ABSTRACT

NATURAL SELECTION ON MRNA SECONDARY STRUCTURE AND ITS CORRELATION WITH PROTEIN FUNCTIONAL GROUPS

by
Suresh Solaimuthu

Natural selection may occur at multiple levels of the biological hierarchy, including at the molecular level. It may occur on any phenotypic trait that evidences variation and that is heritable. This research uses computational methods to investigate whether the stability of the secondary structures of mRNAs has been the subject of natural selection.

The DNA sequence that codes for a particular target protein is only partially determined by that protein, since the redundancy of the genetic code permits multiple possible synonymous codons for each peptide. An RNA transcript of a DNA protein template (gene) folds back on itself through complementary base pairing, resulting in an mRNA secondary structure. This mRNA secondary structure tends to have a configuration that minimizes free energy. Two synonymous mRNAs, coding for the identical protein with different sets of synonymous codons, will in general fold into different secondary structures with different minimum free energies (MFEs). The secondary structure of an mRNA is therefore a phenotypic trait that could be a target of natural selection.

Several related questions were investigated: 1) Is there natural selection on the stability of RNA secondary structure, across various types of organisms? 2) Does the MFE of microbial mRNAs correlate with the function of the target protein? 3) Is there evidence of natural selection on the nucleotide composition and/or secondary structure of the prefixes and suffixes of bacterial mRNAs? 4) Is there natural selection on the secondary structures and substructures of subviral RNAs?

These questions were investigated using large-scale simulations, based on the generation of sets of randomized synthetic mRNAs for particular genes. The secondary structure of each mRNA (naturally occuring and synthetic) was then computationally predicted. The experiments were performed on the complete sets of genes of a number of prokaryotes

and eukaryotes. Two types of randomized experiments were performed on each genetic data set, providing an independent confirmation of the results. In the first method of randomization, synonymous mRNAs were generated for each gene, creating sequences that code for the identical protein, with a frequency of codon use characteristic of the organism. In the second method of randomization, the nucleotides of the mRNA were permuted in manner that does not preserve the mRNA sequence's target protein, but exactly preserves the mRNA sequence's nucleotide and dinucleotide frequencies.

The MFE of each naturally occuring mRNA sequence is then compared with the MFEs of the corresponding randomized sequences. A pattern of deviation, across an entire organism, of the value of the MFE of the naturally occurring sequence from that of the corresponding randomized sequences is evidence of natural selection on the stability of the mRNA transcript.

This research establishes that:

1) In all prokaryotes studied, natural selection has favored of highly stable (lower MFE) mRNAs. In some prokaryotes, natural selection has also favored highly unstable mRNAs. No statistically significant evidence of such selection was found in eukaryotes.

2) The distributions of MFEs of mRNAs of 25 broad functional classes of proteins (COGs – Clusters of Orthologous Groups) of five microbes and yeast correlate to functional class.

3) mRNA prefixes have a distinctive MFE signature. The naturally occurring prefixes display more structure, on average, than randomized sequences with identical nucleotide and dinucleotide content, suggesting that natural selection favors secondary structure in the prefix of mRNA.

4) Viroids (with RNA genomes) have highly stable secondary structures and the structures are similar among the viroids belonging to the same family.

The results indicate that natural selection on the MFE of mRNA is widespread in the evolution of the genome.

NATURAL SELECTION ON MRNA SECONDARY STRUCTURE AND ITS CORRELATION WITH PROTEIN FUNCTIONAL GROUPS

by
Suresh Solaimuthu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science
Department of Computer Science

August 2010

APPROVAL PAGE

NATURAL SELECTION ON MRNA SECONDARY STRUCTURE AND ITS CORRELATION WITH PROTEIN FUNCTIONAL GROUPS

Suresh Solaimuthu

Dr. Barry Cohen, Dissertation Co-Advisor — Date
Associate Dean, College of Computing Sciences; Director, Bioinformatics Program, NJIT

Dr. Usman Roshan, Dissertation Co-Advisor — Date
Assistant Professor of Computer Science, NJIT

Dr. Narain Gehani, Committee Member — Date
Professor of Computer Science; Dean, College of Computing Sciences, NJIT

Dr. Vincent Oria, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Michael Halper, Committee Member — Date
Professor of Computer Science, Kean University

BIOGRAPHICAL SKETCH

Author:	Suresh Solaimuthu
Degree:	Doctor of Philosophy
Date:	August 2010

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, New Jersey, USA, 2010
- Master of Science in Computer Science,
  New Jersey Institute of Technology, Newark, New Jersey, USA, 2003
- Bachelor of Engineering in Computer Science and Engineering,
  University of Madras, Chennai, Tamil Nadu, India, 2001

Major:	Computer Science

Presentations and Publications:

Suresh Solaimuthu, and Barry Cohen, "Folding energy of mRNA transcripts in bacteria correlates with their gene functional group", *Systems Biology Symposium, CSHL*, 2006.

To My Beloved Parents,
M. Solaimuthu and M. Mani

ACKNOWLEDGMENT

The journey to the completion of this dissertation would not have happened without the support of numerous people at various levels. They encouraged, morally supported, and tolerated me through thick and thin.

First and foremost, I would like to thank my advisor, Dr. Barry Cohen for accepting me to work under him. He is an advisor that every student would dream of having. I have learnt a lot from him and he has helped me a lot, both professionally and personally. The success of my research work was due to his constant guidance, encouragement and feedback. I learnt about patience, to be self critical, and to have an eye for detail from him.

Special thanks to my dissertation committee members, Dr. Usman Roshan, Dr. Narain Gehani, Dr. Vincent Oria, and Dr. Michael Halper for their guidance and encouragement. I would also like to thank the Computer Science PhD program director Dr. David Nassimi for his support. I would like to take this oppurtunity to thank Dr. Yehoshua Perl for encouraging me to pursue PhD.

I would like to thank Dr. Ronald Kane, Dean of Graduate Studies, who has been a source of encouragement all throughout my graduate life at NJIT. He has always gone out of his way to help graduate students. This dissertation would not have been completed without his help. I would also like to thank Ms. Clarisa Gonzalez, who reviewed this dissertation. Her keen eye made this dissertation a neat and fair one. I would also like to thank Ms. Lillian Quiles, who has always been helpful at Department of Graduate Studies.

This dissertation would not have completed without the help of System Administrators. I would like to thank Dr. David Perel, Dr. Gedaliah Wolosh, Mr. Dean Knape, and Mr.

Kevin Walsh for providing service for Kong and AFS systems. Their continuous help and fast service helped me in doing my research smoothly.

I would like to thank Dr. George Olsen, who was always helpful throughout my PhD program. His constant support helped me in gaining wonderful experience. I would also like to thank Ms. Angel Bell, who is an angel to the students in our department.

I would also like to thank the staff of Office of International Students, Mr. Jeffery Grundy, Mr. Scott Kline, Ms. Tiffany Small, and Ms. Annie Brown, for helping in making the international student life at NJIT a happy one.

I would like to thank my family and friends for their constant support, understanding, and patience.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter Page

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Overview of The Dissertation

The focus of the dissertation is to find if there is a natural selection on the stability of mRNA secondary structures. Several related questions were raised and analysis performed.

The next section of this chapter gives a brief introduction to molecular biology, evolution and genetics. Then the dataset and the randomization methods are discussed. The major differences between the two methods and reasons to use them are also discussed.

Chapter 2 deals with the question: Is there natural selection on the stability of RNA secondary structure, across various types of organisms? The chapter gives a brief background of various other teams that have done similar research, their methods and their results. Then it discusses the experiments and the results obtained.

Chapter 3 deals with the question: Does the MFE of microbial mRNAs correlate with the function of the target protein? The basics of COG and the various functional classes in it are discussed. The correlation between the COG functional classes and the MFE were analyzed.

Chapter 4 deals with the question: Is there evidence of natural selection on the nucleotide composition and/or secondary structure of the prefixes and suffixes of bacterial mRNAs? The subsequences in the form of prefixes, suffixes, and windows were analyzed. The results are then discussed.

Chapter 5 deals with the question: Is there natural selection on the secondary structures and substructures of subviral RNAs? The entire viroid family sequences were folded and the stability was analyzed. Also the optimal substructures were analyzed to find if there is a selection for a particular kind.

1.2 Overview of Molecular Biology, Evolution, Genetics

The fundamental unit of life which forms the basic building blocks is called the cell. All living beings can be classified into Prokaryotes and Eukaryotes. The key difference between these two is the prokaryotic cells do not have nuclei whereas the eukaryotic cells have. Each cell contains thread-like structures which is the hereditary material called chromosome. A chromosome consists of the macromolecule called DNA. Some of these DNA can also be found in mitochondria. All living beings are made up of the basic blocks called macromolecules. The nucleic acids (DNA and RNA) and proteins together are called macromolecules.

1.2.1 Macromolecules

DNA Deoxyribonucleic acid is the important hereditary material that carries information. DNA can either be single stranded or double stranded. The DNA is made of four chemical bases called nucleotides (bases). The nucleotides are grouped into two types, namely, purines and pyramidines. Adenosine (A) and Guanine (G) belong to purine; Cytosine (C) and Thymine (T) belong to pyramidines. The single stranded DNA is made of chain of nucleotides and is called a polynucleotide.

The bases form chemical bonds with each other called base pairs. The pairing is specific: A pairs with T and G pairs with C. These nucleotide pairs are arranged as strands forming a spiral, which is called a double helix. Since the strands are complementary to each other, the replication of DNA is simple. Using one strand the other can be obtained easily.

1.2.2 Genes and Genomes

The smallest inheritable unit is called the gene. The complete set of genes that an organism inherits from its parents is called the genotype. The physical characteristic of the organism because of the genotype is called the phenotype.

The complete genetic material present in an organism is called the genome. The chromosomal and mitochondrial DNA together forms the genome of an organism.

1.2.3 Evolution

Darwin's theory of evolution states that all life descended from a common ancestor. Some random mutations stay over generations because of their usefulness for survival of the organisms. The process of evolution takes place using various mechanisms: descent, mutation, genetic drift, natural selection.

Descent: Evolution occurs when there is a change in gene frequency within a population over time. These genetic differences are heritable and can be passed on to the next generation.

Mutation: Mutations are random changes in DNA.

Genetic Drift: is the change of gene frequency in a population.

Natural Selection: Natural selection acts to preserve and accumulate minor advantageous genetic mutations. Selection occurs whenever individuals with a particular genotype enjoy an advantage in survival or reproduction over other genotypes.

1.2.4 Central Dogma of Molecular Biology

”The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.” - Francis Crick.

It primarily states that the general flow of genetic information is from DNA to RNA to protein. Crick proposed that once the information becomes a protein it can not take any other form i.e. the transfer of information from protein to nucleic acid is not possible.

1.2.5 Ribonucleic Acid

RNAs are an important class of molecules in the biological world, serving two distinct classes of functions. mRNAs serves as informational molecules – templates for proteins. Functional non-coding RNAs (ncRNAs) catalyze biochemical reactions. These two groups of functions suggest that RNA might have played an important role in the prebiotic evolution of replicating systems. The sequence of RNA is its primary structure. RNA molecules tend to fold back on themselves to form secondary structure. The secondary structure is generally made up of Watson-Crick GC and AU pairs, separated by nonhelical segments.

Four major classes of RNA exist, and can be found in most organisms:

1. mRNA: messenger RNA is a sequence which codes for formation of one or more proteins. They vary considerably in size, which reflects the variation in the size of the protein encoded by mRNA as well as the gene serving as the template for transcription of mRNA.

2. tRNA: transfer RNA are small sequences which bring amino acids to the ribosome, where they translate mRNA into amino acid sequences. Because more than one

tRNA molecule interacts simultaneously with the ribosome, the molecule's smaller size facilitates these interactions.

3. rRNA: ribosomal RNA sequences form ribosomes. This usually constitutes 80 percent of all RNA in the cell. The various forms of rRNA found in prokaryotes and eukaryotes differ distinctly in size.

4. viral RNA: A virus that has RNA as its genetic material is called viral RNAs.

A gene or cistron is defined as the region of DNA that is transcribed into functional RNA. The transcript functions either as such (e.g. tRNA, rRNA) or as a messenger (mRNA), which codes for a single polypeptide chain in the translation process. A polypeptide is a polymer made of amino acids. A polynucleotide such as RNA is an asymmetrical polymer that is assembled from nucleoside triphosphates by a stepwise mechanism linking the 3' position of one nucleotide by a phosphate bride to the 5' position of the adjacent nucleotide. In the finished polynucleotide chain, the first nucleotide residue has a 5' position that is not linked to another nucleotide, whereas the last nucleotide has an unlinked 3' position. Thus, polynucleotide synthesis proceeds from the 5' to the 3' terminus and the polymer is said to have a 5'-to-3' polarity. Usually, linear RNA sequences are written with the 5' terminus on the left and the 3' terminus on the right. The genetic information stored in DNA is not usable directly for making proteins but must be copied first into mRNA by an enzymatic transcription of segments of DNA containing the genes. Messenger RNA serves as template for protein synthesis, that is, the linear nucleotide sequence of the mRNA dictates the amino acid sequence of the polypeptide encoded originally by the gene. The mechanism for translating RNA into protein is complex, and the cell devotes considerable resources to the translational machinery. The components include 20 different amino acids, transfer RNAs, aminoacyl-tRNA synthetases, ribosomes and a number of protein factors that cycle on and off the ribosomes and facilitate various steps in initiation of translation, elongation of the nascent polypeptide chain, and termination of synthesis with release of the completed polypeptide from the ribosome.

1.2.6 mRNA Structure

The sequence information of a gene is copied (transcribed) into the nucleotide sequence of RNA using one strand of DNA (called the coding strand) as the template. The primary transcript is a single strand of RNA , which is a faithful copy of the other strand of DNA (the non-coding strand), with substitution of U residues in place of T residues found in DNA. Sometimes, the primary transcript is altered, before it functions as mRNA. In these cases the original unmodified transcript is the precursor or pre-mRNA. The decoding process involves base pairing between three bases (i.e. codon) in the mRNA and the three base anticodon of a transfer RNA. In a separate reaction, each tRNA is first linked to a particular amino acid, and thus the pairing of mRNA with tRNA determines the sequence of amino acids in the resulting protein.

Prokaryotic mRNA In organisms that do not have a nucleus (prokaryotes), pre-mRNA usually undergoes little or no modification, with the result that pre-mRNA and mRNA are very similar if not identical. Since mRNA is collinear with DNA, DNA and proteins are usually collinear in these organisms. Gene expression in prokaryotes usually involves the cotranscription of several adjacent genes and translation of mRNA sequences into polypeptides may begin at the 5' end of mRNA while transcription is still in progress at the 3' end.

Eukaryotic mRNA In cells with a nucleus (eukaryotes), the genetic information is stored mainly in the nucleus and to a minor degree in some organelles (mitochondria and chloroplasts). The description that follows pertains only to nuclear genes. Eukaryotic genes are more complicated than prokaryotic genes because the coding region in the former is often discontinuous: the coding sequences or exons are interrupted by intervening sequences (introns). Thus, genes and proteins are usually not collinear in eukaryotes. In the nucleus, a complicated set of splicing reactions removes all the introns and fuses the exons into a continuous coding sequence. Other processing steps involve adding a cap to the 5' end of the mRNA adding a polyadenulated tail to the 3' end. After completion of these nuclear maturation steps the mRNA is transported to the cytoplasm, where it is translated. As with prokaryotic mRNA, the coding region is flanked by 5' and 3' nontranslated sequences.

Transcription into mRNA The coding information contained in a gene (a DNA coding region) is transcribed into mRNA from one DNA strand (the coding strand). The mRNA is a copy of the DNA with U residues in place of T residues. In prokaryotes, the mRNA transcript undergoes little or no modification before being translated into a protein. However, in eukaryotes, the mRNA may be extensively processed before translation.

In prokaryotes, adjacent genes are often coregulated – that is, simultaneously transcribed into mRNA.

1.2.7 Genetic Code

The relationship between coding regions of DNA or RNA and the proteins that are formed from these regions is called the genetic code. With minor variations, the genetic code is the same for all organisms. It consists of the 64 possible DNA (or RNA) triplets (codons) and the corresponding amino acid peptides.

The process of translation of mRNA into protein is usually initiated at an AUG or GUG codon, which are called start codons. Of the 64 possible DNA or RNA triplets, 61 correspond to one of the 20 amino acids. The remaining three triplets, called nonsense codons, serve as stop signals for the process of translation. Each codon specifies a single amino acid, but most amino acids are coded for by from two to six codons. The codons that code for the same amino acid are called synonymous codons.

Table 1.1 Universal Genetic Code For Ribonucleic Acid

Amino Acid	Codons
Alanine	GCU GCC GCA GCG
Arginine	CGU CGC CGA CGG AGA AGG
Asparagine	GAU GAC
Aspartic acid	GAU GAC
Cysteine	UGU UGC
Glutamic acid	GAA GAG
Glutamine	CAA CAG
Glycine	GGU GGC GGA GGG
Histidine	CAU CAC
Isoleucine	AUU AUC AUA
Leucine	UUA UUG CUU CUC CUA CUG
Lysine	AAA AAG
Methionine	AUG
Phenylalanine	UUU UUC
Proline	CCU CCC CCA CCG
Serine	UCU UCC UCA UCG
Threonine	ACU ACC ACA ACG
Tryptophan	UGG
Tyrosine	UAU UAC
Valine	GUU GUC GUA GUG

The Table 1.1 gives the universal genetic code. The start and stop codons are also the same in most organisms.

1.2.8 Genetic Mechanisms in Prokaryotes and Eukaryotes

Prokaryotes and eukaryotes were studied separately because there are significant differences in their translation mechanisms:

1. Transcription in eukaryotes occurs within the nucleus; prokaryotes have no nucleus. In prokaryotes translation and transcription overlap in time. In eukaryotes, the RNA transcript migrates out of the nucleus before translation.

2. The eukaryotic mechanism regulating initiation of transcription is more complex, involving various DNA sequences and protein factors.

3. Eukaryotic mRNA undergoes multiple processing steps before translation; prokaryotic mRNA is generally directly translated into protein.

4. Eukaryote translation occurs on ribosomes that are larger and more complex than those of prokaryotes.

5. Eukaryotic mRNAs have longer half lives than prokaryotic mRNAs (hours rather than minutes).

Differences of transcription in prokaryotes and eukaryotes The major differences are:

1. Transcription in eukaryotes occurs within the nucleus under the direction of three separate forms of RNA polymerase. Unlike prokaryotes, in eukaryotes the RNA transcript is not free to associate with ribosomes prior to the completion of transcription. For the mRNA to be translated, it must move out of the nucleus into the cytoplasm.

2. Initiation and regulation for transcription involve a more extensive interaction between upstream DNA sequences and protein factors involved in stimulating and initiating

transcription. In addition to promoters, other control units called enhancers may be located in the 5' regulatory region upstream from the initiation point, but they have also been found within the gene or even in the 3' downstream region beyond the coding sequence.

3. Maturation of eukaryotic mRNA form the primary RNA transcript involves many complex stages called processing. An initial processing step involves the addition of a 5'-cap and a 3'-tail to most transcripts destined to become mRNAs. Other extensive modifications occur to the internal nucleotide sequence of eukaryotic RNA transcripts that eventually serve as mRNAs.

Differences in translation in prokaryotes and eukaryotes The major differences are:

1. In eukaryotes the translation occurs on ribosomes that are larger and whose rRNA and protein components are more complex than those of prokaryotes.
2. Eukaryotic mRNAs are much longer-lived than the prokaryotic mRNAs. Most exist for hours rather than minutes prior to their degradation by nucleases in the cell, remaining available much longer to orchestrate protein synthesis.
3. The initiation of translation is different in eukaryotes compared to prokaryotes. The 5'-cap is present in eukaryotes, which is essential for efficient translation, as RNAs lacing the cap are translated poorly, whereas prokaryotes don't have it. Most eukaryotic mRNAs contain a short recognition sequence that surrounds the initiating AUG codon, 5'-ACCAUFF.
4. Amino acid formulmethionine is not required to initiate eukaryotic translation. However, as in prokaryotes, the AUG triplet, which encodes methonine, is essential to the

formation for the translational complex and a unique transfer RNA is used during initiation.

5. In eukaryotes a large proportion of the ribosomes are found in association with the membranes that make up the endoplasmic reticulum. Such membranes are absent from the cytoplasm of prokaryotic cells.

1.2.9 RNA Secondary Structure

mRNA is a single stranded molecule that forms intra-strand base pairs to produce secondary structures.

The stability of a secondary structure is the sum of the free energies that are released by the formation of its base pairs. The lower the free energy of a structure, the more likely is its formation and the greater is its stability. Laboratory measurements have determined the free energy changes associated with a variety of possible configurations that constitute the great majority of actually occurring secondary structures, including stacked base pairs, internal loops, bulges and hairpin loops. These empirically determined free energy values are used in secondary structure prediction.

There are five types of secondary structural elements: hairpin loops, internal loops, multibranched loops, bulges and stacks or stem loops.

Hairpin loops: The unpaired region formed when an RNA folds back upon itself to form a helix. It occurs at the end of a helix when the sugar phosphate backbone reveals a hairpin like structure. Comparisons of small subunit ribosomal RNA structures reveal an uneven distribution of hairpin loop sizes: four base loops are the most common. Larger hairpin loops can pair into complex structures involving non-Watson-Crick interactions.

Hairpin loops are important for mRNA stability, RNA tertiary interactions, and protein binding sites.

Internal loops: Two or more opposing unpaired bases between two helical segments; internal loops can be symmetric (the same number of unpaired bases on each side of the loop) or asymmetric (a different number of unpaired bases on each side of the loop). Two base internal loops are often called mismatches. Common small internal loops have increased stability due to base tacking and non-Watson-Crick hydrogen bonding. Internal loops are important sites of RNA-protein interaction in 5S rRNA and proposed RNA-RNA tertiary and quaternary interactions in group I introns.

Multiloop: Region in which three or more helices join to form a closed loop. The crystal structure of tRNA has a four-helix multibranched loop stabilized by helix-helix stacking as well as significant non-Watson-Crick secondary and tertiary interactions. These interaction probably stabilize other multiloops.

Bulge loop: Regions in which there are unpaired bases on only one side of a helix. They can bend RNA backbones. Bulges are important recognition sites for many regulatory and structural proteins. For this study the right and left bulges are taken separately.

Stack: Also called stem loops, they contribute most to the stability of the RNA secondary structure through hydrogen bonds and base stacking. The base stacking is the interaction between the pi orbitals of the bases' aromatic rings. The Watson-Crick pairs G-C and A-U, as well as some of the mismatches, such as G-U, stabilize the stacks. Base stacking is an important stabilizing effect since a single base stacking on the 3' side of a helix can add as much stability to the structure as a base pair.

The other types of the RNA secondary structural elements include pseudoknots, which are too unstable to be considered here. Pseudoknots are structures that result when any single-stranded loop forms a helix with another single-stranded region.

RNA Secondary Structure Prediction The secondary structure prediction method employed in this work assumes that the structure formed is the one with the most negative Gibbs free energy ΔG°. Due to simplifying assumptions made, RNA secondary structure prediction algorithms achieve only a first order approximation of actual RNA structures. Among the excluded factors are the kinetics of folding during transcription, the existence of pseudoknots and other nonplanar secondary structures, the role of chaperone proteins and the role of modified bases (e.g. inosine or methylated bases).

Factors Influencing RNA Secondary Structure Prediction The major factors influencing the secondary structure prediction are the nucleotide content, dinucleotide content and the codon composition of amino acids in genetic code. [1] found that there is a pronounced periodic pattern of nucleotide involvement in mRNA secondary structure. This pattern was created by the structure of genetic code and the dinucleotide relative abundances are important for the maintenance of mRNA secondary structure. Although synonymous codon usage contributes to this pattern, it is intrinsic to the structure of the genetic code and manifests itself even in the absence of synonymous codon usage bias at the 4-fold degenerate sites. While all codon sites are important for the maintenance of mRNA secondary structure, degeneracy of the code allows regulation of stability and periodicity of mRNA secondary structure. The third degenerate codon sites contribute most strongly to mRNA stability. This shows that the redundancies in the genetic code allows transcripts to satisfy requirements for both protein structure and RNA structure. The selection may be operating on synonymous codons to maintain a more stable and ordered mRNA secondary structure, which is likely to be important for transcript stability and translation.

It was shown that under GC pressure, in most of the quartet codon groups there is a preferential choice of the C-ending codon, except in leucine and valine codon groups where the choice is the G-ending codon is preferred. Among the duet groups, the choice of codons specifying phenylalanine and glutamate shows the strongest dependence on GC content. A high correlation is found between the GC content at the third codon position of exons and the neighboring introns and flanking sequences. These relationships indicate the existence of compositional constraints operating on both coding and noncoding sequences.

The dinucleotide content in a coding sequence plays a major role in the secondary structure prediction based on the thermodynamic principle. So the dinucleotide energy is very important. A modest electron-transfer effect is found in the Watson-Crick AT , GC pairs and Hoogsteen AT pair, confirming the weak covalence in the hydrogen bonds. The electrostatic attraction and polarization effects account for most of the binding energies, particularly in GC pair. Both theoretical and experimental data show that he GC pair has a binding energy of -25.4 kcal twice that of the AT with -12.4 kcal and H-AT -12.8 kcal. The GC has three H-bonds compared to two in the other pairs. A strong binding between the guanine and cytosine bases benefits from the opposite orientations of the dipole moments in these two bases assisted by the pi-electron delocalization from the amine groups to the carbonyl groups, model calculations demonstrate that pi-resonance has very limited influence on the covalence of the hydrogen bonds.

1.3 Methods and Materials

1.3.1 Monte Carlo Methods

Monte Carlo experiments were performed on the genes of eight prokaryotes (eubacteria and archaea) and three eukaryotes (yeast, worm, fruit fly). The experiments were done on the entire set of genes of the eight prokaryotes and yeast.

Monte Carlo experiments were performed by two independent methods. The first method, which is called codon preference randomization, preserves the coding function of the sequence – that is, it codes for the same protein. The second method, which is called the shufflet method, exactly preserves the nucleotide and dinucleotide composition of each sequence.

Codon Preference Randomization In this method, a randomized sequence is generated that codes for the same sequence as the naturally occurring sequence. That is, if S is a natural sequence and T is a corresponding randomized sequence, each codon T_i in T is a synonymous codon of S_i in S.

For example, consider the sequence ATG-CTA-GGC (hyphens inserted only to indicate codon boundaries) which codes for the amino acids argenine, leucine and glycine (see the synonymous codon Table 1.2). A synonymous sequence is ATG-TTG-GAA. It codes for the same three amino acids, even though two of the codons are different.

The biological motivation for this constraint is that protein sequence is known to be more strongly conserved, in the course of evolution, than nucleotide sequence in coding regions of the genome [2].

Further, in selecting among synonymous codons while constructing the randomized sequence, codon preference randomization uses probabilities established by the pattern of

Table 1.2 Amino Acids And Synonymous Codons For ATGCTAGGC

Amino acid	Codons
Arginine	ATG
Leucine	TTG TTA CTA CTG CTT CTC
Glycine	GGG GGA GGC GGT GAG GAA

codon frequency in the organism as a whole. For example, in the above sequence, C_1 is CTA, a leucine-coding codon. If the synonymous codon TTG accounts for 10 percent of the leucine-coding codons in the organism, then the probability that T_1 will be TTB is 10 percent.

The biological motivation for this is that codon usage is fairly consistent within an organism, but differs among organisms.

Shufflet Randomization Shufflet randomization uses a method devised by Kandel [3] to construct randomized sequences is such a way that both the counts of nucleotides and the counts of adjacent pairs of nucleotides (dinucleotides) are exactly preserved. Further, this algorithm uniformly samples the set of all possible such shufflings. The algorithm, which works in linear time, constructs an Euler path on a directed graph. The implementation used was written by [4].

Consider the sequences ATGACG, which has the amino acids methionine and threonine. A shufflet randomized version of this sequence is ACGATG.

The biological motivation of this randomization method is that it maintains exactly both the GC content and dinucleotide content of the randomized sequences, which parameters significantly influence the MFE of an RNA sequence.

Concurring Results of the Two Methods of Randomization Each of the above methods provides an independent test of the hypothesis that evolution selects for high- or low-MFE mRNAs, and each identifies a set of mRNAs whose MFEs appears to have been shaped by natural selection. Having two independent methods of identifying such genes provides yet another test of the hypothesis.

1.3.2 RNA Folding Software

A computational method was used to find the predicted lowest energy secondary structure. The ViennaRNA package [5] was used to fold each sequence (both naturally occurring and synthetic) to get the minimum free energy and RNA secondary structure. The algorithm has been improved by a number of contributors [6]. The program minimizes a free energy function, which sums contributions from different secondary structure motifs. For any given RNA sequence length, the lower the energy estimate the more stable the predicted fold. The minimization is done be a dynamic programming method that always finds the secondary structure with the minimum free energy under a simplified secondary structure model.

1.3.3 Z Score Analysis and Quantile Analysis

The stability of each sequence was analyzed by Z score and quantile. Each analysis was performed for experiments performed using both methods of randomization.

The Z score standardizes or normalizes the results for each gene, expressing the stability of the naturally occurring sequence in terms of how many standard deviations it is above or below the mean MFE of the corresponding synthetic sequences.

$$Z = \frac{x - \mu}{\sigma}$$

where x is the MFE of the naturally occurring sequence; μ is the mean MFE of the corresponding synthetic sequences and σ is the standard deviation MFEs of the synthetic sequences

If a sequence has a Z score of -2 or less it is considered highly stable. If a sequence has a Z score of +2 or more it is considered highly unstable.

The quantile analysis ranks the MFE of a naturally occurring sequence relative to the population of 50 ordered MFEs of synthetic sequences. Let the MFE of a natural sequence be $E(S)$ and the artificial sequences be $E(S_j)$. The quantile of the natural sequence is the number of E(S_j) such that E(S) $\leq$ E(S_j). Hence the quantile is 0 if the natural sequence is more stable than all of the synthetic sequences. In the absence of selective pressure, quantile scores are expected to be evenly distributed among values 0- 50.

CHAPTER 2

EVIDENCE OF SELECTION FOR SECONDARY STRUCTURE STABILITY IN PROKARYOTES AND EUKARYOTES

2.1 Introduction

2.1.1 Background

One of the interesting questions in evolution is why nature chose a particular one of exponentially many possible RNA encodings for a protein. Did the nature preferentially ”choose” some types of encodings or is the encoding we see just a snapshot of a random mutational walk among possible equivalent encodings? This question is prompted by the redundancy of the genetic code: the $4^3 = 64$ codons map to only 20 amino acids.

Like any other phenotypic trait of an organism, the shape and stability of a mRNA molecule might enhance or diminish the survival and reproductive prospects of the organism. In the case of RNA secondary structure, for example, it might impede the chemical machinery which translates it into a protein, causing a selective pressure on mRNA sequences to form secondary structures, or to avoid them.

There is some biological evidence to support such a notion. Although the study of mRNA has focused largely on its protein coding function, as the putative aboriginal biotic material [7], RNA would have been subject to selection for structure well before its protein-coding role evolved [8]. The transcription and translation of mRNA in the course of protein production expose it to varied processes and environments. Its life cycle may require depending on the organism formation and breaking of secondary and tertiary structure, the excision of introns, passage through an organelle membrane, and persistence in the cytoplasm. Each phase in its life cycle offers possibilities for structure-based selection. RNA structure is also known to play a regulatory role [9]. [10] have recently demonstrated the existence of bistable RNAs which are easily accessible in evolution and which could serve as conformational switches.

2.1.2 Related Work

[11] compared the free energies of 51 randomly selected sequences from prokaryotes, plants, invertebrates and higher animals with randomized versions of those sequences. The 51 sequences were less than 1,200 bases in length. Each was compared with 10 random sequences. Six randomizing methods were used. SHUFFLE randomizes the nucleotide bases keeping their composition constant. The CDS-random technique randomizes within the coding region. The codon-shuffled technique randomizes by shuffling the codons within the coding sequence. The codon-random technique randomizes the codon choice but keeps the nucleotide base composition and the final protein product same. The codon-flat technique which does not constrain the nucleotide base composition. The UTR-random technique randomizes by shuffling the UTRs but leaves the CDS unchanged. None of the randomization methods preserved the dinucleotide content of the sequence, which exerts significant influence on its minimum free energy. Also these randomizing techniques do not maintain the end- protein product, which is more conserved than the sequence. The shuffling technique does not maintain the GC, content which plays a major role in stability. The authors concluded that natural mRNA sequences are more stable than the randomized sequences. The authors also concluded that the mRNA secondary structures favors codons that contribute to higher stability.

[12] used 48 sequences from the above. For each sequence, they generated 10 random sequences using several different methods. The methods used were zero order markov, mononucleotide shuffled, first order Markov, dinucleotide shufflet. The zero order Markov technique generates random sequences based on the mononucleotide frequencies of each base. The mononucleotide shuffle technique randomizes by drawing at random, weighted by the nucleotide proportions based on the length of sequence and the nucleotide base counts. The first order Markov technique randomizes based on the conditional probability P(a —b) of nucleotide a given b from all the possible combinations of the four nucleotides. The dinucleotide shuffled technique randomizes, by selecting a random trinucleotide at each iteration and then by shuffling all the non- overlapping trinucleotides that being and end with the same nucleotide base. The authors concluded did not detect significant difference between the stability of natural and randomized sequences when the dinucleotide content was held constant.

[13] performed an analysis based on windows of 50 bases rather than entire coding sequences. The methods used were codon shuffle (preserves the protein encoded and codon usage), dicodon shuffle (sequences generated by preserving the dinucleotide frequencies, encoded protein, codon usage) and the dishuffle (sequences preserving the dinucleotide frequencies). The codon shuffle method does not preserve the dinucleotide content which plays a very important role in minimum free energy prediction. The dicodon shuffle is not random enough as it has lots of constraints. The dishuffle method does not preserve the end protein product and the codon usage. The authors concluded that there is a strong bias towards the local RNA structure in the majority of the eubacterial species studied.

The first large-scale experiments performed on sequences from a variety of organisms to compare the stability of mRNA sequences to random synonymous sequences were [14, 15]. The randomization method is the same as the codon preference method used as one of the techniques in this paper. The experiment was conducted on over 27,000 sequences from 34 microbial species. It showed that in all organisms highly stable sequences occur more frequently than would be expected by chance.

2.1.3 Dataset

Publicly available whole genome data were used for the experiments. The data were obtained from NCBI and TIGR.

Bacteria Bacteria are microscopic unicellular organisms that reproduce by binary fission. They are widely distributed in soil, air, water, and within more complex organisms. Bacteria are prokaryotes; they do not have a nucleus.

The bacterial genome is usually a single chromosome – a double-stranded, circular molecule of DNA. Some bacteria have more than one such chromosome, and many bacteria also contain plasmids – small double-stranded rings of RNA having a small number of genes.

Yeast Yeasts (order Saccharomycetales) are unicellular fungi, commonly found on plant, in soil and salt water, and on the skin and in the intestinal tracts of warm- blooded animals.

Fruitfly The fruitfly Drosophila melanogaster is one of the most widely researched organisms, particularly in genetics and developmental biology. It is a small animal with a life cycle of just two weeks. Mutant flies, with defects in any of several thousand genes are available. It has four pairs of chromosomes, the X/Y sex chromosomes and the autosomes 2, 3, and

4. The size of the genome is about 165 million bases and contains about 18,000 coding sequences.

2.1.4 Bacterial Datasets

The bacterial datasets used were: i) 500 coding sequences selected at random from 160 bacterial genomes and ii) the complete set of coding sequences of eight bacteria (Synechocystis, Chlamydia trachomatis, Haemophilus influenzae, Mycoplasma genitalium, Pseudomonas aeruginosa, Halobacterium sp. NRC-1, Methanosarcina acetivorans, Escherichia coli).

For each gene in the dataset, 50 synthetic sequences were generated by the codon preference method and 50 synthetic sequences were generated by the shufflet method. For each naturally occurring sequence and each synthetic sequence, a predicted MFE and secondary structure was computed.

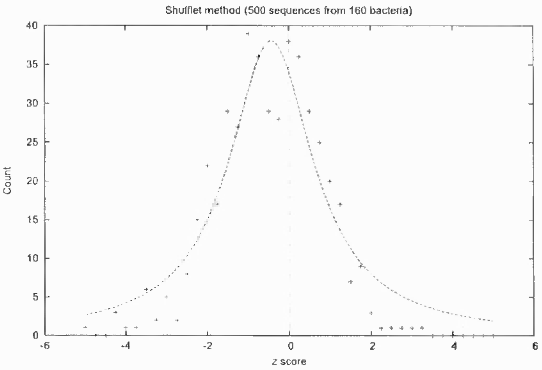
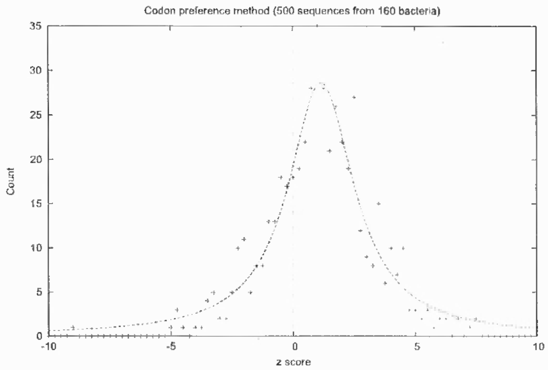
2.2 Results

2.2.1 Results for 500 Bacterial Sequences

A bias towards highly stable sequences is evident from the results for the analyzed sequences. The 500 randomly selected bacterial sequences are skewed toward low MFE compared to shufflet sequences. They show an overrepresentation of both low and high MFE compared to codon preference randomized sequences.

Table 2.1(a) shows the stability (MFE) of a random sample of 500 natural bacterial coding sequences compared to corresponding sets of randomized sequences. For each natural sequence, the set of randomized sequences is drawn uniformly at random from the universe of sequences with identical dinucleotide content (shufflet method). The natural

Table 2.1 Z Scores of MFE of 500 Bacterial Sequences Based on Shufflet and Codon Preference Methods

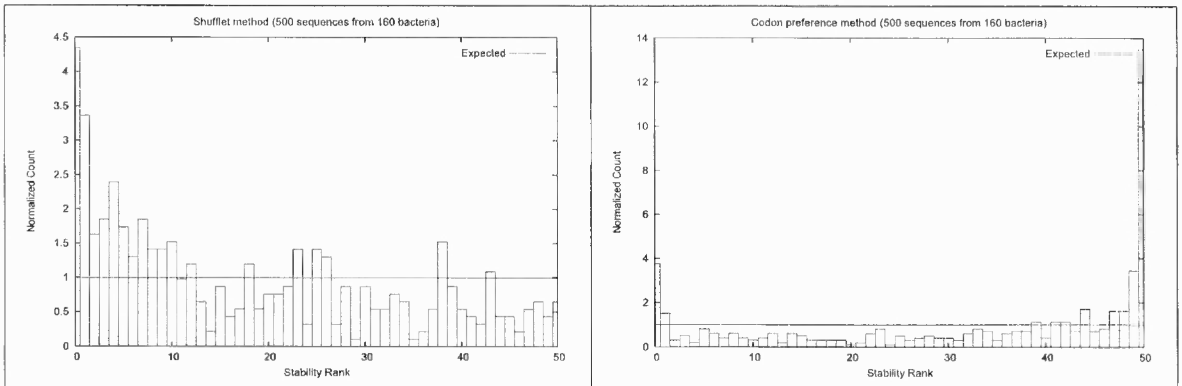
 (a) Naturally occurring sequences show a bias toward low MFE (high stability) compared to sets of randomized sequences with identical dinucleotide composition. In 66 of 500 (13.2 percent) cases, Z ≤ -2.	 (b) Both very low MFE (Z ≤ -2) and very high MFE (Z ≥ +2) occur with greater than normal frequency compared to codon preference randomized sequences. In 47 of 500 (9.4 percent) cases, Z ≤ -2; in 145 of 500 (29.0 percent) of cases, Z ≥ +2.

sequences display a strong bias towards low MFE (high stability). Sixty six of 500 (13.2 percent) natural sequences have a Z score ≤ -2.

Table 2.1(b) shows the stability of the same 500 natural sequences relative to sets of random sequences that maintain the protein product (codon preference method). There is a bias toward both very low MFE (high stability – Z ≤ -2) and very high MFE (low stability – Z ≥ +2). Forty seven of 500 (9.4 percent) wildtype sequences are at least 2 SD more stable than the mean MFE of the corresponding set of randomized synonymous sequences. One hundred and forty five of 500 (29.0 percent) natural sequences are at least 2 SD less stable that the mean MFE of corresponding set of randomized synonymous sequences.

The stability of these sequences were also analyzed by quantiles. (Table 2.2 (a) and (b)). In the absence of any selective force it is expected to be evenly distributed across quantiles. A normalized frequency for each quantile may be calculated by dividing

Table 2.2 Quantile Analysis of MFE of 500 Bacterial Sequences Based on Shufflet and Codon Preference Methods



(a) The large number in quantiles 0 and 1 show a strong bias toward highly stable secondary structure.

(b) The large number in quantile 0 shows a bias toward highly stable secondary structure; the large number in quantiles 49 and 50 shows a bias toward very low stability secondary structure.

the observed number of occurrences by the expected number. A normalized frequency significantly greater than 1.0 for low quantiles indicates selection for high stability. A normalized frequency significantly greater than 1.0 for high quantiles indicates selection for low stability. A bimodal distribution indicates selection for both very high and very low stability.

The quantile analysis shows normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of the natural sequences. These results are presented in Table 2.2.

Table 2.2 (a) shows the results of experiments that use shufflet randomization. The values for quantiles 0 and 1 are 4.35 and 3.37, respectively, showing a strong bias towards the highly stable secondary structures.

Table 2.2 (b) shows the results of experiments that use codon preference randomization. The normalized frequencies of quantiles 0 and 1 and 47-50 are significantly greater than 1.0. This shows selection for highly stable and unstable structures.

Table 2.3 MFE of Natural mRNA Sequences of Synechocystis Compared to Synthetic Sequences

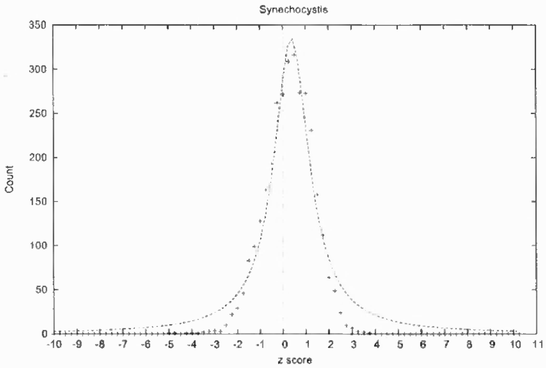
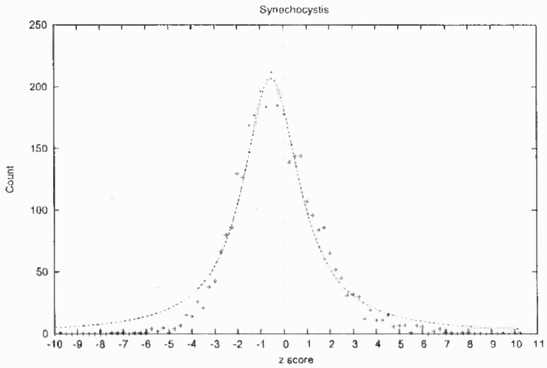
	
(a) No evident bias toward either low or high stability.	(b) Evident bias toward both very low MFE and high MFE. In 549 of 3169 (17.32 percent) cases, Z $\leq$ -2.In 305 cases (9.6 percent) cases, Z $\geq$ +2.

Table 2.4 MFE of Natural mRNA Sequences of Chlamydia trachomatis Compared to Synthetic Sequences

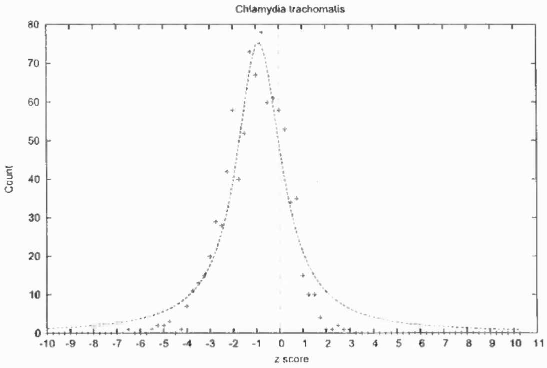
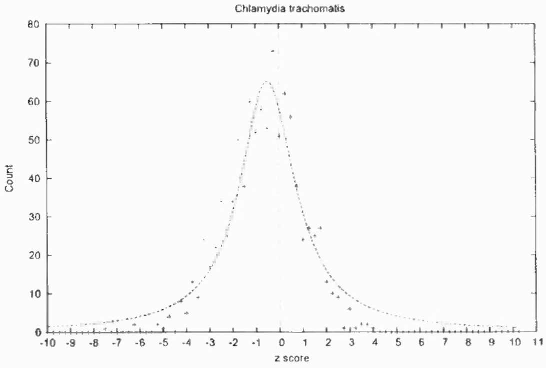
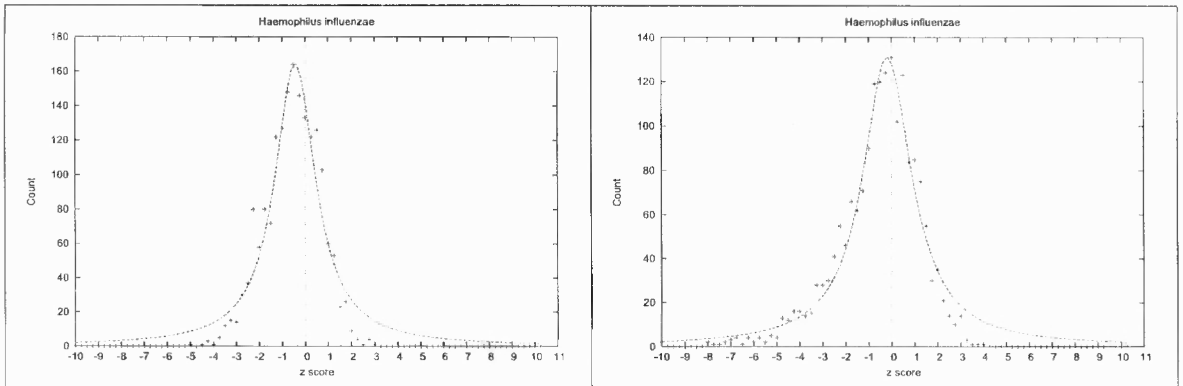
	
(a) Evident bias toward very low MFE. In 233 of 940 (24.78 percent) cases, Z $\leq$ -2.	(b) Evident bias toward very low MFE. In 201 of 940 (21.38 percent) cases, Z $\leq$ -2

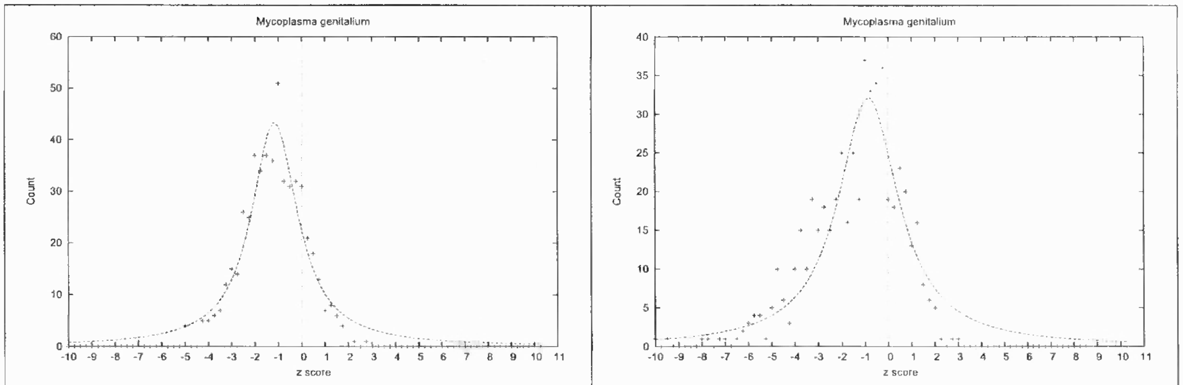
Table 2.5 MFE of Natural mRNA Sequences of Haemophilus influenzae Compared to Synthetic Sequences



(a) Evident bias toward low MFE. In 257 of 1788 (14.37 percent) cases, Z ≤ -2.

(b) Evident bias toward low MFE. In 191 of 1788 (19.63 percent) cases, Z ≤ -2.

Table 2.6 MFE of Natural mRNA Sequences of Mycoplasma genitalium Compared to Synthetic Sequences

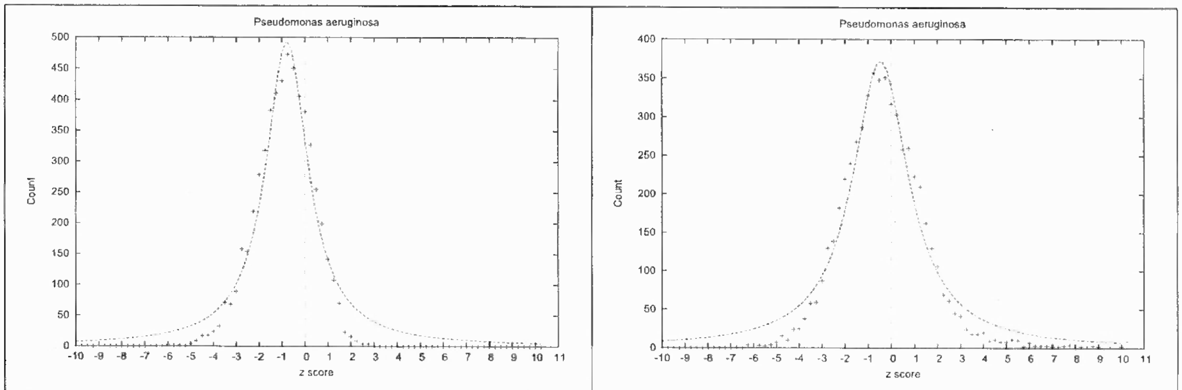


(a) Evident bias toward low MFE. In 157 of 523 (30.0 percent) cases, Z ≤ -2.

(b) Evident bias toward low MFE. In 191 of 523 (36.52 percent) cases, Z ≤ -2.

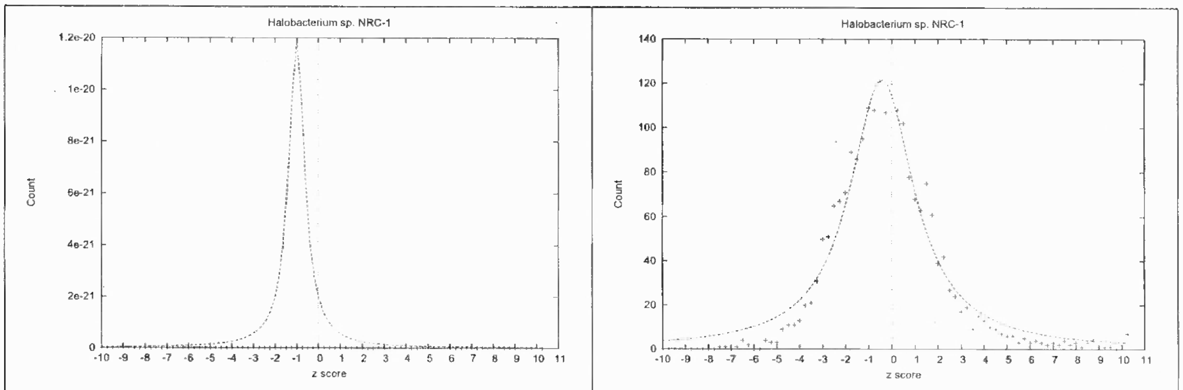
Synechocystis Table 2.3(a) shows the stability (MFE) of the entire population of 3,169 coding sequences of Synechocystis compared to corresponding sets of shufflet randomized sequences. The natural sequences do not show any bias towards either low or high MFE. Only 78 of 3169 (2.0 percent) sequences have Z ≤ -2. Only 96 of 3169 (3.0 percent) of sequences have Z ≥ +2. Table 2.3(b) shows the stability of the same population relative

Table 2.7 MFE of Natural mRNA Sequences of Pseudomonas aeruginosa Compared to Synthetic Sequences



(a) Evident bias toward low MFE. In 1149 of 5571 (20.62 percent) cases, Z ≤ -2	(b) Evident bias toward low MFE. In 1009 of 5571 (18.11 percent) cases, Z ≤ -2

Table 2.8 MFE of Natural mRNA Sequences of Halobacterium sp. NRC-1 Compared to Synthetic Sequences

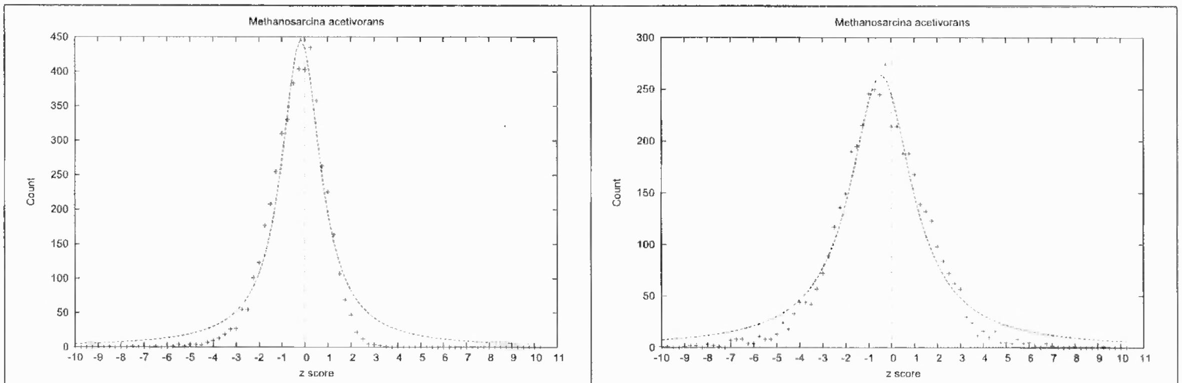


(a) Evident bias toward low MFE. In 367 of 2127 (17.25 percent) cases, Z ≤ -2.	(b) Evident bias toward low MFE. In 439 of 2127 (20.06 percent) cases, Z ≤ -2.

to sets of codon preference randomized synonymous sequences. The sequences display a strong bias towards low MFE. For 549 of 3,169 sequences (17.32 percent), Z ≤ -2. For 305 of 3,169 (9.6 percent) sequences, Z ≥ +2.

Analyzed by quantiles, the 3169 sequences of Synechocystis show normalized frequencies differ systematically from the expected frequencies, and therefore provide evidence for a

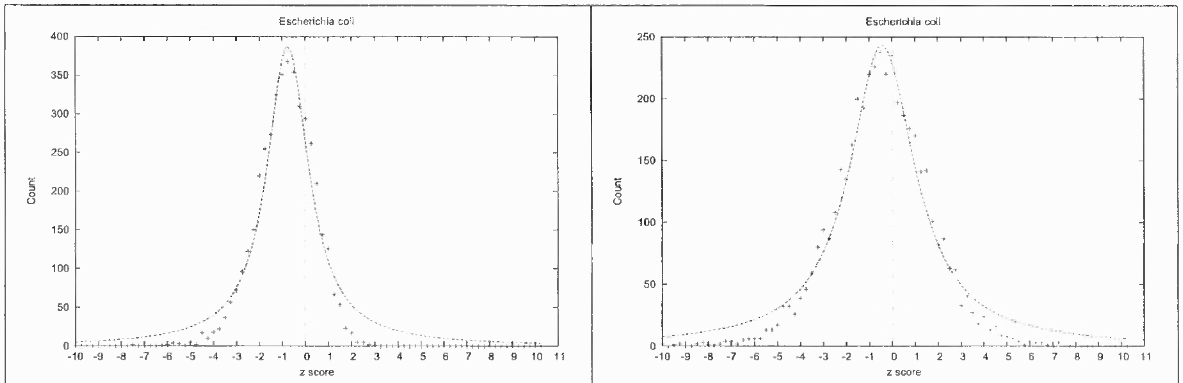
Table 2.9 MFE of Natural mRNA Sequences of Methanosarcina acetivorans Compared to Synthetic Sequences



(a) Evident bias toward low MFE. In 438 of 4662 (9.39 percent) cases, Z ≤ -2.

(b) Evident bias toward low MFE. In 909 of 4662 (19.49 percent) cases, Z ≤ -2.

Table 2.10 MFE of Natural mRNA Sequences of Escherichia coli Compared to Synthetic Sequences

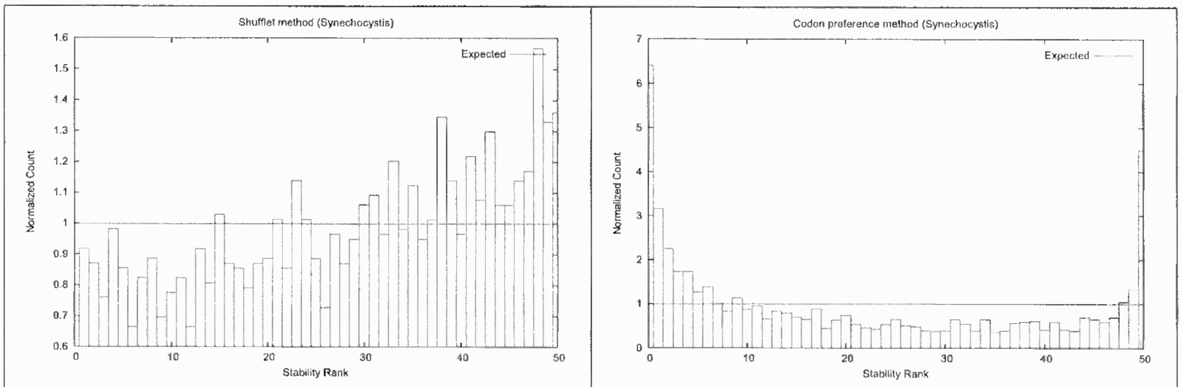


(a) Evident bias toward low MFE. In 842 of 4289 (19.63 percent) cases, Z ≤ -2.

(b) Evident bias toward low MFE. In 971 of 4289 (22.64 percent) cases, Z ≤ -2.

selective force. Table 2.11 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 41 to 50 is more than 1.0. This shows selection for very high MFE. Table 2.11 (b) shows the results of experiments that use codon preference randomization. The normalized frequency of quantiles 0 to 7 and 48 to 50 is significantly higher than 1.0. This shows selection of very high MFE. The values for bins 0, 1, and 2

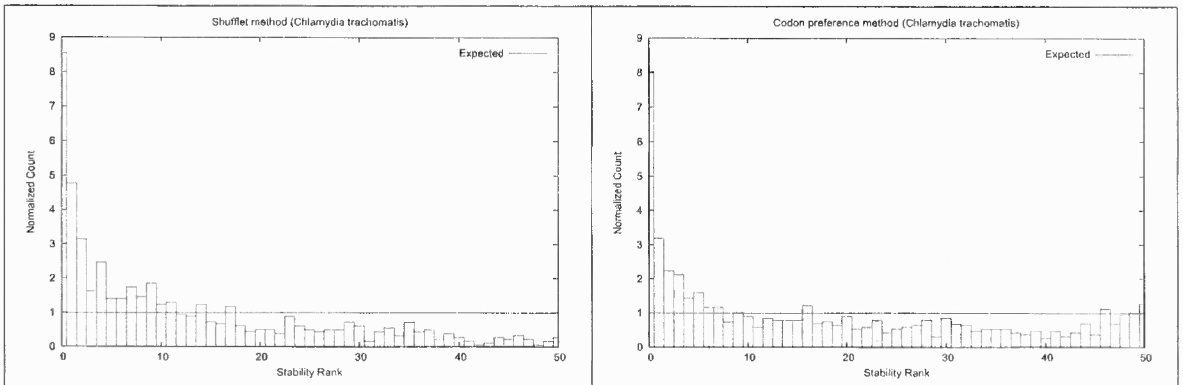
Table 2.11 Quantile Analysis of MFE of Natural mRNA Sequences of Synechocystis Relative to Randomized Sequences



(a) The large number in quantiles 48, 49, and 50 show a bias toward higher MFE.

(b) The large number in quantiles 0, 1, and 2 shows a bias toward very low MFE; the large number in quantile 50 shows a bias toward very high MFE.

Table 2.12 Quantile Analysis of MFE of Natural mRNA Sequences of Chlamydia trachomatis Relative to Randomized Sequences

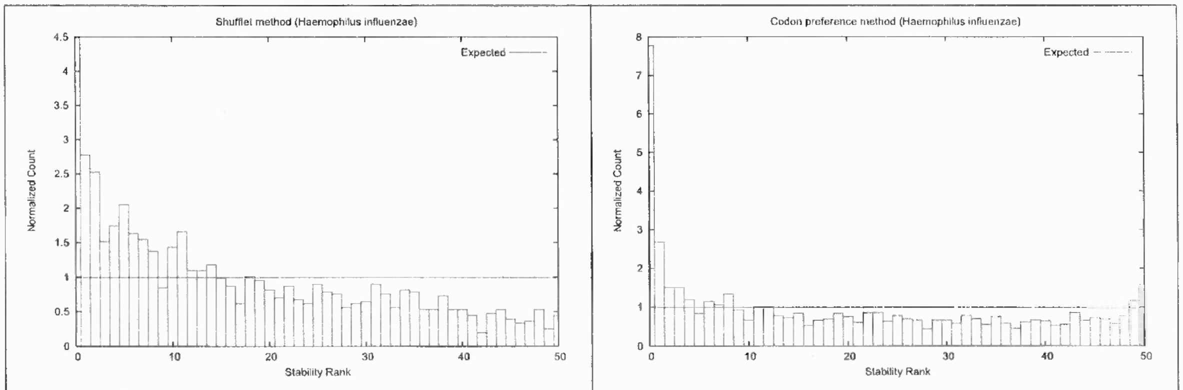


(a) The large number in quantiles 0, 1, and 2 shows a bias toward high MFE.

(b) The large number in quantiles 0 and 1 shows a bias toward very high MFE.

are 6.4, 3.16, and 2.25, respectively. The values for bins 48, 49, and 50 are 1.04, 1.33, and 4.49, respectively. There is a bimodal distribution with the primary mode on the low MFE side.

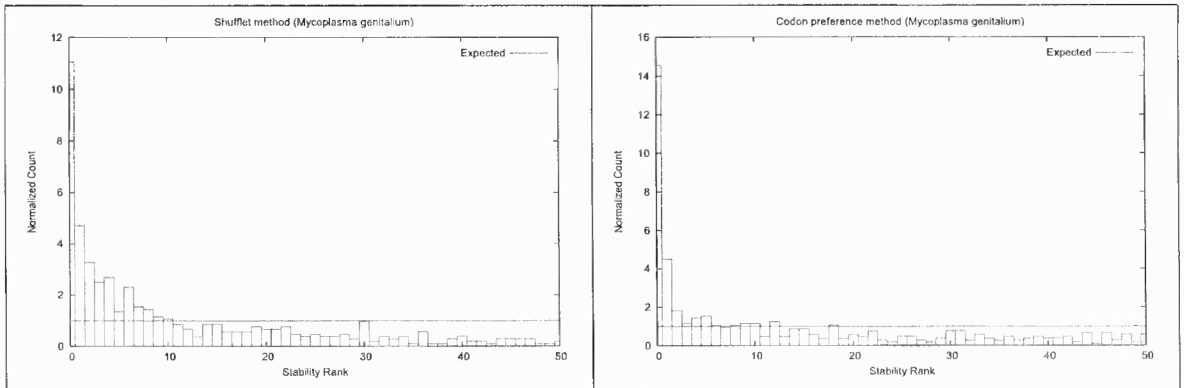
Table 2.13 Quantile Analysis of MFE of Natural mRNA Sequences of Haemophilus influenzae Relative to Randomized Sequences



(a) The large number in quantiles 0, 1, and 2 shows a bias toward very low MFE.

(b) The large number in quantiles 0 and 1 shows a bias toward very low MFE.

Table 2.14 Quantile Analysis of MFE of Natural mRNA Sequences of Mycoplasma genitalium Relative to Randomized Sequences

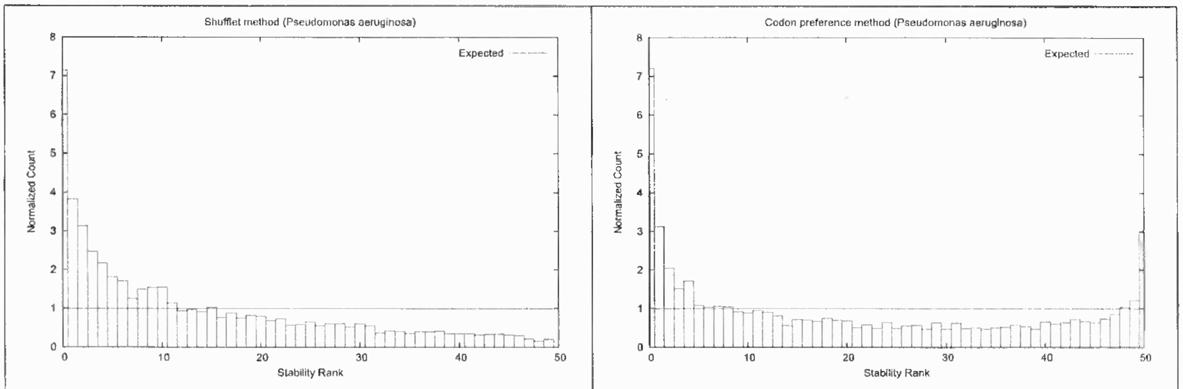


(a) The large number in quantiles 0, 1, and 2 show a strong bias toward very low MFE.

(b) The large number in quantiles 0 and 1 shows a bias toward very high MFE.

Chlamydia trachomatis Table 2.4(a) shows the stability (MFE) of the entire population of 940 natural coding sequences of Chlamydia trachomatis compared to corresponding shufflet randomized sequences. The sequences show a bias towards low MFE. For 233 of 940 (24.78 percent) sequences, $Z \leq -2$. For only 5 of 940 (0.53 percent) sequences is $Z \geq +2$. Table 2.4(b) shows the same population relative to sets of codon preference randomized

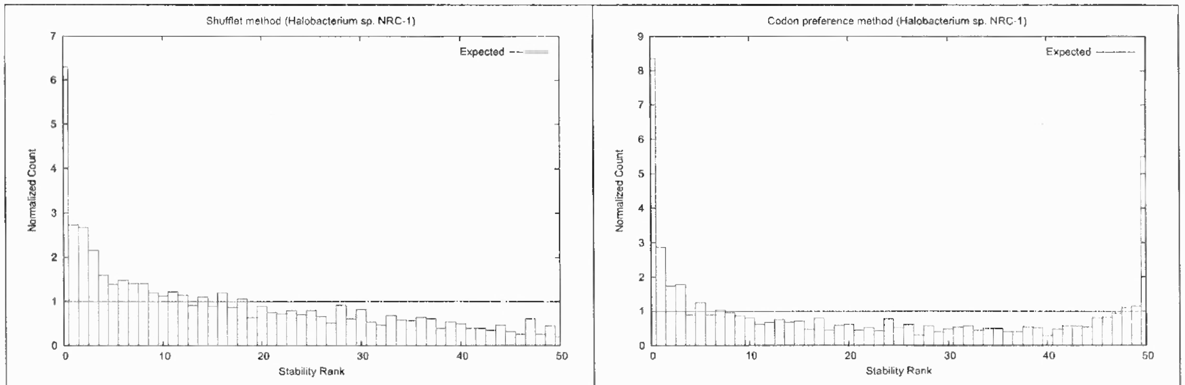
Table 2.15 Quantile Analysis of MFE of Natural mRNA Sequences of Pseudomonas aeruginosa Relative to Randomized Sequences



(a) The large number in quantiles 0, 1, and 3 show a bias toward very low MFE.

(b) The large number in quantiles 0 and 1 shows a bias toward very low MFE; the large number in stability rank 50 shows a bias toward very high MFE.

Table 2.16 Quantile Analysis of MFE of Natural mRNA Sequences of Halobacterium sp. NRC-1 Relative to Randomized Sequences

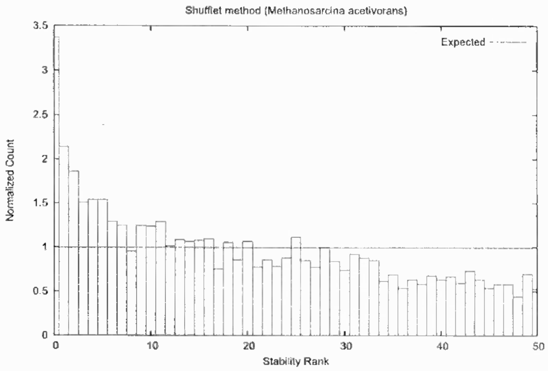


(a) The large number in quantiles 0, 1, and 2 show a bias toward very low MFE.

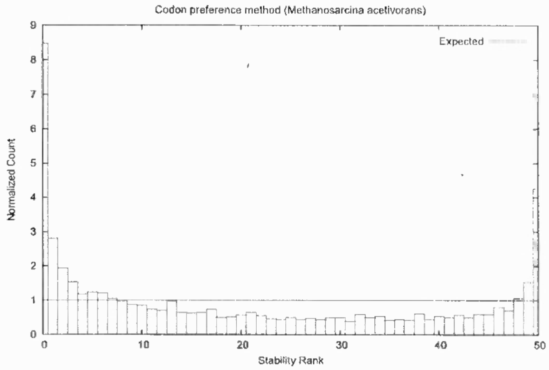
(b) The large number in quantiles 0 and 1 shows a bias toward low MFE; the large number in quantile 50 shows a bias toward very high MFE.

sequences. The sequences display a bias towards low MFE. For 201 of 940 (21.38 percent) sequences, $Z \leq -2$. For only 32 of 940 (3.4 percent) sequences is $Z \geq +2$.

Table 2.17 Quantile Analysis of MFE of Natural mRNA Sequences of Methanosarcina acetivorans Relative to Randomized Sequences

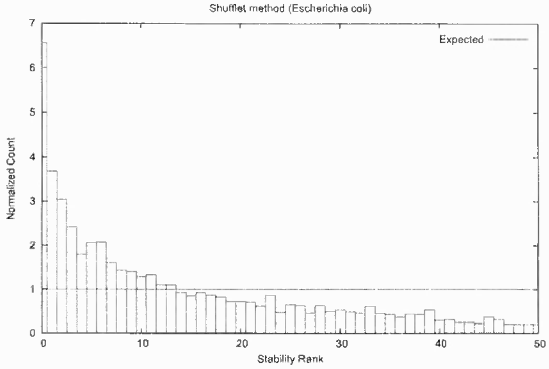


(a) The large number in quantiles 0 and 1 show a bias toward very low MFE.

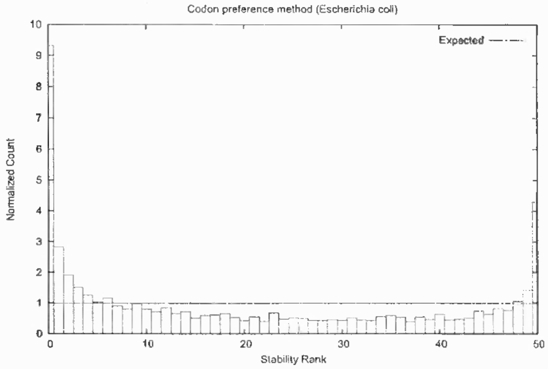


(b) The large number in quantiles 0 and 1 shows a bias toward very low MFE; the large number in quantile 50 shows a bias toward very high MFE.

Table 2.18 Quantile Analysis of MFE of Natural mRNA Sequences of Escherichia coli Relative to Randomized Sequences



(a) The large number in quantiles 0, 1, and 2 show a bias toward very low MFE.



(b) The large number in quantiles 0 and 1 shows a bias toward very low MFE; the large number in stability rank 50 shows a strong bias toward high MFE.

The stability of the same sequences of analyzed by quantile shows normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences. Table 2.12 (a) shows the results of experiments that used the shufflet method of randomization, preserving dinucleotide composition. The quantiles 0 to 11 are greater than 1.0, showing selection for highly stable structures. The values for bins 0, 1, and 3 are 8.54, 4.78 and 3.15, respectively.

Table 2.12 (b) shows the results of experiments that used the codon preference method of randomization that preserves the gene product and uses the frequency of synonymous codons in the organism. Quantiles 0 to 7 and 49-50 are significantly higher than 1.0, showing selection for very low and very high MFE. The values for bins 0, 1, and 2 are 6.4, 3.16, and 2.25, respectively. The values for bins 49 and 50 are 1.33, and 4.49, respectively. This shows the bimodal distribution with the primary mode on the low MFE side.

Haemophilus influenzae Table 2.5(a) shows the stability (MFE) of the entire population of 1,788 coding sequences of Haemophilus influenzae compared to corresponding sets of shufflet randomized sequences. The natural sequences show a strong bias towards low MFE. For 257 of 1,788 (14.37 percent) natural sequences, $Z \leq -2$. For only 10 of 1,788 (0.56 percent) sequences is $Z \geq +2$.

Table 2.5(b) shows the stability of the same population relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 351 of 1,788 (19.63 percent) of sequences, $Z \leq -2$. For only 65 of 1788 (3.6 percent) of sequences is $Z \geq +2$.

The stability of the same population analyzed by quantile show normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences. Table 2.13 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 0 to 14 (except 9) is significantly greater than 1.0, showing selection for highly stable structures. The values for bins 0, 1, and 2 are 4.49, 2.78, and 2.52, respectively.

Table 2.13 (b) shows the results of experiments that use codon preference randomization. The normalized frequency of quantiles 0 to 8 (except 5) and 49 to 50 is significantly higher than 1.0, showing selection for both very low and very high MFE. The values for quantiles 0, 1, and 2 are 7.77, 2.68, and 1.51, respectively. The values for quantiles 49 and 50 are 1.17, and 1.59, respectively.

Mycoplasma genitalium Table 2.6(a) shows the stability (MFE) of the entire population of 523 coding sequences of Mycoplasma genitalium compared to corresponding sets of shuffles randomized sequences. The sequences show a bias towards low MFE. For 157 of 523 (30.0 percent) sequences, Z $\leq$ -2. For only 2 of 523 (0.38 percent) sequences is Z $\geq$ +2.

Table 2.6(b) shows the stability of the same population relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 191 of 523 sequences (36.52 percent), Z $\leq$ -2. For only 3 of 523 (0.57 percent) is Z $\geq$ +2.

The stability of the same sequences analyzed by quantile show normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences. Table 2.14 (a) shows the results of experiments that used shufflet randomization. The normalized frequency of stability rankings 0 to 10 is significantly higher than 1.0. The values for quantiles 0, 1, 2, 3, and 4 are 11.05, 4.71, 3.26, 2.5, and 2.69, respectively

Table 2.14 (b) shows the results of experiments that used codon preference randomization. The normalized frequency of quantiles 0 to 10 (except 7) is significantly higher than 1.0.

This shows the selection of highly stable structures. The values for bins 0, 1, and 2 are 14.55, 4.5, and 1.81, respectively.

Pseudomonas aeruginosa Table 2.7(a) shows the stability (MFE) of the entire population of 5,571 coding sequences of Pseudomonas aeruginosa compared to corresponding sets of shufflet randomized sequences. The sequences show a bias towards low MFE. For 1,149 of 5571 (20.62 percent) sequences, $Z \leq -2$.

Table 2.7(b) shows the stability of the same population relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 1,009 of 5571 (18.11 percent) of sequences, $Z \leq -2$.

The stability of the same sequences analyzed by quantile relative to the codon preference randomized sequences show that normalized frequencies differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences.

Table 2.15 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 0 to 11 is significantly higher than 1.0. This shows selection for highly stable structures. The values for quantiles 0, 1, and 2 are 7.14, 3.83, and 3.13, respectively.

Table 2.15 (b) shows the results of experiments that use codon preference randomization. The normalized frequency of quantiles 0 to 8 and 48 to 50 is significantly higher than 1.0. This shows the selection of highly stable and unstable structures. The values for quantiles 0, 1, and 2 are 7.20, 3.12, and 2.05, respectively. The value for quantile 50 is 2.98.

Halobacterium sp. NRC-1 Table 2.8(a) shows the stability (MFE) of the entire population of 2,127 coding sequences of Halobacterium sp. NRC-1 compared to corresponding sets of shufflet randomized sequences. The sequences show a bias towards low MFE. For 367 of 2127 (17.25 percent) sequences, $Z \leq -2$.

Table 2.8(b) shows the stability of the same population relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 439 of 2,127 (20.06 percent) of sequences, $Z \leq -2$.

The stability of the same sequences analyzed by quantile shows normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences.

Table 2.16 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 0 to 12 is more than 1.0. This shows selection for highly stable structures. The values for quantiles 0, 1, 2, and 3 are 6.29, 2.72, 2.67 and 2.16, respectively

Table 2.16 (b) shows the results of experiments that used the codon preference method of randomization that preserves the gene product and uses the frequency of synonymous codons in the organism. The normalized frequency of quantiles 0 to 3 and 48 to 50 is significantly higher than 1.0. This shows the selection of both very low and very high MFE. The values for quantiles 0, 1, 2, and 3 are 8.35, 2.87, 1.74, and 1.78, respectively. The values for quantile 50 is 5.50.

Methanosarcina acetivorans Table 2.9(a) shows the stability (MFE) of the entire population of 4,662 coding sequences of Methanosarcina acetivorans compared to corresponding sets of shufflet randomized sequences. The sequences show a strong bias towards low MFE. For 438 of 4,662 (9.39 percent) of sequences, Z ≤ -2.

Table 2.9(b) shows the stability of the same population sequences relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 909 of 4,662 (19.49 percent) sequences, Z ≤ -2. For 424 of 4,662 (9.09 percent) of sequences, Z ≥ +2.

The stability of the same sequences analyzed by quantiles show normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences.

Table 2.17 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 0 to 16 (except 8) is more than 1.0. This shows selection for highly stable structures. The values for quantiles 0, 1, and 2 are 3.37, 2.14 and 1.86, respectively.

Table 2.17 (b) shows the results of experiments that used the codon preference method of randomization that preserves the gene product and uses the frequency of synonymous codons in the organism. The normalized frequency of quantiles 0 to 7 and 49 to 50 is significantly higher than 1.0. This shows the selection of very low MFE and very high MFE. The values for bins 0, 1, and 2 are 8.48, 2.81, and 1.93, respectively. The values for quantiles 49 and 50 are 1.54 and 4.27, respectively.

Escherichia coli Table 2.10(a) shows the stability (MFE) of the entire population of 4,289 natural coding sequences of Escherichia coli compared to corresponding sets of shufflet randomized sequences. The sequences show a strong bias towards low MFE. For 842 of 4289 (19.63 percent) of sequences, Z ≤ -2.

Table 2.10(b) shows the stability of the same population relative to sets of codon preference randomized sequences. The sequences display a bias towards low MFE. For 971 of 4,289 (22.64 percent) of sequences, $Z \leq -2$. For 426 of 4,289 (9.9 percent) of sequences, $Z \geq +2$.

The stability of the same sequences analyzed by quantile shows normalized frequencies that differ systematically from the expected frequencies, and therefore provide evidence for a selective force acting on the stability of these natural sequences.

Table 2.18 (a) shows the results of experiments that used the shufflet method of randomization, preserving dinucleotide composition. The normalized frequency of quantiles 0 to 13 is significantly higher than 1.0. This shows selection for highly stable structures. The values for quantiles 0, 1, and 2 are 6.56, 3.68 and 3.04, respectively.

Table 2.18 (b) shows the results of experiments that use codon preference randomization. The normalized frequency of stability rankings 0 to 6 and 49 to 50 is significantly higher than 1.0. This shows the selection of highly stable and unstable structures. The values for bins 0, 1, and 2 are 9.31, 2.83, and 1.91, respectively. The values for bins 49 and 50 are 1.43, and 4.29, respectively.

2.2.2 Corroboration of Results by Independent Methods

Each of the randomization methods identifies a set of sequences in each organism that have been selected for low or high MFE. By comparing the intersection of these sets, it can be established whether each of the two methods corroborate the other's results.

The Figure 2.19 gives the number of genes in each bacteria and the number of genes with Z scores less than -2 and greater than +2. The intersection of genes based on the is statistically significant for each data set

Table 2.19 Count of Low and High MFE Bacterial Genes

	$Z \leq -2$			$Z \geq +2$			
Organism	Shufflet	Codon Preference	Overlap	Shufflet	Codon Preference	Overlap	Total Sequences
Random Bacteria Sequences	66	47	22	5	145	4	500
AB001339	78	549	26	96	305	16	3169
NC_000117	233	201	111	5	32	3	940
NC_000907	257	351	142	10	65	2	1788
NC_000908	157	191	105	2	3	0	523
NC_002516	1149	1009	502	17	403	9	5571
NC_002607	367	439	170	15	257	6	2127
NC_003552	438	909	207	46	424	22	4662
U00096	842	971	433	13	426	8	4289

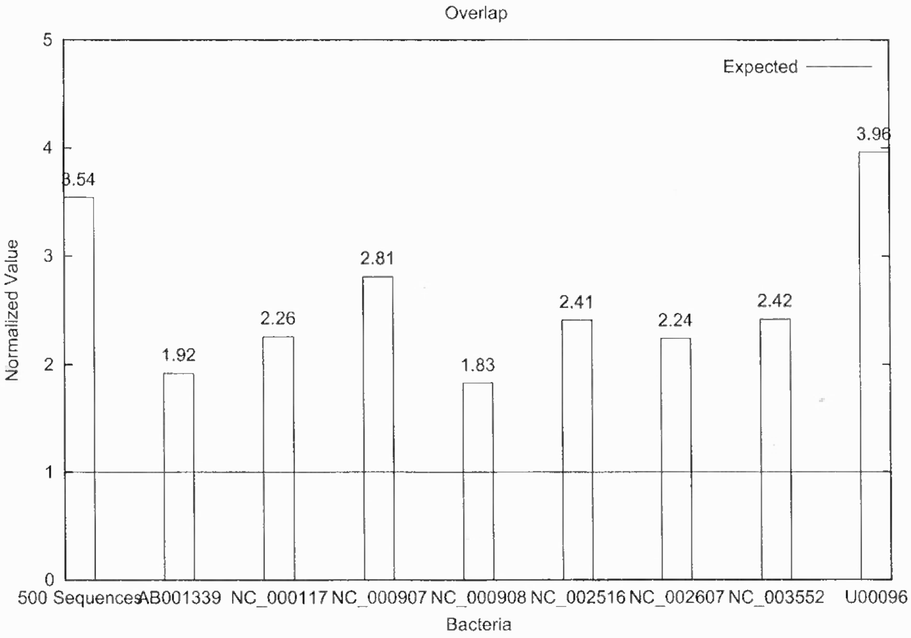


Figure 2.1 Expected value of intersection between Shufflet and Codon Preference methods.

2.2.3 Fruitfly

Drosophila melanogaster is a eukaryotic species. It has four chromosomes with 18,312 coding sequences, as long as 194,916 bases. Figure 2.2 shows the stability (MFE) of a

Table 2.20 Count of Bacterial Genes in Extremal Quantiles

	Quantile 0			Quantile 50			
Organism	Shufflet	Codon Preference	Overlap	Shufflet	Codon Preference	Overlap	Total Sequences
AB001339	43	403	13	86	283	10	3169
NC_000117	152	151	73	5	24	3	940
NC_000907	160	278	83	16	57	2	1788
NC_000908	115	152	74	2	6	0	523
NC_002516	794	801	319	16	332	6	5571
NC_002607	268	355	115	9	234	4	2127
NC_003552	313	748	122	49	377	25	4662
U00096	563	799	279	17	368	9	4289

Table 2.21 Statistical Significance (p-value) of the Intersection from Table 2.19

Organism	$Z \leq -2$	$Z \geq +2$
Random Bacteria Sequences	≤ 0.0001	0.031
AB001339	0.995	0.992
NC_000117	≤ 0.0001	≤ 0.0001
NC_000907	≤ 0.0001	0.124
NC_000908	≤ 0.0001	0.1
NC_002516	≤ 0.0001	0.176
NC_002607	≤ 0.0001	0.152
NC_003552	0.005	0.269
U00096	≤ 0.0001	0.076

random sample of 500 coding sequences from this population. For each natural sequence, the set of 50 randomized sequences was generated by the shufflet method. The sequences do not display statistically significant bias either towards low MFE or high MFE. There are two peaks, one at -0.75 and +0.25. For 19 of 500 (3.8 percent) sequences, $Z \leq -2$. For 14 of 500 (2.8 percent) sequences, $Z \geq +2$.

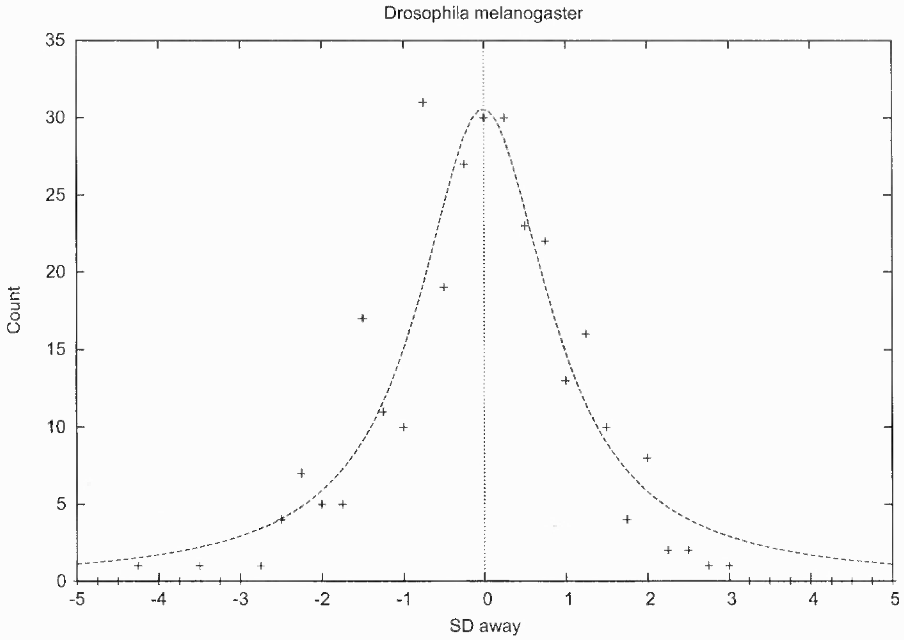


Figure 2.2 Drosophila melanogaster MFE based on Shufflet method shows no specific bias.

2.2.4 Yeast

Saccharomyces cerevisiae is a eukaryote with 16 chromosomes and 1 mitochondria. The total number of genes is about 6,226. The complete set of genes of Saccharomyces cerevisiae was folded, along with 50 randomly generated shufflet sequences for each of the natural sequences. Table 2.22 gives the stability of genes in each of the chromosomes. Seven of the chromosomes shows bias towards high MFE and three of the chromosomes shows bias towards low MFE. Six of the chromosomes show a bimodal distribution with one of them having a bigger peak in the unstable side.

Table 2.23 shows the stability analysis of yeast genes based on shufflet and codon preference randomization methods. In the shufflet experiment, there are 251 genes (0.4 percent) with $Z \leq -2$ and 195 genes (3.1 percent) with $Z \geq +2$. In the codon preference

Table 2.22 Stability of Saccharomyces cerevisiae Chromosomes

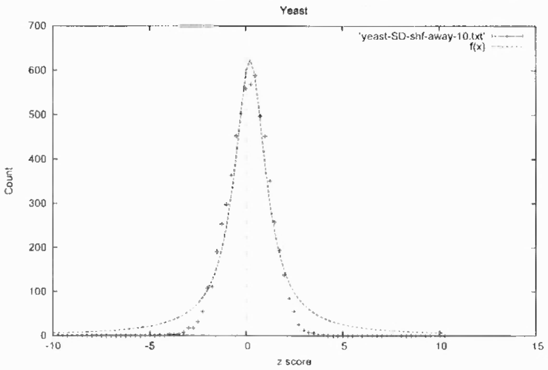
Chromosome	**Stability**
1	Unstable
2	Stable
3	Stable
4	Unstable
5	Unstable
6	Bimodal
7	Bimodal
8	Stable
9	Bimodal
10	Slightly bimodal
11	Bimodal
12	Unstable
13	Unstable
14	Unstable
15	Unstable
16	Bimodal with a bigger peak in unstable region
All	Unstable

experiment, there are 821 genes (13.2 percent) with $Z \leq -2$ and 358 (5.7 percent) genes with $Z \geq +2$.

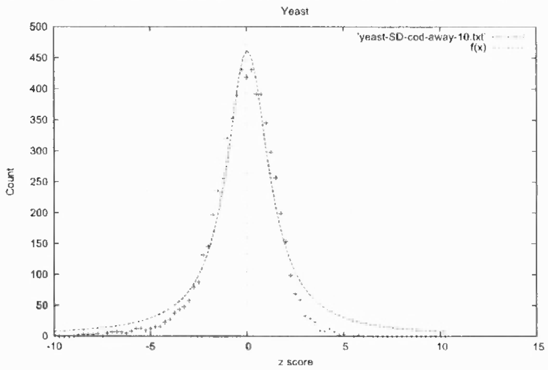
Table 2.24 (a) shows the results of experiments that use shufflet randomization. The normalized frequency of quantiles 0 to 6 and 46 to 50 is more than 1.0. This shows selection of very low MFE and very high MFE.

Table 2.24 (b) shows the results of experiments that used the codon preference method of randomization that preserves the gene product and uses the frequency of synonymous codons in the organism. The normalized frequency of quantiles 0 to 5 and 47 to 50 is significantly higher than 1.0. This shows the selection of very low MFE and very high MFE.

Table 2.23 Yeast Stability Analysis Based on Z Scores

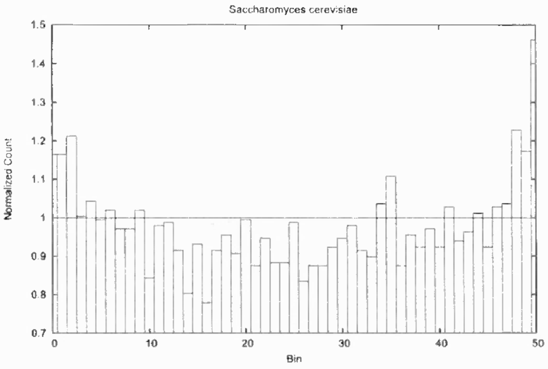


(a) The graph shows the peak is on the unstable side but there are 251 sequences where Z ≤ -2 and 195 sequences that where Z ≥ +2.

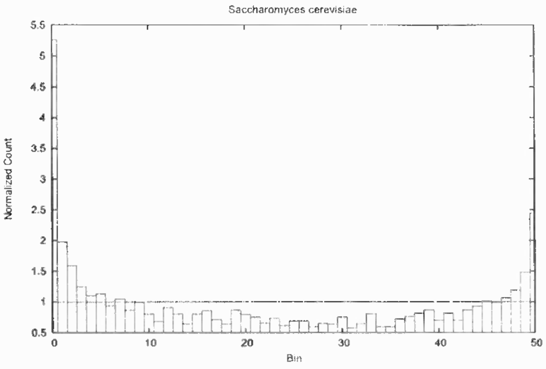


(b) There are 821 sequences where Z ≤ -2 and 358 sequences where Z ≥ +2, based on codon preference randomization.

Table 2.24 Yeast Stability Analysis Based on Quantile Analysis



(a) The large number in quantiles 0, 1, and 2 shows a bias toward very low MFE; the large number in quantile 48, 49, and 50 shows a bias toward very high MFE.



(b) The large number in quantiles 0, 1, and 2 shows a bias toward very low MFE; the large number in quantile 48, 49, and 50 shows a bias toward very high MFE.

CHAPTER 3

FREE ENERGY OF BACTERIAL AND YEAST MRNA CORRELATES TO GENE FUNCTIONAL GROUP

3.1 Introduction

3.1.1 Background and Related Work

Messenger RNA (mRNA) is a polymer molecule comprised of four types of bases, denoted A, C, G, and U. It serves as a template, coding for a protein. This genetic code has a triplet form, with each set of three adjacent nucleotides (codons) coding for a single peptide. Since there are $3^4 = 64$ possible codons and only 20 peptides, the genetic code is redundant. Most peptides are coded for by multiple codons, called synonymous codons. Each protein, consisting of many peptides, has exponentially many possible mRNA encodings.

mRNA is a single stranded molecule that folds back on itself and forms characteristic base pairs (GC, AU, GU). Each base pair, together with the bases enclosed within them, is called a secondary structure, and together the secondary structures define the molecule's secondary structure. The stability of an mRNA molecule is the sum of the minimum free energy (MFE) of its component secondary structures [6]. The more the negative the MFE of a molecule, the more stable it is. Each mRNA molecule is assumed to settle into its thermodynamically minimal (most stable) state.

Therefore, nature, when "choosing" among the exponentially many mRNA encodings for a given protein, is also choosing a particular molecular secondary structure. But why is this encoding and its corresponding structure chosen by nature? As can be seen in Chapter 1, mRNA evolution is not indifferent to the structure of mRNA. It was found that across a wide selection of eubacteria and archaea, encodings and structures have evolved that are more likely to be very stable or very unstable than could be accounted for by chance. From a biological point of view, when one finds evidence for selection for a particular trait, it is natural to ask what selective advantage it grants to the organism. In this case, mRNA stability is viewed as a kind of molecular phenotype, and the traits of the microbe that it is associated with are investigated. Traits that are selected for and against at the molecular level are being looked at, that might impact the survival and/or replication of the organism. In particular, correlations between the stability of the mRNA and the functional class of the target protein are examined.

With the increase in the number of genomes sequenced and the identification of a large number of genes and their protein products, the COG database [16,17] was established to group proteins with similar functionality within an organism and from different organisms. COG classifies genes by delineating clusters of orthologous groups (COG) of proteins. In the COG the conserved genes are classified according to homologous relationship, both paralog and ortholog. Paralogs are distinct genes in the same organism with a common ancestry (and often related function). Orthologs are genes from different organisms that evolved from a common ancestor, and which often have related functionality. When entire proteomes of two organisms are available, orthologs and paralogs may be identified by their sequence similarity. The objective of COG is to identify all matching proteins in the organism, defined as an orthologous group related by speciation or gene duplication. Related orthologous groups are clustered to form functional classes. These clusters correspond to classes of metabolic functions. The proteins encoded by many prokaryotic organism have been analyzed for COG relationships; however, not all proteins and genes have been so classified.

Genes are also grouped into operons, or coregulated genes. An operon is the set of one or more genes along with an operator and promotor that switch the set of genes on and off to produce mRNA. In this work, it is also examined to see whether the genes in a multigenic operon have correlated secondary structure and MFE.

3.2 Material and Methods

3.2.1 Organisms

The organisms were chosen for diversity of features and characteristics, in terms of GC content and gram stain of the genome, metabolism, environment and other characteristics. The GC content of their genomes varies from 42 to 67 percent. Their environments are also quite different. Pseudomonas aeruginosa lives in multiple habitats and is an aerobic bacteria; Methanosacina aetivorans lives in an aquatic environment and is anaerobic. Escherichia coli is host associated and can live with or without oxygen. Synechocystis lives in an aquatic habitat. The bacteria's shape also varies from coccus to rod.

Table 3.1 Number of Genes of Bacteria in Clusters of Orthologous Groups (COG) Database

Organism	Number of Genes in COG database	Number of Genes in Organism
Pseudomonas aeruginosa (PA)	4894	5571
Methanosarcina acetivorans (MA)	2998	4721
Escherichia Coli (EC)	3762	4289
Synechocystis (Syn)	3167	3169
Saccharomyces cerevisiae (SC)	3167	6305

Pseudomonas aeruginosa is an opportunistic pathogen. It causes urinary tract, respiratory, skin and soft tissue infections, bone and joint and gastrointestinal infections and a variety of systemic infections. It is a Gram- negative rod. Almost all strains propel by means of a single polar flagellum. The bacterium is ubiquitous in soil and water, and on surfaces

Table 3.2 Number of Guanine and Cytosine in the Bacteria Analyzed

Name	Number of GC	Number of Nucleotides	Percentage of GC
PA	3761499	5602564	.67
MA	2415626	5751492	.42
Syn	1510700	3109122	.49
EC	2119563	4089837	.52

Table 3.3 Cellular Features of Bacteria

Name	Gram Stain	Shape	Arrangement	Motility	Pathogenic in
PA	Negative	Rod	Singles	Yes	Human
MA	N/A	Irregular coccus	Singles, Aggregates	No	No
Syn	N/A	Coccus	Aggregates	N/A	No
EC	Negative	Rod	Singles Pairs	Yes	Human

in contact with soil or water. Its metabolism is respiratory and never fermentative, but it will grow in the absence of O2 if NO3 is available as a respiratory electron acceptor. Its optimum temperature for growth is 37 degrees Centigrade, and it is able to grow at temperatures as high as 42 degrees.

Methanosarcina species live in oil wells, sewage lagoons, trash dumps, decaying leaves, stream sediments, and similar environments. Only Methanosarcina species possess all three known pathways for methanogenesis. They releases methane into the global carbon cycle. M. acetivorans is unique among archaea in forming multicellular structures

Table 3.4 Environmental Features of Bacteria

Name	Oxygen Req	Habitat
PA	Aerobic	Multiple
MA	Anaerobic	Aquatic
Syn	N/A	Aquatic
EC	Facultative	Host-associated

Table 3.5 Survival Temperature of Bacteria

Name	Optimal temp	Range
PA	25-30 C	Mesophilic
MA	35-40C	Mesophilic
Syn	N/A	Mesophilic
EC	37 C	Mesophilic

or colonies. The complete genome of Methanosarcina acetivorans str. C2A should provide some clues to the organism's capacity to adapt and break down a variety of waste products. At the time of sequencing, the genome of Methanosarcina acetivorans was by far the largest of all sequenced archaeal genomes.

E. coli belongs to the large bacterial family Enterobacteriaceae. They are anaerobic, Gram-negative rods that live in the intestinal tracts of animals in health and disease. E. coli can grow in media with glucose as the sole organic constituent. It can grow in the presence or absence of O2. Under anaerobic conditions it will grow by means of fermentation; however, it can also live by anaerobic respiration, utilizing NO3, NO2 or fumarate.

Synechocystiae are unicellular, photoautotrophic, facultative glucose- heterotrophic cyanobacteria. They are oxygenic photosynthetic with two photosystems, and they can fix nitrogen. Synechocystis sp. PCC 6803 has developed into a model cyanobacterium that scientists around the world are using. Synechocystis sp. PCC 6803 can grow in the absence of photosynthesis if a suitable fixed-carbon source such as glucose is provided.

Saccharomyces cerevisiae is a species of budding yeast. It is the most intensively studied eukaryotic organism. It is the microorganism behind the most common type of fermentation. Saccharomyces cerevisiae cells are round to ovoid, 5-10 micrometers in diameter. It reproduces by a division process known as budding.

3.2.2 COG Database

The COG is the Clusters of Orthologous Groups of proteins and the database has the homologous proteins from completely sequenced genomes or groups of orthologs from different lineages and corresponds to conserved domain. The COG database serves as functional annotation of completely sequenced genomes and as a platform to study genome evolution. The COGs are classified into 17 broad functional categories and consist of 138,458 proteins, which are divided into 4, 873 COGs. The eukaryotes are represented in the Eukaryotic orthologous groups (KOGs). The KOG currently has 4,852 COGs, with 59,838 proteins.

3.2.3 Randomization

Sets of randomized sequences were generated by two different randomization processes for each natural sequence, as a basis of comparison of the stability of the natural sequence. One method (referred to here as the shufflet method) of randomization preserves the nucleotide and adjacent pair (dinucleotide) frequencies. The other method (referred to here as the codon preference method) preserves the protein that the RNA is coding for. The methods are discussed in detail in Chapter 1.

3.2.4 Analysis of Minimum Free Energy (MFE)

The natural and the random sequences were computationally folded to predict their minimum free energy (MFE) and secondary structure. The ViennaRNA package, implementing the nearest neighbor thermodynamic algorithm, was used for this. For each natural mRNA sequence, 50 random sequences were generated by the shufflet method and 50 random sequences were generated by the codon preference method.

Two methods are used to analyze the stability of the mRNA structures of an organism:

- Quantiles – The randomized sequences generated by a given method are ordered by MFE. The natural sequence is then given a rank, or quantile, in this order, based on its MFE. For example, if the natural sequence has a MFE lower than any of the randomized sequences, it has rank or quantile of 0. The quantiles of a set of natural sequences were examined – those belonging to a common COG functional class – to detect a bias in MFEs toward high, low or median values.

- Z scores – A Z score characterizes a particular value relative to the mean and standard deviation of a reference population. For example, if a natural sequence has a MFE value X that is one standard deviation less than the mean MFE value of the corresponding randomized set of sequences, the natural sequence value X has a Z score of -1. Z scores may be expected, in the absence of any selective pressure, to be normally distributed. Therefore the Z scores of a set of mRNAs belonging to a common functional class to detect bias in MFEs were used.

3.3 Results and Observation

3.3.1 Selection of Structures

Tables 3.6 and 3.7 show the stability (MFE) of the entire population of the four organisms (discussed in Materials and Methods section) compared to corresponding sets of randomized sequences. The randomized sequences were generated using two methods: shufflet and codon preference.

Synechocystis does not show any bias towards either low MFE (high stability) or high MFE (low stability) compared to the shufflet sequences, but does show a bias towards the

Table 3.6 Z Scores Using Shufflet Randomization

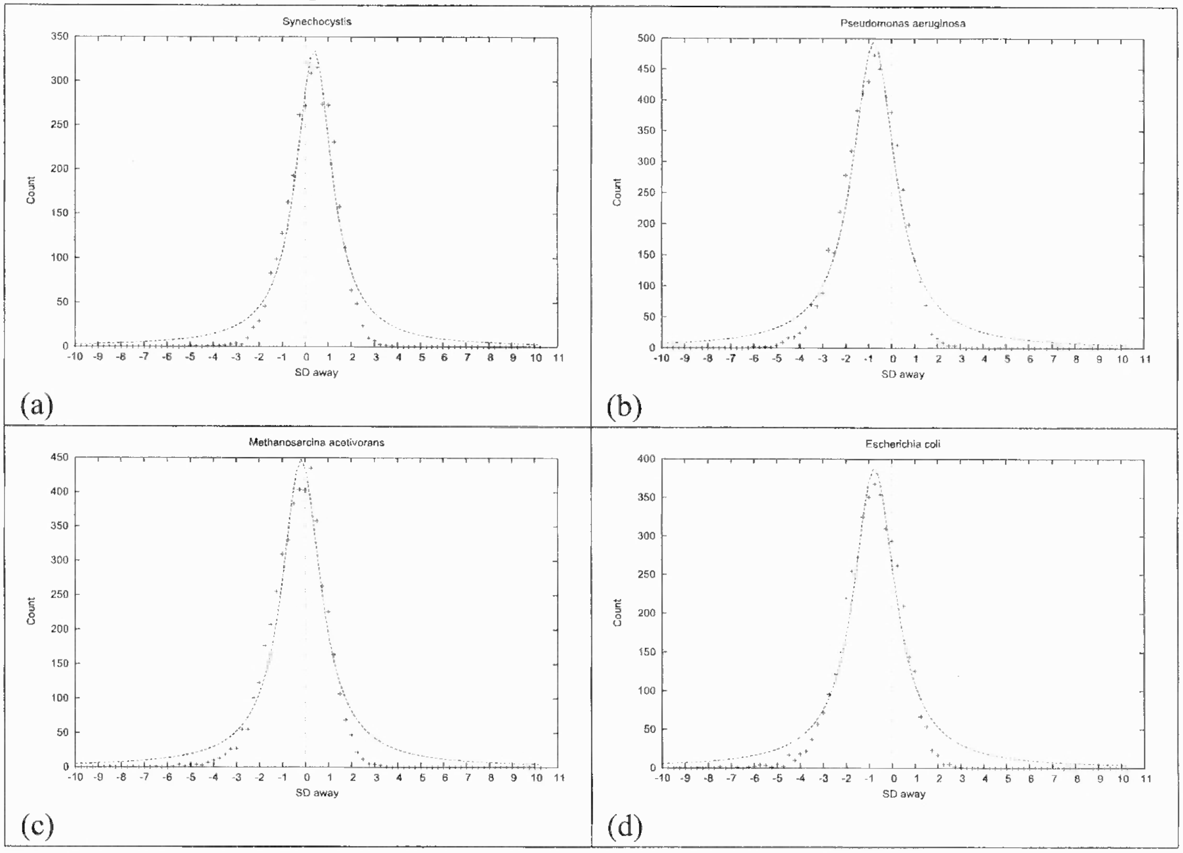
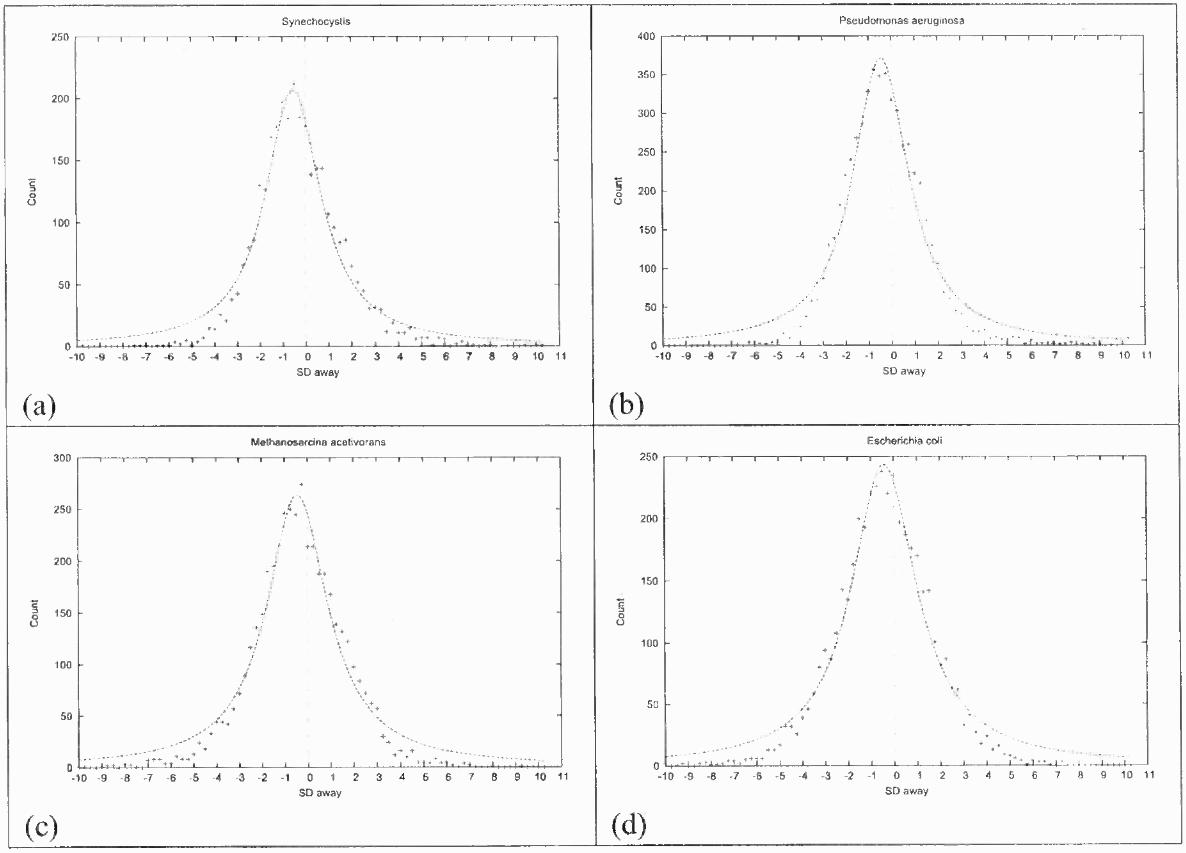


Table 3.7 Z Scores Using Codon Preference Randomization



low MFE (high stability) based on the codon preference method. Pseudomonas aeruginosa, Methanosarcina acetivorans, and Escherichia coli show a strong bias toward low MFE (high stability) based on both shufflet and codon preference methods.

3.3.2 Pattern of Greater Structure in COG Groups Holds Across Organisms

Table 3.8 COG Functional Groups Correlate With Low MFE by Organism Using Z scores

Organism	Shufflet	Codon	Both

Synechocystis	J(0.001)
	V(0.07)
	T(0.04)
		E($\ll$ 0.001)
		P(0.024)
		N(0.066)
		T($\ll$ 0.001)
			T(0.04, 0.0)

Pseudomonas aeruginosa	T(0.0)
	N(0.038)
	L(0.048)
	C(0.092)
	V($\ll$ 0.001)
	G(0.002)
		E(0.001)
		P(0.013)
		H(0.045)
		V($\ll$ 0.001)
		G(0.001)
			V($\ll$ 0.001, $\ll$ 0.001)
			G(0.002, 0.001)

Methanosarcina acetivorans	L($\ll$ 0.001)
	J($\ll$ 0.001)
		C($\ll$ 0.001)
		E($\ll$ 0.001)
		P($\ll$ 0.001)
		H(0.005)
		J($\ll$ 0.001)
			J(0.014, $\ll$ 0.001)

Escherichia coli	J(0.010)
	L($\ll$ 0.001)
	M($\ll$ 0.001)
	C($\ll$ 0.001)
	P($\ll$ 0.001)
	D(0.015)
		E($\ll$ 0.001)
		V(0.01)
		L(0.059)
		M(0.098)
		C($\ll$ 0.001)
		P($\ll$ 0.001)
		D(0.036)
			L($\ll$ 0.001, 0.059)
			M($\ll$ 0.001, 0.098)
			C($\ll$ 0.001, $\ll$ 0.001)
			P(0.06, $\ll$ 0.001)
			D(0.015, 0.036)

The mRNA sequences of the subject organisms were folded and their structure and MFE obtained. The sequences were then grouped into COG functional classes and the correlation of MFEs of the mRNAs with each functional class analyzed. A p- value was calculated for each correlation. The most significant correlations are shown in Table 3.8.

Table 3.9 COG Functional Groups Correlate With MFE by Organism Using Quantile Evaluation

Organism	Shufflet	Codon	Both

Synechocystis	J(1.493,0.008)
	L(1.574,$\ll$ 0.001)
	D(2.652,0.031)		

Pseudomonas aeruginosa	V(1.27,0.029)
	M(1.147,0.006)
	N(1.216,0.008)
	U(1.229,0.007)
	G(1.203,$\ll$ 0.001)
	E(1.048,0.05)
	P(1.154,0.002)
		L(1.051,0.023)
		V(1.273,0.003)
		M(1.128,0.001)
		U(1.078,0.009)
		G(1.292,$\ll$ 0.001)
		E(1.118,$\ll$ 0.001)
		H(1.113,0.002)
		P(1.201,$\ll$ 0.001)
			V
			M
			U
			G
			E
			P

Methanosarcina acetivorans	T (1.0355, 0) R (1.019578313, $\ll$ 0.001) S (1.022877919, 0.001)	K (1.198, 0.05) O (1.046, 0.012) G (1.189, 0.037) H (1.118, 0.013) Q (1.311, 0.071) R (1.016, $\ll$ 0.001)	

Escherichia coli	J(1.308,0)
	M(1.189,0.007)
	E(1.068,0.06)
		J(1.309,0)
		M(1.204,0)
		E(1.076,0.009)
		H(1.099,0.079)
		P(1.07,0.048)
			J
			M
			E

The functional classes A, B, W, X, Y, Z were not included in the analysis because the population sizes of these classes, across organisms, were insufficient to generate statistically significant results.

The functional class E shows more than expected structures (low MFE) in all the bacteria analyzed, based on codon preference randomization, but does not show this when based on shufflet randomization. The functional class E is involved in amino acid transport and metabolism. The same characteristics are displayed by the functional class P, which is also involved in inorganic ion transport and metabolism.

The functional classes M (cell wall/membrane/envelope biogenesis) and D (cell cycle control and mitosis) have more structure (low MFE) in the organism E. coli, by both codon preference and shufflet methods. It is interesting to note that these two functional classes show low MFE only in E. coli and not in the other bacteria.

The functional class G (carbohydrated metabolism and transport) shows lower MFE using both randomization methods in P. aeruginosa.

The functional classes E (amino acid transport and metabolism) and P (inorganic ion transport and metabolism) show low MFE in all the bacteria. The functional class J (translation, ribosomal structure and biogenesis), V (defense mechanisms), L (replication, recombination and repair) and C (energy production and conversion) have low MFE in three of the bacteria analyzed. The functional classes T (signal transduction mechanisms) and H (coenzyme transport and metabolism) have low MFE in just two of the bacteria analyzed.

3.3.3 Intersection of Low MFE Genes Identified by Shufflet and Codon Preference Randomization Methods

The Table 3.10 shows the statistical significance (p-value) of the overlap of genes that have low MFE (high stability), using both the methods of randomization. Most have a very high confidence ($p \ll 0.001$).

3.3.4 Selection For High MFE

The evidence for correlation of high MFE (less secondary structure) with functional class was also looked into. The Table 3.12 shows the functional classes that tend to have high MFE (less stable) genes. The functional classes M and L of Synechocystis show

Table 3.10 Statistical Significance (p-value) of The Overlap Between Shufflet and Codon Preference Methods

Organism	Stable	Unstable
Synechocystis	$\ll 0.001$	0.016
Pseudomonas aeuginosa	$\ll 0.001$	$\ll 0.001$
Methanosarcina acetivorans	0.244	0.009
Escherichia coli	$\ll 0.001$	$\ll 0.001$

Table 3.11 COG Functional Groups Correlated With Low MFE Using Both Shufflet and Codon Preference Methods, by Organism

Organism	Functions
Synechocystis	T (0.038), K (0.06)
Pseudomonas aeuginosa	C, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V (all $\ll$ 0.001)
Methanosarcina acetivorans	C, J, K, P, S, T (all $\ll$ 0.001), H (0.001), R (0.002), E (0.004), F (0.008), G (0.019), O (0.022), L (0.032), M (0.043), U (0.076)
Escherichia coli	C, E, F, G, H, J, K, L, M, O, P, R, S, T, U, D (0.004), N (0.009), Q (0.016)

this phenomenon. None of the other organisms have functional classes with a significant correlation with high MFE.

Table 3.12 COG Functional Groups Correlated With Decreased Negative MFE by Organism

Organism	Shufflet	Codon
Synechocystis	M (0.006)	L (0.0)
Pseudomonas aeruginosa	none	none
Methanosarcina acetivorans	none	none
Escherichia coli	none	none

3.3.5 Enriched GC Alone Does Not Account For Greater Structure

The GC content of the COG functional classes were analyzed to determine if GC content plays a role in some functions tendency to have low MFE. The Figure 3.1 shows that the stability is not only based on GC but also other factors.

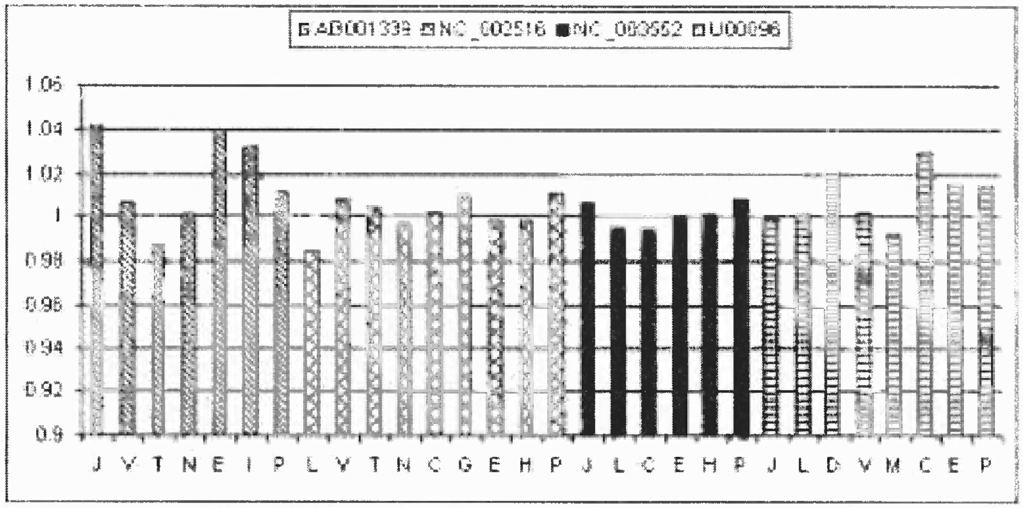


Figure 3.1 Role of GC towards stability based on COG functional classes.

3.3.6 Functional Classes Showing Mean MFE Bias

This section discusses the functional classes of the mRNA sequences that are close to the median MFE values (Tables 3.6 and 3.7).

The mRNA functional group sequences that are close to the median (Z score range -2 to +2) do not have low MFE (high stability) or high MFE (low stability). Genes in functional groups that have small numbers of very high and very low MFE values may be subject to selection for near-median values and against very high or very low values.

The functional classes of mRNA sequences that appear more frequently ($p < 0.05$) in the median range were identified for each method of randomization (Table 3.13).

Table 3.13 COG Functional Groups Correlated With Mean MFE by Organism

Organism	Shufflet	Codon
Synechocystis	none	S, O (0.006)
Pseudomonas aeruginosa	O (0.035), S (0.059)	O (0.051)
Methanosarcina acetivorans	V (0.059), I (0.051), Q (0.056)	none
Escherichia coli	J, K, L, D, V, T, M, N, U, O, C, G, E, F, H, I, P, Q, R, S (all $\ll$ 0.001)	J, K, L, D, V, T, M, N, U, O, C, G, E, F, H, I, P, Q, R, S (all $\ll$ 0.001)

Correlation of Mean MFEs to Functional Class by Organism and Randomization Method The following summarizes the functional classes that show mean bias in MFE, by organism and randomization method.

Synechocystis: Based on codon preference randomization, the functional classes S and O are significantly overrepresented. Based on shufflet randomization, no functional classes are significantly overrepresented.

P. aeruginosa: Based on the shufflet method, the functional classes S and O are significantly overrepresented. Based on the codon preference randomization, the functional group O is significantly overrepresented.

M. acetivorans: Based on the shufflet method, the functional classes V, I, Q are significantly overrepresented. Based on codon preference randomization, no functional classes are significantly overrepresented.

Escherichia coli: Based on both shufflet and codon preference randomization methods, all functional classes are significantly overrepresented.

3.3.7 Yeast COG Functional Classes Show Low MFEs

Table 3.14 shows the COG functional classes that have low MFEs in yeast. The functional classes C, Q, and I show low MFE based on shufflet randomization method. The functional classes G, Q, and I show low MFE based on codon preference method. Thus it can be seen that the functional classes Q and I are there based on both the methods. Also these functional classes are different from the classes that were show low MFE in bacteria.

Table 3.14 COG Functional Groups Correlated With Low MFE For Yeast

Organism	Shufflet	Codon	Both
Yeast	C (0.002) Q (0.066) I (0.034)	G (0.019) Q (0.039) I (0.021)	Q (0.066, 0.039) I (0.034, 0.021)

3.3.8 No Significant Correlation of MFEs of Co-regulated Genes

The operons of E Coli were analyzed to find if there is a correlation between the stability and the operons.

Table 3.15 gives the number of coregulated genes in each of the operons in E Coli. As can be seen, the operons with one and two genes are most common. There are operons with as many as 15 coregulated genes.

The analysis was done on the operons by the following method. First the standard deviation of MFEs all the adjacent genes in operons in the entire organism was calculated. This process was then repeated, using genes chosen at random order for both the forward and reverse orientation of the strands. The results are summarized in the following tables. The MFE values of coding sequences based on shufflet randomization was chosen, since

Table 3.15 Number of Operons in Escherichia coli

Genes in Operon	Number of Operons
1	256
2	144
3	72
4	53
5	35
6	13
7	9
8	4
9	5
10	3
11	2
12	1
13	1
14	0
15	3

the conservation of nucleotide and dinucleotide content excludes any influence of local variation in the nucleotide composition and noise in the randomization process.

Table 3.16 MFEs of Operons Analysis Based on Averaging The SD

Operon Size	Num	All	Operons	Random
2	144	0.98	1.06	1.67
3	72	1.09	1.13	1.5
4	53	1.13	1.27	1.56
5	35	1.16	1.4	1.38
6	13	1.17	1.19	1.32
7	9	1.18	1.03	1.31
8	4	1.19	0.97	1.54

The analysis shows that the MFEs of adjacent genes are significantly correlated. However, as can be seen in the Tables 3.19, 3.18, there is no significant correlation of the MFEs of coregulated genes.

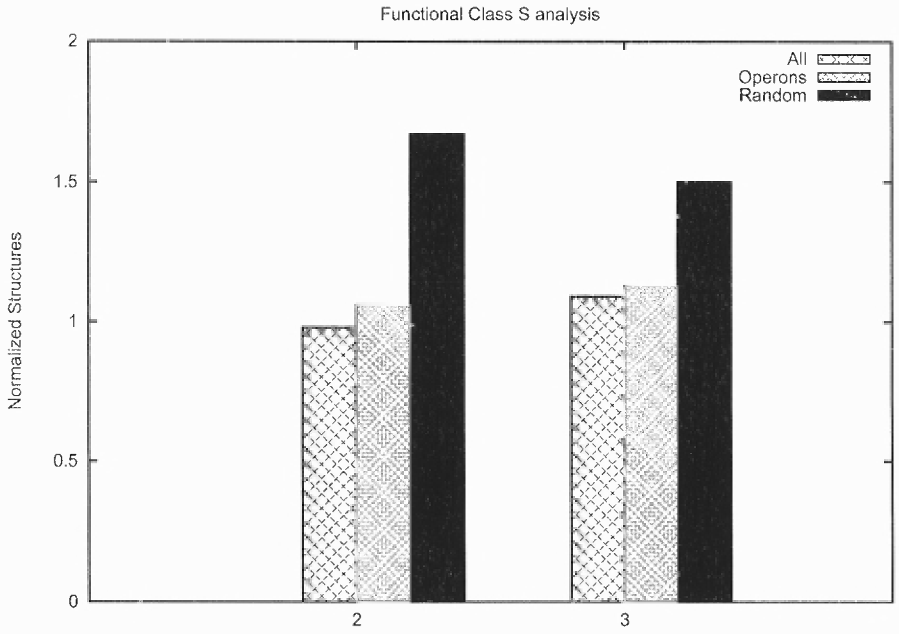


Figure 3.2 Average of the standard deviations of the MFE of E. Coli for operons with two and three genes.

3.3.9 Very Highly Stable Genes

In this section, genes that have Z scores of -5 or less are identified. The tail probability of a Z score of -5 is 2.867 E-7; consequently the probability of encountering such a score by chance is negligible. Further research on the reasons such extreme structural characteristics have evolved in these molecules could be fruitful.

The following excludes results obtained from codon preference randomization; Z scores are based only on shufflet randomized sequences.

The reason for excluding codon preference results is that this method tends to inflate the size of the extreme tails of the distribution in two ways. The first reason flows from a biological fact: the nucleotide and dinucleotide composition of an organism is not entirely uniform, and consequently any particular natural sequence will likely differ in these respects

Table 3.17 E. Coli Operons Analysis Based on SD of The Operons

Klet	Num	All	Operons	Random
2	144	0.75	0.7	0.76
3	72	0.58	0.74	0.6
4	53	0.49	0.64	0.51
5	35	0.44	0.44	0.44
6	13	0.4	0.37	0.41
7	9	0.37	0.27	0.38
8	4	0.34	0.34	0.35

Table 3.18 E. Coli Operons Analysis Based on SD of Operons With Forward Orientation

Klet	Num	All	Operons	Random
2	66	0.75	0.68	0.78
3	33	0.58	0.66	0.61
4	31	0.49	0.69	0.5
5	15	0.44	0.53	0.45
6	6	0.4	0.25	0.4
9	4	0.32	0.23	0.33

from the organism as a whole. This affects the raw MFE value to be scored. The second reason flows from the character of random generation: a codon randomized sequence will likely differ in nucleotide and dinucleotide composition from the organism as a whole. This affects the mean and standard deviation of the standard population. Shufflet randomization exactly preserves nucleotide and dinucleotide content, and therefore is not subject to either of these distorting factors.

The following genes are extreme low energy outliers, with Z scores of -5 or less. The bacteria studied have a considerable number of such structurally extremal genes; none was found in yeast.

Synechocystis gene sll1441, which belongs to the COG functional class I (Lipid transport and metabolism), has Z score -5.14. In P. aeruginosa genes that have Z score

Table 3.19 E Coli Operons Analysis Based on SD of Operons With Reverse Orientation

Klet	Num	All	Operons	Random
2	81	0.75	0.72	0.76
3	39	0.58	0.82	0.58
4	22	0.49	0.58	0.52
5	20	0.44	0.39	0.46
6	7	0.4	0.49	0.49

of -5 or less belongs mostly to COG functional class Metabolism. In M. acetivorans COG functional class L genes predominate. In E. Coli the COG functional classes C and G predominate.

Table 3.20 Very Highly Stable Genes in Synechocystis With Z $\leq$ -5

Gene Name	COG Functional Class	Protein
sll1441	I	delta 15 desaturase

Table 3.21 Very Highly Stable Genes in Pseudomonas aeruginosa With Z $\leq$ -5

Gene Name	COG Functional Class	Protein
PA1124	F	deoxyguanosinetriphosphate triphosphohydrolase
PA1278	H	adenosylcobinamide kinase/adenosylcobinamide-phosphate
PA1487	C	carbohydrate kinase
PA1549	P	cation-transporting P-type ATPase
PA1833	CR	putative oxidoreductase
PA2158	ER	putative alcohol dehydrogenase (Zn-dependent)
PA2439	O	hypothetical protein
PA3636	M	2-dehydro-3-deoxyphosphooctonate aldolase
PA4921	IR	hypothetical protein
PA5036	E	glutamate synthase subunit alpha

Table 3.22 Very Highly Stable Genes in Methanosarcina acetivorans With Z ≤ -5

Gene Name	COG Functional Class	Protein
MA1050	L	transposase
MA1093	J	50S ribosomal protein L30P
MA1459	N/A	hypothetical protein
MA1621	N/A	hypothetical protein
MA2157	N/A	hypothetical protein
MA4678	N/A	hypothetical protein
MA3645	L	reverse transcriptase
MA4036	V	ABC transporter, ATP-binding protein
MA4530	L	transposase
MA4615	E	2-isopropylmalate synthase
MA0785	N/A	proteophosphoglycan
MA0089	S	hypothetical protein
MA0901	R	sodium/chloride-dependent transporter

Table 3.23 Very Highly Stable Genes in Escherichia coli With Z ≤ -5

Gene Name	COG Functional Class	Protein
rnhB	L	ribonuclease HII (rnhB)
yahF	C	putative enzyme with acyl-CoA domain
prpR	KT	prpR
yaiD	N/A	DNA-binding protein, non-specific
sucA	C	2-oxoglutarate dehydrogenase E1 component
tolB	U	periplasmic protein
yccY	T	phosphotyrosine-protein phosphatase
ynjI	N/A	predicted inner membrane protein
mglB	G	methyl-galactoside transporter subunit MglB
yfiC	R	YfiC
uxaC	G	uxaC
rplR	J	50S ribosomal protein L18
gntU	GE	luconate transporter GntU, low affinity GNT 1 system
yhjA	P	YhjA
yidG	N/A	predicted inner membrane protein
thiF	H	thiazole biosynthesis adenylyltransferase ThiF
alsK	KG	D-allose kinase

CHAPTER 4

EVIDENCE OF NATURAL SELECTION ON MFE OF SUFFIXES AND PREFIXES OF BACTERIAL MRNAS

4.1 Background and Introduction

RNA secondary structure is formed through side-chain hydrogen bonds betweem canonical Watson-Crick pairs (GC, AU) between short stretches of RNA [18]. RNA molecules are characterized by unique folding pathways and structural motifs. The pathways involve multiple transitions and stable intermediates. The structure is formed in the presence of divalent metal ions and high ionic strength, both of which minimize the electrostatic repulsion between phosphate groups. The hydrophobic effect, hydrogen bonding, metal ion coordination and vander Waals forces contribute to the formation of the structure. The hydrophobic effects in RNA occcur mainly at the level of secondary structure, making a contribution to the vertical stacking of purine and pyrimidine bases [18, 19]. The RNA folding is opposd by a large configurational entropy due to the reduction in local backbone rotations and the compactness of native states and also the electrostatic repulsion from the negatively charged phosphate backbone. The loops and bulges decrease the entropy of the single strand, so they form only if the free energy decrease of base pair formation more than balances the cost of loop closure. Complementary base pairs will collide randomly, but a single base-pair is never stable in aqueous solution [20].

Levinthal's paradox [21] – the fact that complex macromolecules are able to fold in minutes or less despite the astronomical time needed to search all potentially accessible conformational states – may relate also to RNA folding. The general solution lies in the presence of intermediate organizational levels of the RNA chain, which create a hierarchical folding pathway leading to the native structure. The directionality of chain elongation from the 5'-to the 3'- terminus imprints next-neighbor interactions in the growing chain. Two mechanisms of folding: one in which the entire molecule is first synthesized as a "random" coil, then folded into a functional conformation; and the other in which the molecule is folded sequentially and with some concomitance to its synthesis. Intermediate folded structures may or may not rearrange as the folding progresses. Given a sequence, next-neighbor interaction leads first to regular structural elements of secondary and 3-D motifs, which subsequently merge into domains as structural units that fold seperately. Performed domains, at the next level, associate to form the compact tertiary structure without much reorganization. Although reasonable, it is not yet entirely proven whether or not the mechanism by which RNA folds proceeds with the formation of such structural nuclei and if the same mechanism is applicable to any RNA molecule. The folding of substructure could lead to kinetic traps and thus to nonnative conformers separated from the native ones by high-energy barriers. The underlying reason is the extraordinarily high stability of RNA secondary structures compared to RNA tertiary structure. [22]

The assumption that RNA folds by first forming secondary structure and then forming tertiary interactions from the unpaired bases is not always true [23]. The intermediates during the folding to the final secondary struture must progress downhill energetically, to allow folding on observed time scales, and the final folded state must sit in a relatively narrow region at the bottom of the energy landscape to ensure that the specific functional state is favored over the ensemble of all other possible conformations. The landscape is a representation of intermediates and their energetic connections with each other and with the native and unfolded states. Beyond this energetic description, structural descriptions of intermediate forms are needed. The folding properties strongly depend on the region of the landscape from which the folding starts. The different starting structures lead to folding along discrete pathways or the different starting structures, instead represent successive intermediates along a single folding pathway. Strong evidence for discrete folding pathways have been found [24]. For RNA it is particularly striking that the starting states are not simply ”unfolded” but rather contain substantial structure. The structual features in the starting states and structural differences between different starting states can have profound effects on folding. The RNA forms a nonspecifically collapsed intermediate and then searches for its tertiary contacts within a highly restricted subset of conformational space. As a macromolecule folds to its functional form, it must undergo compaction from a disordered chain to a specific structure [25].

4.2 Normalized Stability of Bacterial mRNA Prefixes

This experiment analyzes bacterial mRNA prefixes of lengths 60, 90, 120 and 150 for evidence of possible natural selection on their MFE and GC content.

4.3 Experiment

The MFE of each prefix is found by computationally folding folding, using the MFOLD implementation of the nearest neighbor secondary structure prediction algorithm, as described in Chapter 1. Additionally, 50 *controlled-shufflets* of the prefixes are generated and folded. The *controlled-shufflets* are generated by shuffling window lengths 30 and then joining them. Consequently, the nucleotide and dinucleotide content of each 30 nucleotide segment is conserved, as is the nucleotide and dinucleotide content of the entire prefix being examined. This joining procedure is employed for prefix lengths of 60, 90, 120 and 150.

4.3.1 Stability Analysis

The *relative stability* of a sequence is found by locating its MFE among the MFEs of the shufflet sequences. If no selective factor is influencing the MFE of the sequence, it is expected that the MFEs of prefixes across the organism will be evenly distributed among the energy quantiles of the corresponding shuffled sequences. If the MFEs of natural sequences is concentrated in the lower energy quantiles of shuffled sequences, then the sequences are highly stable. Quantile 0 is the most stable and quantile 50 is least stable.

The graphs below show the MFEs of the folded bacterial mRNA prefixes of various sizes.

4.3.2 GC Content Analysis

GC content is significant in the investigation of RNA MFE because the pair bonds that form RNA secondary structure are of varying strength: the GC bond is formed by three shared electrons, the AU bond by two, and the GU bond by one.

This GC content analysis compared the GC content of a sequence to its computationally calculated stability. The graph plotted is between the energy quantiles and the *normalizedGC*(normGCbin), where

$$normGCbin = (numGC_{binx}/numNT_{binx})/(numGC_{allbins}/totNT_{allbins})$$

Normalized GC content highlights local GC content compared to the GC content of the entire organism.

The graphs were plotted for the folded bacterial prefixes of various lengths.

4.4 Results and Observation

4.4.1 Prefix 90

This experiment was performed on the complete genomes of 195 bacteria. As observed from the Figure 1, mRNA prefixes of length 90 are skewed toward the stable side of the energy quantile scale. Intuitively, one might expect that GC content is strongly correlated with low MFE. However, no such consistent relationship exists. For example, quantile 43 has a relatively small number of coding sequences in it but it shows higher GC content.

4.4.2 Prefix 120

The experiment was done for 194 bacteria. This has a single highly stable (low MFE) mode. As above, GC content is not correlated with stability.

4.4.3 Prefix 150

The experiment was done for 120 bacteria. The graph shows that this is also unimodal highly stable.

This experiment is to analyze the prefix length 120 of 165 bacterias. The following figure shows the stability of all the bacteria involved.

This experiment is to analyze the prefix length 150 of 169 bacterias. The following figure shows the stability of all the bacterias involved.

4.4.4 Suffix Length 120

This experiment is to analyze the suffix length 120 of 78 bacterias. The following figure shows the stability of all the bacterias involved.

4.4.5 Suffix Length 90

This experiment is to analyze the suffix length 90 of 108 bacterias. The following figure shows the stability of all the bacterias involved.

4.5 Stability Analysis Based on Windows For Bacteria

There was a variation in normalized MFE along the length of sequence. Natural sequence window has less structure than the shufflet window at the beginning and end of the sequence but more structure in the middle. This depends on window size. Normalized MFE in windows covering the first 1/5 of the sequence tends to be positive. This is already evident with window size 100 and is apparent at all windows up to length 950. Normalized MFE at the end of sequence also tends to be positive. This is not visible with window size 100 but with 350 and most marked with 550 and then disappears in 700. Normalized MFE in the middle region of sequence tends to be negative. This is dependent on window size and is clearly evident at 550 and peaks at 750. Window steps experiments were run on a highly stable sequence from a bimodal bacteria of length 2000 with window sizes 50-950. As window size increases the MFE of the window tends to MFE of the sequence.

4.6 Results and Observation

As can be seen, the prefix and suffix follow the same pattern; they tend to have more structures as the size of the window increases.

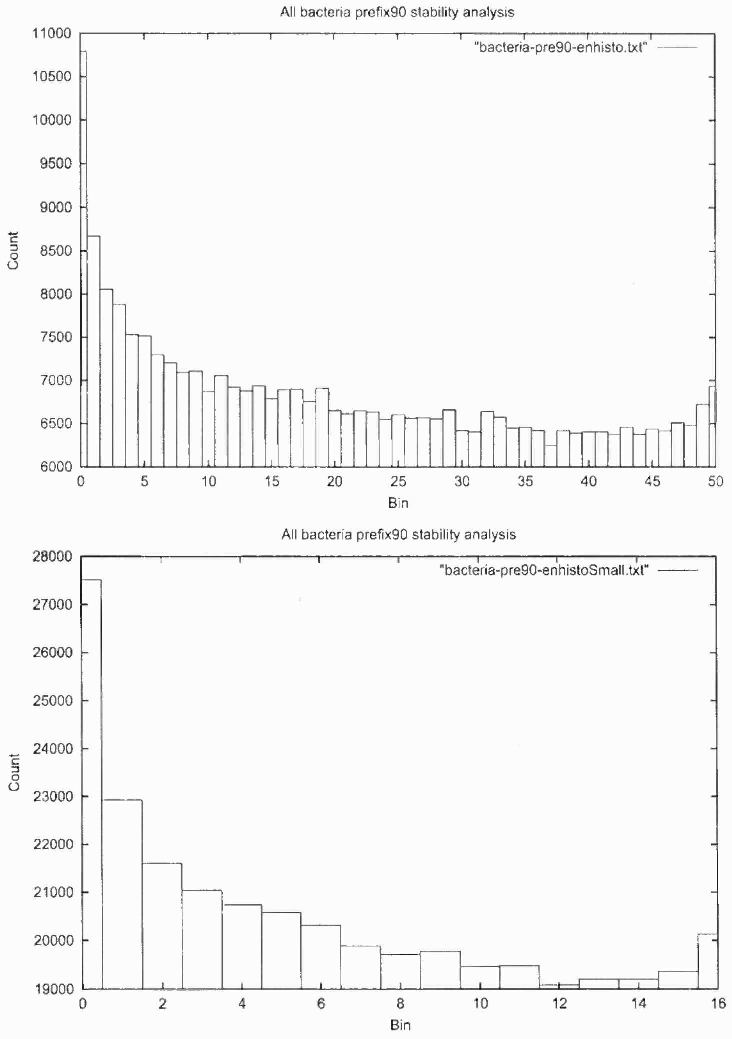


Figure 4.1 Stability analysis of bacteria of prefix length 90.

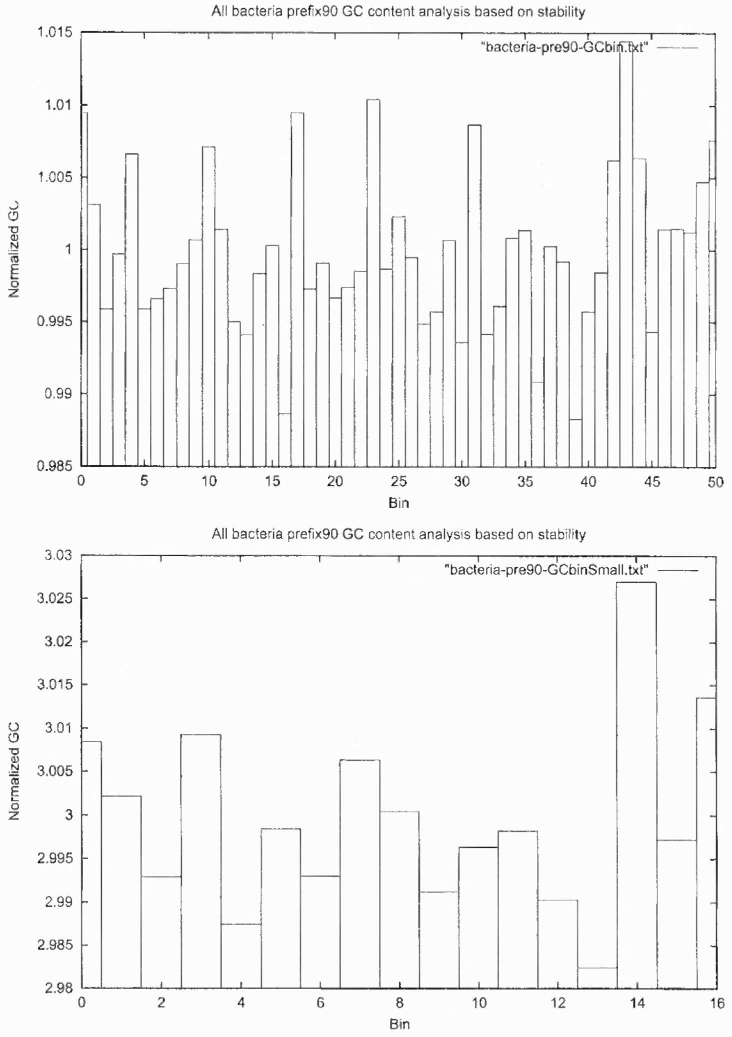


Figure 4.2 GC content analysis of bacterial prefixes of length 90.

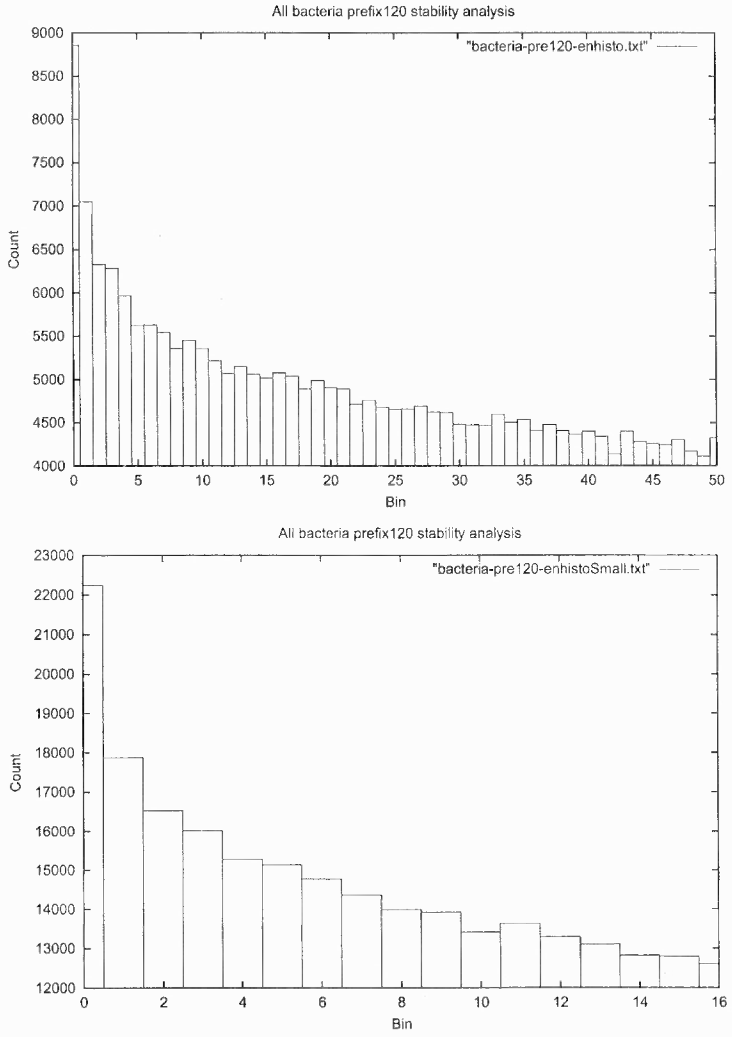


Figure 4.3 Stability analysis of bacterial prefixes of length 120.

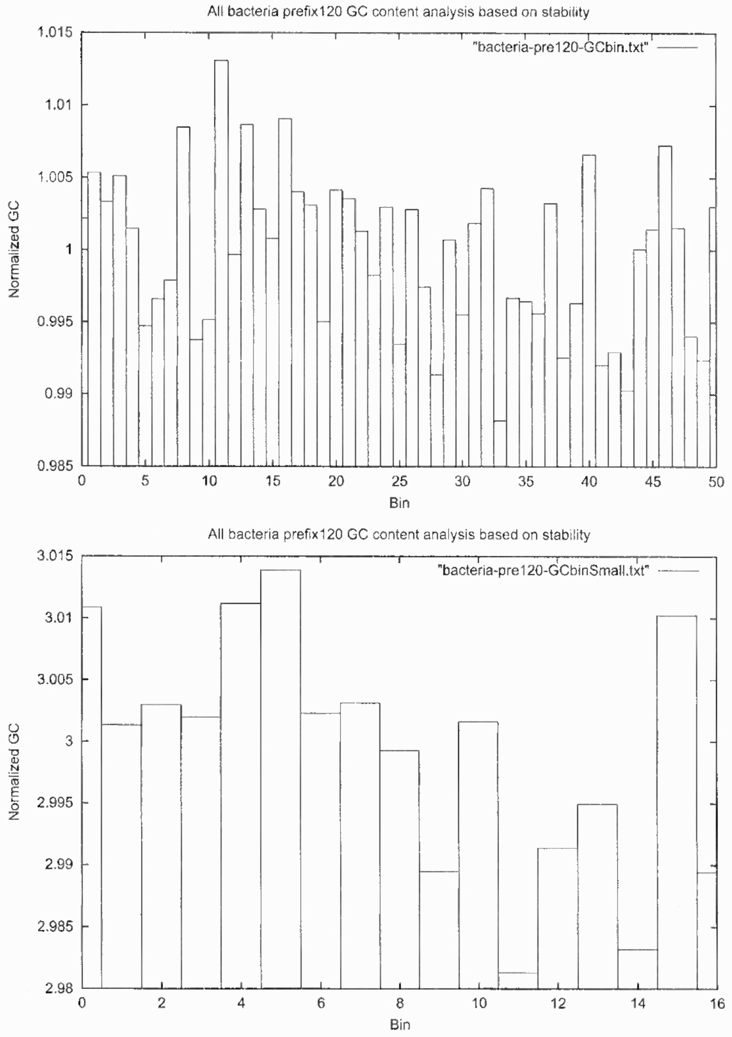


Figure 4.4 GC content analysis of bacterial prefixes of length 120.

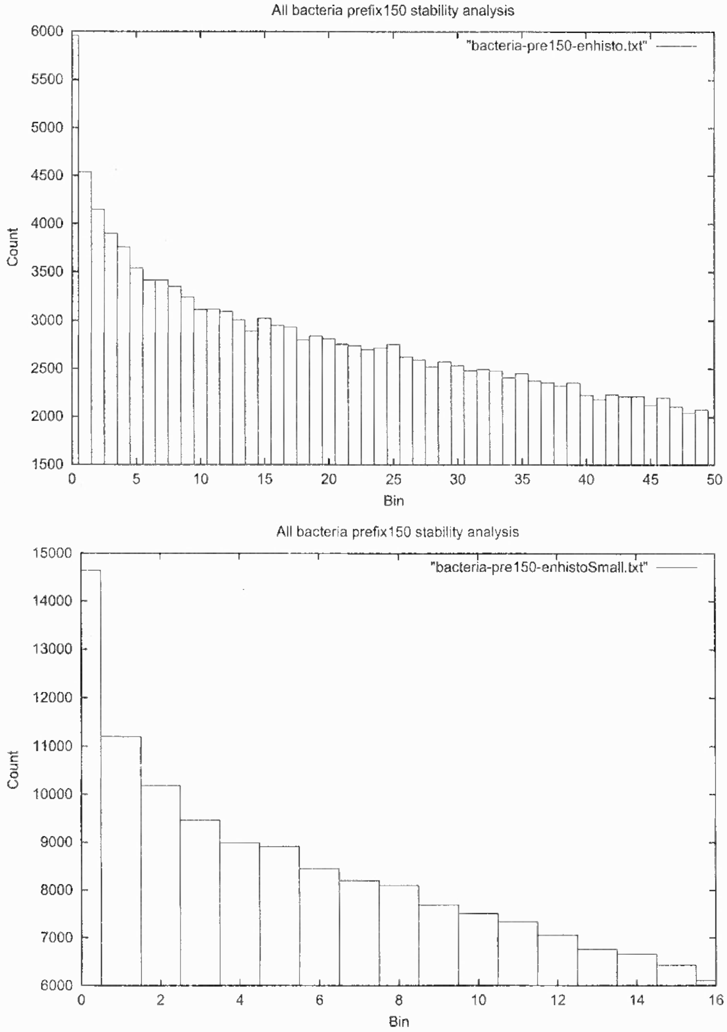


Figure 4.5 Stability analysis of bacterial prefixes of length 150.

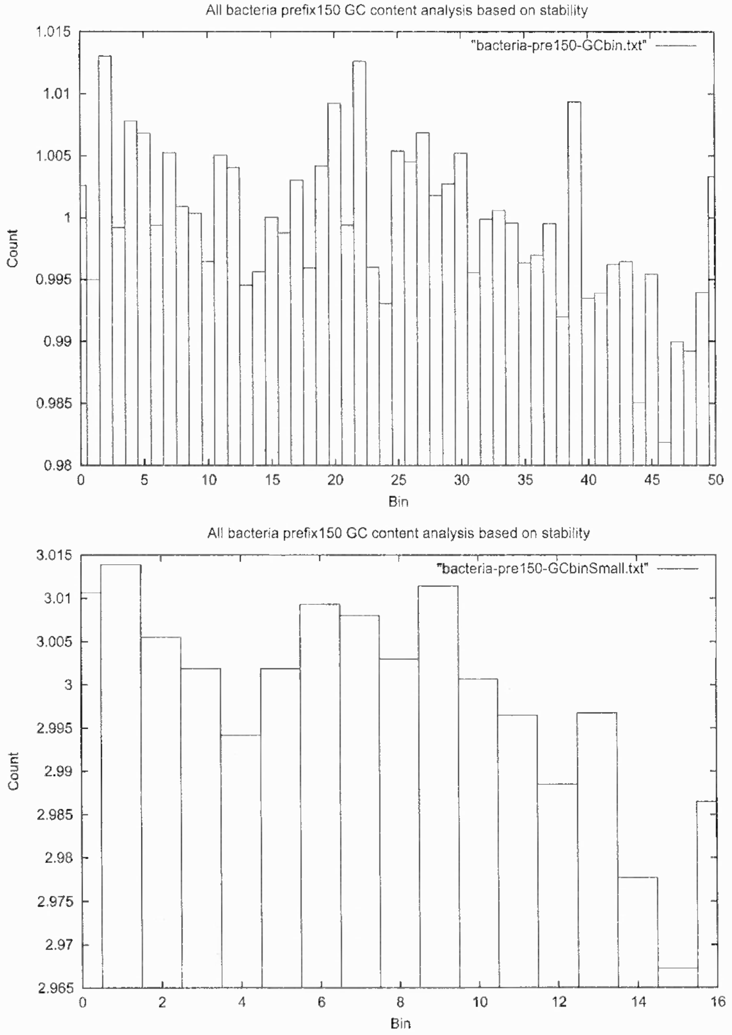


Figure 4.6 GC content analysis of bacterial prefixes of length 150.

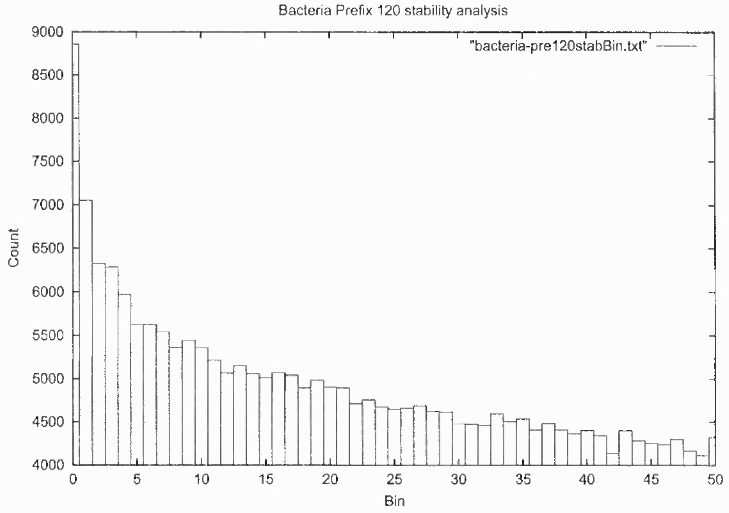


Figure 4.7 Stability analysis of bacterias of prefix length 120.

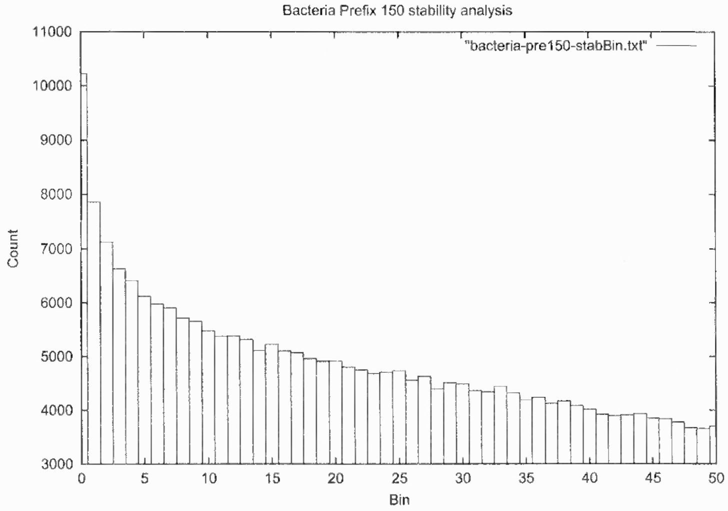


Figure 4.8 Stability analysis of bacterias of prefix length 150.

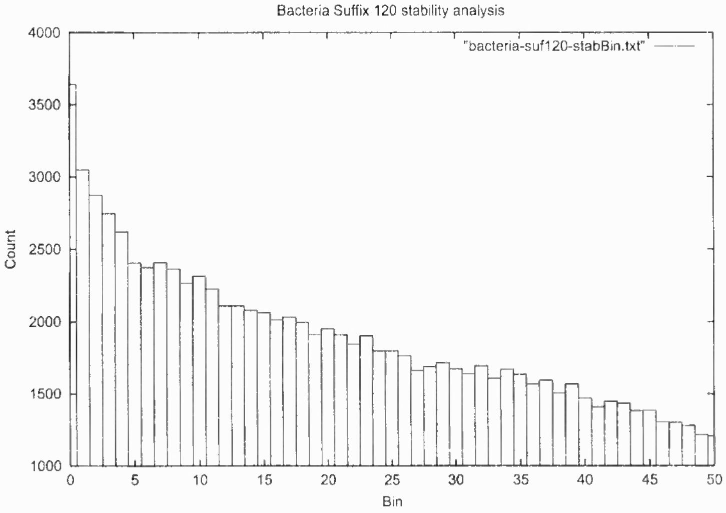


Figure 4.9 Stability analysis of bacterias of suffix length 120.

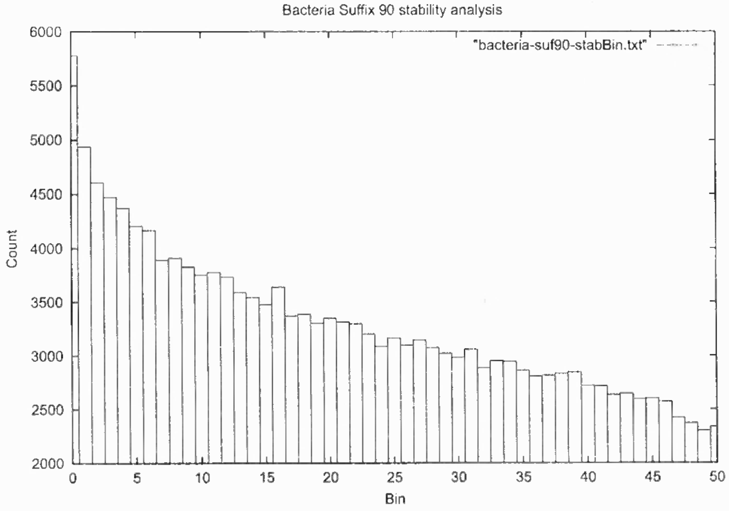


Figure 4.10 Stability analysis of bacterias of suffix length 120.

CHAPTER 5

DISTINCTIVE ENERGY AND SECONDARY STRUCTURE SIGNATURES OF SUBVIRAL RNAS

5.1 Introduction

5.1.1 Background and Related work

Viroids are small, single-stranded, unencapsidated, covalently closed-circular RNAs. They are infectious agents that affect many plants. They are much smaller and simpler than viruses and lack the protein cover that is typical for viruses. Viroids use higher plants (such as potatoes, tomatoes and cucumbers) to reproduce, inserting themselves into the nucleus of a plant cell to be replicated there. Viroids are usually transmitted by seed or pollen. Infected plants can show distorted growth. The first viroid to be identified was the Potato spindle tuber viroid in the early 1970's [26].

Viroids and viroidlike satellite RNAs are of importance for the following two reasons as stated by [27]:

1. Viroid RNAs are the smallest and simplest replicons known; elucidation of their mechanisms of replication and pathogenesis is therefore of considerable significance. [28]

2 Viroid RNAs are of potential evolutionary importance, as they may represent relics of precellular evolution in an RNA world [29]. The compelling evidence of RNA world is the recognition that RNA is the only known macromolecule that can function both as genotype and phenotype - thus permitting Darwinian evolution to occur at the molecular level in the absence of DNA or functional proteins [30].

Viroids are the etiologic agents of a number of diseases affecting economically important herbaceous and ligneous plants including potato, tomato, cucumber, hop, coconut, grapevine, several subtropical and temperate fruit trees (avocado, peach, apple, pear, citrus, and plum), and some ornamentals (chrysanthemum and coleus). Coconut cadang-cadang viroid (CCCVd) and Coconut tinangaja viroid (CTiVd) infect monocotyledons, whereas the others infect dicotyledons. Some viroids, among which the most instructive example is Hop stunt viroid (HSVd), have wide host ranges but others, exemplified by those forming the family Avsunviroidae, are mainly restricted to their natural hosts. A single nucleotide substitution converts PSTVd from noninfectious to infectious for Nicotiana tabacum.

Although most viroids are transmitted mechanically and some through seed or pollen, with only Tomato planta macho viroid (TPMVd) known to be aphid-transmissible under specific ecological conditions, the most efficient transmission route for viroids is vegetative propagation of infected material. This explains why certain grapevine and, specifically, citrus cultivars propagated on infected cultivars or rootstocks contain complex mixtures of different viroids.

5.1.2 RNA Secondary Structure

There are five types of secondary structural elements: hairpin loops, internal loops, multibranched loops, bulges and stacks or stem loops.

Hairpin loops: The unpaired region formed when an RNA folds back upon itself to form a helix. It occurs at the end of a helix when the sugar phosphate backbone reveals a hairpinlike structure. Comparisons of small subunit ribosomal RNA structures reveal an uneven distribution of hairpin loop sizes: four base loops are the most common. Larger hairpin loops can pair into complex structures involving non-Watson-Crick interactions.

Hairpin loops are important for mRNA stability, RNA tertiary interactions, and protein binding sites.

Internal loops: Two or more opposing unpaired bases between two helical segments; internal loops can be symmetric (the same number of unpaired bases on each side of the loop) or asymmetric (a different number of unpaired bases on each side of the loop). Two base internal loops are often called mismatches. Common small internal loops have increased stability due to base tacking and non-Watson-Crick hydrogen bonding. Internal loops are important sites of RNA-protein interaction in 5S rRNA and proposed RNA-RNA tertiary and quaternary interactions in group I introns.

Multiloop: Region in which three or more helices join to form a closed loop. The crystal structure of tRNA has a four-helix multibranched loop stabilized by helix-helix stacking as well as significant non-Watson-Crick secondary and tertiary interactions. These interaction probably stabilize other multiloops.

Bulge loop: Regions in which there are unpaired bases on only one side of a helix. They can bend RNA backbones. Bulges are important recognition sites for many regulatory and structural proteins. For this study the right and left bulges are taken separately.

Stack: Also called stem loops, they contribute most to the stability of the RNA secondary structure through hydrogen bonds and base stacking. The base stacking is the interaction between the pi orbitals of the bases' aromatic rings. The Watson-Crick pairs G-C and A-U, as well as some of the mismatches, such as G-U, stabilize the stacks. Base stacking is an important stabilizing effect since a single base stacking on the 3' side of a helix can add as much stability to the structure as a base pair.

The other types of the RNA secondary structural elements include pseudoknots, which are too unstable to be considered here. Pseudoknots are structures that result when any single-stranded loop forms a helix with another single-stranded region.

5.2 Dataset

The data for these experiments were obtained from Subviral RNA Database [31]. The viroids analyzed are listed below, along with the length of the viroid sequence and the number of variants. Viroids, like viruses, propagate in their hosts as populations of closely related sequence variants (quasi-species), although one or more may predominate in the population.

A large number of viroid sequences are now known and their classification is good. The major classification criterion is the type of Central Conserved Region (CCR). Based on this viroids can be classified into two families: Pospiviroidae (with CCR and without hammerhead self-cleavage) and Avsunviroidae (without CCR and with hammerhead self-cleavage). The inclusion of existence or lack of hammerhead structures in the primary classification criteria is due to the following: (i) it is connected with replication, and the characteristics of replication are one of the criteria recommended for virus – and, by extension, for viroid – classification and (ii) it leads to the same grouping as the presence or absence of CCR (III). It establishes an evolutionary link between viroids and viroid-like satellite RNAs, which in all cases contain hammerhead structures in one or in both polarity strands. The subfamily taxon was introduced because the members of genera Pospiviroid, Hostuviroid and Cocadviroid share an identical subset of nucleotides within their CCRs which are more closely related with each other than any is with CCRs of apsca and coleviroids.

Table 5.1 Viroid Classification and Details

Family	Subfamily	Genus	Species
Pospiviroidae	Pospiviroinae	Pospoviroids	PSTVd (potato spindle tuber)
			TCDVd (Tomato chlorotic dwarf)
			MPVd (Mexican papita)
			TPMVd (tomato planta macho)
			CEVd (citrus exocortis)
			CSVd (chrysanthemum stunt)
			TASVd (tomato apical stunt)
			IrVd-1 (iresine 1)
			CLVd (columnea latent)
		Hostuviroid	HSVd (hop stunt)
		Cocadviroid	CCCVd (coconut cadang-cadang)
			CTiVd (coconut tinangaja)
			HLVd (hop latent)
			CVd-IV (citrus IV)
	Apscaviroinae	Apscaviroids	ASSVd (apple scar skin)
			CVd-III (citrus III)
			ADFVd (apple dimple fruit)
			GYSVd-1 (grapevine yellow speckle 1)
			GYSVd-2 (grapevine yellow speckle 2)
			CBLVd (citrus bent leaf)
			PBCVd (pear blister canker)
			AGVd (Australian grapevine)
	Coleviroinae	Coleviroids	CbVd-1 (cleus blumei 1)
			CbVd-2 (cleus blumei 2)
			CbVd-3 (cleus blumei 3)
Avsunviroidae		Avsunviroid	ASBVd (avocado sunblotch)
		Pelamoviroid	PLMVd (peach latent mosaic)
			CChMVd (chrysanthemum chlorotic mottle)
		Elaviroid	ELVd (Eggplant latent)

Table 5.2 Number of Variants and Nucleotides in Viroids

Species	Variants	Nucleotides
PSTVd	134	359
TCDVd	4	360
MPVd	10	360
TPMVd	2	360
CEVd	123	371
CSVd	31	356
TASVd	7	360
IrVd-1	4	370
CLVd	26	370
HSVd	11	256
CCCVd	12	246
CTiVd	3	254
HLVd	11	256
CVd-IV	8	284
ASSVd	10	329
CVd-III	69	297
ADFVd	11	306
GYSVd-1	68	367
GYSVd-2	11	363
CBLVd	26	318
PBCVd	24	315
AGVd	10	369
CbVd-1	10	248
CbVd-2	3	301
CbVd-3	4	361
ASBVd	88	247
PLMVd	189	337
CChMVd	25	399
ELVd	10	335

Table 5.2 shows that there are many variants. The peach latent mosaic viroid has the maximum number of variants – 189. The RNA sequence size varies from 254 nucleotides to 399 nucleotides.

Most of the nearly 30 viroid species known belong to the family Pospiviroidae. They adopt in vitro a rod-like or quasi-rod-like secondary structure of minimal free energy with five structural-functional domains. The CCR, within the C domain, is formed by two stretches of conserved nucleotides, in which those of the upper strand are flanked by an inverted repeat of the nature of the CCR, and on the presence or absence of a terminal conserved region (TCR) and a terminal conserved hairpin (TCH), members of this family are allocated to five genera. The other four viroids, Avocado sunblotch viroid (ASBVd), Peach latent mosaic viroid (PLMVd), Chrysanthemum chlorotic mottle viroid (CChMVd), and Eggplant latent viroid (ELVd), do not have the conserved CCR, TCR, and TCH motifs but, remarkably, both their polarity strands self-cleave through hammerhead ribozymes; they form the second family, Avsunviroidae, whose type species is ASBVd (formal inclusion of ELVd in this family is pending ICTV approval). Apart from the core nucleotides conserved in their hammerhead structures, no extensive sequence similarities exist between them, but PLMVd and CChMVd are grouped in one genus because of their branched secondary structure, which is stabilized by a pseudoknot and their insolubility in 2 M LiCl. ASBVd, the only viroid with a high A + U content (62 percent), forms a monospecific genus, and ELVd, whose properties fall between those of the members of the other two genera, has been proposed to constitute its own genus. This classification scheme is further supported by phylogenetic reconstructions with entire viroid sequences and by the different subcellular replication (and accumulation) sites of the type members of both families, with available data indicating that in this respect other viroids behave like their corresponding type species. Within each genus, the criteria to demarcate viroid species are an arbitrary level of below 90sequence similarity and distinct biological properties. Viroids, like viruses, propagate in their hosts as populations of closely related sequence variants (quasi-species), although one or more may predominate in the population. Heat stress may significantly alter the structure of viroid quasi-species. Some viroid variants with minor changes affecting certain regions are directly related to specific diseases or to dramatic alterations in symptom severity.

5.3 Methodology

The natural and the random sequences were folded to predict the minimum free energy of secondary structure. The ViennaRNA package that implements Zuker's RNA prediction algorithm was used for this. For each natural mRNA sequence, 1000 random sequences were generated using the shufflet method. Program was written to analyze the structure of RNA sequences. The program finds the position of opening and closing parenthesis, index of base pair at opening and closing parenthesis, type of structure, size of the structure. The common secondary structure motifs include hairpin loops, stems and bulges.

5.4 Skewing of Viroids Structure

5.4.1 Stability of Viroids

The Figure 5.4.1 shows that the viroids are highly stable. As can be seen most of the wildtype sequences have more energy than the synthetic sequences. There are 10 sequences for which Z = -10. Generally, a Z score of -2 is considered highly stable. For around 73 percent of the sequences, in this experiment, $Z \leq -2$.

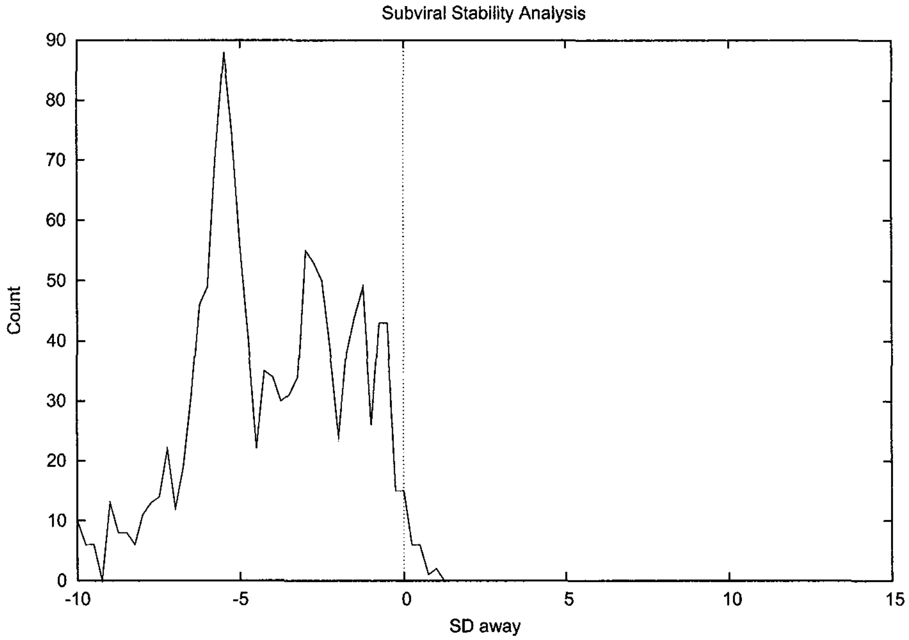


Figure 5.1 Stability of viroids using shufflet randomization method based on Z score.

Table 5.3 Maximum Size of Secondary Structure Motifs

	Hairpin	Internal loop	Multibranch	Left bulge	Right bulge	Stack
seq	19	20	31	12	17	18
shf	49	30	59	27	29	23

5.4.2 Secondary Structure Analysis

Maximum Size of Secondary Structure Motifs The following table shows the maximum size of the secondary structure motifs for all the viroids. As can be seen the synthetic sequences have motifs of larger size than the wildtype sequences.

Number of Structures in Wildtype And Synthetic Sequences The number of structures in each of the secondary structures predicted was calculated for each of the viroids.

Table 5.4 Number of Hairpin Loops in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	11	60.734
AFCVd	58	200.269
AGVd	12	64.884
ASBVd	113	372.788
ASSVd	11	59.983
CBLVd	48	145.502
CCCVd	13	61.84
CChMVd	209	175.756
CEVd	218	815.104
CLVd	48	170.001
CSVd	69	190.244
CTiVd	3	14.569
CVd-III	140	378.019
CVd-IV	8	42.882
CVd-LSS	9	46.275
CVd-OS	16	42.343
CbVd	1	5.332
CbVd-1	33	45.752
CbVd-2	7	16.223
CbVd-3	4	25.378
ELVd	40	58.688
GYSVd-1	182	428.874
GYSVd-2	64	70.009
HLVd	15	52.992
HSVdalm	325	1432.607
IrVd	6	26.689
JCVd1	8	12.11
MPVd	12	63.225
PBCVd	86	131.061
PLMVd	1172	1156.642
PSTVd	138	863.332
TASVd	10	43.543
TCDVd	7	25.232
TPMVd	2	12.754

Hairpin: The number of hairpin loops in the natural viroid sequences and the shufflet sequences are shown in the Table 5.4.2. It shows that the randomized sequences have more hairpin loops than the natural sequences.

Internal Loop: The number of internal loops in the natural viriod sequences and the shufflet sequences are shown in the Table 5.4.2. It shows that the natural sequences have more internal loops than the randomized sequnces.

Multi loop: The number of multi loops in the natural viroid sequences and the shufflet sequences are shown in Table 5.4.2. It shows that the randomized sequences have more multi loops than the natural sequences.

Left Bulge: The number of left bulges in the natural viroid sequences and the shufflet sequences are shown in Table 5.4.2. The randomized sequnces in some viroids have more left bulges than the natural sequences but the majority of the viroids are other way round.

Right Bulge: The number of right bulges in the natural viroid sequences and the shufflet sequences are shown in Table 5.4.2. The randomized sequnces in some viroids have more right bulges than the natural sequences but the majority of the viroids are other way round.

Stack: The number of stacks in the natural viroid sequences and the shufflet sequnces are shown in Table 5.4.2. The natural sequences have more stacks than the randomized sequences.

Wildtype and Synthetic Sequence Secondary Structure Comparison The size of the various substructures were also analyzed.

Table 5.5 Number of Internal Loops in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	182	105.162
AFCVd	514	341.927
AGVd	187	109.579
ASBVd	997	609.854
ASSVd	211	99.327
CBLVd	477	264.628
CCCVd	154	94.533
CChMVd	193	291.419
CEVd	2524	1314.727
CLVd	562	278.918
CSVd	590	300.859
CTiVd	40	21.716
CVd-III	1080	613.259
CVd-IV	81	60.854
CVd-LSS	126	80.925
CVd-CS	120	71.978
CbVd	13	8.235
CbVd-1	88	74.643
CbVd-2	35	25.844
CbVd-3	67	42.225
ELVd	143	96.28
GYSVd-1	1173	746.155
GYSVd-2	123	116.753
HLVd	134	84.171
HSVdalm	4369	2251.89
IrVd	80	41.597
JCVd1	30	20.716
MPVd	208	103.943
PBCVd	281	233.481
PLMVd	2066	1887.84
PSTVd	2661	1326.299
TASVd	122	71.739
TCDVd	76	39.875
TPMVd	44	20.428

Table 5.6 Number of Multi Loops in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	0	34.944
AFCVd	26	123.232
AGVd	1	40.07
ASBVd	25	187.927
ASSVd	1	36.228
CBLVd	21	83.589
CCCVd	1	34.696
CChMVd	97	111.743
CEVd	59	505.764
CLVd	20	105.667
CSVd	19	114.952
CTiVd	0	7.963
CVd-III	65	218.966
CVd-IV	0	24.757
CVd-LSS	1	27.238
CVd-OS	8	25.568
CbVd	0	3.049
CbVd-1	22	23.891
CbVd-2	2	9.383
CbVd-3	0	15.626
ELVd	20	34.47
GYSVd-1	95	261.392
GYSVd-2	32	42.826
HLVd	3	28.904
HSVdalm	53	822.532
IrVd	1	16.781
JCVd1	2	7.263
MPVd	2	38.336
PBCVd	61	74.227
PLMVd	452	704.436
PSTVd	2	532.746
TASVd	3	26.171
TCDVd	2	15.505
TPMVd	0	7.832

Table 5.7 Number of Left Bulge Loops in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	90	29.522
AFCVd	213	100.319
AGVd	66	32.21
ASBVd	323	135.885
ASSVd	48	29.584
CBLVd	128	81.015
CCCVd	69	27.797
CChMVd	146	79.562
CEVd	645	384.715
CLVd	145	81.366
CSVd	155	77.326
CTiVd	12	6.3
CVd-III	398	162.771
CVd-IV	64	16.984
CVd-LSS	69	24.697
CVd-OS	25	22.15
CbVd	7	2.27
CbVd-1	35	21.006
CbVd-2	15	7.636
CbVd-3	27	12.334
ELVd	10	24.777
GYSVd-1	349	225.445
GYSVd-2	26	34.184
HLVd	42	24.577
HSVdalm	1568	659.821
IrVd	30	12.434
JCVd1	4	6.488
MPVd	39	29.042
PBCVd	33	71.214
PLMVd	318	511.605
PSTVd	581	369.992
TASVd	50	19.595
TCDVd	14	10.997
TPMVd	8	5.332

Table 5.8 Number of Right Bulge Loops in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	92	29.964
AFCVd	305	100.727
AGVd	66	32.231
ASBVd	109	136.208
ASSVd	74	29.854
CBLVd	152	80.986
CCCVd	63	27.691
CChMVd	86	79.306
CEVd	575	385.27
CLVd	116	80.964
CSVd	57	77.887
CTiVd	17	6.404
CVd-III	317	161.223
CVd-IV	40	16.918
CVd-LSS	64	24.714
CVd-OS	36	21.682
CbVd	3	2.308
CbVd-1	26	21.021
CbVd-2	13	7.778
CbVd-3	27	12.218
ELVd	19	24.638
GYSVd-1	604	225.834
GYSVd-2	90	34.544
HLVd	73	24.357
HSVdalm	995	659.769
IrVd	26	12.282
JCVd1	12	6.395
MPVd	61	28.748
PBCVd	197	71.093
PLMVd	412	510.62
PSTVd	449	369.259
TASVd	52	19.69
TCDVd	21	10.811
TPMVd	14	5.409

Table 5.9 Number of Stack in Natural And Shufflet Sequences of Viroids

Viroid	Number in sequences	Number in Shufflets
ADFVD	325	278.419
AFCVd	1283	1135.57
AGVd	328	291.142
ASBVd	3969	3859.565
ASSVd	280	267.21
CBLVd	888	834.648
CCCVd	316	275.223
CChMVd	882	901.294
CEVd	8321	8054.471
CLVd	1041	900.709
CSVd	1058	1053.688
CTiVd	61	55.245
CVd-III	3216	3053.02
CVd-IV	193	172.628
CVd-LSS	225	207.459
CVd-OS	191	182.937
CbVd	23	19.839
CbVd-1	217	204.877
CbVd-2	66	64.381
CbVd-3	120	104.748
ELVd	239	255.782
GYSVd-1	3396	3368.612
GYSVd-2	316	316.963
HLVd	253	237.174
HSVdalm	17461	16832.304
IrVd	124	106.706
JCVd1	48	49.292
MPVd	304	278.007
PBCVd	790	735.664
PLMVd	14178	12604.205
PSTVd	9687	8695.657
TASVd	209	184.408
TCDVd	104	99.912
TPMVd	52	49.067

Table 5.10 Hairpin Size in Wildtype and Shufflet Sequences Based on Z Score

Size of Hairpin	Number in sequences	Number in Shufflets
0	0.000	0.000
1	0.000	0.000
2	0.000	0.000
3	294.000	537.482
4	1468.000	2652.069
5	575.000	1450.491
6	232.000	1003.258
7	181.000	519.584
8	252.000	525.942
9	65.000	199.175
10	20.000	135.130
11	3.000	90.699
12	2.000	59.551
13	1.000	42.680
14	0.000	28.317
15	3.000	19.756
16	0.000	13.114
17	0.000	10.050
18	1.000	6.811
19	1.000	5.089
20	0.000	3.534
21	0.000	2.593
22	0.000	1.798
23	0.000	1.338
24	0.000	0.897
25	0.000	0.688
26	0.000	0.456
27	0.000	0.313
28	0.000	0.240
29	0.000	0.158
30	0.000	0.099
31	0.000	0.090
32	0.000	0.067
33	0.000	0.044
34	0.000	0.030
35	0.000	0.025
36	0.000	0.025
37	0.000	0.012
38	0.000	0.007
39	0.000	0.006
40	0.000	0.004

Table 5.11 Internal Size in Wildtype and Shufflet Sequences Based on Z Score

Size of Internal loop	Number in sequences	Number in Shufflets
0	0.000	0.000
1	0.000	0.000
2	5627.000	4289.461
3	2420.000	1163.296
4	6676.000	3052.781
5	1421.000	1046.406
6	1341.000	810.222
7	1084.000	439.320
8	303.000	341.213
9	264.000	209.213
10	319.000	162.214
11	228.000	97.368
12	13.000	76.963
13	11.000	49.595
14	13.000	39.044
15	4.000	21.694
16	9.000	17.089
17	2.000	10.651
18	0.000	8.437
19	0.000	4.879
20	16.000	4.182
21	0.000	2.220
22	0.000	2.000
23	0.000	1.025
24	0.000	0.882
25	0.000	0.482
26	0.000	0.478
27	0.000	0.228
28	0.000	0.213
29	0.000	0.099
30	0.000	0.124

Table 5.12 Multibranch Size in Wildtype and Shufflet Sequences Using Z Score

Size of Multibranch	Number in sequences	Number in Shufflets
0	0.000	0.000
1	8.000	19.305
2	45.000	83.106
3	17.000	171.280
4	46.000	252.083
5	146.000	309.912
6	179.000	346.486
7	65.000	359.070
8	104.000	355.543
9	97.000	339.336
10	79.000	314.934
11	52.000	283.594
12	46.000	251.945
13	30.000	220.153
14	27.000	188.775
15	21.000	160.583
16	29.000	134.318
17	18.000	111.404
18	12.000	91.385
19	38.000	74.279
20	11.000	60.235
21	8.000	48.006
22	5.000	38.396
23	2.000	30.380
24	1.000	23.790
25	1.000	18.795
26	0.000	14.433
27	2.000	11.413
28	0.000	8.661
29	1.000	6.602
30	0.000	5.148
31	6.000	3.999
32	0.000	2.854
33	0.000	2.190
34	0.000	1.633
35	0.000	1.183
36	0.000	0.905
37	0.000	0.689
38	0.000	0.471
39	0.000	0.360
40	0.000	0.282

Table 5.13 Left Bulge Size in Wildtype and Shufflet Sequences Using Z Score

Size of Left bulge	Number in sequences	Number in Shufflets
0	0.000	0.000
1	4284.000	2263.993
2	582.000	418.619
3	330.000	260.244
4	526.000	154.495
5	15.000	89.050
6	3.000	49.542
7	1.000	34.954
8	0.000	23.965
9	2.000	16.318
10	8.000	11.079
11	0.000	7.280
12	1.000	4.587
13	0.000	2.899
14	0.000	1.622
15	0.000	0.897
16	0.000	0.553
17	0.000	0.366
18	0.000	0.186
19	0.000	0.140
20	0.000	0.067
21	0.000	0.033
22	0.000	0.029
23	0.000	0.014
24	0.000	0.006
25	0.000	0.009
26	0.000	0.002
27	0.000	0.003

Table 5.14 Right Bulge Size in Wildtype and Shufflet Sequences Using Z Score

Size of Right bulge	Number in sequences	Number in Shufflets
0	0.000	0.000
1	3910.000	2239.333
2	921.000	420.089
3	258.000	260.502
4	77.000	154.813
5	56.000	88.887
6	9.000	49.922
7	7.000	35.002
8	0.000	23.931
9	0.000	16.329
10	0.000	11.105
11	0.000	7.384
12	0.000	4.489
13	1.000	2.889
14	0.000	1.758
15	0.000	0.964
16	0.000	0.560
17	3.000	0.367
18	0.000	0.226
19	0.000	0.136
20	0.000	0.083
21	0.000	0.039
22	0.000	0.030
23	0.000	0.015
24	0.000	0.011
25	0.000	0.010
26	0.000	0.003
27	0.000	0.000
28	0.000	0.001
29	0.000	0.001

Table 5.15 Stack Size in Wildtype and Shufflet Sequences Using Z Score

Size of Stack	Number in sequences	Number in Shufflets
0	0.000	0.000
1	23060.000	23573.686
2	3372.000	2692.696
3	3154.000	2514.419
4	2850.000	1949.903
5	1175.000	1325.633
6	647.000	816.330
7	440.000	466.598
8	354.000	250.054
9	64.000	126.939
10	11.000	62.291
11	29.000	28.959
12	15.000	13.248
13	1.000	5.845
14	2.000	2.474
15	1.000	1.159
16	0.000	0.451
17	0.000	0.195
18	0.000	0.071
19	0.000	0.035
20	0.000	0.010
21	0.000	0.005
22	0.000	0.002
23	0.000	0.003

From the Tables 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, it can be observed that the size of the structures in the wildtype sequence is more highly conserved than in the shufflet sequences.

Most of the natural sequences have only one hairpin. Overall the shufflet sequences have more hairpins than the wildtype, which shows that the wildtype is more stable. The wildtype has more internal loops than the shufflet. The wild type has more left bulges than the shufflet. The wild type has more right bulges than the shufflet. Most of the wildtype sequences do not have multiple loops. Also the shufflet sequences have more multiple loops. The number of stacks in wildtype is slightly higher than the shufflet.

CHAPTER 6

CONCLUSION

The dissertation found evidence for natural selection on mRNA secondary structures and correlation between mRNA stability and COG functional classes.

Is there natural selection on the stability of RNA secondary structure, across various types of organisms?

The mRNA sequences folds back on itself and complementary bases form pairs resulting in mRNA secondary structure. The stability of a secondary structure is quantied as the amount of free energy released or used by forming base pairs. The more negative the free energy of a structure, the more likely is formation of that structure, because more stored energy is released. The goal is to find if there is a selection for formation of RNA secondary structure among various types of organisms. In order to find if a structure is stable or not, randomized sequences are generated based on the natural sequences. The free energy of the structures formed by the generated randomized sequnces are compared with the free energy of the structures formed by the corresponding natural sequence. Although many groups have worked on similar problem, the method of randomization has always been the point of contention. So, two completely different methods of randomization was used. One maintains the nucleotide and dinucletide composition, which plays a major role in the RNA secondary structure prediction. This was achieved by using the implementation of the algorithm which is based on euler algorithm. The other method maintains the end protein product. This is achieved by generating synonomous sequences by selecting synonomous codons based on the uniform probability. Also the data was analyzed using two different methods. The Z score standardizes or normalizes the results for each gene, expressing the stability of the naturally occurring sequence in terms of how many standard deviations it is above or below the mean MFE of the corresponding synthetic sequences. The quantile analysis ranks the MFE of a naturally occurring sequence relative to the population of 50 ordered MFEs of synthetic sequences.

Experiments were run on both prokaryotes and eukaryotes. Complete genomes of eight bacteria and yeast were folded. Also 500 coding seqeuences were chosen randomly from all the bacterial sequences and fruitfly. In prokaryotes, natural selection has favored highly stable sequences. Only one organism showed bimodal tendency. No statistically significant result for natural selection was found in eukaryotes. Since two completely different methods of randomization was used. Many genes were found to be highly stable based on both the methods with high statistical significance.

Does the MFE of microbial mRNAs correlate with the function of the target protein?

Widespread deviation of the MFE of naturally occurring sequences from the average MFE of the corresponding randomized sequences, provide evidence of natural selection on the stability of the mRNA. Such selection is typically the consequence of improved functionality or adaptivity of some phenotypic characteristic. This raises several questions, why is there a selection, are there any characteristics that are affected by the selection. Experiments were performed to see if there is a correlation between the stability and functional classes. mRNAs of 25 broad functional classes of COGs were computationally predicted to determine whether the deviation of their MFE's from the expected values correlate to the functional class of the protein product.

A clear correlation was found between the MFEs of mRNA sequences and the COG functional group. Certain functional classes were found to select mRNA secondary structures that were highly stable. Similar analysis was done to find if there was a correlation between the MFEs of mRNA sequences and co-regulated genes, but none were found.

Is there evidence of natural selection on the nucleotide composition and/or secondary structure of the prefixes and suffixes of bacterial mRNAs?

Further exploring the reasons for nature's selection for highly stable secondary structures. Looking into how the RNA folds, the folding mechanism moves from 5' to 3' of the coding sequence. Also it does not fold the whole coding sequence but by windows. The beginning and end of the coding plays a major role in the final secondary structure because mRNA folds onto itself and hence there will be bonds that are formed between the nucleotides in the beginning and end of the coding sequences. So experiments were performed to find the stability of the prefixes and suffixes of the mRNA coding sequences. Prefixes and suffixes of various sizes ranging from 30 to 150 nucleotides and the respective randomized synonomous sequences were folded.

The prefixes are highly stable. Except the small prefixes that are bimodal but highly stable for a most of the sequences. No correlation between the GC Content and Structures. It was conjectured that the tendency to have higher GC content and the tendency to have more relative structure would be correlated and vice versa

mRNA prefixes and suffixes have a distinctive MFE signature. The naturally occurring prefixes display more structure, on average, than randomized sequences with identical nucleotide and dinucleotide content, suggesting that natural selection favors secondary structure in the prefix and suffix of mRNA.

Is there natural selection on the secondary structures and substructures of subviral RNAs?

Viroids are special organisms because, they do not code for proteins. The RNA secondary structure is what they have and hence all the functions are based on the RNA secondary structure. Analyzing the stability of the secondary structures and the substructures and finding if there is a selection for any particular structure will be helpful.

The stability of substructures were analyzed. It was found that viroids are highly stable. Structures were similar among viroids in the same family. Significant differences were found in the stability based on the size of the substructures.

The biological reasons and significance of these results are to be analyzed.

REFERENCES

[1] S. Shabalina and et.al., "A periodic pattern of mRNA secondary structure created by the genetic code," *Nucleic Acids Research*, 2006.

[2] D. Forsdyke and J. Mortimer, "Chargaffs legacy," *Gene*, vol. 261, 2000.

[3] D. Kandel, Y. Matias, T. Unger, and P. Winkler, "Shuffling biological sequences," *Discrete Applied Mathematics*, vol. 71, 1996.

[4] E. Coward, "Shufflet: shuffling sequences while conserving the k-let counts," *Bioinformatics Applications Note*, vol. 15, 1999.

[5] I. Hofacker, W. Fontana, P. Stadler, and L. Bonhoeffer, "Fast folding and comparison of RNA secondary structures," *Monatsheftefur Chemie*, vol. 125, 1994.

[6] M. Zuker, D. Mathews, and D. Turner, "Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide," *JCB*, vol. 5, 1989.

[7] R. Gesteland, T. Cech, and A. J.F., *The RNA World.* Cold Spring Harbor Laboratory Press, 1999.

[8] J. Smith and E. Szathmary, *The Origins of Life: From the Birth of Life to the Origin of Language.* Oxford University Press, 1999.

[9] D. Oxender, G. Zurawski, and C. Yanofsky, "Attenuation in the escherichia coli tryptophan operon: Role of RNA secondary structure involving the tryptophan codon region," in *Proc. Natl. Acad. Sci.*, 1979.

[10] C. Flamm, I. Hofacker, S. Maurer-Stroh, P. Stadler, and M. Zehl, "Design of multistable RNA molecules," *RNA*, vol. 7, 2000.

[11] W. Seffens and D. Digby, "mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences," *Nucleic Acids Research*, vol. 27, no. 158, 1999.

[12] C. Workman and A. Krogh, "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution," *Nucleic Acids Research*, vol. 27, 1999.

[13] L. Katz and C. Burge, "Widespread selection for local RNA secondary structure in coding regions of bacterial genes," 2003.

[14] B. Cohen and S. Skiena, "Designing RNA structures: Natural and artificial selection," in *RECOMB 02*, April 2002.

[15] ——, "Natural selection and algorithmic design of mRNA," *Journal of Computational Biology Volume 10, Numbers 3 4, Pp. 419 432*, 2003.

[16] R. Tatusov, E. Koonin, and D. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, 1997.

[17] R. Tatusov and et al, "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 41, 2003.

[18] S. Price and K. Nagai, "Secrets of RNA folding revealed," *Structure*, 1996.

[19] J. Doudna and E. Doherty, "Emerging themes in RNA folding," *Folding and Design*, 1997.

[20] I. Tinoco and C. Bustamante, "How RNA folds," *Journal of Molecular Biology*, 1999.

[21] C. Levinthal, "Are there pathways for protein folding?" *Journal of Chemical Physics*, 1968.

[22] P. Brion and E. Westhof, "Hierarchy and dynamics of RNA folding," *Annual Review of Biophysics and Biomolecular Structure*, 1997.

[23] M. Wu and I. Tinoco, "RNA folding causes secondary structure rearrangement," *Proceedings of the National Academy of Sciences*, 1998.

[24] R. Russell, X. Zhuang, H. Babcock, I. Millett, S. Doniach, S. Chu, and D. Herschlag, "Exploring the folding landscape of a structured RNA," *Proceedings of the National Academy of Sciences*, 2002.

[25] R. Russell, I. Millett, M. Tate, L. Kwok, B. Nakatani, S. Gruner, S. Mochrie, V. Pande, S. Doniach, D. Herschlag, and L. Pollack, "Rapid compaction during RNA folding," *Proceedings of the National Academy of Sciences*, 2002.

[26] T. Diener, "Potato spindle tuber virus iv. a replication, low molecular weight RNA," *Virology*, 1971.

[27] ——, "Subviral pathogens of plants: viroids and viroidlike satellite RNAs," *The FASEB Journal*, 1991.

[28] ——, *Viroids and Viroid Diseases*. Wiley and Sons, 1979.

[29] ——, "Ciroular RNAs: relics of precellular evolution?" *PNAS*, 1989.

[30] G. Jouce, "RNA evolution and the origin of life," *Nature*, 1989.

[31] L. Rocheleau and P. M., "The subviral RNA database: a toolbox for viroids, the hepatitis delta virus and satellite RNAs research," *BMC Microbiol.*, 2006.