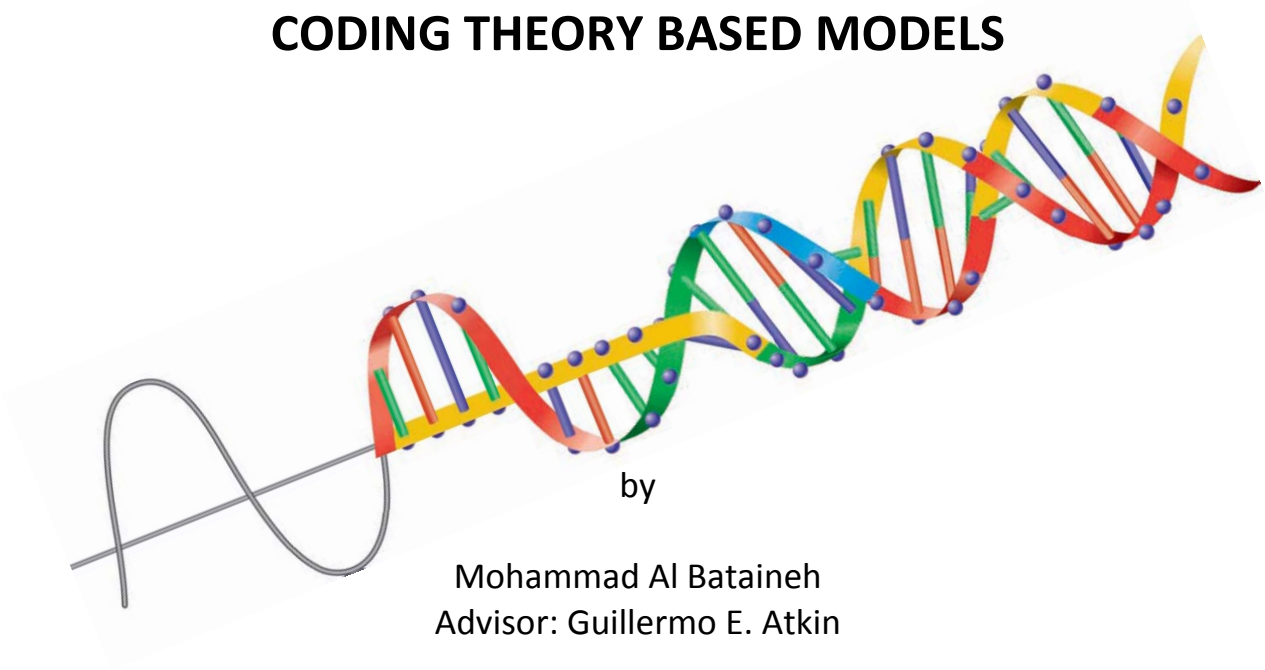
ILLINOIS INSTITUTE OF TECHNOLOGY
Transforming lives.Inventing the Future.www.iit.edu

GENE AND REGULATORY SEQUENCE IDENTIFICATION USING COMMUNICATIONS, INFORMATION AND CODING THEORY BASED MODELS



by

Mohammad Al Bataineh
Advisor: Guillermo E. Atkin

A PhD proposal submitted to the Graduate Faculty of
Illinois Institute of Technology
In partial fulfilment of the
requirements for the Degree of
Doctor of Philosophy

Department of
Electrical and Computer Engineering

Chicago, IL
December 4, 2008

CONTENTS

1 Introduction

Identification and annotation of all the functional elements in the genome, including genes and regulatory sequences, is 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 modeling and detection of DNA regulatory elements. My PhD research proposes a novel use of techniques and principles from communications engineering, coding and information theory for modeling, identification and analysis of genomic regulatory elements and other biological sequences of a particular significance in genomics and computational biology. Such techniques include the ones used in source and channel coding, frame synchronization, pattern recognition, wavelet analysis, and discrete Fourier Transform. This research will be initially applied to the complete genomes of prokaryotic organisms and later will be extended to eukaryotic organisms.

An emerging paradigm in biology in the postgenomic era is the emergence of a "systems biology" approach to understanding life. This is in contrast to the reductionist approach in vogue leading up to the genomic era which was used with great success in producing the very detailed but very low level knowledge that we have about many biological processes. The reductionist approach produced such detailed information such as:

- Detailed knowledge of metabolic (50's and 60's) and signaling pathways (70's – 00's).
- Atomic level structures of many proteins and other biological macromolecules (60's – 00's).
- Detailed knowledge of gene structure, culminating in genome level knowledge (90's – 00's).

However, we are still not exactly sure what life is or how it works. The simplest indication of this is that all – or at least a very large preponderance - of these characteristics (pathways and metabolites; structure of macromolecules, gene structure) are identical in a living person compared to recently deceased person – yet something crucial is clearly different.

One conception of what this might be is that the key to living processes is not in the substance of living organisms, but in the system of interactions of these substances. Reading a genome does not inform us very much at all about the system by which genes encoded within interact in a complex system to produce the various biochemical machines that make up the cell – or how this system is modified in different cell types, or under different environmental conditions. We seek to provide tools and methods to detect and understand the punctuation of the genetic code: the regulatory sequences, such as promoters, terminators, transcriptional and translational regulation signals such as repressors or inducer binding sites. This list can be expanded in eukaryotic genomes to include exon spicing signals, enhancers, and noncoding RNA genes.

Communications and information theory has proved to provide powerful tools for the analysis of these signals [1]–[7]. An up-to-date summary of ongoing research can be found in [8]. 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 $\bar{X} = \{A, C, G, T\}$. An important field of interest is gene expression, the process during which this information stored in the DNA is transformed into cell functions like oxygen transport etc., largely by coding for the expression of specific proteins that carry out and regulate these processes. Protein gene expression takes place in two steps: transcription and translation (see Figure 1).

Figure 1: The process of protein synthesis (gene expression)

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.

This work contributes to the field of bioengineering and biology through the use of information theory, communications theory and coding theory principles. Initially, our research will study and analyze transcription and translation initiation mechanisms in prokaryotes (e.g. E. coli, and well as other bacteria), and then will be extended to study other types of organisms (eukaryotes such as us).

The main goals of my PhD research are to:

i) develop analogies between information transmission in communications engineering and gene expression. Find models for prokaryotic and eukaryotic organisms that represent the genetic and molecular mechanisms that different organisms use to regulate their genome expression,

ii) validate these biologically-motivated coding models for the processes of transcription and translation, and use these models to gain new insights on the biological interactions between the RNA Polymerase and DNA, and ribosome and mRNA,

iii) develop models considering mutations in regulatory sequences and the genomic structure (coding and noncoding regions) and study their effects on protein synthesis. The models developed can be associated to communication channels with noise. Entropies of the source (DNA) and the output (modified gene and ultimately protein synthesis) and mutual information between them can be used to develop informatics models for these processes,

iv) initially analyze gene structure using a variable-length codes (VLC) approach and iterative decoding algorithm to detect genes and regulatory sequences. This approach will have to be adjusted for organisms that do not exhibit the prefix condition. This will lead to a better understanding of the structure and correlations between coding and non-coding regions of the whole genome. Mutation will produce path deviation that can be quantified using proper measures of distortion that must be defined,

v) introduce an improved gene and regulatory sequences identification approach that will provide a solution for current limitations that exist in gene-finding programs by using pattern recognition [9], Discrete Fourier Transform (DFT) [10], and Wavelet analysis [11],

vi) develop new computational algorithms and databases for systematic identification of transcriptional regulators and regulons in new genomes as they become available; and integrate genome expression data with known and predicted regulons and metabolic pathways. These algorithms will improve the detection of the effect of mutations in organisms. The proposed models will be also used in future work to test the effect of mutations in the ribosome on protein synthesis, and predict the effect of other possible mutations,

vii) use principles of error control coding theory to interpret the genetic translation and transcriptions mechanisms with and without mutations,

viii) apply and extend the proposed models to prokaryotic and eukaryotic organisms to uncover the genetic and molecular mechanisms that different organisms use to regulate their genome expression in response to the stimuli and stresses.

This research will allow for the analysis of various interactions that take place in gene expression using communications models that will allow savings in laboratory resources and time-consuming laboratory experimentations. Moreover, it will lead to better understanding of these complex processes.

The research focuses initially on developing Communications and Information Theory models for the process of translation in Gene Expression using different E. coli bacteria strains. Such models can be tested to analyze many biological problems related to Gene Expression (like mutations), and hence save time and cost of laboratory experimentation. These models will be extended to other organisms.

Future work will focus on developing new models to analyze gene and regulatory sequence identification. These models will be based on communications, information theory and coding theory principles. Analysis will be initially applied to prokaryotic genomes and later will be extended to study eukaryotic genomes. Mutations affecting the level of protein synthesis in the organisms under study will be tested and analyzed using these developed models.

Future work will also focus on applying principles of coding theory to map between gene and protein sequences (i.e. mapping of codons to amino acids). The well-known genetic code has 64 codons that uniquely map to 20 amino acids which is a redundant mapping. This redundancy suggests that an embedded structure may exist (code). According to coding theory principles, the genetic code can be viewed as a quaternary alphabet (A, U, C and G). Analogies with variable length codes theory, source and channel coding, pattern recognition can be utilized to establish a reasonable representation of the genetic code. Entropy and distance metrics between different codons and different amino acids can be defined as well.

Another suggested direction of research is to study the level of Gene Expression in E. coli under different kinds of stress (like temperature and Chlorine concentration). Markov models and other mathematical approximations can be used to analyze such a study. Laboratory data required for the study are available and ready for analysis. Such analysis will save time and cost of laboratory experimentation and will allow gaining new insights on the biological interactions related to bacterial growth under such kinds of stress.

2 Background and Significance

Regulatory elements are often short and variable, their identification and discovery using computational algorithms is difficult. However, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements. The availability of complete genome sequence from multiple organisms, as well as mRNA profiling and high-throughput experimental methods for mapping protein-binding sites in DNA, have contributed to developing methods that use these auxiliary data to inform the detection of transcriptional regulatory elements. Progress is also being made in identifying cis-regulatory modules and higher order structures of the regulatory sequences, which is essential to understand transcription regulation in the metazoan genomes.

Here I briefly describe the process of gene expression (transcription and translation) and some of the regulatory sequences that we will use in our research.

2.1 Gene Expression

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 2).

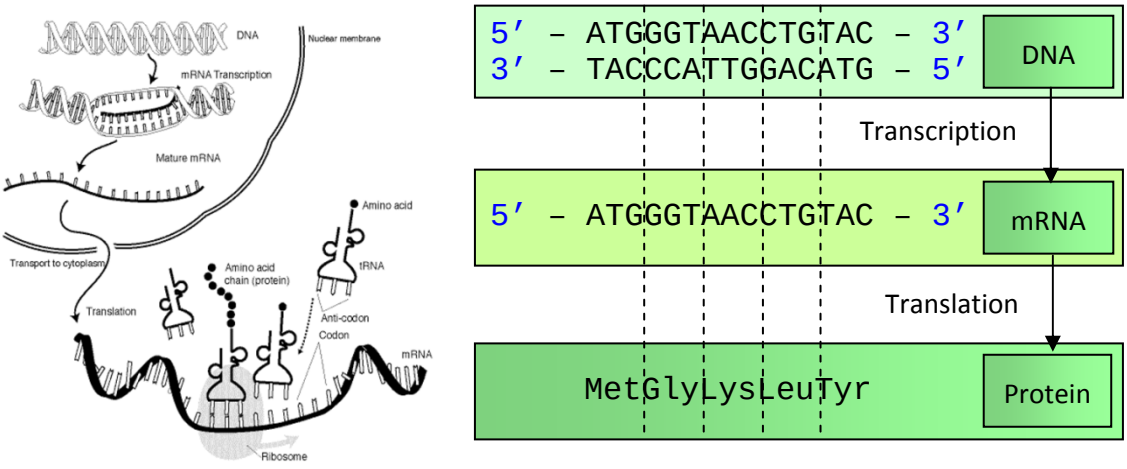


Figure 2: Protein Synthesis (Gene Expression)

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 (see

Figure 2) into a chain of amino acids that forms a protein according to the mapping rule described by the genetic code [12]. In prokaryotes, the RNA is essentially competent to do this immediately; however in eukaryotes, there is an intermediate step in which the massage 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.

Regulatory process operates at each step. In the transcriptional step, individual messages need to be identified, often only under specified circumstances, and sent (RNA synthesized). This process involves signals termed promoters, which initiate this process. There are many types of promoters and one of the most common studied types in *E. coli* is illustrated in Figure 3.

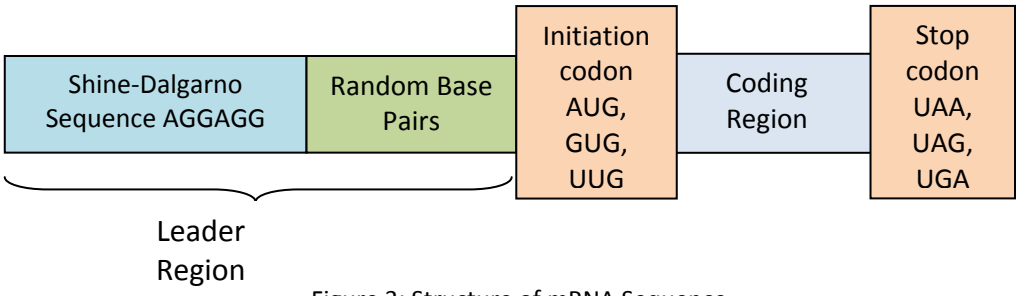


Figure 3: Structure of mRNA Sequence

2.2 Regulatory Sequences

A regulatory sequence (also called a regulatory region or a regulatory element, RE) is a segment of DNA or RNA which exerts some control over the process whereby information in the sequence is communicated or utilized. The usual fashion by which these REs act is by binding some regulatory proteins, which then affects some cellular process involving this information. For instance,

- transcription factors bind to promoters and recruit RNA polymerase to be available to transcribe the information downstream of the promoter, and so cause the information in the gene to be moved from the genome to mRNA.
- the ribosome binds to ribosome binding sites (Shine Dalgarno sites in bacteria) and help initiate transcription, which processes this information into a different form, from RNA to protein, in a process called translation.

In my preliminary work, I will detect regulatory sequences (e.g. promoters, enhancers, silencers, locus-control regions, Shine-Dalgarno, etc) that are involved in the process of gene expression (transcription and translation). Preliminary results shows that in prokaryotes the detection of these sequences can be helped using initially the algorithms described in sec 4.1 and a variable length code (VLC) model approach and iterative decoding algorithm (section 4.2). In the case of eukaryotes we will develop similar algorithms that will allow gaining knowledge in the gene structure and identifying regulatory sequences.

2.3 Biological Significance

To a very good approximation, every cell of a given species has the same DNA – yet they can appear and function very differently. This is most obvious in multicellular organisms, such as higher eukaryotes, in which different tissue types comprise the body. These cell types typically have their own subset of genes expressed, and their own subset of regulatory signals. Even in unicellular organisms, such as bacteria, cells can exist in various states, depending on environmental cues. This is often mediated through changes in the metabolism which are controlled by complex regulatory mechanisms. Functional characterization of individual transcriptional regulators at nucleic acid sequence levels is a first step to elucidate such regulatory mechanisms that coordinate the activity of different metabolic and signaling pathways.

To uncover the global transcriptional regulatory architecture of metabolic networks we propose to develop new computational tools that will integrate microarray expression data from this study with known or predicted regulatory elements in fully sequenced genomes. Initially we will target E. coli as a simple prokaryotic model organism, but will expand this to other bacteria and eukaryotes. An outline of our computational approach is shown in Figure 4.

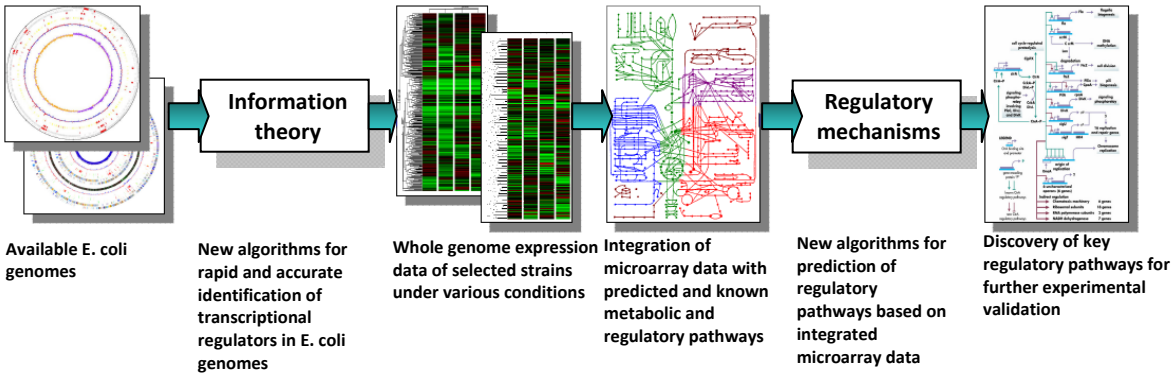


Figure 4: An outline of our computational approach

Detection of transcriptional units and their promoter sites is one of the keys to understanding the regulon structure of bacterial genomes. Predicting regulons, in turn, gives us strong hints about gene function. Computational detection of promoter and terminator sequences is the only practical means of systematically identifying large numbers of regulons today, and few experimentally verified regulons exist outside of *B. subtilis* and *E. coli*. Eukaryotic transcription factor sites are much more variable, and less well understood. The criteria by which Transcription Factors (TFs) recognize these signals is not entirely clear; so that an exact description of these signals in not possible. Rather, consensus binding sequences based upon known example binding sequences have been built up. There are two ways in which this confounds a simple identification of new such TF binding sites:

- The redundancy of the recognition sequence means that the signal is not one specific code, but rather a subset of codes

- Our knowledge of the requirements of this code is only approximate. It is largely build up by consensus analysis of a known subset of codes for each TF. These are typically some of the strongest activating codes, but some of the other weaker codes, or other cryptic codes, are exactly what we are looking to detect.

Several previous computational methods (Carafa *et al.* 1990; de Hoon *et al.* 2005) have relied on simple decision boundaries to separate promoters from non-promoters after training on experimentally known terminating and non-terminating sequences. Other studies have considered only the DNA binding portion of potential promoters (Washio *et al.* 1998; Unniraman *et al.* 2002). Due to lack of sequence data, previous systems (e.g. Carafa *et al.* 1990; Lesnik *et al.* 2001) have tended to focus on *E. coli* or on only a portion of the now-available genomes. In this study, we will develop a computational system for rapid and accurate predictions of transcriptional regulators in any genomic data, starting with E. coli and extending our results to eukaryotes.

The algorithm will search genomic DNA for specific regulatory signals and assign each candidate a score related to the likelihood that it arose by chance. We will utilize existing data bases of regulatory protein binding sites as well as compiling new information as it becomes available, and then use our detection algorithms to search entire genomes of these regulatory sequences. The relative organization of these signals will then be used to detect specific putative genes, as well as the conditions under which these genes would be expressed. Examples of this organization include heuristic rules such as:

- promoter sequences occur 5' to genes.
- the message transcribed by these genes should be sensible:
  - if it is a protein coding gene, is should contain other signals for ribosome binding and translation initiation, and an open reading frame.
  - in eukaryotes, other signals for RNA processing should be present, including exon splicing signals.
  - if it is a noncoding gene, appropriate RNA structure and sequence should be present
- in bacteria, appropriate terminators should be present at the 3' end.

As has been done with TransTermHP (Kingsford *et al.* 2007), we will assess the sensitivity and specificity of our predictions using a set of experimentally verified regulons (both from the literature and from this study). The algorithm will be based on sequence characteristics of all known bacterial transcriptional regulator families. The new system will be easily portable, user-friendly, and will be released as free, open-source software. The speed of our search algorithm facilitates interactive experimentation and refinement and allows us to add more genomes easily; it also includes (1) a more accurate scoring scheme; (2) more informative output; (3) the ability to handle overlapping genes; (4) better handling of gaps in hairpin structures; (5) the ability to handle gene annotations as either a simple list or in NCBI's ptt format.

Initially we will develop these tools in prokaryotic systems, using E. coli as a test organism to validate the system. This will involve the following major components:

- Identification of consensus sequences for promoters i.e. transcriptional start sites

- Identification of translational signals such as Shine-Dalgarno and S1 protein ribosome binding sites; as well as terminators.
- Identification of noncoding RNA (ncRNA) genes

We will then expand this to eukaryotic organisms, namely humans. This is a substantially more complex task for several reasons:

- Eukaryotic regulatory elements, especially promoters, are much more complex and heterogeneous, composed of several independent parts as well as unique elements specific for only one or a few genes. In this case homology modeling using known promoters from related species can be a useful tool.
- Eukaryotic RNA processing is a complex, and as yet incompletely understood process, which requires detection of both processing (e.g. poly adenylation) signals as well as exon splicing signals (5'- and 3' splice sites; branch point sites; as well as exon splicing enhancers and silences ESE and ESS).

3 Preliminary Studies

The following section portrays our preliminary research work, models, algorithms and techniques that we used to model and analyze the process of translation in gene expression. Current and future research direction are presented and described as well.

3.1 Coding Theory, Communications and Information Theory Based Modeling

Sec 3.1.1 describes my preliminary work to model the process of translation in gene expression. A variable length codebook, an energy table, and a specially designed metric were used to analyze the mechanism that the ribosome uses to decode the mRNA sequence. The codebook and metric are common elements used in the detection process of communication systems. 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. Sec 3.1.2 highlights the five new models developed to analyze gene and regulatory sequence identifications. Preliminary results are shown. These models are described in details in sec 4.1.2.

3.1.1 Coding Theory Based Models

Preliminary work of the study has been to validate the work done by "[1]Z. Dawy[1, 3], F. Gonzalez[1], Joachim Hagenauer[1] and Jacob Mueller[2]," in their paper "Modeling and Analysis of Gene expression Mechanism: A communication Theory Approach" [6]. This work deals with modeling gene expression (information contained in the DNA molecule when 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.

[1] Munich University of Technology, Germany[1]; National Research Center for Environment and Health, Germany[2] and American University of Beirut, Lebanon[3]

Gene expression involves two main stages (See Figure 2). The first one is transcription (related to coding theory) where the information stored in the DNA is transformed into the messenger RNA (mRNA). The second one is translation (related to detection theory), where the mRNA molecule serves as an instructive for protein synthesis. Analyzing gene expression, many similarities with the way engineers send 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. 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 prompted by the increased availability of genetic data. In our preliminary work we study a communication theory based model for translation in genome expression. The model uses the assumption that the ribosome decodes the mRNA sequences using the 3' end of the 16SrRNA molecule as a one-dimensional embedded codebook. The biological consistency of the model is proven in detecting the Shine-Dalgarno sequence and the initiation and stop codons for translation initiation and termination. Results obtained using these models have 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.

In the proposed model, 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, Figure 5. During transcription the DNA sequence is decoded to produce the mRNA sequence. This sequence is thought to be a decoder output because the mRNA sequence is shorter than the 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 take the mRNA sequence to start the protein synthesis. The protein output of our model is the final recovered message.

This initial work focuses on modeling translation, specifically in the E. coli bacteria. During translation 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. Figure 5 shows a general model of gene expression from a communication theory point of view.

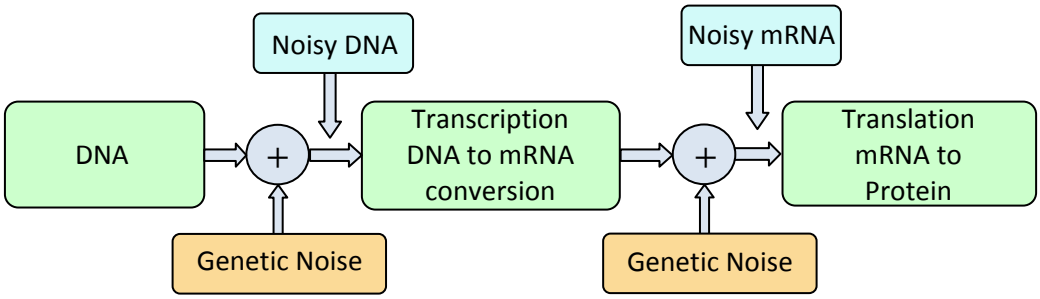


Figure 5: Transcription and Translation as a Communication Model

Figure 6 shows a typical mRNA sequence [6]. 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.

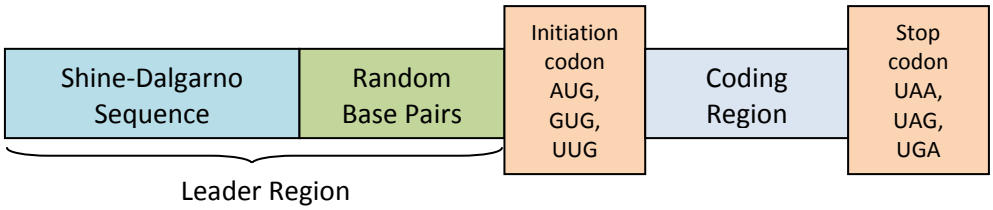


Figure 6: mRNA Sequence

The process of translation in prokaryotes is triggered by the detection of an RE known as the Shine-Dalgarno (SD) sequence. Physically, this detection operates by homology mediated binding of the RE to the last 13 bases of the 16S rRNA in the ribosome [13]. In our work [1] and [2], we have modeled this detection/recognition system by designing a one dimensional variable-length codebook and a metric. The codebook uses a variable codeword length N between 2 and 13 using the Watson-Crick complement of the last 13 bases of the 16S rRNA molecule, i.e. we obtain (13-N+1) codewords; $\overline{c}_i = [s_1, s_2, ..., s_{i+N-1}]$; $i \in [1, 13-N+1]$ where $\overline{s} = [s_1, s_2, ..., s_{13}]$ denotes the complemented sequence of the last 13 bases (i.e. $\overline{s} = [\text{UAAGGAGGUGAUC}]$).

The input to our proposed model is the noisy mRNA and the last 13 bases of the molecule 16S rRNA (in the ribosome) interact with the leader region of the mRNA to start translation. The mRNA is a noisy version of the mRNA produced in the transcription process due to the addition of genetic noise. The codebook used has variable length N between 2 and 13. Then 13-N+1 codewords are produced by taking a sliding window through the Watson-Crick complement of the sequence of 13 bases (shift one base at a time), [6]. This sequence (UAAGGAGGUGAUC) and the resulting codebook for a value N = 5 are shown in Figure 7 and Table 1 (notice the SD sequence is AGGAGG):

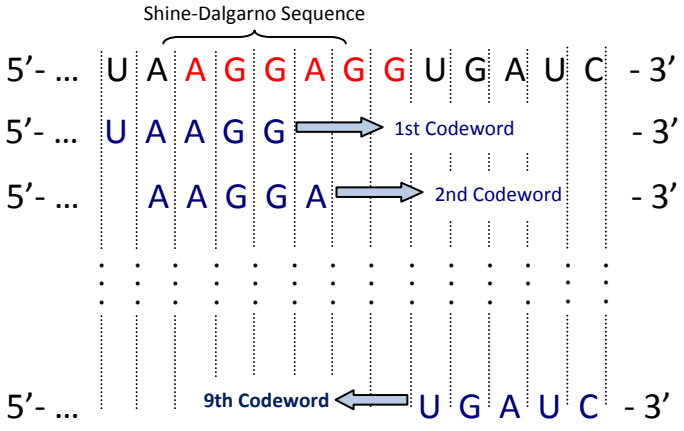


Figure 7: Codebook Structure Length N = 5

Cl	Codeword
C1	UAAGG
C2	AAGGA
C3	AGGAG
C4	GGAGG
C5	GAGGU
C6	AGGUG
C7	GGUGA
C8	GUGAU
C9	UGAUC

Table I: Codebook length N = 5

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

Table II: Energy Table

A moving window of size N is applied 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) in kcal/mol is selected as the correct codeword (Table 2). Biologically, the ribosome achieves this by means of the complementary principle. The energetics 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. The minimum energies are evaluated and plotted to determine the performance of the proposed algorithm.

To test the model, and obtain the results shown in Figure 9, 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) position for each gene were obtained and saved
4. The consensus Shine-Dalgarno signal (“AGGAGG” or “AGGA” or “GGAG” or “GAGG”) was located in the noncoding 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 obeyed the following structure shown in Figure 8.

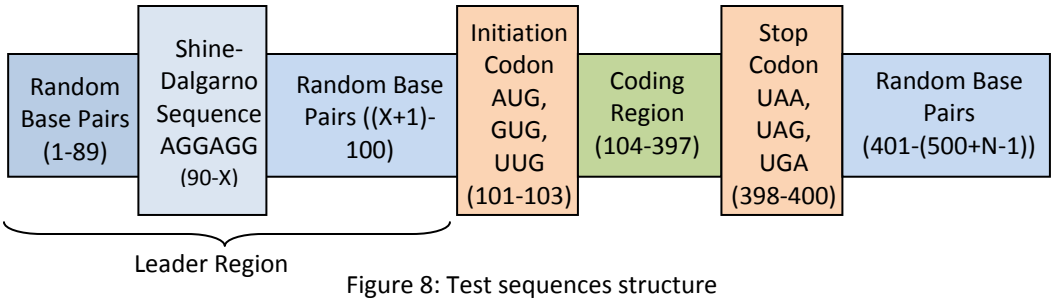


Figure 8: Test sequences structure

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.

The results obtained matched the ones obtained in the previous research. We considered also mutations and the results also matched previous results. These are consistent with published experimental results. This shows the relevance of the model, its biological accuracy, and its flexibility to incorporate and study structural changes. Also, the proposed algorithm allows testing various combinations of mutations 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.

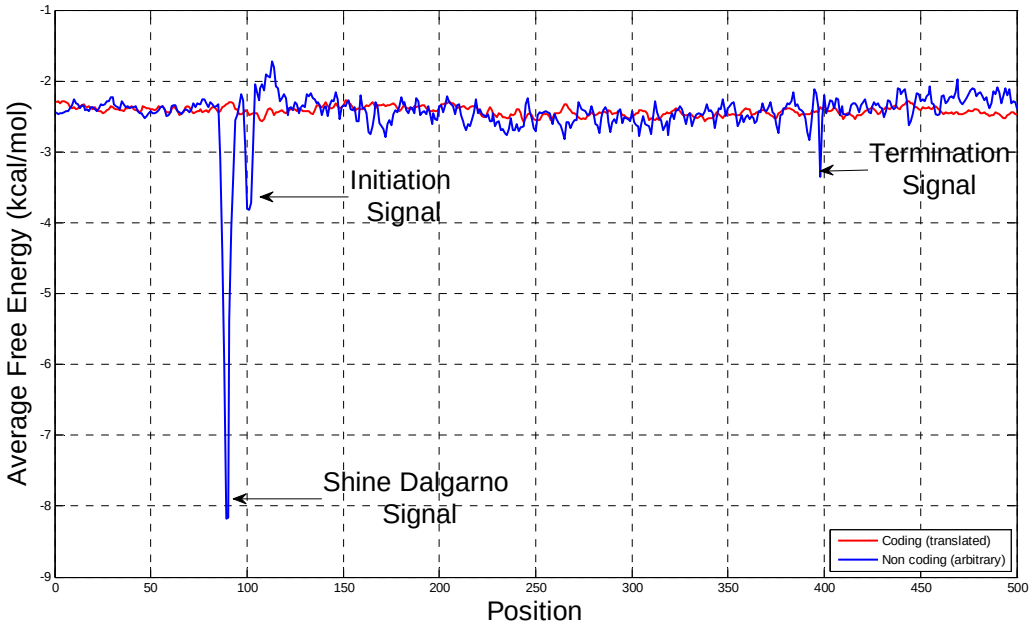


Figure 9: Detected translation Signals

3.1.1.1 Detection Algorithm Optimization

In our model, a modified version of the method of free energy doublets presented in [14] is adopted to calculate an energy function (see equation 1) that represents a free energy distance metric in kcal/mol instead of minimum distance (see Tables 2) [6]. Our algorithm assigns weights to the doublets such that the total energy of the codeword is increased with a match and decreased if a mismatch occurs, and stresses or de-emphasizes the value when consecutive matches or mismatches occur. The energy function has the following form:

$$E = \sum_{k=1}^{N} w_k \, \delta_k \tag{1}$$

where δ_k means a match ($\delta_k = 1$) or a mismatch ($\delta_k = 0$) and w_k is the weight applied to the doublet in the k^{ith} position. The weights are given by:

$$w_k = \begin{cases} \rho + a^{\sigma} & \text{if } \delta_k = 1 \\ \max\left\{ w_{k-1} - \left(a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}} \right), 0 \right\} & \text{if } \delta_k = 0 \end{cases} \tag{2}$$

where σ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches and ρ is an offset variable updated as follows:

$$\rho = \begin{cases} \rho & \text{if } \delta_k = 1 \\ 0 & \text{if } \delta_k = 0 \;\&\; \rho \le a \\ \max\left\{ w_{k-1} - \left(a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}} \right), 0 \right\} & \text{otherwise} \end{cases} \tag{3}$$

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

Detection of translation signals (Shine-Dalgarno, Initiation and termination signals) has been optimized using this exponentially weighted free energy decoding algorithm. This algorithm was used to improve the resolution and flexibility of translation signals detection. According to this proposed algorithm, free energy introduced in [6] can be modified to follow this form:

Algorithm I: Exponentially-Weighted Free Energy Ribosome Decoding (EWFERD)

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 nth symbol of codeword k, s_n is the nth symbol of S, E_k is the exponentially weighted free energy metric when codeword *k* is used (E_k is initialized to *0, 0 ≤ k ≤ L*), and *Energy(a,b)* is the energy dissipated on binding with the nucleotide doublets *ab* (see Table II, e.g. the energy dissipated by binding with AC is −1.8 kcal/mol). 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.

This algorithm allows detection of the exact position of the Shine-Dalgarno sequence on the genes rather than using an average. 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 Shine-Dalgarno in the genes under study.

EWFERD Algorithm

for $k = 1 \ldots L$ *do*
  Initialize $\sigma = 0, \tilde{\sigma} = 0, \rho = 0, w_1 = a;$
  for $n = 1 \ldots N-1$ *do*
    if $c_n^k c_{n+1}^k$ *and* $s_n s_{n+1}$ *are matching then*
      Increment $\sigma = \sigma + 1;$
      set $\tilde{\sigma} = 0;$
      $w_n = \rho + a^{\sigma};$
    else
      Increment $\tilde{\sigma} = \tilde{\sigma} + 1;$
      set $\sigma = 0;$
      $v = w_1 - (a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}});$
      if $n >= 2$
        $v = w_{n-1} - (a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}});$
      end
      $w_n = \max(0, v);$
      if $\rho \le a$
        $\rho = 0;$
      else
        $\rho = \max\left\{w_{n-1} - \left(a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}}\right)\right\};$
      end of if
    end of if
    $E_k = E_k + w_n . Energy(c_n^k c_{n+1}^k);$

3.1.1.2 Analysis and Results

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. 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. Figure 10 shows average results for the detection of the SD, start and stop codons being compared to previous work [6]. 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. Different mutations were tested using our algorithm and the results obtained further certified the correctness and the biological relevance of our model.

To detect the Shine Dalgarno signal in a single gene, the proposed weighting algorithm was applied and compared to the algorithm used in [6]. Figure 11 illustrate that if the parameter a is further increased the resolution of the peak corresponding to the SD sequence will be larger. It can be observed that our algorithm performs much better than the codebook alone. Not only does the proposed 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.

This shows that our proposed weighted algorithm results in a better resolution of the SD sequence. It allows detecting this sequence in the genes without the need of averaging over a large set of them. The algorithm is sensitive to the parameter a, and by properly choosing this value the accuracy of the previous work can be improved. This also shows the relevance of the model, its biological accuracy, and its flexibility to incorporate and study structural changes. Also, the proposed algorithm allows testing various combinations of mutations reducing the need for time and cost consuming laboratory experimentation.

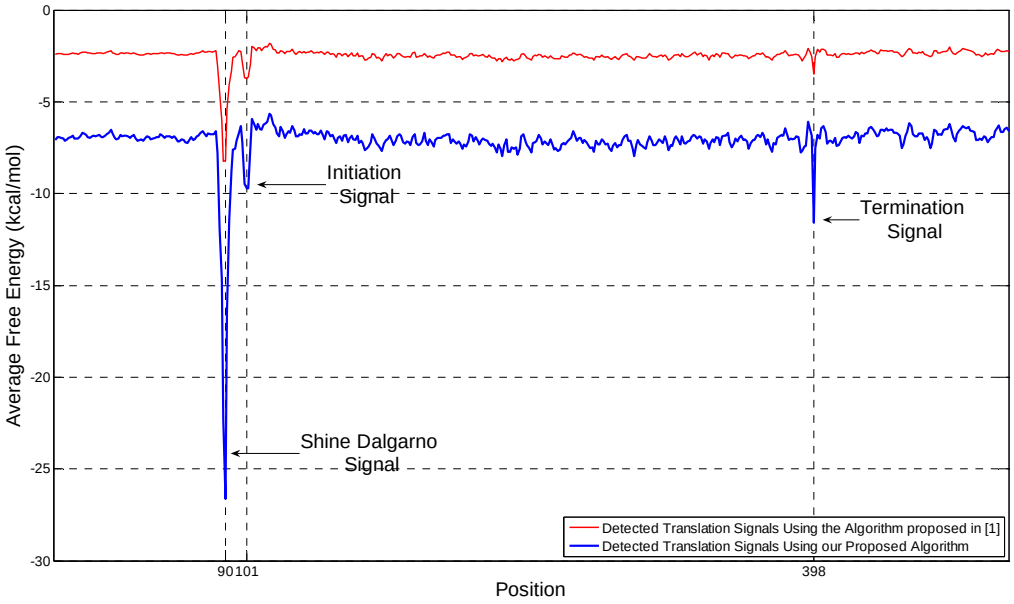


Figure 10: Comparison of SD signal (position 90), start (position 101) and termination (position 398) codon between the algorithm used in [6] and the weighted algorithm (N=5, a=1.5)

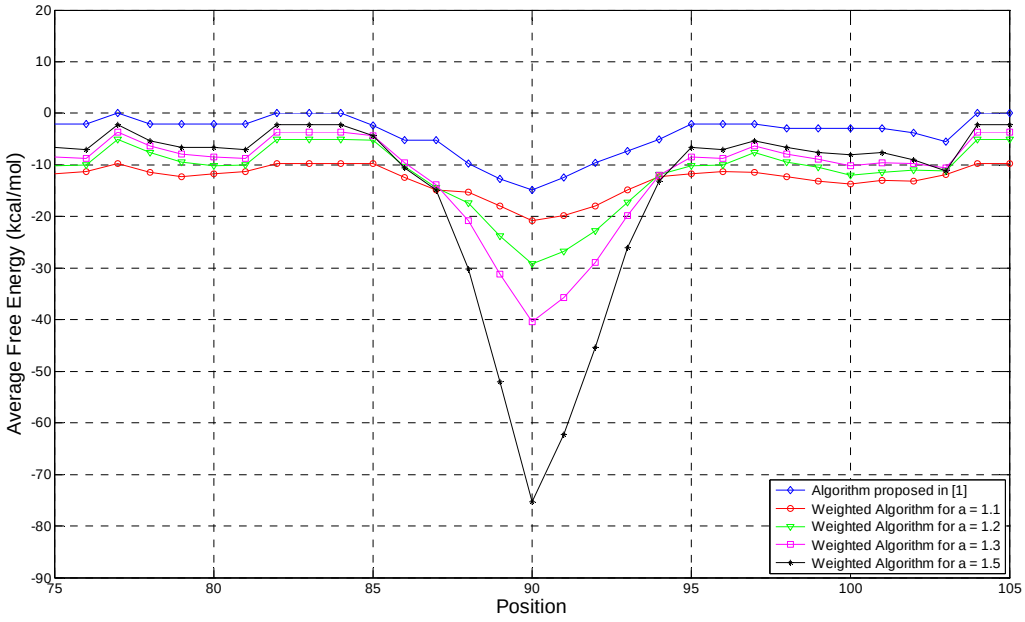


Figure 11: Detected SD signal using algorithm in [6] compared to our weighting algorithm with different values of a

3.1.2 Communications and Information Theory Based Models

The previous models discussed in sec 3.1.1 are based on coding theory (codebook). We have also developed other five different models (sec 4.1.2) for detection of regulatory sequences. These models are based on basic concepts in communications and information theory as correlation (model I), Euclidean distance (model II), matched filter (model III), correlation based exponential metric (model IV) and free energy doublets (model V). Applying these methods to detect the last 13 bases of the 16S rRNA molecule in the given mRNA sequence allows not only detecting the translational signals at their exact corresponding locations but also distinguishing coding from noncoding regions. This new finding suggests that the last 13 bases of 16S rRNA molecule has a higher correlation with coding regions. Preliminary results for method I (sec 4.1.2.) are shown in Figure 12 from which coding and noncoding regions can be apparently identified. This interesting result will be further analyzed and investigated using other binding sequences.

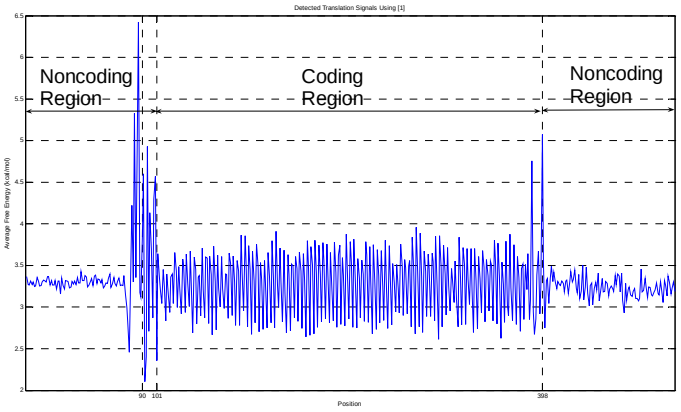


Figure 12: Method I result

3.2 Mutation Analysis

In our preliminary work [1] [2] based on the codebook model, we have applied our proposed algorithm 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 of our 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 the 5th position of the 16SrRNA [16]. Specifically, the 5th position in the arrangement illustrated below:

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
Base	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 consisted in a change of the nucleotide C → U in the ribosome small subunit. This is equivalent to make a mutation from G → A in the complement sequence shown above. The result of this mutation was a reduction in the level of protein synthesis. Another published record of the behavior of the protein synthesis under mutations in the 3' end of the 16SrRNA, was done by Hui and De Boer [17]. In this experiment, the mutations were done in

positions 4 to 8 (GGAGG → CCUCC) and positions 5 to 7 (GAG → UGU). The results of both mutations were lethal for the organism in the sense that the production of proteins stopped.

These published mutations are tested using our model. First the mutations as specified in [16] and [17] are performed in the 13 bases. For each case, 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 13 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. However, Jacob mutation does not affect the detection of the termination signal. This means that protein synthesis process is normally terminated.

After introducing the mutations as in [17], the results showed a complete loss of the SD signal. Hence, it can be inferred that the translation will never take place. This is illustrated in Figure 14. Note that results obtained by mutations in the 16S rRNA also apply to scenarios with mutations in the mRNA at corresponding positions.

It is noted in Figure 14 that the SD peak (position 90) almost complete disappear due to the mutations. The large peak corresponds where no mutations are present. The same mutations are tested using our model 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 15.

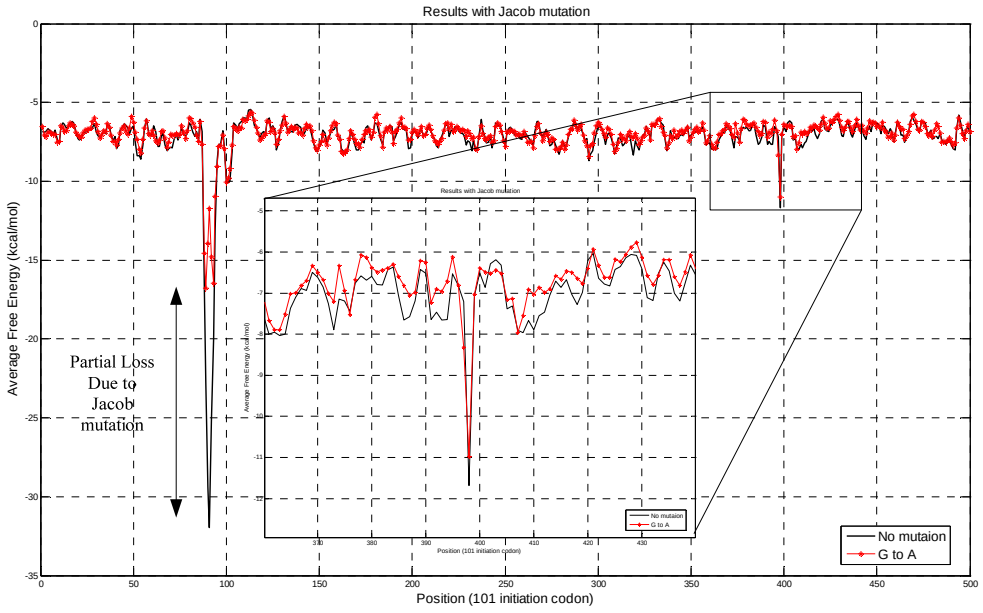


Figure 13: Results with Jacob mutation

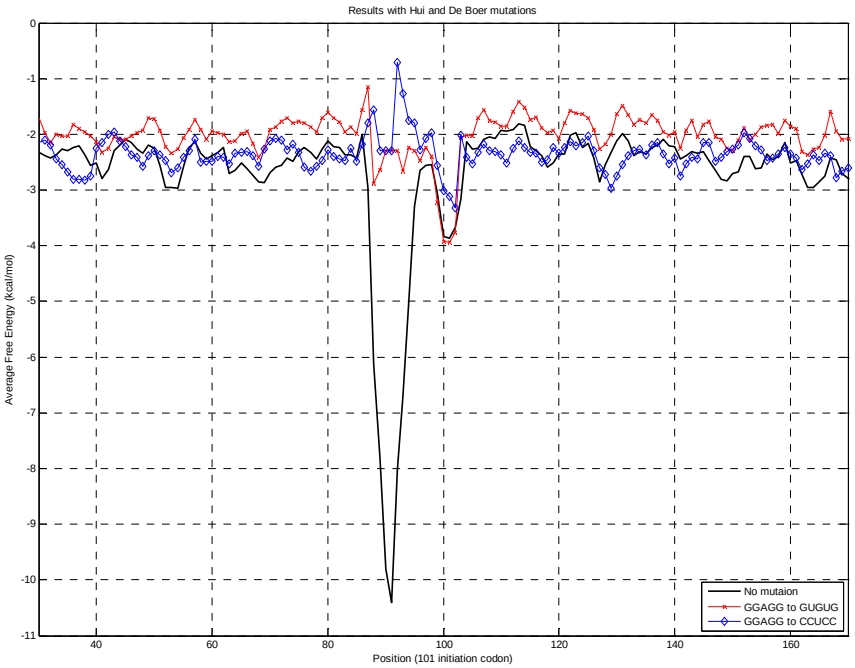


Figure 14: Results with Hui and De Boer mutations using algorithm in [6]

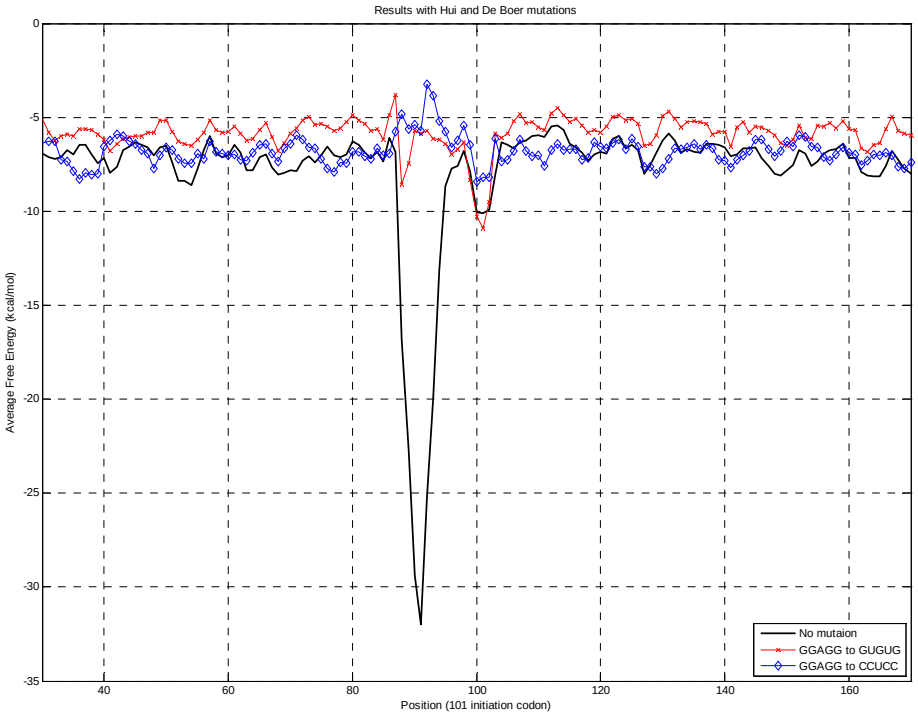


Figure 15: Results with Hui and De Boer mutations using the our algorithm

These results are completely consistent with the published experimental results. This demonstrates the relevance of our 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.

To further exploit the model, point mutations have been introduced in all positions of the last 13 bases of the 16SrRNA molecule in order to study their influence on the process of translation. The obtained results are summarized in Table 3 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.

POSITION	1	2	3	4	5	6	7	8	9	10	11	12	13
SD	–	–	–	↓	⇓	⇓	↓	–	–	–	–	–	–
Initiation	–	–	↓	↑	↓	↓	↓	–	⇓	↓	↓	↓	↑
Stop	↓	⇓	↓	–	–	–	–	–	↓	⇓	⇓	–	↑

Table III: point mutations in the last 13 bases of 16SrRNA molecule

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 will be further carried out using the other four methods discussed in sec 3.1.2 and 4.1.2.

3.3 Variable Length Code Modeling

Preliminary results show that genes (the coding regions) can be modeled as prefix codes (i.e. no gene is a prefix of any other gene in the whole genome). Adding up the non-coding regions we can still have the prefix condition satisfied. This can be proved using the fact that prefix codes

should satisfy the Kraft's inequality which characterizes the sets of codeword lengths that are possible in a prefix code. For clarification, let each source symbol from the alphabet $\overline{S} = [s_1, s_2, ..., s_n]$ be encoded into a uniquely decodable code over an alphabet of size r with codeword lengths $l_1, l_2, ..., l_n$, then

$$\sum_{i=1}^{n} \left(\frac{1}{r}\right)^{l_i} \leq 1, \qquad i \in \{1,2,3,...,n\} \tag{4}$$

where $\overline{S}$ denotes the set of all genes, n is the number of genes, l_i is the length of the ith gene (in codons), and r is the alphabet size and here is equal to 64 denoting the number of all possible codons.

Figure 16 shows a general proposed tree diagram representation of all possible genetic sequences of any length. Here we have mapped the 64 codons to the numeric alphabet {1, 2… 64}. Hence any genetic sequence (coding + non-coding regions) can be mapped to a certain branch in the tree. The terminal node in each branch is the stop codon. In a prefix code, the codewords are only associated with the terminal nodes. The code for any gene can be obtained by traversing the tree from the root to the terminal node corresponding to that gene. In Figure 16, the orange (upper) branch corresponds to the "gene code" {1, 3, 64, 64, 1}, and the blue branch corresponds the "gene code" {64, 2, 64}.

Since prefix codes are uniquely decodable, a message (DNA) can be transmitted as a sequence of concatenated codewords (coding and noncoding regions) and hence can be decoded instantaneously. An iterative decoding algorithm based on VLC decoding techniques [18] can be developed for gene identification. If a gene of length i (which corresponds to a certain branch in the tree diagram) is identified, then all genes of length j ($j > i$) that branch out from this specific gene (i.e. $64^{(j-i)}$ genes) will be eliminated (out of the search). This will speed up the finding of genes by eliminating in the search the genes that have the detected gene as a prefix (Our proposed gene identification algorithm is described in section 4.2).

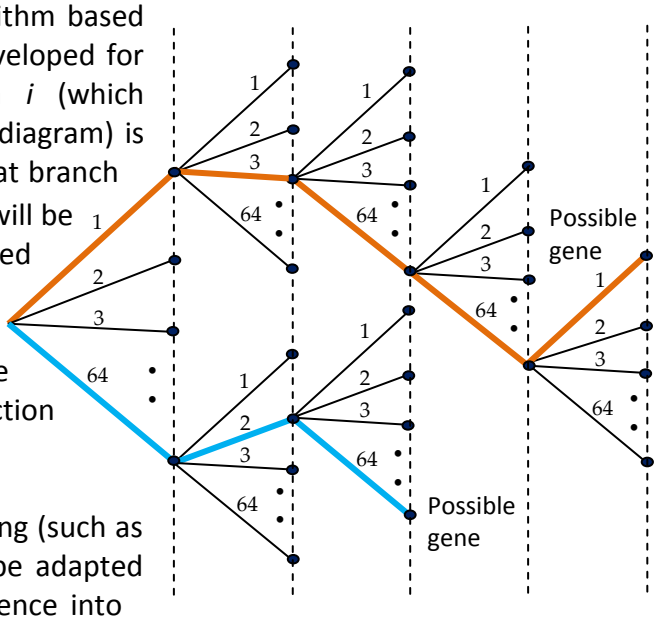


Figure 16: Tree representation of genes

Some of the algorithms used in prefix decoding (such as conventional look-up table approach), can be adapted here to be used in decoding the DNA sequence into the set of all genes. A table of all possible genes that code for proteins (# of proteins is ~$10^{4.5}$) can be assumed to be our look-up table. Moreover, tree search algorithms can be utilized here as well. The basic principle is that a node is taken from a data structure, its successors examined and added to the data structure. By manipulating the data structure (the DNA in our case); the tree is explored in different orders for instance level by level (breadth-first search [19]) or reaching a

leaf node first and backtracking (depth-first search [20]). Other examples of tree-searches include iterative-deepening search [21], depth-limited search [21], bidirectional search [22], and uniform-cost search [21]. We also can make use of the information that regulatory sequences corresponds to specific transitions in the tree (trellis) path and these sequences are found at relative positions with respect to the start/stop codons.

Prefix property should be also verified when using an alphabet of 20 amino acids instead of an alphabet of 64 codons (more compact representation).

3.4 Coding Theory and Genetic Code

The discovery of the mapping of codons to amino acids (known as the genetic code) was a major advance in the field of molecular biology [12]. The genetic code has 64 codons that uniquely map to 20 amino acids which is a redundant mapping. This redundancy suggests that an embedded structure may exist (code). There are many research efforts trying to study the evolution of the genetic code and its optimality properties. The approach used to test optimality is based on generating other mappings of codons to amino acids and trying to compare them with the natural genetic code using physio-chemical metrics such as polarity and hydrophobicity (e.g. see [27]). We have initially mapped the genetic alphabet {A, U, G, C} to a numeric alphabet {0, 1, 2, 3} respectively. In Figure 17 we show a 3D colored graph of the 64 codons being mapped to the known 20 amino acids. Here, we have assigned codons to amino acids in an arbitrary manner. Other mappings should be studied that might show different relationships (metrics) between codons and amino acids.

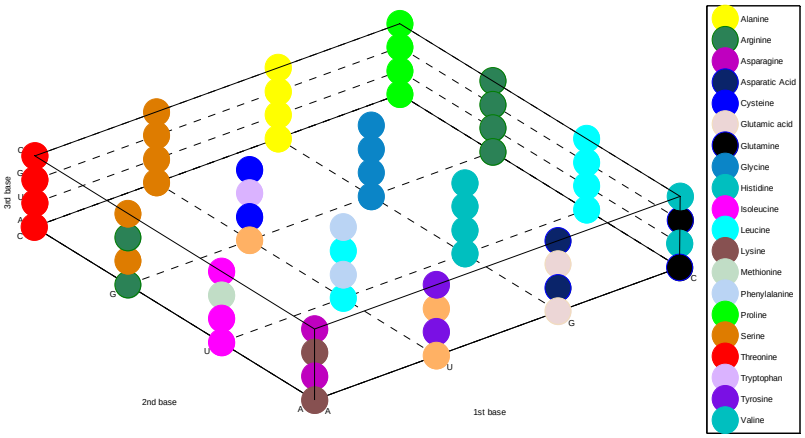


Figure 17: 3D plot of all amino acids (stop codons are UAA, UAG, and UGA)

One can easily show that codons which code for one amino acid are more closely related to one another (in sequence) than they are related to codons that code for other amino acids. In other words, codons that code for one amino acid differ in several cases by just one nucleotide. Thus, single nucleotide mutations will often not change the resulting amino acid rather than lead to an error. Investigating protein substitution matrices, another interesting observation is that the smaller the number of codons per amino acid, the higher the self substitution scores for that amino acid. A higher self substitution score implies that the amino acid was more often conserved in its location within evolutionary related protein sequences.

3.5 Level of Gene Expression under Different kinds of Stress

Another direction of current research is to study the level of gene expression in E.coli when subjected to different kinds of stress. Figure 19 below shows a number of growth curves for three different E.coli strains under different levels of Chlorine concentration. Basically the horizontal axis stands for all the time points, with an interval of 30 min, and a total length of 49 hours. The vertical axis represents the OD value (count of bacteria) for different E.coli strains with different treatments. The E.coli strains are: Sakai, K12 and TW 14359 (a spinach outbreak strain). The treatments include control (no chlorine), and chlorine with a concentration ranging from 500ppm to 1300ppm.

The OD value actually stands for the "optical density" of the bacteria culture, this density will increase if the amount of bacteria increases. Therefore, it can represent the amount of bacteria (biologically called "CFU", which stands for "Colony-Forming Unit"). In our case, an OD value 0.15 is equal to approximately 10^4-10^5 CFUs.

Laboratory data required for the study was provided by study done in the National Center for Food Safety and Technology (NCFST in Chicago, IL). A mathematical model used to analyze such a study is described in sec 4.5. The proposed solution is designed as a variation of the logistic equation published by Pierre Verhulst. The differential equation is modified empirically in order to get a good approximation of the original curves.

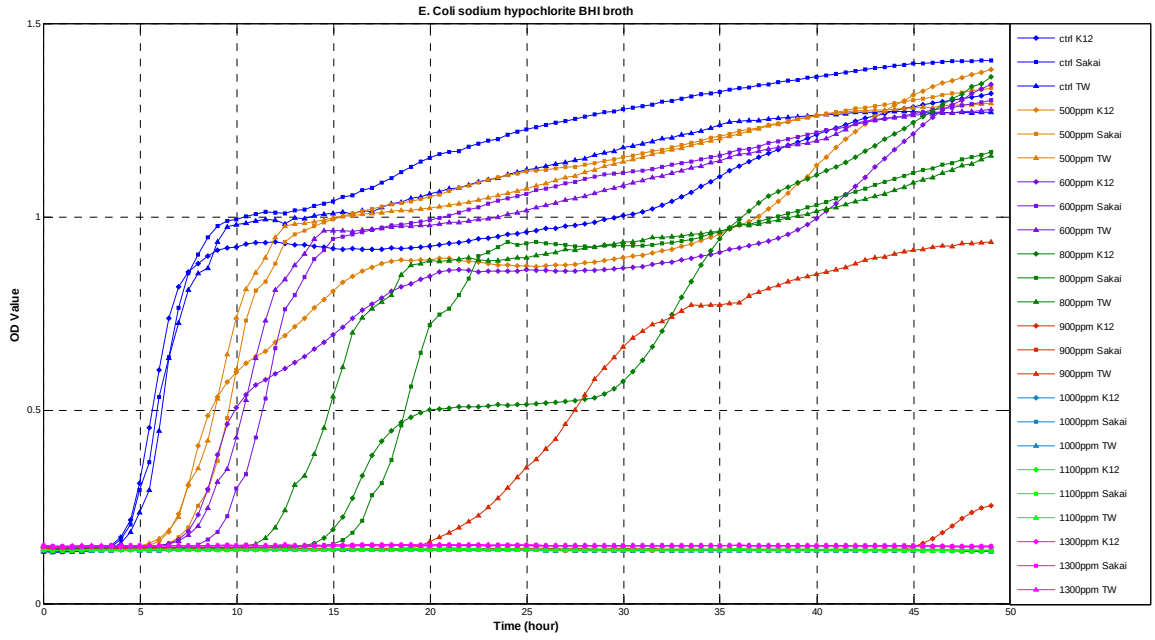


Figure 19: Growth Curves for E.coli under different levels of Stress

4 Research Design and Future Work

Analyzing DNA processing in gene expression, many similarities with the way engineers send digital information in communication systems come into view. 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. Transcription initiation corresponds to a process of frame synchronization where the RNA polymerase detects the promoter sequences (biological sync words). 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 20 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 [23].

Transcription involves decoding the noisy DNA sequence into an mRNA sequence. Mapping this decoding into a decoding matrix (parity check matrix) will provide insight of the error correction or detection in this conversion. Results will provide invaluable information about transcription and its ability of processing the correct decoded sequence. The work of May [23] established the first concrete ideas for modeling gene expression interactions based on algorithms inspired from coding theory [25] [26] [27].

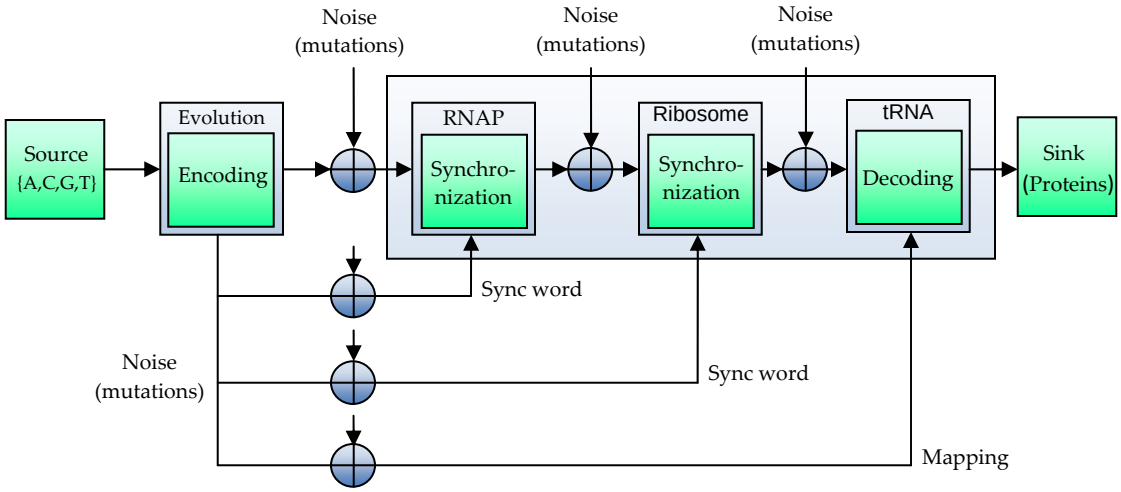


Figure 20: Communication theory model for gene expression

In continuous and packet data transmission, successful decoding of a transmitted data stream at the receiver side strongly depends on the choice of the synchronization (sync) word that indicates the beginning of the message and thus needs to be detected reliably. Analogously, biological sync words indicate the beginning of a gene, i.e. they mark the sequence in the DNA that needs to be copied during transcription. These biological sync words are the promoter (and other transcription factor binding regions) and terminator regions, which identify the limits of the gene (message). In protein coding genes, this message goes through another cycle, in which

it is transmitted to the translational machinery, which has to identify translational start and stop signals (Shine-Dalgarno or Kazak sequences in prokaryotes and eukaryotes respectively; and start and stop codons; as well as other signals such as IRES sequences). This analogy between frame synchronization in digital data transmission and transcription and translation initiation provides a powerful tool for promoter analysis. Promoters can be seen as biological sync words that need to be detected reliably by the protein sigma factor. Research in molecular biology has focused on bacterial promoter regions for decades, however, without addressing the presented aspects of a sequence's detectability. Our approach helps to bridge this gap which demonstrates once more the importance of communications theory for the interpretation of processes in molecular biology.

Table 3 summarizes the comparison of digital communication systems and transcription and translation initiation.

Table IV: Comparison of Frame Synchronization and Bacterial Transcription and Translation Initiation

	Digital Communications	Transcription initiation	Translation Initiation
Data	binary, quaternary or larger alphabet data streams	quaternary DNA sequence (can be mapped to a larger alphabet)	quaternary mRNA sequence (can be mapped to a larger alphabet)
Marker	binary or quaternary synchronization word	two quaternary promoter regions	quaternary Shine-Dalgarno region
Detection	correlator	sigma subunit of RNAP	16s rRNA molecule
Decision Criteria	correlation between sync word and data	binding energy between sigma factor and DNA	Binding energy between ribosome and mRNA

Our research will address the goals described in section 1 (Introduction). The following sections will describe our research and design methods that are going to be considered in this work. Sec 4.1 describes the five new models proposed, presents a simple example to verify their performance, and address future work. Sec 4.2 describes the mutation analysis that is going to be carried out using the proposed models and address future work as well. Sec 4.3 presents a variable length code (VLC) based algorithm for gene identification. Sec 4.4 addresses our future work related to the application of coding theory principles to model and analyzes the structure of the genetic code. Sec 4.5 proposes a mathematical approximation to study the level of gene expression under different kinds of stress. Sec 4.6 introduces an improved gene and regulatory sequences identification approach that will provide a solution for current limitations that exist in gene-finding programs by using pattern recognition, Discrete Fourier Transform, and Wavelet analysis. Finally Sec 4.7 addresses our future work to extend the proposed study other prokaryotic and eukaryotic genomes.

4.1 Coding Theory, Communications and Information Theory Based Modeling

4.1.1 Coding Theory Based Modeling

Our research is directed to use the models developed in our preliminary work and a variation of it to gain new insights on the biological interactions between the RNA polymerase and DNA on one side, and ribosome and mRNA on the other side. We have used an exponential metric with a one-dimensional variable length codebook. Our future work will consider:

1. Applying different algorithms for regulatory sequence detection that will be adapted to detect start and stop codon locations as well.
2. Using autocorrelation and cross-correlation functions to analyze coding and non-coding regions in DNA sequence. This will allow for detecting common patterns that repeat along DNA sequence.

4.1.2 Communications and Information Theory Based Modeling

The process of detecting a regulatory sequence in the DNA sequence can be achieved using the detection techniques used in communications engineering. Based on this analogy, concepts like correlation, convolution, Euclidean distance, matched filter, and certain metrics can be utilized in this detection process. The following five modls are based on these concepts:

Model I: Number of Matches vs. Position

This method locates a binding sequence by sliding it through the mRNA sequence and counting the number of matches at each alignment as a function of position. If the number of matches is equal or close to the length of the binding sequence under study (i.e. if a total or partial match occurs), a peak will occur. This will account for the mismatches that might happen during alignment. Applying a threshold to the number of matches will control the resolution of binding sequence detection. This method is not only able to detect the binding sequences at their exact locations but also results in peaks with amplitude equal to the number of matches available at each alignment. This allows for a more informative output.

Model II: Euclidean Distance Based Algorithm

In this method, a Euclidean distance measure can be used to detect a given binding sequence in the mRNA sequence. This measure is calculated at each alignment as follows:

1. Map both mRNA sequence and the binding sequence under study 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 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.

Model III: Cross Correlation (Matched Filter)

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. Model III is based on using a matched filter of an impulse response equal to y(-n) and an input of x(n) (see Figure 21) where y(n) is the binding sequence and x(n) is the mRNA sequence.

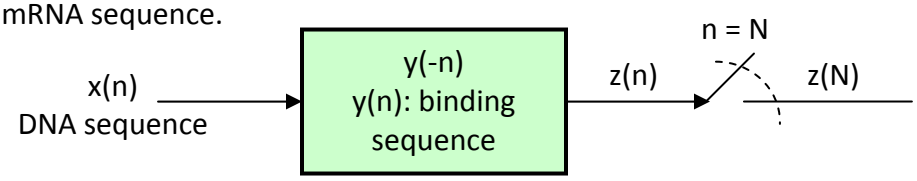


Figure 21: Matched Filter

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

$$z(n) = x(n) \otimes y(n) = x(n) * y^*(-n) = \sum_{n=-\infty}^{\infty} x(k)y(n+k)\,. \tag{5}$$

   where ($\otimes$) corresponds to cross correlation and ($*$) corresponds to convolution. 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:

$$DP_{mRNA} = \lceil \left(DP_{plot} - 2\,BSL + 1 \right) / 2 \rceil \tag{6}$$

   where DP_{mRNA} is the detected position in the mRNA sequence, DP_{plot} is the detected position in the plot, BSL is binding sequence length, and $\lceil X \rceil$ rounds the value X to the nearest integer larger than X.

Model IV: Exponential Detection Metric

This method detects a TFBS based on aligning the binding sequence with the DNA sequence. An exponential metric related to the number of matches at each alignment is evaluated as follows:

1. Slide the binding sequence under study along the DNA sequence one base at a time.

2. At the i[th] alignment, compute an exponential weighting function ($W(i)$) using the equations:

$$W(i)=\sum_{n=1}^{N} w(n), \tag{7}$$

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:

$$w(n)=\begin{cases} a^{\sigma} & , \ if\ match \\ 0 & , if\ mismatch \end{cases}, \tag{8}$$

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

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

$$\overline{W}=\left[w(1),w(2),...,w(L-N+1)\right], \tag{9}$$

where L is the length of the DNA sequence under study.

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

Model V. Free Energy Metric

In this method we use the free energy table (see Table II) to calculate a free energy distance metric in kcal/mol. This metric is calculated 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 ith alignment, calculate the free energy metric using the equation:

$$\mathrm{E}(i)=\sum_{n=1}^{N-1} E\left(y_n y_{n+1}\right).\,\delta\left(n\right) \tag{10}$$

where N is the length of the binding sequence. $\overline{y}$ denotes the binding sequence vector and is given by $\overline{y}=[y_1,y_2,...,y_N]$. Let $\overline{x}$ denote the mRNA sequence vector where $\overline{x}=[x_1,x_2,...,x_L]$.

$E(y_n y_{n+1})$ is the energy dissipated on binding with the nucleotide doublets $y_n y_{n+1}$ and is calculated from Table II. $\delta(n)$ is given by:

$$\delta(n)=\begin{cases} 1 & , if\ y_n y_{n+1}=x_n x_{n+1}\ (match) \\ 0 & , if\ y_n y_{n+1}\neq x_n x_{n+1}\ (mismatch) \end{cases} \tag{11},$$

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

To show how the five previous models behave, we arbitrarily selected a 71-bases-long mRNA sequence as a test sequence. Then, we chose an 11-bases-long sequence starting at position 13 to be a hypothetical binding sequence. This binding sequence was also inserted at position 53 with two bases being changed to get a partial match of the original sequence. The five previous

models were applied to detect these binding sequences. Figures 22-26 show that these methods are accurately detecting the binding sequence as expected. A total match occurs at position 13 (longer peak/dip), and a partial match occurs at position 53 (shorter peak/ dip).

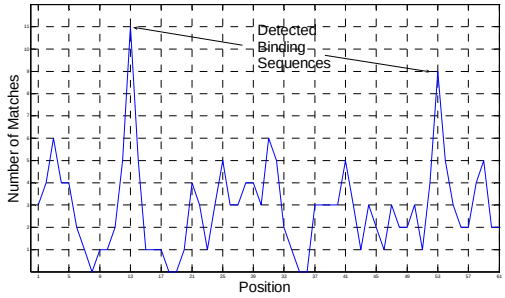


Figure 22: Model I: Number of Matches vs. Position

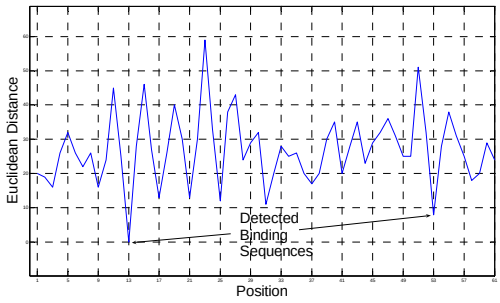


Figure 23: Model II: Euclidean Distance Metric

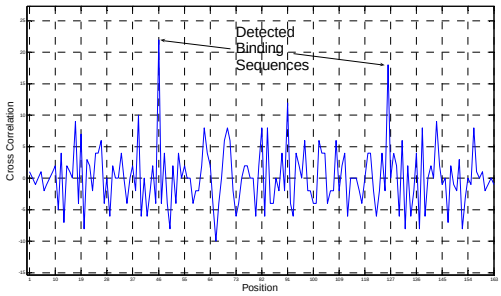


Figure 24: Model III: Cross Correlation (Matched Filter)

Figure 25 shows that the binding sequence has been detected at positions 46 and 126. According to equation 5, these positions correspond to positions 13 ($\lceil (46 - 22 + 1)/2 \rceil$=13) and 53 ($\lceil (126 - 22 + 1)/2 \rceil$=53) in the original mRNA sequence, respectively.

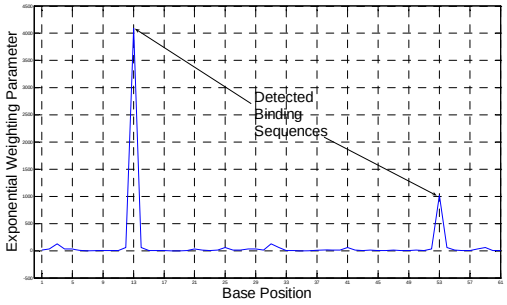


Figure 25: Model IV: Exponential Detection

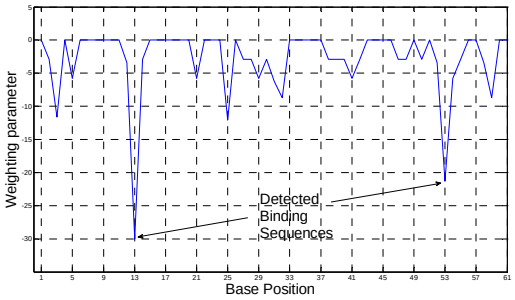


Figure 26: Model V: Free Energy Metric

In future work, these five proposed models described above are going to be applied to detect the last 13 bases of the 16S rRNA molecule in the given mRNA sequence. Preliminary simulation results show that the proposed models allow detecting the translational signals at their exact corresponding locations. Furthermore, they allow identifying coding regions and the and noncoding regions. This new finding suggests that the last 13 bases of 16S rRNA molecule has a higher correlation with coding regions as compared with noncoding regions. This also suggests that the proposed models, which were originally designed for regulatory sequence identification, can be utilized for gene identification as well. Other binding sequences will be considered for further analysis. Transcription Factor Binding Sites (TFBS) can be also detected using the five models proposed in this paper as we will move to study eukaryotes (like us) instead of prokaryotes (like E. coli).

4.2 Mutation Analysis

In our work [1] [2] based on the codebook model, we have applied our proposed algorithm to test the effect of single point mutations in the ribosome on protein synthesis. To do this, we have introduced point 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. The obtained results totally agreed with published experimental results in terms of their effect on the level of gene expression. Another published record of the behavior of protein synthesis under mutations in the 3' end of the 16S rRNA, was done by Hui and De Boer [17]. These two mutations were also tested using our proposed model and results totally matched laboratory experimentation as well.

Jacob mutation, a mutation in the 5th position of the last 13 bases of 16S rRNA molecule [16], results in a reduction in the level of protein synthesis. This mutation was tested using the codebook model proposed in sec 3.1.1. Simulation result showed a reduction in the amplitude of the Shine-Dalgarno signal compared to the non-mutation case. 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.

Hui and De Boer mutations occur in positions 4 to 8 (GGAGG → CCUCC) and positions 5 to 7 (GAG → UGU). The results of both mutations are lethal for the organism in the sense that the production of proteins stop. Simulation results showed a complete loss of the Shine-Dalgarno (SD) signal. Hence, it can be inferred that the translation will never take place.

In future work, this mutation analysis will be carried out using the new five models proposed in sec 4.1. Other types of mutation in the ribosome will be investigated as well. Preliminary results show a total agreement with experimental work that has published records.

Our proposed work will extend mutation analysis results obtained in preliminary work to: 1) design similar models for the process of transcription in prokaryotes, 2) design similar models for gene expression in eukaryotes including translation, transcription, and splicing, and 3) apply the developed models to genomes of different organisms.

4.3 Variable Length Modeling - Gene Identification Algorithm

Based on the analogy between DNA and variable length codes (VLC), genes can be viewed as branches in a tree diagram (Figure 16) and hence can be identified (located) using the following procedural steps:

1- Design a sequence search algorithm based on correlation, matched filters, or codebooks to identify a regulatory elements (REs) (e.g. promoters, ribosome binding sites, start codons, stop codon, transcription factor binding sites etc.) in a data stream (e.g. DNA) with a well-defined resolution.
2- Decide which groups of REs (identified using algorithm developed in step 1) and data are organized in a fashion that suggest a functional gene. This includes proper placement of regulatory sequences such as promoters, enhancers, ribosome binding sites (Shine-Dalgarno sequence in prokaryotes), exon structure including splice site recognition (in Eukaryotes), or any other transcription factor (TF) binding sites that occur in proximity to start codons. This will require building a data base of all known promoters and TF binding sites. This process will be iterative in nature, and additional information obtained in the iterations will be used to improve posteriori decisions (turbo decoder principle).
3- Assign all detected genetic sequences (coding + noncoding) to their corresponding branches in the tree diagram representation described in Figure 16. This will help eliminate some wrongly detected genes.
4- Study correlations between coding and non-coding regions for every sequence, correlations among coding regions, and correlations among non-coding regions. This will help identify characteristics to the organism under study and detect new possible regulatory sequences.

This algorithm of gene identification will be first applied to prokaryotes (like E. coli) as a start and then will be directed toward eukaryotes (like us). The prefix structure will have to be verified for all organisms that we will be dealing with. If this condition doesn't hold true; still the searching algorithm will be based on detection methods used in communications (correlators, matched filters, codebooks, soft and/or hard decisions, etc. [26]). The specific

method to be used in the different cases will be adapted depending on the general characteristics of the organisms under study.

The Decision in step 2 can be made using the following approaches:

1) A block code model can be used to distinguish translated from non-translated genes. The messenger RNA (mRNA) can be modeled as a noisy (with errors), encoded signal and the ribosome as a minimum Hamming distance decoder, where the 16S ribosomal RNA (rRNA) serves as a template for generating a set of valid codewords (the codebook) [5].
2) Block-code-based Bayesian classifiers can be used to distinguish translation initiation sites from non-initiation sites. This classification is based on the average minimum Hamming distance values in the −15 to −11 alignment window (this includes mRNA bases from position −15 to −7). The −15 to −11 window appears to provide the greatest distinction between the mean minimum Hamming distance values of leader (contains valid initiation site) and non-leader (contains invalid initiation site) sequences in E. coli K-12 [28].

4.4 Coding Theory and Genetic Code

Based on the given observations and analysis in sec 3.4, our research will be directed to

1. uncover the structure of the genetic code using channel coding theory,
2. prove the optimality of the genetic code using channel coding theory,
3. study the relationship between the number of possible codons that result in a given amino acid and the importance of the amino acid,
4. investigate the relationship between the redundant structure of the genetic code and DNA repair mechanisms, and
5. check if there is some kind of structure or pattern when mapping codons (alphabet size = 64) to amino acids (alphabet size = 20).

4.5 Level of Gene Expression under Different kinds of Stress

The Verhulst approximation

To build a mathematical model of bacterial growth under different levels of stress as given in sec 3.5, Verhulst approximation can be utilized. The logistic equation (or Verhulst equation) is a model of population growth first published by Pierre Verhulst (1845-1847), which is given by

$$\frac{dy}{dt} = \big(r - ay(t)\big)y(t)\,, \tag{12}$$

where r and a are constants. This equation was first introduced by the Belgian mathematician Pierre Verhulst to study population growth. The logistic equation differs from the Malthus model in that the term $r - ay(t)$ is not constant. This equation can be written as

$$\frac{dy}{dt} = \big(r - ay\big)y = ry - ay^2\,, \tag{13}$$

where the term $-ay^2$ represents an inhibitive factor. Under these assumptions the population is neither allowed to grow out of control nor grow or decay constantly as it was with the Malthus model.

The logistic equation is separable, and thus, can be solved by separation of variables. We solve the equation subject to the condition $y(0) = y_0$. Separating variables and using partial fractions to integrate with respect to y, we have

$$\frac{1}{(r-ay)y}dy = dt, \tag{14}$$

$$\left(\frac{a}{r}\frac{1}{r-ay} + \frac{1}{r}\frac{1}{y}\right)dy = rdt, \tag{15}$$

$$\left(a\frac{1}{r-ay} + \frac{1}{y}\right)dy = rdt, \tag{16}$$

$$-\ln|r-ay| + \ln|y| = rt + C, \tag{17}$$

Using the properties of logarithms to solve this equation for y yields

$$\ln\left|\frac{y}{r-ay}\right| = rt + C, \tag{18}$$

$$\frac{y}{r-ay} = \pm e^{rt+C} = Ke^{rt}, \tag{19}$$

$$y = r\left(\frac{1}{K}e^{-rt} + a\right)^{-1}, \tag{20}$$

Applying the initial condition $y(0) = y_0$ and solving for K, we find that

$$K = \frac{y_0}{r-ay_0}, \tag{21}$$

After substituting this value into the general solution and simplifying, the solution of the equation that satisfies the initial condition $y(0) = y_0$ can be written as

$$y = \frac{ry_0}{ay_0 + (r-ay_0)e^{-rt}}. \tag{22}$$

Notice that if $r > 0$,

$$\lim_{t\to\infty} y(t) = r/a, \text{ because } \lim_{t\to\infty} e^{-rt} = 0. \tag{23}$$

This makes the solution to the logistic equation different from that of the Malthus model in that the solution to the logistic equation approaches a finite nonzero limit as $t \to \infty$ while that of the Malthus model approaches either infinity or zero as $t \to \infty$.

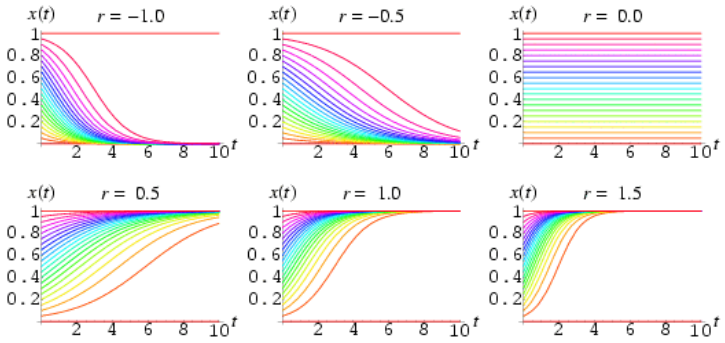


Figure 35: Logistic model - The model is continuous in time, but a modification of the continuous equation to a discrete quadratic recurrence equation known as the logistic map is also widely used

As it has been explained before, the logistic equation given by the equation:

$$\frac{dN}{dt} = r_{\max} N\left(\frac{K-N}{K}\right), \tag{24}$$

where $dN/dt = r_{\max} N$ represents the exponential growth which is unlimited, $(K-N)$ represents how many individuals can be added to the population, and $(K-N)/K$ is the fraction of K that is still available.

$$\frac{dN}{dt} = r_{\max} N\left(1-\frac{N}{K}\right), \tag{25}$$

An example is shown with $r_{\max} = 1.0$ and $K = 1{,}500$.

Adjusting the parameters, the theoretical model can be adjusted to the experimental data. It is proven, therefore, that Verhulst curves can be used to model the e-coli OD (count of bacteria) value under different kinds of stress. This mathematical model can save time and cost of laboratory experimentation. Markov models can be used to analyze such a study. Laboratory data required for the study are available and ready for analysis.

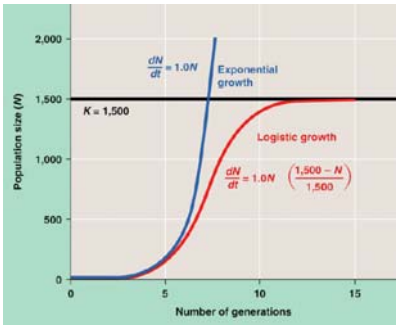


Figure 36: Verhulst Approximation

4.6 Pattern recognition in gene identification using DFT

The essential underlying assumption for pattern recognition in biological sequences is that strings carrying information will be different from random strings, which have no information. So if a hidden pattern can be identified in a string, it must be carrying information. This task needs to be automated because of the large sizes of the genome. To have an estimate of the size of the DNA strings for computational analysis, it may be noted that the species with the smallest genome, Mycoplasma genitalium – a parasite genome, originally isolated from urethral specimens of patients is about 6×10^5. The human genome is about 3×10^9 bps long.

The various approaches used in gene prediction are:

- Detecting appropriate groups of REs and Open Reading Frames (ORFs) as described

above in Sec 4.1.

- Homology search (involves pair-wise alignment) against known genes.
- Content-based methods: Ab initio methods based on statistics, nucleotide distribution, periodicity in base occurrence, their dependencies on the characters preceding it (i.e., how often an A is followed by a C, etc.), frequency of occurrence of codons (triplets), di-codons (hexamers), amino acids, etc.
- Signal-based methods: look for signals in the vicinity of coding region, viz., CpG islands, promoter sequences, translational signals, poly-A signal, splice sites, etc.
- In the feature generation stage orthogonal transforms are used. Techniques such as DCT, DST, Hadamard and Haar transforms are commonly used. Other novel techniques such as DFT and Wavelet transforms are also employed now in this area. The incentive is to give all the necessary information so that you are able to develop software, based on filter banks, in order to generate features.

Prediction of multiple genes in a sequence is still difficult and most programs only predict protein coding genes and not genes whose products function exclusively at the RNA level. In our research, a signal-based approach will be the approach. DFT and wavelet transform techniques will be used to correlate the synchronizing words and the data, in our case; these represent the binding energy between the sigma factor and DNA. Based on the basic properties of the DFT like periodicity, cross-correlation and circular convolution, an improved detection approach will be introduced. This method can provide useful information about gene locations and it will provide a solution for current limitations that exists in gene-finding programs.

4.7 Application and Extension to other Organisms

The proposed models will be extended to other prokaryotic and eukaryotic genomes to understand the mechanisms of transcriptional regulation in different spatial and temporal contexts. Given the complex pattern of regulatory interactions, the motif discovery tools and comparative genomics approaches will also be integrated to detect regulatory elements in many genomes, including the accurate location of transcriptional start sites, DNase hypersensitive sequences within nuclear chromatin that represent regulatory regions (including promoters, enhancers, silencers, locus-control regions), and TF binding locations from the ChIP–chip experiments.

5 References

[1] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo Atkin and Wei Zhang, “Ribosome Binding Model Using a Codebook and Exponential Metric,” IEEE EIT 2007 Proceedings, Chicago, IL, USA, May 17 – 20, 2007.

[2] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo Atkin and Wei Zhang “An Optimized Ribosome Binding Model Using Communication Theory Concepts,”; In: Proceedings of 2007 International Conference for Bioinformatics and Computational Biology, Las Vegas, June 25 – 27, 2007.

[3] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo Atkin and Wei Zhang, “Regulatory Sequence Identification using Communications, Coding and Information Theory Based Models”,; submitted to Proceedings of 2009 AMIA (American Medical Informatics Association) Summit on Translational Bioinformatics, San Francisco, March 15 – 17, 2009.

[4] Mohammad Al Bataineh, Maria Alonso, Lun Huang, Ismaeel Muhamed, Nick Menhart and Guillermo Atkin, “Novel Gene Finding Approach Based on Regulatory Sequence Detection”; submitted to Proceedings of 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nashville, TN, USA, March 30–April 2, 2009.

[5] E. E. May, M. A. Vouk, D. L. Blitzer, and D. I. Rosnick, “Coding theory based models for protein translation initiation in prokaryotic organisms,” BioSystems, vol. 76, pp. 249–260, August-October 2004.
[6] Z. Dawy, F. Gonzalez, J. Hagenauer, and J. C. Mueller, “Modeling and analysis of gene expression mechanisms: a communication theory approach,” proceedings of the IEEE International Conference on Communications (ICC), May 2005.
[7] Z. Dawy, B. Goebel, J. Hagenauer, et al., “Gene mapping and marker clustering using Shannon’s mutual information,” IEEE Transactions on Computational Biology and Bioinformatics, vol. 3, no. 1, pp. 47–56, January-March 2006.
[8] “DNA as Digital Data - Communication Theory and Molecular Biology,” IEEE Engineering in Medicine and Biology, vol. 25, no. 1, January/February 2006.
[9] A.W.-C. Liew, H. Yan, and M. Yang, “Pattern recognition techniques for the emerging field of bioinformatics”, Pattern Recognition, vol. 38, pp. 2055—2073, 2005.
[10] E. A. Cheever, G. C. Overton, and D. B. Searls. Fast Fourier Transform-based Correlation of DNA Sequences Using Complex Plane Encoding. Comput. Applic. Biosci. 7(2)143-159, 1991.
[11] Lio, P. 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19:2-9.
[12] B. Hayes. The invention of the genetic code. American Scientist, 86(1):8–14, 1998
[13] J. Steitz and K. Jakes. How ribosomes select initiator regions in mRNA: base pairformation between the 3’ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in E. coli. Proc. Natl. Acad. Sci., 72:4734–4738, 1975.
[14] D. Rosnick, Free Energy Periodicity and Memory Model for Genetic Coding. PhD thesis, North Carolina State Univesity, Raleigh, 2001.
[15] “NCBI: National Center for Biotechnology Information.” http://www.ncbi.nlm.nih.gov/.
[16] W. Jacob et al., “A single base change in the Shine Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins,” Proc. Natl. Acad. Sci., vol. 84, pp. 4757–4761, 1987.
[17] A. Hui and H. D. Boer, “Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli,” Proc. Natl. Acad. Sci., vol. 84, pp. 4762– 4766, 1987.
[18] PAJAROLA R., Fast prefix code processing, Proc. IEEE ITCC Conference,Las Vegas, Nevada, USA, (2003) 206-211.
[19] Knuth, Donald E. (1997), The Art Of Computer Programming Vol 1. 3rd ed., Boston: Addison-Wesley, ISBN 0-201-89683-4.
[20] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Section 22.3: Depth-first search, pp.540–549.
[21] Russell, Stuart J. & Norvig, Peter (2003), Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, NJ: Prentice Hall, ISBN 0-13-790395-2.
[22] Dennis de Champeaux & Lenie Sint, An Improved Bi-directional Heuristic Search Algorithm, Journal ACM, vol 24, no 2, 1977 May, pp 177-191.
[23] E. May, M. Vouk, D. Bitzer, and D. Rosnick. An errorcorrecting code framework for genetic sequence analysis. Journal of the Franklin Institute, 34:89–109, January-March 2004.
[24] J. Shine and L. Dalgarno, “The 3’ terminal sequence of Escherichia coli 16S ribosomal RNA:complementarity to nonsense triplets and ribosome binding sites,” Proc. Natl. Acad. Sci., vol. 71, pp. 1342–1346, 1974.
[25] S. Lin and D. J. Costello, Jr., Error Control Coding. 2nd Edition: Prentice-Hall, 2004.
[26] J. G. Proakis, Digital Communications, 5th ed. New York: McGraw- Hill, 2007
[27] S. J. Freeland, T. Wu, and N. Keulmann. The case for an error minimizing standard genetic code. Orig. Life Evol. Biosph., 33(4-5):457–77, 2003.
[28] M. Tompa, “An exact method for finding short motifs in sequences, with application to the ribosome binding site problem,” in Proc. ISMB, 1999.
[29] M. T.D. Schneider, “Measuring molecular information,” J. Theor. Biol., vol. 201 pp. 87–92, 1999.
[30] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick. Coding Model for Translation in E. coli K-12. In First Joint Conference of EMBS-BMES., 1999.
[31] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick. The Ribosome as a Table-Driven Convolutional Decoder for the Escherichia coli K-12 Translation Initiation System . In World Congress on Medical Physics and Biomedical Engineering Conference., 2000.
[32] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick. Coding Theory Based Maximum-Likelihood Classification of Translation Initiation Regions in Escherichia coli K-12 . In 2000 Biomedical Engineering Society Annual Meeting., 2000.
[33] H. Yockey, Information theory and molecular biology. Cambridge: Cambridge University Press, 1992.
[34] T. Schneider, “Theory of molecular machines I. Channel capacity of molecular machines,” Journal Theoretical Biology, vol. 148, pp. 83–123, 1991.
[35] T. Schneider, “Theory of molecular machines II. Energy dissipation from molecular machines,” Journal Theoretical Biology, vol. 148, pp. 125–137, 1991.