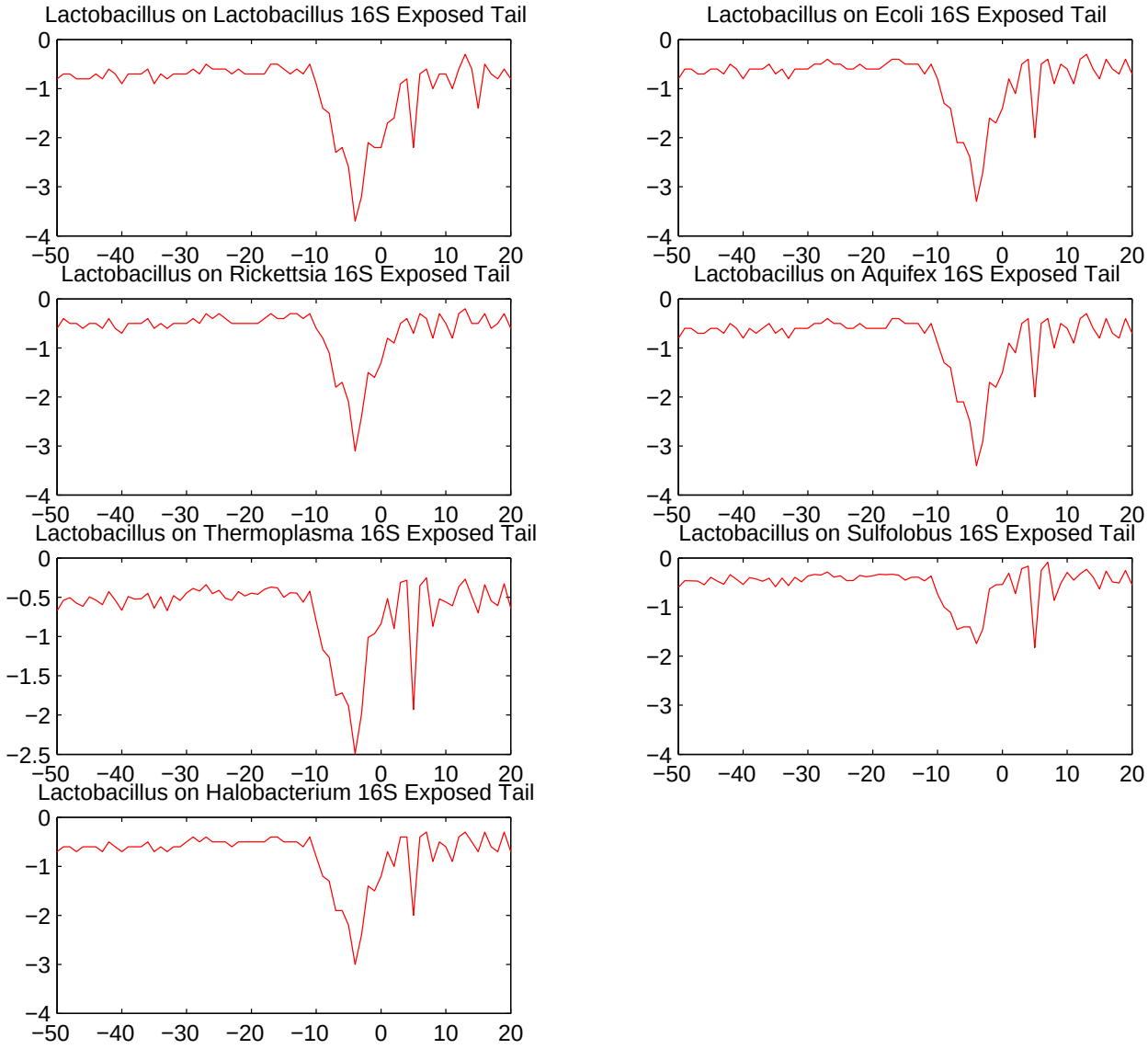

Figure A.11: Average Free Energy over 200 genes for Lactobacillus mRNA with all 16S Tails: Lock Signal

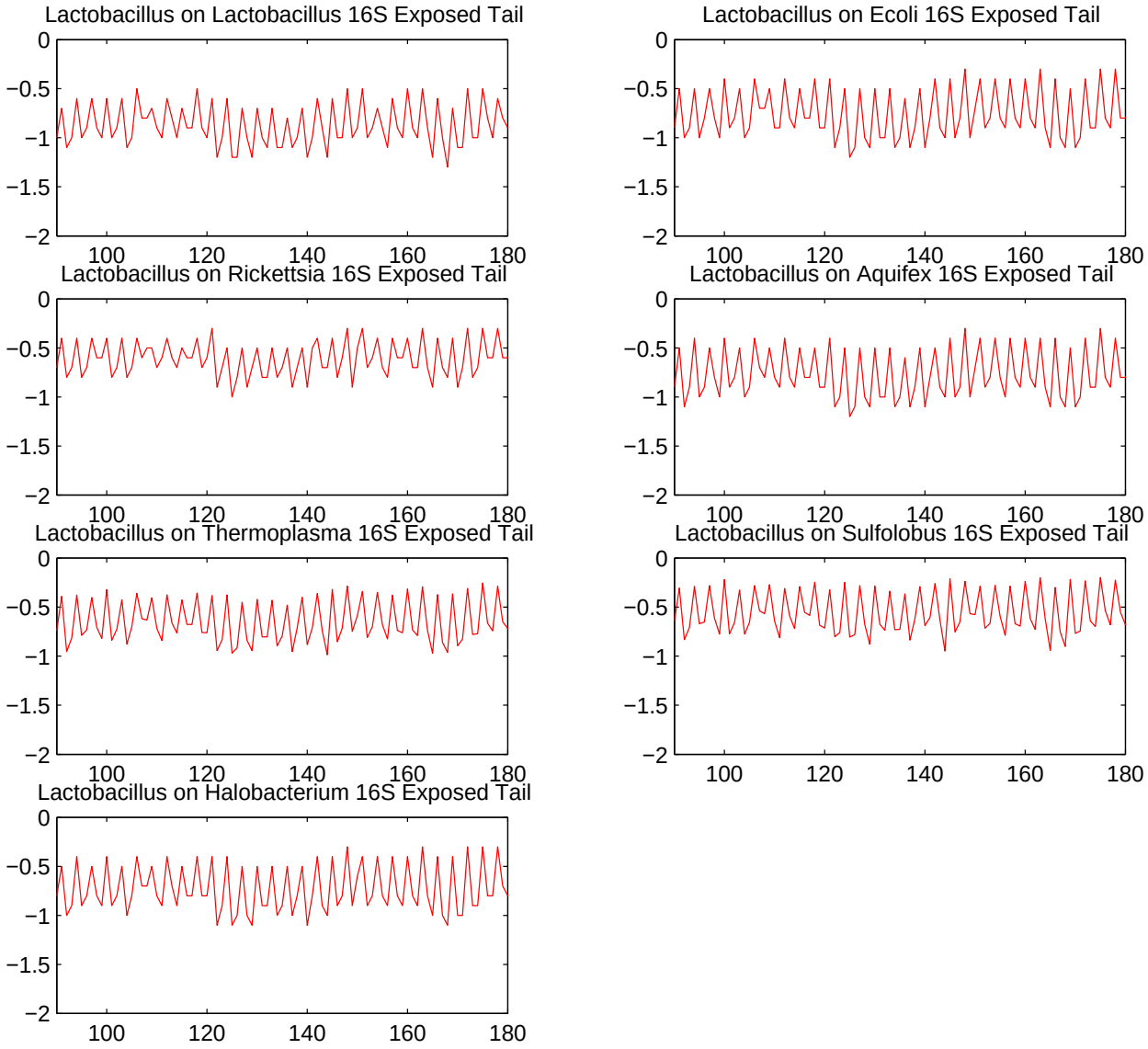

Figure A.12: Average Free Energy over 200 genes for Lactobacillus mRNA with all 16S Tails: Synchronization Signal

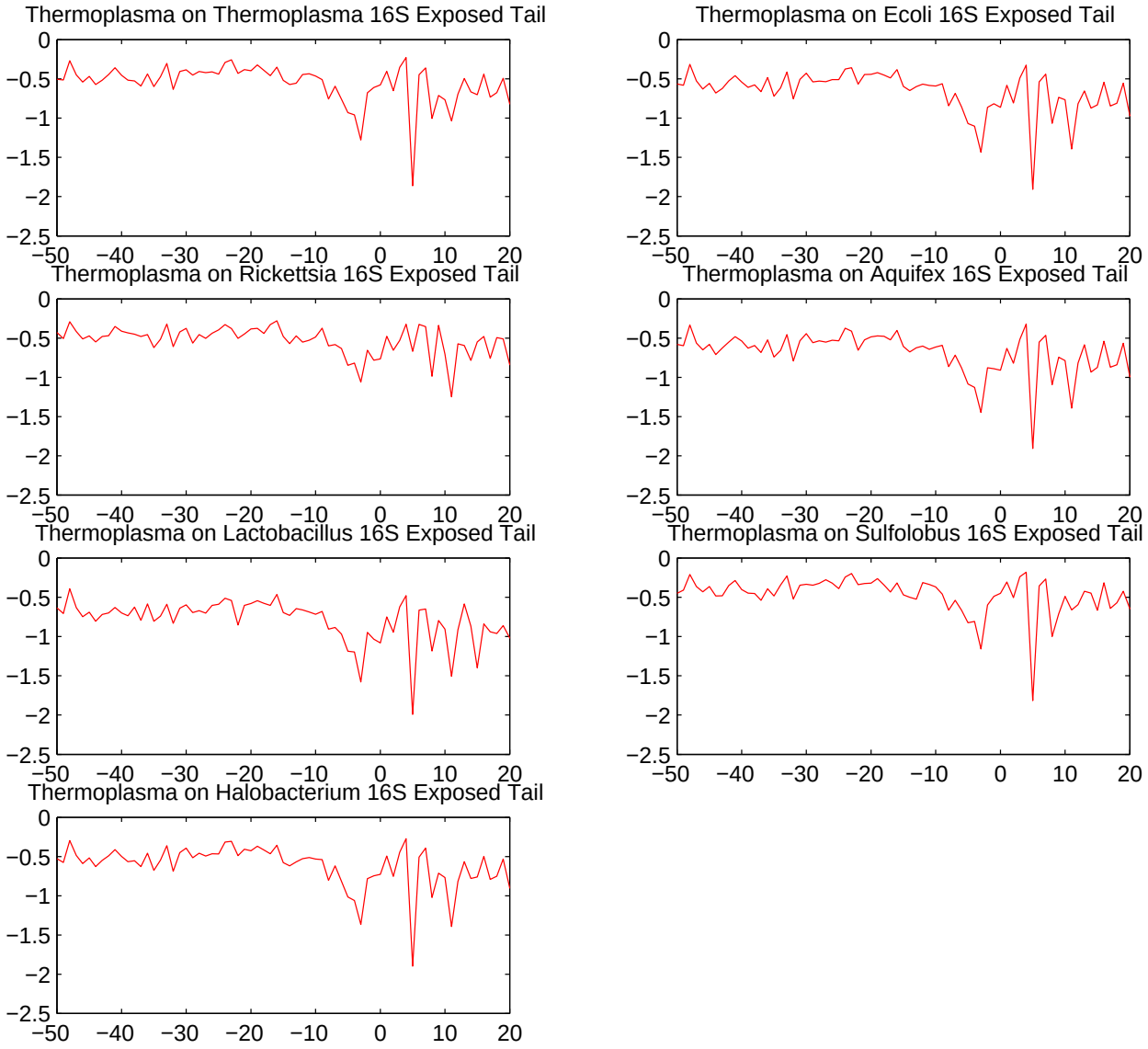

Figure A.13: Average Free Energy over 200 genes for Thermoplasma mRNA with all 16S Tails: Lock Signal

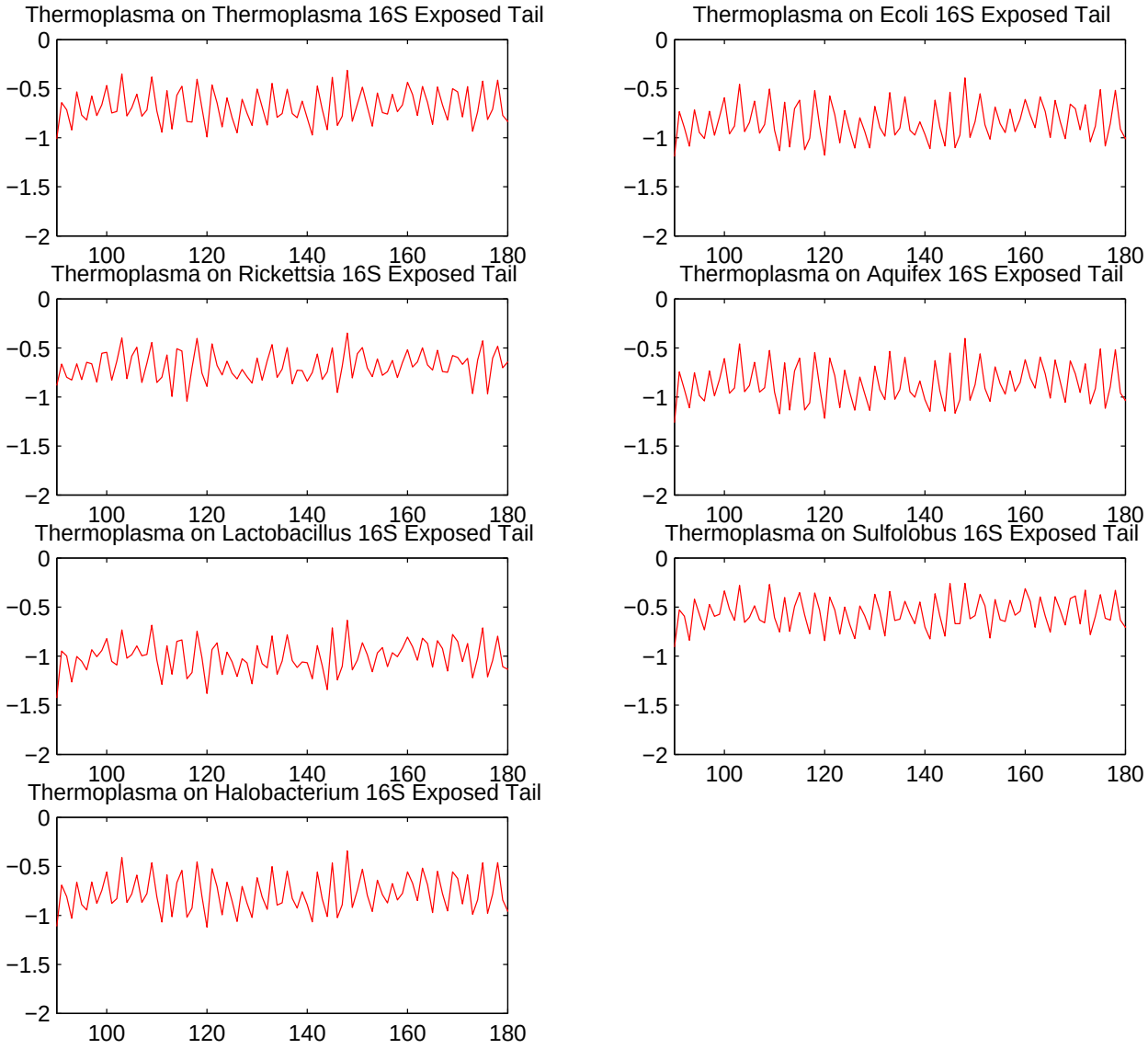

Figure A.14: Average Free Energy over 200 genes for Thermoplasma mRNA with all 16S Tails: Synchronization Signal

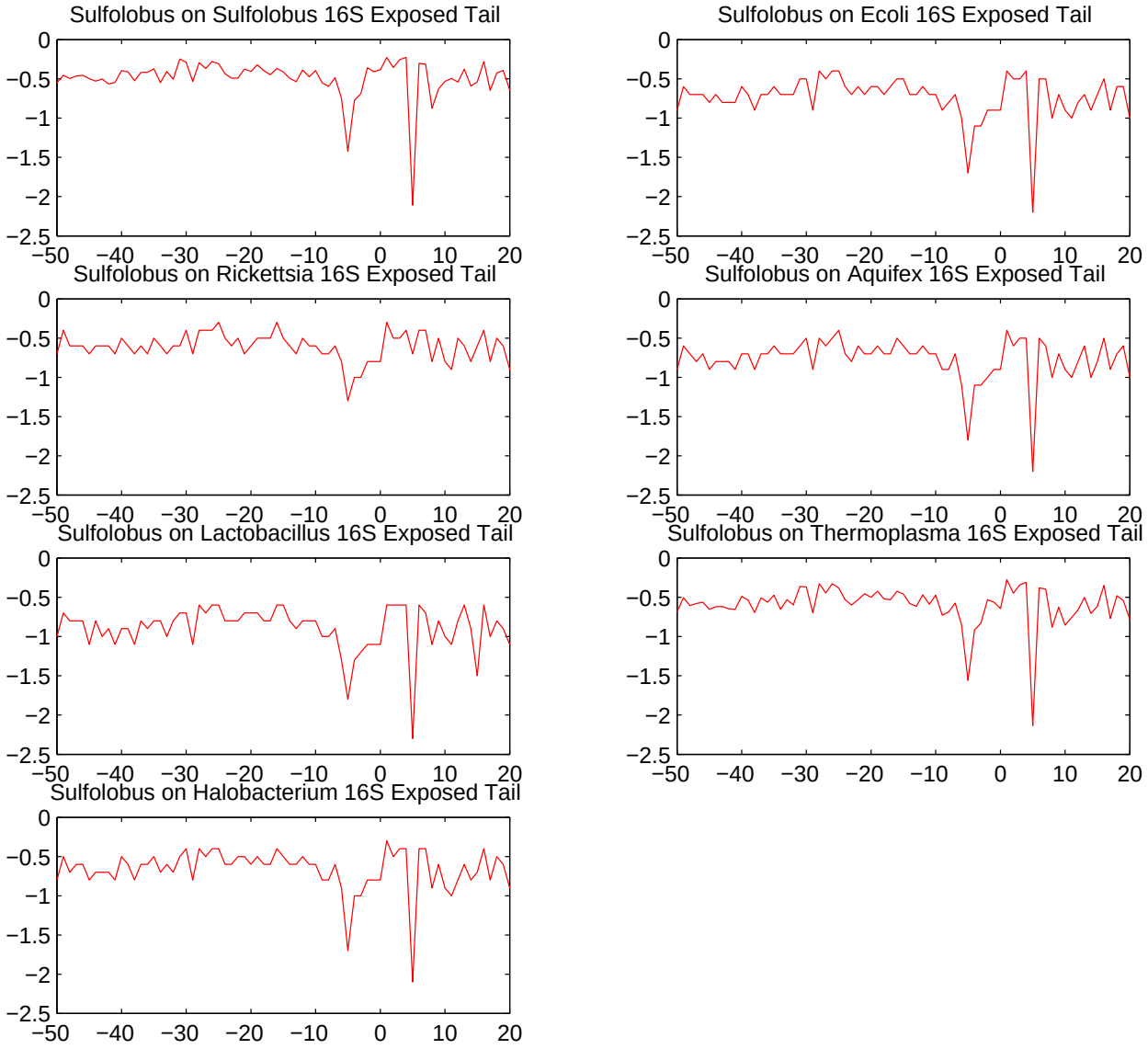

Figure A.15: Average Free Energy over 200 genes for Sulfolobus mRNA with all 16S Tails: Lock Signal

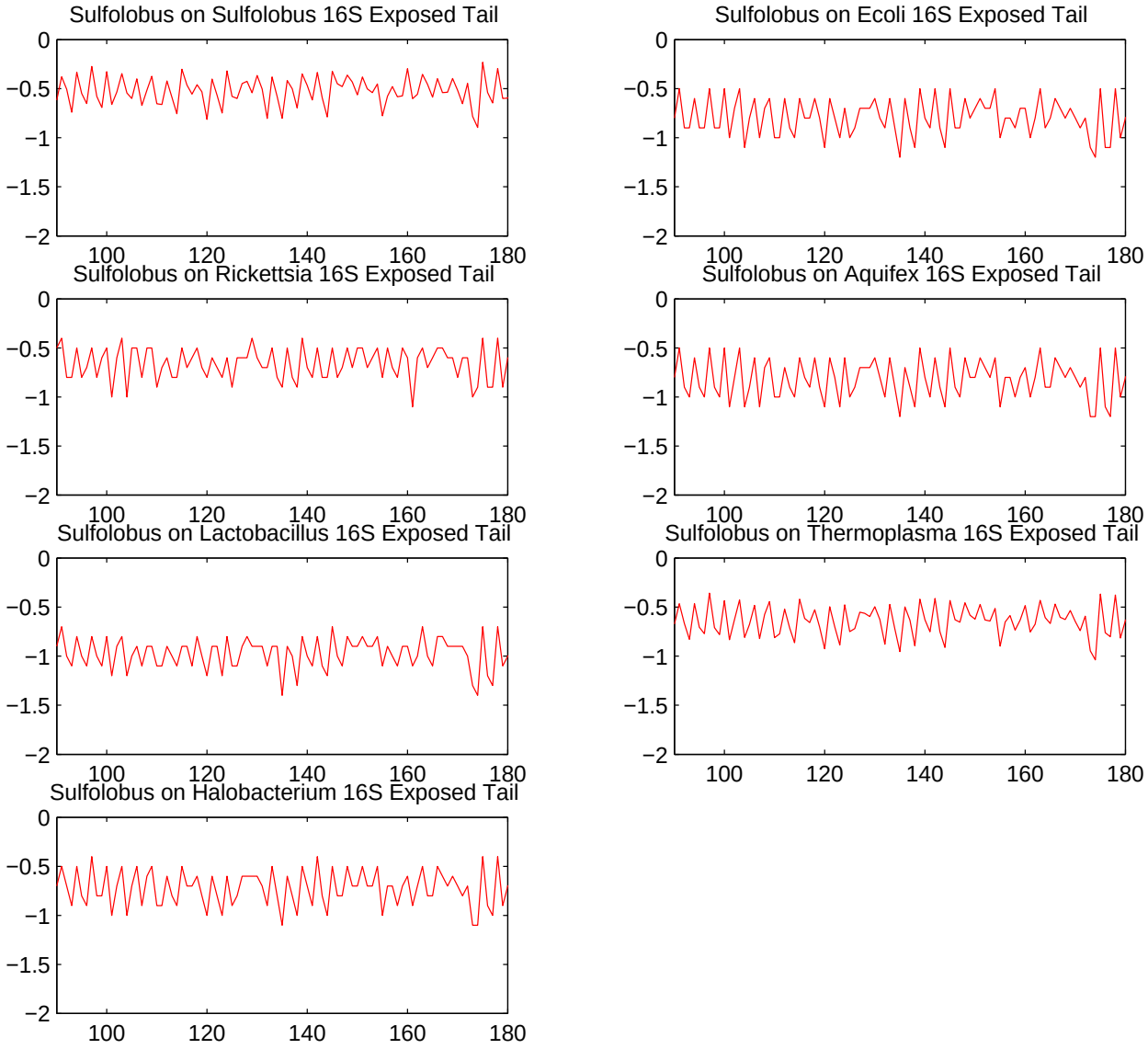

Figure A.16: Average Free Energy over 200 genes for Sulfolobus mRNA with all 16S Tails: Synchronization Signal

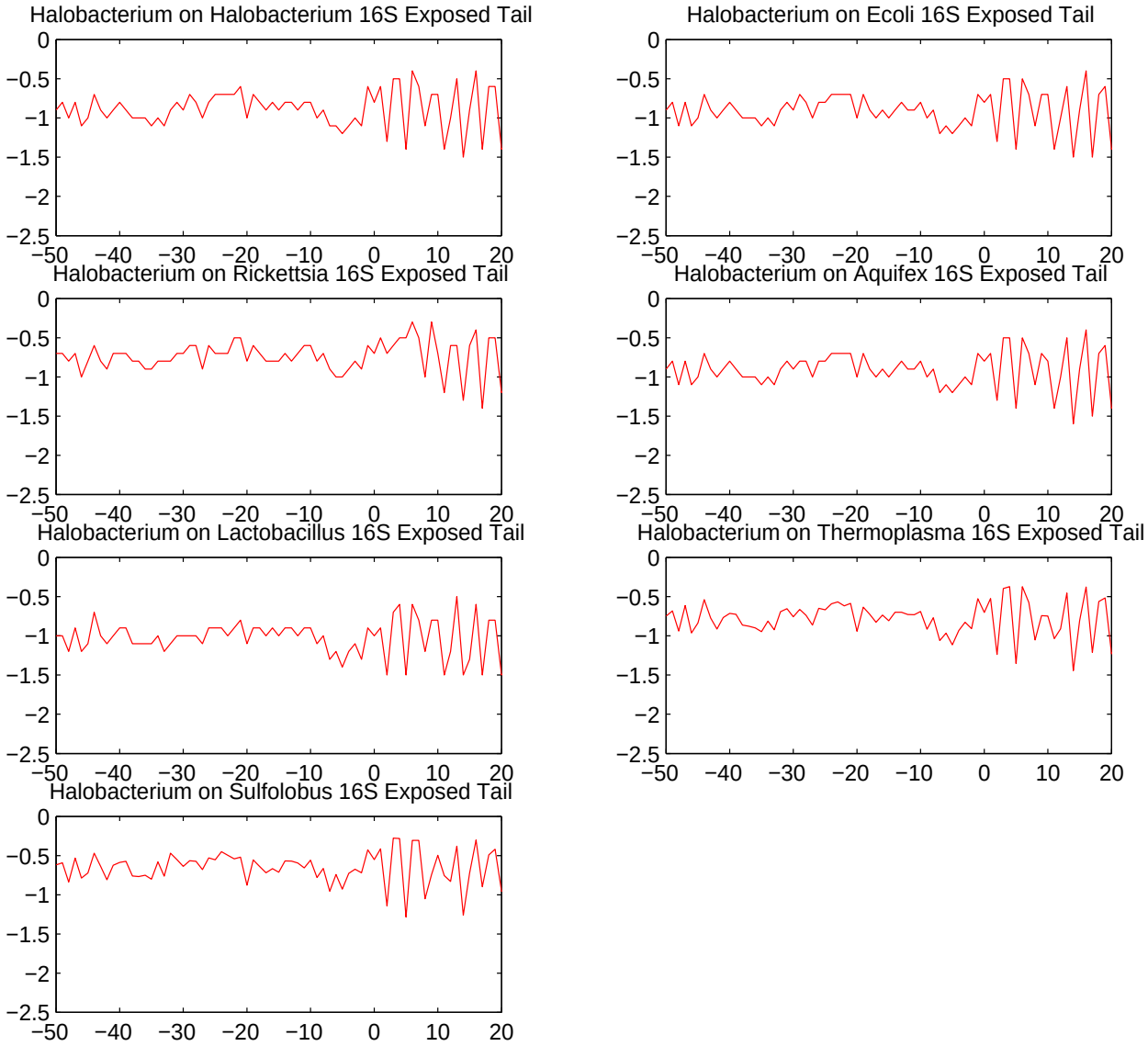

Figure A.17: Average Free Energy over 200 genes for Halobacterium mRNA with all 16S Tails: Lock Signal

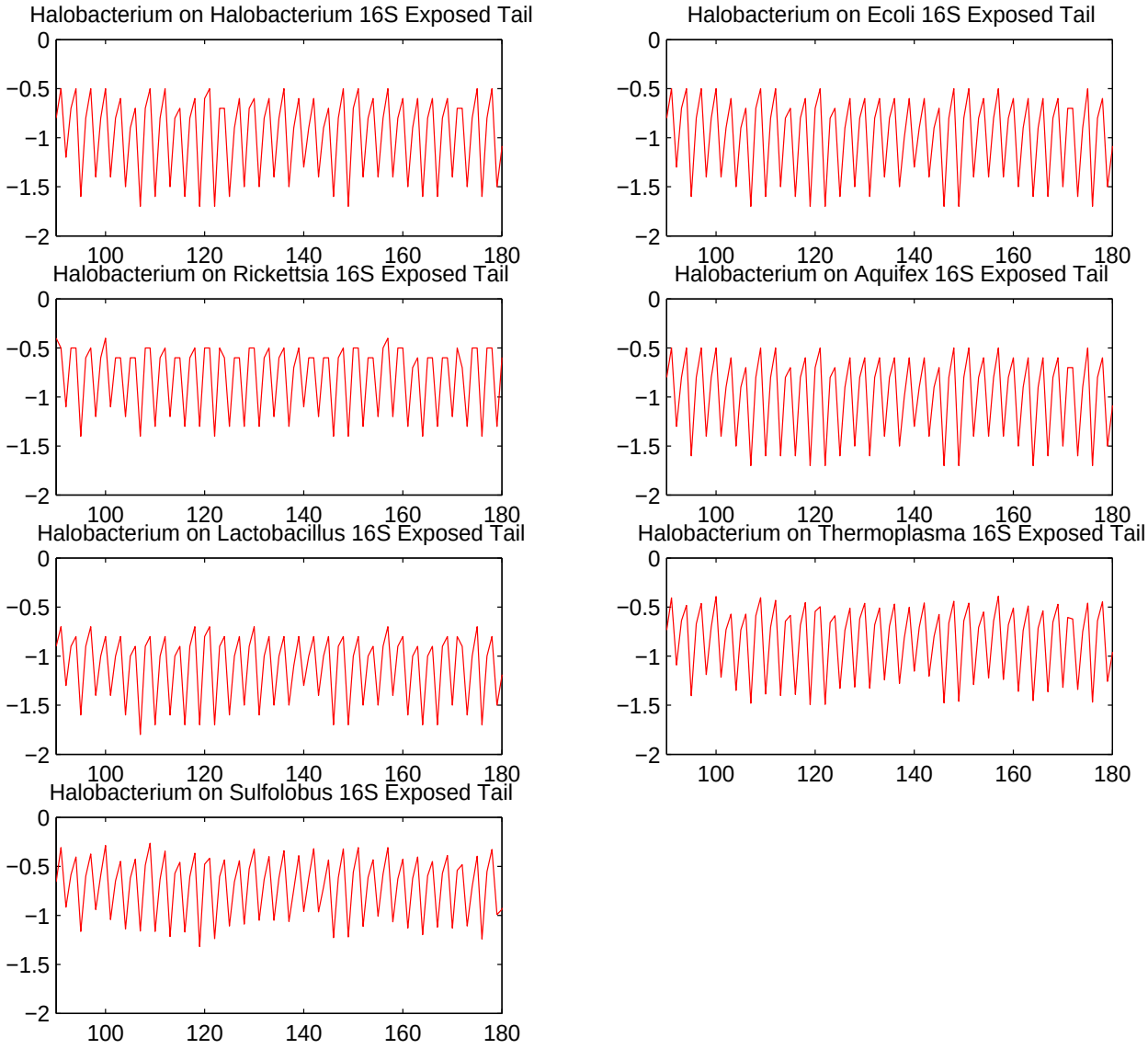

Figure A.18: Average Free Energy over 200 genes for Halobacterium mRNA with all 16S Tails: Synchronization Signal

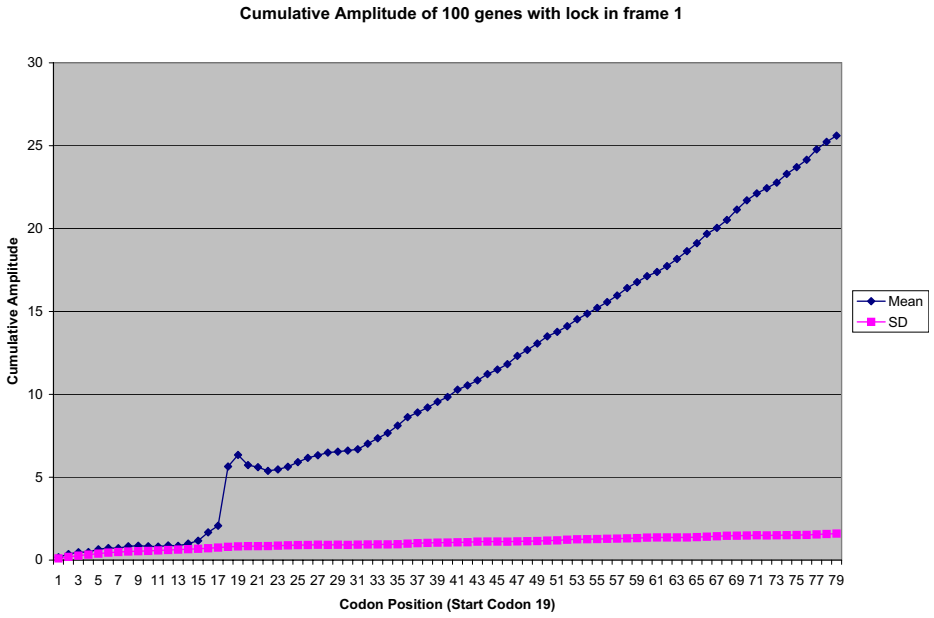

Figure A.19: Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 1

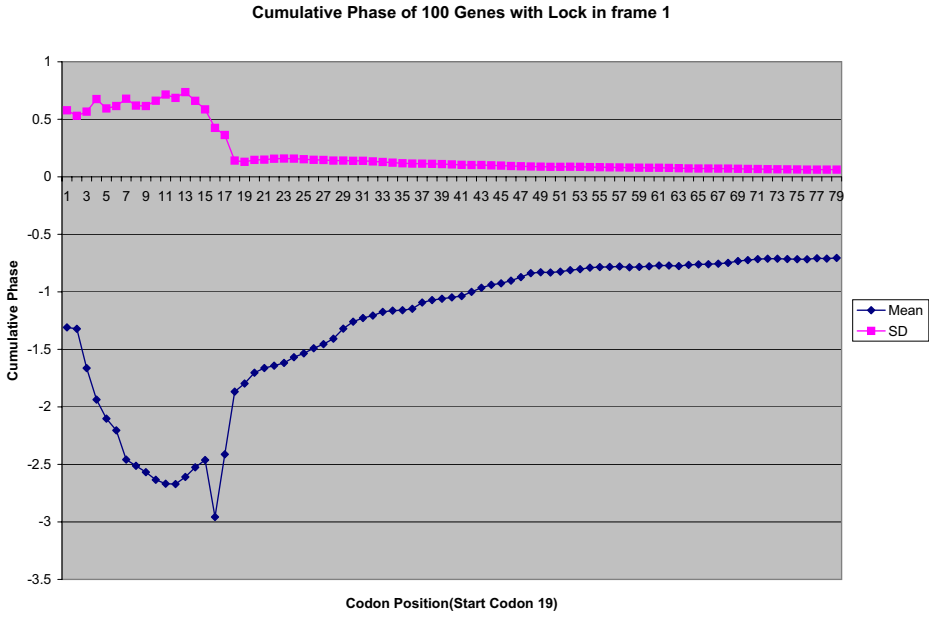

Figure A.20: Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 1

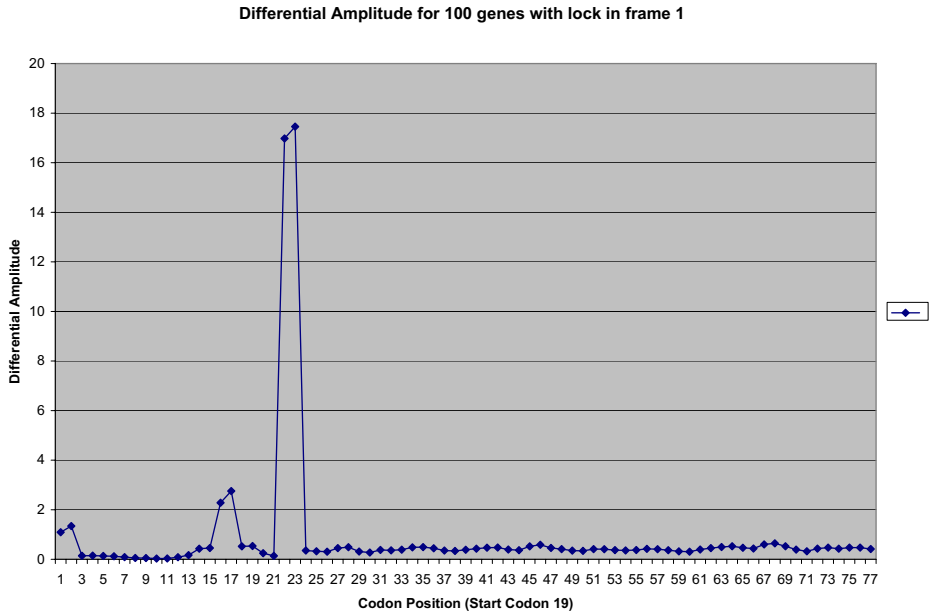

Figure A.21: Differential Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 1

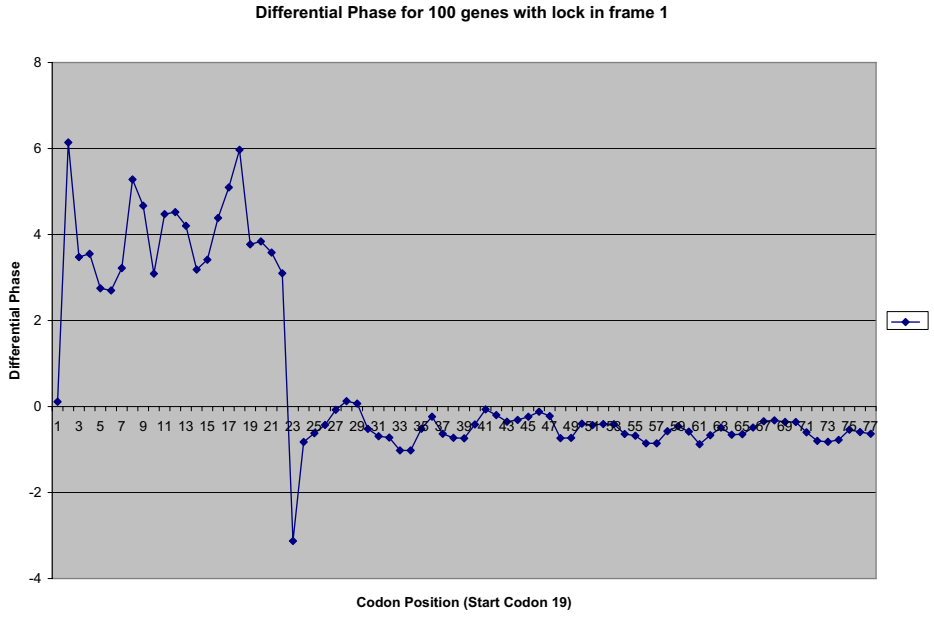

Figure A.22: Differential Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 1

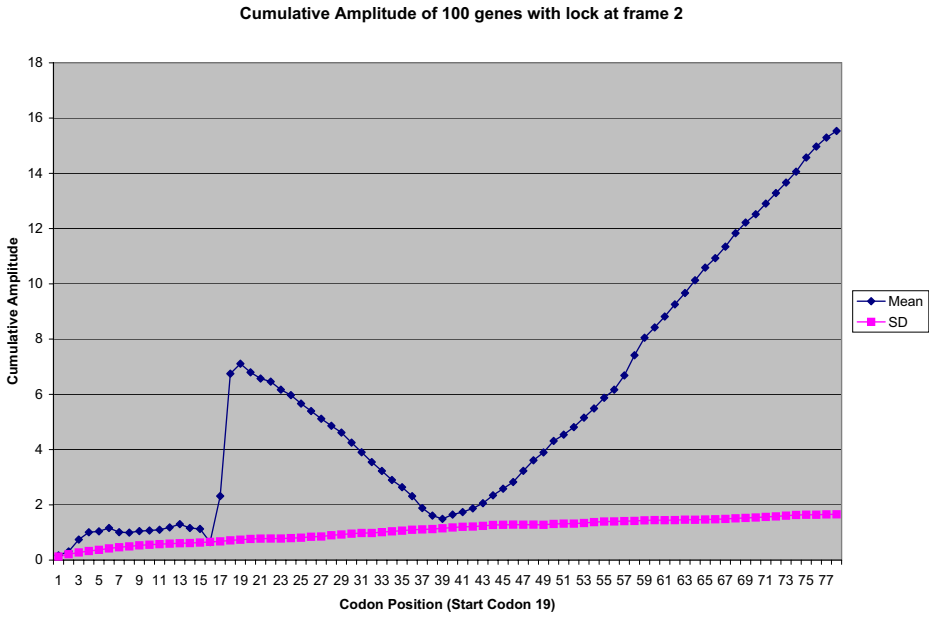

Figure A.23: Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 2

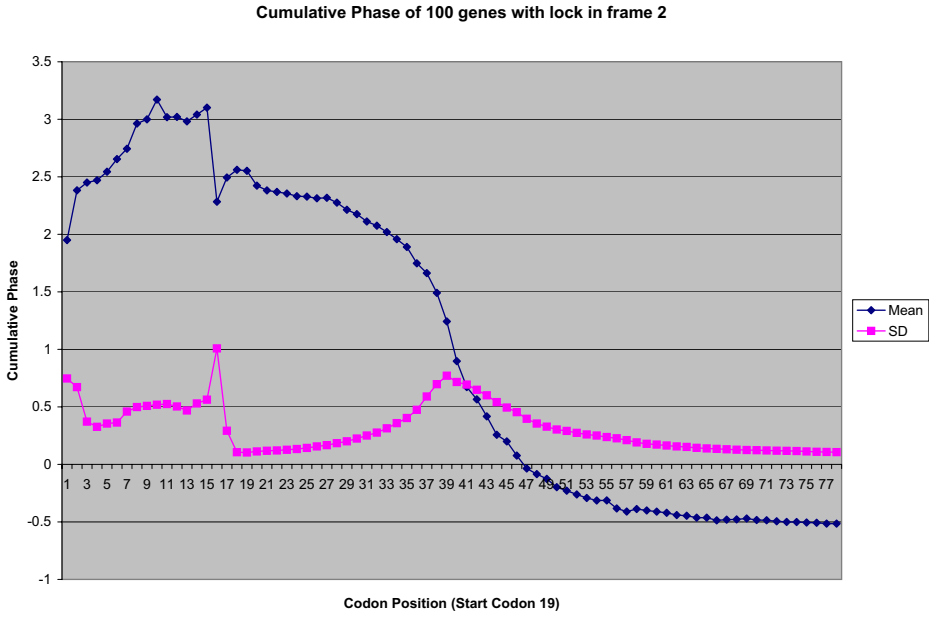

Figure A.24: Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 2

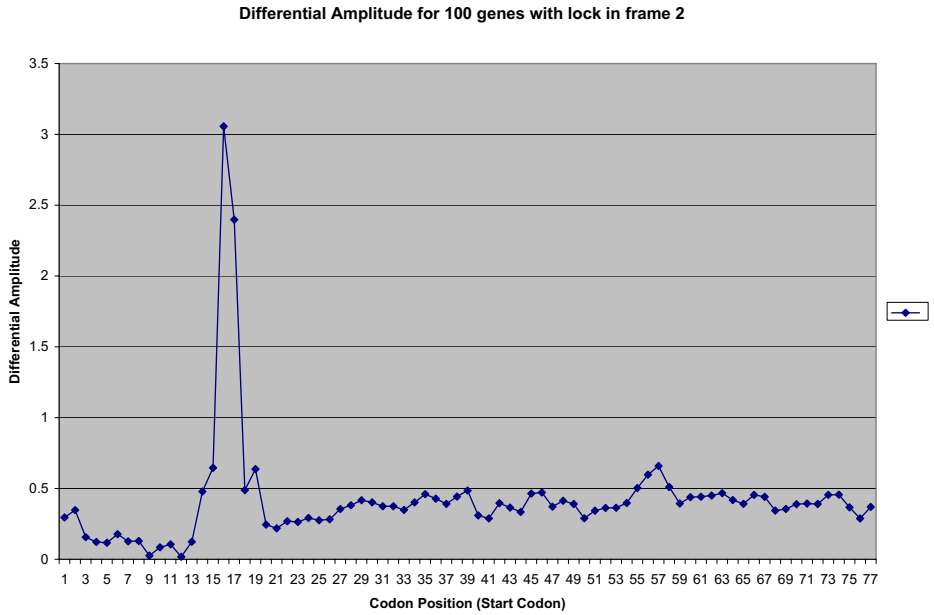

Figure A.25: Differential Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 2

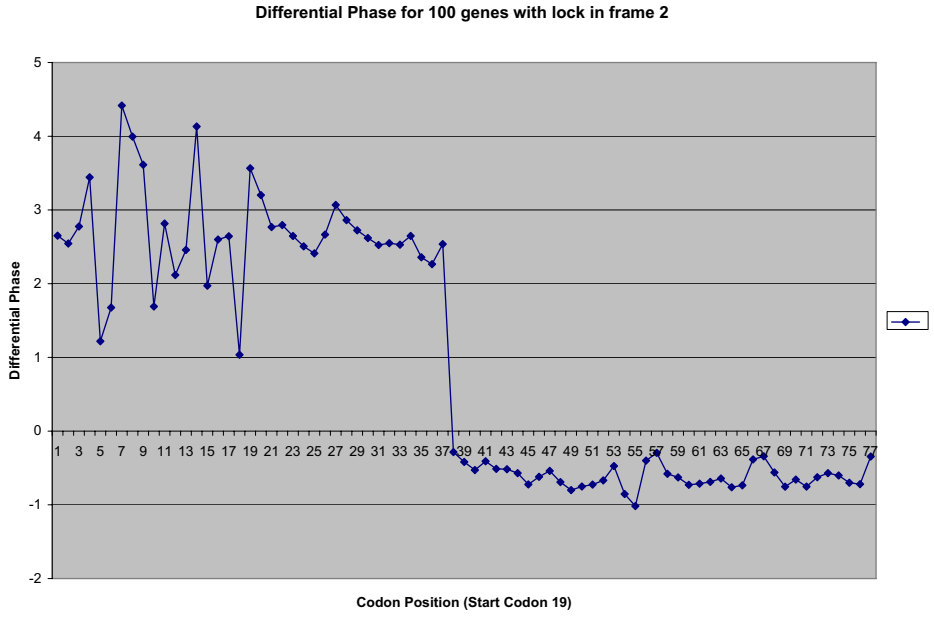

Figure A.26: Differential Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 2

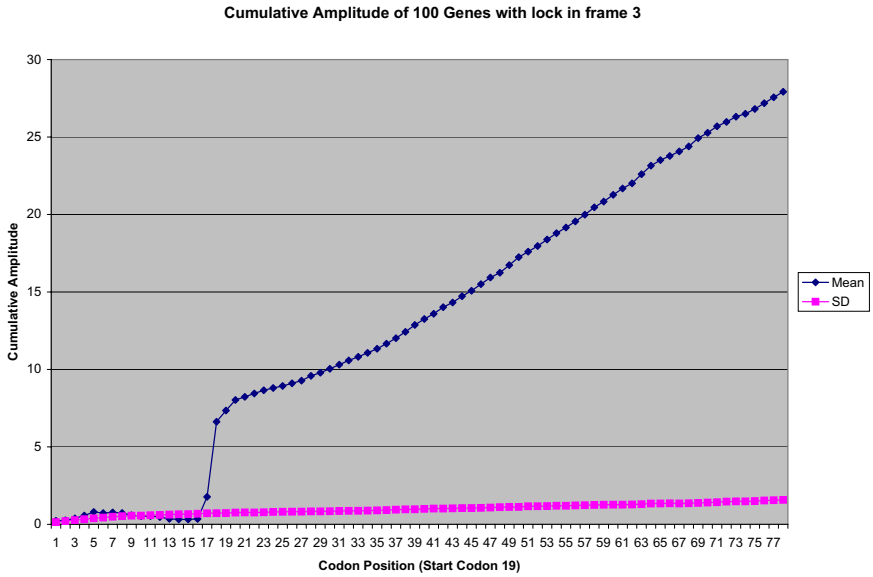

Figure A.27: Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 3

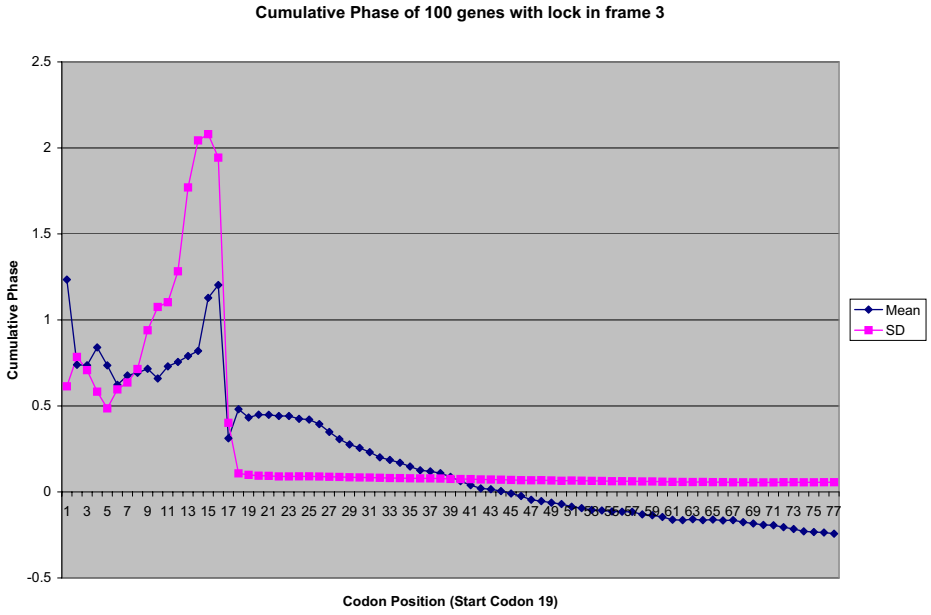

Figure A.28: Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 3

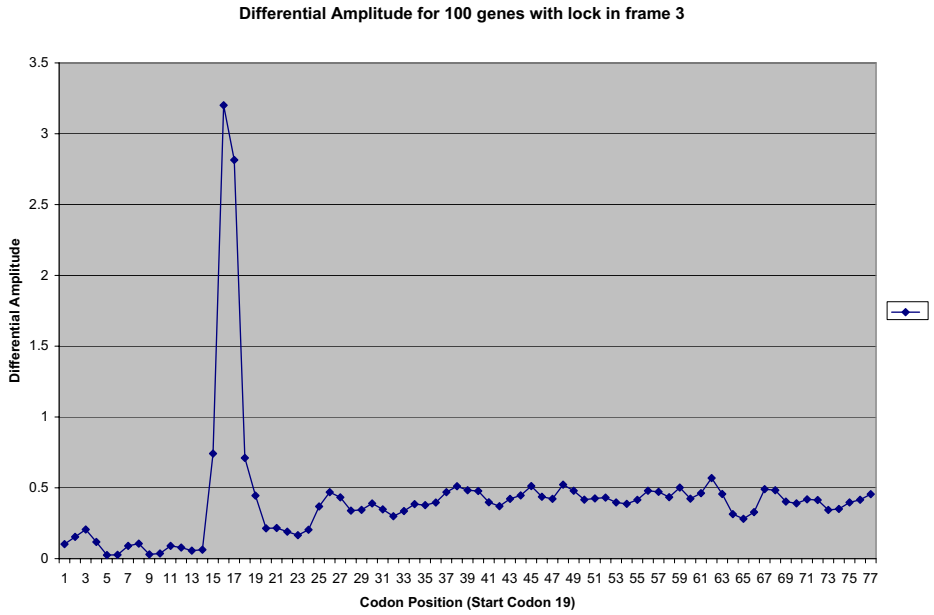

Figure A.29: Differential Magnitude accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 3

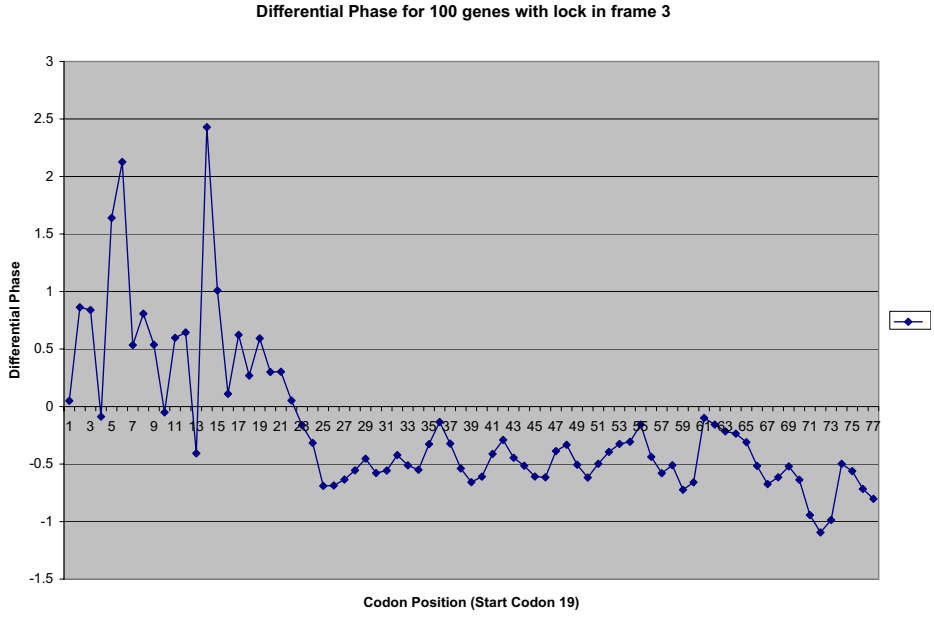

Figure A.30: Differential Phase accumulated over from -18 codons to +60 codon position for an average of 100 verified genes with SD lock at frame 3

Table A.1: Clockwise verified genes which fail the Binding-Index Based Locking Model for k = 0.19 and $F_t = 0.9$.

Start..Stop	Possible reason for failing
90094..91035	Adjacent gene
200971..201996	Adjacent gene
203348..204496	Adjacent gene
817793..818278	Adjacent gene
2620254..2620892	Adjacent gene
2623135..2624676	Adjacent gene
3341585..3342310	Adjacent gene
4074595..4075032	Adjacent gene
14168..15298	Locks at -7 or -8
784856..785908	Locks at +1
1108558..1110093	Alternate Start:1108567..1110093
1157092..1158525	Locks at -6
1299206..1300837	Locks at -6
1494910..1496535	Locks at +1
2459320..2460666	Alternate Start:2459326..2460666
3057773..3058666	Locks at -7
3084725..3085879	Locks at -7
3408908..3409204	Locks at -6
100765..102240	Wobble c to a Locks at -1
108279..110984	Wobble t to g Locks at -3
167484..169727	Wobble t to g Locks at -2;
3208748..3210493	Wobble g to t Locks at -1
3432844..3434133	Wobble c to a Locks at -6
4630329..4630655	Wobble c to a Locks at -2

Table A.2: Counter-clockwise verified genes which fail the Binding-Index Based Locking Model for k = 0.19 and $F_t = 0.9$.

Start..Stop	Possible reason for failing
177662..178462	Adjacent gene
1396798..1397550	Adjacent gene
1843023..1844984	Adjacent gene
1990897..1992729	Adjacent gene
2432844..2433656	Adjacent gene
2698638..2699018	Adjacent gene
2868278..2869327	Adjacent gene
2962383..2963177	Adjacent gene
3755644..3757488	Adjacent gene
3878849..3879949	Adjacent gene
4478550..4481405	Adjacent gene
53416..54702	Locks at +1
134788..135582	Locks at -6(WrontCAAT)
235535..236002	Locks at -6
551814..552323	Locks at -6
2444408..2445493	Locks at +1
2658337..2659551	Locks at -6
3159273..3160685	Locks at +1
3313680..3315167	Locks at -9
3397681..3398724	Locks at -6
3919688..3920068	Locks at -6
556098..556964	Wobble a to t Locks at -6
1056485..1057177	Wobble a to c Locks at -7
1140405..1143590	Wobble a to c Locks at -4
2245083..2246552	Wobble a to t Locks at -6
2733051..2734031	Wobble g to c Locks at +1
4061182..4061874	Wobble a to t Locks at +1
4156969..4158369	Wobble t to c Locks at -7
4481405..4481848	Wobble t to c Locks at -3

Table A.3: Verified non-locking genes.

Clockwise Genes		
197928..200360	491316..493247	499349..500653
612038..613162	961218..962891	1147982..1148935
1753722..1755134	2192320..2194353	2632252..2633622
2708440..2710062	2720747..2722102	2752917..2753399
3964032..3965291	4261893..4263308	
Counter-clockwise Genes		
54755..57109	60358..63264	440325..440567
643420..644226	661975..663186	854047..854967
988377..989579	1041253..1043433	1062078..1062998
1444402..1445307	1715375..1716031	1924803..1926863
1939675..1940607	2277808..2278503	2742592..2743359
2888122..2889921	3033204..3034302	3516181..3516702
3558255..3559085	3961980..3963245	4251622..4254045
4360923..4362620		