Modeling and Analysis of Genome Structure using Information Theory, Communications and Coding Theory Concepts

By

Guillermo Atkin (Associate Professor, Department of Electrical & Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, United States)

Wei Zhang (Assistant Professor, Department of Biological, Chemical, Physical Sciences, College of Science and Letters; Joint Assistant Professor, National Center for Food Safety and Technology, U. S. Food and Drug Administration)

1 Specific Aims

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. This research proposes a novel use of techniques and principles from communications engineering, coding and information theory such as the ones used in source and channel coding, frame synchronization, pattern recognition, wavelet analysis, and discrete Fourier Transform for modeling, identification and analysis of genomic regulatory elements and biological sequences.

It has become increasingly evident that the *Escherichia coli* species is comprised of clonal lineages that show biased distribution among environmental, food, and human clinical samples. The past knowledge of serotype- or strain-specific prevalence in foods and human infections substantiates the need to elucidate the unique genetic, physiological, and ecological characteristics of this pathogen. In the proposed study, we will combine our experimental data from functional genomics based approaches (i.e. DNA microarrays) with the *in silico* analysis as described above to uncover the genetic and molecular mechanisms that different *Escherichia coli* species use to regulate their genome expression in response to the stimuli and stresses in the natural environment, foods and human or animal species. The proposed experiments build logically from our knowledge of transcription factors and comparative genome analysis of diverse *Escherichia coli* populations. The combination of the experience of our investigators and the studies presented in the preliminary data section underscore the likelihood that the proposed project will yield highly useful results. This proposal represents one of the first attempts to explore information theory and correlate to the functional consequences in the genomes of prokaryotic pathogens.

Communications and information theory has proven to provide powerful tools for the analysis of biological signals [1]–[5]. An up-to-date summary of ongoing research can be found in [6]. 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).

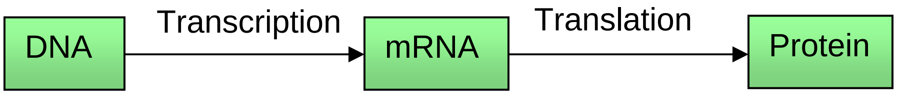


Figure 1: The process of protein synthesis (gene expression)

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*, as well as other bacteria), and then will be extended to study other types of organisms (e.g. eukaryotes).

The main goals of this work are to:

i) develop an analogy 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) 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 modified 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;

iv) 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 [18], Discrete Fourier Transform (DFT) [19], and Wavelet analysis [20];

v) 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;

vi) use principles of error control coding theory to interpret the genetic translation and transcriptions mechanisms;

vii) use the proposed models to test the effect of mutations in the ribosome on protein synthesis, and predict the effect of other possible 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;

ix) and most importantly, integrate research findings from this project with educational and extension programs and activities at Illinois Institute of Technology**.** This is one of the key goals in this proposal in support of NSF's goals to foster integration of research and education through the programs, projects, and activities. PIs of this proposal are actively engaged in various teaching and educational programs and are dedicated to providing diverse learning opportunities to students and general public with different educational backgrounds. We plan to integrate our research with different types of educational and extension programs. The extension activities will include a wider dissemination of findings at appropriate professional scientific meetings as well as the development of more targeted training and educational materials that could be used through a number of different communication routes. Specifically, we will (1) Present our findings in seminar lectures in the Electrical and Computer Engineering (ECE), Biology, Computer Science (CS), Math, and Bio-Medical Engineering (BME) departments at Illinois Institute of Technology; (2) Implement new research findings as teaching materials (such as applications of new computational algorithms in identifying genomic regulatory elements) into the current undergraduate and graduate core curriculum including BIOL562 Functional Genomics currently taught by Dr. Zhang; (3) Develop a new interdisciplinary course “Computational Biology and Bioinformatics” for senior undergrad and entry-level grad in ECE, Biology, CS and BME majors; (4) Provide Special Projects (these are courses with research credits) to minority undergraduate and graduate students in ECE, BME and Biology majors; (5) Develop joint educational program for high school students in the Chicago area (we have hosted such programs at NCFST every year); (6) Organize educational activities and participate in Science Fairs for the general public through the Chicago Council on Science and Technology; (7) develop IPRO (Interprofessional Project program) that joins together students from various academic disciplines to work as a team. Furthermore, we plan to collaborate with other centers of bioinformatics (including the Center for Computational

Biology and Bioinformatics at the University of Maryland), Bioengineering Departments and Research Institutes (such as the Pritzker Institute) to foster education by applying engineering principles to cell biology, integrated with applied mathematics, computational science, bioengineering and medical sciences.

x) encourage the participation of students (women/men) from underrepresented and minority groups, and people with disabilities in our educational and extension programs.

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.

2 Background and Significance

Here we 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).

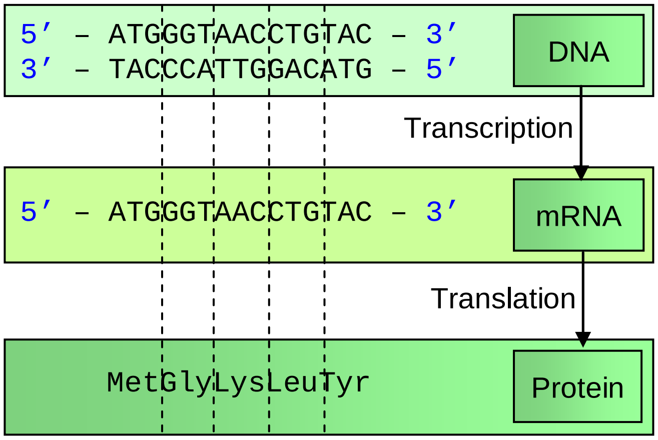

5′ – ATGGGTAACCTGTAC – 3′
3′ – TACCCATTGGACATG – 5′
DNA
Transcription
5′ – ATGGGTAACCTGTAC – 3′
mRNA
Translation
MetGlyLysLeuTyr
Protein

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 [10]. 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.

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 and most studied types in *E. coli* is illustrated in Figure 3.

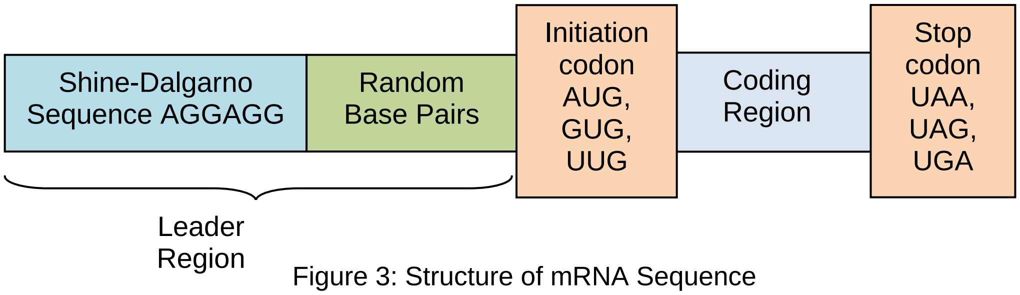


Figure 3: Structure of mRNA Sequence

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 our preliminary work, we use 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.

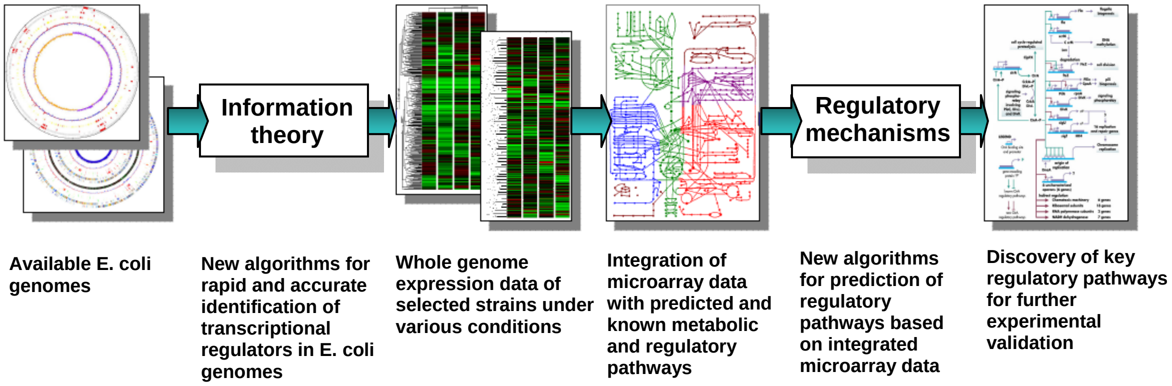

Information theory
Regulatory mechanisms
Available E. coli genomes
New algorithms for rapid and accurate identification of transcriptional regulators in E. coli genomes
Whole genome expression data of selected strains under various conditions
Integration of microarray data with predicted and known metabolic and regulatory pathways
New algorithms for prediction of regulatory pathways based on integrated microarray data
Discovery of key regulatory pathways for further experimental validation

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 are 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 built 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 then extending our results to eukaryotes.

The algorithms developed 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 new developed 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 algorithms developed 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

- Study relationships (correlations, distance metrics, etc) between coding regions and noncoding regions and regulatory sequences

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.

3.1 Coding Theory, Communications and Information Theory Based Modeling

3.1.1 Coding Theory Based Models

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

Table 1: 16SrRNA Codebook

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 2: Energy Doublets [17]

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 [8]. 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 [UAAGGAGGUGAUC]. A sliding 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 (see Table 1). 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 energetics involved in the rRNA-mRNA interaction tells the ribosome when a signal is detected and, thus, when the start of the process of translation should take place. In our model, the a modified version of the method of free energy doublets presented in [17] 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) [4]. 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^{th} 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.

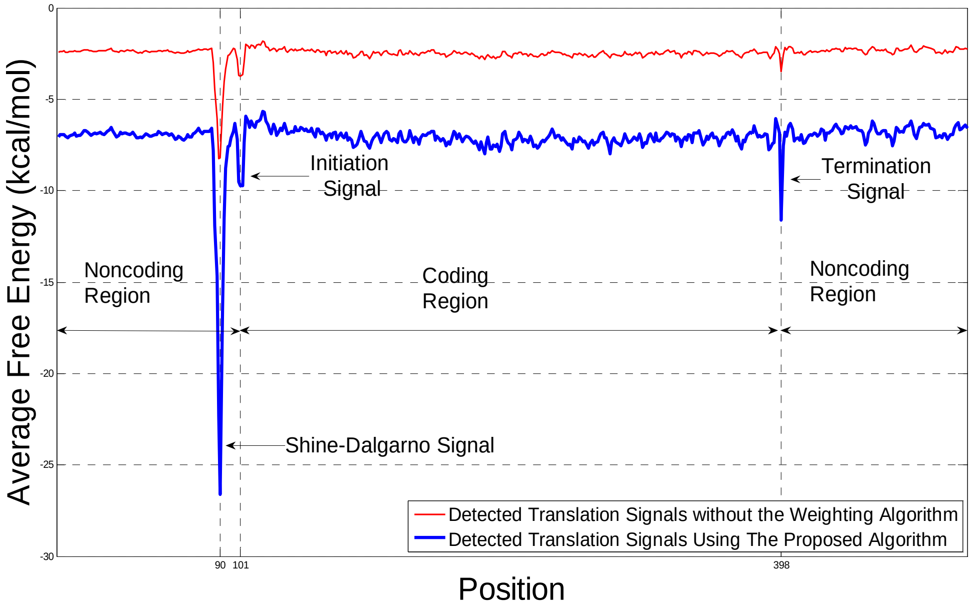


Figure 5: Comparison of SD signal (position 90), start (position 101) and termination (position 398) codon between the algorithm used in [6] and the weighted algorithm (N=5, a=1.5)

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 deciding the exact position of the Shine-Dalgarno on the genes rather than using an average.

For the analysis, 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 alone (without weighting). Figure 5 shows average results for the detection of the SD, start and stop codons being compared to previous work [4]; it can be seen that the proposed algorithm is able to identify the Shine-Dalgarno (peak at position 90) and the start codon (peak at position 101) and the stop codon (peak 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 (section 3.3) and the results obtained further certified the correctness and the biological relevance of our model.

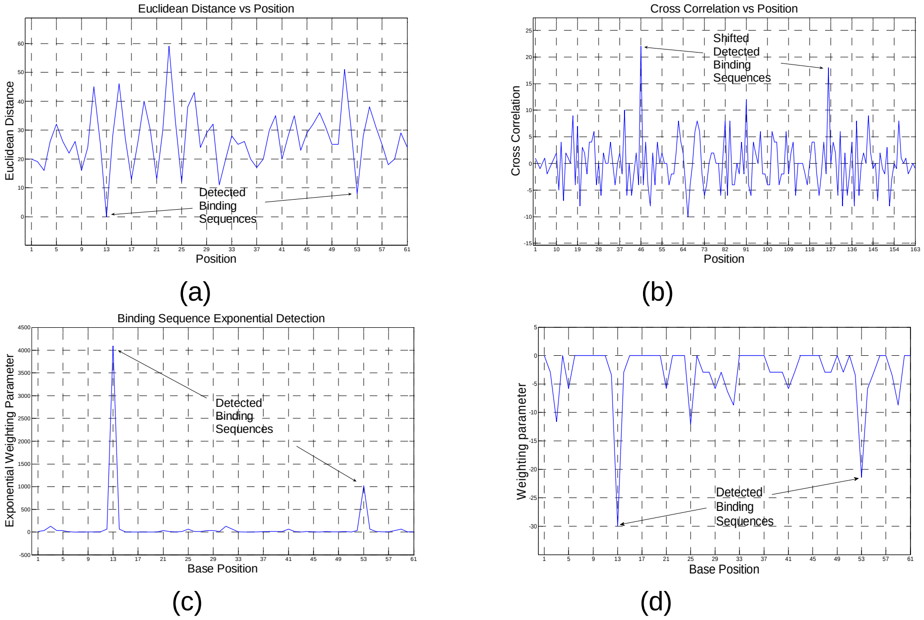
3.1.2 Communications and Information Theory Based Models

The previous model discussed in sec 3.1.1 is based on coding theory (codebook). We have also developed other four different methods (sec 4.1.2) for detection of transcription factor binding sequences, (TFBS). These methods are also based on concepts in communications and information theory such as correlation (method I), Euclidean distance (method II), matched filter (method III), and correlation based exponential metric (method IV). These and the previous method will be used to study the effects of mutations in different parts of the coding and non-coding regions.

To show how the four previous methods behave, we arbitrarily selected a 71-bases-long DNA sequence as a test sequence. Then, we chose an 11-bases-long sequence starting at position 13 to be a hypothetical binding sequence. We inserted this binding sequence at position 53 with two bases being changed to get a partial match of the original sequence. We applied the four previous methods to detect this binding sequence. Figures 6 show these methods are accurately detecting the binding sequence as expected. A total match occurs at position 13 (longer peak or dip), and a partial match occurs at position 53 (shorter peak or dip).

There are many different ways that DNA can be changed, resulting in different types of mutation. Examples include substitution, Insertion, deletion, and frameshift. In Figure 6 , we inserted a sequence of bases at position 53 to be detected later using the proposed methods. This can be viewed as an insertion type of mutation. The proposed methods were able to detect these sequences at their exact positions. Other mutations types will be analyzed using the proposed methods.

Substitution is a mutation that exchanges one base for another (that is, a change in a single "chemical letter" such as switching an A to a G). Insertions are mutations in which extra base pairs are inserted into a new place in the DNA. Frameshift: Since protein-coding DNA is divided into codons three bases long, insertions and deletions can alter a gene so that its message is no longer correctly parsed. These changes are called frameshifts.



(a) (b)

(c) (d)

Figure 6: The Four Proposed Methods Results (a) Euclidean Distance (b) Cross Correlation (c) Exponential detection, (d) Free Energy.

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 matching locations (as previous methods in sec 3.1.1, Figure 5), but also “interestingly” distinguishing coding from noncoding regions. This new finding suggests the last 13 bases of 16S rRNA molecule have a higher correlation with the coding regions. Preliminary results for method I (sec 4.1.2.) are shown in Figures 7 and 8 from which coding and noncoding regions can be identified. This interesting result will be further analyzed and researched using other binding sequences. We will need to define quantitative measures to analyze these results. Such measures can be based on the mean Square Error (MSE), correlation, Hamming or Euclidean distances. Overall, we will study how regulatory sequences correlate between themselves and the different segments of the noncoding and coding regions. We will also need to correlate these results with their biological significance.

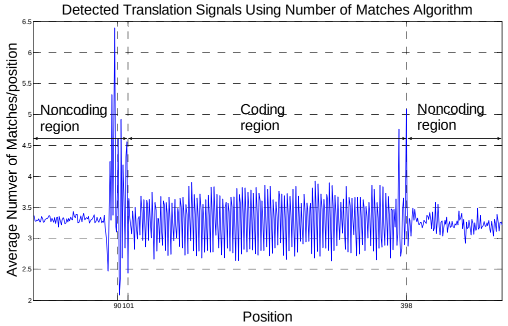


Figure 7: Method I (Number of Matches) Result

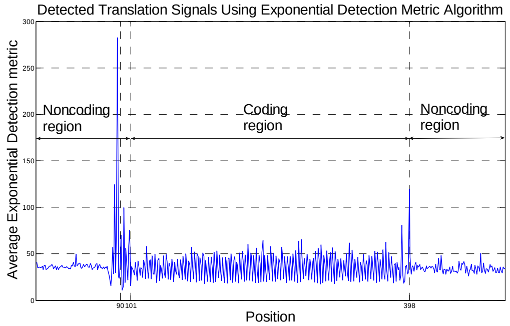


Figure 8: Exponential Detection Metric Results

Using the four developed methods, we will study the effect of mutations in last 13 bases of the 16S rRNA molecule on the correlations with coding and non-coding regions.

3.2 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 proven 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

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

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 9 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

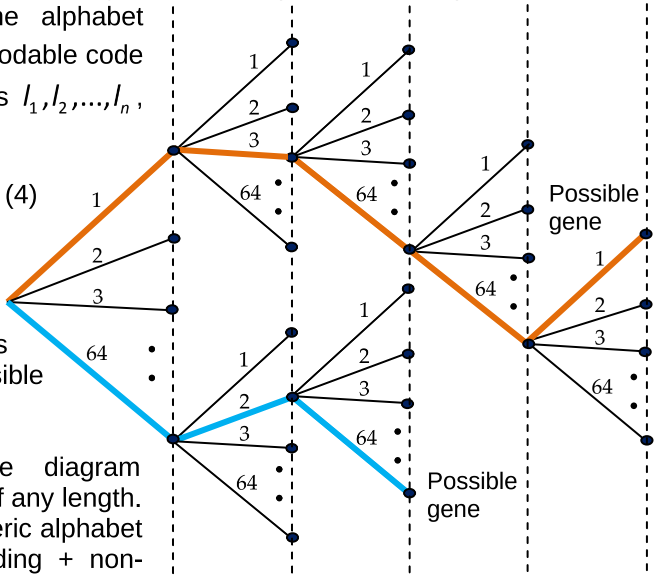


Figure 9: A tree representation of genes

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 9, the orange (upper) branch corresponds to the “gene code” {1, 3, 64, 64, 1}, and the blue branch corresponds to 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 [21] 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).

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 [22]) or reaching a leaf node first and backtracking (depth-first search [23]). Other examples of tree-searches include iterative-deepening search [24], depth-limited search [24], bidirectional search [25], and uniform-cost search [24]. 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).

3.3 Mutation Analysis

In our preliminary 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 the protein synthesis under mutations in the 3’ end of the 16S rRNA, was done by Hui and De Boer [9]. These two mutations were also tested using our proposed model and results (Figures 10 and 11) totally matched laboratory experimentation as well [1] [2]. This in turn certifies the correctness and the biological relevance of our proposed model.

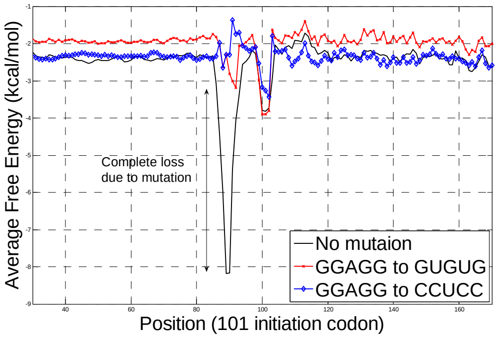


Figure 10: Hui and De Boer mutation

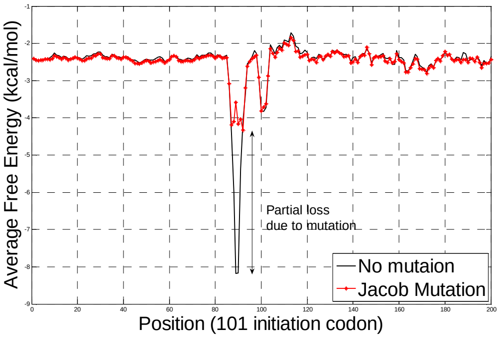


Figure 11: Jacob mutation

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

4 Research Design and Methods

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 12 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 [7].

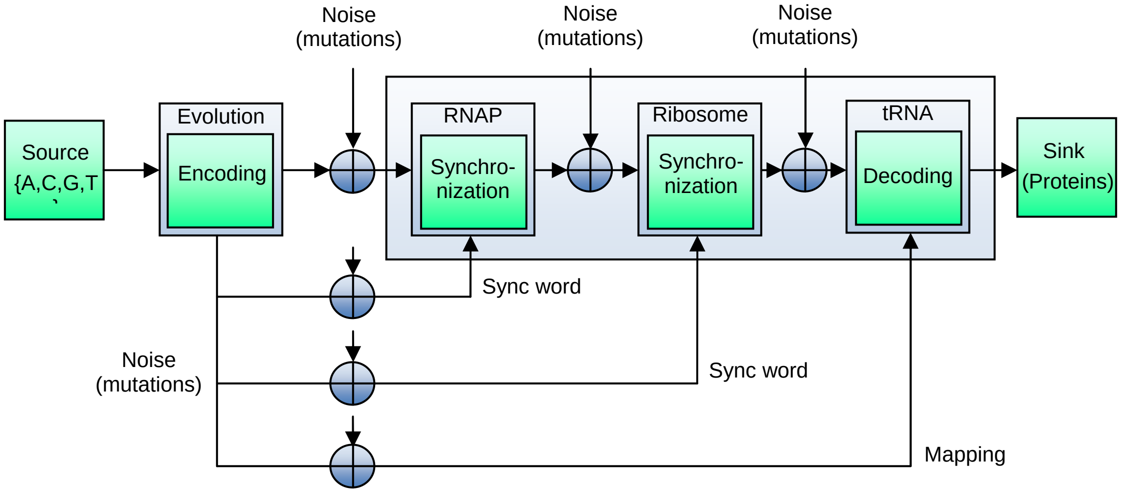


Figure 12: Communication theory model for gene expression

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 [7] established the first concrete ideas for modeling gene expression interactions based on algorithms inspired from coding theory [27] [28] [29].

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 3: 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 a larger alphabet)	quaternary mRNA sequence (can be 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 (Specific Aims) with a special emphasis on goals ix and x. The following sections will describe our research and design methods that are going to be considered in this work.

3.4 Coding Theory, Communications and Information Theory Based Modeling

3.4.1 Coding Theory Based Modeling

Our research is directed to use the models developed in our preliminary work and variations of them 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.
3. Studying the relationships between coding and noncoding regions and regulatory sequences

3.4.2 Communications and Information Theory Based Modeling

The process of detecting a Transcription Factor Binding Sequence (TFBS) 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 four methods are based on these concepts:

Method I: Euclidean Distance Based Algorithm

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

1. Map both DNA sequence and the binding sequence under study to their equivalent numerical quaternary representations using (A = 0, C = 1, G = 2, and T = 3).
2. Slide the binding sequence along the DNA 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. Hence, this method is able to detect the binding sequences in their exact location and accounts for gabs (mismatches as well).

Method II: 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. Method III can be done using a matched filter of an impulse response equal to y(-n) and an input of x(n) (y(n) is the binding sequence and x(n) is the DNA sequence) as follows (see Figure 13):

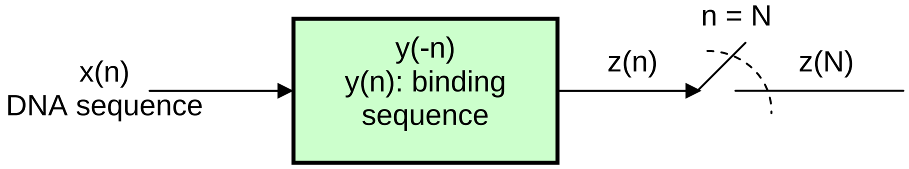


Figure 13: Matched Filter

1. Map both the DNA 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) for a better correlation form.
3. Correlate both sequences using

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

   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 position (a maximal point in the plot) to their corresponding locations in the original DNA sequence using

$$\begin{Bmatrix}\text{Detected Position}\\ \text{In the DNA sequence}\end{Bmatrix} = \left\lceil \left(\begin{Bmatrix}\text{Detected position}\\ \text{in the Plot}\end{Bmatrix} - \begin{Bmatrix}\text{length of the}\\ \text{binding sequence}\end{Bmatrix} + 1 \right) \Big/ 2 \right\rceil$$

   Where $\lceil X \rceil$ rounds the value X to the nearest integer larger than X.

Method III: 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 ith alignment, compute an exponential weighting function ($W(i)$) using the equations:

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

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:

$$w(n)=\begin{cases} a^{\sigma} & if\ \delta(n)=1 \\ 0 & if\ \delta(n)=0 \end{cases},\ \delta(n)=\begin{cases} 1, & if\ match \\ 0, & if\ mismatch \end{cases}$$

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 $\overline{W}$:
   $\overline{W}=\left[w(1),w(2),...,w(L-N+1)\right]$,
   where L is the length of the DNA sequence under study.
4. Plot the weighting vector $\overline{W}$, and detect peaks.

Model IV: 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:

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

(6)

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:

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

(7)

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.

The four previous models can be modified to utilize the energy table given in table 2 as well.

3.5 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 9) 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 9. 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.

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. [23]). The specific method to be used in the different cases will be adapted depending on the general characteristics of the organisms under study.

3.6 Mutation Analysis

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.

3.7 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 Timetable

During the first year, work will be directed to study prokaryotic genomes using *E. coil* as a test organism to validate the system. We will model the genome structure using models in sec 4.1 and 4.2. This will involve identification of (1) consensus sequences for promoters i.e. transcriptional start sites, (2) translational signals such as Shine-Dalgarno and S1 protein ribosome binding sites; as well as terminators, (3) ncRNA genes. 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. The detection of the genes and regulatory elements (REs) will be done using an iterative decoding algorithm analogous to turbo decoders. In the following years, work will be expanded to eukaryotic organisms, namely humans. This will be a substantially more complex task than in prokaryotes. It will require building a data base of all known promoters and TF binding sites. Work will involve using the previous methods and developing algorithms based on pattern recognition, Discrete Fourier Transform (DFT), and wavelet analysis. Moreover, we will develop computational algorithms and databases for systematic identification of transcriptional regulators and regulons in new organisms; and integrate genome expression data with known and predicted regulons and metabolic pathways. Throughout our work we will integrate our research with various educational and extension activities paying special attention to the goals ix and x described in section 1.