ANALYSIS OF GENOMIC TRANSLATION USING A COMMUNICATIONS THEORY APPROACH

BY

MOHAMMAD AL BATAINEH

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
Illinois Institute of Technology

Approved _____
Adviser

Chicago, Illinois
July 2010

ACKNOWLEDGEMENT

Before all, all thanks and praises are due to Allah (SWT) whom without his mercy and constant support, this work would've never been possible. I would like to thank my advisor, Professor Guillermo E. Atkin, for his selfless guidance, constant support and for always being within reach when I needed help. I am indebted to him for introducing me to this interdisciplinary challenging research. His unbridled enthusiasm for this work was contagious and inspiring. I would also like to thank Professor Saniie, Professor Wernick and Professor Menhart for serving on my oral examination committee. Their careful reading and insightful feedback are greatly appreciated. Special thanks also go to Professor Wei Zhang from the Biology department at IIT for his cooperation and support.

Funding of this research was partially provided by the Pritzker Institute and Illinois Institute of Technology. Their support is gratefully acknowledged. I would specifically thank Professor Mohammad Shahidehpour for supporting this research in its early beginning.

I owe special thanks to Professor Ilias Belharouak from Argonne National Laboratories and to Mr. Alan ElShafei from Microsun Technologies. Their help and altruistic support in my last year of research were a true reason to finalize my work. Working with them was a priceless experience for me. I would not also forget to thank my lab mates, colleagues, and friends who made the long hours of study at IIT enjoyable.

I owe everything to my parents – my dearly missed late father Fayez Al Bataineh and my beloved mother Fawzia Ghunaimat. In addition to their endless support, and unwavering love, they have always been a source of inspiration and encouragement. I

cannot find the words to express my appreciation to them. Although it is insignificant in comparison to what they have given me, I dedicate this dissertation to both of them.

My sincere gratitude goes to my dearest sisters (Rana, Sahar, Ebtesam, Rania, my twin Asma, Somaia, Ekhlass, and Rowaida) for their prayers, love, and support over the years.

Finally, I will be forever beholden to the one love of my life, my wife Enas, whom without her constant encouragement, selfless sacrifices, quiet patience, and insightful criticism; this work would've never been accomplished. Besides bringing me two gorgeous sons, Fayez and Omar, she was always beside me and encouraged my every step toward PhD. I will always cherish all what she has tolerated in this long journey to see me graduating. Living to her and to my family's expectation is my true success and fulfillment. I love you all.

Mohammad F. Al Bataineh

Chicago, May 2010

To my beautiful sons

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Definition
a	the exponential growth parameter in the codebook model
a_i	the i^{th} element of the vector $\boldsymbol{A}_{1\times n}$
A_i	the i^{th} amino acid
$\boldsymbol{\mathcal{A}}$	genomic alphabet $\{A, G, C, T/U\}$
$\boldsymbol{A}_{1\times n}$	a $(1 \times n)$ distance weight vector assigned to the set of conserved binding sites belonging to a given transcription factor
b_i	the i^{th} base
$\boldsymbol{B}_{1\times(L-m+1)}$	a $(1 \times L - m + 1)$ distance metric vector assigned to the set of sequences in the genome with the same length as the binding site sequence (m)
c_n^k	the n^{th} symbol in the k^{th} codeword
cm_i	the i^{th} value in the center of mass vector $\boldsymbol{CM}_{1\times L}$
$\boldsymbol{CM}_{1\times L}$	a $(1 \times L)$ center of mass for a set of binding sites belonging to a given transcription factor
d_H^k	the Hamming distance when codeword k is used
$d_{min_i}^j$	the minimum distance metric for the j^{th} received sequence at the i^{th} alignment
$\boldsymbol{D}_{1\times p}$	a $(1 \times p)$ thresholded distance vector
δ_n	the n^{th} match/mismatch parameter
e_n	the approximation for the small-sample correction in a sequence logo
E_n	the n^{th} free doublet energy in (KCAL/MOL)
$\boldsymbol{E}_{1\times r}$	a $(1 \times r)$ detected distance vector
$Energy(ab)$	the energy dissipated on binding with the nucleotide doublets ab

$g_i^{(j)}$	the i^{th} symbol in the j^{th} generator of a convolutional code
g_L	the g-mask length in a convolutional code
$\boldsymbol{g}^{(i)}$	the i^{th} generator in a convolutional code
$\boldsymbol{G}_{1\times L}$	A given genome of length L
h_i	the polyphase value for the i^{th} base
H	Shannon entropy in bit/symbol
$\boldsymbol{H}_{1\times L}$	The polyphase mapping of the genome $\boldsymbol{G}_{1\times L}$
H_i	the entropy at position i
N_{bi}	the number of times the base $b \in \{A, G, C, T\}$ occurs in column i of the matrix $\boldsymbol{X}_{n\times m}$
M	number of sequences is a dataset
r_{bi}	the probability of the base $b \in \{A, G, C, T\}$ occurring at base position i in a given binding site
R	signal information content in (bit/symbol)
$\boldsymbol{R}_{4\times m}$	A $(4 \times m)$ position frequency matrix for given set of binding sites (each of m bases long) of a given transcription factor
$R_{frequency}$	the amount of information needed to locate a binding site in (bit/site)
ρ_n	the n^{th} offset variable in the codebook model
$R_{sequence}$	informational content of a binding site sequence
$R_{iw}(b, l)$	a $(4 \times L)$ information weight matrix for a base b at position l, L is the number of binding site sequences
$R_i(j)$	individual information content of the j^{th} binding site sequence
S_{ijk}	The j^{th} syndrome value of the i^{th} $mRNA$ subsequence using the k^{th} g-mask
$\boldsymbol{S}_{mRNA}$	the $mRNA$ test sequence
$\boldsymbol{SM}^{(k)}$	the syndrome matrix calculated using the k^{th} g-mask

$\boldsymbol{SM}_{avg}^{(k)}$	the syndrome matrix calculated using the k^{th} g-mask normalized to the number of $mRNA$ test sequences
$\boldsymbol{SM}_{avg}$	$\boldsymbol{SM}_{avg}^{(k)}$ normalized to the number of g-masks
s_n	the n^{th} symbol of the sequence S
σ_n	number of consecutive matches at the n^{th} alignment
$\tilde{\sigma}_n$	number of consecutive mismatches at the n^{th} alignment
T	threshold
u_i	the i^{th} information base in the k-base information vector
$\boldsymbol{u}$	the input information sequence in a block or convolutional code
v_i	the i^{th} encoded base in the $(n-k)$ parity bases
$v_i^{(j)}$	the i^{th} encoded symbol using the j^{th} generator in a convolutional code
$\boldsymbol{v}$	the encoded sequence in a block or convolutional code
w	the information block length that satisfies the one-to-one mapping in table-based decoding of a convolutional code
w_n	the n^{th} weight applied to the energy doublet at the n^{th} alignment in the codebook model
$\boldsymbol{W}$	the weighting vector
$\boldsymbol{X}_{n\times m}$	a $(n \times m)$ matrix representing n binding sites (each of m bases long) of a given transcription factor
$\boldsymbol{Y}_{n\times m}$	The $\boldsymbol{X}_{n\times m}$ matrix mapped to its corresponding polyphase equivalent

LIST OF ABBREVIATIONS

Abbreviation	Definition
A	Adenine
AGL	Actual Gene Locations
BSL	Binding Sequence Length
C	Cytosine
CoM	Center of Mass
DNA	deoxyribonucleic acid
EC	Error Correction
E.coli	Escherichia coli
EWFERD	Exponentially Weighted Free Energy Ribosome Decoding
FWM	Frequency Weight Matrix
G	Guanine
GGL	GLIMMER Gene Locations
GMGL	GeneMark Gene Locations
HGGL	Hypothetical GLIMMER Gene Locations
HMGL	Hypothetical Gene Locations
HMM	Hidden Markov Model
I	Inosine
MDI	Minimum Distance Information
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NTGL	Non-Translated Gene Locations

ORF	Open Reading Frame
ORFGL	Open Reading Frame Gene Locations
PFM	Position Frequency Matrix
RBS	Ribosome Binding Site
RNA	ribonucleic acid
RNAP	RNA Polymerase
rRNA	ribosomal RNA
SD	Shine-Dalgarno sequnece
SNR	Signal to Noise Ratio
T	thymine
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
tRNA	transfer RNA
U	Uracil
UTR	Untranslated Region
VEIL	Viterbi Exon-Intron Locator
WTMM	Wavelet Transform Modulus Maxima

ABSTRACT

The increase in genetic data during the last years has prompted the efforts to use methods and techniques from the engineering fields for their interpretation. These fields include information theory, communications, coding theory, signal processing, machine learning, and various statistical methods. This thesis discusses the role and contributions of communications theory in the field of genetic data analysis. Novel use of techniques and principles from the latter field are proposed for examining and analyzing the genomic structure including coding and non-coding regions. The developed analyses allow testing different biological aspects related to the process of gene translation and determining whether regions of a specified genome are protein-producing sequences. This analysis promotes new interdisciplinary collaborations in research and education, integrating biomedical engineering, electrical engineering and life sciences. The knowledge gain can help address fundamentally important issues that cannot be explored systematically and quantitatively by experimentation alone. Moreover, it can allow savings in laboratory resources and time-consuming laboratory experimentations and leads to better understanding of the complex genetic processes.

This thesis deals with modeling the process of translation in gene expression using a communications theory approach. Gene expression involves two main stages. The first one is transcription (related to coding theory) where the information stored in the DNA is transformed into $mRNA$. The second is translation (related to detection theory), where the noisy $mRNA$ molecule serves as an instructive for protein synthesis. The accuracy of this process is vital to the survival of the living organism.

First, this thesis investigates a variable length codebook model for the process of translation with an exponentially weighted algorithm to optimize its detection mechanism. A mutational analysis is carried out to test the model and certify its correctness and biological relevance. Different communications theory based metrics are used to quantify the mechanism that the ribosome uses to detect the translational signal. Second, a block code model for the process of translation is proposed. The mRNA signal is modeled as a block encoded signal and the ribosome as a block decoder. The last 13 bases of the $3'$ end of the $16S\ rRNA$ molecule were used to build the codebook. Several minimum distance decoders are used to verify the proposed model based on the free energies involved in the binding between the ribosome and the $mRNA$ sequence. Third, a convolutional code model for the process of translation is proposed. The mRNA is modeled as a convolutional encoded signal and the ribosome as a table-based convolutional decoder. Finally, this thesis proposes two novel methods for regulatory sequence detection. The purpose is to examine the distribution of the regulatory sequences in coding and non-coding regions in an attempt to utilize that in gene identification.

The developed models were tested on different bacterial genomes. The obtained results prove the validity, significance and biological relevance of the models being confirmed with experimental published data. This further verifies the relevance of using communications theory to approach several biological problems and hence encourages the research cooperation between the two communities of communications engineering and molecular biology.

As a broader impact, successful development of biological information and coding theory models can provide a theoretical basis for understanding and quantifying error-control mechanisms in natural and synthesized biosystems. In the future, we can envisage diseases, as cancer, AIDS, and geriatric maladies quantified by failures of the genetic system as a whole, rather than simply failures of various components. The interacting of engineering ideas and molecular biology will potentially yield to a quantitative framework for analyzing these anomalies and find solutions for them. The proposed models will provide tools for advancing our understanding of different cellular mechanisms, and has therefore a strong transformative potential for discovering new strategies to mitigate many diseases where the interplay of engineering and biology is critical.

CHAPTER 1

INTRODUCTION

The rapid advances in both genomic data acquisition and computational technology have encouraged development and use of engineering based methods in the field of genetic data analysis. Techniques from engineering fields such as information theory [2, 11, 54, 62, 124, 125, 180], communications [62, 86, 96, 99-103, 178], coding theory [13, 29, 59, 73, 88, 90, 92, 94, 95, 97, 98, 127, 128, 130, 160, 161], signal processing [7, 10, 28, 87, 118, 174], machine learning [171], and various statistical methods [23, 39, 41, 58, 65, 70, 120] are now being actively researched for use in gene and regulatory sequence identification. This identification problem is considered a fundamental challenge in genomics and computational biology. Since regulatory elements are frequently short and variable, their identification and discovery using computational algorithms is difficult. However, significant advances have been made in the computational methods for identification of latter elements. This thesis discusses the role and contributions of communications, and coding theory in the field of genetic data analysis. Novel use of techniques and principles from the latter fields are proposed for examining and analyzing the genomic structure including coding and non-coding regions and studying the interactions between these regions and the different regulatory elements that affect the gene expression mechanism. The developed analyses will facilitate testing different biological aspects related to the process of gene expression. Moreover, it will allow determining whether regions of a specified genome are protein-producing sequences. The approaches used here differ in terms of "methodology and objectives" from the widespread bioinformatics approaches used for gene and regulatory sequence

identification. The latter methods are usually based on hidden Markov models (HMM) and machine learning methods which are mainly concerned with the identification problem. The approaches presented in this thesis are also concerned with the identification process but with a main emphasis on studying genomic structure as well as the relationships and interactions between different regulatory sequences and coding and non-coding regions. Such analyses will lead to better understanding of the biological implications of the gene expression process, discover the signals and functions in the coding and non-coding regions of the genome, and study the various factors affecting the entire process. It will also give an explanation of the detection mechanism that the ribosome adopts to identify the translational signals needed to accurately initiate/terminate the process of translation (protein synthesis). This analysis will promote new interdisciplinary collaborations in research and education, integrating biomedical engineering, electrical engineering and life sciences. The knowledge gain here will help address fundamentally important issues that cannot be explored systematically and quantitatively by experimentation alone. The developed analyses will allow savings in laboratory resources and time-consuming laboratory experimentations and will lead to better understanding of the complex genetic processes.

In this thesis, the last 13 bases sequence of the 3' end of the $16S\ rRNA$ molecule was used as a test sequence with the proposed models. Results show that the proposed models are not only able to identify translational signals in the $mRNA$ sequence, but also can distinguish coding from non-coding regions. Therefore, the proposed models can be utilized to identify genes in the $mRNA$ sequence. Moreover, mutations in the last 13 bases sequence of the $16S\ rRNA$ molecule were also investigated. The obtained results

showed total agreement with published investigations on mutations which further validate the biological relevance of the proposed models. The proposed models were applied to the complete genome sequences of several bacterial genomes including *Escherichia coli* $K-12\ MG1655$, *Escherichia coli* $O157{:}H7$, *Bacillus Subtilis*, *Salmonella Typhimurium* $LT2$, and *Staphylococcus Aureus* $Mu50$.

1.1 Motivation

In 1948, Claude E. Shannon's idea that images, text, and various types of data can be transmitted using a series of binary digits transformed the communication industry and society as a whole. Several years later in 1953, James Watson and Francis Crick [182] announced the discovery of the DNA double helix. (In addition to Watson and Crick's work, Rosalind Franklin's [137] research contributed significantly to the discovery of the structure of DNA). Their discovery led to the eventual realization that proteins and regulatory signals can be represented using a series of quaternary symbols, A, T, G, and C, corresponding to the nucleic acid bases adenine, thymine, guanine, and cytosine. Whereas the quantification of Shannon's treatise on information theory led to the birth of coding theory and has promoted advances in digital communication, satellite communication, storage technology, and biomedical imaging, the parallels and the intersections of Shannon's 1948 ideas with Watson and Crick's 1953 discovery are still being realized.

As evidenced by many articles in this area, researchers are increasingly curious about the communication protocols of molecular systems. In this special area we endeavor to explore ideas at the crossroads of communication theory and molecular

biology from various disciplinary backgrounds and vantage points, providing an overview of the state of research and making compelling observations regarding the nature of biological information transmission in light of the principles of communications, coding theory and information theory.

Communications engineering as well as genetics have both experienced a major breakthrough in the mid 20th century. From this point on it was clear that the genetic information is stored in form of two complementary directed strands composed of letters from a four symbol alphabet. Until the discovery of the molecular basis of genetics, the research was concentrating on classical genetics, based on the rules of Mendelian inheritance of traits. Shannon himself was using mathematics to study how different trait combinations propagated through several generations of breeding in his Ph.D. thesis completed in 1940 [154]. He devised a general expression for the distribution of several linked traits in a population after multiple generations under a random mating system, which was original at that time, but went largely unnoticed, since he did not publish his work. After completing his Ph.D. thesis, Shannon shifted his focus towards digital communications and cryptography. In 1948 Shannon established the theoretical fundamentals of digital communication systems [153]. He introduced the concept of information based solely on the statistical characteristics of the information source. He defined information in an abstract way independent of semantics that does not differentiate between text, video or audio as was generally being done when studying communication systems at that time. Using such information definition, Shannon proved that a message generated by an information source can be compressed to the entropy of the source (source coding theorem) and that it is possible to code the information in a

way, such that one can transmit it error-free at the maximum rate that the channel allows (channel coding theorem). Ever since, communications engineers have been devising algorithms to achieve the limits of these two theorems. The definition of information based solely on statistical characteristics of the information source also applies to genetic data. Recent advances in *DNA* sequencing technology supply enough data to apply Shannon's general information concept to molecular biology.

The original involvement of information theorists with molecular genetics goes back to the discovery of the genetic code. In the period between the discovery of the *DNA* structure in 1953 and the decipherment of the genetic code 1961- 1969, when no actual *DNA* sequences and only very few amino acid sequences were known, several different coding schemes describing the mapping of the *DNA* sequence (size-4 alphabet) to a protein (amino acid sequence from a size-20 alphabet) were proposed by coding theory experts. Some of them have high information density, while others have foreseen error correction capabilities. The experimental discovery of the actual genetic code (the mapping rule of the $4^3 = 64$ *DNA* sequence triplets to the 20 amino acids and a stop symbol) was a disappointment for the coding community since it does not seem to implement any of the two. A review of the proposed codes can be found in [63]. From this point, there has been little interaction between the two communities until recently. We believe that with all the newly available sequence data further interactions would be highly beneficial as our research suggests. The question why the genetic code has evolved the way it is remains open. There seems to be evidence for the optimality of the code in terms of error minimization using metrics based on physio-chemical properties of the resulting amino acids like their hydrophobicity [50]. Apparently, evolution imposes

additional constraints on the optimization of how the genetic information is being stored, which makes the modeling rather peculiar. This has to be accounted for by communications engineers modeling evolution and the molecular processing of genetic information in the cell as a communication system.

This work deals with modeling the process of gene expression (information contained in the DNA molecule when transformed into proteins). Gene expression involves two main stages. The first one is transcription (related to coding theory) where the information stored in the DNA that has been contaminated with genetic noise is transformed into the messenger $(mRNA)$. The second one is translation (related to detection theory), where the noisy $mRNA$ molecule serves as an instructive for protein synthesis (See Figure 1.1). The accuracy of this process is vital to the survival of the organisms.

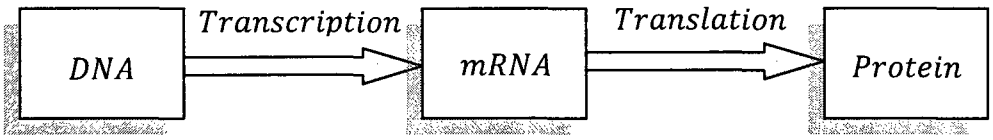


Figure 1.1. The process of protein synthesis (gene expression)

Analyzing gene expression, many similarities with the way engineers transmit digital information come into view. Concepts of information theory, communications, detection theory, pattern recognition and source and channel coding can be used to find out analogies between these fields [30, 87, 98-100, 103, 104, 183]. At the same time the analysis of the results made possible by developing these models can serve as a way to introduce new lines of biological research. In practice, these results can lead to better recognition of signals in gene expression. The use of communications engineering ideas

for understanding genetic information has been lately promoted by the increased availability of genetic data.

The genetic information of an organism is stored in the *DNA*, which can be seen as a digital signal of the quaternary alphabet of nucleotides $\mathcal{A}$ = {A, C, G, T}. Motivated by the redundant structure of the genetic code, the existence of large evolutionary conserved non-coding regions among species, and the existence of special sequences in coding regions, several researchers are trying to apply communications and coding theory models to understand the structure of the *DNA* and the operation of various genetic processes. An summary of ongoing research can be found in [61]. Yockey [180] proposed one of the first models for gene expression using encoding/decoding concepts from communication theory. Liebovitch et al. [79] developed the first efficient method to scan through *DNA* sequences to determine whether some linear block code structure is present. Years later, Rosen [129] developed a method for the detection of linear block codes that accounts for possible insertions and deletions in the *DNA* sequences. However, neither work was able to support the existence of such simple error correcting codes in the *DNA*. Battail [11] argued about the existence of nested error correcting codes in the *DNA* supported by several biological observations such as the size of the genome being far larger than the size needed to specify every characteristic of any given individual. On other fronts, Mac Donaill [89] proposed a parity check code interpretation of nucleotide composition, and May et al. [97] proposed the use of block and convolutional codes to model the process of translation initiation in prokaryotic organisms. Yockey and Schneider pointed out the ambivalence between biological interactions in genetic systems and certain aspects in communication systems [145, 146, 180]. Schneider states that

modeling biological systems using concepts from information theory can lead to a better understanding of the accuracy, mechanisms, and evolution of the molecular machines. In order for the living system to survive, Battail and Eigen suggest the necessity of error correcting capabilities in the replication of the DNA [16, 42]. Taking all this into consideration, May introduced channel coding models for the process of translation initiation [92, 93]. The work of May established some first ideas of modeling the DNA interactions based on coding theory. Additionally, the presented approach paved the way to continue working in the search for new and better models to describe the coding-decoding relations between $DNA - mRNA - rRNA$ molecules. Dawy [30] proposed a communication theory based model for the process of gene expression based on a codebook. Hagenauer [177] used an analogy to frame synchronization to gain more insights into the process of transcription in gene expression.

The main contributions of this work are: i) develop analogies between information transmission in communications engineering and gene expression, ii) develop and validate novel biologically-motivated communications theory based models for the process of prokaryotic translation initiation, iii) gain new insights on the biological interactions between the ribosome and the $mRNA$, and iv) identify coding and non-coding regions in prokaryotic genomes, and v) test the effect of different types of mutations in the ribosome on protein synthesis. This work is an interdisciplinary research that involves cooperation of researchers from different backgrounds. The key goal is to apply techniques from communications engineering to problems in the area of biology.

1.2 Thesis Outline

The rest of this thesis is organized as follows. Chapter 2 presents a theoretical background and a literature survey of the methods, techniques and models that attempted to capture the information theoretic aspects of the genetic system. It also highlights the various connections between information theory and computational molecular biology and defines the central dogma of molecular biology. The chapter then gives a review of the various algorithms and methods used for the gene identification problem, and then a review of information and communications theory based models used for genetic sequence analysis. After that, the chapter describes different coding theory based approaches used for genetic sequence analysis.

Chapter 3 provides the biological background needed to understand the communication theoretic models in later chapters. It is written such that the reader is not required to have any prior knowledge. Important literature references are supplied that provide a broader introduction to the topic than possible in this thesis. The focus lies on gene expression, the process of protein synthesis vital to all organisms. It is covered separately for prokaryotic organisms (specifically for *herichia coli*) with a special emphasis on the translation process.

The following chapters (Chapter 4 to Chapter 7) detail the communications theoretic modeling for the process of translation in prokaryotic organisms. A novel use of techniques and principles from communications engineering for modeling, identification and analysis of genomic regulatory elements and biological sequences is presented.

Chapter 4 investigates a model based on a variable length codebook and a metric for the process of translation in gene expression. In this model it is assumed the ribosome

decodes the $mRNA$ sequence by using the 3' end of the $16S\ rRNA$ molecule as an embedded codebook. The metric uses an exponentially weighting free energy decoding algorithm to identify the $Shine\text{-}Dalgarno\ (SD)$ sequence. This algorithm allows for better resolution and flexibility in detecting the translational signals by a simple change of parameter values. The validity and biological relevance of this model is verified by testing the effect of different types of mutations in the ribosome on protein synthesis. Results are compared to biological experimental data and prove to be consistent. The chapter then presents different analyses for the detection mechanism that the ribosome uses to identify the translational signals based on a communications theory approach. Concepts of Euclidean distance and cross correlation are utilized to emulate the ribosome detection mechanism. Results show that the proposed approach is able to identify coding and non-coding regions by a clear difference in the detection signal ripple in both regions.

Chapter 5 and 6 investigate the use of coding theory based models that quantitatively describe the behavior of the ribosome during translation process in gene expression. Chapter 5 presents a code model for the process of translation in prokaryotic organisms. The model employs several minimum distance decoders to verify the validity of the model based on the free energies involved in the binding between the ribosome and the mRNA sequence. The model was tested on five different bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs (GeneMark and Glimmer).

Chapter 6 investigates a convolutional code model to analyze the process of translation using table-based decoding. In the investigated model, the messenger RNA ($mRNA$) sequence can be viewed as a noisy convolutional encoded signal, and ribosome as a table-based convolutional decoder. The $16S$ ribosomal RNA ($16S\ rRNA$) sequence is used to form decoding masks for table-based decoding. A syndrome matrix is calculated. The model was tested on five different bacterial genomes. The obtained results verify the validity and the significance of the investigated convolutional code model and its biological relevance by being able to identify the right start codons from the false ones. Most importantly, the results also show that the convolutional code model investigated here can be used to interpret some of the aspects related to the process of translation in prokaryotic organisms.

Chapter 7 deals with transcription factor binding site detection in an attempt to utilize that in gene identification. Two novel techniques for $TFBS$ detection are proposed. While the first detection technique is based on a frequency weight matrix (FWM) concept, the second one uses center of mass based metrics and polyphase mapping. Results show that around 85% of the detected TFBSs are totally located in the non-coding region, 6.5% are totally located in the coding regions, and only 8.5% are overlapped between coding and non-coding regions.

Finally, Chapter 8 concludes this thesis. The main achievements are detailed along with future research directions.

CHAPTER 2

THEORITICAL BACKGROUND AND LITERATURE SURVEY

This chapter presents a theoretical background and a literature survey of the methods, techniques and models that analyzed the information related aspects of the genetic system. It also highlights the various connections between information theory and computational molecular biology and defines the central dogma of molecular biology. The chapter then gives a review of the various algorithms and methods used for the gene identification problem, and then a review of information and communications theory based models used for genetic sequence analysis. Finally, the chapter describes different coding theory based approaches used for genetic sequence analysis.

2.1 Informational Genetic Sequence Analysis

The study of the information processing capabilities of living systems was revived in the later part of the 1980s [48, 126, 135], due to the increase in genomic data which spurred a renewed interest in the use of information theory in the study of genomics. Information measures, based on the Shannon entropy [152], have been used in recognition of *DNA* patterns, classification of genetic sequences, and other computational studies of genetic processes [83, 112, 140, 141]. Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communication engineering [42, 52, 53, 180]. Several researchers have explored the central dogma of genetics (i.e. the gene expression process) from a communications theory point of view. A review of the information theoretic models developed by different researchers in genetic sequence

analysis is given in the following sections. These researchers include Gatlin, Yockey, Liebovitch, Roman-Roldan, Rosen, Battail, Eigen, Mac Donaill, Schneider, May, Hagenauer, Dawy and others. The ability to extend information theory concepts to study information transmission in biological systems will contribute to the general understanding of biological communication mechanisms and extend the field of coding theory into the biological domain.

The genetic information of an organism is stored in the DNA, which can be seen as a digital signal of the quaternary alphabet of nucleotides $\mathcal{A} = \{A, C, G, T\}$. Motivated by the redundant structure of the genetic code, the existence of large evolutionary conserved non-coding regions among species, and the existence of special sequences in coding regions, several researchers are trying to apply communications and coding theory models to understand the structure of the DNA and the operation of various genetic processes. A summary of ongoing research can be found in [62].

Roman-Roldan et al. suggest that living beings can be characterized by their information processing ability and hence information based analysis can be used in their study. Viewing protein synthesis as an information processing system allows nucleotide sequences to be analyzed as messages without considering the physical-chemical elements for information processing. Transfer of biological information can be modeled as a communication channel with the DNA sequence as the input and the amino acid sequence which forms protein as the channel output [126]. Yockey [180] proposed one of the first models for gene expression using encoding/decoding concepts from communication theory. Liebovitch et al. [73] developed the first efficient method to scan through DNA sequences to determine whether some linear block code structure is

present. Years later, Rosen [128] developed a method for the detection of linear block codes that accounts for possible insertions and deletions in the DNA sequences. However, neither work was able to support the existence of error correcting codes in the DNA. Battail [12] argued about the existence of nested error correcting codes in the DNA supported by several biological observations such as the size of the genome being far larger than the size needed to specify every characteristic of any given individual. On other fronts, Mac Donaill [88] proposed a parity check code interpretation of nucleotide composition. Yockey and Schneider pointed out the ambivalence between biological interactions in genetic systems and certain aspects in communication systems [144, 146, 180, 181]. Schneider states that modeling biological systems using concepts from information theory can lead to a better understanding of the accuracy, mechanisms, and evolution of the molecular machines. In order for the living system to survive, Battail and Eigen suggest the necessity of error correcting capabilities in the replication of the DNA [14, 15, 42]. Taking all this into consideration, May et al. [97] proposed the use of block and convolutional codes to model the process of translation initiation in prokaryotic organisms [38, 92]. Additionally, the previous research paved the way to continue working in the search for new and better models to describe the coding-decoding relations between $DNA - mRNA - rRNA$ molecules. Dawy et al. [184] proposed a communication theory based model for the process of gene expression based on a codebook. Weindl and Hagenauer [178] used an analogy to frame synchronization to gain more insights into the process of transcription in gene expression.

2.2 Biology as an Information Science

There are many connections between information theory and computational molecular biology [180]. We can think of cells not only as transducers of energy but also as small computers that deal with discrete symbols occurring in DNA, RNA, protein, etc.

The information that is held in the genome and transcribed and translated into proteins to govern metabolic and regulatory processes has become a key focus of biology. Because of the algorithmic and control theoretic aspects of such a system, fields such as dynamical systems, control theory, information theory and algorithm design will play a central role in the "new system biology".

The work of the entropy related to the compressibility of $exons$ (that comprise the expressed parts of our genes) as compared with the $introns$ (that intervene between those exons) represents one of the many connections between information theory and computational biology. A famous early misstep in the history of molecular biology was the hypothesis that the genetic code is a comma-free code[1], which would very neatly have explained why we have exactly 20 amino acids, but turned out to be quite contrary to the biological realities. There is also the general notion that meaningful patterns are highly compressible or have low Kolmogorov complexity [78]. This suggests a number of information theoretic approaches to finding structure in discrete data such as genomic sequences and gene expression data. Many methods for clustering and classification of biological data and feature selection to describe dependencies between various variables are available. These methods are based on information theoretic approaches like minimum description length principle, information bottleneck concept, and conditional

[1] A code constructed so that any partial codeword, beginning at the start of a codeword but terminating prior to the end of that codeword, is not a valid codeword.

entropy (used in selecting relevant features explaining some observed variables). The information theory bounds in computer science have been applied to various probing experiments to identify biological sequences. Finally, it can be productive to think of evolution as a noisy channel where particular biological characters, mutations, and rearrangements are passed from species to descendent species. One can take this point of view to ask how easily or efficiently one can reconstruct the genomes of ancestral species given the genomic information about extinct species.

The most basic principle in molecular biology is the fact that *DNA* is the passive container for the genomic information. Proteins are created into two-stage process: transcription and translation (see Figure 2.1). In transcription, *DNA* gets transcribed symbol by symbol to an intermediate form called messenger *RNA* (*mRNA*) . In translation, *mRNA* is then transported out of the nucleolus of eukaryotic cell to molecular machines called *ribosomes* where it gets translated into protein according to the universal genetic code [165].

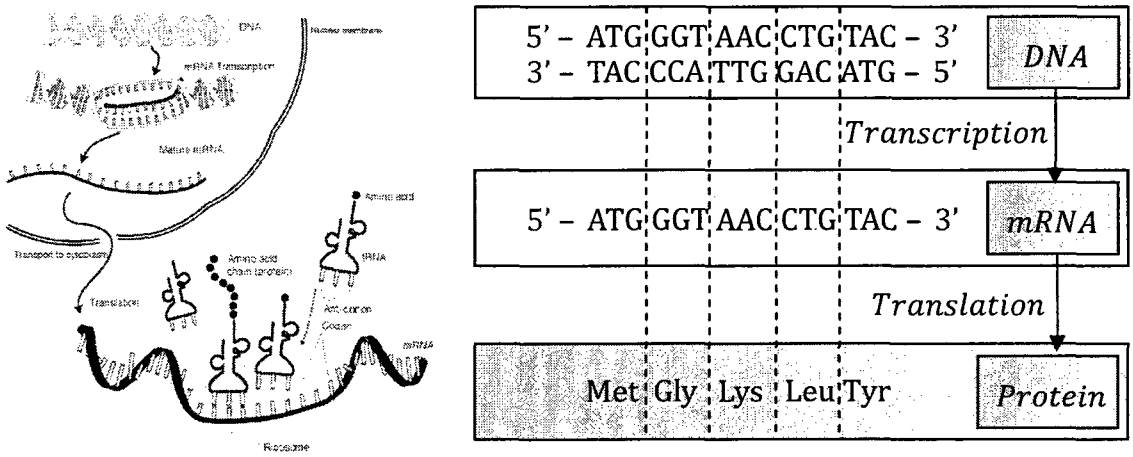


Figure 2.1. Protein Synthesis (Gene Expression)

2.3 Computational Methods used for Gene Identification

Advances in genomic sequencing have provided large amounts of data and have spurred computational tools for recognition and modeling of protein coding regions and accurate identification of exact translation start sites [28, 51, 60, 106, 109, 151, 166, 170, 175, 185]. For example, probabilistic methods, such as Suzek et al.'s *RBSFinder* and Yada et al.'s *GeneHacker Plus* [179] return the location of the initiation codon for prokaryotic genes. Besemer and Borodovsky's *GeneMarkS* uses iterative hidden Markov models (*HMMs*) to locate translation start sites with relatively high accuracy [20]. Walker, et al. couple statistical methods with comparative genomics methods to identify start sites. Hannenhalli et al. incorporate several biological factors into their quantitative description of translational start sites, including the binding energy at the ribosome binding site (*RBS*), distance between *RBS* and initiator, and the initiator codon. They use a mixed integer linear program to determine parameters for their discriminatory model. *GeneLook* identifies protein-coding sequences using a two-stage, *ab initio* process [106]. Classification is based on structural characteristics of the sequence such as properties of the ribosome binding site (*RBS*), operon structure, and codon and nucleotide frequency. Other computational techniques including support vector machine, machine learning, combinatorial approaches, free energy calculations, Bayesian methods, and information theory have also been used in quantifying and classifying translational start sites [28, 64, 108, 139, 185].

Though current computational methods have provided tools for locating start sites and increased the overall accuracy of gene locater systems such as *GLIMMER* and *GeneMark*, they usually require larger sequence windows for classification of initiation

sites. Several initiation site classification tools such as *RBSfinder* use prior gene classification knowledge to aid in start site identification, hence functioning more like a postprocessor. The ribosome, the protein translation machine, makes initiation decisions based on "real-time" processing of a single messenger *RNA* leader region. In this thesis, we are trying to analyze the analogies between genetic information processing in living organisms and the processing of communications data. The conducted analyses are meant to develop several communications theory based models for the process of translation. This is done in an attempt to simulate the detection mechanisms that the ribosome uses to identify the right start site and hence correctly initiate/terminate the translation process. The next section provides an overview of different algorithms and methods used for gene identification.

2.4 Review of Gene Identification Algorithms and Methods

Fickett [46] defines the gene identification dilemma (from the developers' point of view) as the problem of using computers to decipher nucleotide sequences for the purpose of locating and defining the structure and the functionality of protein producing genes. The goal of gene identification is the automatic annotation of genomic sequences. This means, given a nucleotide sequence, the identification system can determine all regions that are biochemically active and describe the reaction that occurs and the products of said reaction [46]. The long-range goals of gene identification include the development of algorithms for the purpose of designing new genes and genomes [46].

Computational techniques used in developing gene identification systems can be categorized as either template methods or lookup methods. Template methods use

prototypes to identify genes. Unknown genes are compared to a prototype and classified based on a pre-defined metric [46]. The lookup methods, or similarity searches, compare the unknown gene or nucleotide sequence to known genes or gene components stored in databases [46]. While the template method adds to our understanding of the gene, it excludes outliers or important exceptions to the gene prototype [46]. Although the lookup method may include exceptions, sequencing errors can result in serious problems when using a look up method [46].

2.4 .1 Template Methods. The majority of computational gene identification techniques reviewed in this section can be classified as template methods. Although some methods do not always fit neatly into a specific category, the following four categories will be used to classify computational techniques discussed:

- Hidden Markov Methods
- Machine Learning Methods
- Signal Processing Methods
- Information Theory Methods

The basis of many gene identification systems is a value Fickett refers to as a coding measure. For a given sequence window, a coding measure assigns to a sequence a value or vector of values which correspond to the probability that the sequence is a protein coding sequence [46]. Some systems employ various combinations of computational methods and coding measures for the purpose of gene identification and classification. Fickett found that combining several measures improves the accuracy of the identification system [46].

Hidden Markov Methods. Advanced gene finders for both prokaryotic and eukaryotic genomes typically use complex probabilistic models, such as hidden Markov models (*HMM*), in order to combine information from a variety of different signal and content measurements. Hidden Markov Models are used to model systems with hidden discrete states [65]. Biological sequences, such as nucleic acid sequences, can be modeled as the output of a process that goes through discrete states,v. The modeling assumption, or Markov assumption, on which *HMMs* are based, is that the states that follow any state, v, depend only on v and are independent of all states which precede v [65]. The *HMM* is defined by a set of possible states and transitions. Each state is associated with a discrete output probability distribution. There are also transition probabilities for each possible transition from a given state. The sum of all transition probabilities must be equal to one for a given state, v [65].

HMMs, which are also used in speech recognition, have been applied to various biological systems [39-41, 58, 65, 70, 85, 119]. Once the *HMM* is developed for a particular sequence or group of sequences, it is then used to classify sequences whose functions are not yet determined. For example, the *HMM* modeling approach was used by Henderson, Salzberg, and Fasman [65] to develop *VEIL* (Viterbi Exon-Intron Locator), an *HMM* system that segments uncharacterized eukaryotic *DNA* sequences into exons, introns, and intergenic regions. This *HMM* uses the Viterbi decoding algorithm, which is also used in coding theory [34, 35], to determine the probability of the *HMM* generating the observed sequence.

Another *HMM* based model, *HMMER*, was designed by Eddy [39-41]. *HMMER* provides tools for constructing an *HMM* model from initially unaligned training

nucleotide sequences. This software provides a useful tool for modeling sequences that may have similar characteristics. Once the model is constructed, it can then be used to determine how well uncharacterized sequences align with the sequences that the *HMM* models.

The *HMM* method has also been used to develop other gene identification tools such as *GeneMARK.hmm* [85] and *Meta − MEME* (a motif-based *HMM* for protein sequences) [58], and *GLIMMER. GeneMark.hmm* is derived from *GENEMARK* [24]. *GENEMARK* uses a Markov chain model in its gene identification algorithm [85]. Krogh, Mian, and Haussler [70] use a hidden Markov model to locate genes in *E.coli* nucleotide sequences. The *GLIMMER* system is a widely used and highly accurate gene finder for prokaryotes [32].

Machine Learning Based Methods. Machine learning techniques such as neural networks are also used in the analysis and classification of nucleic acid sequences. For example, a multiple sensor neural network was used by Uberbacher and Mural [172] to elucidate protein-coding regions in human *DNA* sequences. They used seven algorithms whose results serve as inputs to the neural network. They evaluate a ninety-nine base window. The seven values indicate the likelihood that a given sequence position is part of a coding region. After training the neural network and extracting the weights, the neural network is used to characterize human *DNA* sequences as coding or non-coding. Neural network methods previously evaluated outperformed most of the statistical methods [46]. Fickett attributes the neural networks performance either to the factors considered by the network or to the integration method employed by the network [46].

Signal Processing Based Methods. Coding measures derived from signal processing techniques such as Fourier Transforms [27] and Wavelet Transforms [7] have been used in developing gene recognition algorithms and analyzing genetic sequences.

Veljkovic [174] and Cosic [26] used Fourier analysis of DNA, RNA, and protein sequences to find a parameter that relates the biological function of nucleic acid and amino acid sequences to their functionality. Using cross-spectral and spectral analysis, the biological signal was analyzed as a finite-length deterministic signal. Since there are examples of biologically unrelated sequences that have great homology, Veljkovic and Cosic's work introduced a new approach for functionally evaluating and classifying biological sequences. They evaluated sequences based on their frequency domain characteristics. Cosic used spectral analysis as the foundation for her resonant recognition model (RRM) which showed a correlation between the biological function of a sequence and the frequencies which are present within the biological signal.

Arneodo et al. [7] developed the Wavelet Transform Modulus Maxima ($WTMM$) method and applied it to human genomic sequences. Arneodo et al. claim that $WTMM$ analysis can provide a definite answer to questions regarding the long-range correlation properties of DNA coding and non-coding sequences. The wavelet transform, which is the basis of the $WTMM$ method, is able to characterize the scaling properties of fractal objects or signals even in the presence of low frequency trends. Arneodo et al. conclude that $introns$ or non-coding subsequences behave as positively correlated fraction Brownian Motion while coding regions, $exons$, behave like uncorrelated ordinary Brownian motion [7]. Results of this Wavelet Transform based analysis of nucleic acid

sequences could be used in designing gene identification algorithms that classify sequences as protein-coding or non-coding.

2.4.2 Lockup Methods (or Similarity Searches). Sequence similarity searches or lookup methods are based on sequence conservations resulting from evolutionarily conserved properties [46, 114]. Many lookup methods use alignment scores as measures for determining protein function or protein coding potential. Two types of alignment methods are used by lookup methods [114]:

- Local Alignment - Finds the region of greatest similarity between two sequences. Differences outside of the region of greatest similarity are ignored.
- Global Alignment - Requires the sequence alignment to start at the beginning of each sequence and to continue to the end of each sequence.

These alignment methods are used in rigorous, similarity-search algorithms (such as the $Needleman-Wunsch$ and $Smith-Waterman$ algorithms) and in rapid, heuristic algorithms (such as the widely used $FASTA$ and $BLAST$ algorithms) [114].

Rigorous algorithms calculate the optimal similarity score between two sequences while rapid heuristic algorithms do not guarantee an optimal score for every element in a sequence library [114]. Although they do not guarantee optimal scores, rapid heuristic methods are five to fifty times faster than rigorous algorithms like the $Smith-Waterman$ algorithm. The faster execution time of rapid algorithms are due to the smaller number of potential alignments analyzed by rapid techniques [114].

Two main rapid heuristic methods, $BLAST$ and $FASTA$, can be characterized as follows [114]: $BLASTP$ (1) is the most widely used program for rapid sequence comparison, (2) accurately estimates statistical significance of similarity scores, (3) looks

at regions with conserved amino acid triplets, and (4) uses a discrete finite automaton to recognize substitutions. $FASTA$ (1) calculates optimal scores and accurately estimates significance of scores, (2) performs, with the above improvements, better than $BLASTP$ and nearly as well as $Smith-Waterman$ methods, (3) looks at regions with high pairwise and single element alignment similarities, and (4) uses $Smith-Waterman$ algorithm to produce the final sequence alignments.

Similarity searches are used mostly for determining the functionality of proteins. They can be used in combination with template methods for annotating genomes. The first pass annotation of the Drosophila melanogaster genome was performed using similarity searches as well as gene finding software [115]. Greater weight was placed on results from the similarity methods than the gene finding software. Though similarity methods provide insight into the functionality and identity of new genes, rarely expressed genes may be difficult to locate with lookup methods.

2.5 Information Theory and Genetic Sequence Analysis

Historically, the application of information theory to genetic analysis began in the $1970s$ [47, 122, 126]. Between 1970 and 1977, in an attempt to quantify and convey the complexity of DNA, methods were developed for estimating information, redundancy or divergence parameters for DNA sequences [126]. These efforts did not prove completely successful. After a ten year hiatus, the increase in genomic data encouraged renewed interest in the use of information theory concepts in the study of genomics. This second research period began in 1987 and continues to the present. In this present period, techniques from the field of signal processing and communications (such as

autocorrelation analysis, Fourier transform, and random walks) have been used in the informational analysis of genetic sequences. Discovering the existence of long-range correlations in *DNA* sequences proved to be a significant result of information-based analysis of genetic sequences. Mutual information, an information theory measure, has also been used to detect long-range correlations in nucleotide sequences. Other information measures, such as entropy based measures, have been used in recognition of *DNA* patterns, classification of genetic sequences, and various other computational studies of genetic sequences [5, 6, 49, 82, 107, 110, 111, 126, 133, 139, 141, 143, 145, 146, 164].

Several researchers have explored the central dogma of genetics from an information transmission point of view. The central premise of genetics is that genetic information is perpetuated in the form of nucleic acid sequences once expressed as proteins [71, 77]. Various investigators have developed models that attempt to capture different information theoretic related aspects of the genetic system. Most of these models are based on the information contained in *DNA* sequences. Some of these models are described in the following subsections.

2.5.1 Gatlin's Model. One of the earliest work on the information theoretic properties of biological systems is presented by Gatlin [52]. Gatlin's interpretation of the biological information processing system is depicted in Figure 2.2. In Gatlin's model, *DNA* base sequences are the encoded message generated by a source encoder. Gatlin suggests that extra bases in *DNA* may be used for error control and correction purposes. The encoded *DNA* goes through a channel (defined in Gatlin's model by transcription and translation) which Gatlin refers to as all the mechanics for protein production. The amino acid

sequence of the protein is the received message. It is unclear where *DNA* replication fits or whether Gatlin considers the replication process as part of the encoder. Although transcriptional and translational errors occur, replication also introduces errors that propagate beyond a single replication event. Replication is a potentially significant source of noise and should be addressed explicitly.

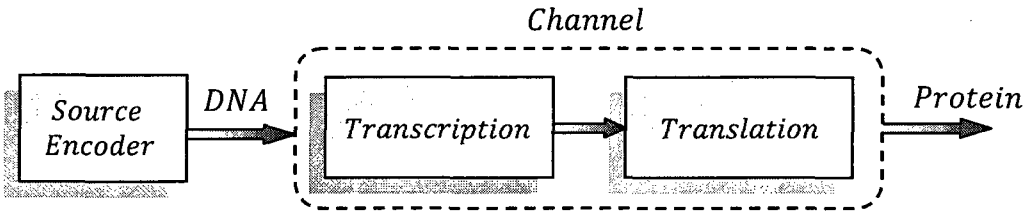


Figure 2.2. Gatlin's communication theory view of the genetic system

In addition to an information theoretic view of genetic processes, Gatlin also parallels the genetic sequence to a computer program. She proposes that the genetic code can be viewed as "part of an informational hierarchy" where the redundant *DNA* regions and the non-coding *DNA* have important programmatic control functions. It is well known that non-protein coding regions of *DNA*, such as promoters and the 5' untranslated leader (*UTR*) of messenger *RNA* (*mRNA*), have regulatory functions in the protein synthesis process [43, 71].

2.5.2 Yockey's Model. Yockey [180] performs a fundamental investigation of biological information theory and lays the foundations for developing theoretical biology from the mathematical principles of information and coding theory. Yockey's biological information framework diverges from the traditional communication system model and is based on a data storage model, where the behavior of the genetic information system is

compared to the logic of a Turing machine[2]. The *DNA* is paralleled to the input tape where the genetic message is the bit string recorded on the tape. The computer program or internal states of the Turing machine are the *RNA* molecules and molecular machines that implement the protein synthesis process. The output tape, similar to Gatlin's model, is the protein families produced from the recorded message in *DNA*.

Error Correcting (*EC*) codes are used in data storage media to ensure data fidelity; hence, Yockey's model incorporates EC coding. Yockey's *DNA–mRNA–protein* communication system is recreated in Figure 2.3 [180]. In Yockey's *DNA–mRNA–protein* communication system, the source code, genetic message in *DNA*, is stored on the *DNA* tape. Transcription is the encoder, transferring *DNA* code into *mRNA* code. Messenger *RNA* is the channel by which the genetic message is communicated to the ribosome, the decoder. Translation represents the decoding step where the information in the *mRNA* code is decoded into the protein message or the protein tape. Genetic noise is introduced by events such as point mutations. Yockey states that while genetic noise can occur throughout the system, all of the noise is represented in the *mRNA* channel. In Yockey's model, the genetic code (the codon to amino acid mapping) is the decoding process and referred to as a block code. He suggests that the redundancy in the codon to amino acid mapping is used as part of the error protection mechanism. Therefore, we can assume that the transcription step would be the *EC* encoding step in Yockey's model. This assumption is consistent with Yockey's inclusion of transcription in the channel code portion of his communication system diagram. Yockey's *DNA–*

[2] A Turing machine is a theoretical device that can manipulate symbols contained on a strip of tape, and can be adapted to simulate the logic of any computer algorithm.

$mRNA-protein$ system is a discrete, memoryless (probability of symbol error is statistically independent of the error history of the preceding symbols) [167, 180].

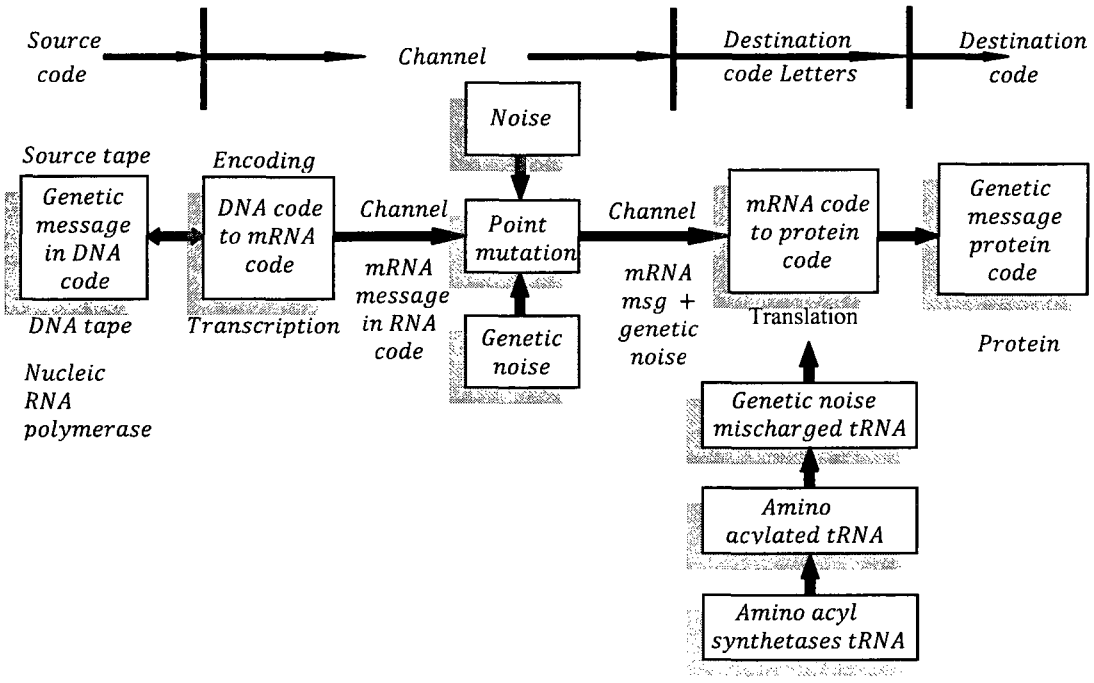


Figure 2.3. Yockey's $DNA–mRNA–protein$ communication system\

2.5.3 Roman-Roldan et al.'s model. Roman-Roldan et al. suggested that living beings can be characterized by their information processing ability and hence information based analyses can be used in their study [126]. The use of information theory in genetic data analysis requires redefinition of the genetic system as an information system. According to Roman-Roldan et al., "the processing of biological information has an artificial parallel: the processing of information by computers". Viewing protein synthesis as an information processing system, allows nucleotide sequences to be analyzed as messages without considering the physical-chemical elements for information processing [126]. Similar to Gatlin, Roman-Roldan et al. modeled the transfer of biological information as a communication channel with the DNA sequence as the input and the amino acid sequence which forms protein as the channel output [126], depicted in Figure 2.4.

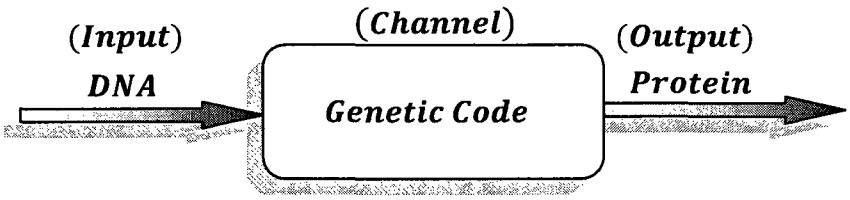


Figure 2.4. Information Theory Based View of Gene Expression (Roman-Roldan)

Roman-Roldan et al. defined the genetic information source as an ergodic source that generates messages from a finite alphabet. An ergodic source is a source that, using random selection criteria, generates typical messages and atypical messages. Typical or statistically homogenous messages are generated with probability close to one while atypical messages are generated with probability close to zero [125]. Roman-Roldan et al. defined the genetic information source with the following parameters:

- Genetic alphabet: $\mathcal{A} = \{A, G, C, U\}$; where the symbols of the alphabet represent Adenine, Cytosine, Guanine, and Uracil, respectively,
- $p(A) + p(G) + p(C) + p(U) = 1$,
- Genetic message source is modeled as a Markov source (bases in a message are not independent) with a stochastic distribution matrix,
- $[p(b_i|b_j)], \sum_i p(b_i|b_j) = 1$, where $b_i \in \mathcal{A}$,
- The Markov source is assumed to be stationary and ergodic.

Similar to Yockey and Gatlin, Roman-Roldan et al. designated the genetic code, the process of mapping codons to amino acids, as the transmission channel through which DNA is transmitted and protein is received. If the genetic channel is noiseless, or free of genetic mutations, in Roman-Roldan et al.'s model the input/output probabilities are specified as follows:

$$p(A_i|b_1b_2b_3) = \begin{cases} 1, & if\ (A_i|b_1b_2b_3)\ is\ a\ part\ of\ the\ genetic\ code \\ 0, & Otherwise \end{cases}, \qquad (2.1)$$

where A_i is the i^{th} amino acid and $b_1b_2b_3$ represents the codon. It is assumed that Roman-Roldan et al.'s model essentially parallels Gatlin's model since DNA rather than $mRNA$ is the input into the channel. As in previous models, Roman-Roldan et al. does not address the role of DNA replication in the genetic information transmission framework. Additionally, their model does not explicitly address the presence or function of redundancy in the DNA input sequence.

2.5.4 May et al.'s model. The communication channel view proposed by Roman-Roldan et al. differs from the initial model presented by May et al. [91]. May's initial model defined the $mRNA$ as the output of a communication channel and incorporates a decoder that translates the $mRNA$ into protein forming amino acid chains. Originally the channel consisted of the DNA replication and transcription process during which errors are introduced into the nucleotide sequence. Based on Battail and Eigen's work, May's initial communication view of the genetic system is modified as follows: (1) the replication process represents the error-introducing channel; (2) a nested genetic encoder is assumed. The genetic decoding process is separated into three phases: transcription, translation initiation, and translation elongation plus termination [17, 42, 91]. Figure 2.5 depicts May's coding theoretic view of information transmission in genetic systems. In the developed genetic communication system, the un-replicated DNA sequence is the output of an error-correcting genetic encoder that adds redundancy to inherently noisy genetic information. The noise in the source can be thought of as mutations transferred from parent to offspring. Drawing from Gatlin's model, May considered the non-coded genetic information that the organism is communicating as instructions for protein production or control of protein production. In contrast to Yockey's model, the encoder

does not emulate any aspect of the central dogma of genetics. Unlike replication, transcription, and translation, the encoding process does not and must not introduce errors into the message. Therefore, none of these genetic processes can adequately emulate an EC encoder. Defining the genetic encoder in our model would require addressing the question of the origin of the genetic code. Perhaps, additional insight into potential biological functions corresponding to the encoder will emerge as researchers continue investigating the evolution of the genetic code and speculating on the origins of the code [168].

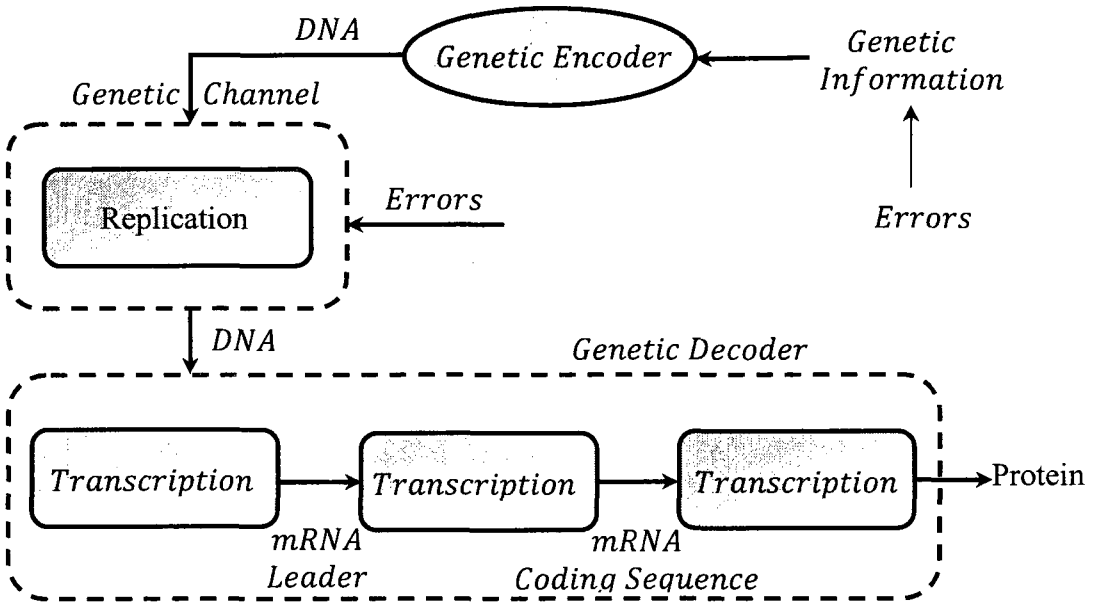


Figure 2.5. May et al.'s coding theoretic view of the central dogma of genetics

Neither of Gatlin, Yockey, or Roman-Roldan et al.'s models explicitly addresses replication in their models. All three frameworks represent the noise introducing channel as the genetic mechanisms responsible for protein synthesis, namely transcription and translation in Gatlin and Roman-Roldan et al.'s frameworks and the $mRNA$ itself in Yockey's framework. May defined the genetic channel as the DNA replication process

during which errors are introduced into the nucleotide sequence [43]. Similar to Roman-Roldan et al., May assumed the transmission channel to be stationary, memoryless and discrete. This is also in line with Yockey's assumptions about the genetic channel. Assuming the channel is memoryless simplifies the model but may not accurately represent biological systems and events such as the existence of mutation hotspots. Further investigation is necessary to determine the channel characteristics of the replication channel.

Incorporating Battail's nested coding idea [12], EC decoding occurs in three phases represented by transcription, translation initiation, and translation elongation plus termination. Similar to Yockey, the ribosome is paralleled to an EC decoder. Given the similarities between the translation and transcription mechanisms, May represents transcription as a decoding step and the RNA polymerase as the EC decoder.

2.5.5 Gail Rosen's Model. In [129], Gail developed a method to uncover an error correction coding structure in the nucleotide sequence, and showed that her framework is efficient for detecting approximate tandem repeats, such as microsatellite regions. According to Gail, communication channel models can be paralleled to DNA processes. In one doctrine, the channel is assumed to be the amino acid translation from nucleotide triplets [52]. In May's model, the channel is the actual replication process, and the DNA is the medium in which genetic information is transmitted from generation to generation [43]. The latter is good for mutation modeling since transcription and copying of DNA is a noisy process.

In Figure 2.6, Gail assumed that the DNA is the sequenced genomic data available in $GenBank$ [116] and that her goal was to examine the dashed-line-encompassed area

and uncover the encoder scheme; in other words, she tried to infer structure from the noisy output to retrieve the original genetic information. According to Gail, nothing is known about the encoder or the original information; therefore, system identification and deconvolution methods cannot be used. Gail assumed that the encoder is linear and tried to characterize it given such output.

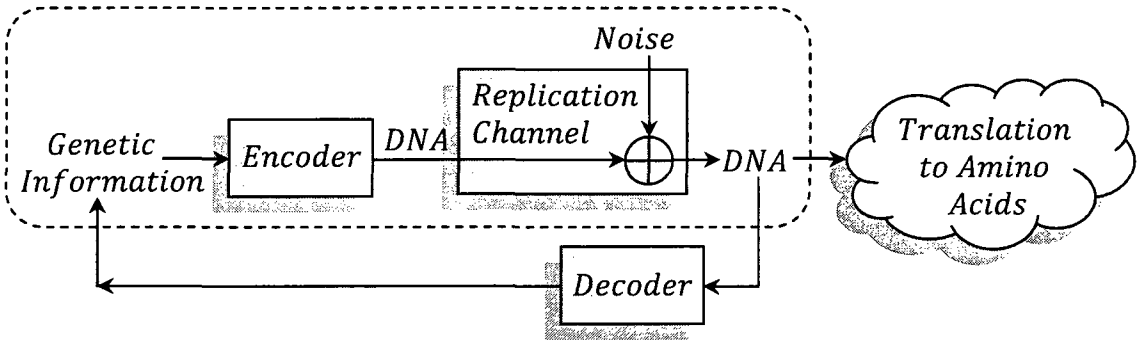


Figure 2.6. Gail Rosen's channel model of genome

2.5.6 Dawy et al.'s Model. Dawy et al. proposed a communication theory based model for the process of translation in gene expression using elements from communications engineering as depicted in Figure 2.7 [30]. His model is based on the assumption that the ribosome decodes the $mRNA$ sequences using the 3' end of the $16S\ rRNA$ molecule as a one dimensional codebook. The biological consistency of the model is proven in the detection of the *Shine-Dalgarno* signal and the initiation codon for translation initiation. Interestingly, the obtained results lay out the possibility of interaction of this part of the ribosome in the process of translation termination. Results obtained via Dawy's model are compared with published experimental results for different mutations of the $rRNA$ molecule. Total agreement between both sets of results proves the validity of the proposed model. By means of simulated mutations in the last 13 bases of the $16S\ rRNA$, Dawy established a global analysis of this part of the ribosome in the process of

translation. This model illustrates the relevance of communication theory based models for genetic regulatory systems.

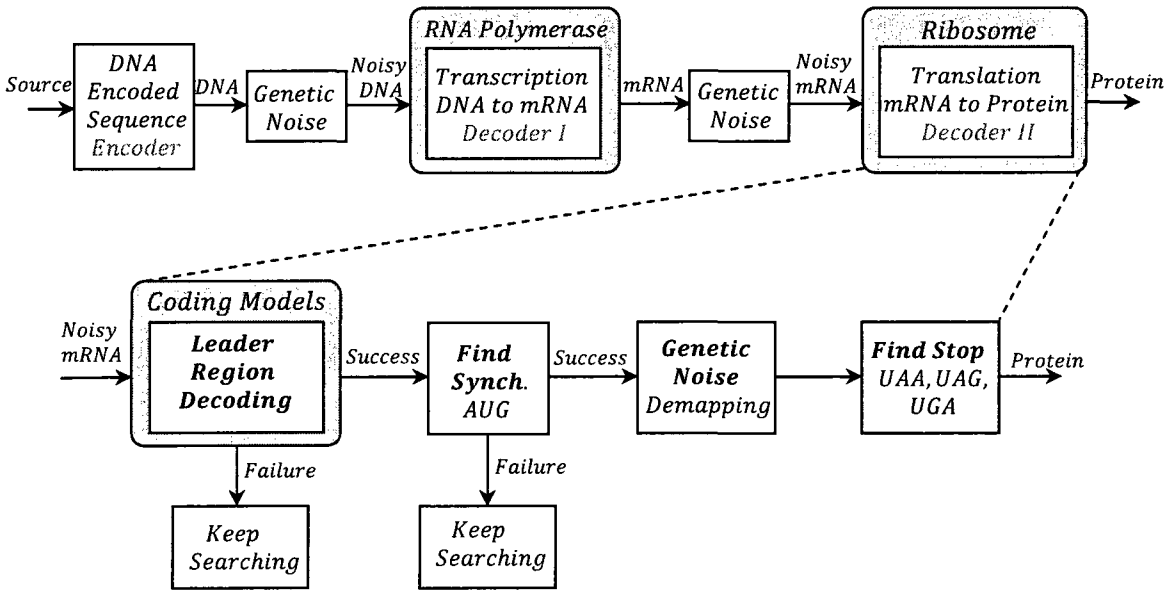


Figure 2.7. Dawy et al's communications model for the process of gene expression

The noisy $mRNA$ is the input of the system. In the leader region decoding, the ribosome decodes the $mRNA$ leader region in order to detect a signal that will enable the ribosome to start translation. If it does not find the signal then the ribosome keeps scanning the $mRNA$. Once the signal is found, the ribosome will try to start translation. When the synchronization signal (AUG) is found, the process of elongation starts where the ribosome uses the genetic code to start demapping (demodulating) the triplets (codons) in the $mRNA$ sequence into a chain of amino acids. In the last stage of the model, the ribosome recognizes one of the stop codons (UAA, UAG, UGA) and then the protein production complex is liberated from the $mRNA$ resulting in the production of the protein. The lower part of Figure 2.7 shows a communications model for the process of translation initiation.

2.5.7 Schneider's Model. The information based model, used by Schneider et al. [149], is based on the information theory concept defined by Shannon. Shannon defined one "bit" as the amount of information needed to distinguish between two equally probable symbols. To distinguish M symbols which have equal probability of occurrence, we need $\log_2 M$ bits [146]. For a given signal, a series of symbols or bases, Schneider et al. defined an associated information based measure R, a variable which indicates the average (in an information theory sense) amount of information in the signal. R is a measure of the information gained and is measured in bits/symbol [146]:

$$R = H_{before} - H_{after}, \tag{2.2}$$

where

$$H = -\sum_{i=1}^{M} p_i \log_2 p_i, \tag{2.3}$$

where p_i is the probability of each symbol i.

Schneider et al. use these information theory measures to analyze nucleotide sequences and identify highly conserved regions. In [141] and [149] Schneider et al. developed methods for analyzing the informational content of binding site sequence groups and individual binding site sequences, respectively. Binding sites are regions on DNA and RNA sequences to which macromolecules such as repressers, polymerases, and ribosomes, bind.

In [149] Schneider et al. analyze $E.coli$ binding site groups using two information based measures derived from the Shannon entropy, H:

- $R_{sequence}$ - A measure of the information in the binding site sequence patterns,

- $R_{frequency}$ - The amount of information needed to locate the binding site, given that the binding site occurs with a certain frequency in the genome.

where $R_{sequence}$ (measured in bits/site) describes how different the binding site sequences are from all the other genomic sequences. The value of $R_{sequence}$ should be related to the binding interaction between the macromolecules that bind to the sites and the binding sites themselves. On the other hand, $R_{frequency}$ (measured in bits/site) measures the amount of information needed to locate the binding sites (i.e. information necessary for site distinction). $R_{frequency}$ value depends on the size of the genome and on the number of binding sites in the genome. The $R_{frequency}$ of a less occurring binding site would be greater than the $R_{frequency}$ of a more prevalent binding site.

Using genetic sequences which contain known binding sites, Schneider et al. calculated $R_{sequence}$ as follows [149]: (1) aligned binding site sequences by the zero base or the first base of the initiation codon, (2) from the aligned sequences, formed the frequency table $f(b,l)$, where $f(b,l)$ represents the frequency of a base (b) at sequence position l, (3) using Shannon's entropy, Equation 2.3, and the calculated positional frequency table, they formed $H_s(L)$, the positional entropy:

$$H_s = -\sum_{B=A}^{T} f(b,L)\log_2 f(b,L) \quad [bit/base]\,, \tag{2.4}$$

The positional entropy takes on values between zero (if only one base appears) and two (if all four bases are equiprobable) bits per base, and (4) positional $R_{sequence}$ is then defined as:

$$R_{sequence}(L) = H_g - H_s(L) \quad [bit/base], \tag{2.5}$$

where $H_g = H_{genome}$ is close to two bits/base for the $E.coli$ organism used in Schneider et al.'s study.

Assuming that the frequency at each position is not influenced by the frequency at another position, $R_{sequence}$ is:

$$R_{sequence} = \sum_{L} R_{sequence}(L) \ [bit/base], \tag{2.6}$$

Results of the $R_{sequence}$ calculations can be graphically displayed using sequence logos. Sequence logos show base conservation at various positions and regions within the sequence [147]. Each position represents the conserved base or bases in bits per symbol. Hence, a completely conserved symbol would be two bits high at the position of conservation. Highly conserved sequence regions (locations with informational spikes) indicate areas of key structural contacts and regions of genetic interactions within a nucleotide sequence [149]. Figure 2.8 shows some aligned sequences and their sequence logo.

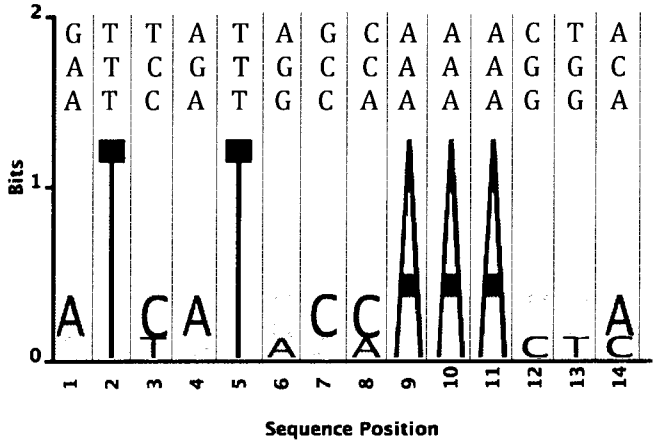


Figure 2.8. Sequences and their sequence logo

The calculation of $R_{frequency}$ is also based on entropy measures similar to those used to calculate $R_{sequence}$. The details for the calculation of $R_{frequency}$ can be found in [149]. $R_{sequence}$ and $R_{frequency}$ serve as quantitative tools for studying how proteins locate their respective binding sites among non-binding site sequences. Schneider et al. found that these two genetic measures are similar for the binding sites evaluated in their work.

Schneider et al. calculated $R_{sequence}$ and $R_{frequency}$ for the ribosome binding sites of 149 $E.coli$ sequences. The highest informational peak in $R_{frequency}$ (L) occurred at the initiation codon. The second highest peak occurred at the $Shine\text{-}Dalgarno$ site. For the ribosomal binding site, $R_{frequency} = 11.0\ bit/site$ [149]. Eleven bits translates into 5.5 bases. This suggests that a window of six bases should be sufficient in ribosome binding site recognition. The ribosome binding site contains more than six bases. Assuming the ribosome binding site is thirty bases long, if each position is equally conserved (contributes the same amount of information) then each base would contribute less than 0.5 bits of information. But, all bases are not equally conserved in the ribosome binding site. Hence the 11.0 bits of information required to distinguish a ribosome binding site are not evenly distributed. Since these 11.0 bits are not necessarily evenly distributed across the ribosomal binding site, a six-base window may be insufficient for distinction unless the ribosome has some type of "memory" mechanism.

According to Schneider et al., the genetic measure $R_{sequence}$ does not reveal anything about the physical mechanism employed by the macromolecule in binding site recognition. Yet, Schneider was able to use the informational measure $R_{sequence}$ in

recognizing individual binding sites based on their individual informational content $R_i(j)$ [141].

Using similar methods as previously described, Schneider created $R_{iw}(b,l)$, a four by L individual information weight matrix, where L is the number of binding site sequences. Using $R_{iw}(b,l)$ and the nucleotide sequence, Schneider calculated the individual information content $R_i(j)$ of the j^{th} test binding site sequence. This value was then compared to a histogram derived from sample binding sites. The expected value of the histogram is $R_{sequence}$. For instance, for $E.coli$ ribosomal binding sites, $R_i = 8.68 \pm 3.42\ bits$. This technique is used to evaluate and search for new binding sites. Schneider et al.'s information based evaluation of binding sites led to two notable discoveries [141]:

- The consensus sequence (or the most "perfect" sequence) is improbable,
- The method proves that there exists an evolutionary relationship between changes or variations of specific control points and the overall cellular control mechanism.

These two ideas indicate that the genetic translation system (most likely the genetic system as a whole) permits, if not requires, some degree of error. Therefore it must provide some method of error detection and error correction.

The next section highlights various coding theory based approaches used for genetic sequence analysis.

2.6 Coding Theory and Genetic Sequence Analysis

Battail makes a plea for increased research for the purpose of identifying the error-correcting process proposed in [17]. Though there is little known research into

error-correcting models for genetic processes (beside the work by May et al. [91], Rosen [128] Schneider's proposed coding model for molecular machines [146]), there is some research into coding theory based approaches to analyzing genetic sequences [8, 72, 160-162]. The following subsections show different areas where coding theory based approaches are used for genetic sequence analysis.

2.6.1 DNA Computing. Kari et al. used circular codes to define heuristics for constructing codewords for DNA computing applications. In DNA computing, the information storage capability of DNA is combined with laboratory techniques that manipulate the DNA to perform computations [72]. A key step in DNA computing is encoding the problem using single-stranded or double stranded DNA. The challenge is to find codewords for encoding that do not form undesirable bonds with itself or other codewords used or produced during the computational process. Kari et al. used coding theory to define rules for constructing "good" codewords for DNA computing.

2.6.2 Reading Frame Identification. Arques et al. [8] statistically analyzed the results of 12,288 autocorrelation functions of protein coding sequences. Based on the results of the autocorrelation analysis, they identified three sets of circular codes X_0, X_1, X_2 which can be used to distinguish the three possible reading frames in a protein coding sequence.

A set of codons X is a circular code, or a code without commas, if the code is able to be read in only one frame without a designated initiation signal. Crick et al. originally introduced the concept of codes without commas in the alphabet $A, C, G,$ and T. It was later successfully addressed and extracted over the alphabet R, Y, N [8]. Arques et al. successfully defines a circular code over the A, C, G, T alphabet. They were able to use the three sets of circular codes to retrieve the correct reading frame for a given protein

sequence in a thirteen base window. The three circular codes are described using a flower automaton. Arques et al. have used their coding based model to analyze the Kozak's scanning mechanism for eukaryotic translation initiation and other models of translation [8].

2.6.3 Genetic Code Analysis. Stambuk also explored circular coding properties of nucleic acid sequences [159, 161]. His approach was based on the combinatorial necklace model which asks: "How many different necklaces of length m can be made from a bead of q given colors [9, 18, 161]." Using $q = [A, C, G, T]$ and $q = [R = Purine;\ Y = Pyrimidine;\ N = R\ or\ Y]$, Stambuk applied the necklace model to genetic sequence analysis, enabling the use of coding theory arithmetic in genetic code analysis [161].

Though Stambuk did not use error control coding in his analysis, his work provided important insight into the structure of DNA sequences:

- Non-protein coding DNA contains properties corresponding to natural language; protein coding DNA has properties that are characteristic of coded language structures.
- A binary nucleotide mapping and a corresponding Gray code mapping can be defined based on chemical properties of the bases. Binary mapping incorporates complementarity of DNA and Gray code mapping ensures error minimization during the translation and transcription process.
- The Hamming distance measure can be used to express the difference between different codon and different amino acid positions.

CHAPTER 3

BIOLOGICAL BACKGROUND

This chapter provides the biological background necessary to understand the communication theory models in later chapters. In Section 3.1, the DNA is described as a digital signal. Section 3.2 deals with basic terms and definitions. This includes the structure of DNA and RNA, genes, mutations as well as the organization of genomes of bacteria (prokaryotes). In Section 3.3, a detailed description of gene expression, the process of protein synthesis, is presented. The main steps and involved components are presented for prokaryotic organisms.

3.1 The DNA as a Digital Signal

The deoxyribonucleic acid (DNA) is the primary carrier of genetic information, which can be seen as a digital signal of the quaternary alphabet $\mathcal{A} = \{A, G, C, T\}$. In digital data transmission, information is processed in numerous steps: it is read out, transformed (modulation, coding), transmitted, possibly altered by transmission errors, corrected and interpreted. The genetic information stored in the DNA comes into effect only after transformation into proteins, molecules that determine many genetic traits of living beings. This process takes place in a series of transformation steps similar to those in digital data transmission: parts of the DNA sequence are read out, transformed into different alphabets, possibly altered by mutations and corrected. In this chapter, the basic steps underlying protein synthesis are detailed.

3.2 Terms and Definition

3.2.1 *DNA* and *RNA*. The *DNA* is formed by two strands, linked together and twisted in the shape of a double helix. The strands consist of a chain of nucleotides, small molecules built up by a nucleobase, a pentose sugar and a phosphate. The nucleobase can be of four types: (A), $cytosine$ (C), $guanine$ (G) and $thymine$ (T). The larger nucleotides adenine and guanine belong to the class of a double-ringed chemical structure called purine. They form hydrogen bonds with their respective complements thymine and cytosine, belonging to the single-ringed pyrimidines. Two nucleotides on opposite complementary *DNA* strands that are connected via hydrogen bonds are called a base pair (in the following abbreviated as bp). Base-pairing in the *DNA* can exclusively occur between A and T as well as between G and C [75]. Other bonds are unfavorable since the patterns of hydrogen acceptors and donors do not match: While adenine and thymine bind via two hydrogen bonds, cytosine and guanine are connected via three hydrogen bonds (see Figure 3.1). The binding between the two strands is called Watson Crick base-pairing. A *DNA* sequence is typically written from its 5'-end to its 3'-end, where the naming originates from the chemical structure of the pentose sugar. With respect to this directionality of the *DNA*, the relative position of a sequence element is either denoted as upstream (towards the 5'-end) or downstream (towards the 3'-end).

During protein synthesis (gene expression), parts of the *DNA* are transformed into *RNA* (ribonucleic acid) in the process of transcription (see Section 3.3). The difference to *DNA* lies in the chemical structure: *RNA* is single-stranded and uracil replaces thymine as the base complement to adenine. Since *RNA* is single-stranded, it often contains short sequences of nucleotides that can base-pair with complementary sequences found

somewhere else on the same molecule [4]. These interactions arrange for an *RNA* molecule to fold into a stable three-dimensional structure, which allows it to play regulatory roles during protein synthesis.

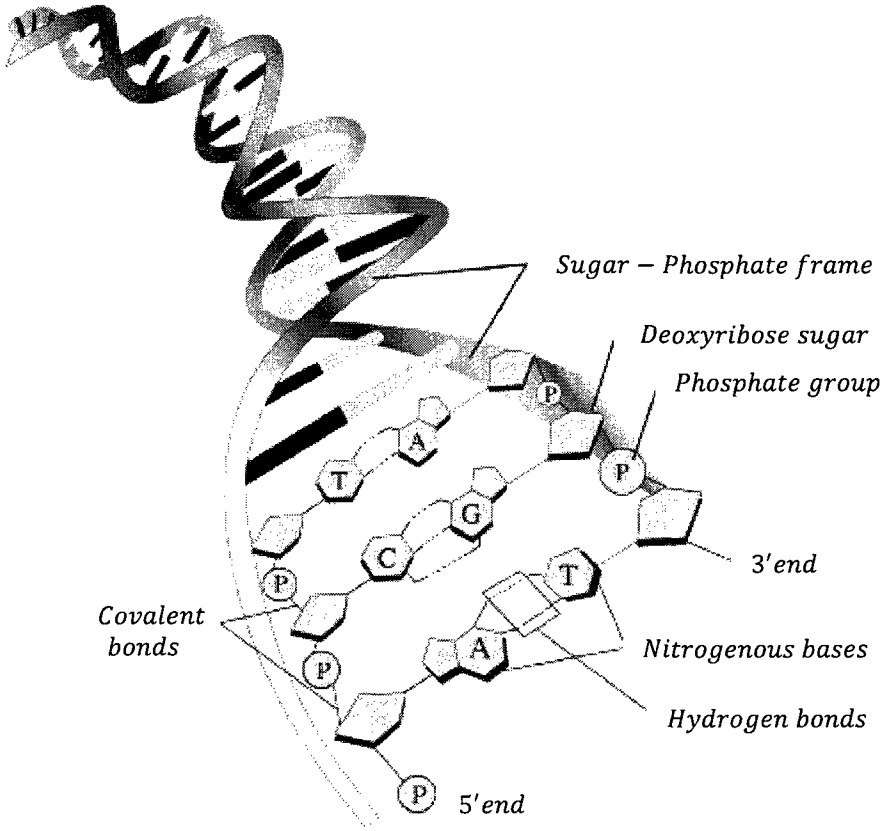


Figure 3.1. Structure of *DNA*[3]

3.2.2 Mutations. Mutations are changes happening to the nucleotide sequence of genetic material (*DNA or RNA*). They correspond to transmission errors and noise in communication systems that alter the signal during transmission. Mutations occur due to both external influences (like radiation) and errors during cell processes (like replication). Three basic types of small scale mutations exist [57] :

[3] Figure 2.1 shows the structure DNA (http://img.sparknotes.com/101s/biology/6 – 1.jpg)

- Point mutation: a single nucleotide is exchanged for another one. The most common point mutation exchanges a purine for a purine ($A \leftrightarrow G$) or a pyrimidine fora pyrimidine ($C \leftrightarrow T$). This type of mutation is called a transition as opposed to a transversion which exchanges a purine for a pyrimidine or vice versa ($A/G \leftrightarrow C/T$).
- Deletion: one or several nucleotides are deleted irreversibly from their position in the DNA.
- Insertion: one or several nucleotides are inserted at a random position in the DNA.

The effect of mutations depends on their position and on whether they affect the synthesis of a protein. In the majority of cases, the effect is either neutral or can be corrected before bringing harm to the organism. In some cases, mutations are harmful or – in rare cases –beneficial and, thus, lead to an evolutionary change through positive or negative selection.

3.2.3 Genes and Protein. A gene is a sequence of nucleic acids containing the information for a functional product (usually a protein). Proteins (Greek $prota$ = 'of primary importance') are large organic compounds that constitute an essential part of all living beings [75]. They are responsible for oxygen transport, cell signaling, catalysis of biochemical reactions, immune response as well as maintaining the cell scaffold. While it was long believed that one gene codes for one protein which is itself responsible for one trait, many exceptions to this rule have been found until today [113]. Nonetheless, in the following, the term gene is used in its traditional definition as those parts of the DNA that are copied into $mRNA$ in the process of transcription (see Section 3.3).

3.2.4 Genome. The term genome refers to the entire genetic information or hereditary material possessed by an organism, i.e. the entirety of genes and extra-genic DNA [4]. The latter refers to those parts of the DNA that are not transformed into mRNA during gene expression. The organization of the genome depends on the organism: While simple organisms carry only a single chromosome organized as a ring, the majority of higher organisms has between 8 and 100 chromosomes organized in an X-shape [75]. In addition to the organization, the length of the genome varies significantly between organisms (bacterium Carsonella ruddii: 1.6×10^5 bases, human: 3×10^9 bases).

3.2.5 Prokaryotic and Eukaryotic Organisms. Organisms are classified into two basic families, namely *prokaryotes* and *eukaryotes* (Greek *pro* = 'before', *eu* = 'true', *karyon* = 'kernel'). *Prokaryotes* comprise all organisms from the families archaea and bacteria. Prokaryotic organisms are in most cases unicellular, and their cells have no cell nucleus, i.e. the genetic material is not membrane-bound but freely floating in the cytoplasm. The *DNA* of prokaryotes consists of one single circular chromosome localized in an area called nucleoid. The single chromosome is densely packed with genes (typically several thousand [169]), only few percent are non-coding and serve regulatory purposes. Research on *prokaryotes* strongly focuses on the bacterium *Escherichia coli* (*E.coli*), which infects the lower intestines of mammals. *Eukaryotes* comprise all unicellular and multicellular organisms whose cells contain a cell nucleus. The genetic information is stored in chromosomes localized inside the membrane-bound nucleus which is surrounded by the cytoplasm. In contrast to prokaryotes, the chromosomes in *eukaryotes* contain a high percentage ($> 90\%$) of *DNA* not coding for proteins. *Eukaryotes* comprise all higher organisms like plants and animals. The best studied

eukaryotes are the human (*Homo sapiens*), the thale cress (*Arabidopsis thaliana*), the fruit fly (*Drosphila melanogaster*) as well as the yeast species *Sacharomyces cerevisiae*.

3.3 Gene Expression

3.3.1 Overview. Gene expression is the process in which the information stored in the ***DNA*** is used to synthesize proteins. It takes place in two basic steps:

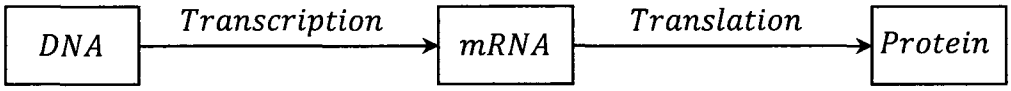


Figure 3.2. Gene expression

During the process of transcription, the double stranded *DNA* serves as a template to synthesize the single stranded *mRNA* (messenger *RNA*, see Figure 3.3, middle). In the second step of gene expression (translation), this *mRNA* is translated into proteins by chaining amino acids (see Figure 3.3, bottom).

The main differences in the gene expression of prokaryotes and eukaryotes are:

- In eukaryotic organisms, the *mRNA* only contains the information to translate one protein (monocistronic *mRNA*), but it may as well contain the information for several proteins in prokaryotes (polycistronic *mRNA*).
- After transcription, the *mRNA* of eukaryotes consists of coding regions (exons) and non-coding regions (introns) separating the exons. In the process of splicing, the introns are cut out of the *mRNA* yielding the so-called mature *mRNA*.
- In contrast to prokaryotes, translation and transcription of eukaryotes are locally separated. Transcription and splicing take place inside the nucleus membrane,

whereas translation takes place in the cytoplasm surrounding the nucleus membrane. Therefore, the mature *mRNA* is exposed to additional radiations and thermal noise during its travel to the less protected cytoplasm.

- Due to the missing separation of transcription and translation in prokaryotic cells, no intermediate step lies between transcription and translation, which allows simultaneous processing, i.e. the *mRNA* can already be translated while still being synthesized through transcription.

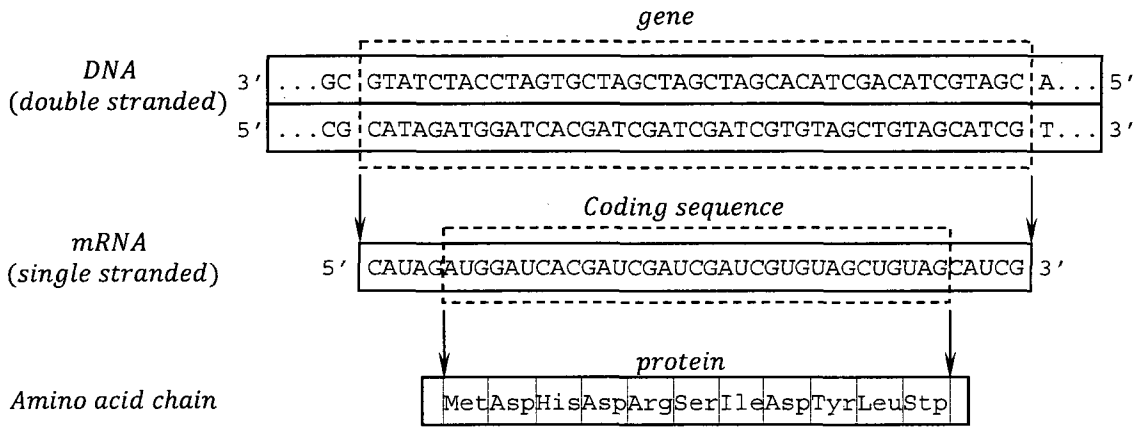


Figure 3.3: Illustration of sequence transformations during gene expression

3.3.2 Prokaryotic Transcription. During transcription, a part of the *DNA* (the gene) is copied into mRNA (see Figure 3.3). This step is performed by the macromolecule RNA polymerase (*RNAP*) and its sigma subunit which first randomly binds to the DNA and move along [75]. Equivalent to frame synchronization in continuous transmission, a short DNA motif (the so-called promoter) informs the *RNA polymerase* about the upcoming gene start. After the sigma factor recognizes the promoter, the *RNA* polymerase opens and unwinds the *DNA* (also called *DNA* melting) on a range of around 12 base pairs to enable the copying of one strand [81]. The sigma factor does not play a role in this copying process: in around 30 % of the cases, it dissociates from the

RNA polymerase directly after initiation, while it otherwise dissociates at random points during transcription [56]. During transcription elongation, the RNA polymerase moves along the DNA, opens the double helix and copies one strand (the so-called coding strand) by building the complement of the template strand (see Figure 3.4). Termination of transcription is either induced by an RNA-binding protein or based on sequences in the RNA that fold into hairpin structures that mechanically interrupt transcription [4]. After dissociating from the DNA, the RNA polymerase can bind to another sigma factor and restart the process.

3.3.3 Prokaryotic Translation. During translation, the $mRNA$ is transformed into a protein. This step is performed by the ribosome, a complex of two large subunits that are themselves made up of protein subunits as well as $(ribosomal\ RNAs)$. The larger subunit is denoted as $50S$ subunit, the smaller one as $30S$ subunit, together building the $70S$ ribosome. It is important to note that not the complete $mRNA$ is translated into a protein but only the so called coding sequence, which is delimited by the start codon $AUG\ or\ GUG\ or\ UUG$ and one of the stop codons UAA, UAG or UGA (a codon is a nucleotide triplet).

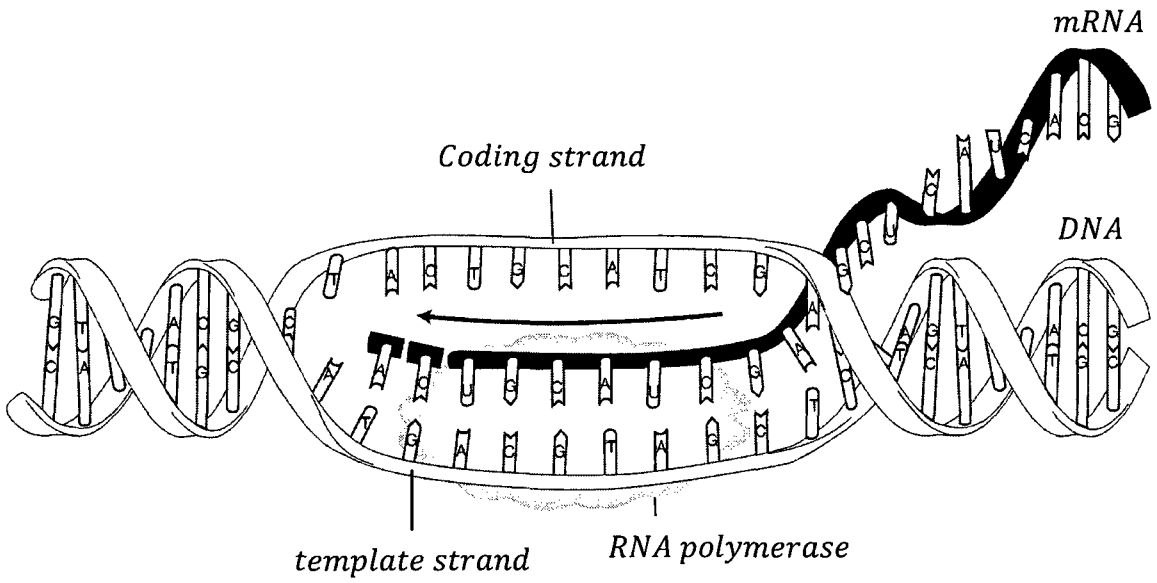


Figure 3.4. Transcription by *RNA* polymerase

Initiation. During initiation, a composite structure of proteins and ribosomal *RNA* (*rRNA*) forms and binds to initiator region of the *mRNA* (the so-called 5′ untranslated region or 5′-*UTR*, see Figure 3.5) to form the ribosome. The length of the 5′-*UTR* varies between 0 and 920 *bp* with the mean length being around 100 *bp* [25, 157]. The ribosome aids protein production in various critical ways, and is considered to be the translation "factory." One portion of the ribosome, called the *16S rRNA*[4] (in prokaryotes - the corresponding eukaryotic structure is the 18*S rRNA*) likely makes the initial contact with the *mRNA* (see Figure 3.6). The 3′ end of the 16S has a high potential for base pairing with *mRNA* close to the region to be translated. The 16*S rRNA* is part of a larger complex called the 30*S* subunit (or small subunit) which covers a space of approximately ten *mRNA* codons. Two of these codons are critically important. *The* 30*S* must cover the

[4] Figure 2.5 shows the structure of the *. coli* 16 *S rRNA*. Structure publically available from the Ribosomal Database Project at Michigan State University (*http://rdp.cme.msu.edu/*)

first codon to represent an amino acid in a special location called the *donor* or *P* site. This first codon is called the *start codon* or the *initiation codon*. Most often, the start codon is "*AUG*", though in *E.coli*, for example, "*GUG*" and "*UUG*" start codons are known to exist. In bacteria, the start codon codes for a special amino acid, formylated methionine. A special transfer *RNA* (*tRNA*) carries the formylated methionine to the *P* site of the 30*S*. The *tRNA* base pairs with the initiation codon at the *P* site, allowing the 50*S* subunit to join the 30*S* and complete the ribosome.

After binding to the 5'-*UTR*, the 30*S* subunit moves rapidly along the *mRNA* until it detects the start codon (*AUG* , position +1) and the *Shine - Dalgarno* sequence (*SD*), a hexamer located shortly before the coding sequence. The 16*S rRNA* is the part of the 30*S* subunit of the ribosome that is responsible for the detection of the *Shine-Dalgarno* sequence via base-pairing [163]. In the communications engineering sense, the *Shine - Dalgarno* sequence corresponds to the synchronization word of translation that needs to be detected by the 16*S rRNA*, the synchronization unit of the ribosome.

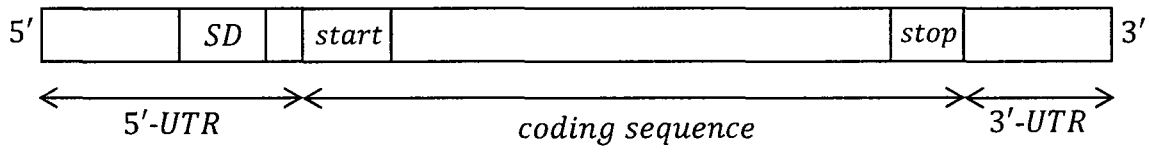


Figure 3.5. *mRNA* structure in prokaryotes

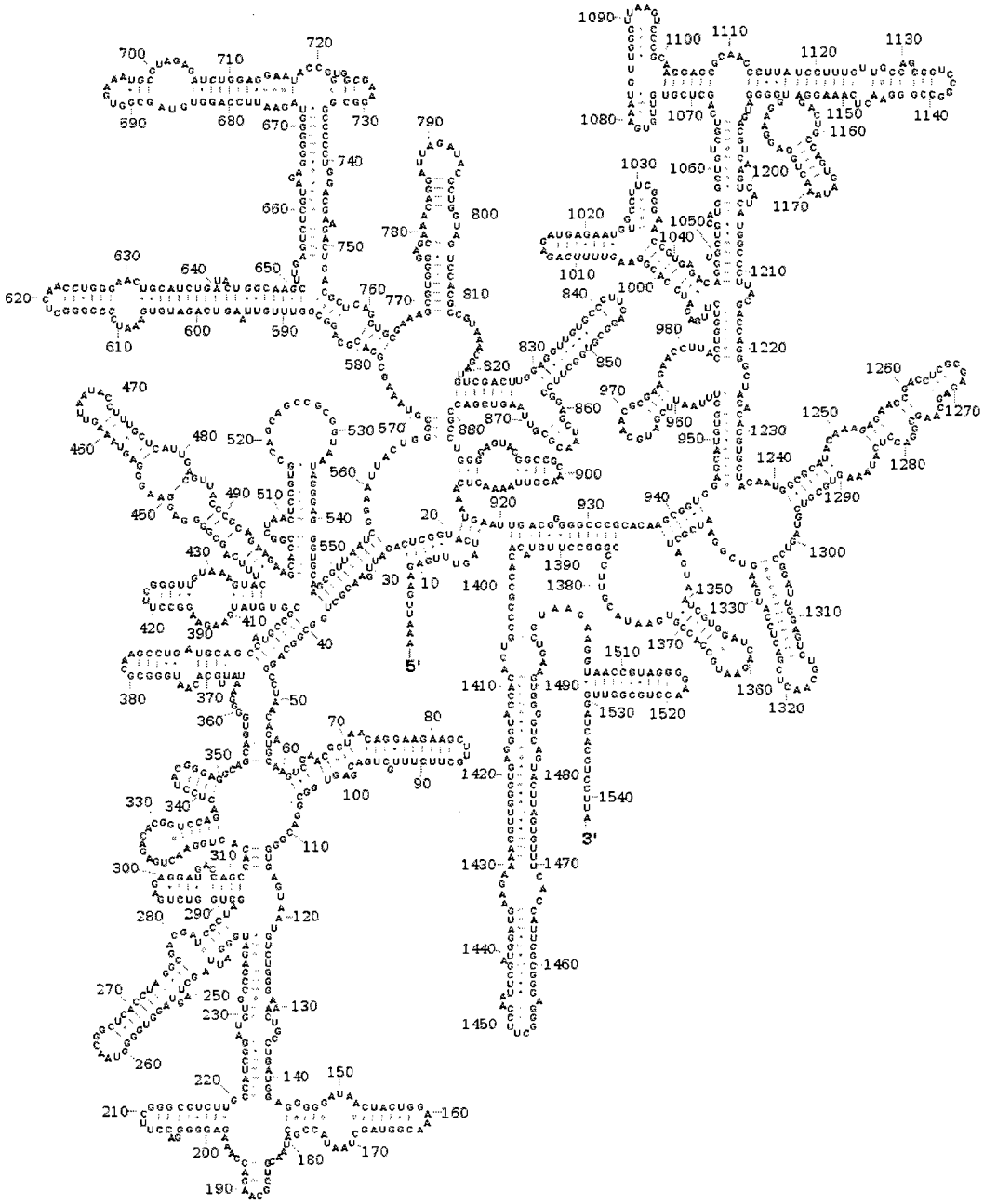


Figure 3.6. Sequence and structure of the *Escherichia coli* 16*S rRNA*

Elongation. Adjoining the *P* site in the ribosome is the *A* site, which covers the next codon in the sequence. Depending on the codon at the *A* site, a specific *tRNA* will bind there, carrying with it the amino acid associated with that *tRNA*. The *tRNA* triplet binding to the codon is called the *anticodon* (see Figure 3.7, left). In order to easily see

the complementary relation to codons, anticodons are written 3'-5'. For clarity, they will be written with a left arrow. Thus, the codon ACG binds perfectly to the anticodon $U\tilde{G}C$.

This relation between the codon, the $tRNA$, and the acid it carries is called the genetic code (see Figure 3.7, right), as the triplet codes uniquely for the next specific amino acid in the protein under construction. However, a single amino acid may be coded by as many as six different codons. The multiplicity leads naturally to the question, "Is there one type of $tRNA$ for each codon, or one $tRNA$ for each amino acid?" According to the wobble hypothesis, the answer lies in between. The wobble hypothesis states that the rules for base pairing change for the third position of a codon (and so the first position of an anticodon). The standard base pairs are summarized in Table 3.1. The primary difference in the "wobble rules" relative to standard base pairing is the pairing of *guanine* and *uracil*. Note the table contains an entry for *inosine* (I), one of many "modified bases." In the case of *inosine*, a *guanine* is modified to permit an additional pairing with *cytosine*. Additionally, a modification of *uracil* to produce 2-thiouracil prevents the nonstandard pairing with *guanine*. All these additional rules help reduce the number of different $tRNAs$ needed to carry the proper amino acids.

Next, the protein at the P site (initially, the single amino acid of the start codon) bonds to the new amino acid at the A site to create a protein one acid longer. Finally, the ribosome shifts by three nucleotides so that the empty $tRNA$ from the P site (the uncharged $tRNA$) is expelled from the ribosome, the $tRNA$ carrying the protein now sits in the P site, and the A site is now ready for a new $tRNA$ charged with the amino acid corresponding to the new codon at the A site. This process of shifting is called translocation, and the overall process of lengthening the protein is called elongation.

Table 3.1. Wobble Base Pairings

tRNA 5' anticodon base	*mRNA 3' codon base*
A	*U*
C	*G*
G	*C or U*
U	*A or G*
I	*A or C or U*

Because the ribosome "sees" *mRNA* in groups of three nucleotides, and shifts by three, there is said to be a frame in which the ribosome "reads." As seen Figure 3.7, there are three frames by which the ribosome can read codons. Note that the three sequences are the same, merely spaced differently to identify the three possible readings by codon.

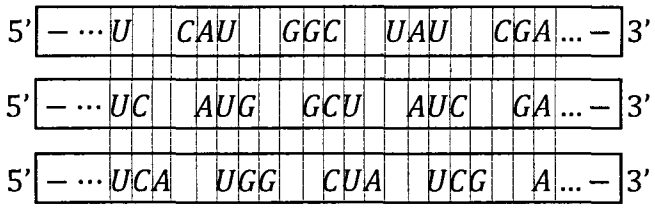


Figure 3.7. The three possible reading frames

Because it is "in synchronization" with the start codon, the second frame is the one read by the ribosome. The top one lags the start codon by one base, so it is called the -1 frame, and the bottom one is appropriately called the $+1$ frame. Occasionally, translocation does not take place exactly as expected. Instead of shifting by three bases, the ribosome may only move two, or move four. When the ribosome translocates by anything other than a multiple of three bases, there is a change in frame. Thus, such an event is called a frameshift. Frameshifts are not always errors in translocation. Sometimes frameshifts are regular events at specific locations along the *mRNA*, and may be used as a biological control mechanism. [45, 75]

Because of the ribosome travels along the $mRNA$ in the $5' - 3'$ direction, areas closer to the $5'$ end are described as upstream while those nearer to the $3'$ end are called downstream. The initiation region where the $16S\ rRNA$ binds, then, is said to be upstream from the start codon, and the rest of the series to be coded is downstream from the start.

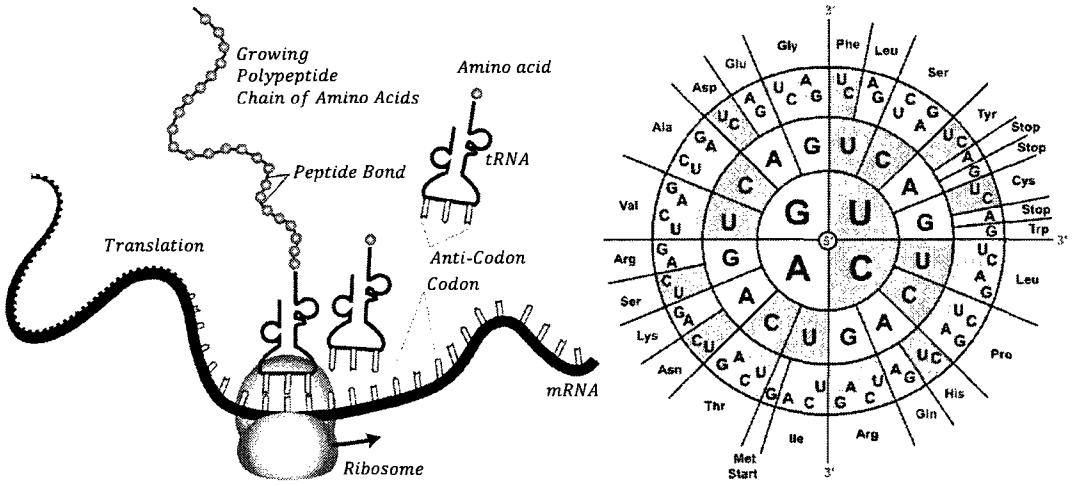


Figure 3.8. The $tRNAs$ map the codons in the $mRNA$ to an amino acid

Termination. The ribosome continues elongating downstream until one of three codons enters into the A site. There are no $tRNAs$ that bind with the codons "UAA," "UAG," or "UGA".Consequently, these nonsense codons stall the protein production. The protein is detached from the $tRNA$ at the P site, and the ribosome breaks back down into the $30S$ and $50S$ subunits.

CHAPTER 4

ANALYSIS OF GENE TRANSLATION USING A COMMUNICATIONS THEORY APPROACH

As evidenced by many articles, researchers are increasingly curious about the communication protocols of molecular systems. In this special area we endeavor to explore ideas at the crossroads of communications theory and molecular biology from various disciplinary backgrounds and vantage points, providing an overview of the state of research and making compelling observations regarding the nature of biological information transmission in light of the principles of communications and coding theory. This chapter deals with modeling the process of translation in gene expression (information contained in the DNA molecule when transformed into proteins). Gene expression involves two main stages. The first one is transcription (related to coding theory) where the information stored in the DNA that has been contaminated with genetic noise is transformed into the $messenger\ RNA$ ($mRNA$). The second one is translation (related to detection theory), where the noisy $mRNA$ molecule serves as an instructive for protein synthesis. The accuracy of this process is vital to the survival of the organisms.

This chapter is organized as follows: Section 4.1 gives a theoretical background establishing an analogy between gene expression and communications engineering. In Section 4.2 we describe our proposed codebook model for the process of translation in gene expression with an algorithm to optimize its detection mechanism. A mutation analysis is being carried out to test the proposed model and certify its correctness and biological relevance. Section 4.3 introduces four different metrics that quantify the

mechanism that the ribosome uses to detect the translational signal. These metrics are based on communication theory concepts. Simulation results and discussion are given.

4.1 Gene Expression as a Communications System

Gene expression is the translation of information encoded in a gene into protein or RNA. It takes place in two basic steps: transcription and translation (see Figure 4.1).

During transcription, a portion of the genomic DNA is copied into RNA ($mRNA$) except that the base T is substituted by U. For protein-coding genes, this RNA is eventually translated into a chain of amino acids that forms a protein according to the mapping rule described by the genetic code [63]. In prokaryotes, the RNA is essentially competent to do this immediately; however in eukaryotes, there is an intermediate step in which the message is processed into a mature $mRNA$ by an editing process, itself dependent on an additional layer of sequences. At all of these stages, regulatory signals need to operate. Once the $mRNA$ is produced, these messages are then interpreted by the cellular machinery (ribosome, etc.) to produce desired effects (the construction of new proteins). On the other hand, there is a large subset of genes that act only at the RNA level, and they have their own signals, such as RNA structural signals (hairpins, etc) or homology to other protein encoding genes that they regulate.

Analyzing DNA processing in gene expression, many similarities with the way engineers transmit digital information in communication systems can be noticed. The DNA can be modeled as an encoded information source that is decoded (processed) in several steps to produce proteins. During these decoding steps, the processed DNA is subjected to genetic noise which results in several types of mutations.

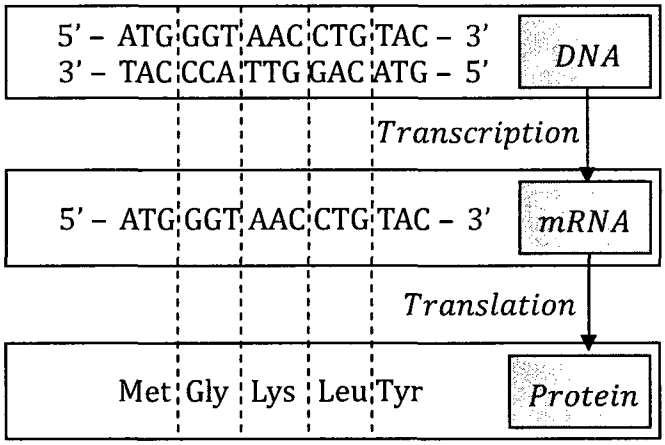


Figure 4.1. Protein Synthesis (Gene Expression)

All approaches of frame synchronization are based on a correlator that compares the known synchronization word with the incoming data stream for each position and decides for the position with the highest correlation. During transcription initiation the *RNA* polymerase and its *sigma* subunit need to recognize two promoter regions that indicate the beginning of the gene. Hence, the process of transcription can be considered as frame synchronization with two synchronization words where the *RNA* polymerase detects the promoter sequences (biological synchronization words). Since the distance between two consecutive promoters varies, it precisely corresponds to the detection of aperiodically inserted synchronization words. The sliding process of the protein along the *DNA* represents the correlation taking place in technical communication systems Translation initiation also corresponds to a process of frame synchronization to detect the translation initiation signals (e.g. for prokaryotes this includes the *Shine-Dalgarno* sequence and the start codon). This is followed by a decoding process to map codons to amino acids. Figure 4.2 shows a model for gene expression based on building blocks from communications theory. In this model, we assume that mutations can also occur in

the involved proteins, i.e. RNA polymerase, ribosome, and $tRNA$. Other similar models for gene expression are summarized in [97].

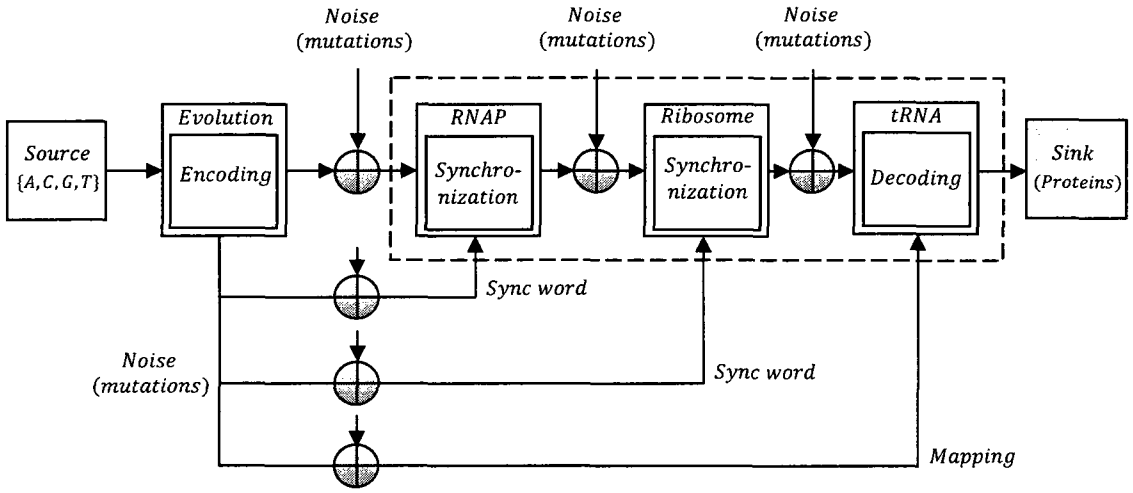


Figure 4.2. Communication theory model for gene expression

The following section describes a coding theory based model of the process of translation in gene expression. A variable length codebook, an energy table, and a specially designed metric are used to analyze the mechanism that the ribosome uses to decode the $mRNA$ sequence. The algorithm developed to optimize the resolution of the detected translational signals is described. Mutations in the ribosome that affect the level of protein synthesis are investigated and results are shown.

4.2 Coding Theory Based Modeling

Our initial work in this chapter has been to validate the work done by Z. Dawy, F. Gonzalez, Joachim Hagenauer and Jacob Mueller, in their paper "Modeling and Analysis of Gene expression Mechanism: A communication Theory Approach" [30]. This work deals with modeling gene expression during which information contained in the DNA molecule is transformed into proteins. These protein products are later used for different

processes in the living system. The accuracy of this process is vital to the survival of the organisms which make modeling such a process something worthwhile.

The model developed uses the assumption that the ribosome decodes the $mRNA$ sequences using the $3'$ end of the $16S\ rRNA$ molecule as a one-dimensional embedded codebook. The ribosome decodes the $mRNA$ leader region in order to detect a signal that will enable the ribosome to start translation. If it does not find the signal then the ribosome keeps scanning the $mRNA$. Once the signal is found, the ribosome will try to start translation. When the synchronization signal (AUG) is found, the process of elongation starts where the ribosome uses the genetic code to start de-mapping (demodulating) the triplets (codons) in the $mRNA$ sequence into a chain of amino acids. In the last stage of the model, the ribosome recognizes one of the stop codons (UAA, UAG, UGA) and then the protein production complex is liberated from the $mRNA$ resulting in the production of the protein. In the sequel, we concentrate on how the ribosome "decodes" the leader region of the $mRNA$ in order to get the signals to start translation.

Figure 4.3 shows a typical $mRNA$ sequence [30]. It is assumed the ribosome binds in the leader region of the $mRNA$ sequence. The leader region is formed by the bases upstream of the initiation codon. These codons, typically AUG, GUG or UUG, are in the start of a coding region that is the part of the $mRNA$ that will translate to a protein.

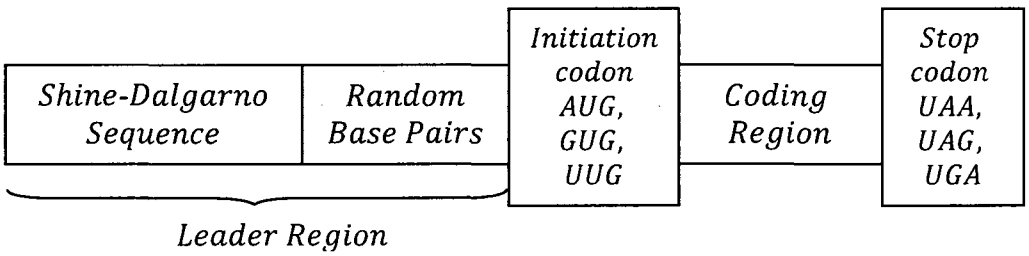


Figure 4.3. $mRNA$ Sequence

In Figure 4.4, a simplified version of the model shown in Figure 4.2 is presented. In this block diagram, an unknown source produces the information in the *DNA* message. A channel encoding process creates the structure of bases of the *DNA* sequence. Once the *DNA* is released (start of the transcription stage), mutations in the sequence are produced by adding noise. During transcription, the *DNA* sequence is decoded to produce the *mRNA* sequence. The *mRNA* sequence is thought to be a decoder output because it is shorter than the original *DNA* sequence, thus, some redundancy is removed. The resulting *mRNA* contains only the exons or protein coding regions (message) whereas the introns (redundancy) are removed. Continuing with the process the *mRNA* molecule is again exposed to noise and radiations, especially when it travels outside the nuclear membrane in eukaryotic organisms. Once the *mRNA* reaches the ribosome, a second decoding process takes place. Here, the ribosome will interact with the *mRNA* sequence to start the protein synthesis. The protein output of this model is the final recovered message.

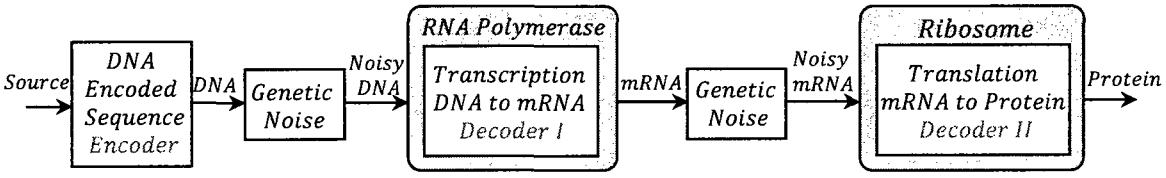


Figure 4.4. Transcription and Translation as a Communication Model

As this work focuses on the modeling of translation in gene expression, it is important to mention that the mechanism of translation is different for prokaryotes and eukaryotes [69]. We are interested in how translation initiates in prokaryotes, more specifically in the organism *Escherichia coli*. Translation involves the chaining of amino acids which form proteins. In order to do this, the ribosome binds to the messenger

RNA to create a closed complex. The ribosome is able to "scan" the *mRNA* in the search for sequences that contain a sign to start translation.

The biological consistency of the developed model is proven in detecting the *Shine-Dalgarno* sequence and the initiation and stop codons for translation initiation and termination. Results obtained using this model has been also compared with published experimental results for different mutations of the *rRNA* molecule. Total agreement between both sets of results proves the validity of the proposed model and show the relevance of communication theory based models for genetic regulatory systems (Section 4.2.3).

4.2.1 Codebook Model. In order to start translation, the last 13 bases of the molecule *16S rRNA* inside the small subunit of the ribosome interact with the leader region of the *mRNA* [163]. This interaction is based on the hydrogen bonding between nucleotide bases in the genetic sequences. This principle allows the bonding of *cytosine* (*C*) with *guanine* (*G*) and *adenine* (*A*) with *thymine* (*T*) (or *uracil* (*U*) in case of *RNA*). This chemical interaction permits the recognition of signals between *DNA* and *RNA* molecules. As in other translation initiation models [93], the coding model presented in this work is based on the structure of the *16S rRNA* and more specifically on its 3' end.

A common assumption made in biology is that the ribosome recognition of the initiation signal in the leader region is achieved when the so called *Shine-Dalgarno* sequence is detected [156]. A consensus sequence of the *Shine-Dalgarno* (*SD*) is *AGGAGG*. It is obtained by calculating the most frequent nucleotide for each position of aligned *mRNA* sequences. In the case of the *SD* sequence, the base *A* in the first position is the base with the highest frequency among all aligned sequences. The approach of

assigning the recognition of binding sites just to one sequence is in general wrong as it discards all the variability of the sequences and introduces "hard" decisions that incur a loss of information [138]. In order to improve the accuracy of the model, a more flexible approach for the translation initiation mechanism is needed. The approach taken in this work is the use of error correction coding theory.

In our work [103] and [99] we have modified Dawy et al.'s model for this detection/recognition system by designing a one-dimensional variable-length codebook (see Table 4.1) and an exponentially weighted metric (Section 4.2.2). The codebook uses a variable codeword length N between 2 and 13 using the Watson-Crick complement of the last 13 bases sequence of the 3' end of the $16S\ rRNA$ molecule. Hence, we obtain $(13 - N + 1)$ codewords each of which has a length N bases and has the form $C_i = (s_i, s_{i+1}, s_{i+2}, \dots, s_{i+N-1})$; where $i \in [1,13 - N + 1]$ and s_i corresponds to the i^{th} base in the Watson-Crick complement of the last 13 bases sequence of the 3' end of the $16S\ rRNA$ molecule given by (UAAGGAGGUGAUC). The model is based on the assumption that the ribosome uses this highly conserved sequence (among prokaryotes) in the $16S\ rRNA$ molecule to do the comparisons needed to detect the initiation signals (*Shine-Dalgarno* signal and initiation codon signal).

To build the codebook, an example of codeword length of $N = 5$ is assumed and the codewords are obtained by taking a sliding window through the Watson-Crick complement of the sequence of the last 13 bases (see Figure 4.5). Having $N = 5$ and a length of 13 bases, the codebook is obtained via a sliding window operation which results in 9 codewords. The codewords are taken from the complement because those are the words that will be found in the $mRNA$ molecule.

The input to our proposed model is the noisy $mRNA$ (due to to the addition of genetic noise) and the last 13 bases of the 3' end of the $16S\ rRNA$ molecule. This sequence interacts with the leader region of the $mRNA$ to start translation. This sequence (UAAGGAGGUGAUC) and the resulting codebook for a value $N = 5$ are shown in Figure 4.5 and Table 4.1 (notice the SD sequence is AGGAGG).

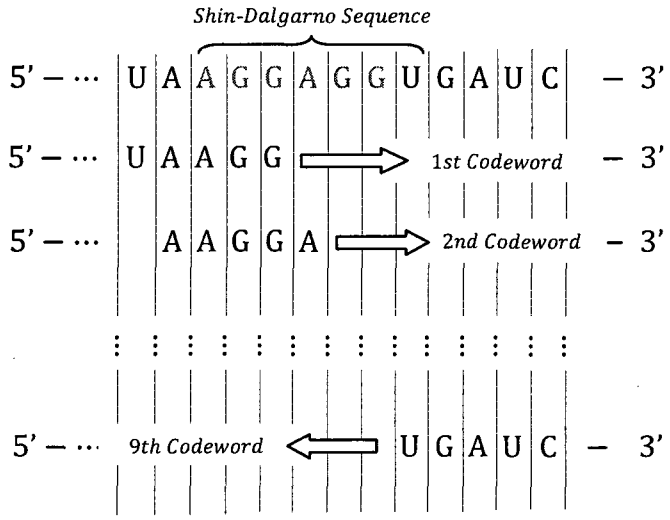


Figure 4.5. Codebook Structure Length $N = 5$

Table 4.1. Codebook

C_l	**Codeword**
C_1	$UAAGG$
C_2	$AAGGA$
C_3	$AGGAG$
C_4	$GGAGG$
C_5	$GAGGU$
C_6	$AGGUG$
C_7	$GGUGA$
C_8	$GUGAU$
C_9	$UGAUC$

Table 4.2.
Energy doublets

Pairs of bases Energy			
AA	-0.9	GA	-2.3
AU	-0.9	GU	-2.1
UA	-1.1	CA	-1.8
UU	-0.9	CU	-1.7
AG	-2.3	GG	-2.9
AC	-1.8	GC	-3.4
UG	-2.1	CG	-3.4
UC	-1.7	CC	-2.9

A sliding window of size N applies to the received noisy $mRNA$ sequence to select subsequences of length N and match them with the codewords in the codebook.

The codeword that results in a minimum weighted free energy exponential metric between doublets (pair of bases) is selected as the correct codeword and the metric value is saved. Biologically, the ribosome achieves this by means of the complementary principle. The energies involved in the $rRNA - mRNA$ interaction tell the ribosome when a signal is detected and, thus, when the start of the process of translation should take place. In our model, a modified version of the method of free energy doublets presented in [131] is adopted to calculate an exponentially weighted $free\ energy$ distance metric in $kcal/mol$. Table 4.2 shows the $free\ energy$ associated with each doplet of bases when the ribosome scans the $mRNA$ sequence looking for the initiation signals. The minimum energies are evaluated and plotted to determine the performance of the proposed algorithm.

To test Dawy's model, and obtain the results shown in Figure 4.7 and 4.8, we proceeded as follows:

1. We obtained the complete genome of the prokaryotic bacteria $E.coli$ strain $MG1655$
2. The $mRNA$ sequence was obtained by replacing the nitrogenous base "$Thymine$" with "$Uracil$". i.e., replacing "T" with "U"
3. We located and identified all genes in the given $mRNA$ sequence by running a searching algorithm developed for this purpose. Start (AUG, GUG, UUG) and stop (UAA, UAG, UGA) positions for each gene were obtained and saved
4. The consensus $Shine\ -\ Dalgarno$ signal ("$AGGAGG$" or "$AGGA$" or "$GGAG$" or "$GAGG$") was located in the non-coding regions
5. We implemented the proposed Free Energy Ribosome Decoding algorithm using a large number of sequences, and the average was calculated. For presentation

purposes, all the tested sequences chosen for analysis follows the structure shown in Figure 4.6.

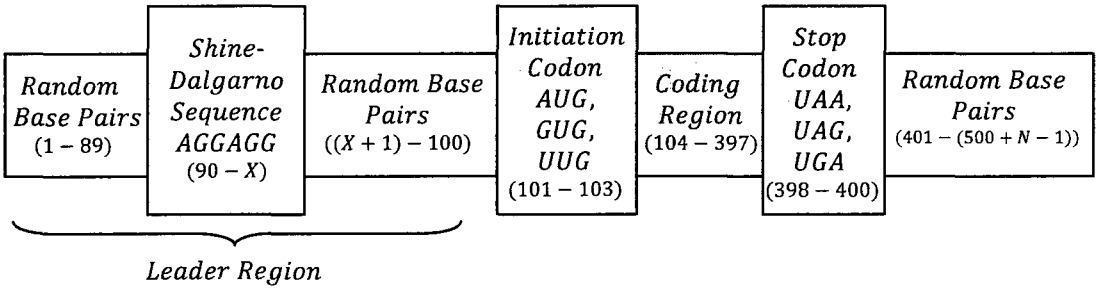


Figure 4.6. Test sequences structure

Where X represents the position of the last G of the *Shine-Dalgarno* sequence in the above sequence structure (i.e. $90 + SD$ length). N is the codeword length used to design the codebook.

Figures 4.8 and 4.9 are the result of applying the model proposed by Dawy [30] before using our weighting algorithm presented in section 4.2.2. Figure 4.8 and 4.9 are obtained for $E.coli\ MG1655$ and $O157{:}H7$, respectively.

In the search for biological accuracy, the proposed model is tested for the whole process of translation to seek insights of how the ribosome performs during the rest of the stages of translation: elongation and termination. By doing so, important observations are obtained on the behavior of the gene expression mechanism. For the analysis, sequences of coding regions greater or equal than 500 base pairs are taken from the complete genome of $MG1655\ E.coli$ strain and $O157{:}H7\ E.coli$ strain. In addition, sequences of the same length taken arbitrarily (non-coding regions) from the genome are used for comparison. The algorithm is applied to these sequences and the average results are plotted in Figure 4.7 and 4.8.

The x-axis represents the position in the selected and aligned sequences. For presentation purposes, the positions of the initiation and termination codons for all coding sequences are fixed at 101 and 398, respectively, thus, only 294 nucleotides from the coding region are kept while the others are removed. The y-axis represents the calculated average free energy measure in $kcal/mol$ for each position in the $mRNA$. The results found are significant. The $Shine-Dalgarno$ and the initiation codon signals are identified in the translated sequences while they are not in the arbitrary sequences, just as is expected from a model for translation initiation. Finally, a remarkable fact is that the model was additionally able to recognize the presence of the termination codon. As a result, the plot presents an overview of the complete process of translation. First a detected signal (the $Shine-Dalgarno$ signal) signal the ribosome to slow down because a coding region is approaching (this was also realized in [69]). Second, a synchronization flag is detected as shown in the initiation signal. Once the translation complex has been constructed, the process of elongation starts and a steady state takes place in the coding region. Finally, a third detected region (the termination codon) informs the ribosome that the protein (message) has been synthesized (decoded).

The translation signals found by this model suggests that the last 13 bases sequence of the $16S\ rRNA$ has a broader role in the process of translation, i.e. not only initiation but also elongation and termination. Although there have been some suggestions that the $16S\ rRNA$ is involved in the process of termination [55], there is no reference that points out the involvement of the 3' end of the $16S\ rRNA$ in the process of translation. Hence, the obtained results are a novel finding that extends the significance of this structure in the gene expression process in prokaryotes.

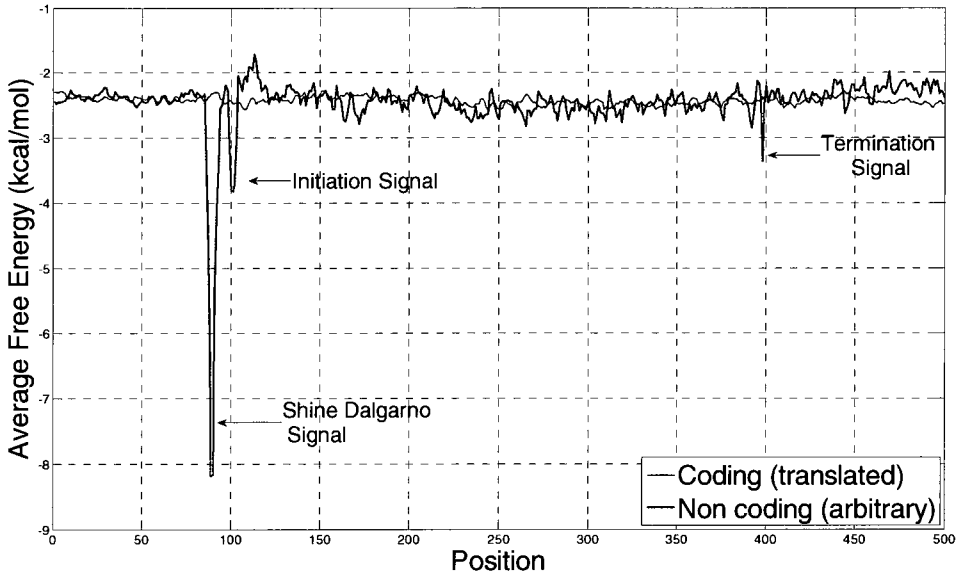


Figure 4.7. Detected translation Signals for MG1655

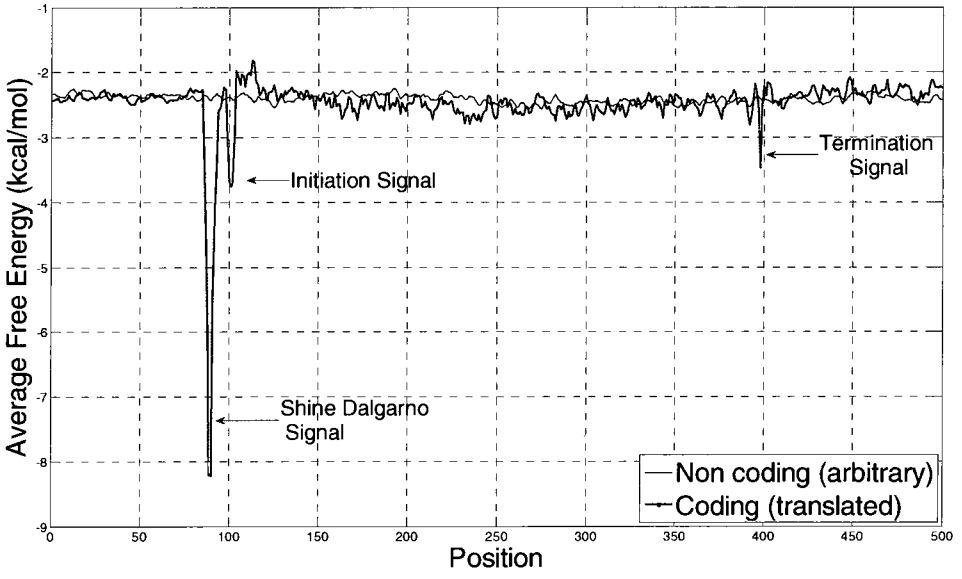


Figure 4.8. Detected translation Signals for O157:H7

4.2.2 Detection Algorithm Optimization. In order to optimize the detection algorithm of the codebook model described before, a modified version of the method of free energy doublets presented in [131] is adopted to calculate the energy function given in equation

(4.1). This function represents a free energy distance metric in $kcal/mol$. Our algorithm assigns weights (w_n) to the n^{th} doublet of the k^{th} codeword such that the total energy (E_k) of that particular codeword increases exponentially as we have consecutive number of matches. At the same time this total energy is decreased if a mismatch occurs. Therefore, the total energy value (E_k) gets emphasized or de-emphasized when consecutive number of matches or mismatches occurs. The free energy metric for the k^{th} codeword is calculated by

$$E_k = \sum_{n=1}^{N-1} w_n \delta_n E_n, \tag{4.1}$$

where δ_n means a match ($\delta_n = 1$) or a mismatch ($\delta_n = 0$) at the n^{th} doublet position when comparing the $\boldsymbol{k^{th}}$ codeword with the corresponding subsequence of length N from the $mRNA$ sequence, E_n is the free energy for the n^{th} doublet, w_n is the weight applied to the doublet energy E_n. The weights (w_n) are given by:

$$w_n = \begin{cases} \rho_n + a^{\sigma_n} & , \quad if\ match \\ max\{w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}), 0\} & , \quad if\ mismatch' \end{cases} \tag{4.2}$$

where (σ_n) and ($\tilde{\sigma}_n$) are the numbers of consecutive matches or mismatches at the n^{th} position (updated at each alignment) and ρ_n is an offset variable updated as follows:

$$\rho_n = \begin{cases} \rho_{n-1} & if\ \delta_n = 1 \\ 0 & if\ \delta_n = 0\ \&\ \rho_n \leq a \\ max\{w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}), 0\} & otherwise \end{cases} \tag{4.3}$$

where a is a constant that will determine the exponential growth of the weighting function.

By using this exponentially weighted free energy decoding algorithm, detection of translation signals ($Shine-Dalgarno$, Initiation and termination signals) has been

optimized. Hence, this algorithm allows for a better resolution and flexibility of translation signals detection. The proposed modeling of ribosome decoding is summarized in Algorithm I.

Algorithm I: Exponentially-Weighted Free Energy Metric Based Ribosome Decoding

Given: Codebook C with L codewords of length N and a subsequence S of length N from the received noisy $mRNA$ sequence. Notation: c_n^k is the n^{th} symbol of k^{th} codeword, s_n is the n^{th} symbol of S, E_k is the exponentially weighted free energy metric when the k^{th} codeword is used (E_k is initialized to $0, 0 \leq k \leq L$), and $Energy\ (a, b)$ is the energy dissipated on binding with the nucleotide doublets ab (see Table 4.2, e.g. the energy dissipated by binding with AC is $-1.8\ kcal/mol$). The parameter w_n is the weight applied to the doublet free energy in the n^{th} position. σ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches respectively, and ρ is an offset variable updated at each step. This offset variable is introduced for the purpose of keeping track of the growing trend that happens when consecutive number of matches occurs followed by a mismatch. When a mismatch occurs we increment the number of mismatches that is initialized to zero by one, reset the number of matches back to zero, calculate the current weighting factor, and finally reevaluate the offset variable to be used in the next alignment. Without the use of this offset variable, we will have several peaks when we come into a good match of the codeword in that particular alignment.

EWFERD Algorithm

Initialize $E_1, E_2, E_3, \ldots, E_L = 0$
for $k = 1 \ldots L$ ***do***
    Initialize $\sigma_0 = 0,\ \tilde{\sigma}_0 = 0,\ \rho_0 = 0, w_1 = a$
    for $n = 1 \ldots N - 1$ ***do***
        if $c_n^k c_{n+1}^k = s_n s_{n+1}$, ***then***
            Increment $\sigma_n = \sigma_{n-1} + 1$

Set $\tilde{\sigma}_n = 0$
$w_n = \rho_{n-1} + a^{\sigma_n}$
$\rho_n = \rho_{n-1}$
else
Increment $\tilde{\sigma}_n = \tilde{\sigma}_{n-1} + 1$
Set $\sigma_n = 0$
$v = w_1 - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n})$
if $n \geq 2$, ***then***
$v = w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n})$
end if
$w_n = max(0, v)$
if $\rho_{n-1} \leq a$, ***then***
$\rho_n = 0$
else
$\rho_n = max\{w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}), 0\}$
end if
end if
Update $E_k = E_k + w_n \times Energy(c_n^k c_{n+1}^k)$
end if
end for
$E_{min} = min(\boldsymbol{E})$, *where* $\boldsymbol{E} = [E_1, E_2, \ldots, E_L]$

For larger values of a, the exponential will grow faster as the number of consecutive matches increases (hence increasing the likelihood that the right sequence is enhanced) making the algorithm more sensible to the correlation in the sequence. Not only does this algorithm allow controlling the resolution of detection (by the choice of the parameter a) but also allows identification of the exact position of the best match of the *Shine-Dalgarno* signal in the genes under study.

In order to test our proposed model, sequences of the complete genome of the prokaryotic bacteria $E.coli$ strain $MG1655$ and $O157{:}H7$ strains were obtained from the National Center for Biotechnology Information. Test sequences were selected to be 500 bases long and follow the structure described in Figure 4.6. The same procedural steps to test the codebook model described in section 4.2.1 are followed here too, with step 5

being replaced by our exponentially weighting algorithm. Our proposed exponentially weighting algorithm was not only able to detect the translational signals (*Shine-Dalgarno*, start codon, and stop codon) but also resulted in a much better resolution than the results obtained when using the codebook without weighting. This improvement in the resolution of detection can be proven by the use of several statistical measures. One of these measures in the peak-to-average values in both cases. Paying a close attention to the average and peak values with/without the use of the exponentially weighting algorithm, reveals this improvement. Moreover, this improvement can be controlled by the choice of the exponential parameter a. In other words, a larger a will results in a larger peak-to-average value, and hence a better resolution. Figure 4.9 shows average results for the detection of the SD, start and stop codons being compared to the work in [30]. It can be observed that the proposed algorithm is able to identify the *Shine-Dalgarno* (dip at position 90) and the start codon (dip at position 101) and the stop codon (dip at position 398). Moreover, these results support the arguments for the importance of the $16S\ rRNA$ in the translation process.

To detect the *Shine-Dalgarno* signal in a single gene, the proposed weighting algorithm was applied and compared again to the algorithm used in [30]. Figure 4.10 illustrates that if the parameter a is further increased, the resolution of the peak corresponding to the SD sequence will be larger. This shows that the algorithm is sensitive to the parameter a, and by properly choosing this value, the accuracy can be improved obtaining a much better performance than the codebook alone. In summary, not only does this weighting algorithm detect the *Shine Dalgarno* in the exact location, but

also provides flexibility in controlling the resolution of detection through the choice of the parameter a.

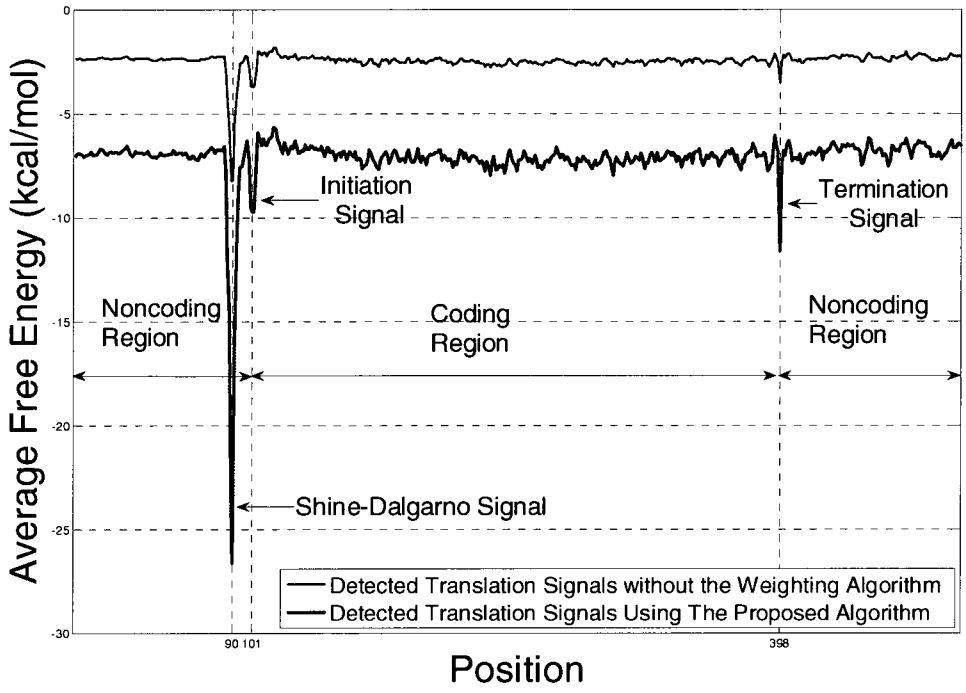


Figure 4.9. Detected translational signals using our weighting algorithm compared to the algorithm in [30] ($N = 5, a = 1.5$)

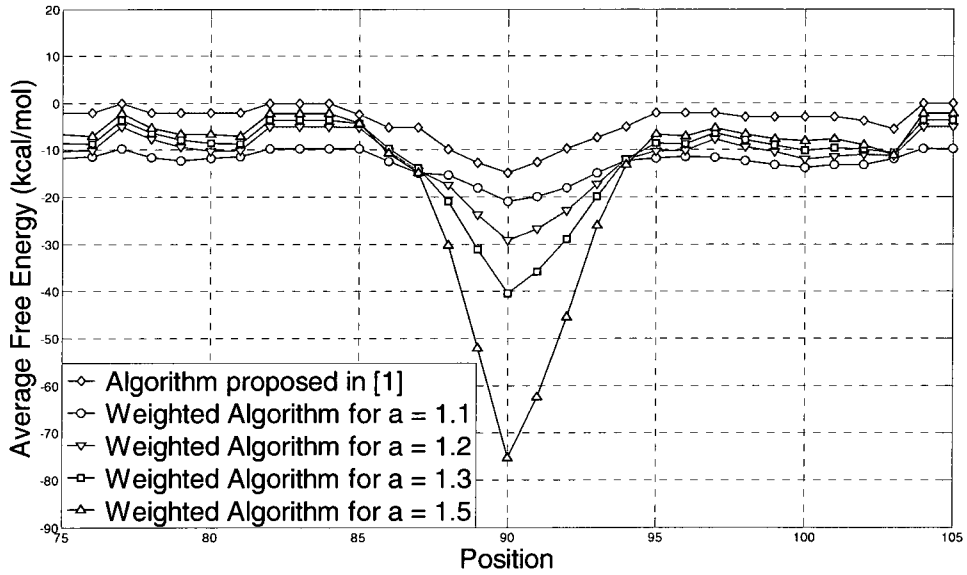


Figure 4.10. Detected SD signal using our weighting algorithm with different values of the parameter a

4.2.3 Mutation Analysis. In our work [103] and [99] based on the codebook model described in [30], we have applied our proposed exponentially weighting algorithm described in section 4.2.2 to test the effect of different types of mutations in the ribosome on protein synthesis. To do this, experimental results obtained by mutating regions of the 3' end of the $16S\ rRNA$ molecule are compared with results obtained by incorporating these mutations in the $16S\ rRNA$ based codebook model. In other words, we have introduced these mutations *in silico* in all positions of the last 13 bases of the $16S\ rRNA$ and executed the proposed algorithm on the $E.coli$ data set.

Jacob introduced a point mutation in the 5^{th} position of the $16S\ rRNA$ [66]. Specifically, the 5^{th} position in the arrangement illustrated in Table 4.3.

Table 4.3. *Jacob* mutation

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
Original	*A*	*U*	*U*	*C*	*C*	*U*	*C*	*C*	*A*	*C*	*U*	*A*	*G*
Compliment	*U*	*A*	*A*	*G*	*G*	*A*	*G*	*G*	*U*	*G*	*A*	*U*	*C*
Mutation	*U*	*A*	*A*	*G*	*A*	*A*	*G*	*G*	*U*	*G*	*A*	*U*	*C*

This point mutation is a change of the nucleotide $C \rightarrow U$ in the ribosome small subunit. This is equivalent to make a mutation from $G \rightarrow A$ in the complement sequence shown above. The result of this mutation was a reduction in the level of protein synthesis as certified by experimental published results. This mutation is tested using the codebook model with our weighing algorithm. First the mutation as specified in [67] is performed in the 13 bases. The codebook is constructed based on the mutated sequence. The resulting "mutated" decoder is used in the algorithm and the response of the system is observed. Figure 4.11 shows how the recognition of the *Shine-Dalgarno* signal is affected for the *Jacob* mutation (notice the partial loss in the amplitude of the *Shine-*

Dalgarno signal). It can be inferred from the plot that the levels of protein production will be reduced but not completely stopped. This can be explained by the fact that the ribosome (while having a mutated $16S\ rRNA$ molecule) will detect a wrong start codon as an initiation signal. Hence, it will end up translating a shorter part of the $mRNA$ sequence. However, $Jacob$ mutation does not affect the detection of the termination signal. This means that protein synthesis process is normally terminated.

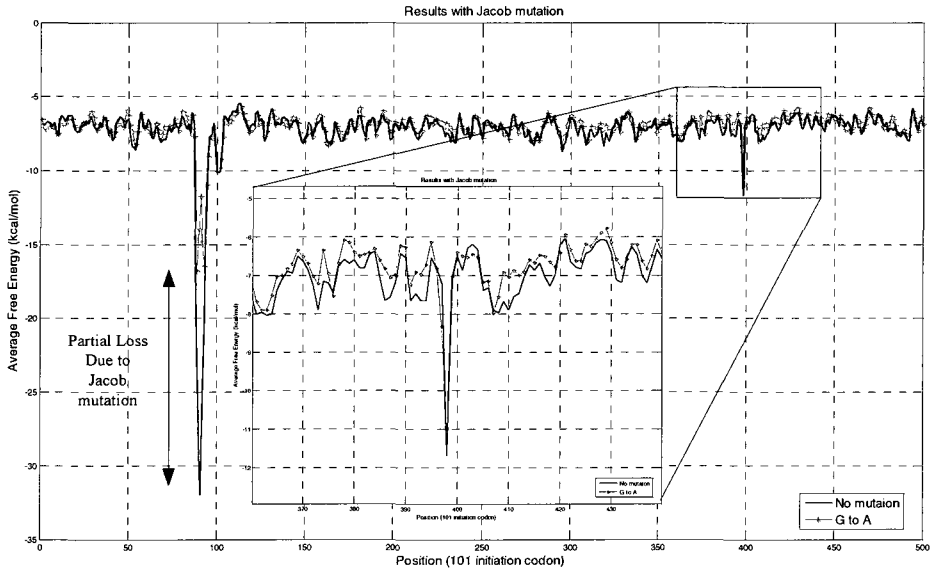


Figure 4.11. $Jacob$ mutation

Another published record of the behavior of the protein synthesis under mutations in the 3' end of the $16S\ rRNA$, was done by *Hui* and *De Boer* [66]. In this experiment, the mutations were done in positions 4 to 8 ($GGAGG \rightarrow CCUCC$) and positions 5 to 7 ($GAG \rightarrow UGU$). The results of both mutations were lethal for the organism in the sense that the production of proteins stopped. In other words, no protein was synthesized and the whole translation process did not take place.

After introducing *Hui* and *De Boer* mutation and testing them using the codebook model and the weighting algorithm, the results showed a complete loss of the *SD* signal. This is illustrated in Figure 4.12 where the black curve corresponds to the case where no mutations are present, the blue curve corresponds to *Hui* mutation, and the red curve corresponds to *De Boer* mutation. Based on this complete loss of the *Shine-Dalgarno* signal and the fact that the *Shine-Dalgarno* signal is an initiation signal, it can be inferred that the translation will never take place. Note that results obtained by mutations in the $16S\ rRNA$ also apply to scenarios with mutations in the *mRNA* at corresponding positions.

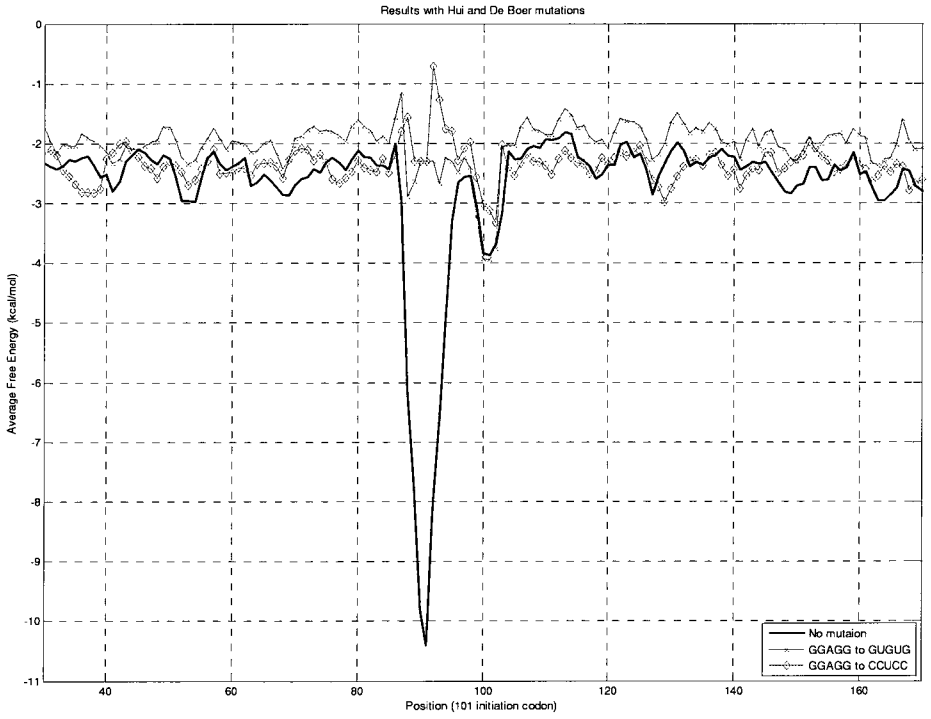


Figure 4.12. *Hui* and *De Boer* mutations without using the weighting algorithm

The same mutations are tested using our exponentially weighting algorithm which resulted in a similar result but with a better resolution (note the difference in the y-axis) of the translation signals as illustrated in Figure 4.13.

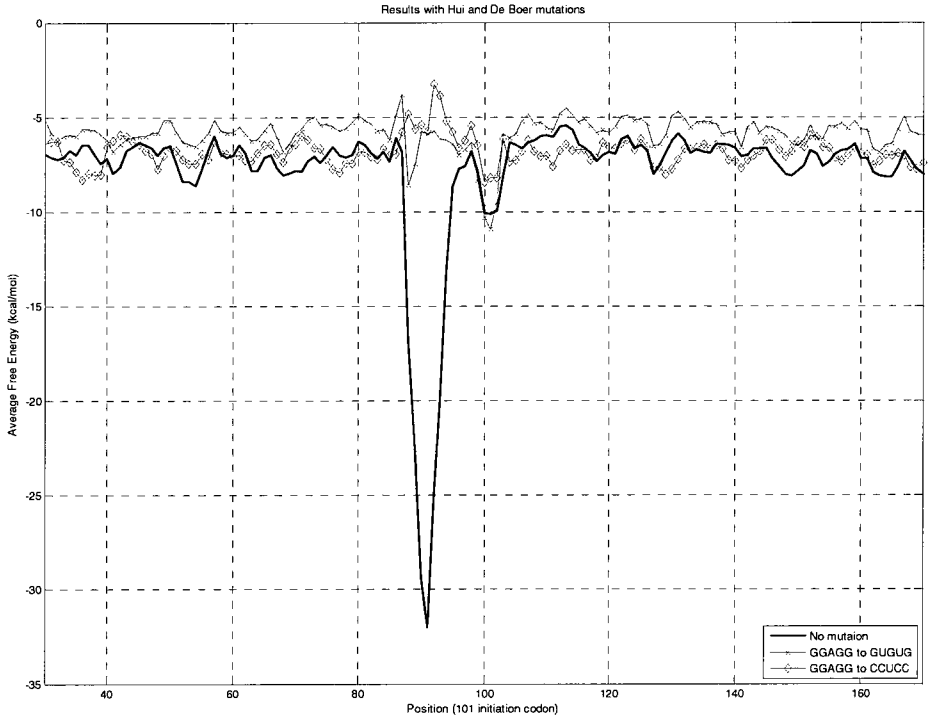


Figure 4.13. *Hui* and *De Boer* mutations using our weighting algorithm

To further exploit the model, point mutations have been introduced in all positions of the last 13 bases of the $16S\ rRNA$ molecule in order to study their influence on the process of translation. The obtained results are summarized in Table 4.4 by quantizing into 5 levels the influence of these mutations on each of the translation signals (SD, initiation, and stop). The levels are: – represents no influence in the recognition of the signal, ⇓ represents a strong negative influence, ↓ a weak influence, ↑ a weak positive influence, and ⇑ a strong positive influence. For example, results show how a mutation in

position 5 has a strong negative influence in the recognition of the SD signal, just as found in the $Jacob$ investigation.

Many other point mutations in the last 13 bases of the 3' end of the $16S\ rRNA$ molecule were incorporated and tested using the proposed model. All results totally agreed with real life which further certifies the correctness of the proposed model.

Table 4.4. Point mutations in the last 13 bases of the $16S\ rRNA$

Base	1	2	3	4	5	6	7	8	9	10	11	12	13
SD	–	–	–	↓	⇓	⇓	↓	–	–	–	–	–	–
Start	–	–	⇓	↓	↓	↓	↓	–	⇓	↓	↓	↓	↓
Stop	↓	⇓	↓	–	–	–	–	–	↓	⇓	⇓	–	↓

Inspecting the results more carefully, several remarkable and "new" findings can be observed. Some of these are:

1. A mutation in position 8 has no influence in the detection of the translation signals, probably the reason is that the role of this nucleotide is to introduce spacing at the moment of decoding the $mRNA$ sequence
2. A mutation at position 6 has nearly the same influence as a mutation at position 5
3. A mutation at position 9 affects the recognition of the initiation codon even if it does not affect the SD signal. This could lead to a wrong initiation of translation or a "frame shift"
4. Exactly the central part of the 13 bases (bases 4-8) which influences the SD is missing in eukaryotes. The rest of the sequence that involves AUG and stop codon recognition are still there.

All mutation analysis done before has been applied to both $MG1655$ and $O157{:}H7\ E.coli$ strains. This of course further strengthens and supports the proposed

model. This mutation analysis can be further investigated using different metrics discussed in section 4.3.

The results obtained are completely consistent with the published experimental results. This demonstrates the relevance of the proposed model, its biological accuracy, and its flexibility to incorporate and study structural changes. Moreover, a laboratory work that usually takes months was simplified through the introduction of mutations to our model. Other combinations of mutations can also be tested using our algorithm without the need for time and cost consuming laboratory experimentation.

The analysis of the results made possible by this model can serve as a way to introduce new lines of biological research. In practice, these results can lead to better recognition of signals in translation, therefore, improving test-tube translation in genetic engineering. On the other hand, these results further certify the relevance of communications and coding theory concepts and principles to better understand biological phenomena. This will also help address other fundamentally important issues that cannot be explored systematically and quantitatively by experiments alone.

4.3 Communications Theory Based Modeling

The previous codebook model discussed in section 4.2.1 is based on the use of a free energy distance metric to quantify the correlation between the mRNA and the last 13 bases sequence. In a broader context, we can use different distances to quantify this rRNA – mRNA interaction by looking at these sequences as numerical vectors. The most important distances for statistics are: Euclidean distance, generalized Euclidean distance (which include χ^2 and $Mahalanobis$ distances), $Minkowsky$ distance (which include

the sorting and the symmetric difference distances), and the $Hellinger$ distances [1]. This section presents four different distance metrics that can be used to quantify the interaction between the received $mRNA$ and the last 13 bases sequence on the $16S\ rRNA$ molecule. The four distance metrics are described below. The reasoning behind the use of several distance metrics is to optimize the detection mechanism that we assume the ribosome uses to detect the $Shine$ - $Dalgarno$ signal and hence accurately initiate/terminate the translation process.

A. Euclidean Distance Metric

A Euclidean distance measure (or a generalized Euclidean distance) can be used to detect the ribosome binding site in the $mRNA$. This measure is calculated at each single base in the $mRNA$ sequence as follows:

1. Map both $mRNA$ sequence and the binding sequence to their equivalent numerical quaternary representations using ($A = 0$, $C = 1$, $G = 2$, and $U = 3$).
2. Slide the binding sequence along the $mRNA$ sequence and find the generalized Euclidean distance at each alignment position.
3. Sum the resulting Euclidean distance vector and save the result as a function of base position.
4. Plot the resulting vector in step 3 and detect minimal points.

A minimal point (dip) of amplitude of zero in the resulting plot corresponds to a total match of the binding sequence. The next minimal point is a partial match of the binding sequence. Therefore, this method is able to detect the binding sequences in their exact location and accounts for mismatches as well.

B. Cross Correlation Metric

In telecommunication, a matched filter is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. This is equivalent to convolving the unknown signal with a time-reversed version of the template. The matched filter is the optimal linear filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. This metric is based on using a matched filter of an impulse response equal to $h(n) = y(-n)$ and an input of $x(n)$ (see Figure 4.14) where $y(n)$ is the binding sequence and $x(n)$ is the $mRNA$ sequence.

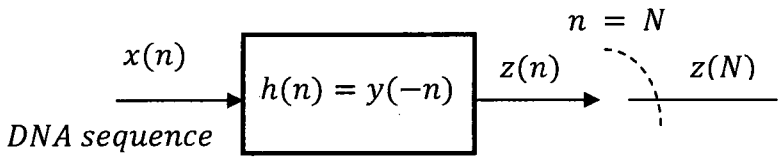


Figure 4.14. Matched Filter

1. Map both the $mRNA$ sequence $x(n)$, and the binding sequence $y(n)$, under study to their equivalent binary representation using ($A = 00$, $C = 01$, $G = 10$, $and\ T = 11$).
2. Convert each zero in the resulting binary sequences to (-1) to get a better correlation form.
3. Correlate both sequences using

$$z(n) = \sum_{n=-\infty}^{\infty} x(k)y(n+k), \tag{4.4}$$

Correlation is equivalent to convolution of the sequence $x(n)$ with an inverted version of the sequence $y(n)$. This can be done by first flipping the sequence $y(n)$ and then convolving it with the sequence $x(n)$.

4. Plot the cross correlation function and detect the maximal points.
5. Convert the binding sequence detected positions (a maximal point in the plot) to their corresponding locations in the original $mRNA$ sequence using:

$$DP_{mRNA} = \lceil (DP_{plot} - 2\,BSL + 1)/2 \rceil, \tag{4.5}$$

where DP_{mRNA} is the **d**etected **p**osition in the $mRNA$ sequence, DP_{plot} is the **d**etected **p**osition in the plot; BSL is **b**inding **s**equence **l**ength, and $\lceil X \rceil$ rounds the value X to the nearest integer larger than X.

C. Exponential Detection Metric

By aligning the binding sequence to be detected with the mRNA sequence, an exponential metric related to the total number of matches at each alignment is evaluated as follows:

1. Slide the binding sequence along the $mRNA$ sequence one base at a time.
2. At the i^{th} alignment, calculate an exponential weighting function ($W(i)$) using the equation:

$$W(i) = \sum_{n=1}^{N} w(n), \tag{4.6}$$

where $w(n)$ is the weight applied to the base in the n^{th} position and N is the length of the binding sequence under study. The weights are given by:

$$w(n) = \begin{cases} a^{\sigma} & , \quad if\ match \\ 0 & , \quad if\ mismatch \end{cases}, \tag{4.7}$$

where a is an input parameter that controls the exponential growth of the weighting function W, and σ is the number of matches at each alignment. .

3. Repeat step 2 for all alignments along the $mRNA$ sequence to get the weighting vector $\boldsymbol{W}$:

$$\boldsymbol{W} = [w(1), w(2), \dots, w(L - N + 1)], \tag{4.8}$$

where L is the length of the $mRNA$ sequence.

4. Plot the weighting vector $\boldsymbol{W}$, and detect peaks.

This metric considers the total number of matches at each alignment rather than the consecutive number of matches and mismatches as in the codebook model discussed in 4.2.1.

D. Free Energy Metric

By using the free energy table (see Table 4.2), a free energy distance metric in $kcal/mol$ can be calculated. This metric is evaluated at each alignment between the $mRNA$ sequence and the binding sequence under study as follows:

1. Align the binding sequence with the $mRNA$ sequence and shift it to the right one base at a time.

2. At the i^{th} alignment, calculate the free energy metric using the equation

$$E(i) = \sum_{n=1}^{N-1} E(y_n y_{n+1}).\delta(n), \tag{4.9}$$

where N is the length of the binding sequence. $\boldsymbol{Y}$ denotes the binding sequence vector and is given by $\boldsymbol{Y} = [y_1, y_2, y_3, \dots, y_N]$. Let $\boldsymbol{X}$ denote the $mRNA$ sequence vector where$\boldsymbol{X} = [x_1, x_2, x_3, \dots, x_L]$. $E(y_n y_{n+1})$is the energy dissipated on

binding with the nucleotide doublets $y_n y_{n+1}$ and is calculated from Table 4.2.

$\delta(n)$ is given by

$$\delta(n) = \begin{cases} 1 , & if \ \ y_n y_{n+1} = x_n x_{n+1} \ \ (match) \\ 0 , & if \ \ y_n y_{n+1} \neq x_n x_{n+1} \ \ (mismatch) \end{cases}, \tag{4.10}$$

3. Repeat step 2 for $i = 1,2,\dots, L - N + 1$, where L is the length of the $mRNA$ sequence vector,
4. Plot the free energy vector $\boldsymbol{E}$ and detect minimal points.

4.3.1 Simulation Results. In this section, we prove the validity of the proposed distance metrics and we demonstrate their usefulness in pointing out interesting and new biological insights related to the process of translation in gene expression.

In order to test our proposed metrics, sequences of the complete genome of the prokaryotic bacteria $E.coli$ strain $MG1655$ and $O157:H7$ were used. These sequences are available in the National Center for Biotechnology Information ($NCBI$) [136]. The four proposed metrics described before were applied to detect the last 13 bases of the $16S\ rRNA$ molecule in the given $mRNA$ sequence. Simulation results (applied to $MG1655\ E.coli$ starin) show that the proposed metrics allow detecting the translational signals at their exact corresponding locations ($Shine\text{-}Dalgarno$ at position 90, start codon at position 101, and stop codon at position 398). Interestingly, they also identify coding regions (marked with higher ripple) and the non-coding regions (marked with lower ripple) as can be observed in figures 4.15-4.18. This new finding suggests that the sequence of the last 13 bases of $16S\ rRNA$ molecule has a higher correlation with coding regions as compared with non-coding regions. Furthermore, this finding suggests that the proposed metrics, which were originally designed for regulatory sequence identification, can be utilized for gene (non-coding + coding regions) identification as well.

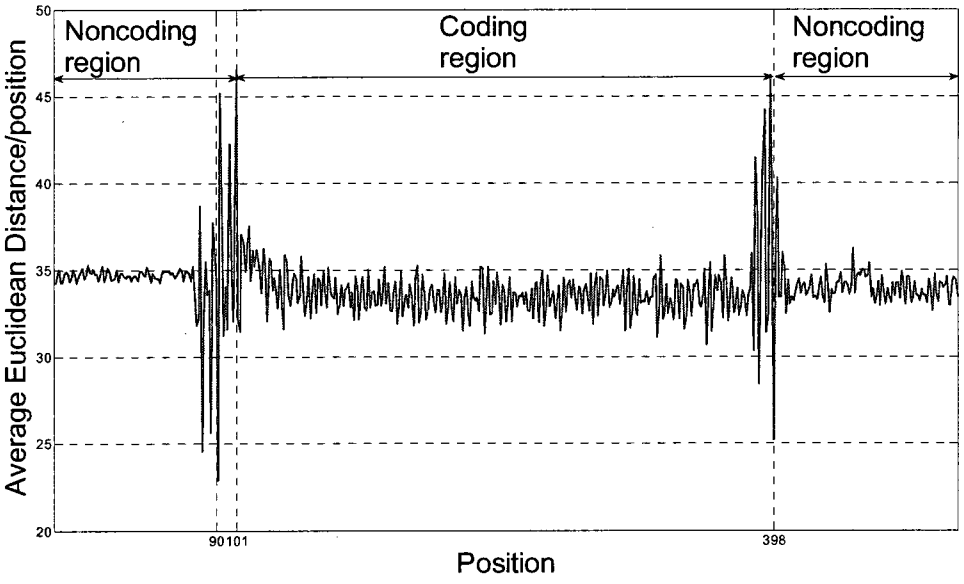


Figure 4.15. Euclidean distance metric applied to $E.coli\ MG1655$

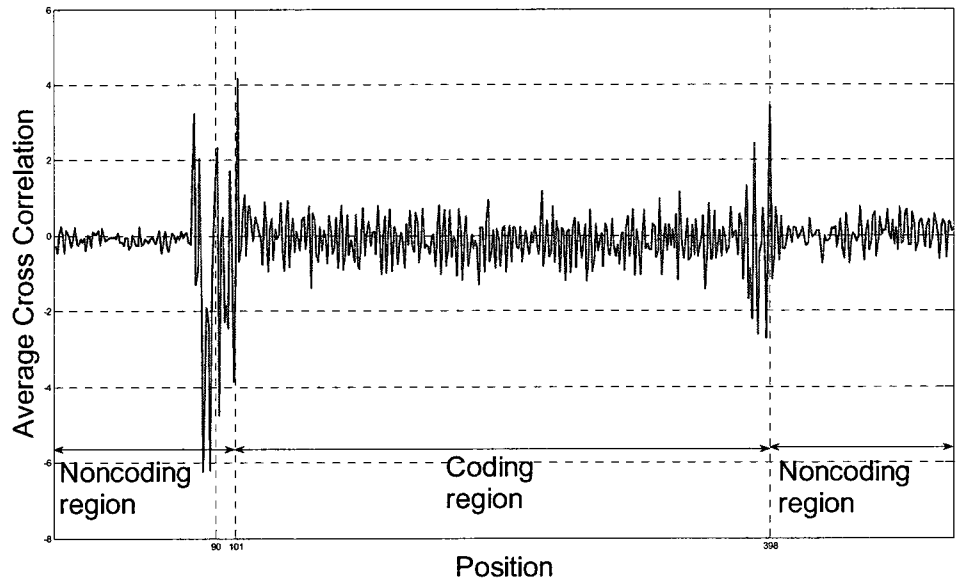


Figure 4.16. Cross correlation metric applied to $E.coli\ MG1655$

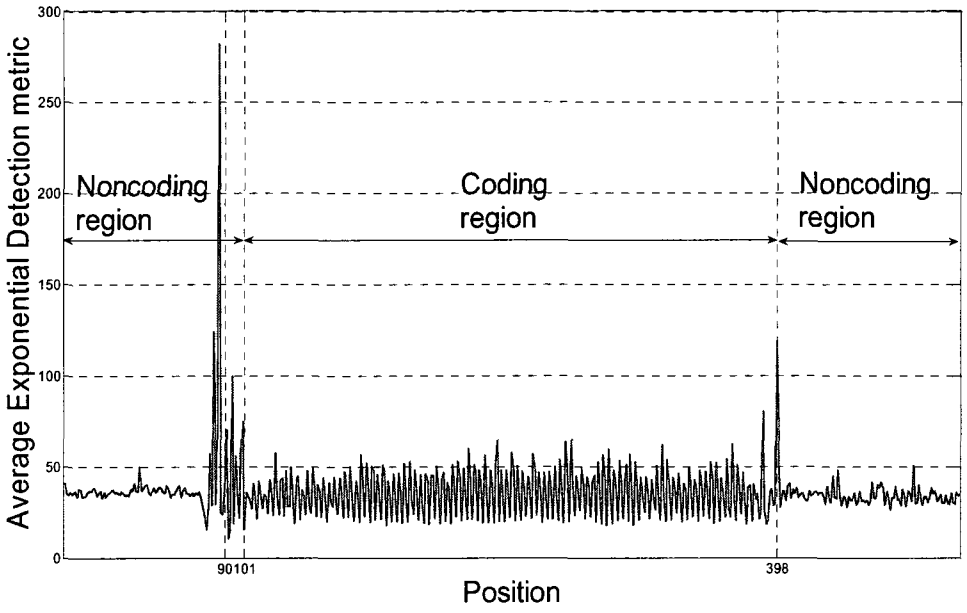


Figure 4.17. Exponential detection metric applied to $E.coli\ MG1655$

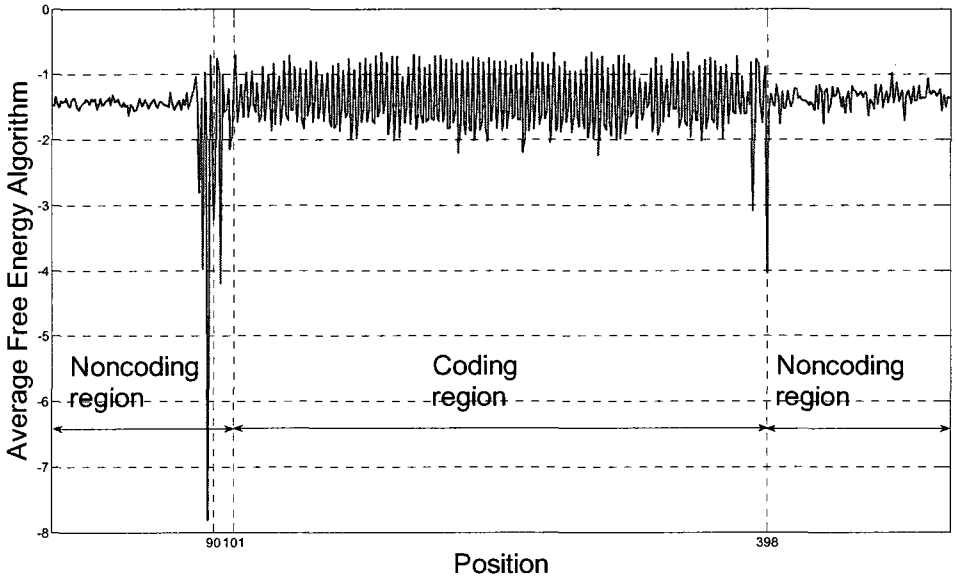


Figure 4.18. Free energy metric applied to $E.coli\ MG1655$

Giving a closer look to the details in figures 4.15-4.18 reveals the existence of a periodic structure in the coding regions. This periodicity is the well known "period-3 property" which has been known phenomenon for eukaryotic exons [44, 87, 158]. *DNA*

periodicity in exons is determined by codon usage frequencies. This periodicity reflects correlations between nucleotide positions along coding sequences which is caused by the asymmetry in base composition at the three coding positions. Figure 4.19 shows a closer look to the coding region in Figure 4.18.

To confirm the previous results shown in figure 4.15-4.18, the four detection algorithms proposed were applied to several other bacterial genomes. Simulation results shown in figure 4.20-4.23 were obtained for *Salmonella Typhimurium LT2* bacterial genome. Period-3 property in the coding regions is also confirmed by the the obtained results which further certifies the correctness of the proposed metrics and their biological relevance.

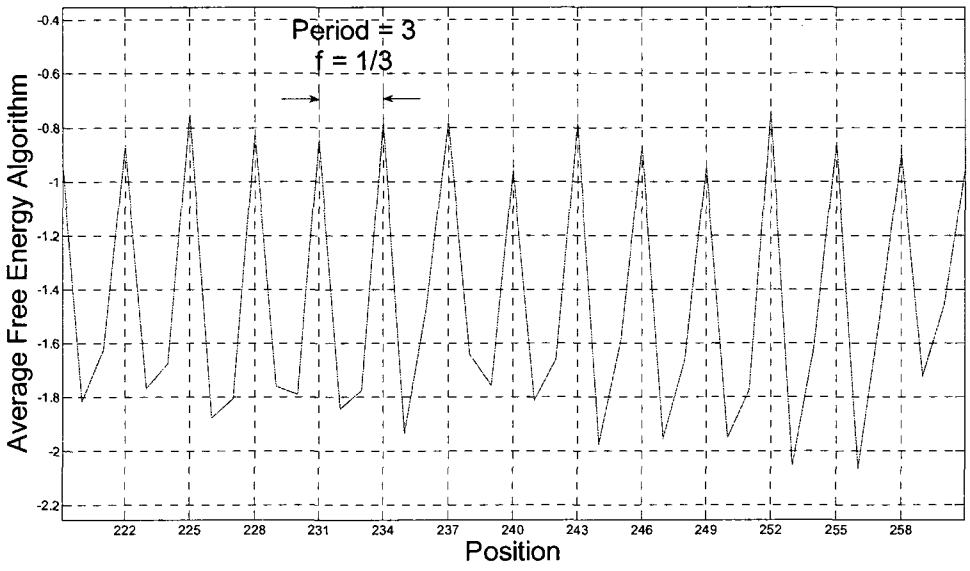


Figure 4.19. Illustration of period-3 property

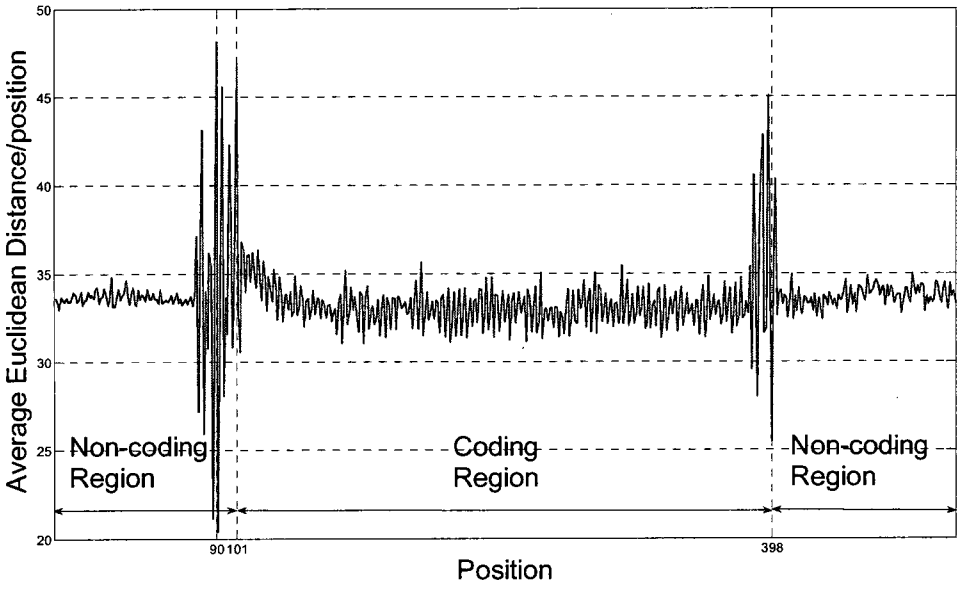


Figure 4.20. Euclidean distance metric applied to *Salmonella Typhimurium LT2*

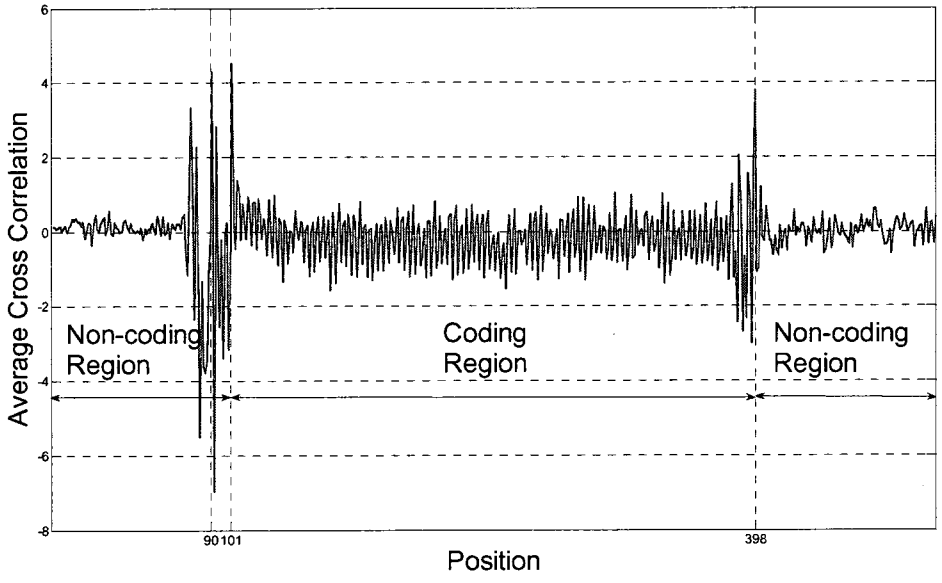


Figure 4.21. Cross correlation metric applied to *Salmonella Typhimurium LT2*

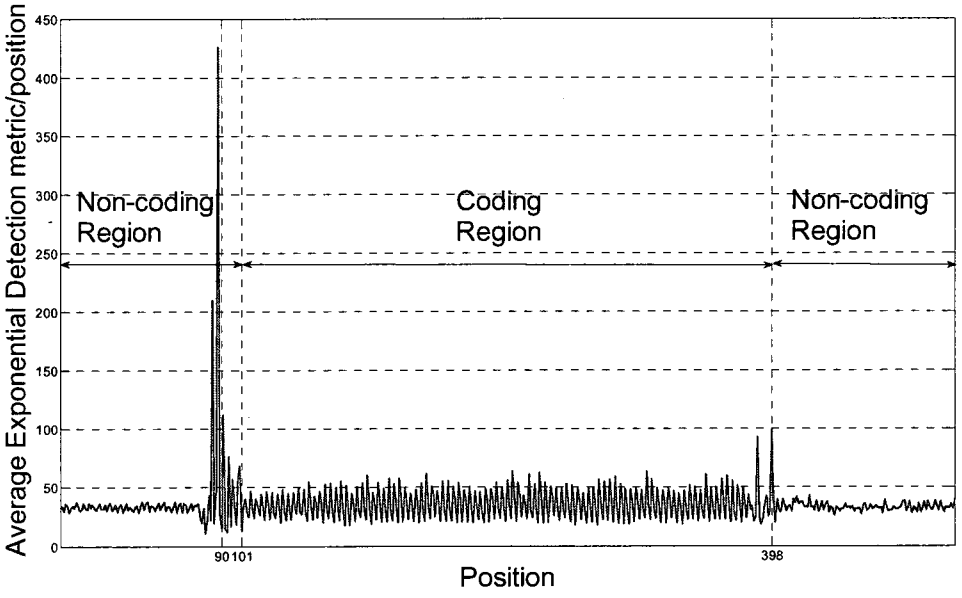


Figure 4.22. Exponential detection metric applied to *Salmonella Typhimurium LT2*

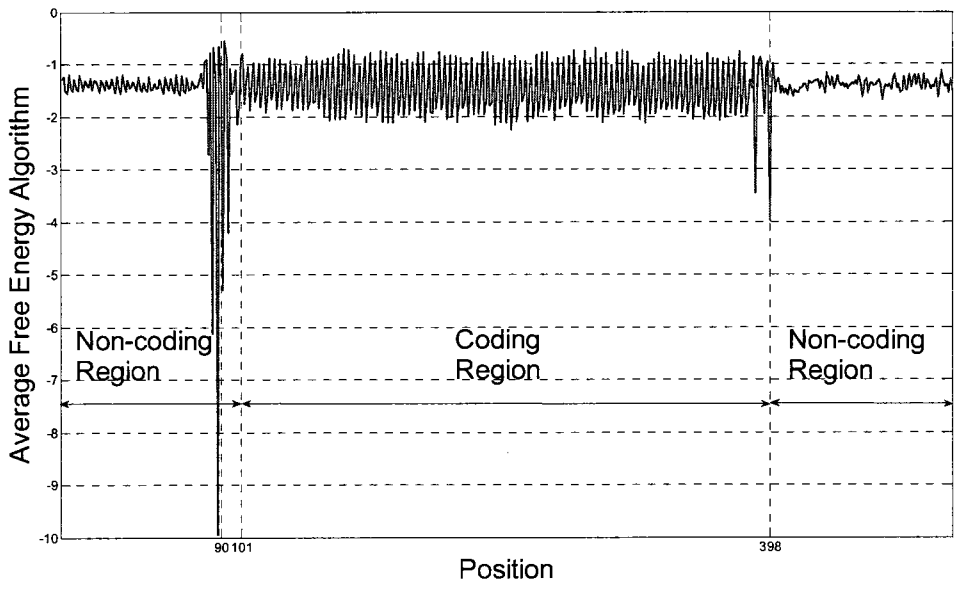


Figure 4.23. Free energy metric applied to *Salmonella Typhimurium LT2*

In section 4.2.3 based on the codebook model and our weighting algorithm, we have tested the effect of different point mutations in the ribosome on protein synthesis. This same mutation analysis was also carried out using the new four metrics discussed

before. For example, *Jacob* mutation, a mutation in the 5^{th} position of the last 13 bases of $16S\ rRNA$ molecule [67], was tested using The exponential detection metric. Simulation result in Figure 4.24 shows a reduction in the amplitude of the *Shine-Dalgarno* signal compared to the non-mutation case in Figure 4.17. This reduction can be interpreted as a reduction in the level of protein synthesis, i.e. the levels of protein production will be reduced but not completely stopped. Moreover, coding and non-coding regions can be identified which is something the codebook model did not offer.

Hui and *De Boer* mutations which occur in positions 4 to 8 ($GGAGG \rightarrow CCUCC$) and positions 5 to 7 ($GAG \rightarrow UGU$) respectively, were also tested. The results of both mutations show a complete loss of the *Shine-Dalgarno* signal. Hence, it can be inferred that the translation will never take place. This is also consistent with published results as well. This is illustrated in Figures 4.25 and 4.26 which also show that the developed metrics allow identifying coding and non-coding regions.

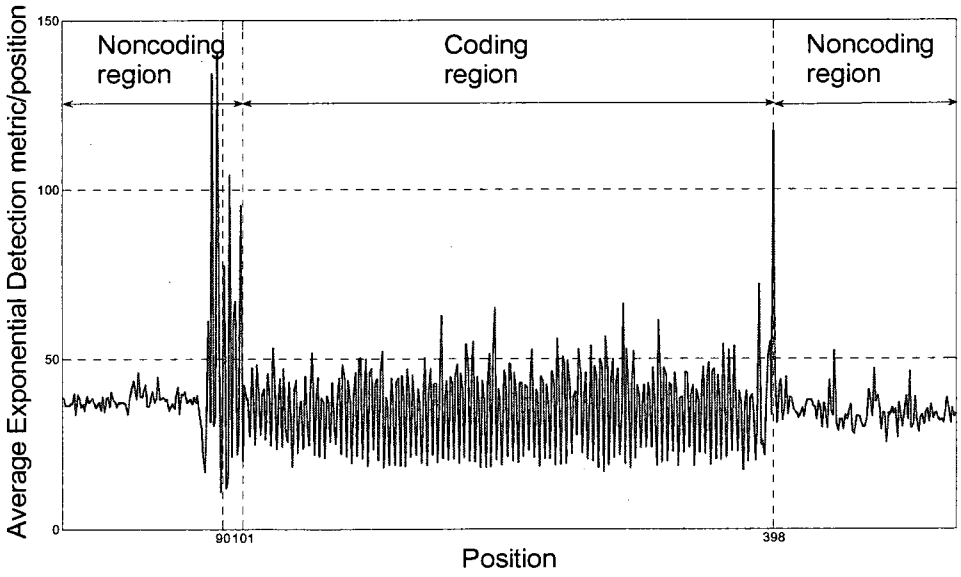


Figure 4.24. *Jacob* Mutation using Exponential Detection

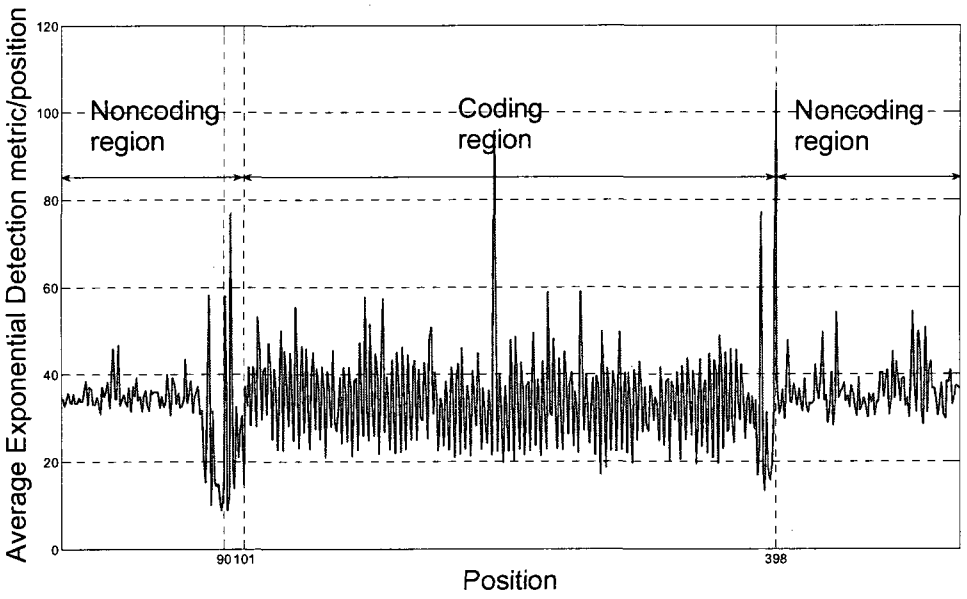


Figure 4.25. *Hui* Mutation using Exponential Detection

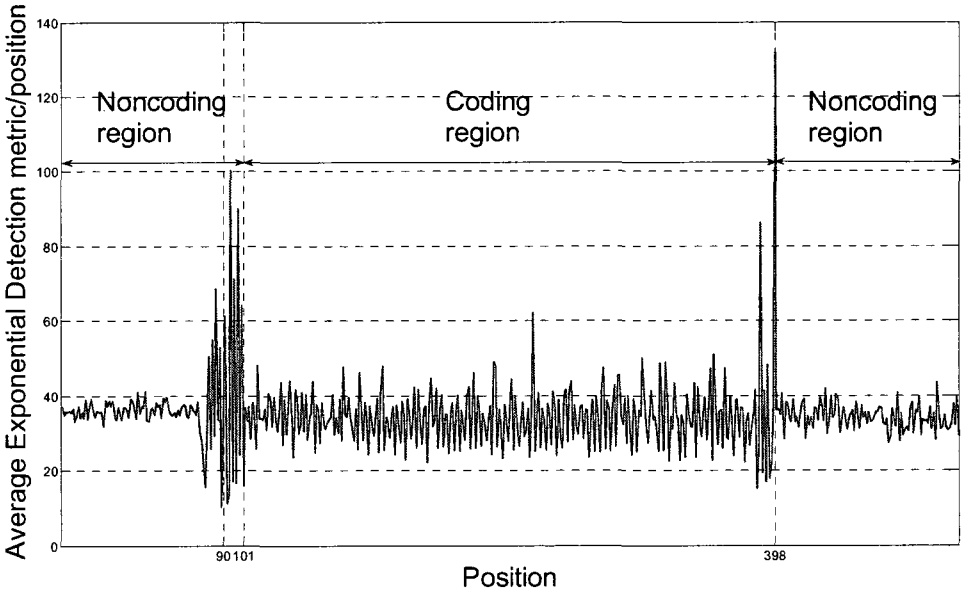


Figure 4.26. *De Boer* Mutation using Exponential Detection

4.4 Conclusion

The increase in genetic data during the last years has prompted the efforts to use advanced techniques for their interpretation. This chapter proposes a novel application of ideas and techniques from communications and coding theory to model and analyze gene expression and gene and regulatory sequence identification. A model based on a variable length codebook and an exponentially weighted metric is used to model the process of translation in gene expression in $E.\,coli$. In this model it is assumed the ribosome decodes the $mRNA$ sequence by using the 3' end of the $16S\ rRNA$ molecule as an embedded codebook. The metric uses an exponential algorithm to recognize the $Shine\text{-}Dalgarno$ (SD) sequence that allows for a better resolution in detecting the initiation signals ($Shine\text{-}Dalgarno$, initiation codon, and termination codon) as compared to the work done in [30]. Four new different metrics for regulatory elements identification are developed and investigated. Simulation results certify the correctness and accuracy of these metrics in detecting regulatory sequences. Moreover, as these metrics are surprisingly capable of distinguishing coding from non-coding regions, they can be utilized for gene identification. Mutations in the 3' end of the $16S\ rRNA$ molecule were investigated. The obtained results totally agree with real life experimentations that have published records. This further certifies the correctness and the biological relevance of the proposed metrics and hence can serve as a way to introduce new lines of biological research.

CHAPTER 5

A CODE MODEL FOR GENE TRANSLATION USING FREE ENERGY BASED DISTANCE DECODERS

Informational analysis of genetic sequences has revealed the existence of significant analogies between the genetic process and information processing systems used in the field of communications engineering. By analyzing key elements involved in the process of gene expression, we have developed several communications and coding theory based models for the process of translation [1-5]. A previous research investigated the use of coding theory based models that quantitatively describe the behavior of the ribosome during translation initiation in prokaryotic organisms [43]. In this chapter, we have investigated a code model for the process of translation in gene expression. We have also employed several minimum distance decoders to verify the proposed model based on the free energies involved in the binding between the ribosome and the $mRNA$ sequence. The key biological elements considered in forming the investigated model are 1) the last 13 bases of the $3'$ end of the $16S\ rRNA$ molecule, 2) the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), 3) the free energies involved in the $rRNA - mRNA$ interaction, and 4) RNA/DNA base-pairing principles. The model was tested on five different bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs (e.g. $GeneMark$ [19] and $Glimmer$) [3].

5.1 Introduction

The increase in the availability of genetic data in the last years has encouraged development and use of novel techniques and principles from the engineering fields for examining and analyzing the genomic structure including coding and non-coding regions [2, 3, 7, 13, 15, 17, 18]. In this chapter, a code model is investigated for the analysis and understanding of the process of translation in gene expression. As a basic analogy, data information is encoded, transmitted and processed in communications, while *DNA* information is replicated, expressed and processed in genetics. The precision and robustness found in molecular biology motivates the quest to try to explain this behavior using concepts from communications and coding theory.

Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communications engineering. Of particular interest are the results from Schneider [141, 150] and Eigen [42]. Drawing from their work and previous work in protein annotation and gene identification, we make several key observations that lead one to hypothesize that similar to engineering, information-processing systems, the genetic system contains mechanisms to protect an organism from errors that occur within its genome.

By analyzing key elements involved in the process of protein gene expression, we have developed several communications and coding theory based models for the process of translation [99-103]. A previous research conducted by E. May investigated the use of coding theory based models that quantitatively describe the behavior of the ribosome during translation initiation in prokaryotic organisms [43]. In this work, the messenger

RNA ($mRNA$) of bacteria is modeled as a noisy encoded signal and the ribosome as a minimum distance decoder, where the $16S$ ribosomal RNA ($16S\ rRNA$) serves as a template for generating a set of valid codewords (the codebook). We have also employed several distance based decoders to verify the proposed model. The key biological elements considered in forming our model are 1) the $3'$ end of the $16S$ ribosomal ($16S\ rRNA$), 2) the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), 3) the free energies involved in the $rRNA - mRNA$ interaction and 4) RNA/DNA base-pairing principles. The model was tested on five different bacterial genomes including $Escherichia\ coli\ \ K$ - $12\ MG1655$, $Escherichia\ coli\ O157{:}H7$, $Salmonella\ \ typhimurium\ \ LT2$, $Bacillus\ Subtilis$, and $Staphylococcus\ Aureus\ Mu50$. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs (e.g. $GeneMark$ and $Glimmer$).

5.2 Code Model for Prokaryotic Translation

In the investigated code model, we verify the significance of using communications and coding theory specifically, for quantitatively modeling the protein translation initiation mechanism in the process of gene expression. Based on this model, the messenger RNA ($mRNA$) can be modeled as a noisy encoded signal and the ribosome as a minimum free energy distance decoder, where the $16S$ ribosomal RNA ($16S\ rRNA$ molecule) serves as a template for generating a set of valid codewords (the codebook).

The RNA bases must be mapped to a numeric representation. The genetic coding alphabet must correspond to a finite field. In binary codes, the finite field consists of 0 and 1. For the genomic code, we know that four bases are found in $mRNA$: $adenine$, $guanine$, $cytosine$, and $uracil$. A fifth base, $inosine$, is found in transfer RNA ($tRNA$) which participates mainly in the elongation phase of translation. During elongation, $inosine$ can wobble pair with more than one of the $mRNA$ bases. These biological characteristics are used to define an alphabet on $GF(5)$ field. Thereby, the RNA bases are thereby mapped as follows: $Inosine\ (I) = 0$, $ne\ \ (A) = 1$, $Guanine\ \ (G) = 2$, $Cytosine\ \ (C) = 3$, and $Uracil\ \ (U) = 4$. The RNA bases are mapped such that in modulo-5 addition of the bases that form hydrogen pairs is zero ($A + U = 1 + 4 = 0, G + C = 2 + 3 = 0$).

The reasoning behind the developed code is as follows. The $3'$ end of the $16S\ rRNA$ is directly involved in binding the $mRNA$ during the initiation phase of protein translation [75]. We use the last thirteen bases of the $16S\ rRNA$ in forming our codewords. The last thirteen bases are used since the hexamer complementary to the $Shine\text{-}Dalgarno$ sequence is found in this region of the $16S\ rRNA$ molecule. The $Shine\text{-}Dalgarno$ sequence is a series of nucleic acid bases on the $5'$ untranslated region (UTR) of prokaryotic $mRNA$. The $Shine\text{-}Dalgarno$ sequence helps attract the ribosome to the initiation site by forming Watson-Crick bonds with the $16S\ rRNA$, part of the $30S$ ribosomal subunit [75, 176]. Specifically, the last thirteen bases of the $16S\ rRNA$ that interact with the $Shine\text{-}Dalgarno$ domain and other sequences on the $5'$ untranslated $mRNA$ leader, are: $3' - AUUCCUCCACUAG \ldots 5'$. Since our received sequence, the $mRNA$, contains the nucleotide sequence which base pairs with the $16S\ rRNA$, we use

the Watson-Crick complement of the last thirteen base sequence in forming our codewords. The complemented sequences is: $5' - UAAGGAGGUGA\ UC\ \ldots 3'$.

We select our parity symbols from all (n − k)-base subsequences of the Watson-Crick complement of the last thirteen base sequence. For example, for a (5,2) code, we select our parity symbols from all 3-base nucleotide subsequences of the thirteen base sequence. Table 5.1 shows these subsequences and their summation values.

Table 5.1. 3-base parity symbols derived from the $16S\ rRNA$

Parity bases	*Modulo-5 sum*	*Parity bases*	*Modulo-5 sum*
UAA	1	*GGU*	3
AAG	4	*GUG*	3
AGG	0	*UGA*	2
GGA	0	*GAU*	2
GAG	0	*AUC*	3

The n-base codewords were selected based on the systematic (a systematic code contains the k information bases at the beginning of the codeword) zero-parity check encoding methodology, such that the following equation is satisfied

$$\sum_{i=1}^{k} u_i + \sum_{i=1}^{n-k} v_i = 0 \tag{5.1}$$

where u_i is the i^{th} base out of $k\text{-}base$ information vector, and v_i is the i^{th} base out of $(n-k)$ parity bases. The addition sign (+) stands for modulo-5 addition. The k information bases are selected to include all the possible combinations of k number of bases (e.g for $k = 2$, there are 4^2 possible information bases: $AA,\ AG,\ AC,\ AU, \ldots, UU$). The parity bases are selected to include all the $(n-k)$-base subsequences of the Watson-Crick complement of the last 13 bases of the 3' end of $16S\ rRNA$ molecule given by $UAAGGAGGUGAUC$.

The input to the investigated code model is the $mRNA$ sequence. For analysis, the tested sequences ($\boldsymbol{S}_{mRNA}$) are selected to be comprised of q bases preceding the initiation codon ($AUG\ or\ GUG\ or\ UUG$) and $(q-3)$ bases from the coding region immediately following the initiation codon. Hence, the analysis sequence has the following form

$$\boldsymbol{S}_{mRNA} = \left[b_{-q},\ b_{-(q-1)}, \dots, A, U, G, b_3, b_4, \dots, b_{q-3}\right] \tag{5.2}$$

where each base is referenced with respect to its relative position from the initiation codon. The $2q$ -bases-long analysis sequence is then split up into $(2q-n+1)$ subsequences each of length (n) bases which are then compared to the set of codewords to decide which codeword is the correct one for each received sequence (r_p). The minimum distance metric used for the j^{th} received sequence at the i^{th} alignment is defined as

$$d^{j}_{min_i} = min\left[d(r_p, C)\right] \tag{5.3}$$

where p is a position relative to the initiation codon and C corresponds to the codebook. The minimum distance is recorded for each received sequence in the analysis stream based on different decoding strategies described in section 5.3. These distance metrics are used to evaluate how well the investigated code model captures the biological aspects of the initiation process.

The decoder (the ribosome) stores the minimum distance information (MDI) for each sequence group in matrices of the form

$$MDI = \begin{bmatrix} d^1_{min_{-q}} & d^1_{min_{-(q-1)}} & d^1_{min_{(q-2)}} & \cdots & d^1_{min_{(q-n)}} \\ d^2_{min_{-q}} & d^2_{min_{-(q-1)}} & d^2_{min_{(q-2)}} & \cdots & d^2_{min_{(q-n)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d^M_{min_{-q}} & d^M_{min_{-(q-1)}} & d^M_{min_{(q-2)}} & \cdots & d^M_{min_{(q-n)}} \end{bmatrix} \tag{5.4}$$

where M is the number of analysis sequences, and $(q - n)$ is the last valid comparison position in the analysis sequence between the received $mRNA$ sequence and the codewords.

To extract the information signal from the noise contained in the developed model, we take the average d_{min} in value by position for each sequence group. This produces one signal (MDI_{avg}) that describes the minimum distance characteristic of each sequence group

$$MDI_{avg} = \frac{1}{M}\sum MDI \tag{5.5}$$

The averaging used in equation (5.5) is a standard signal processing technique used to amplify a signal in the presence of noise. Averaging suppresses the noise in individual sequences and amplifies the common characteristics among all the sequences in a sequence group. Smaller distance values in the MDI_{avg} vector indicate stronger hydrogen bond formations between the $16S\ rRNA$ molecule and the $mRNA$ sequence.

5.3 Minimum Distance Decoders

A decoder provides a strategy for selecting the transmitted codeword for a given received sequence. One method, maximum likelihood decoding, compares the received sequence with every possible codeword sequence in the codebook and selects the most likely sequence. In this paper, we have used the following three distance decoders (i.e. three different strategies to calculate the minimum distance in equation (3)):

1. **Minimum Hamming Distance Decoder**

The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. Based on this definition, the

minimum Hamming distance is recorded for each received sequence (r_p) in the analysis base stream. This decoder strategy is described in Algorithm I.

Algorithm I Minimum Hamming Distance Decoding

Given: Codebook C with L codewords of length N and a received sequence S of length N from the received noisy $mRNA$ sequence. Notation: c_n^k is the n^{th} symbol of codeword k, s_n is the n^{th} symbol of S, d_H^k $(0 \leq k \leq L)$ is the hamming distance when codeword k is used (d_H^0 is initialized to 0). The decoder selects the codeword that results in the minimum hamming distance.

Minimum Hamming Distance

for $k = 1 \ldots L$ ***do***
  for $n = 1 \ldots N$ ***do***
    if $c_n^k \neq s_n$, **then**
      $d_H^k = d_H^{k-1} + 1$
    end if
  end for
end for
$d_{min} = \min(\boldsymbol{d_H})$, where $\boldsymbol{d_H} = [d_H^1, d_H^2, \ldots, d_H^L]$

2. Minimum Free Energy Distance Decoder

The proposed minimum free energy distance decoding strategy is summarized in Algorithm II.

Algorithm II Free Energy Distance Decoding

Given: Codebook C with L codewords of length N and a received sequence S of length N from the received noisy $mRNA$ sequence. Notation: c_n^k is the n^{th} symbol of codeword k, s_n is the n^{th} symbol of S, E_k $(0 \leq k \leq L)$ is the free energy metric when codeword k is used (E_0 is initialized to 0), and $Energy(a, b)$ is the energy dissipated on

binding with the nucleotide doublets ab. The decoder selects the codeword that results in the minimum free energy distance.

Minimum Free Energy

for $k = 1 \ldots L$ **do**
  for $n = 1 \ldots N - 1$ **do**
    if $c_n^k c_{n+1}^k = s_n s_{n+1}$, **then**
      $E_k = E_{k-1} + Energy(c_n^k c_{n+1}^k)$
    else
      $E_k = E_{k-1}$
    end if
  end for
end for
$E_{min} = \min(\boldsymbol{E})$, where $\boldsymbol{E} = [E_1, E_2, \ldots, E_L]$

The parameter E_{min} is the minimum free energy distance. $Energy(ab)$ is calculated using Table 5.2. The values refer to the free binding energy resulting from a bond between the listed dinucleotides and their complement. For example, the Watson–Crick bond between AA in the $mRNA$ and UU in the $16S\ rRNA$ yields a free energy of $-0.9\ kcal/mol$.

Table 5.2. Energy table (KCAL/MOL) [131]

Free Energy Doublets							
AA	-0.9	AG	-2.3	GA	-2.3	GG	-2.1
AU	-0.9	AC	-1.8	GU	-2.1	GC	-3.4
UA	-1.1	UG	-2.1	CA	-1.8	CG	-3.4
UU	-0.9	UC	-1.7	CU	-1.7	CC	-2.9

3. Minimum Exponentially-Weighted Free Energy Distance Decoder

The decoding strategy devised here represents an enhanced version of the one proposed in *B*. This strategy is summarized Algorithm III.

Algorithm III Exponentially-Weighted Free Energy Distance Ribosome Decoding

Given: Codebook C with L codewords of length N and a received sequence S of length N from the received noisy $mRNA$ sequence. Notation: c_n^k is the n^{th} symbol of codeword k, s_n is the n^{th} symbol of S, E_k $(0 \leq k \leq L)$ is the exponentially-weighted free energy when codeword k is used (E_0 is initialized to 0), and $Energy(a,b)$ is the energy dissipated on binding with the nucleotide doublets ab. w_k is the weight applied to the doublet in the k^{th} position. σ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches respectively, and ρ is an offset variable updated at each step. The parameter a is a constant that controls the exponential growth of the free energy E_k.

Minimum Exponentially-Weighted Free Energy

for $k = 1 \ldots L$ ***do***
  Initialize $\sigma_0 = 0, \tilde{\sigma} = 0, \rho_o = 0, w_1 = a$
  for $n = 1 \ldots N-1$ ***do***
    if $c_n^k c_{n+1}^k = s_n s_{n+1}$, **then**
      Increment $\sigma_n = \sigma_{n-1} + 1$
      Set $\tilde{\sigma}_n = 0$
      $w_n = \rho_n + a^{\sigma_n}$
      $\rho_n = \rho_{n-1}$
    else
      Increment $\tilde{\sigma}_n = \tilde{\sigma}_{n-1} + 1$
      Set $\sigma_n = 0$
      $v = w_1 - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n})$
      if $n \geq 2$, ***then***
        $v = w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n})$
      end if
      $w_n = max(0, v)$
      if $\rho_{n-1} \leq a$, ***then***
        $\rho_n = 0$
      else
        $\rho_n = \max\{w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}), 0\}$
      end if
    end if
    $E_n = E_{n-1} + w_n \times Energy(c_n^k c_{n+1}^k)$
  end if
end for

$E_{min} = \min(\boldsymbol{E})$, where $\boldsymbol{E} = [E_1, E_2, \ldots, E_L]$

$Energy(ab)$ is calculated using Table 5.1.

5.4 Input Data Preparation

The complete prokaryotic genome sequences required for the analysis in this chpater were obtained from the National Center for Biotechnology Information ($NCBI$) [117]. Using MATLAB, we developed a toolbox to extract and manipulate the data required and put it in a format suitable to our analysis. This toolbox will be disseminated through our research lab website to be publically available. Using this toolbox, we extracted the following data from the $NCBI$ for each tested genome: 1) the complete DNA sequence, 2) the exact locations of all known genes in the forward and reverse strands, 3) gene predictions obtained by $GeneMark$, 4) gene predictions obtained by $Glimmer$, and 5) the set of all possible open reading frames based on a pre-specified criteria. Based on this analysis, we were able to classify the tested data into four different groups (See Figure 5.1):

1. Actual Translated Sequences: Open reading frames which GenBank indicates as sequences that translate into proteins,
2. GeneMark Hypothetically Translated Sequences: Open reading frames which GeneMark indicates as genes but are actually not (GeneMark false positives),
3. Glimmer Hypothetically Translated Sequences: Open reading frames which Glimmer indicates as genes but are actually not (Glimmer false positives),

4. Non-Translated Sequences: Open reading frames which do not appear on the list of actually translated or hypothetically translated sequences. For this work, the open reading frame had to have: 1) A valid initiation codon; 2) A valid termination codon; 3) A sequence length greater than or equal to 99 bases.

The following block diagrams give an illustrative description of the approach used to prepare the data required for analysis.

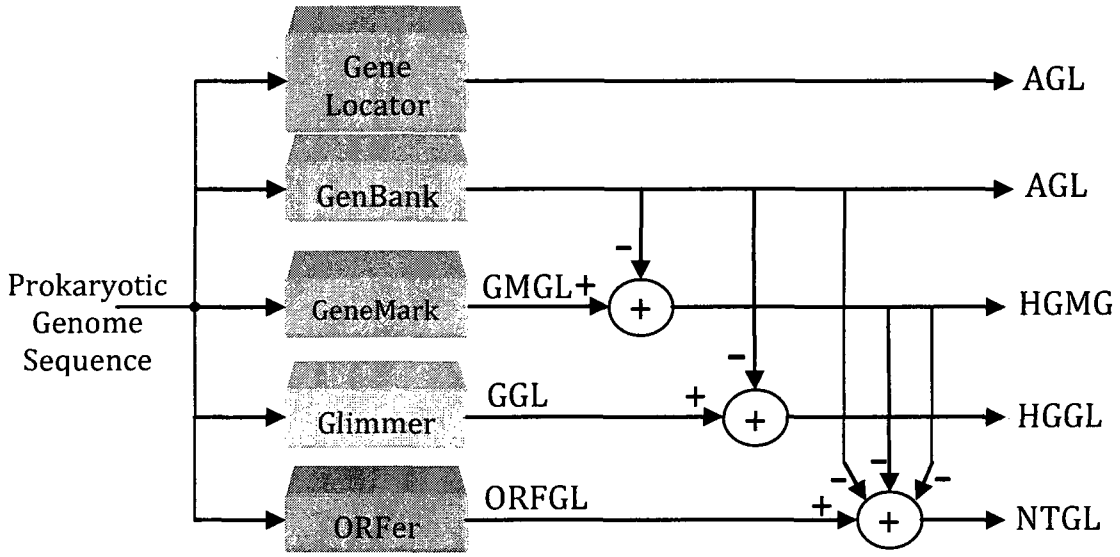


Figure 5.1. Schematic diagram of the analysis data preparation

The notations used is Figure 5.1 are: *AGL* (Actual Gene Locations) corresponds to group 1, *HGMGL* (Hypothetical GeneMark Gene Locations) corresponds to group 2, *HGGL* (Hypothetical Glimmer Gene Locations) corresponds to group 3, and *NTGL* (Non-Translated Gene Locations) corresponds to group 4. The intermediate parameter *GMGL* (*GeneMark* Gene Locations) corresponds to the gene predictions obtained by *GeneMark*, *GGL* (*Glimmer* Gene Locations) corresponds to the gene predictions obtained by Glimmer, and *ORFGL* (Open Reading Frame Gene Locations) corresponds to the set of all possible genes that are greater than 99 bases long and start with a valid start

codon ($AUG\ or\ GUG\ or\ UUG$) and end up with a valid stop codon ($UUA\ or\ UGA$ or UAG).

Figure 5.2 shows a schematic diagram of the proposed code model. The input parameters are 1) n and k: the code parameters, 2) AGL, $HGMGL$, $HGGL$, and $NTGL$ obtained in figure 5.1) the last 13 bases of the 3' end of $16S\ rRNA$ molecule. The parameter q is the number of bases upstream of the initiation codon where the beginning of our analysis tested sequences was selected, m is an input parameter to determine which of the four inputs ($AGL/\ HGMGL/\ HGGL/\ NTGL$) to select, N is the number of codewords in the codebook, M is the number of analysis sequences being processed (different for each input).

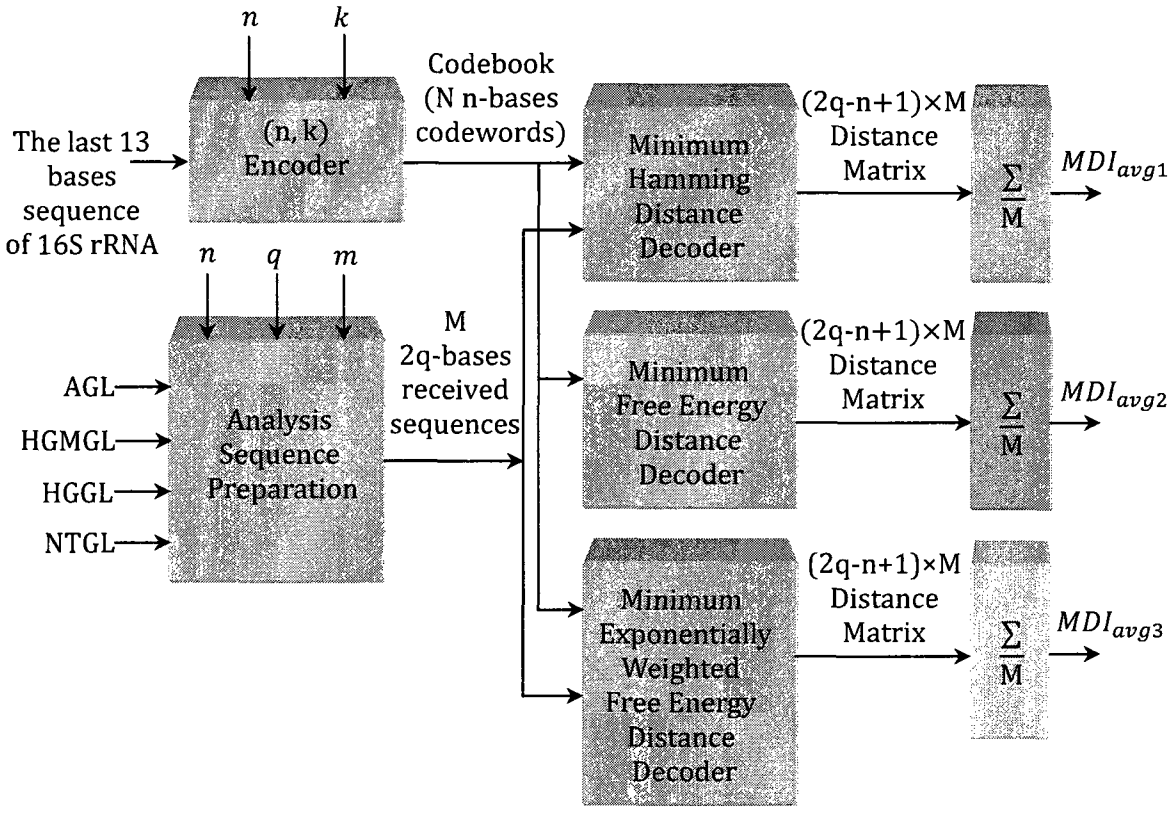


Figure 5.2. Schematic diagram of the proposed augmented block code model

5.5 Simulation Results

For analysis, we have applied the developed model to *Escherichia coli* $K-12$ $MG1655$ strain, *Escherichia coli* $O157:H7$ strain, and to other prokaryotic organisms of varying taxonomical relation to *E. coli* including: *Salmonella Typhimurium LT2*, *Bacillus Subtilis*, and *Staphylococcus Aureus Mu50* (The simulation results shown in Figures 3 and 4 are obtained for *Salmonella Typhimurium LT2*). The model was able to successfully identify and distinguish the regions on the 5' untranslated leader regions where the minimum free energy distance values of translated *mRNA* subsequences, hypothetically translated subsequences (obtained by *GeneMark* and *Glimmer*), and non-translated subsequences differ the most. These regions correspond to the *Shine − Dalgarno* domain and the non-random domain that exists in the genomic structure of the tested sequences.

Figures 5.3-5.8 clearly show significant differences between the translated, hypothetically translated and the non-translated sequence groups. The horizontal axis in figures 5.3-5.8 represent the position relative to the first base of the initiation codon. The vertical axis shows the mean of the aligned minimum distance values of the sequences in each of the four sequence groups. All the sequence groups in the (5,2) and (8,2) models achieve a global minimum distance value in the -5 to 0 region. The -15 to 0 region contains large synchronization signals which can be used to determine valid protein coding sequences or frames. There are also smaller synchronization signals outside the -15 to 0 region which seem to oscillate with a frequency of three. The results of the longer (8,2) block code model in Figure 5.4 illustrate the effect of two or more codons while the (5,2) block code model is affected by at most two codons.

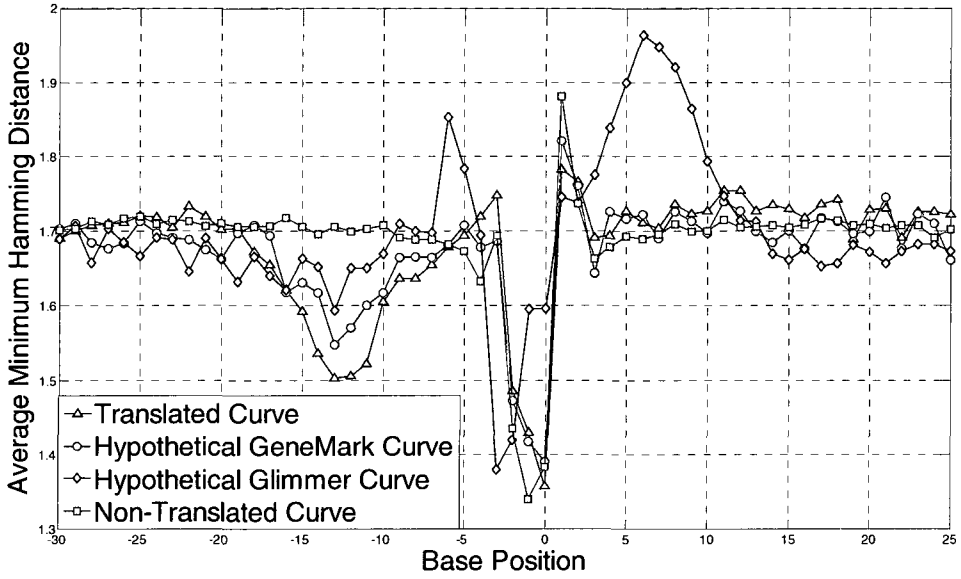


Figure 5.3. (5,2) code model output for Salmonella Typhimurium LT2 using minimum Hamming distance decoder

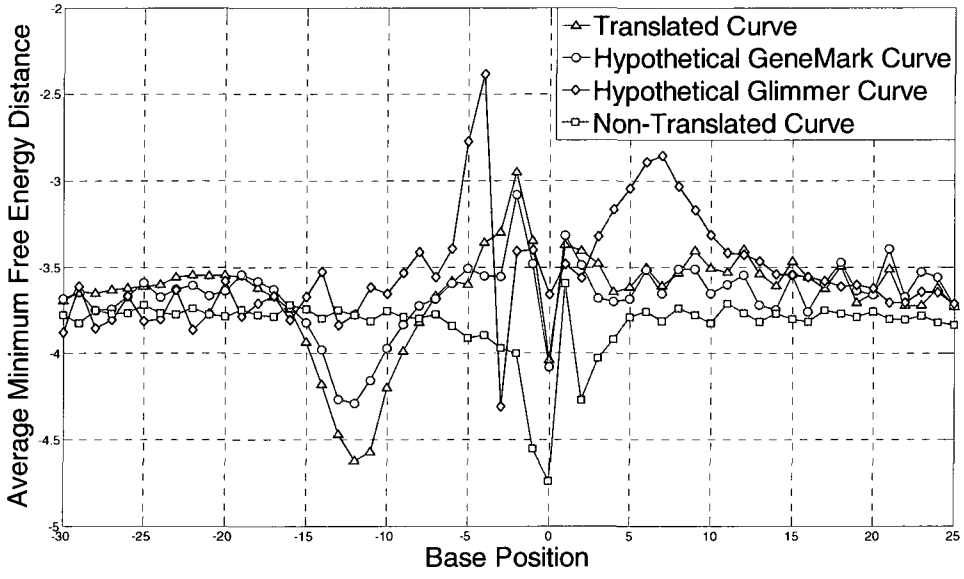


Figure 5.4. (5,2) code model output for Salmonella Typhimurium LT2 using minimum free energy distance

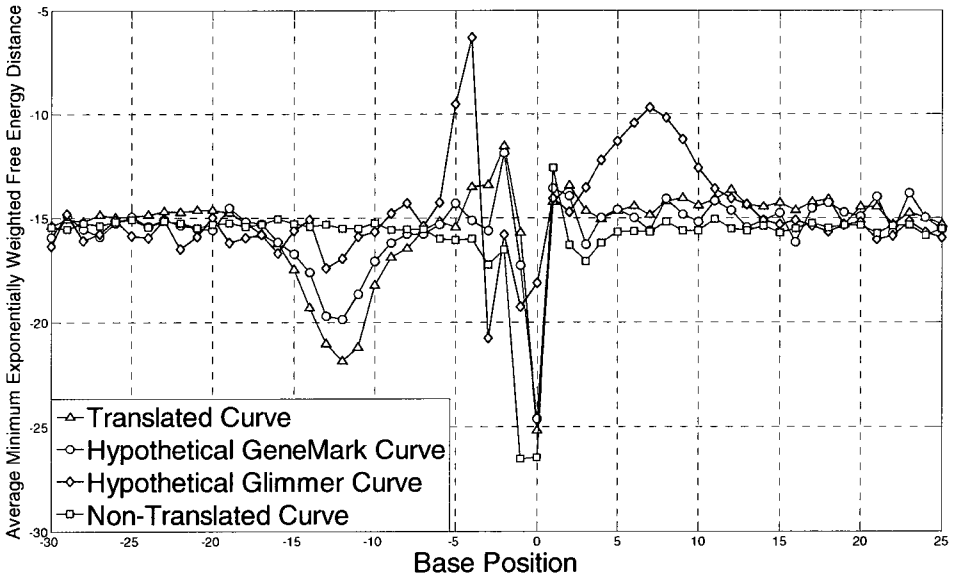


Figure 5.5. (5,2) code model output for Salmonella Typhimurium LT2 using minimum exponentially-weighted free energy distance decoder

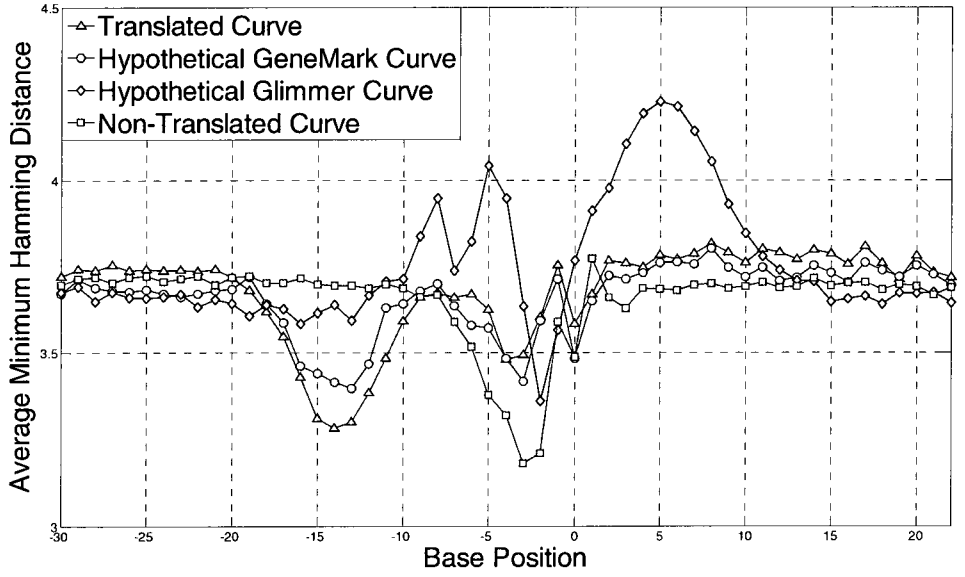


Figure 5.6. (8,2) code model output for Salmonella Typhimurium LT2 using minimum Hamming distance decoder

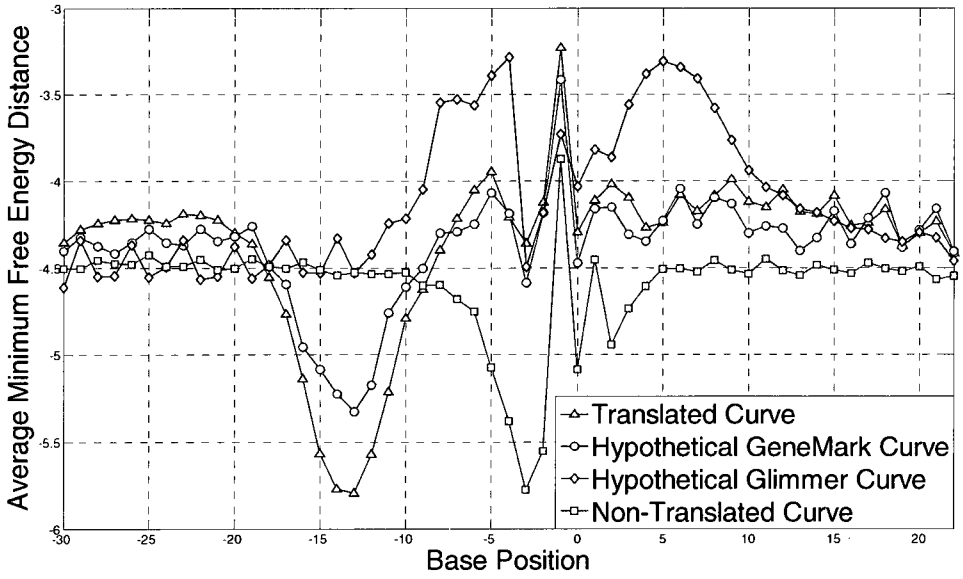
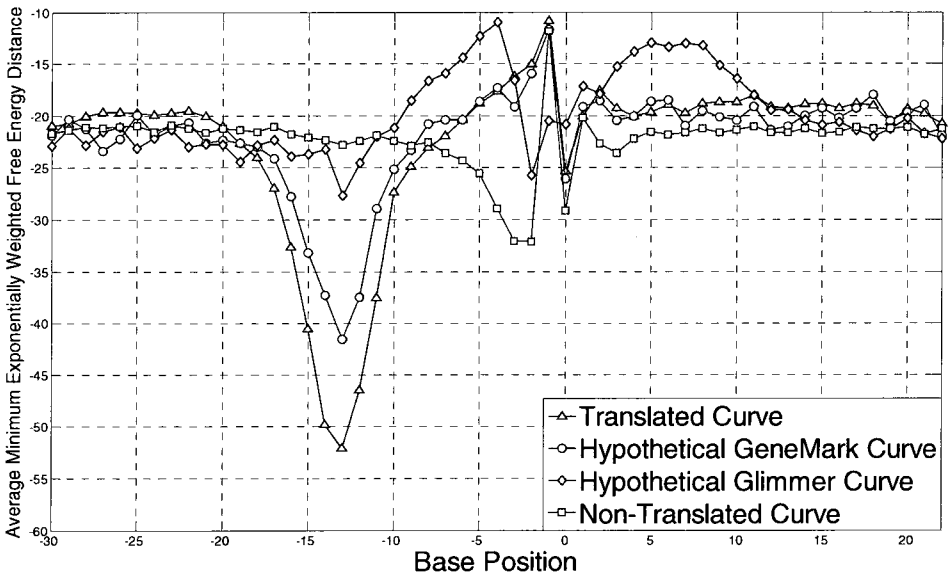


Figure 5.7. (8,2) code model output for Salmonella Typhimurium LT2 using minimum free energy distance decoder



(f)

Figure 5.8. (8,2) code model output for Salmonella Typhimurium LT2 using minimum exponentially-weighted free energy distance decoder

The simulation results obtained for the other prokaryotic organisms tested in this work are quite similar to the ones shown in figures 5.3-5.8. These organisms include *Escherichia coli K − 12 MG165*, *Escherichia coli O157:H7*, *Salmonella Typhimurium LT2*, *Subtilis*, and *Staphylococcus Aureus Mu50*. Simulation results are shown is figures 5.9-5.13 for the five organisms respectively. This further certifies the correctness and the biological relevance of the proposed block code model to identify and distinguish the four different test groups (translated, hypothetically translated, and non-translated test groups).

The results of this work suggest that it is possible to design a coding based algorithm for distinguishing between protein coding and non-protein coding genomic sequences by "decoding" the *mRNA* leader region. The success of this work can lead to the development of improved methods for identifying the precise location of translation initiation start sites. Additionally, design of effective coding-based models for genetic regulatory systems can potentially help researchers determine how to incorporate deliberate, sequence-controlled regulation into engineered proteins. Such a tool would be useful for designing regulatory sequences for transgenic organisms, as well as further our understanding of the translation regulatory mechanisms.

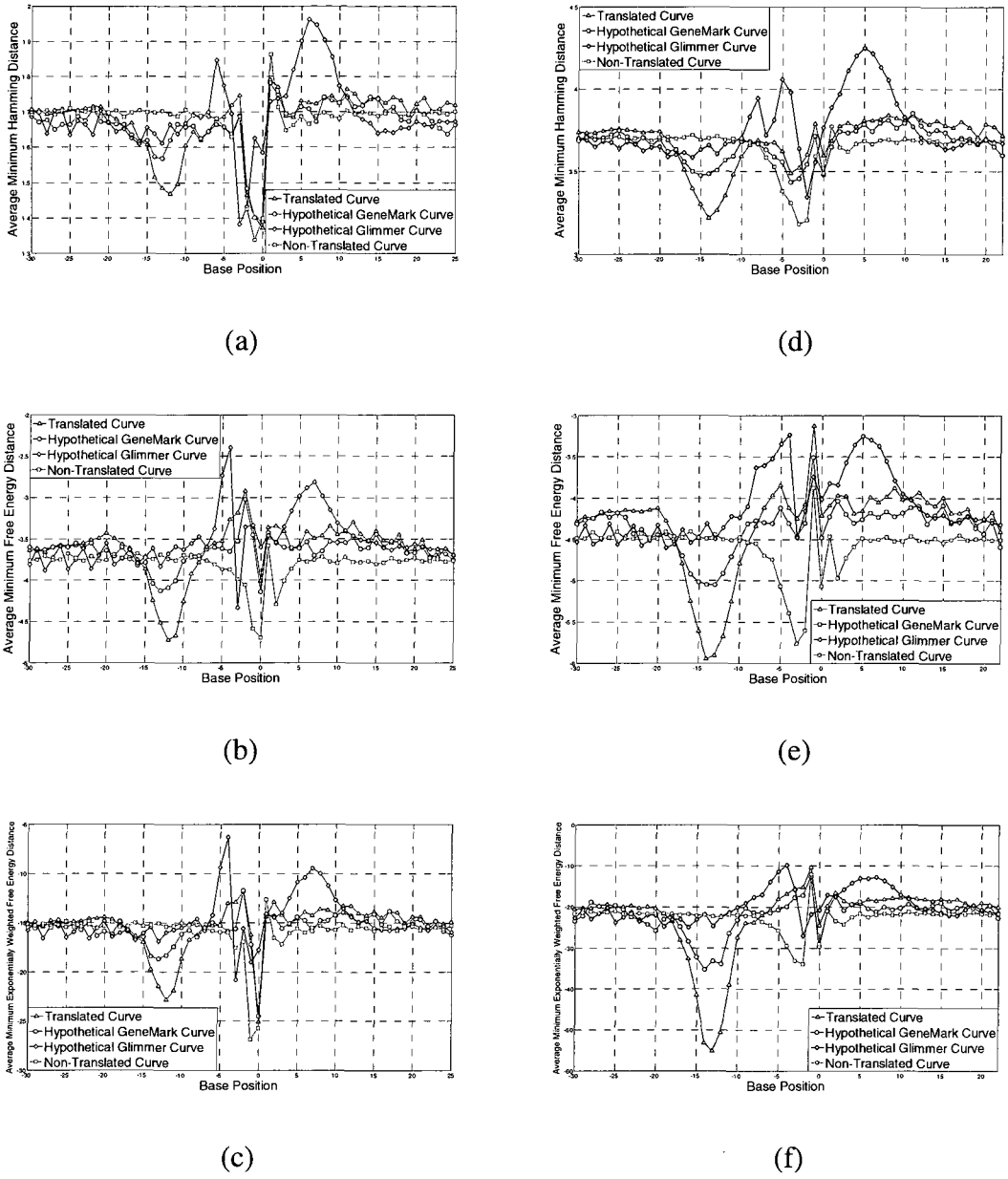


Figure 5.9. Block Code Model output applied to *Escherichia Coli K*-12 substrain MG1655 using (a) a (5,2) code with minimum Hamming distance decoder, (b) a (5,2) code with minimum free energy distance decoder, (c) a (5,2) code with minimum exponentially weighted free energy Distance Decoder, (d) a (8,2) code with minimum Hamming distance decoder, (e) a (8,2) code with minimum free energy distance decoder, and (f) a (8,2) code with minimum exponentially weighted free energy Distance Decoder

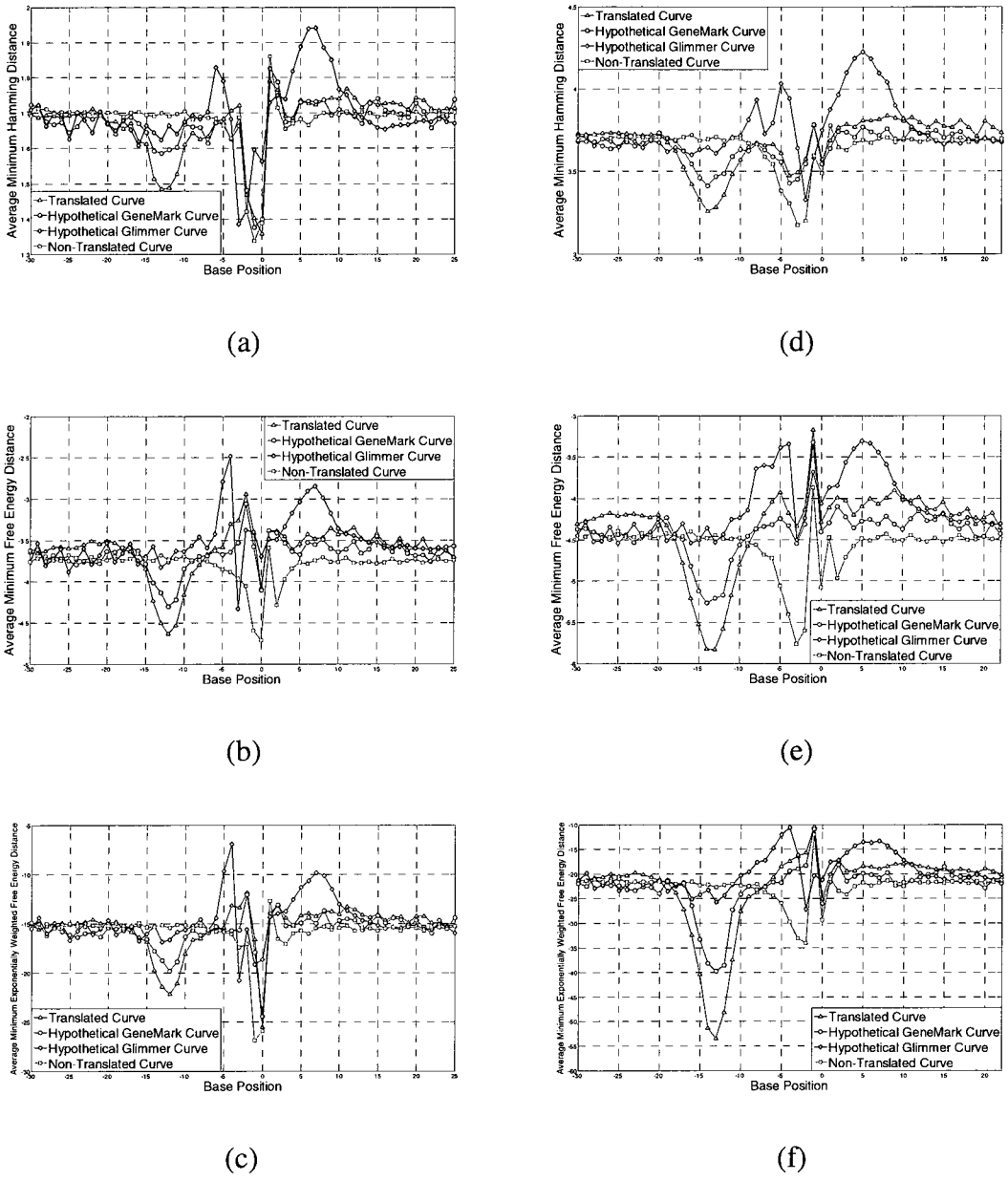


Figure 5.10. Block Code Model output applied to *Escherichia Coli O157H7* substrain using (a) a (5,2) code with minimum Hamming distance decoder, (b) a (5,2) code with minimum free energy distance decoder, (c) a (5,2) code with minimum exponentially weighted free energy Distance Decoder, (d) a (8,2) code with minimum Hamming distance decoder, (e) a (8,2) code with minimum free energy distance decoder, and (f) a (8,2) code with minimum exponentially weighted free energy Distance Decoder

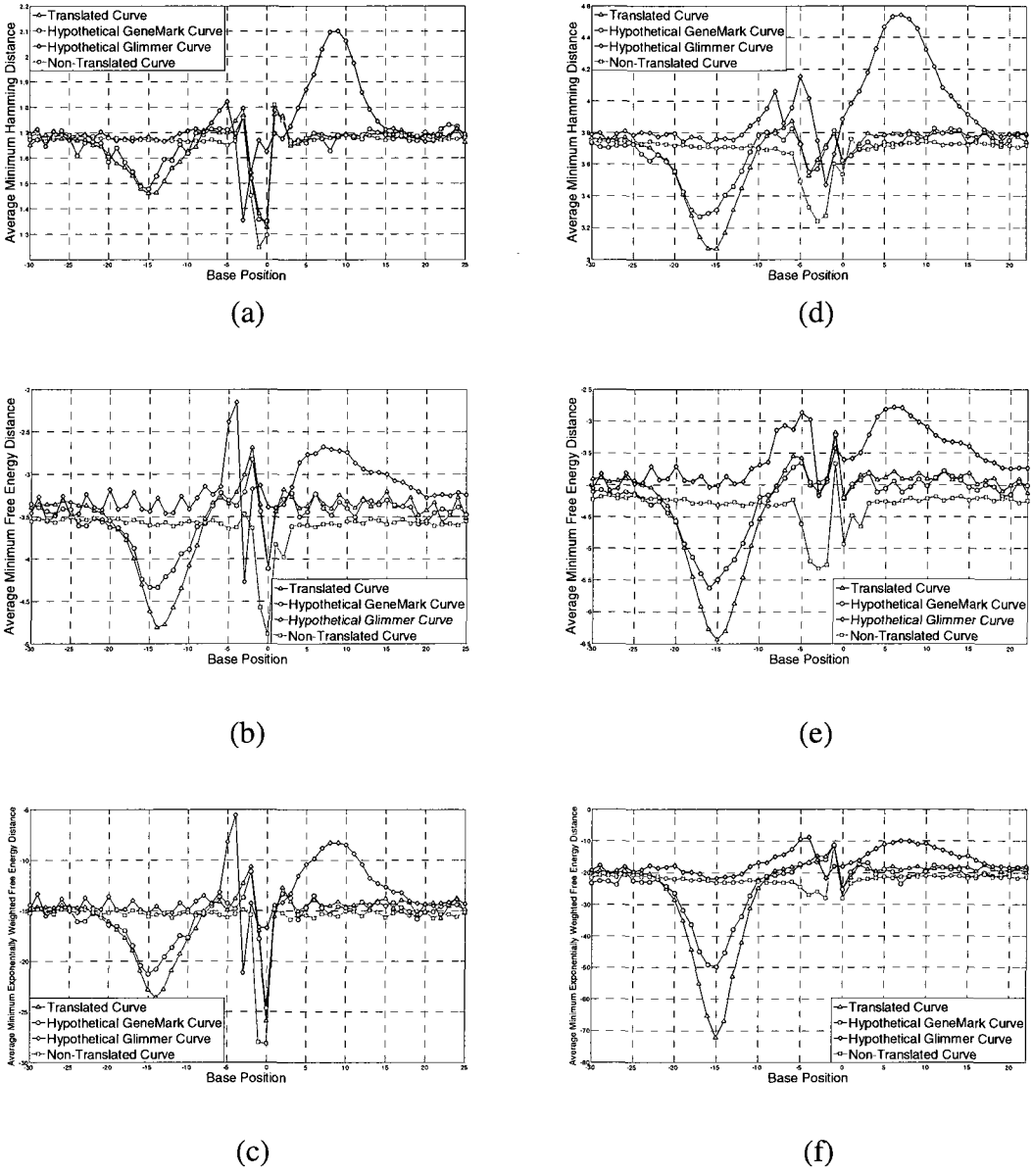


Figure 5.11. Block Code Model output applied to *Bacillus Subtilis* using (a) a (5,2) code with minimum Hamming distance decoder, (b) a (5,2) code with minimum free energy distance decoder, (c) a (5,2) code with minimum exponentially weighted free energy Distance Decoder, (d) a (8,2) code with minimum Hamming distance decoder, (e) a (8,2) code with minimum free energy distance decoder, and (f) a (8,2) code with minimum exponentially weighted free energy Distance Decoder

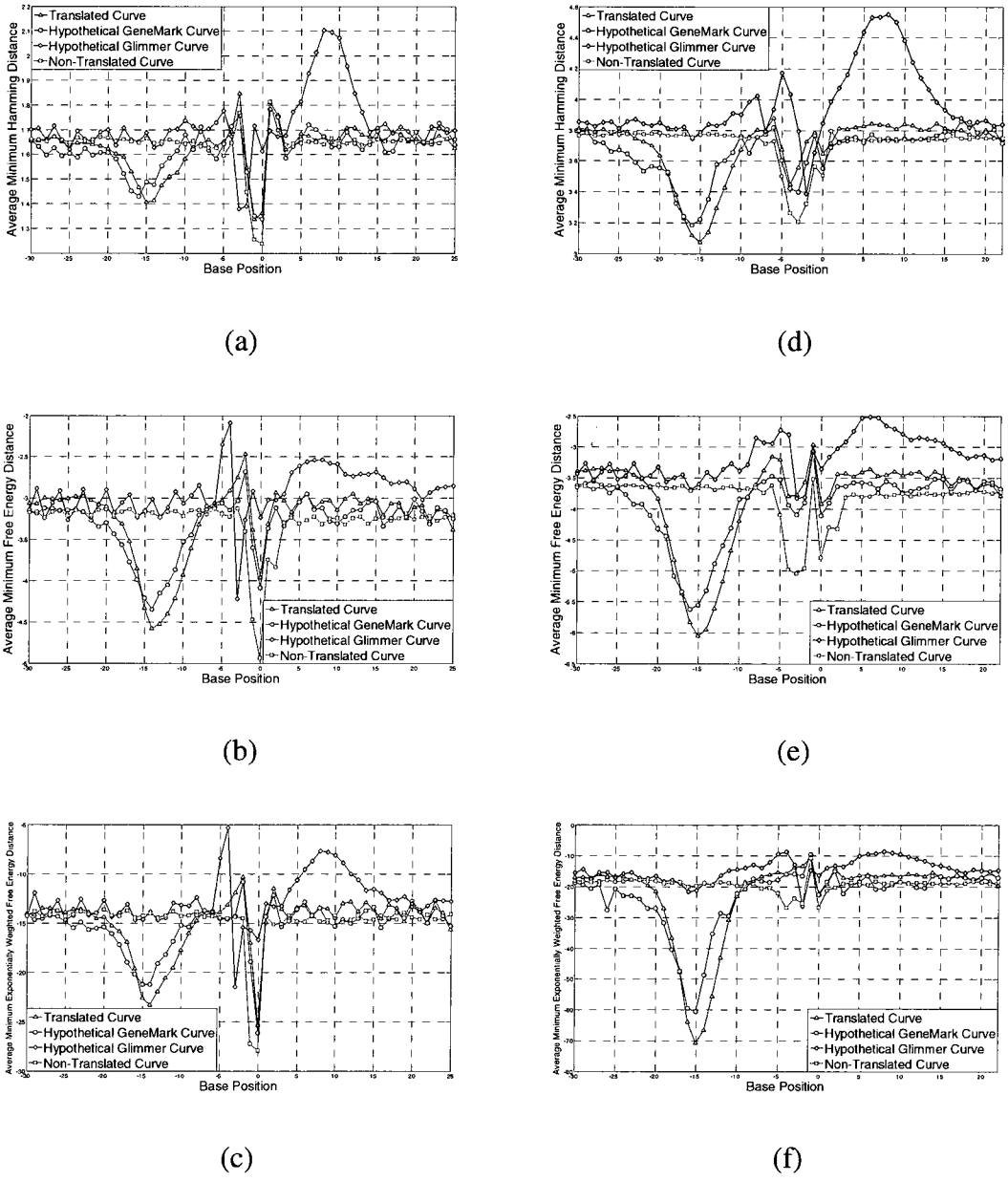


Figure 5.12. Block Code Model output applied to *Staphylococcus Aureus Mu*50 using (a) a (5,2) code with minimum Hamming distance decoder, (b) a (5,2) code with minimum free energy distance decoder, (c) a (5,2) code with minimum exponentially weighted free energy Distance Decoder, (d) a (8,2) code with minimum Hamming distance decoder, (e) a (8,2) code with minimum free energy distance decoder, and (f) a (8,2) code with minimum exponentially weighted free energy Distance Decoder

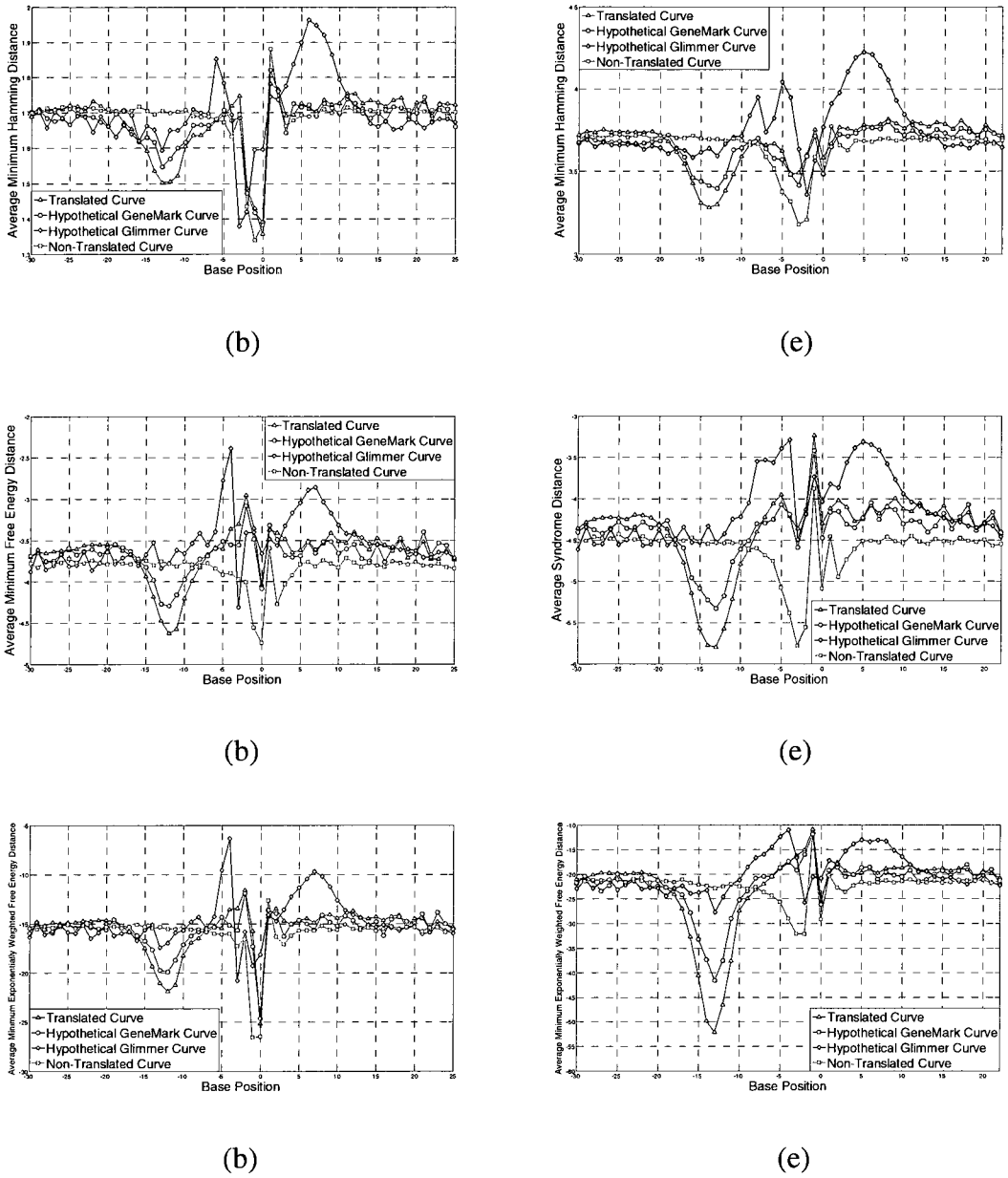


Figure 5.13. Block Code Model output applied to *Salmonella Typhimurium LT2* using (a) a (5,2) code with minimum Hamming distance decoder, (b) a (5,2) code with minimum free energy distance decoder, (c) a (5,2) code with minimum exponentially weighted free energy Distance Decoder, (d) a (8,2) code with minimum Hamming distance decoder, (e) a (8,2) code with minimum free energy distance decoder, and (f) a (8,2) code with minimum exponentially weighted free energy Distance Decode

CHAPTER 6

A CONVOLUTIONAL CODE BASED ANALYSIS FOR GENE TRANSLATION USING TABLE- BASED DECODING

Redundancy occurs naturally within RNA and DNA sequences [76]. The existence of tandem repeats, and sequences such as the $Shine\text{-}Dalgarno$ sequence, the $Pribnow\ box$ and the $TATA\ box$, leads to the assumption that biological communication systems use some method of coding to recognize specific information regions within a nucleotide sequence and possibly correct for "transmission" errors such as mutations. In this chapter, we investigate a convolutional code model to analyze the process of translation using table-based decoding [21, 22, 33, 37]. In the investigated model, the messenger RNA ($mRNA$) sequence can be viewed as a noisy convolutional encoded signal, and ribosome as a table-based convolutional decoder. The 16S ribosomal RNA ($16S\ rRNA$) sequence is used to form decoding masks for table-based decoding. The model was tested on five different $E.coli$ bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs (e.g. $GeneMark$ [19] and $Glimmer$ [31]).

6.1 Introduction

Viewing protein synthesis as an information-processing system allows nucleotide sequences to be analyzed as messages [125]. May's model defines a genetic channel as the DNA replication, transcription and translation process during which errors may be introduced, detected and possibly corrected [91]. The transfer of biological information

can be modeled as a communication channel, with the DNA sequence as input, and the polypeptide amino acid sequence as the output. As in error-protected information channels, redundancy occurs within genomic sequences. The DNA encoded message, in its double-helix form, is doubly redundant. Hence, the DNA sequence can be viewed as a half-rate systematic code. At the translation level, messenger RNA ($mRNA$) sequences are mapped to chains of amino-acids by grouping every three nucleotides (codon) together to form an amino acid. This process can be viewed as the decoding of a rate one-third code, i.e., each amino acid is encoded using three "parity" items, or it could be viewed as a code of different (non-systematic) rate depending on how much error detection capability is assumed to be present.

It is also known that leader regions of the messenger RNA ($mRNA$), and other prokaryotic regulatory regions contain consensus sequences which in some way signal or control translation. Examples are Shine-Dalgarno (SD) sequence in bacterial $mRNA$, and the $TATA\ box$ and the $Pribnow\ box$ in double helix DNA [76]. Specifically, the SD sequence always occurs before the start codon (usually AUG) of the $mRNA$ sequence in $E.coli$. If there is a method in place that verifies for the validity of the leader sequence (which includes SD), the ribosome should have a mechanism of recognizing it, and hence be able to correctly initiate the translation process. Assuming there is a validating relationship among the leader sequence bases and/or codons, and assuming that the ribosome has an exposed region which is in contact with the $mRNA$ leader for validation purposes, one may speculate that the leader sequence may have embedded in it (or may be modeled as) an error detecting code such as a block code (discussed in Chapter 5) or as a convolutional code (investigated in this chapter). In *Escherichia Coli*, the exposed

part of the 16S ribosomal RNA $(16S\,rRNA)$ binds with the $mRNA$ leader sequence during the initiation phase of the translation process. The same exposed part appears to remain in contact with $mRNA$ during the translation process.

In the sections which follow, we give a brief overview of convolutional coding and table-based decoding. We also discuss the relationship to the genetic process, and present the methodology for forming a table-based decoder. Simulation results and possible implications of the model are analyzed and discussed.

6.2 Theoretical Background

In Convolutional encoding, an n-bit encoded block at time i depends on the k-bit information block at time i and on m previous information blocks [36]. Hence, a convolutional encoder requires memory. Convolutional codes are referred to as (n, k, m) codes or (n, k) codes. A codeword is the output of the convolutional encoder for a given input data block [80]. A decoder provides a strategy for selecting an estimated codeword for each possible received sequence.

6.2.1 Table-Based Convolutional Coding. A rate $(\boldsymbol{k/n})$ convolutional code will have $\boldsymbol{n}$ generators, each of which will operate on $\boldsymbol{L}$ input symbols at a time. The length $(\boldsymbol{L = m + 1})$, which is also the length of each generator, is called the constraint length. The encoded data is also referred to as parity. Each generator is applied on the input data and the encoded symbols are calculated. Then the generators are shifted by $\boldsymbol{k}$ positions to the right, and the procedure is repeated until the whole input data are being encoded.

The table-based decoding method is based on the existence of a one-to-one mapping between the information symbols and the encoded symbols. Though the

encoded symbols are generated at a rate different from the information rate, it is possible to find a relationship between the two which involves the same number of symbols. For a rate (k/n) code with L constraint length, the first L information symbols produce n encoded symbols. Each additional k information symbols produce n encoded symbols as shown in Table 6.1. A necessary condition for the one-to-one mapping is that the number of information symbols and encoded symbols are equal. By having such requirement in the j^{th} iteration, we have

$$L + jk = n + jn. \tag{6.1}$$

Solving for j yields

$$j = \frac{L - n}{n - k}. \tag{6.2}$$

Thus, a set of w-symbol information block must correspond uniquely to a set of w-symbol encoded block, where w (the window length) is defined as

$$w = L + jk = n\frac{L - n}{n - k}. \tag{6.3}$$

Table 6.1. One-to-one mapping in table-based coding

Iteration	*Number of Information symbols*	*Number of encoded symbols*
0	L	n
1	$L + k$	$n + n$
2	$L + 2k$	$n + 2n$
⋮	⋮	⋮
j	$L + jk$	$n + jn$

To illustrate the table-based encoding procedure, let the generator vectors of a rate $(1/2)$ code with $(L = 3)$ constraint length be given by

$$\boldsymbol{g}^{(1)} = \left[g_0^{(1)}, g_1^{(1)}, g_2^{(1)} \right], \tag{6.4}$$

$$\boldsymbol{g}^{(2)} = \left[g_0^{(2)}, g_1^{(2)}, g_2^{(2)} \right]. \tag{6.5}$$

Let the input information sequence be given by

$$\boldsymbol{u} = (u_0, u_1, u_2, \dots). \tag{6.6}$$

Each generator is then applied to the input information sequence ($\boldsymbol{u}$) and the first two encoded symbols are calculated as

$$v_0^{(1)} = u_0 g_0^{(1)} + u_1 g_1^{(1)} + u_2 g_2^{(1)}, \tag{6.7}$$

$$v_0^{(2)} = u_0 g_0^{(2)} + u_1 g_1^{(2)} + u_2 g_2^{(2)}. \tag{6.8}$$

Next, the generators are shifted by ($k = 1$) positions to the right along the input information sequence, and the next two encoded symbols are calculated. This procedure is repeated until all the information symbols are being encoded. The next encoded symbols are given by

$$v_1^{(1)} = u_1 g_0^{(1)} + u_2 g_1^{(1)} + u_3 g_2^{(1)}, \tag{6.9}$$

$$v_1^{(2)} = u_1 g_0^{(2)} + u_2 g_1^{(2)} + u_3 g_2^{(2)}, \tag{6.10}$$

$$v_2^{(1)} = u_2 g_0^{(1)} + u_3 g_1^{(1)} + u_4 g_2^{(1)}, \tag{6.11}$$

$$v_2^{(2)} = u_2 g_0^{(2)} + u_3 g_1^{(2)} + u_4 g_2^{(2)}. \tag{6.12}$$

At the end of the encoding process, the encoded sequence or the parity stream ($\boldsymbol{v}$) will look like

$$\boldsymbol{v} = \left[v_0^{(1)}, v_0^{(2)}, v_1^{(1)}, v_1^{(2)}, v_2^{(1)}, v_2^{(2)}, \dots \right]. \tag{6.13}$$

It is worth mentioning that the arithmetic operations involved in evaluating the encoded symbols before should be carried out in the corresponding Galois field. For example,

Table 6.2 shows the corresponding arithmetic operations in $GF(5)$ field, where "NaN" is a flag indicating division by zero.

Table 6.2. Arithmetic operation of $GF(5)$

+	**0**	**1**	**2**	**3**	**4**
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

−	**0**	**1**	**2**	**3**	**4**
0	0	4	3	2	1
1	1	0	4	3	2
2	2	1	0	4	3
3	3	2	1	0	4
4	4	3	2	1	0

×	**0**	**1**	**2**	**3**	**4**
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

÷	**0**	**1**	**2**	**3**	**4**
0	NaN	0	0	0	0
1	NaN	1	3	2	4
2	NaN	2	1	4	3
3	NaN	3	4	1	2
4	NaN	4	2	3	1

A way to recognize a correct sequence is to form the so called syndrome. The syndrome vector is zero if there are no detectable errors in the parity; otherwise, the syndrome value will be nonzero. The g-mask provides an efficient method for syndrome vector generation [22]. The g-mask is $(w + n)$ symbols long. Once the g-mask has been constructed, it can be used to calculate the syndrome vector for the parity stream. The g-mask is first ANDed with the first $(w + n)$ parity symbols. The result is then XORed to produce the first syndrome value. Next, the received parity stream is shifted by n bits, and the process is repeated until all syndrome values are produced. Based on the value of the syndrome vector, the received parity sequences can be used to estimate the transmitted sequence or used to detect errors in the transmission.

6.3 Convolutional Code Model

Figure 6.1 gives a general block diagram of the convolutional code model investigated in this work. Figure 6.2 gives a detailed description of Figure 6.1.

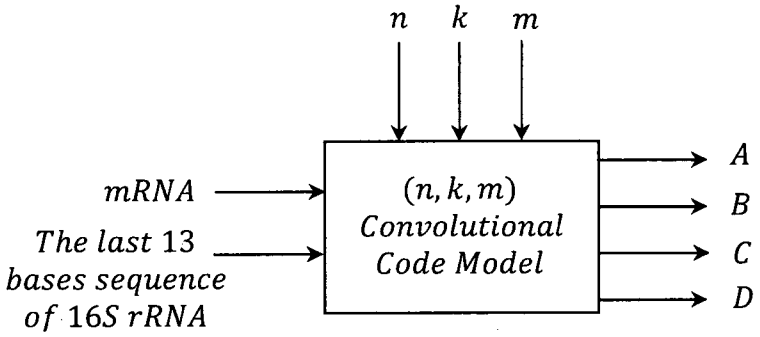


Figure 6.1. A schematic diagram of the investigated convolutional code model

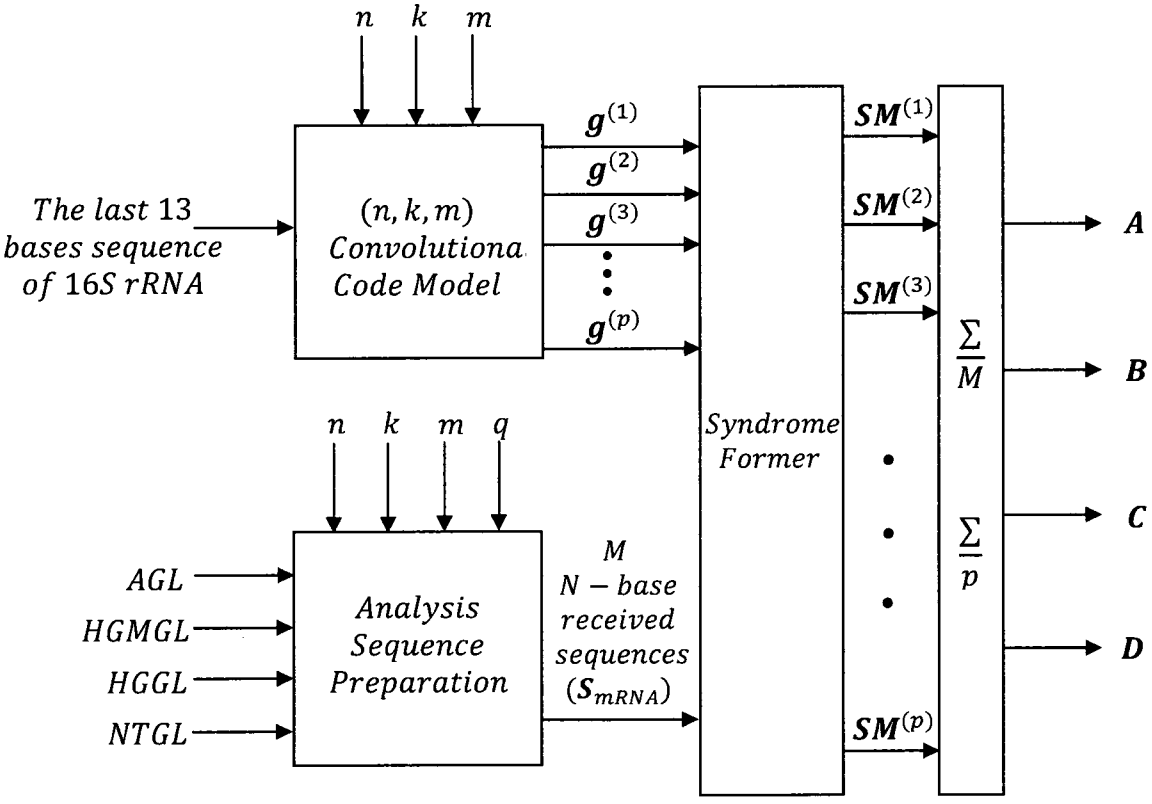


Figure 6.2. A detailed schematic diagram of Figure 6.1

In the investigated convolutional code model for the process of translation, the messenger RNA ($mRNA$) sequence is considered as a received parity sequence of a convolutional encoded data stream. The coding alphabet is defined using principles of base pairing, wobble pairing, and translation initiation information [76]. The RNA bases are mapped to the field of five a follows: $Inosine(I) = 0$, $Adenine(A) = 1$, $Guanine(G) = 2, Cytosine(C) = 3, Uracil(U) = 4$. Multiplication and addition are modulo-5. The RNA bases are defined so that in modulo-5 addition of the bases that pair is zero.

The interaction between $mRNA$ and the $16S\ rRNA$ molecule significantly influences the initiation of the translation process. When a base on the $mRNA$ pairs with a base on the $16S\ rRNA$, hydrogen bonds are formed. The greater the number of consecutive pairings formed between these two RNA molecules, the greater the probability that the translation process with be initiated. Every time the 16S ribosomal subunit attaches to the $mRNA$, a bonding pattern is formed. The bonding pattern that results in a positive signal is the bonding pattern with high numbers of consecutive hydrogen bonds. This process of locating regions on the $mRNA$ which form high numbers of consecutive hydrogen bonds can be modeled as locating parity blocks which produce zero syndrome values for a received parity stream. Hence, the ribosome can be modeled as a convolutional decoder that follows a certain decoding strategy. In this work, we will be using the table-based decoding strategy described in section 6.2.1 to model the mechanism that the ribosome uses to bind to the $mRNA$ sequence and correctly initiate translation.

6.3.1 Syndrome Calculation. For the investigated model, the syndrome-former, or the g-mask, will be derived from the last 13 bases of the 3′ end of the $16S\ rRNA$ sequence given by$(3' - AUUCCUCCACUAG \ldots 5')$. Considering a (n, k, m) code, the length of the g-mask is given by

$$g_L = w + n, \tag{6.14}$$

where w is given by equation (6.3), and the constraint length (L) is given by

$$L = m + 1. \tag{6.15}$$

Table 6.3 shows all the (n, k, m) possibilities that have been considered in this work for analysis. The table also shows the corresponding g-mask sequences. To illustrate how to calculate the syndrome values for a given g-mask sequence, let the received $mRNA$ sequence be selected to follow the form

$$\boldsymbol{S_{mRNA}} = \left[b_{-q},\ b_{-(q-1)}, \ldots, b_{-1}, A, U, G, b_3, b_4, \ldots, b_{q-3}\right], \tag{6.16}$$

where each base (b) is referenced with respect to its relative position from the initiation codon. The $2q$-bases-long analysis sequence $(\boldsymbol{S_{mRNA}})$ is then decoded by the k^{th} g-mask $\boldsymbol{g}^{(k)}$ which operates on $(w + n)$ bases at a time and then shifts by (n) bases. At the j^{th} alignment of the k^{th} g-mask with the i^{th} $mRNA$ sequence, the corresponding $mRNA$ $(w + n)$-bases subsequence is given by

$$\boldsymbol{r}_i^j = \left[b^{(i)}_{-q+(j-1)n},\ b^{(i)}_{-(q-1)+(j-1)n}, \ldots,\ b^{(i)}_{-(q-g_L+1)+(j-1)n}\right], \tag{6.17}$$

and the k^{th} g-mask is given by

$$\boldsymbol{g}^{(k)} = \left[g_0^{(k)}, g_1^{(k)}, \ldots, g_{g_L-1}^{(k)}\right], \tag{6.18}$$

where the elements of both $\boldsymbol{r}_i^j$ and $\boldsymbol{g}^{(k)}$ belong to $\{A, G, C, U\}$.

Table 6.3. Different (n, k, m) convolutional code models

(n,k,m)	L	w	$w+n$	$Number\ of\ g-masks$	G_mask
(2,1,2)	3	4	6	8	'AUUCCU' 'UUCCUC' 'UCCUCC' 'CCUCCA' 'CUCCAC' 'UCCACU' 'CCACUA' 'CACUAG'
(2,1,3)	4	6	8	6	'AUUCCUCC' 'UUCCUCCA' 'UCCUCCAC' 'CCUCCACU' 'CUCCACUA' 'UCCACUAG'
(2,1,4)	5	8	10	4	'AUUCCUCCAC' 'UUCCUCCACU' 'UCCUCCACUA' 'CCUCCACUAG'
(2,1,5)	6	10	12	2	'AUUCCUCCACUA' 'UUCCUCCACUAG'
(3,1,2)	3	3	6	8	'AUUCCU' 'UUCCUC' 'UCCUCC' 'CCUCCA' 'CUCCAC' 'UCCACU' 'CCACUA' 'CACUAG'
(3,1,4)	5	6	9	5	'AUUCCUCCA' 'UUCCUCCAC' 'UCCUCCACU' 'CCUCCACUA' 'CUCCACUAG'
(3,1,6)	7	9	12	2	'AUUCCUCCACUA' 'UUCCUCCACUAG'

Equations $(6.16 - 6.17)$ are then mapped to their corresponding $GF(5)$ equivalents. Hence, the corresponding vector elements will now belong to $\{0,1,2,3\}$ instead of $\{A, G, C, U\}$. The j^{th} syndrome value of the i^{th} $mRNA$ subsequence using the k^{th} g-mask is then calculated as

$$S_{ijk} = \sum_{l=0}^{g_L-1} b^{(i)}_{-(q-l)+(j-1)n}\, g^{(k)}_l \tag{6.19}$$

Figure 6.3 illustrates the calculation of the j^{th} syndrome value as described by equation (6.19) where the multiplication and the addition operations are carried out in $GF(5)$ as given in table (6.2).

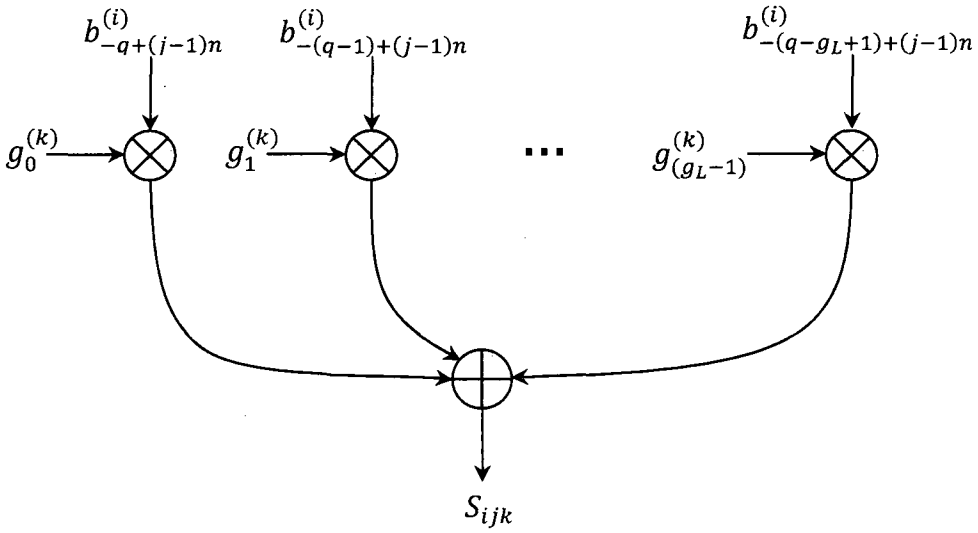


Figure 6.3. Syndrome calculation using the g-mask

After calculating the syndrome values for all $mRNA$ analysis sequences, the resulting k^{th} syndrome matrix can be written as

$$\boldsymbol{SM}^{(k)} = \begin{bmatrix} S_{11k} & S_{12k} & S_{13k} & \cdots & S_{1(N-1)k} & S_{1Nk} \\ S_{21k} & S_{22k} & S_{23k} & \cdots & S_{2(N-1)k} & S_{2Nk} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ S_{M1k} & S_{M2k} & S_{M3k} & \cdots & S_{M(N-1)k} & S_{MNk} \end{bmatrix}_{M\times N} \tag{6.20}$$

where M is the number of $mRNA$ analysis sequences and N is the number of alignments between the k^{th} g-mask and the $mRNA$ analysis sequence.

By averaging the k^{th} syndrome matrix $\boldsymbol{SM}^{(k)}$ obtained in equation (6.20) over the number of analysis sequences (M), we get

$$\boldsymbol{SM}_{avg}^{(k)} = \frac{\boldsymbol{SM}^{(k)}}{M}, \tag{6.21}$$

And by averaging $\boldsymbol{SM}_{avg}^{(k)}$ over the number of g-masks, we get

$$\boldsymbol{SM}_{avg} = \frac{1}{p}\sum_{k=1}^{p} SM_{avg}^{(k)}. \tag{6.7}$$

where p is the number of g-masks considered.

The ribosome covers approximately thirty bases of the *mRNA* at a time [76]. Therefore, a sixty base received parity sequence should be sufficient in representing which part of the *mRNA* is exposed to the g-mask of the ribosomal decoder prior to the initiation of translation.

6.4 Simulation Results

The input data required for analysis is obtained in the same way described in the block code model (see section 5.5). The input data in our analysis consist of five different bacterial genomes including *Escherichia coli K-12 MG1655*, *Escherichia coli* O157:H7, *Bacillus Subtilis, Salmonella Typhimurium LT2*, and *Staphylococcus Aureus Mu50.*

Simulation results are shown in Figures (6.4 – 6.13). The investigated model was able to successfully identify and distinguish the group of translated genes (as obtained from GenBank database) from the other groups of hypothetically translated genes (predicted by GeneMark and Glimmer gene prediction programs), and the non-translated genes. In other words, this model is able to filter out the false positives provided by GeneMark and Glimmer gene prediction programs, as well as the group of open reading frames that were selected to follow certain criteria (start with a valid start codon, end with a valid stop codon, and are greater than 99 bases long).

The horizontal axes in figures ($6.4 - 6.13$) represent the position relative to the first base of the initiation codon. The vertical axis shows the mean syndrome value being normalized to the number of analysis $mRNA$ subsequences (M) and also to the number of the decoding masks considered in the analysis (p). The conducted analysis was applied to different (n, k, m) combinations as described in Table 6.3. Only (2,1,2) and (3,1,2) convolutional code model were able to identify the right initiation codons from the false one in the four test groups (described in details in section 5.5). This can be explained by the fact that for both of (2,1,2) and (3,1,2) cases, the length of the obtained g-masks is equal to six bases which is equal to the so called anti-Shine-Dalgarno sequence ($UCCUCC$) located at the 3' end of the $16S\ rRNA$ in the ribosome. This sequence is only available in prokaryotes and generally located eight basepairs upstream of the start codon. All the sequence groups in the (2,1,2) and (3,1,2) models achieve a global minimum syndrome value at the zero position which corresponds to the first base in the initiation codon.

It can also be noticed from figures that the group of hypothetical gene locations obtained by $GeneMark$ program (marked by circles) also show a local minimum at the zero position. The amplitude of this minimum is still less than the global minimum obtained for the translated group (marked by triangles).

The obtained results verify the validity and the significance of the investigated convolutional code model and its biological relevance by being able to identify the right start codons from the false ones. Most importantly, the results also show that the convolutional code model investigated here can be used to interpret some of the aspects related to the process of translation in prokaryotic organisms.

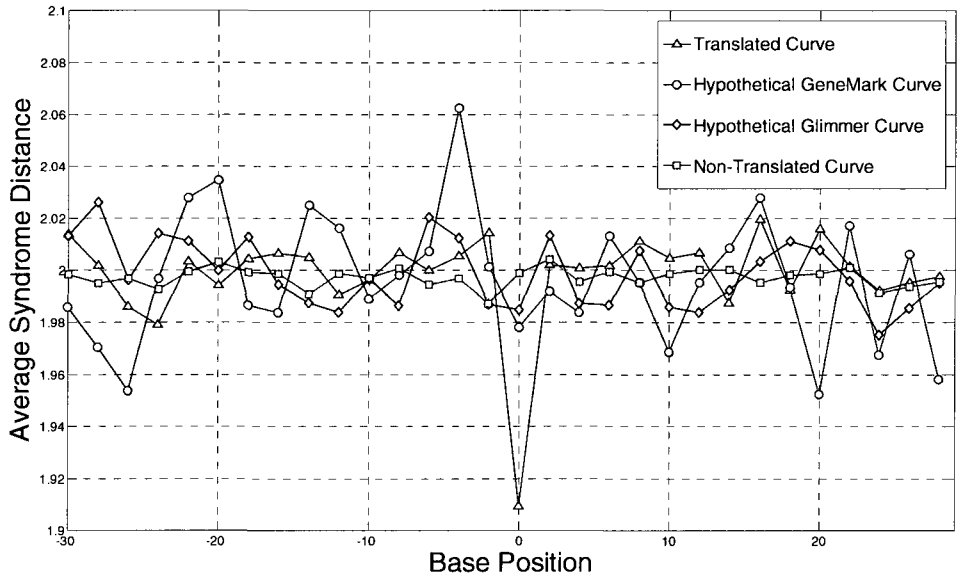


Figure 6.4. Average syndrome distance for Escherichia Coli K-12 MG1655 strain fo a (2,1,2) convolutionl code model

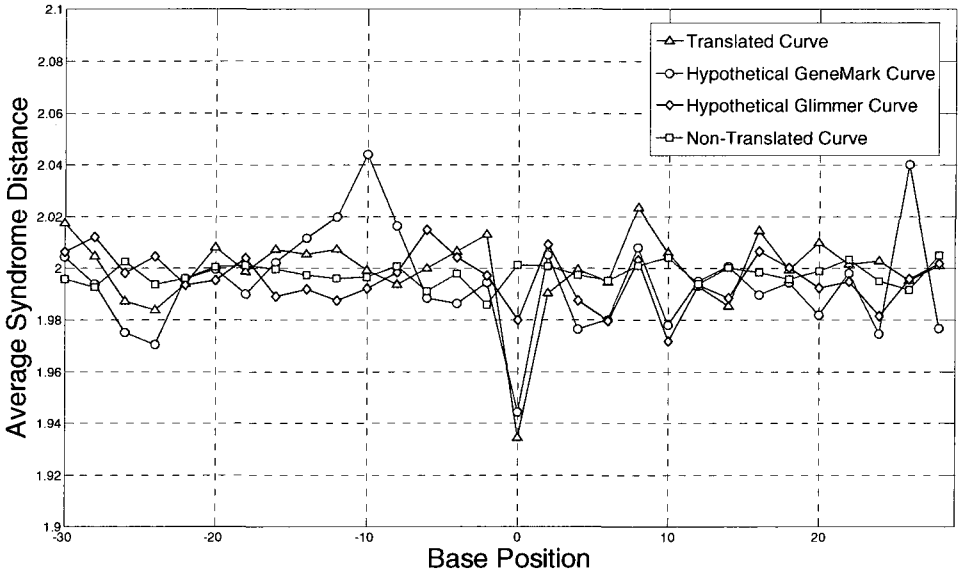


Figure 6.5. Average syndrome distance for Escherichia Coli O157:H7 strain for a (2,1,2) convolutionl code model

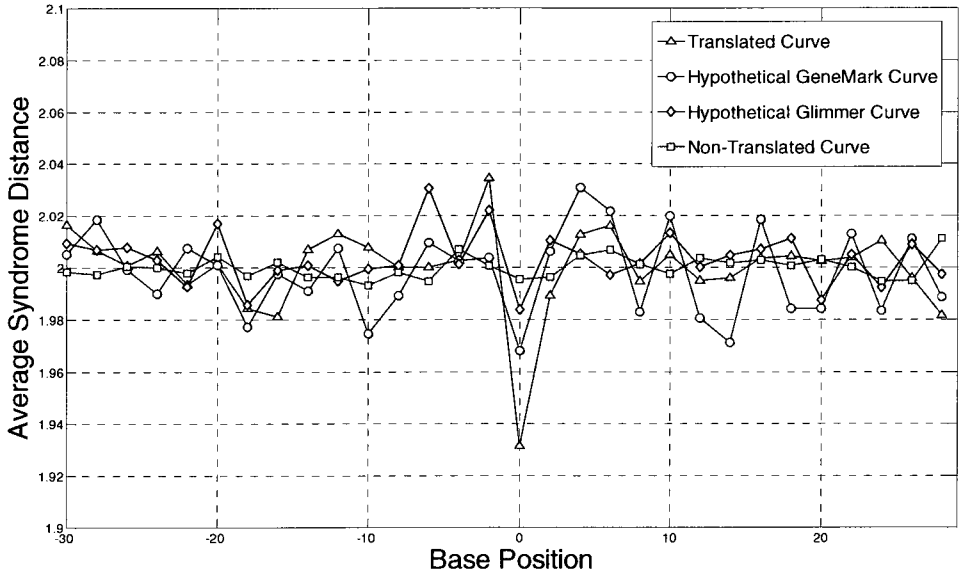


Figure 6.6. Average syndrome distance for Bacillus Subtilis for a (2,1,2) convolutionl code model

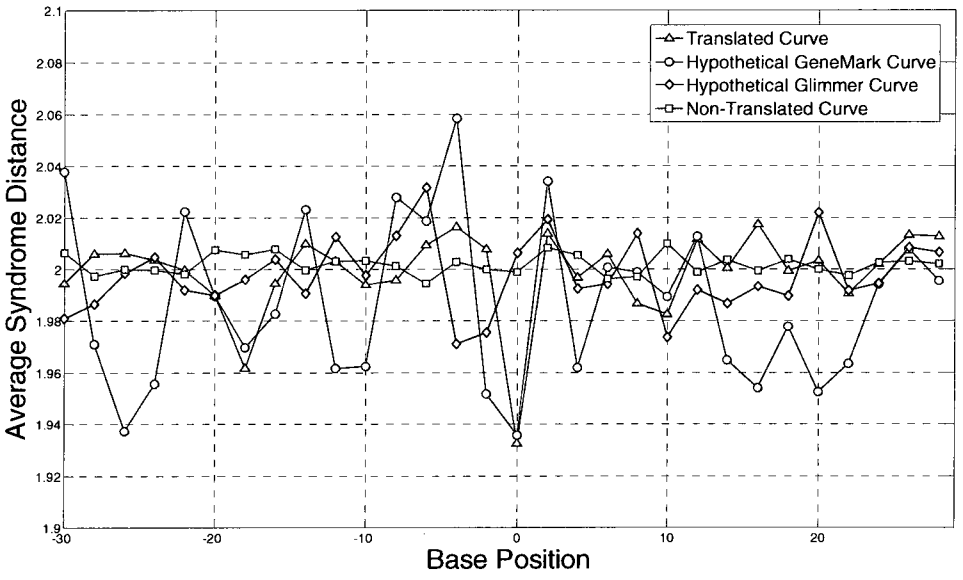


Figure 6.7. Average syndrome distance for Staphylococcus Aureus Mu50 for a (2,1,2) convolutionl code model

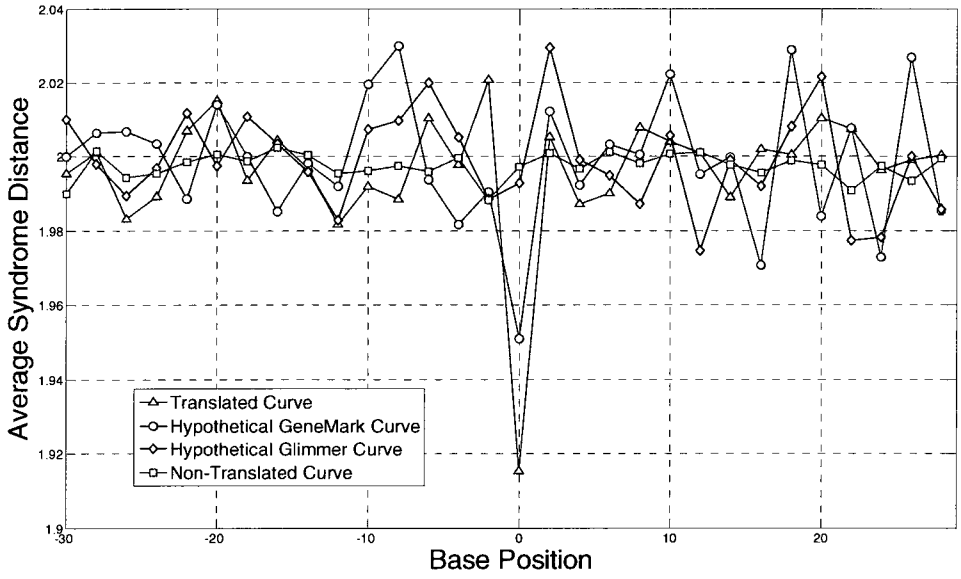


Figure 6.8. Average syndrome distance for Salmonella Typhimurium LT2 for a (2,1,2) convolutionl code model

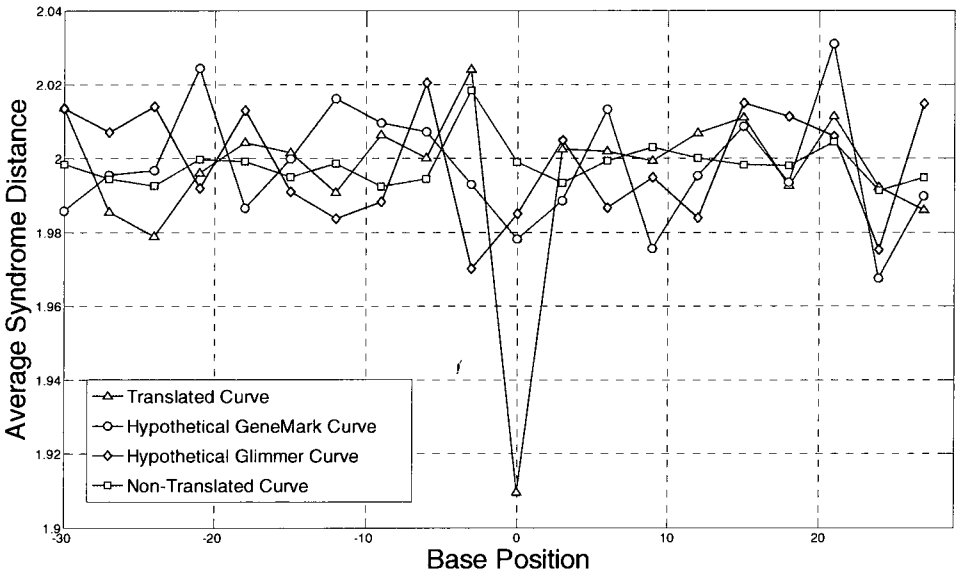


Figure 6.9. Average syndrome distance for Escherichia Coli K-12 MG1655 strain for a (3,1,2) convolutionl code model

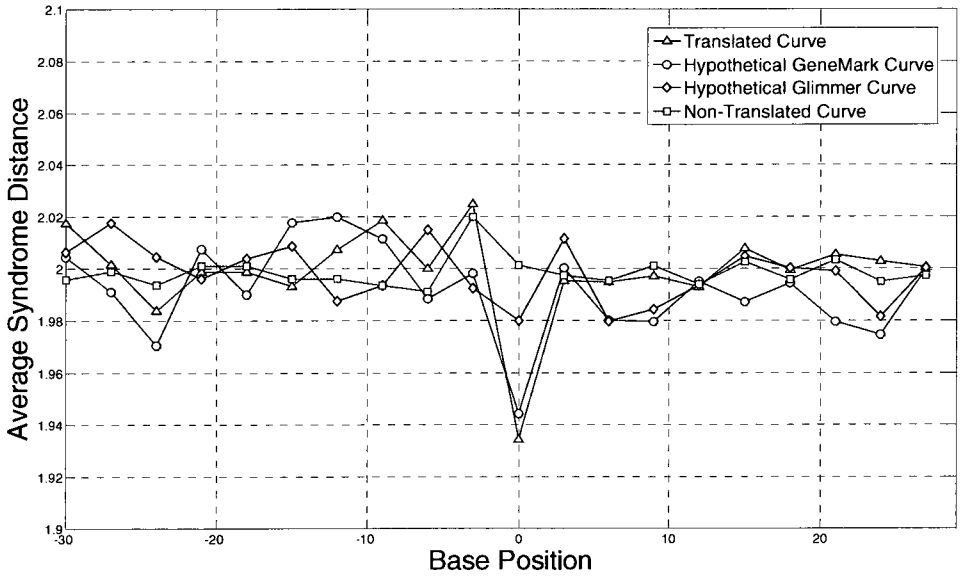


Figure 6.10. Average syndrome distance for Escherichia Coli O157:H7 strain for a (3,1,2) convolutionl code model

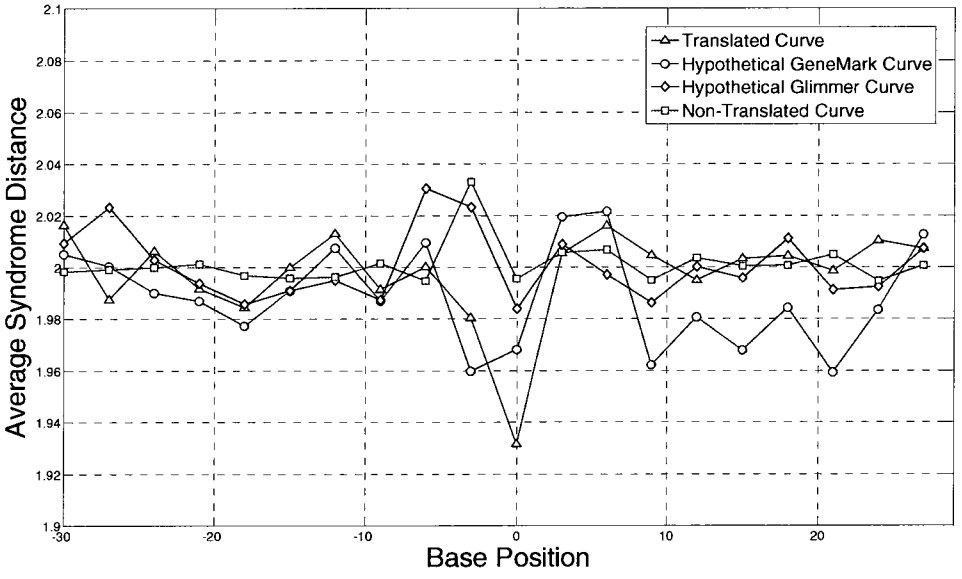


Figure 6.11. Average syndrome distance for Bacillus Subtilis for a (3,1,2) convolutionl code model

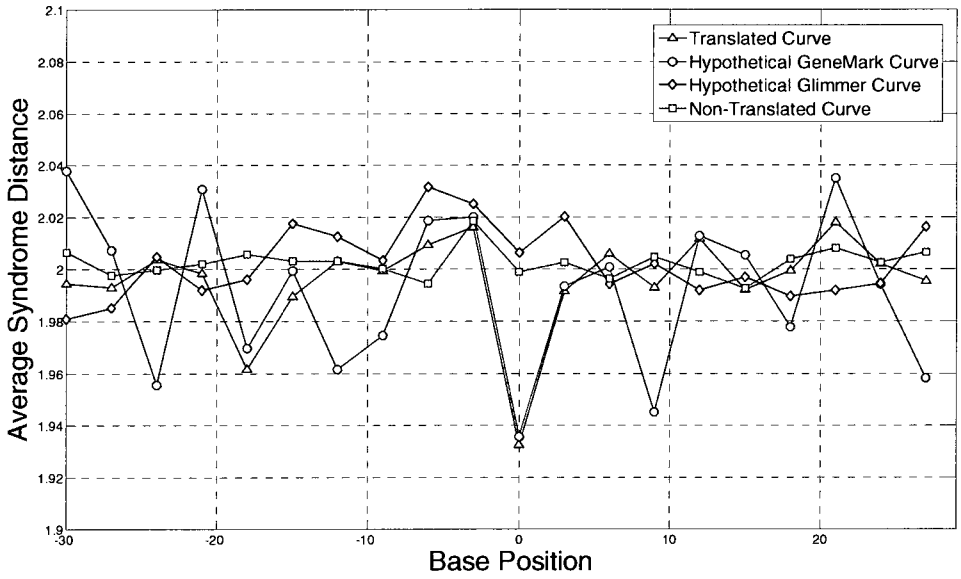


Figure 6.12. Average syndrome distance for Staphylococcus Aureus Mu50 for a (3,1,2) convolutionl code model

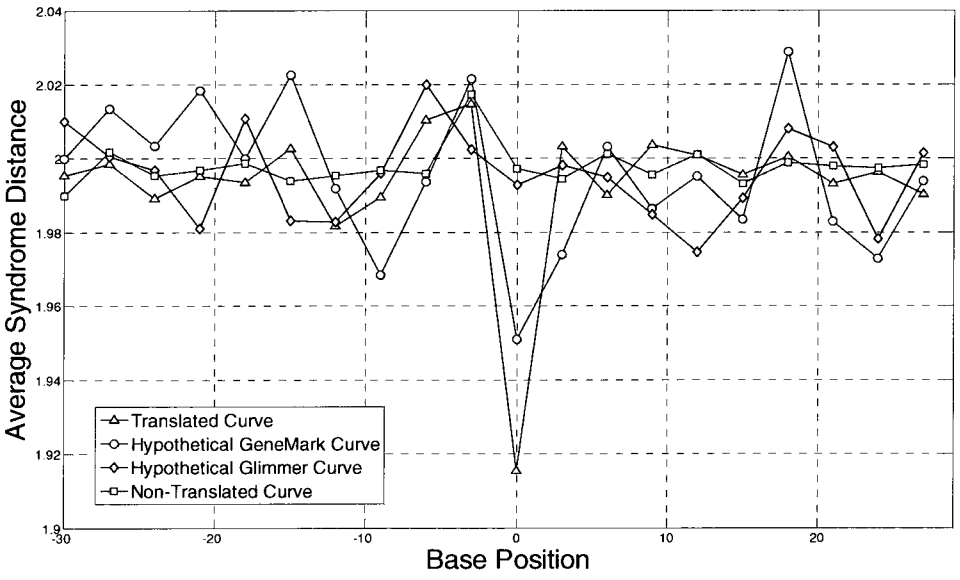


Figure 6.13. Average syndrome distance for Salmonella Typhimurium LT2 for a (3,1,2) convolutionl code model

6.5 Comparison of the Previous Models

In this thesis, all the models that are developed for the process of translation in prokaryotic organisms are generally based on the same key biological elements. These biological elements include: 1) The last 13 bases of the 3' end of the 16S rRNA molecule, 2) The common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), 3) Base-pairing principles, and 4) The free energies involved in the mRNA-rRNA interaction (not used in the convolutional code model). However, even the developed models are almost based on the same key biological elements; they differ in the following aspects:

1. **Model Construction**: The codebook model is based on the assumption that the ribosome uses the last 13 bases sequence as an embedded codebook to identify the initiation signals. In the block code model, the ribosome is modeled as a minimum distance decoder where the mRNA sequence is modeled as a noisy encoded signal, and the last 13 bases sequence as a template to generate a set of valid codewords. However, the main differences between the codebook model and the block code model are in the strategy used to build the codebook and the strategy of decoding the received mRNA sequence. On the other hand, in the convolutional code model, the ribosome is modeled as a table-based convolutional decoder and the mRNA as a noisy convolutional encoded signal. The 16S rRNA sequence is used to form decoding masks for table-based decoding.
2. **Bases Mapping**: While no mapping was used in the codebook model, Galois field of five, GF(5), was used in both of the block code model and the convolutional

code model. It is only in the communications theory based modeling in section 4.3 where the four based were sometimes mapped to their equivalent quaternary representation ($A = 0$, $C = 1$, $G = 2$, and $U = 3$) or to their equivalent binary representation ($A = 00$, $C = 01$, $G = 10$, $and\ T = 11$). In GF(5) the bases were mapped as $Inosine\ (I) = 0$, $Adenine\ (A) = 1$, $Guanine\ (G) = 2$, $Cytosine\ (C) = 3$, and $Uracil\ (U) = 4$.

3. **Model Significance**: The significance of the codebook model lies in its ability to identify the translational signals including the Shine-Dalgarno signal, the initiation signal and the termination signal. Moreover, it allows testing several types of ribosomal mutation and studying their effects of the level of gene expression. In both of the block code model and the convolutional code model, the significance lies in their ability to distinguish four different sequence groups of ORFs (Open Reading Frames) and hence decide whether a given start codon is a valid initiation site or not. In the block code model, the main distinction between the four sequence groups happens in the Shine-Dalgarno domain in the region between -20 to -10 upstream of the initiation codon. In the convolutional code model, the main distinction happens in the region representing the close neighborhood of the initiation codon based on the value of the syndrome in this particular region.

CHAPTER 7

REGULATORY SEQUENCES DISTRIBUTION IN CODING AND NON-CODING REGIONS OF E.COLI BACTERIA

Regulatory sequence detection is a fundamental challenge in computational biology. The transcription process in protein synthesis starts with the binding of the transcription factor to its binding site. Different sites can bind to the same factor. This variability in binding sequences increases the difficulty of their detection using computational algorithms. This chapter proposes two novel methods for detecting transcription factor binding sites in the entire genomic structure by using 1) a frequency weight matrix (FWM), and 2) a distance metric based on a center of mass concept. While the FWM method does not use any mapping for the nucleobases, the second method uses polyphase mapping. The purpose of proposing these two methods is not meant to investigate the $TFBS$ detection problem, rather, it is meant to examine the distribution of the regulatory sequences in coding and non-coding regions in an attempt to utilize that in gene identification. Additionally, using the said methods allow discovering new potential motif sequences that need to be subjected to biological experiments.

7.1 Introduction

7.1.1 Regulatory Sequences. A regulatory sequence (also called a regulatory region or a regulatory area or a regulatory element) is a segment of DNA where regulatory proteins such as transcription factor bind preferentially. These regulatory proteins bind to short stretches of DNA called regulatory regions, which are appropriately positioned in the

genome, usually a short distance *upstream* of the gene being regulated. By doing so, these regulatory proteins can recruit another protein complex, called the *polymerase*. In this way, they control gene expression and thus protein expression.

A transcription factor (TF) is a protein that binds to specific DNA sequences and thereby controls the transcription of genetic information from DNA to $mRNA$ [68, 74]. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes [105, 123]. Transcription factors are essential for the regulation of gene expression and are, as a consequence, found in all living organisms. The number of transcription factors found within an organism increases with genome size, and larger genomes tend to have more transcription factors per gene [173]. Due to the large amount of data stored in the DNA strands, $TFBS$ detection is an arduous work. This chapter proposes two methods to detect $TFBSs$ in an attempt to investigate their distribution in coding and non coding regions.

7.1.2 Conserved Sequences and Consensus sequences. A conserved sequence is a regulatory sequence in a DNA molecule that has remained unchanged throughout the evolution. These conserved sequences have been classified in different families depending on the gene that they regulate. All of the conserved sequences that belong to a family can be represented by a sequence called consensus sequence. As the conserved sequences that form a family are different, this representative sequence can have more than one possible base in some positions.

Table 7.1. E.Coli Transcription Factors

Family #	*TF*	*Family #*	*TF*	*Family #*	*TF*	*Family #*	*TF*
1	*AcrR*	32	*ExuR*	63	*MalI*	94	*PspF*
2	*Ada*	33	*FNR*	64	*MalT*	95	*PurR*
3	*AgaR*	34	*FabR*	65	*MarA*	96	*QseB*
4	*AlaS*	35	*FadR*	66	*MarR*	97	*RbsR*
5	*AllR*	36	*FhlA*	67	*MelR*	98	*RcsAB*
6	*AraC*	37	*Fis*	68	*MetJ*	99	*RhaR*
7	*ArcA*	38	*FlhDC*	69	*MetR*	100	*RhaS*
8	*ArcR*	39	*FruR*	70	*MhpR*	101	*Rob*
9	*ArgR*	40	*Fur*	71	*MngR*	102	*RstA*
10	*AtoC*	41	*GadE*	72	*ModE*	103	*RutR*
11	*BaeR*	42	*GalR*	73	*MprA*	104	*SdiA*
12	*BetI*	43	*GcvA*	74	*MtlR*	105	*SgrR*
13	*BirA*	44	*GlcC*	75	*Nac*	106	*SlyA*
14	*CRP*	45	*GlpR*	76	*NagC*	107	*SoxR*
15	*CaiF*	46	*GntR*	77	*NanR*	108	*SoxS*
16	*Cbl*	47	*H-NS*	78	*NarL*	109	*TdcA*
17	*ChbR*	48	*HU*	79	*NarP*	110	*TdcR*
18	*CpxR*	49	*HcaR*	80	*NhaR*	111	*TorR*
19	*CsgD*	50	*HipB*	81	*NikR*	112	*TreR*
20	*CspA*	51	*HyfR*	82	*NorR*	113	*TrpR*
21	*CueR*	52	*IHF*	83	*NrdR*	114	*TyrR*
22	*CusR*	53	*IclR*	84	*NsrR*	115	*UhpA*
23	*CynR*	54	*IdnR*	85	*NtrC*	116	*UidR*
24	*CysB*	55	*IscR*	86	*OmpR*	117	*UlaR*
25	*CytR*	56	*KdgR*	87	*OxyR*	118	*UxuR*
26	*DcuR*	57	*KdpE*	88	*PaaX*	119	*XapR*
27	*DeoR*	58	*LacI*	89	*PdhR*	120	*XylR*
28	*DgsA*	59	*LexA*	90	*PepA*	121	*YiaJ*
29	*DnaA*	60	*LldR*	91	*PhoB*	122	*ZntR*
30	*EnvY*	61	*LrhA*	92	*PhoP*	123	*ZraR*
31	*EvgA*	62	*Lrp*	93	*PrpR*	124	*Zur*

In order to be able to handle the conserved sequences of all transcription factors, TF, a database (a cell array in MATLAB) is built where each TF is assigned a number. Since *Escherichia coli* ($E.coli$) is a well-studied organism with several highly accurately annotated genome sequences, it is used here as a test case. As such, a database of $E.coli\ TF\ DNA$ binding sequences is produced. Table 7.1 shows 124 different $E.coli$

transcription factors obtained from regulon database [132], and ecogene database [134]. Sequences were collated, redundancies eliminated, and validated by identifying them in the $E.coli$ genome in $ecogene$.

7.1.3 Sequence Logo. A sequence logo is a very descriptive graphical representation of the sequence conservation of nucleotides (in a strand of DNA/RNA) or amino acids (in protein sequences) [142, 148]. To create sequence logos, related DNA, RNA or protein sequences, or DNA sequences that have common conserved binding sites, are aligned so that the most conserved parts create good alignments. A sequence logo can then be created from the conserved multiple sequence alignment. The sequence logo will show how well residues are conserved at each position: the fewer the number of residues, the higher the letters will be, because the better the conservation is at that position. Different residues at the same position will be scaled according to their frequency. Sequence logos can be used to represent conserved DNA binding sites, where transcription factors bind. Figure 7.1 shows some aligned sequences and their sequence logo.

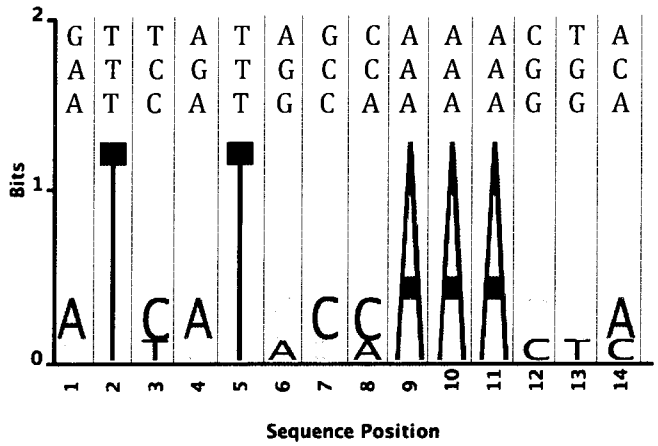


Figure 7.1. Sequences and their sequence logo

The information content (y-axis) of position i is given by

$$R_i = \log_2(s) - (H_i + e_n), \tag{7.1}$$

Where s is 4 for nucleotides and 20 $for\ amino\ acid$. The entropy (H_i) of position i is given by

$$H_i = -\left(\sum f(b,i) * \log_2 f(b,i)\right). \tag{7.2}$$

Here, $f(a,i)$ is the frequency of base b at position i, and e_n is the small-sample correction for an alignment of n letters. The height of letter b in column i is given by $f(b,i) * R_i$. The approximation for the small-sample correction, e_n, is given by:

$$e_n = \frac{s-1}{2 * \ln(2) * n}, \tag{7.3}$$

where s is again 4 for nucleotides and 20 $for\ amino\ acid$. n is the number of sequences in the alignment.

Each family of transcription factor binding sites ($TFBSs$) consists of n conserved sequences where each binding site consists of m nucleobases. Each family of transcription factor binding sites can be represented as a $(n \times m)$ matrix, $\boldsymbol{X}_{n\times m}$. Each row corresponds to a binding site of a given transcription factor. This matrix is given by

$$\boldsymbol{X}_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}, \tag{7.4}$$

where $x_{ij} \in \{A, G, C, T\}; i = 1, 2, \ldots, n; j = 1, 2, \ldots, m.$

In a set of sequences, the j^{th} sequence can be represented by a matrix $S(b,l,j)$ that contains only 1's and 0's. In other words, the j^{th} row in the nucleobases matrix $\boldsymbol{X}_{n\times m}$ will be represented by the matrix $S(b,l,j)$. For example, if the 5^{th} row is given by the sequence $5' - CAGGTCTGCA - 3'$, then the $S(b,l,j)$ matrix is given by

b	C	A	G	G	T	C	T	G	C	A
l	1	2	3	4	5	6	7	8	9	10
A	0	1	0	0	0	0	0	0	0	1
C	1	0	0	0	0	1	0	0	1	0
G	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	1	0	1	0	0	0

	A
	G
	C
	T

Figure 7.2. $S(b, l, 5)$ matrix of the fifth row sequence in $\boldsymbol{X}_{n\times m}$

The position frequency matrix $\boldsymbol{R}_{4\times m}$ can be obtained by

$$\boldsymbol{R}_{4\times m} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{S}(b, l, j)\,, \tag{7.5}$$

where $b \in \{A, G, C, T\}$, l is the nucleobase position, and j corresponds to the j^{th} conserved sequence in $\mathrm{X_{n\times m}}$. The position frequency matrix $\boldsymbol{R}_{4\times m}$ can be written as

$$\boldsymbol{R}_{4\times m} = \begin{bmatrix} r_{A1} & r_{A2} & r_{A3} & \cdots & r_{Am} \\ r_{G1} & r_{G2} & r_{G3} & \cdots & r_{Gm} \\ r_{C1} & r_{C2} & r_{C3} & \cdots & r_{Cm} \\ r_{T1} & r_{T2} & r_{T3} & \cdots & r_{Tm} \end{bmatrix}. \tag{7.6}$$

The entries of $\boldsymbol{R}_{4\times m}$ can be alternatively calculated as

$$r_{bi} = \frac{N_{bi}}{n}, i = 1, 2, 3, \ldots, m, \tag{7.7}$$

where N_{bi} stands for the number of times the base $b \in \{A, G, C, T\}$ occurs in column i of the matrix $\boldsymbol{X}_{n\times m}$. Therefore, r_{bi} is the probability of the base b occurring at position i.

Based on the position frequency matrix $\mathrm{X_{n\times m}}$, an intuitive graphical representation of each family of *TFBSs* (motif sequences) can be obtained. Figure 7.3 is obtained for *XylR* (family 120 in Table 7.1). The x-axis represents the base position and

the y-axis represents the frequency of occurrence of each base being normalized to the number of sequences in each family.

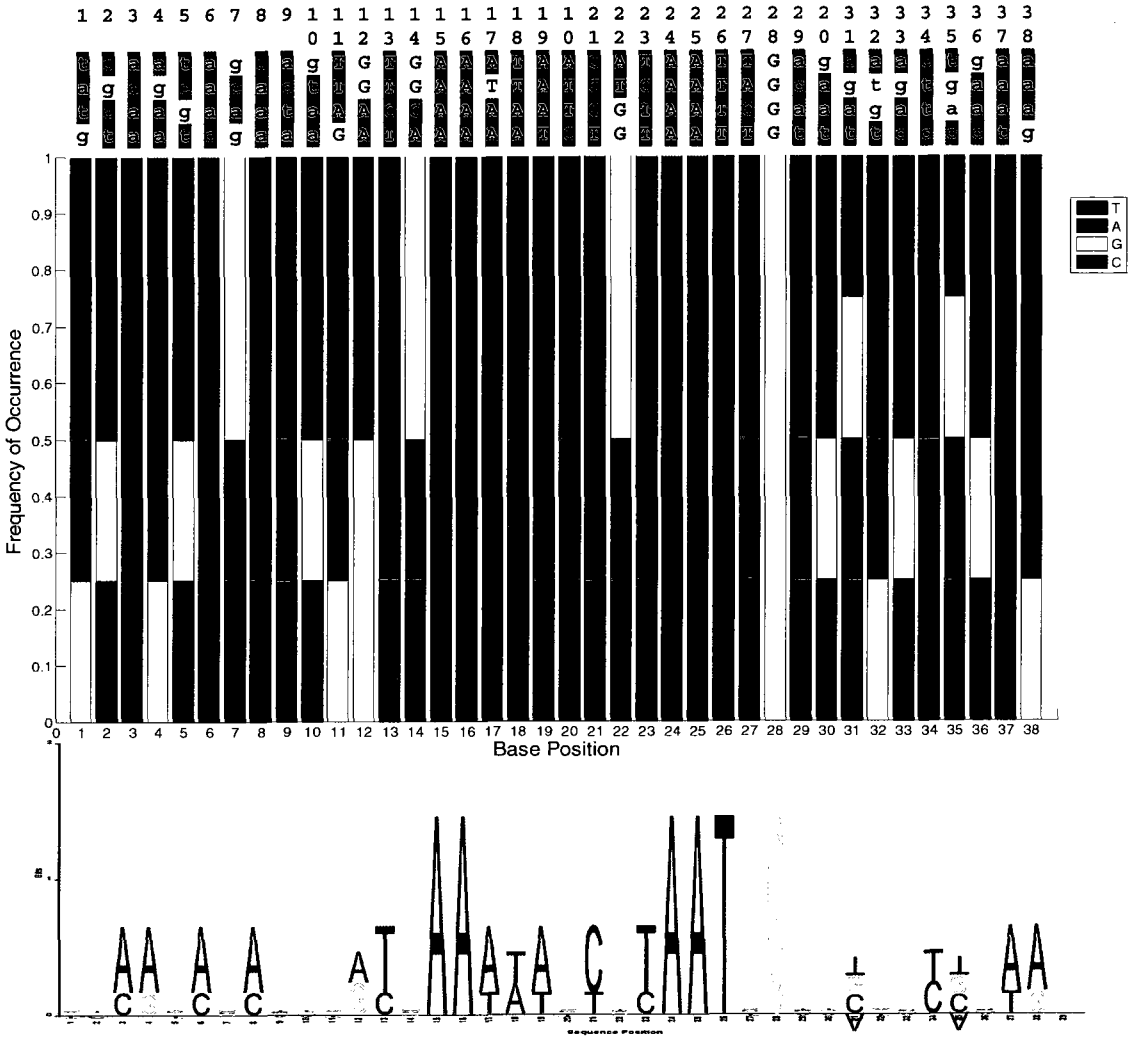


Figure 7.3. A graphical representation of *XylR* compared to the sequence logo

The four bases are represented with different colors and the length of the bars indicates the probability of each base at each position. It can be observed that the higher the probability, the longer are their corresponding bars in the plot. The bars at each position are being sorted in a descending order with longer bars located at the top of the plot. As we are dealing with probabilities we can notice that the sum of the length of the bars at each position is one.

$$\sum_{i=1}^{4} r_{bi} = r_{Ai} + r_{Gi} + r_{Ci} + r_{Ti} = 1, \tag{7.8}$$

As can be observed from Figure 7.3, if all the bases at a given base position are all the same, say "A", then the graphical representation will show a single bar (colored with green) of an amplitude equal to one. This means that having "A" at that particular position is a certain event (probability = 1), and having all other bases ("G" *or* "C" or "T") is an impossible event (Probability = 0). The bars at the top of the graph will correspond to the consensus sequence of the family under consideration.

The goal of the methods developed in this chapter is to be able to efficiently detect the binding sites of all transcription factors given in Table 7.1, and also detect the possible conserved sequences not yet discovered that could be present in the genome. Several researches have stated that $TFBSs$ appear in the promoters of the gene [84]. Thus, most of the possible motif sequences found by this method should be in non-coding regions.

7.2 TFBS Detection Algorithms

In this chapter, two novel methods for $TFBS$ detection have been developed. The first method utilizes the position frequency matrix described before to design distance metrics between the set of conserved sequences and the DNA under study. This method does not use any type of mapping for the nucleobases and hence does not distort the information provided by the set of conserved sequences that belong to the same transcription factor. The second method utilizes a center of mass concept, Euclidean distance, and polyphase mapping to achieve the same goal.

7.2.1 *TFBS* Detection algorithm using the Frequency Weight Matrix. This algorithm is illustrated by a schematic representation as shown in figures 7.4 and 7.5. The inputs to the algorithm are the genome under study $\boldsymbol{G}_{1\times L}$ (L is the length of the genome) and the set of binding site sequences belonging to a particular transcription factor $\boldsymbol{X}_{n\times m}$. The distance metric vector $\boldsymbol{A}_{1\times n}$ is assigned to the set of conserved binding sites belonging to the same transcription factor with a length equal to the number of binding sites (n), while $\boldsymbol{B}_{1\times(L-m+1)}$ is another distance metric vector assigned to set of sequences in the genome with the same length as the conserved sequence (m). The output of the algorithm is distance vector $\boldsymbol{E}_{1\times r}$ that corresponds to the locations of actual binding sites given in table 7.1.

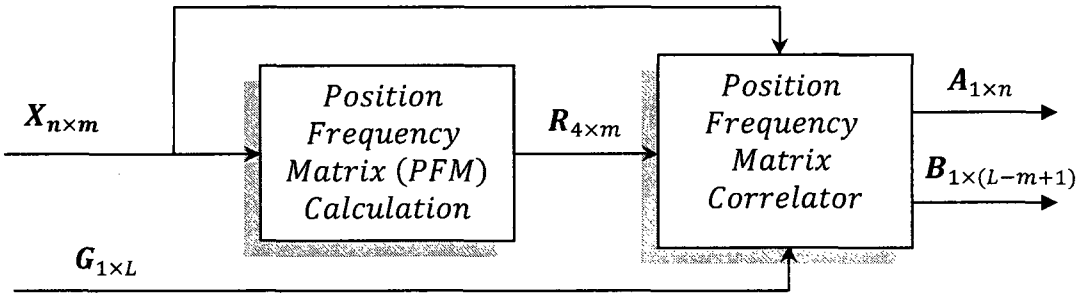


Figure 7.4. A block diagram of *TFBS* detection algorithm (part 1)

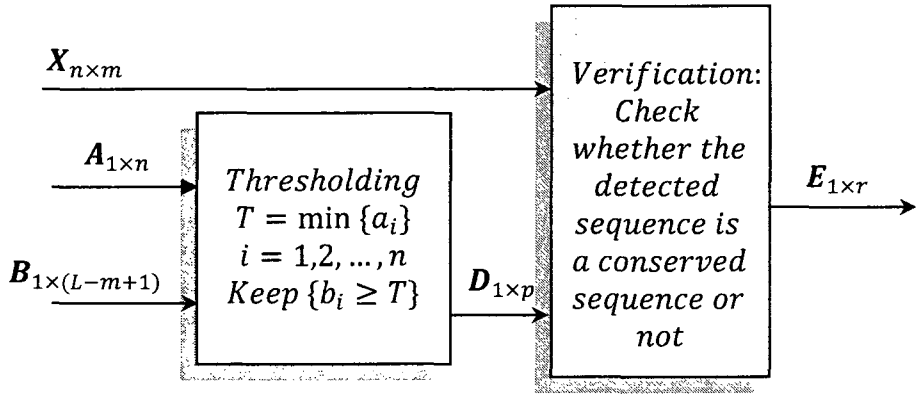


Figure 7.5. A block diagram of *TFBS* detection algorithm (part 2)

According to these schematic diagrams, the algorithm is able to detect $TFBSs$ in the genome by the following steps:

Step 1: Position frequency matrix calculation

The Position frequency matrix (PFM) (denoted as $\boldsymbol{R}_{4\times m}$) can be calculated using equation 7.5 or alternatively using equations 7.6 and 7.7 as described in section 7.1.3.

Step 2: Distance metric vectors calculation

After calculating the position frequency matrix (PFM), the next step is to assign each conserved sequence (i.e. each row in $\boldsymbol{X}_{n\times m}$) a distance metric obtained by first aligning that particular row with the position frequency matrix and then adding up the values in the matrix that correspond to each nucleobase in that row. If any of these values is zero (this means that this particular base does not happen at that position), then we add up the values occurring before zero and stop and then move to the next row. Based on this, we get a distance vector ($\boldsymbol{A}_{1\times n}$) which can be written as

$$\boldsymbol{A}_{1\times n} = [a_1, a_2, a_3, \dots, a_n], \tag{7.9}$$

where a_i is a distance metric associated with the i^{th} row in $\boldsymbol{X}_{n\times m}$ and is obtained by

$$a_i = \sum_{j=1}^{m' \leq m} r_{x_{ij}j}, i = 1, 2, 3, \dots, n. \tag{7.10}$$

$x_{ij} \in \{A, G, C, T\}$ and m' is the index of the first base to map to a zero in the Position Frequency matrix. Given that the genome under study can be written as

$$\boldsymbol{G}_{1\times L} = [g_1, g_2, g_3, \dots, g_L], \tag{7.11}$$

where L is the length of the genome in nucleobases, $g_i \in \{A, G, C, T\}$, $i = 1, 2, 3, \dots, L$. The next step is to compare the whole genome sequence to the set of conserved sequences represented by $\boldsymbol{X}_{n\times m}$. To achieve this, a sliding window of length equal to m

(conserved sequence length) is translated all over the genome ($\boldsymbol{G}_{1\times L}$) one base at a time. This will divide the genome sequence into $(L - m + 1)$ subsequences each of length m. These subsequences are then assigned distance metrics following the same procedure used to get the distance vector $\boldsymbol{A}_{1\times n}$. Hence, this will yield another distance vector, $\boldsymbol{B}_{1\times(L-m+1)}$, which can be written as

$$\boldsymbol{B}_{1\times(L-m+1)} = [b_1, b_2, b_3, \dots, b_{L-m+1}], \tag{7.12}$$

where b_i is a distance metric associated with the sequence of length m obtained at the i^{th} nucleobase position of the genome and is obtained by

$$b_i = \sum_{j=1}^{m' \le m} r_{g_i j}, \qquad i = 1, 2, 3, \dots, L - m + 1. \tag{7.13}$$

where $g_i \in \{A, G, C, T\}$ is a base in the genome being considered. $r_{g_i j}$ is the j^{th} element in $\boldsymbol{R}_{4\times m}$ that corresponds to the i^{th} base in the genome (g_i).

Step 3: Thresholding of the distance metric vector $\mathbf{B}_{1\times(L-m+1)}$

At this point, the weighting vector $\mathrm{B}_{1\times(\mathrm{L-m+1})}$ contains all the possible weights corresponding to all possible conserved sequences in the genome sequence. The higher the weight value of a given sequence, the higher is the probability of that particular sequence to be a real conserved sequence. To eliminate those subsequences in the genome with low weights, a threshold T is calculated as the minimum weight in the distance vector $\boldsymbol{A}_{1\times n}$ vector obtained in the previous steps. All the weights less than T are discarded. The resulting weighting vector after Thresholding $(\boldsymbol{C}_{1\times k})$ is defined as

$$\boldsymbol{C}_{1\times k} = \{c_i = b_j; b_j \ge T, i = 1, 2, 3, \dots, k, j = 1, 2, 3, \dots, L - m + 1\}, \tag{7.14}$$

where

$$T = min\{a_i\}, i = 1, 2, 3, \dots, n. \tag{7.15}$$

Now, the vector $\boldsymbol{C}_{1\times k}$ contains all the possible weights greater than or equal to the threshold T. In other words, the vector $\boldsymbol{C}_{1\times k}$ gives the locations of the subsequences in the genome that are of exact or of high similarity to the given set of binding sites. As we want to only keep the locations corresponding to exact conserved sequences, all the values in $\boldsymbol{C}_{1\times k}$ that are different from the set of weights in the $\boldsymbol{A}_{1\times n}$ vector are discarded. The resulting weighting vector can be written as

$$\boldsymbol{D}_{1\times p} = \{d_i = c_j; c_j = a_j, i = 1, 2, 3, \dots, p; j = 1, 2, 3, \dots, k\}, \qquad (7.16)$$

where $p \leq k$.

Step 4: *TFBS* identification

Although at this point of the algorithm we have drastically decreased the number of possible conserved sequences, we still have too many possible sequences belonging to the genome that are not real conserved sequences (false positives). The vector $\boldsymbol{D}_{1\times p}$ contains all the possible weights that are exactly the same as the weights of the original conserved sequences. Some of these weights are false positive, i.e. they do not correspond to an actual conserved sequence. To filter these false positives out, each one of the detected sequences is compared to the original conserved sequences and only the verified ones are kept. The resulting distance vector $\overline{E}_{1\times r}$ is evaluated as

$$\boldsymbol{E}_{1\times r} = \{e_i = d_j; \boldsymbol{S}(d_j) = \boldsymbol{X}(d_j), i = 1, 2, 3, \dots, r; j = 1, 2, 3, \dots, p\}, \qquad (7.17)$$

where $\boldsymbol{S}(d_j)$ is the sequence of length m in the genome corresponding to the j^{th} weight in the $\boldsymbol{D}_{1\times p}$ vector. $\boldsymbol{X}(d_j)$ is the conserved sequence in $\boldsymbol{X}_{n\times m}$ having the same weight d_j since if these two sequences are the same, they have to have the same weight value.

Illustrative Example. In order to test the performance of the *TFBS* detection algorithm described before, it was applied to all of the *E.coli* transcription factors provided in Table 7.1. The following example gives a clear picture of how this algorithm works.

Step 1: Position frequency matrix calculation

The *XylR* transcription factor (entry # 120 in Table 7.1) was selected as an input to the algorithm. The binding sites of this transcription factor are shown in Table 7.2. The Position Frequency Weight matrix of this transcription factor is shown in Figure 7.6.

Figure 7.7 shows a graphical representation of the Position Frequency matrix with the elements (probabilities at each base position) being sorted column wise in a descending order. In other words, the top colored bars in Figure 7.7 represent the most probable bases at each position of the *TFBSs*, hence the consensus sequence.

Table 7.2. *XylR* transcription factor binding sites

BS #	*TFBS*
1	TCAATAGCAGTGTGAAATAACATAATTGAGCAACTGAA
2	AGCGCACACTTGTGAATTATCTCAATAGCAGTGTGAAA
3	TCAAGAAATAAACCAAAAATCGTAATCGAAAGATAAAA
4	GTAATCGAAAGATAAAAATC TGTAATTGTTTTCCCCTG

l	1	2	3	4	5	6	7	8	. . .	36	37	38
	T	*C*	*A*	*A*	*T*	*A*	*G*	*C*	. . .	*G*	*A*	*A*
	A	*G*	*C*	*G*	*C*	*A*	*C*	*A*	. . .	*A*	*A*	*A*
	T	*C*	*A*	*A*	*G*	*A*	*A*	*A*	. . .	*A*	*A*	*A*
	G	*T*	*A*	*A*	*T*	*C*	*G*	*A*	. . .	*C*	*T*	*G*

A	1/4	0	3/4	3/4	0	3/4	1/4	3/4	. . .	1/2	3/4	3/4
G	1/4	1/4	0	1/4	1/4	0	1/2	0	. . .	1/4	0	1/4
C	0	1/2	1/4	0	1/4	1/4	1/4	1/4	. . .	0	0	0
T	1/2	1/4	0	0	1/2	0	0	0	. . .	0	1/4	0

Figure 7.6. *XylR* Position Frequency Matrix

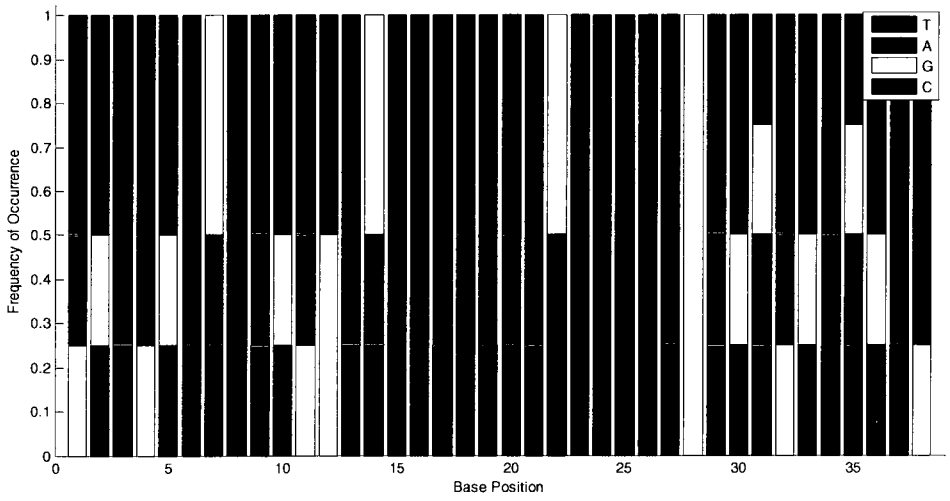


Figure 7.7. Sequence Logo based representation of $XyIR$

Step 2: Distance metric vectors calculation

The distance metric vectors ($\boldsymbol{A}_{1\times n}$) and ($\boldsymbol{B}_{1\times(L-m+1)}$) for $XyIR$ $(n = 4, m = 38, L = 4639221, L - m + 1 = 4639184)$ are given by

$$\boldsymbol{A}_{1\times n} = [22.25, 19.75, 22.00, 19.50], \tag{7.18}$$

$$\boldsymbol{B}_{1\times(L-m+1)} = [0.75, 0.75, 0, 0.75, 0.75, \ldots, 0.25, 0.25, 1.25]. \tag{7.19}$$

Step 5: Thresholding of the distance metric vector $\boldsymbol{B}_{1\times(L-m+1)}$

According to equation 7.15, the threshold being selected as the minimum of the distance metric vector ($\boldsymbol{A}_{1\times n}$) is 19.50. The distance metric vector ($\boldsymbol{B}_{1\times(L-m+1)}$) will then be compared to the threshold such that all the values less than the threshold (19.50) will be discarded (set to zero). In other words, only the values greater than or equal to 19.50 will be kept where their index indicates the base position of the possible conserved sequences. This will yield the distance vector ($\boldsymbol{C}_{1\times k}$). For example, the first subsequence of the genome is assigned the value 0.75 which is lower than 19.50. Therefore, this sequence cannot be a conserved sequence and hence eliminated as an option.

Step 6: *TFBS* identification

At this step, we have a set of all the base positions in the genome under study where a possible conserved sequence might occur (those sequences assigned values greater than or equal to the threshold). These possible sequences still need to be checked to be real *TFBSs*. Based on *XylR* transcription factor, Table 7.3 shows two possible sequences being confirmed as actual *TFBSs* with their distance metric values assigned (b_i) and their occurrence base positions (i).

Table 7.3. *XylR* confirmed *TFBSs*

	Confirmed TFBSs	b_i	i
1	UCAAGAAAUAAACCAAAAAUCGUAAUCGAAAGAUAAAA	22.00	3728618
2	GUAAUCGAAAGAUAAAAAUCUGUAAUUGUUUUCCCCUG	19.50	3728639

Usually, the number of possible *TFBSs* available after thresholding is greater than the number of confirmed ones. This is due to the possibility of having more than one sequence sharing the same distance metric just because of their similarity (not being identical) to the set of conserved sequences belonging to the same transcription factor.

Simulation Results. The previous algorithm for *TFBS* detection was applied to $MG1655$ and $O157{:}H7$ *E. coli* strains (both forward and reverse strands were investigated). All of the transcription factors mentioned in Table 7.1 were considered. Simulation results indicated that most of the *TFBSs* detected occur in the non-coding regions. Figures 7.8-7.11 show the simulation results obtained when the algorithm was applied to $MG1655$ positive strand, $MG1655$ negative strand, $O157{:}H7$ positive strand and $O157{:}H7$ negative strand, respectively. The red color in these figures corresponds to the set of *TFBSs* detected in the non-coding regions, the blue color to the ones detected in the coding regions, and the green color to the ones overlapping between non-coding and

coding regions. The y-axis represents the transcription factor number as referenced in Table 7.1 with some offset values introduced to distinguish the three set of detected $TFBSs$ as located in the non-coding regions (marked in red) or in the coding regions (marked in green) or overlapping in between (marked in blue). The offset values are 124 and 248. In other words, any horizontal line below 124 will pass through all the $TFBSs$ related to the transcription factor indexed by the y-axis value as referenced in Table 7.1. If this horizontal line is between 124 and 248 (i.e. the middle region) a value of 124 should be subtracted from the y-axis value to know what transcription factor is being referred to. Finally, if the horizontal line is above 248 then a value of 248 should be subtracted. In this way, Figures 7.8-7.11 not only classify the detected $TFBSs$ into three different sets but also tell which transcription factor is being referred to at each level.

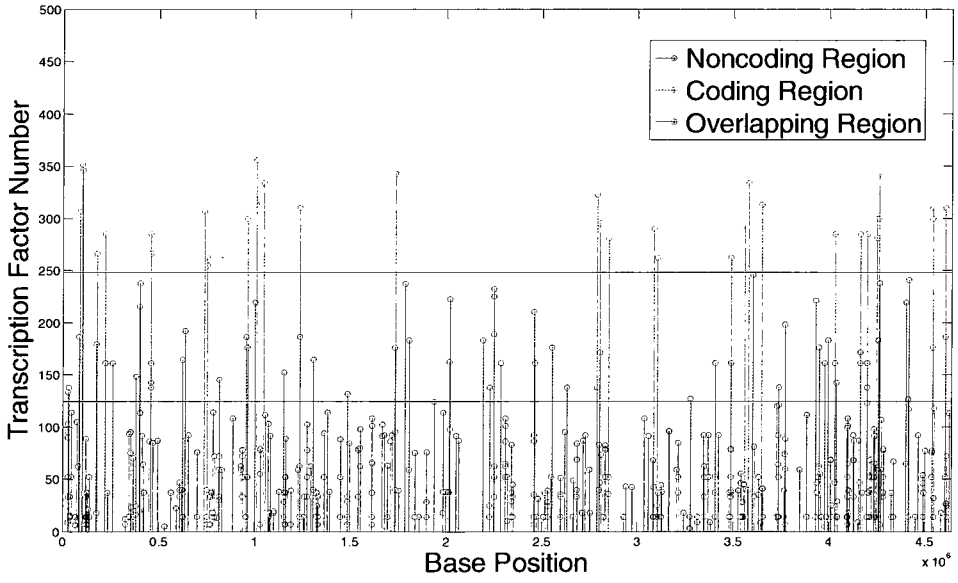


Figure 7.8. $TFBS$ detection using $MG1655$ positive strand

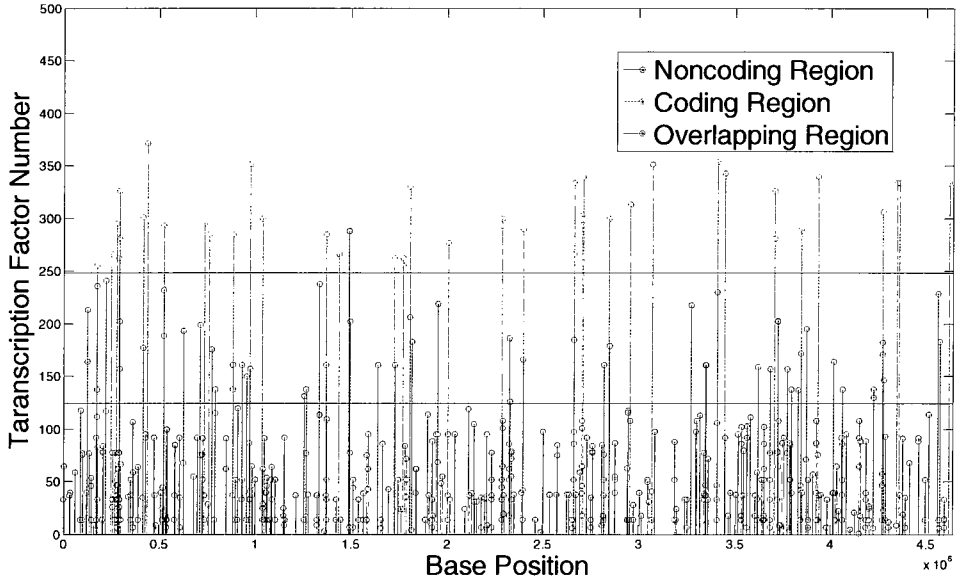


Figure 7.9. *TFBS* detection using $MG1655$ negative strand

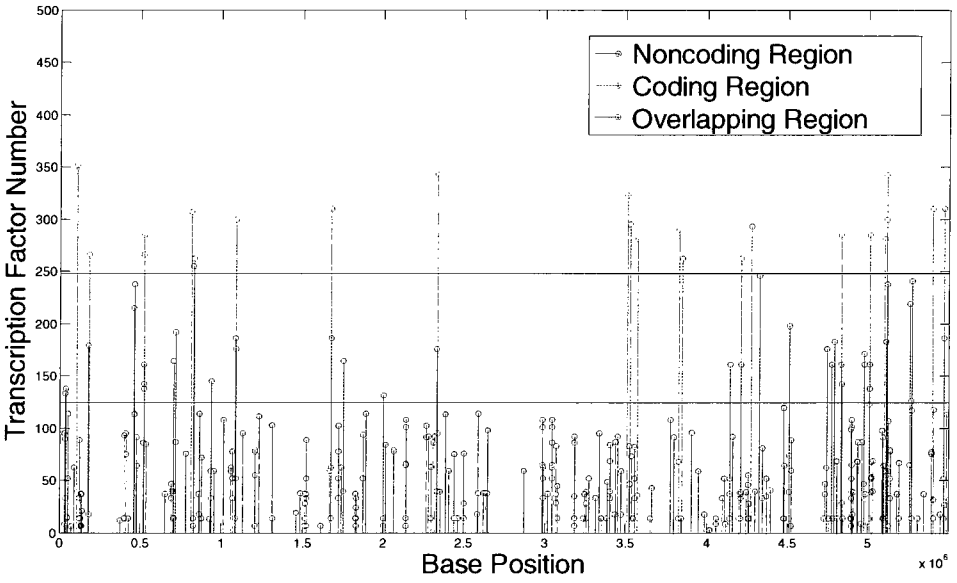


Figure 7.10. *TFBS* detection using $O157{:}H7$ positive strand

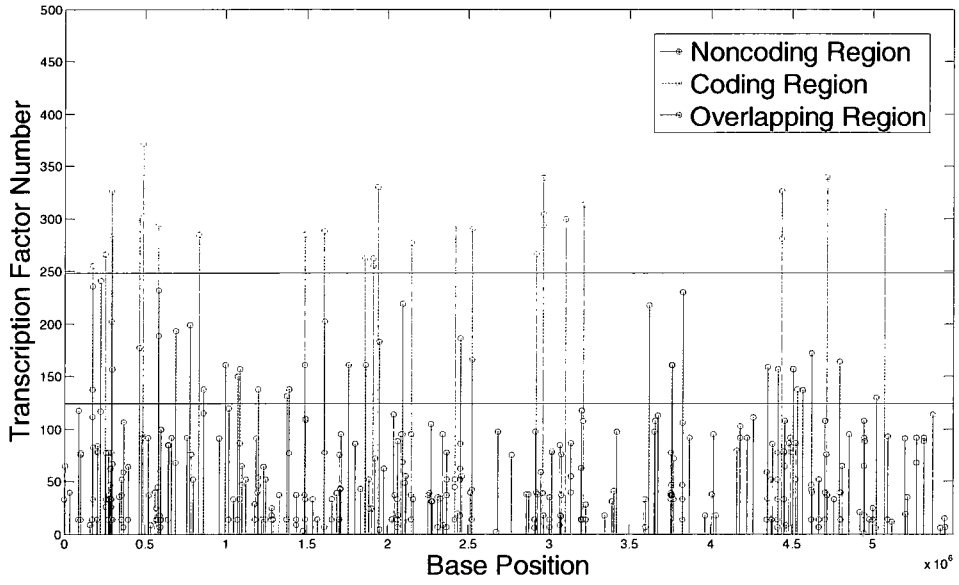


Figure 7.11. *TFBS* detection using $O157{:}H7$ negative strand

According to the simulation results shown previously in Figures 7.8-7.11, Table 7.4 gives some statistical information of the detected *TFBSs* in terms of percentages. As can be observed, most of the detected *TFBSs* are located in the non-coding regions which is biologically relevant and agrees with theory [84] as the transcription factors which bind to the detected sites regulate the transcription of the adjacent genes located ahead.

Table 7.4. *TFBSs* detected in MG1655 and O175:H7 *E. coli* strains

E. coli Strain	*MG1655*		*O157:H7*	
Strand Orientation	*Positive Strand*	*Negative Strand*	*Positive Strand*	*Negative Strand*
Total number of TFBSs	956	1000	642	616
% of *TFBSs in non-coding regions*	85.46	85.90	86.15	85.23
% *of TFBSs in the coding regions*	6.49	6.20	6.85	5.84
% *of TFBSs overlapping in between*	8.05	7.90	7.00	8.93

In the gene expression process, it is very common that a single transcription factor binds not to an exact motif sequence but rather to a consensus motif. This fact is due to

some bases belonging to the motif sequence are not as important as others when a transcription factor binds in the binding site. To consider this in the developed algorithm, the search for possible $TFBSs$ was modified such that if the subsequence being tested is more that 80% (this number can be modified as required) similar to the original corresponding set of $TFBSs$, then it can be considered as a possible $TFBS$. For example, applying the algorithm to $MG16655\,E.coli$ genome (forward strand) to locate subsequences of similitude greater than or equal to 80% of their original corresponding sets of $TFBSs$, yields six other possible sequences. Table 7.5 gives a detailed description of these six sequences with their corresponding location in the genome, the transcription factor they belong to, the number of matching bases and the percentage of similitude of these sequences with their original corresponding transcription factor binding sites.

Table 7.5. Possible motif sequences based on greater than 80% similitude to their original sets of TF conserved sequences

#	Possible *TFBSs*	Base Position	TF #	TF	# of Matches	Similarity Percentage
1	'AGGCCUACGUUAAUUCUGCAAUAUAUUGAAUCU'	5593	52	FNR	29	87.879
2	'AGGCCUACGUUAUUUCAGCAAUAUAUUGAAUUU'	4.3021e+005	52	FNR	28	84.848
3	'AGGCCUACGUGAACUCUGCAAUAUAUUGAAUUU'	8.3153e+005	52	FNR	29	87.879
4	'AGGCCUACAUGAUUUCUGCAAUAUAUUGAAUUU'	3.7386e+006	52	FNR	28	84.848
5	'UGGUAUAACAGGUAUAAAGGUAUACA'	4.5012e+006	77	DcuR	24	92.308
6	'UUAUACCUGUUAUACCAGAUCAAUUA'	3.3713e+006	77	DcuR	24	92.308

Table 7.6 show similar simulation results but with the percentage of similitude being set to 70%. By doing so, thirty two subsequences result. Therefore, this algorithm does not only provide locations of the sets of verified conserved sequences but also locations of all consensus sequences that are partially similar as well. Moreover, with this algorithm, the level of similitude between the target sequences to be located and the original transcription factor binding sites can be controlled accordingly as required.

Table 7.6. Possible motif sequences based on greater than 70% similitude to their original sets of binding sites

#	Possible TFBSs	Base Position	TF #	TF	# of Matches	Similarity Percentage
1	'AGGCCUACGUUAAUUCUGCAAUAUAUUGAAUCU'	5593	52	*IHF*	29	87.879
2	'AGGCCUACGUUAUUUCAGCAAUAUAUUGAAUUU'	4.3021*e* + 005	52	*IHF*	28	84.848
3	'AGGCCUACAUGAUCUCUGCAAUAUAUUGAGUUU'	7.0708*e* + 005	52	*IHF*	26	78.788
4	'AGGCCUACAUUUUCUCCGCAAUAUAUUGAAUUU'	8.1483*e* + 005	52	*IHF*	25	75.758
5	'AGGCCUACGUGAACUCUGCAAUAUAUUGAAUUU'	8.3153*e* + 005	52	*IHF*	29	87.879
6	'AGGCCUACAUGAUUUCUGCAAUAUAUUGAAUUU'	3.7386*e* + 006	52	*IHF*	28	84.848
7	'AGGCCAACGGUAGAAUUGUAAUCUAUUGAAUUU'	4.0476*e* + 006	52	*IHF*	24	72.727
8	'UUAAUUAAAAUGUUACGUGUUUAAUGU'	5.2862*e* + 005	53	*IclR*	19	70.37
9	'UAUUUAAAUUUUUUGUGCUUUUGUUUU'	1.2187*e* + 006	53	*IclR*	19	70.37
10	'ACAUUAAAAAUGAAACUUAUUAAAUUG'	1.5962*e* + 006	53	*IclR*	19	70.37
11	'UUGAUUAAAAAGGUAAAUAUUUAAAAU'	2.9025*e* + 006	53	*IclR*	19	70.37
12	'UAUUUUUUGUUUUUAUUUUUUAAAGGA'	3.4946*e* + 006	53	*IclR*	20	74.074
13	'UUAAUAAAAAUGAUGAAUGAUUUAGAC'	3.5807*e* + 006	53	*IclR*	19	70.37
14	'CUGUACAUAUCAUAGACUAAACGGAUAC'	1.7994*e* + 006	68	*MetJ*	20	71.429
15	'UGGUAUAACAGGUAUAAAGGUAUACA'	4.5012*e* + 006	77	*NanR*	24	92.308
16	'UUAUACCUGUUAUACCAGAUCAAUUA'	3.3713*e* + 006	77	*NanR*	24	92.308
17	'CAUAGAGGUUUAAUCCUUAUUCAGAGU'	2.4967*e* + 006	78	*NarL*	19	70.37
18	'GUCACUAUACAACGGACGGGGGAAGGA'	4.363*e* + 006	78	*NarL*	19	70.37
19	'AAUUGCAUUUAAAAAAUAUGUUCUGUG'	3.8409*e* + 005	78	*NarL*	19	70.37
20	'UGAUAUCACUAUAGAUAUUGAUCAUUA'	1.3899*e* + 006	78	*NarL*	19	70.37
21	'UAUUACAAUGUAAUCAUUAAUUGCUAA'	1.9563*e* + 006	78	*NarL*	19	70.37
22	'UGAACUCAGUAAGAGCAGUGAUAAUCA'	2.1632*e* + 006	78	*NarL*	19	70.37
23	'CUUAAUGAAGUACUCAAUAGAUUUGUU'	2.4831*e* + 006	78	*NarL*	19	70.37
24	'AUUGCUAAUGAAAAACAUCAAUCCAAC'	4.2313*e* + 006	78	*NarL*	19	70.37
25	'UUAGGUAUUGAUAACAAUCAAUAGUAC'	4.5015*e* + 006	78	*NarL*	19	70.37
26	'GUCAUGAUGGCGCAUUAUUUUGUGGUG'	5.3802*e* + 005	78	*NarL*	19	70.37
27	'UUACCGCUGGUGCCGCAGGUCAGUUUC'	1.3878*e* + 006	78	*NarL*	19	70.37
28	'UUACCCAUGAAGCGGUAGGUAAAUGUG'	1.8717*e* + 006	78	*NarL*	20	74.074
29	'GGUAUCAAACUUCUCUUUAAACAGAUA'	4.2953*e* + 006	78	*NarL*	19	70.37
30	'CUAUUGCUUGUGCGGUAUUUGCCAAAA'	1.8083*e* + 006	78	*NarL*	20	74.074
31	'UUUAAUAAAAAAAAAGAUUAAGGGAUGA'	5.834*e* + 005	78	*NarL*	19	70.37
32	'AUACUAUCACUACCCUUUUUUUACACA'	3.6487*e* + 006	79	*NarP*	19	70.37

7.2.2 TFBS Detection algorithm using Distance Metrics Based on Center of Mass and Polyphase Mapping. This algorithm looks for sequences that can be considered as candidates of a TFBS based on their similitude with the available TFBSs. To do this, the algorithm applies a polyphase mapping to both the DNA sequence and the set of conserved sequences associated with a given transcription factor (TF). The center of Mass (CoM) of each set of these conserved sequences can be thought of as a consensus

sequence as it provides a distinctive representation of each family of conserved sequences as does a consensus sequence.

Having defined a CoM for each family of the conserved sequences, the algorithm then calculates distances of each one these conserved sequences to the center of mass. This is also applied to the genome under study by taking a sliding window of the DNA sequence of a length equal to the CoM. This window is translated one base at a time until the whole genome is being considered. At each alignment, the distance between the window sequence and the CoM is obtained. The maximum distance from the CoM to each of the conserved sequence is set as a threshold to locate those positions of the DNA sequence qualified to have a $TFBS$ candidate. This is done by selecting those positions of the sequences that have distances from the CoM less than or equal to the previous threshold. Out of those selected sequences the algorithm then looks for the ones having the same exact distances to the CoM as the distances from the conserved sequences. These sequences are the only sequences that are qualified to be $TFBS$ candidates based on their similitude of the conserved sequences that we already know. Figures 7.12 and 7.13 show a block diagram of the proposed algorithm.

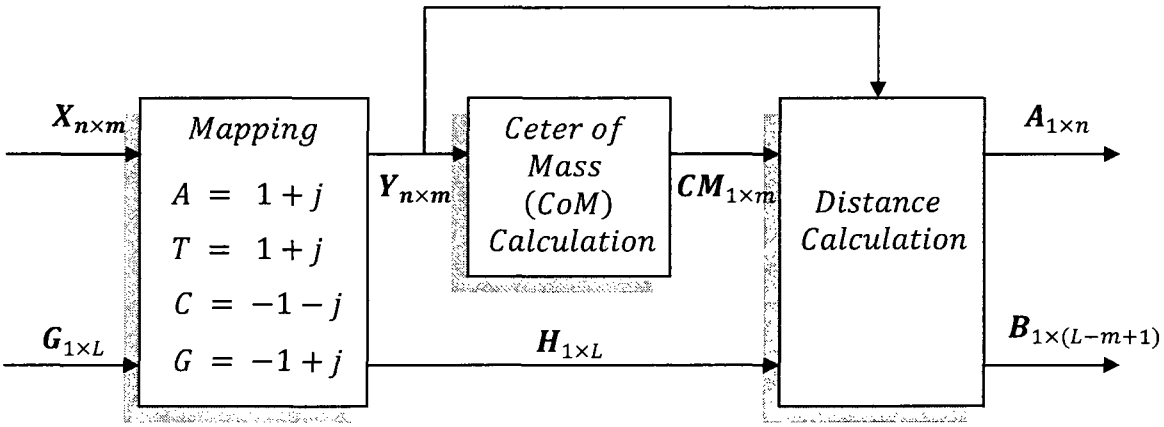


Figure 7.12. A block diagram of $TFBS$ detection algorithm (Part 1)

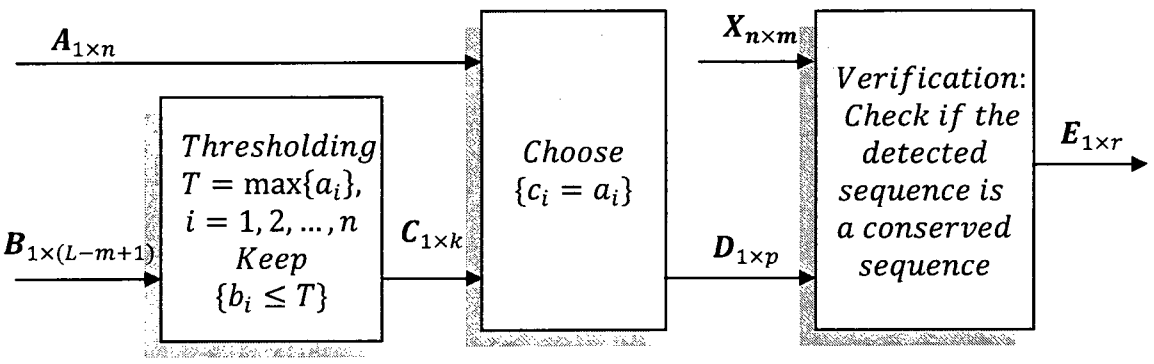


Figure 7.13. A block diagram of $TFBS$ detection algorithm (Part 2)

Mathematical Description. The genome under study is given by

$$\boldsymbol{G}_{1\times L} = [g_1, g_2, g_3, \ldots, g_L], \tag{7.20}$$

where L is the length of the genome in nucleobases.

Each family of transcription factor binding sites ($TFBSs$) consists of n conserved sequences with each conserved sequence consisting of m nucleobases. Each family is can be represented by a matrix of the form

$$\boldsymbol{X}_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}, \tag{7.21}$$

where $g_i \in \{A, G, C, T\}$, $i = 1, 2, \ldots, L$. $x_{kj} \in \{A, G, C, T\}$; $k = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

Using polyphase mapping, the genome $\boldsymbol{G}_{1\times L}$ becomes

$$\boldsymbol{H}_{1\times L} = [h_1, h_2, h_3, \ldots, h_L], \tag{7.22}$$

and the nucleobases matrix $\boldsymbol{X}_{n\times m}$ becomes

$$\boldsymbol{Y}_{n\times m} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1m} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & \cdots & y_{nm} \end{bmatrix}, \tag{7.23}$$

where

$$h_i = \begin{cases} 1+j \ , & if\ g_i = A \\ 1-j \ , & if\ g_i = T \\ -1+j \ , & if\ g_i = C \\ -1-j \ , & if\ g_i = G \end{cases}. \tag{7.24}$$

Similarly,

$$y_{ij} = \begin{cases} 1+j \ , & if\ x_{ij} = A \\ 1-j \ , & if\ x_{ij} = T \\ -1+j \ , & if\ x_{ij} = C \\ -1-j \ , & if\ x_{ij} = G \end{cases}. \tag{7.25}$$

Given $\boldsymbol{Y}_{n\times m}$, the center of mass ($CoM$) can be obtained as the mean of all rows of $\boldsymbol{Y}_{n\times m}$ that correspond to all conserved sequences. Hence, the center of mass is given by

$$\boldsymbol{CM}_{1\times L} = [cm_1, cm_2, cm_3, \dots, cm_L], \tag{7.26}$$

where

$$cm_i = \frac{1}{n}\sum_{k=1}^{n} y_{ki}. \tag{7.20}$$

The next step is to calculate the Euclidean distance of all conserved sequences of a given $TFBS$ family ($\boldsymbol{Y}_{n\times m}$) from the center of mass ($\boldsymbol{CM}_{1\times L}$). The resulting distance vector is given by

$$\boldsymbol{A}_{1\times n} = [a_1, a_2, a_3, \dots, a_n], \tag{7.21}$$

where a_i is obtained by

$$a_i = \left| \sqrt{\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{CM})^2} \right|, \tag{7.21}$$

$\boldsymbol{Y}_i$ is the i^{th} row of the matrix $\boldsymbol{Y}_{n\times m}$, i.e. the i^{th} binding site.

The same Euclidean distance measure is also used to calculate the distance between the center of mass ($\boldsymbol{CM}_{1\times L}$) and a window of the same length as the center of

mass being slid all over the genome one base at a time. This will yield a distance vector $\boldsymbol{B}_{1\times L-m+1}$ of length equal to $(L-m+1)$ given by

$$\boldsymbol{B}_{1\times L-m+1} = [b_1, b_2, b_3, \dots, b_{L-m+1}], \tag{7.22}$$

where b_i is obtained by

$$b_i = \left| \sqrt{\sum_{i=1}^{m} (\boldsymbol{H}_i - \boldsymbol{CM})^2} \right|, \tag{7.23}$$

$\boldsymbol{H}_i$ corresponds to the i^{th} window of the genome $\boldsymbol{H}_{1\times L}$ starting at position i and ending at position $i + m$.

At this point, the distance vector $\boldsymbol{B}_{1\times L-m+1}$ contains all the possible distances that the genome has from the center of mass ($\boldsymbol{CM}_{1\times L}$). To cut that down, a threshold T is calculated as the maximum distance in the $\boldsymbol{A}_{1\times n}$ vector obtained in the previous steps. All the distances greater than T are discarded. The resulting distance vector $\boldsymbol{C}_{1\times k}$ is defined as

$$\boldsymbol{C}_{1\times k} = \{c_i = b_j; b_j \leq T, i = 1,2,3, \dots, k; j = 1,2,3, \dots, L-m+1\}, \tag{7.24}$$

Now, the vector $\boldsymbol{C}_{1\times k}$ contains all the possible distances less than or equal to the threshold T. The next step is to discard all the values in $\boldsymbol{C}_{1\times k}$ that are different from the set of distances in the $\boldsymbol{A}_{1\times n}$ vector. The resulting distance vector $\boldsymbol{D}_{1\times p}$ is then defined as

$$\boldsymbol{D}_{1\times p} = \{d_i = c_j; c_j \in \boldsymbol{A}_{1\times n}, i = 1,2,3, \dots, p; j = 1,2,3, \dots, L-m+1\}, \tag{7.25}$$

where $p \leq k$.

Finally, the vector $\boldsymbol{D}_{1\times p}$ contains all the possible distances that are exactly the same as the distances of the original conserved sequences from their center of mass.

Some of these distances are false positive, i.e. they do not correspond to an actual conserved sequence. To filter out these false positives, each one of the detected sequences is compared to the original conserved sequences and only the verified ones are kept. The resulting distance vector $\boldsymbol{E}_{1\times r}$ is defined as

$$\boldsymbol{E}_{1\times r} = \{e_i = d_j; \boldsymbol{S}(d_j) = \boldsymbol{X}(d_j), i = 1,2,3,\dots,r; j = 1,2,3,\dots,p\}, \qquad (7.26)$$

where $\boldsymbol{S}(d_j)$ is the sequence of length m in the genome corresponding to the j^{th} distance in the $\boldsymbol{D}_{1\times p}$ vector. $\boldsymbol{X}(d_j)$ is the binding site in $\boldsymbol{X}_{n\times m}$ having the same distance d_j.

Simulation Results. The algorithm was applied to detect *TFBSs* in two different *Escherichia coli* genomes (both positive strand and negative strand were used). Table 7.7 summarizes the obtained simulation results in terms of the total number of *TFBSs* detected in both *E. coli genomes* (forward and reverse strands). The table gives similar results as Table 7.4 which further certifies the correctness of the obtained results.

Table 7.7. *TFBSs* detected in MG1655 and O175:H7 *E. coli* strains

E. coli Strain	*MG1655*		*O157:H7*	
Strand Orientation	*Positive Strand*	*Negative Strand*	*Positive Strand*	*Negative Strand*
TFBSs in the genome	956	1000	642	616
TFBSs in non-coding regions	817	59	540	527
TFBSs in coding regions	66	6.20	45	33
TFBSs overlapped in between	73	79	57	56

In Figure 7.14-7.17, the green areas represent the coding regions. Figures 7.14 and 7.15 represent the detected *TFBSs* in *MG1655 E. coli* genome (956 *TFBSs* are detected in the positive strand as shown in Figure 7.14, and 1000 *TFBSs* are detected in the negative strand as shown in Figure 7.15). The x-axis represents the base positions of the genome being considered, the y-axis represents the percentages that the detected *TFBSs* occur in the coding regions. In other words, a value of "100" in the y-axis means

that the detected *TFBS* totally occurs in the coding regions and the value of "0" means that the detected *TFBS* totally occurs in the non-coding regions. For all other values between 0 and 100, the detected *TFBS* occurs in between.

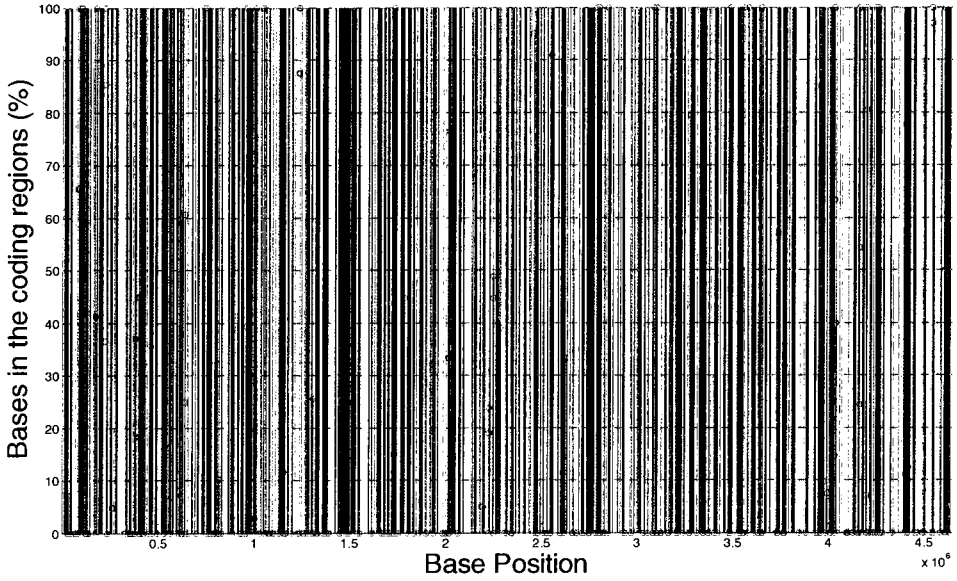


Figure 7.14. *TFBS* detection using MG1655 positive strand

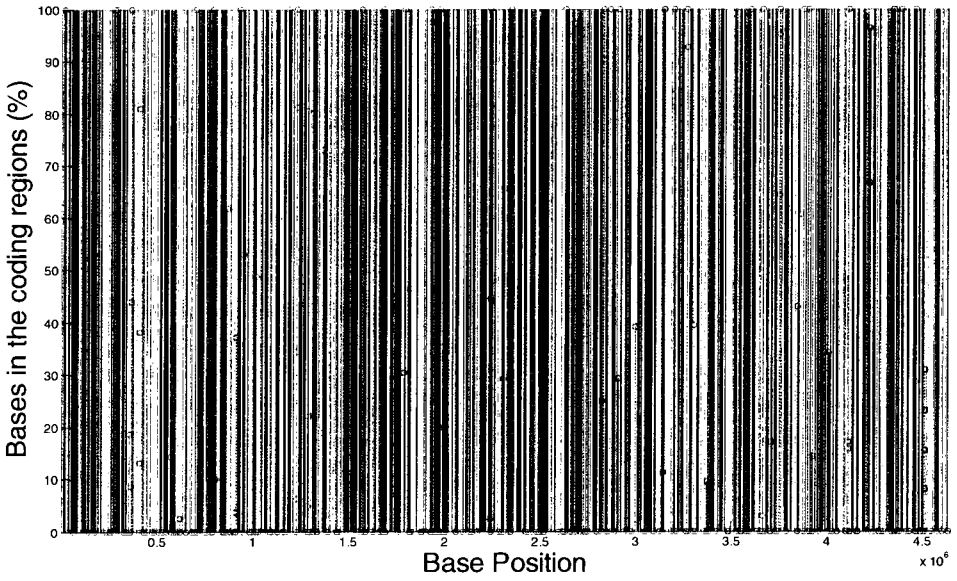


Figure 7.15. *TFBS* detection using MG1655 negative strand

The same analysis was also applied to $O157{:}H7\ E.coil$ genome (both positive and negative strands). The corresponding simulation results are shown in Figures 7.16-7.17 (642 $TFBSs$ detected in the positive strand, and 616 $TFBSs$ detected in the negative strand).

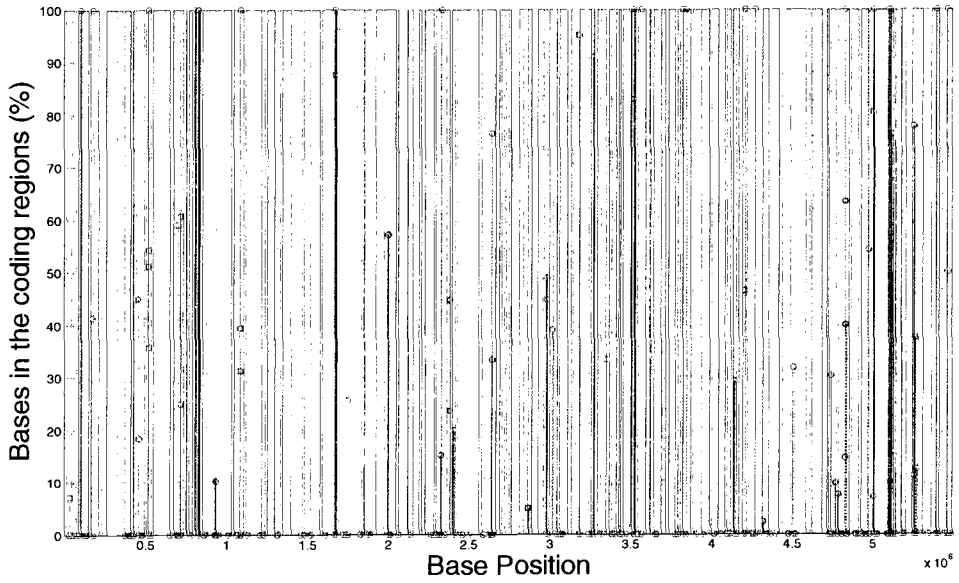


Figure 7.16. $TFBS$ detection using $O157{:}H7$ Positive strand

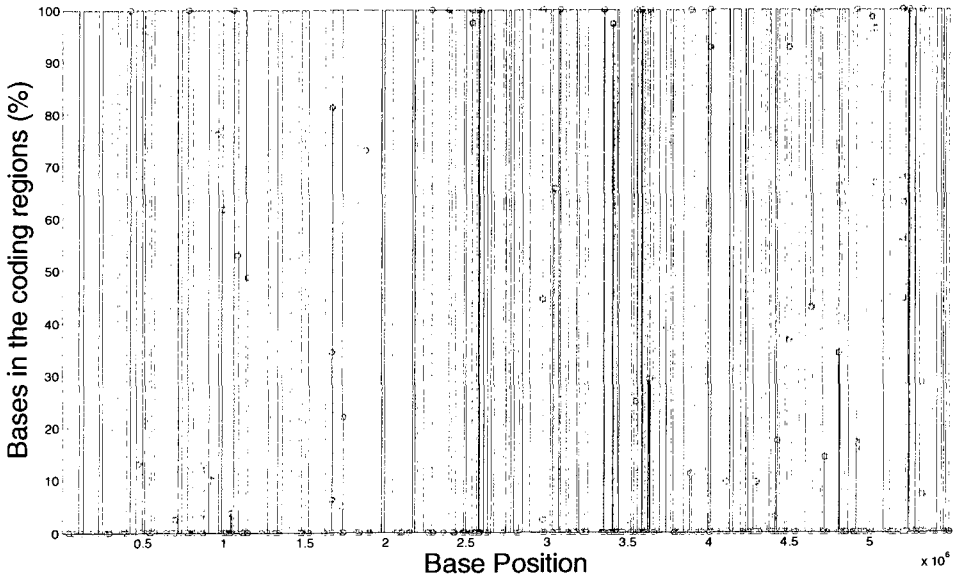


Figure 7.17. $TFBS$ detection using $O157{:}H7$ negative strand

7.3 Discovery of New Motif Sequences

The Motif finding problem involves finding short DNA sequences or consensus patterns that occur surprisingly often, without any prior knowledge of what these patterns looks like. A visualization technique and algorithm for finding motifs in DNA sequences is proposed in [121]. Due to the fact that human vision is not able to discern patterns from millions of strings of the four nucleobases $\{A, G, C\ and\ T\}$; we need to develop a method capable of doing that. This method is designed to identify any substring of bases that occur surprisingly very often in the DNA sequence under study. If a sequence occurs more frequently than expected, then it can be assumed that this sequence is more probable to have a particular significance in the biological sense. Such sequences can be subjected to biological research to verify their significance.

Out of the set of possible $TFBSs$ resulted from applying the detection algorithm described in section 7.2.2 (12150 possible sequences), this method searches if there is any sequence in the whole set of possible $TFBSs$ appears more than once. Based on this, the method classifies each possible sequence depending on the number of times it appears in the DNA sequence as well as the TF that this particular sequence belongs to.

Once the sequences that appear more than once are identified, they can be considered as possible motif sequences. These sequences can be tested to see if they appear in other $E.coli$ genomes. If yes, this will increase the chance that these sequences are potential $TFBSs$. Hence, they will need to be subjected to biological experiments to verify their significance. This type of analysis was applied to $MG1655$ and $O157{:}H7$ $E.coli$ strains. Tables 7.8, 7.9, and 7.10 show the sequences that appeared five, four, and three times in $MG1655\ E.coil$ genome (the positive strand is considered here)

respectively. They also show the TFs that the sequences detected belong to, and the positions at which they were detected.

Table 7.8. Sequences that appear five times in $MG1655$ $E.coli$ genome

#	*Sequence*	*TF*	P_1	P_2	P_3	P_4	P_5
1	UUUGUUCAUGCCGGAUGCGGCGUGAACG	64	698639	1550086	1550264	3636830	4406722
2	AUGGGUCAGCGACUUAUAUUCUGUAGC	78	226333	3941901	4035673	4166794	4208196
3	UGAAGUCGGAAUCGCUAGUAAUCGUGG	78	225100	3940759	4034449	4165567	4207054

Table 7.9. Sequences that appear four times in $MG1655$ $E.coli$ genome

#	*Sequence*	*TF*	P_1	P_2	P_3	P_4
1	AAAAAAUUGAUGGAACAUAUUUCUAUU	53	735677	736057	737324	3622018
2	GUGUUUACCGGAUCAUGCGCCAGAAUGCGC	64	381169	1648927	3184797	4496434
3	CUGUUGACUAUUUGUGCCGUUAUUUCUG	68	735758	737405	1529930	3622099
4	AGUGGCCGUGAAAGAAAGCAAUCAGCG	78	381254	1649012	3184882	4496519
5	AAAUGGAACUUAAAAAAAUUGAUGGAAC	78	735666	736046	737313	3622007
6	AAUUCGCAUUUUAUGUUUAAAAAUUGA	78	735944	737210	1529735	3621904
7	UUUUAUGUUUAAAAAUUGAGAUAUUCC	78	735952	737218	1529743	3621912
8	UGAAAUGAUGAAAUGAUGAAAUGAUGA	78	1197678	1197686	1197694	1197702
9	UAACAUGCUGGAUAUUAAAGGAAAAAU	78	736524	737791	1530316	3622485

Table 7.10. Sequences that appear three times in $MG1655$ $E.coli$ genome

#	*Sequence*	*TF*	P_1	P_2	P_3
1	AUAUGGACGGGCAAAAAAGUGGAUAGCGCACGCGC	7	380880	3184508	4496145
2	UUUUUAUGUUUUUAAAUUAAGUUAUAAAAAUUU	52	4569516	4569516	4569516
3	ACCAAUAUCACUUAAAUAAGUGAUAGUCUUAAU	52	380511	3184139	4495776
4	CCUGUUUCCCGGUGAGGACGGUUACAGCCG	64	729180	3617197	3760184
5	CACUGUGGCUGGUGCGCGGCGGCGUGGCGA	64	729218	3617235	3760222
6	GACGGGUCAUCCGCCUGACCAGUGAAAACG	64	730727	3618744	3761731
7	GGCGUUCACGCCGCAUCCGGCAUGAACAAC	64	138776	2682214	2943994
8	UUAAAGGAAAAAUCAUCACAACUGAUGC	68	736538	1530330	3622499
9	CCGGAGAACAGGUUGUUCCUGCCUCU	77	380762	3184390	4496027
10	AGACAUGUCUUGCCCAAUAUACUCAAC	78	270845	279195	1468400
11	AUAGUCUUAAUACUAGUUUUUAGACUA	78	380533	3184161	4495798
12	AGAACGGUGAAACCACUGUAAUGAUAG	78	16109	607952	2513015
13	AAACACCGAAAGAGAUAAUUGAAAGGG	79	270932	279282	1468487

In order to check whether the 25 sequences detected in Tables 7.8-7.10 are potential motif sequences or not, a different $E.coil$ genome ($O157H7$) was scanned to see if any of those sequences appear any time. Tables 7.11-7.13 show the obtained results.

Table 7.11. Sequences that appear five times in $O157{:}H7$ *E. coli* genome

#	*Sequence*	*TF*	P_1	P_2	P_3	P_4	P_5
1	UUUGUUCAUGCCGGAUGCGGCGUGAACG	64	2073484	5262871	–	–	–
2	AUGGGUCAGCGACUUAUAUUCUGUAGC	78	229664	4737724	4834218	4978399	5019425
3	UGAAGUCGGAAUCGCUAGUAAUCGUGG	78	228431	4736581	4832983	4977256	5018282

Table 7.12. Sequences that appear four times in $O157{:}H7$ *E. coli* genome

#	*Sequence*	*TF*	P_1	P_2	P_3	P_4
1	AAAAAAUUGAUGGAACAUAUUUCUAUU	53	272661	814371	2051038	272661
2	GUGUUUACCGGAUCAUGCGCCAGAAUGCGC	64	–	–	–	–
3	CUGUUGACUAUUUGUGCCGUUAUUUCUG	68	2051119	–	–	2051119
4	AGUGGCCGUGAAAGAAAGCAAUCAGCG	78	–	–	–	–
5	AAAUGGAACUUAAAAAAAUUGAUGGAAC	78	272650	814360	2051027	272650
6	AAUUCGCAUUUUAUGUUUAAAAAUUGA	78	814257	2050924	–	814257
7	UUUUAUGUUUAAAAAUUGAGAUAUUCC	78	2050932	–	–	2050932
8	UGAAAUGAUGAAAUGAUGAAAUGAUGA	78	–	–	–	–
9	UAACAUGCUGGAUAUUAAAGGAAAAAU	78	2051505	–	–	2051505

Table 7.13. Sequences that appear three times in $O157{:}H7$ *E. coli* genome

#	*Sequence*	*TF*	P_1	P_2	P_3
1	AUAUGGACGGGCAAAAAAGUGGAUAGCGCACGCGC	7	–	–	–
2	UUUUUAUGUUUUUAAAUUAAGUUAUAAAAAUUU	52	–	–	–
3	ACCAAUAUCACUUAAAUAAGUGAUAGUCUUAAU	52	–	–	–
4	CCUGUUUCCCGGUGAGGACGGUUACAGCCG	64	808383	4505258	4929251
5	CACUGUGGCUGGUGCGCGGCGGCGUGGCGA	64	–	–	–
6	GACGGGUCAUCCGCCUGACCAGUGAAAACG	64	809930	4506805	4930798
7	GGCGUUCACGCCGCAUCCGGCAUGAACAAC	64	143148	3668372	–
8	UUAAAGGAAAAAUCAUCACAACUGAUGC	68	273142	814852	2051519
9	CCGGAGAACAGGUUGUUCCUGCCUCU	77	–	–	–
10	AGACAUGUCUUGCCCAAUAUACUCAAC	78	–	–	–
11	AUAGUCUUAAUACUAGUUUUUAGACUA	78	–	–	–
12	AGAACGGUGAAACCACUGUAAUGAUAG	78	–	–	–
13	AAACACCGAAAGAGAUAAUUGAAAGGG	79	–	–	–

Based on Table 7.11, it can be observed that all of sequences that appeared five times in the *E. coli* $MG1655$ genome, appeared in the *E. coli* $O157H7$ genome as well. In addition, they also appear several times, what increases the chance that these three sequences are more probable to be considered potential motif sequences.

According to Table 7.12, it can be observed that the new scanned genome ($O157{:}H7$) contains 6 out of the 9 sequences that appeared four times in the *E. coli*

$MG1655$ genome. Although some of them have not been found in the new genome, the six identified sequences can be considered as potential motif sequences and hence should be further investigated in biological experimentation to check their significance.

Table 7.13 shows that approximately 70% of the set of sequences that appeared three times in the $MG1655$ genome are not present in $O157{:}H7$ genome. This will decrease the chance that these sequences are potential $TFBSs$.

The remaining sequences can still be considered for further biological analysis.

In conclusion, we can state that while the sequences that appeared five or four times have a higher probability of being in other genomes and hence must be considered as potential motif sequences, the ones that appear three, two and one times can be discarded. Table 7.14 shows the set of sequences that can be considered as potential motif sequences based on their frequency of occurrence in both $E.coli$ genomes considered in this analysis. These sequences are recommended for further laboratory experimentation to see if they possess any biological significance.

Table 7.14. Possible potential motif sequences in E.coli genome

#	*Sequence*	*TF*	*Frequency of occurrence*	
			in MG1655	*in O157:H7*
1	UUUGUUCAUGCCGGAUGCGGCGUGAACGCC	64	5	2
2	AUGGGUCAGCGACUUAUAUUCUGUAGC	78	5	5
3	UGAAGUCGGAAUCGCUAGUAAUCGUGG	78	5	5
4	AAAAAAUUGAUGGAACAUAUUUCUAUU	53	4	3
5	CUGUUGACUAUUUGUGCCGUUAUUUCUG	68	4	1
6	AAAUGGAACUUAAAAAAAUUGAUGGAAC	78	4	3
7	AAUUCGCAUUUUAUGUUUAAAAAUUGA	78	4	2
8	UUUUAUGUUUAAAAAUUGAGAUAUUCC	78	4	1
9	UAACAUGCUGGAUAUUAAAGGAAAAAU	78	4	1

Having analyzed only two *E. coli* genomes, nine potential motif sequences were detected (as shown in Table 3.14). Therefore, if this method of motif finding is applied to more other genomes, other potential motif sequences can be identified. Hence, the

previous analysis of using the frequency weight to detect and identify new motif sequences can efficiently yield other potential motif candidates that are recommended to be subjected for further analysis.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

In this thesis, we have analyzed the process of translation in gene expression using a communications theory approach. The two fields of communications engineering and molecular biology have only in the last few years started a cautious rapprochement, 50 years after Claude E. Shannon's "An Algebra for Theoretical Genetics" [155]. This thesis endeavors at fostering the cooperation between the two fields and to call attention to this interdisciplinary topic in both communities. This cooperation will provide a wide range of expertise to solving new research problems, and will produce new tools and techniques to explore and analyze these problems from different perspectives. In the following sections, the main aspects and achievements of this thesis are summarized. Moreover, possible future research directions are presented.

8.1 Summary

The two introductory chapters detailed the theoretical background and the literature survey, of the methods, techniques and models that are used to analyze the information theoretic aspects of the genetic system. The following chapters investigated the process of translation in gene expression for prokaryotic bacterial organisms.

Chapter 3 provided the biological background needed to understand the communication theoretic models in later chapters. Important literature references were provided to give a broader introduction to the topic than possible in this thesis. The focus was on gene translation, the process of protein synthesis vital to all living organisms.

The following chapters (Chapter 4 to Chapter 7) detailed the communications theoretic modeling for the process of translation in prokaryotic organisms. A novel use of techniques and principles from communications engineering for modeling, identification and analysis of genomic regulatory elements and biological sequences is presented.

The codebook model investigated in Chapter 4 is based on a variable length codebook and a metric for the process of translation in gene expression. In this model it was assumed the ribosome decodes the $mRNA$ sequence by using the 3′ end of the $16S$ $rRNA$ molecule as an embedded codebook. The metric used an exponentially weighting free energy decoding algorithm to identify the $Shine\text{-}Dalgarno$ (SD) sequence. This algorithm allowed for better resolution and flexibility in detecting the translational signals by a simple change of parameter values. The validity and biological relevance of this model was verified by testing the effect of different types of mutations in the ribosome on protein synthesis. Results were compared to biological experimental data with published records and proved to be consistent. The analysis of the results made possible by this model can support, and sometimes replace or reduce time and cost consuming laboratory experimentation. It also can serve as a way to introduce new lines of biological research. In practice, these results can lead to better recognition of signals in translation, hence, enhancing $in\ vitro$ translation systems in genetic engineering. The chapter then presented different analyses for the detection mechanism that the ribosome uses to identify the translational signals based on a communications theory approach. Concepts of Euclidean distance and cross correlation were utilized to emulate the ribosome detection mechanism. Results showed that the proposed approach is able to

identify coding and non-coding regions by a clear difference in the detection signal ripple in both regions.

The coding theory based models investigated in Chapters 5 and 6 described quantitatively the behavior of the ribosome during the translation process in gene expression. Chapter 5 presented a block code model for the process of translation in prokaryotic organisms. We have employed several minimum distance decoders to verify the validity of the model based on the free energies involved in the binding between the ribosome and the $mRNA$ sequence. The model was tested on five different bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs ($GeneMark$ and $Glimmer$). The exponentially weighted free energy distance decoder resulted in the best distinction between the translated (obtained from $GenBank$), hypothetically translated ($GLIMMER$ and $GeneMark$ false positives), and non-translated (all open reading frames that follow certain criteria with both previous groups excluded) sequence groups.

Chapter 6 investigated a convolutional code model to analyze the process of translation using table-based decoding. In the investigated model, the messenger RNA ($mRNA$) sequence was viewed as a noisy convolutional encoded signal, and ribosome as a table-based convolutional decoder. The $16S$ ribosomal RNA ($16S\ rRNA$) sequence was used to form decoding masks for table-based decoding. A syndrome matrix was calculated. The model was tested on five different bacterial genomes. The obtained results verify the validity and the significance of the investigated convolutional code model and its biological relevance by being able to identify the right start codons from

the false ones. The conducted analysis was applied to different (n, k, m) combinations as described in Table 6.3. Only (2,1,2) and (3,1,2) convolutional code model were able to identify the right initiation codons from the false one in the four test groups (described in details in section 5.5). This can be explained by the fact that for both of (2,1,2) and (3,1,2) cases, the length of the obtained g-masks is equal to six bases which is equal to the so called anti-Shine-Dalgarno sequence ($UCCUCC$) located at the 3' end of the $16S\ rRNA$ in the ribosome. This sequence is only available in prokaryotes and generally located eight basepairs upstream of the start codon. All the sequence groups in the (2,1,2) and (3,1,2) models achieve a global minimum syndrome value at the zero position which corresponds to the first base in the initiation codon. In general, the block code model yielded better results than the convolutional code model in terms of its ability to detect the *Shine-Dalgarno* domain, the non-random domain, and the initiation site.

Chapter 7 investigated the problem of transcription factor binding site detection in an attempt to utilize that in gene identification. Two novel techniques for $TFBS$ detection were presented. While the first detection technique was based on a frequency weight matrix (FWM) concept, the second one used center of mass based metrics and polyphase mapping. Simulation results show that around 85% of the detected TFBSs are located in the non-coding region, 6.5% are located in the coding regions, and only 8.5% are overlapped between coding and non-coding regions.

8.2 Future Research Directions

The analysis conducted in the codebook model is based on the last 13 bases sequence of the $16S\ rRNA$ molecule (a part of the ribosome small subunit, $30S$, in

prokaryots). The analysis also incorporated a study for certain types of point mutations in the latter sequence in almost all base positions. *Jacob*, *Hui* and *De Boer* mutations are examples of these point mutations. Future work can target the analysis of other types of mutations in the ribosome structure to verify the correctness and the biological relevance of the codebook model. The same analysis can be extended to study eukaryotes. The equivalent ribosome structure in eukaryotes (corresponding to $16S\ rRNA$ in prokaryotes) is called $18S\ ribosomal\ RNA$ ($18S\ rRNA$). This structure makes the initial contact with the $mRNA$ sequence to initiate the translation process.

Based on the results obtained from the block code model in chapter 5, a minimum distance trough occurs in the region between the -20 and -10 base position for both translated and hypothetically translated sequence groups. The exponentially weighted free energy distance decoder resulted in the best resolution (in terms of bigger trough amplitude) compared to results of both minimum Hamming distance decoder and minimum free energy decoder. In other words, the results of the exponentially weighted free energy decoder provided the greatest distinction between the translated, hypothetically translated (obtained by *GLIMMER* and *GeneMark*), and non-translated sequence groups. Based on these results, we can design a 10-bases classification system that can make sufficiently correct real-time decisions of weather a given start codon is a valid initiation codon or not.

The results obtained in the convolutional code model (Chapter 6) are based on using the last 13 bases sequence to form decoding masks for table-based decoding. These decoding masks were used to calculate the syndrome assuming that the $mRNA$ sequence is a received noisy convolutional encoded signal. Based on table-based decoding theory,

we can use the g-mask vector to calculate the corresponding generators. Assuming that the $mRNA$ sequence is the information sequence to be transmitted, these generators can be applied to the $mRNA$ sequence to get the corresponding convolutionally encoded sequence. The same analysis can be applied to the new encoded sequence to calculate the syndrome and check if that will yield better results compared to the obtained ones in terms of ability to distinguish the four test groups presented.

The two methods proposed in Chapter 7 for TFBS detection can be further investigated on different prokaryotic genome to detect new TFBS. The distribution of transcription factor binding sites in coding and non-coding region obtained from the analysis conducted, can be utilized to check if a given start codon is a valid initiation codon or not. The decision made based on this analysis can be combined with the decision made based on the 10-bases classification system suggested in the block code model above. This will improve the accuracy of the classification system as it will yield less number of false positives.

In general, the most interesting focus of future research would be a detailed analysis of gene expression in eukaryotes, especially in the human genome. Even though the work in this thesis was restricted on prokaryotic organisms (five different bacterial genomes were investigated), the conducted analysis can be extended to study eukaryotes as well. This will require a broader understanding of underlying interactions in the gene expression mechanism in eukaryotes.

Protein-DNA interactions constitute a basic step of all cellular processes. For example, transcription factors bind to positions close to the gene start site. The binding sites of these proteins could be investigated in a way similar to the last 13 bases

sequences analyses presented in this thesis. Transcription factor binding sites are available in databases like TRANSFAC[5].

[5] BIOBASE Biological Databases. TRANSFAC: Gene Transcription Factor Database. http://www.biobase-international.com/pages/index.php?id=transfac

BIBLIOGRAPHY

[1] Abdi, H. "Distance." Encyclopedia of Measurement and Statistics. 2007.

[2] Adami, C. "Information theory in molecular biology." *Physics of Life Reviews*. 1.1 (2004): 3-22.

[3] Aggarwal, G. and R. Ramaswamy. "Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER." *J Biosci*. 27.1 Suppl 1 (2002): 7-14.

[4] Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. fourth edition ed. New York: *Garland Science*, 2002.

[5] Almagor, H. "Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach." *J Theor Biol*. 117.1 (1985): 127-36.

[6] Altschul, S. F. "Amino acid substitution matrices from an information theoretic perspective." *J Mol Biol*. 219.3 (1991): 555-65.

[7] Arneodo, A., Y. dAubentonCarafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes. "Wavelet based fractal analysis of DNA sequences." *Physica D-Nonlinear Phenomena*. 96.1-4 (1996): 291-320.

[8] Arques, D. G. and C. J. Michel. "A code in the protein coding genes." *BioSystems*. 44 (1997): 107-34.

[9] Ashikhmin, A., A. Barg, and DIMACS (Group). *Algebraic coding theory and information theory : DIMACS workshop, algebraic coding theory and information theory, December 15-18, 2003, Rutgers University, Piscataway, New Jersey*. DIMACS series in discrete mathematics and theoretical computer science,. Providence, R.I.: *American Mathematical Society*, 2005.

[10] Baisnee, P. F., S. Hampson, and P. Baldi. "Why are complementary DNA strands symmetric?" *Bioinformatics*. 18.8 (2002): 1021-33.

[11] Battail, G. "Information theory and error correcting codes in genetics and biological evolution. Introduction to Biosemiotics." Thursday, May 10, 2007 ed.: *Springer Netherlands*, 2007. 299-345.

[12] Battail, G. "Information theory and error correcting codes in genetics and biological evolution. Introduction to Biosemiotics." Thursday, May 10, 2007 ed: *Springer Netherlands*, 2007. 299-345.

[13] Battail, G. *Information theory and error correcting codes in genetics and biological evolution. Introduction to Biosemiotics. Springer Netherlands*, 2007.

[14] Battail, G. "Should genetics get an information-theoretic education?" *Ieee Engineering in Medicine and Biology Magazine.* 25.1 (2006): 34-45.

[15] Battail, G. "An engineer's view on genetic information and biological evolution." *Biosystems.* 76.1-3 (2004): 279-90.

[16] Battail, G. "An engineer's view on genetic information and biological evolution." *Biosystems.* 76.1-3 (2004): 279-90.

[17] Battail, G. "Does information theory explain biological evolution?" *Europhysics Letters.* 40.3 (1997): 343-48.

[18] Berlekamp, E. R. *Algebraic coding theory.* Rev. ed. Laguna Hills, Calif.: *Aegean Park Press*, 1984.

[19] Besemer, J. and M. Borodovsky. "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." *Nucleic Acids Res.* 33.Web Server issue (2005): W451-4.

[20] Besemer, J., A. Lomsadze, and M. Borodovsky. "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." *Nucleic Acids Research.* 29.12 (2001): 2607-18.

[21] Bitzer, D. L., A. Dholakia, H. Koorapaty, and M. A. Vouk. "On Locally Invertible Rate-1/n Convolutional Encoders." *IEEE Transactions on Information Theory.* 44.1 (1998): 420 - 22.

[22] Bitzer, D. L. and M. A. Vouk. "A table-driven (feedback) decoder." *Proceedings of the Tenth Annual International Phoenix Conference on Computers and Communications.* (1991): 385-92.

[23] Borodovsky, M. and J. McIninch. "GENEMARK: Parallel Gene Recognition for Both DNA Strands." *Computers & Chemistry.* 17.2 (1993): 123-33.

[24] Borodovsky, M. and J. McIninch. "GENEMARK: Parallel Gene Recognition for Both DNA Strands." *Computers and Chemistry.* 17.2 (1993): 123-33.

[25] Burden, S., Y. X. Lin, and R. Zhang. "Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences." *Bioinformatics.* 21.5 (2005): 601-7.

[26] Cosic, I. *The resonant recognition model of macromolecular bioactivity : theory and applications*. BioMethods v. 8. Basel ; Boston: *Birkhäuser Verlag*, 1997.

[27] Cosic, I. *The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications*. Basel, Switzerland: *Birkhauser Verlag*, 1997.

[28] Crowley, E. M. "A Bayesian method for finding regulatory segments in DNA." *Biopolymers*. 58.2 (2001): 165-74.

[29] Dawy, Z., P. Hanus, J. Weindl, J. Dingel, and F. Morcos. "On genomic coding theory." *European Transactions on Telecommunications*. 18.8 (2007): 873-79.

[30] Dawy, Z., F. Gonzalez, J. Hagenauer, and J. C. Mueller. "Modeling and analysis of gene expression mechanisms: a communication theory approach." *IEEE International Conference on Communications (ICC)*. 2 (2005): 815- 19.

[31] Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics*. 23.6 (2007): 673-9.

[32] Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. "Improved microbial gene identification with GLIMMER." *Nucleic Acids Res*. 27.23 (1999): 4636-41.

[33] Dholakia, A., M. A. Vouk, and D. L. Bitzer. "Table based decoding of rate one-half convolutional codes." *IEEE Transactions on Communications*. 43.234 (1995): 681 - 86.

[34] Dholakia, A., M. A. Vouk, and D. L. Bitzer. "Table Based Decoding of Rate One-Half Convolutional-Codes." *Ieee Transactions on Communications*. 43.2-4 (1995): 681-86.

[35] Dholakia, A. *Introduction to Convolutional Codes with Applications*. Norwell, Massachusetts: *Kluwer Academic Publishers*, 1994.

[36] Dholakia, A. "Introduction to Convolutional Codes with Applications." *Kluwer Academic Publishers, Norwell, Massachusetts*. (1994).

[37] Dholakia, A., T. M. Lee, D. L. Bitzer, M. A. Vouk, L. Wang, and P. D. Franzon. "An efficient table-driven decoder for one-half rate convolutional codes." *Proceedings of the 30th ACM Annual Southeast Regional Conference*. (1992): 116 - 23.

[38] E. May, M. V., D. Bitzer, and D. Rosnick. "Coding model for translation in E.coli K-12." *Proceedings of The First Joint BMES/EMBS Conference*. 2 (1999): 1178.

[39] Eddy, S. R. "Hidden Markov models and genome sequence analysis." *Faseb Journal.* 12.8 (1998): A1327-A27.

[40] Eddy, S. R. "Profile hidden Markov models." *Bioinformatics.* 14.9 (1998): 755-63.

[41] Eddy, S. R. "Hidden Markov models." *Current Opinion in Structural Biology.* 6.3 (1996): 361-65.

[42] Eigen, M. "The origin of genetic information: viruses as models." *Gene.* 135.1-2 (1993): 37-47.

[43] Elebeoba E. May. "Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms." North Carolina State University, March 2002.

[44] Eskesen, S. T., F. N. Eskesen, B. Kinghorn, and A. Ruvinsky. "Periodicity of DNA in exons." *BMC Mol Biol.* 5 (2004): 12.

[45] Farabaugh, P. J. "Programmed translational frameshifting." *Annu Rev Genet.* 30 (1996): 507-28.

[46] Fickett, J. W. "The gene identication problem: an overview for developers." *Computers and Chemistry.* 20.1 (1996): 103-18.

[47] Fowler, T. B. "Computation as a thermodynamic process applied to biological systems." *Int J Biomed Comput.* 10.6 (1979): 477-89.

[48] Fowler, T. B. "Computation as a thermodynamic process applied to biological systems." *Int J Biomed Comput.* 10.6 (1979): 477-89.

[49] Francisco M. De La vega, Carlos Cerpa, and G. Guarneros. "A mutual information analysis of tRNA sequence and modification patterns distinctive of species and phylogenetic domain." *In Pacific Symposium on Biocomputing.* (1996): 710-11.

[50] Freeland, S. J., T. Wu, and N. Keulmann. "The case for an error minimizing standard genetic code." *Origins of Life and Evolution of the Biosphere.* 33.4-5 (2003): 457-77.

[51] Frishman, D., A. Mironov, and M. Gelfand. "Starts of bacterial genes: estimating the reliability of computer predictions." *Gene.* 234.2 (1999): 257-65.

[52] Gatlin, L. L. *Information Theory and the Living System.* New York, NY: *Columbia University Press*, 1972.

[53] Gatlin, L. L. "The information content of DNA. II." *J Theor Biol.* 18.2 (1968): 181-94.

[54] Gong, L., N. Bouaynaya, and D. Schonfeld. "Information-Theoretic Bounds of Evolutionary Processes Modeled as a Protein Communication System." *IEEE/SP 14th Workshop on Statistical Signal Processing.* (2007): 1-5.

[55] Goringer, H., K. Hujazif, E. Murgolat, and A. Dahlberg. "Mutations in 16S rRNA that affect UGA (stop codon)-directed translation termination." *Proc. Natl. Acad. Sci.* 88 (August 1991): 6603-07.

[56] Greive, S. J. and P. H. v. Hippel. "Thinking quantitatively about transcriptional regulation." *Nature Reviews Molecular Cell Biology.* 6.3 (March 2005): 221-32.

[57] Griffiths, A. J. F., W. M. Gelbart, R. C. Lewontin, and J. H. Miller. New York: *W. H. Freeman Publishers*, 2002.

[58] Grundy, W. N., T. L. Bailey, C. P. Elkan, and M. E. Baker. "Meta-MEME: Motif-based hidden Markov models of protein families." *Computer Applications in the Biosciences.* 13.4 (1997): 397-406.

[59] Gupta, M. K. "The quest for error correction in biology. Recent developments in codes and biology." *IEEE Eng Med Biol Mag.* 25.1 (2006): 46-53.

[60] Hannenhalli, S. S., W. S. Hayes, A. G. Hatzigeorgiou, and J. W. Fickett. "Bacterial start site prediction." *Nucleic Acids Research.* 27.17 (1999): 3577-82.

[61] Hanus, P., B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J. C. Mueller. "Information and communication theory in molecular biology." *Electrical Engineering (Archiv fur Elektrotechnik).* 90 / 2 (December 2007): 161-73.

[62] Hanus, P., B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J. C. Mueller. "Information and communication theory in molecular biology." *Electrical Engineering.* 90.2 (2007): 161-73.

[63] Hayes, B. "The invention of the genetic code." *American Scientist.* 86.1 (1998): 8-14.

[64] Hayes, W. S. and M. Borodovsky. "How to interpret an anonymous bacterial genome: Machine learning approach to gene identification." *Genome Research.* 8.11 (1998): 1154-71.

[65] Henderson, J., S. Salzberg, and K. H. Fasman. "Finding genes in DNA with a Hidden Markov Model." *Journal of Computational Biology.* 4.2 (1997): 127-41.

[66] Hui, A. and H. A. de Boer. "Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli." *Proc Natl Acad Sci U S A.* 84.14 (1987): 4762-6.

[67] Jacob, W. F., M. Santer, and A. E. Dahlberg. "A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins." *Proc Natl Acad Sci U S A.* 84.14 (1987): 4757-61.

[68] Karin, M. "Too many transcription factors: positive and negative interactions." *New Biol.* 2.2 (1990): 126-31.

[69] Kozak, M. "Initiation of translation in prokaryotes and eukaryotes." *Gene.* 234.2 (1999): 187-208.

[70] Krogh, A., I. S. Mian, and D. Haussler. "A Hidden Markov Model That Finds Genes in Escherichia-Coli DNA." *Nucleic Acids Research.* 22.22 (1994): 4768-78.

[71] Kumar, P. V. "Error Control Coding - Fundamentals and Applications - Lin,S, Costello,Dj." *Ieee Communications Magazine.* 21.6 (1983): 48-49.

[72] L. Kari, R. Kitto, and G. Thierrin. "Codes, involutions and DNA encodings." *Formal and Natural Computing, W. Brauer, H. Ehrig, J. Karhumaki, A. Salomaa Eds., LNCS 2300, Springer-Verlag.* (2002): 376–93.

[73] Larry S. Liebovitch, Y. T., Angelo T. Todorov, and Leo Levine. "Is There an Error Correcting Code in the Base Sequence in DNA?" *BIOPHYSICAL JOURNAL.* 71.3 (1996): 1539-44.

[74] Latchman, D. S. "Transcription factors: an overview." *Int J Biochem Cell Biol.* 29.12 (1997): 1305-12.

[75] Lewin, B. *Genes IX.* New York, NY: *Jones & Bartlett Publishers*, 2007.

[76] Lewin, B. *Genes VII.* New York, NY: *Oxford University Press Inc.*, 2000.

[77] Lewin, B. *Genes V.* New York, NY: *Oxford University Press*, 1995.

[78] Li, M. *An introduction to kolmogorov complexity and its applications.* Texts in computer science. 3rd ed. New York: *Springer*, 2008.

[79] Liebovitch, L. S., Y. Tao, A. T. Todorov, and L. Levine. "Is There an Error Correcting Code in the Base Sequence in DNA?" *BIOPHYSICAL JOURNAL.* 71.3 (1996): 1539-44.

[80] Lin, S. and D. J. Costello Jr. *Error Control Coding*. 2 edition ed: *Prentice Hall*, 2004.

[81] Lodish, H., A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology*. 6th edition ed. New York: *W. H. Freeman Publishers*, 2007.

[82] Loewenstern, D. and P. N. Yianilos. "Significantly lower entropy estimates for natural DNA sequences." *Journal of Computational Biology*. 6.1 (1999): 125-42.

[83] Loewenstern, D. and P. N. Yianilos. "Significantly lower entropy estimates for natural DNA sequences." *J Comput Biol*. 6.1 (1999): 125-42.

[84] Long, F., H. Liu, C. Hahn, P. Sumazin, M. Q. Zhang, and A. Zilberstein. "Genome-wide prediction and analysis of function-specific transcription factor binding sites." *In Silico Biol*. 4.4 (2004): 395-410.

[85] Lukashin, A. V. and M. Borodovsky. "GeneMark.hmm: new solutions for gene finding." *Nucleic Acids Research*. 26.4 (1998): 1107-15.

[86] Lun Huang, Mohammad Al Bataineh, Guillermo E. Atkin, Maria Parra, Maria del Mar Perez, and Wei Zhang. "Identification of Transcription Factor Binding Sites Based on the Chi-Square (X^2) distance of a Probabilistic Vector Model." *2009 International Conference on Future BioMedical Information Engineering (FBIE 2009)*. (December 13-14, 2009).

[87] Lun Huang, Mohammad Al Bataineh, Guillermo E. Atkin, Siyun Wang, and Wei Zhang. "A Novel Gene Detection Method Based on Period-3 Property." *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'09), Minneapolis, Minnesota, USA, 2nd - 6th September*. 1 (2009): 3857 - 60.

[88] Mac Donaill, D. A. "Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition." *Origins of Life and Evolution of the Biosphere*. 33.4-5 (2003): 433-55.

[89] Mac Donaill, D. A. "Why nature chose A, C, G and U/T: an error-coding perspective of nucleotide alphabet composition." *Orig Life Evol Biosph*. 33.4-5 (2003): 433-55.

[90] MacDonaill, D. A. "Digital parity and the composition of the nucleotide alphabet. Shaping the alphabet with error coding." *IEEE Eng Med Biol Mag*. 25.1 (2006): 54-61.

[91] May, E. "Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12." NCSU, December 1998.

[92] May, E. "Analysis of Coding Theory Bases Models for Initiating Protein Translation in Prokaryotic Organisms." North Carolina State University, 2002.

[93] May, E., M. Vouk, D. Bitzer, and D. Rosnick. "Coding model for translation in E.coli K-12." *Proceedings of The First Joint BMES/EMBS Conference.* 2 (1999): 1178.

[94] May, E. E. "Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12." NCSU, December 1998.

[95] May, E. E., M. A. Vouk, and D. L. Bitzer. "Classification of Escherichia coli K-12 ribosome binding sites. An error-control coding model." *IEEE Eng Med Biol Mag.* 25.1 (2006): 90-7.

[96] May, E. E. "Communication theory and molecular biology at the crossroads." *IEEE Eng Med Biol Mag.* 25.1 (2006): 28-9.

[97] May, E. E., M. A. Vouk, D. L. Bitzer, and D. I. Rosnick. "An error-correcting code framework for genetic sequence analysis." *Journal of the Franklin Institute-Engineering and Applied Mathematics.* 341.1-2 (2004): 89-109.

[98] May, E. E., M. A. Vouk, D. L. Bitzer, and D. I. Rosnick. "Coding theory based models for protein translation initiation in prokaryotic organisms." *Biosystems.* 76.1-3 (2004): 249-60.

[99] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo E. Atkin, and Wei Zhang. "An Optimized Ribosome Binding Model Using Communication Theory Concepts." *Proceedings of 2007 International Conference for Bioinformatics and Computational Biology.* (June 25 – 27, 2007): 345-48.

[100] Mohammad Al Bataineh, Lun Hauang, Ismaeel Muhamed, Nick Menhart, and Guillermo E. Atkin. "Gene Expression Analysis using Communications, Coding and Information Theory Based Models." *BIOCOMP'09 - The 2009 International Conference on Bioinformatics & Computational Biology.* (July 13-16, 2009): 181-85.

[101] Mohammad Al Bataineh, Lun Huang, Maria Alonso, Nick Menhart, and Guillermo E. Atkin. "Analysis of Gene Translation Using a Communications Theory Approach." Advances in Computational Biology. *Springer*, December, 2009.

[102] Mohammad Al Bataineh, Maria Alonso, Lun Huang, Guillermo E. Atkin, and Nick Menhart. "Effect of mutations on the detection of translational signals based on a communications theory approach." *Conf Proc IEEE Eng Med Biol Soc.* 1 (2009): 3853-6.

[103] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Wei Zhang, and Guillermo E. Atkin. "Ribosome Binding Model Using a Codebook and Exponential Metric." *2007 IEEE International Conference on Electro/Information Technology*. (17-20 May 2007): 438-42.

[104] Mohammad Al Bataineh, Maria Alonso, Lun Huang, Nick Menhart, and Guillermo E. Atkin. "Effect of Mutations on the Detection of Translational Signals Based on Communications Theory Concepts." *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'09), Minneapolis, Minnesota*. 1 (3 - 6 September, 2009): 3853 - 56.

[105] Nikolov, D. B. and S. K. Burley. "RNA polymerase II transcription initiation: a structural view." *Proc Natl Acad Sci U S A*. 94.1 (1997): 15-22.

[106] Nishi, T., T. Ikemura, and S. Kanaya. "GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences." *Gene*. 346 (2005): 115-25.

[107] Oliver, J. L., P. Bernaolagalvan, J. Guerrerogarcia, and R. Romanroldan. "Entropic Profiles of DNA-Sequences through Chaos-Game-Derived Images." *Journal of Theoretical Biology*. 160.4 (1993): 457-70.

[108] Osada, Y., R. Saito, and M. Tomita. "Analysis of base-pairing potentials between 16S rRNA and 5 ' UTR for translation initiation in various prokaryotes." *Bioinformatics*. 15.7-8 (1999): 578-81.

[109] Ou, H. Y., F. B. Guo, and C. T. Zhang. "GS-finder: a program to find bacterial gene start sites with a self-training method." *International Journal of Biochemistry & Cell Biology*. 36.3 (2004): 535-44.

[110] Palaniappan, K. and M. E. Jernigan. "Pattern analysis of biological sequences." *In Proceedings of the 1984 IEEE International Conference on Systems, Man, and Cybernetics*. (1984).

[111] Pavesi, A., B. Delaco, M. I. Granero, and A. Porati. "On the informational content of overlapping genes in prokaryotic and eukaryotic viruses." *Journal of Molecular Evolution*. 44.6 (1997): 625-31.

[112] Pavesi, A., B. De Iaco, M. I. Granero, and A. Porati. "On the informational content of overlapping genes in prokaryotic and eukaryotic viruses." *J Mol Evol*. 44.6 (1997): 625-31.

[113] Pearson, H. "Genetics: what is a gene?" *Nature*. 441.7092 (2006): 398-401.

[114] Pearson, W. R. "Protein sequence comparison and protein evolution." *In Intelligent Systems for Molecular Biology*. (2001).

[115] Pennisi, E. "Ideas fly at gene-finding jamboree." *Science*. 287.5461 (2000): 2182-+.

[116] Pruitt, K. D., T. Tatusova, and D. R. Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Res*. 35.Database issue (2007): D61-5.

[117] Pruitt, K. D., T. Tatusova, and D. R. Maglott. "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Res*. 33.Database issue (2005): D501-4.

[118] R. Kakumani, V. Devabhaktuni, and M. O. Ahmad. "Prediction of protein-coding regions in DNA sequences using a model-based approach." *ISCAS 2008*. 18.21 (2008): 1918-21.

[119] Raman, R. and G. C. Overton. "Application of hidden Markov modeling to the characterization oftranscription factor binding sites." *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*. 5 (1994): 275-83.

[120] Raman, R. and G. C. Overton. "Application of hidden markov modeling in the characterization of transcription factor binding sites." *In Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences*. 5 (1994): 275-83.

[121] Rambally, G. "A Visualization Approach to Motif Discovery in DNA Sequences." *Praire View A&M University: Pairie View, Texas*. (2007).

[122] Rina Sarkar, A. B. Roy, and P. K. Sarkar. "Topological Information Content of Genetic Molecules - I." *Mathematical Biosciences*. 39 (1978): 299-312.

[123] Roeder, R. G. "The role of general initiation factors in transcription by RNA polymerase II." *Trends Biochem Sci*. 21.9 (1996): 327-35.

[124] Rojdestvenski, I. and M. G. Cottam. "Mapping of statistical physics to information theory with application to biological systems." *J Theor Biol*. 202.1 (2000): 43-54.

[125] Roman-Roldan, R., P. Bernaola-Galvan, and J. L. Oliver. "Application of information theory to DNA sequence analysis: A review." *Pattern Recognition*. 29.7 (1996): 1187-94.

[126] RomanRoldan, R., P. BernaolaGalvan, and J. L. Oliver. "Application of information theory to DNA sequence analysis: A review." *Pattern Recognition*. 29.7 (1996): 1187-94.

[127] Rosen, G. "Finding near-periodic DNA regions using a finite-field framework." *2nd IEEE Workshop Genomic Signal Processing Stat.* (May 2004).

[128] Rosen, G. "Examining coding structure and redundancy in DNA." *IEEE Engineering in Medicine and Biology* 25.1 (2006): 62-68.

[129] Rosen, G. L. "Examining coding structure and redundancy in DNA. How does DNA protect itself from life's uncertainty?" *IEEE Eng Med Biol Mag.* 25.1 (2006): 62-8.

[130] Rosen, G. L. and J. D. Moore. "Investigation of coding structure in DNA." *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP '03).* . 2 (6-10 April 2003): 361-64.

[131] Rosnick, D. "Free Energy Periodicity and Memory Model for Genetic Coding." North Carolina State Univesity, 2001.

[132] Rudd, K. E. "EcoGene: a genome sequence database for Escherichia coli K-12." *Nucleic Acids Res.* 28.1 (2000): 60-4.

[133] Salamon, P. and A. K. Konopka. "A maximum entropy principle for the distribution of local complexity in naturally occuring nucleotide sequences." *Computers and Chemistry.* 16.2 (1992): 117-24.

[134] Salgado, H., S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. "RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions." *Nucleic Acids Res.* 34.Database issue (2006): D394-7.

[135] Sarkar, R., A. B. Roy, and P. K. Sarkar. "Topological information content of genetic molecules." *Math. Biosci.* 39 (1978): 299–312.

[136] Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Res.* 37.Database issue (2009): D5-15.

[137] Sayre, A. *Rosalind Franklin and DNA.* New York: *W.W. Norton and Company,* 1975.

[138] Schneider, T. D. "Consensus sequence Zen." *Appl Bioinformatics*. 1.3 (2002): 111-9.

[139] Schneider, T. D. "Measuring molecular information." *Journal of Theoretical Biology*. 201.1 (1999): 87-92.

[140] Schneider, T. D. "Measuring molecular information." *J Theor Biol*. 201.1 (1999): 87-92.

[141] Schneider, T. D. "Information content of individual genetic sequences." *J Theor Biol*. 189.4 (1997): 427-41.

[142] Schneider, T. D. "Information content of individual genetic sequences." *Journal of Theoretical Biology*. 189.4 (1997): 427-41.

[143] Schneider, T. D. and D. N. Mastronarde. "Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method." *Discrete Applied Mathematics*. 71.1-3 (1996): 259-68.

[144] Schneider, T. D. "Theory of Molecular Machines .2. Energy-Dissipation from Molecular Machines." *Journal of Theoretical Biology*. 148.1 (1991): 125-37.

[145] Schneider, T. D. "Theory of molecular machines. II. Energy dissipation from molecular machines." *J Theor Biol*. 148.1 (1991): 125-37.

[146] Schneider, T. D. "Theory of molecular machines. I. Channel capacity of molecular machines." *J Theor Biol*. 148.1 (1991): 83-123.

[147] Schneider, T. D. and R. M. Stephens. "Sequence logos: a new way to display consensus sequences." *Nucleic Acids Research*. 18.20 (1990): 6097-100.

[148] Schneider, T. D. and R. M. Stephens. "Sequence logos: a new way to display consensus sequences." *Nucleic Acids Res*. 18.20 (1990): 6097-100.

[149] Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht. "Information content of binding sites on nucleotide sequences." *Journal of Molecular Biology*. 188.3 (1986): 415-31.

[150] Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht. "Information content of binding sites on nucleotide sequences." *J Mol Biol*. 188.3 (1986): 415-31.

[151] Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht. "Information-Content of Binding-Sites on Nucleotide-Sequences." *Journal of Molecular Biology*. 188.3 (1986): 415-31.

[152] Shannon, C. E. "The mathematical theory of communication (Reprinted)." *M D Computing*. 14.4 (1997): 306-17.

[153] Shannon, C. E. "A mathematical theory of communication." *Bell Systems Technical Journal*. **27** (1948): 379-423.

[154] Shannon, C. E. "An algebra for theoretical genetics." Massachusetts Institute of Technology, 1940.

[155] Shannon, C. E. "An algebra for theoretical genetics.". Massachusetts Institute of Technology, 1940.

[156] Shine, J. and L. Dalgarno. "The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites." *Proc Natl Acad Sci U S A*. 71.4 (1974): 1342-6.

[157] Shultzaberger, R. K., Z. Chen, K. A. Lewis, and T. D. Schneider. "Anatomy of Escherichia coli sigma70 promoters." *Nucleic Acids Res*. 35.3 (2007): 771-88.

[158] Silverman, B. D. and R. Linsker. "A measure of DNA periodicity." *J Theor Biol*. 118.3 (1986): 295-300.

[159] Stambuk, N. "Symbolic Cantor Algorithm (SCA): A method for analysis of gene and protein coding " *Periodicum Biologorum*. 101.4 (1999): 355-61.

[160] Stambuk, N. "Symbolic Cantor Algorithm (SCA): A method for analysis of gene and protein coding." *Periodicum Biologorum*. 101.4 (1999): 355-61.

[161] Stambuk, N. "On circular coding properties of gene and protein sequences." *Croatica Chemica ACTA*. 72.4 (1999): 999-1008.

[162] Stambuk, N. "On the genetic origin of complementary protein coding." *Croatica Chemica ACTA*. 71.3 (1998): 573-89.

[163] Steitz, J. A. and K. Jakes. "How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli." *Proc Natl Acad Sci U S A*. 72.12 (1975): 4734-8.

[164] Strait, B. J. and T. G. Dewey. "The Shannon Information entropy of protein sequences." *Biophysical Journal*. 71.1 (1996): 148-55.

[165] Sustar, P. "Crick's notion of genetic information and the 'central dogma' of molecular biology." *British Journal for the Philosophy of Science*. 58.1 (2007): 13-24.

[166] Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg. "A probabilistic, method for identifying start codons in bacterial genomes." *Bioinformatics*. 17.12 (2001): 1123-30.

[167] Sweeney, P. *Error Control Coding: An Introduction. Prentice Hall*, 1991

[168] Szathmary, E. "The origin of the genetic code: amino acids as cofactors in an RNA world." *Trends Genet.* 16.1 (2000): 17–19.

[169] Taft, R. J., M. Pheasant, and J. S. Mattick. "The relationship between non-protein-coding DNA and eukaryotic complexity." *Bioessays*. 29.3 (2007): 288-99.

[170] Tompa, M. "An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem." *Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology.* (1999): 262–71.

[171] Uberbacher, E. C. and R. J. Mural. "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach." *Proc Natl Acad Sci U S A*. 88.24 (1991): 11261-5.

[172] Uberbacher, E. C. and R. J. Mural. "Locating Protein-Coding Regions in Human DNA-Sequences by a Multiple Sensor Neural Network Approach." *Proceedings of the National Academy of Sciences of the United States of America.* 88.24 (1991): 11261-65.

[173] van Nimwegen, E. "Scaling laws in the functional content of genomes." *Trends Genet.* 19.9 (2003): 479-84.

[174] Veljkovic, V., I. Cosic, B. Dimitrijevic, and D. Lalovic. "Is It Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal-Processing." *Ieee Transactions on Biomedical Engineering*. 32.5 (1985): 337-41.

[175] Walker, M., V. Pavlovic, and S. Kasif. "A comparative genomic method for computational identification of prokaryotic translation initiation sites." *Nucleic Acids Research*. 30.14 (2002): 3181-91.

[176] Watson, J., N. Hopkins, J. Roberts, J. Steitz, and A. Weiner. *Molecular Biology of the Gene*. Menlo Park, CA.: *The Benjamin Cummings Publishing Company, Inc.*, 1987.

[177] Weindl, J. and J. Hagenauer. "Applying Techniques from Frame Synchronization for Biological Sequence Analysis." *IEEE International Conference on Communications (ICC).* (2007): 833 - 38

[178] Weindl, J. and J. Hagenauer. "Applying Techniques from Frame Synchronization for Biological Sequence Analysis." *IEEE International Conference on Communications (ICC)*. (2007): 833-38.

[179] Yada, T., Y. Totoki, T. Takagi, and K. Naka. "A novel bacterial gene-finding system with improved accuracy in locating start codons." *DNA Research*. 8.3 (2001): 97-106.

[180] Yockey, H. *Information Theory and Molecular Biology*. New York, NY: *Cambridge University Press*, 1992.

[181] Yockey, H. P. "Information in Bits and Bytes - Reply to Lifsons Review of Information-Theory and Molecular-Biology." *Bioessays*. 17.1 (1995): 85-88.

[182] Yoxen, E. "The Double Helix - a Personal Account of the Discovery of the Structure of DNA - Watson,Jd." *British Journal for the History of Science*. 16.54 (1983): 278-81.

[183] Z. Dawy, B. Goebel, and J. Hagenauer. "Gene mapping and marker clustering using Shannon's mutual information." *IEEE Transactions on Computational Biology and Bioinformatics*. 3.1 (January-March 2006): 47–56.

[184] Z. Dawy, F. G., J. Hagenauer, and J. C. Mueller. "Modeling and analysis of gene expression mechanisms: a communication theory approach." *IEEE International Conference on Communications (ICC)*. 2 (2005): 815- 19.

[185] Zien, A., G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, and K. R. Muller. "Engineering support vector machine kernels that recognize translation initiation sites." *Bioinformatics*. 16.9 (2000): 799-807.