# Modeling and Analysis of ChIP-on-Chip data using Estimation and Detection Theory

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, detection and estimation theory, 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.

Estimation and detection theory has proven to provide powerful tools for the analysis of biological 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 $\overline{X} = \{A, C, G, T\}$. 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 estimation and detection 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 new methods by combining estimation and detection techniques used in communications and ChIP-on-chip (integrating chromatin immunoprecipitation (*"ChIP"*) with microarray technology (*"chip"*)) techniques to identify the Transcription Factor Binding Sites (TFBS) and related regulatory sequences in the genomes. Find models for prokaryotic and in the future eukaryotic organisms that represent the genetic and molecular mechanisms that 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) introduce an improved gene and regulatory sequences identification approach that will provide a solution for current limitations that exist in TFBS identification algorithms by using pattern recognition [9], Discrete Fourier Transform (DFT) [10], Wavelet analysis [11], Maximum Likelihood Sequence Estimation (MLSE), etc.;

iv) develop new computational algorithms and databases for systematic identification of transcriptional regulators and regulons in new genomes as they become available [12]; and integrate genome expression data with known and predicted regulons and metabolic pathways;

v) use the proposed models to detect variations of conserved sequence families and consensus sequences;

vi) apply and extend the proposed models to different prokaryotic and in the future to 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 [13-17]. The roadmap could be *E. coli*, listeria, yeast and eukaryotic strains in the future;

vii) 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 NIH's goals to develop, maintain, and renew scientific human and physical resources that will assure the Nation's capability to prevent disease. 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 and other Institutions; (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; and ECE 597 Special Projects taught by Dr. Atkin (3) Develop a new interdisciplinary course “Computational Biology and Bioinformatics” for senior undergraduates and entry-level graduates in ECE, Biology, CS and BME majors; (4) Encourage the participation of minority undergraduate and graduate students in ECE, BME and Biology majors thru Research Projects; (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.

viii) encourage the participation of students (women/men) from underrepresented and minority groups, and people with disabilities in our educational and extension programs (in Spring 2009 we had in our group 3 female students and new ones are joining in the Fall 2009. Another one graduated in the fall 2008; this is a quite significant number and represents more than 80% of the female student population in the ECE Department).

The main trust of this research is not only identifying the TFBS in the genome but also the analysis of the various interactions that take place in gene expression using communications models. It will contribute tremendously to the approaches on the whole genome analysis of histone modification patterns and to our understanding of the histone code and epigenetic. This will lead to a better understanding of these complex processes and will allow savings in laboratory resources and time-consuming laboratory experimentations.

## 2 Background and Significance

A ChIP-on-chip experiment can be divided into three major stages [18,19]: The first is to select the appropriate array and probe type. Second, the wet-lab experiment is performed. Last, during the dry-lab portion of the cycle, the data is gathered and analyzed to roughly locate the TFBS so that the cycle can start again to identify the accurate location of TFBS. Here we briefly describe the process of ChIP-on-chip that we will use in our research.

### 2.1 Wet-lab portion of the work flow for ChIP-on-chip technique

The procedure of wet-lab portion of ChIP-on-chip experiment is shown in figure 2.

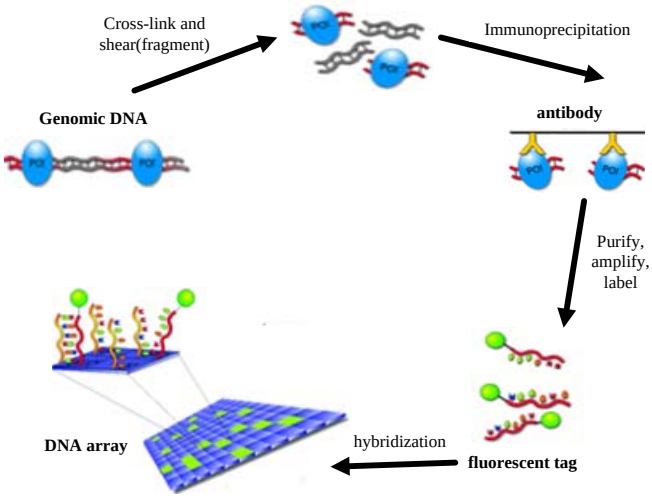


Figure 2: Wet-lab portion [18]

- In the first step, usually by a gentle formaldehyde fixation that is reversible with heat, the protein of interest (POI) is cross-linked with the DNA site, which it would bind to in an *in vivo* environment. Then, the cells are lysed and the DNA is sheared (fragmented) by sonication or micrococcal nuclease. This procedure results in double-stranded chunks of DNA fragments, normally 1 kb or less in length. Among these chunks, those cross-linked to the POI form a POI-DNA complex.
- In the next step, these POI-DNA complexes are filtered out of the DNA fragments, using an antibody specific to the POI. The antibodies attached to a solid surface (by having a magnetic bead, or some other physical property) bind the POI and allow isolating cross-linked complexes from unbound fragments. This procedure is essentially an immunoprecipitation (IP) of the tagged protein with an antibody against the tag (for example, FLAG, HA, c-myc). There an alternative

ways to implement this filtering step: affinity purification that does not require antibodies, for example, the Tandem Affinity Purification (TAP).

- After amplification and denaturation, the single-stranded DNA fragments are labeled with a fluorescent tag such as Cy5 or Alexa 647. The cross-linked of POI-DNA complexes are reversed and the DNA are purified. The POI is not necessary any more.
- Finally, the fragments are poured over the surface of the DNA microarray which is spotted with probes. The probes of the microarray are short, single-stranded sequences that cover the genomic portion of interest. Whenever a labeled fragment meets a complementary fragment on the array, they will hybridize and form a double-stranded DNA fragment.

## 2.2 Dry-lab portion of the work flow

After sufficient hybridization, the array is illuminated with fluorescence light. For the probes on the array, those that are hybridized to one of the labeled fragments will emit a light signal which can be captured by a camera. This image contains all raw data encoded as false-color image, and needs to be converted to numerical values before the analysis algorithm can be used. The analysis and information extraction of the raw data often remains the most challenging part for ChIP-on-chip experiments. Generally, the dry-lab portion, as shown in figure 3, can be divided into three major steps:

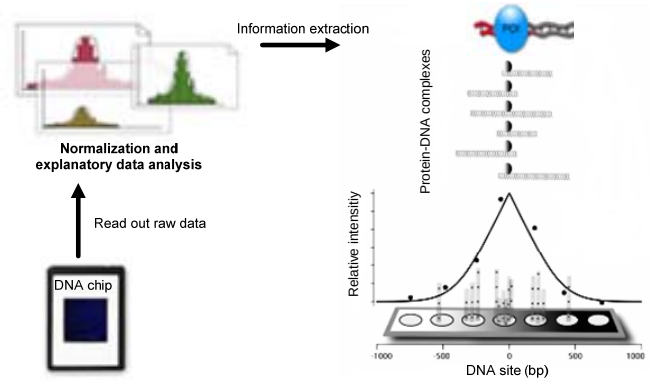


Figure 3: Dry-lab portion [18]

- During the first step, the captured fluorescence signals from the array are normalized by control signals which can be derived from the same or another chip. The control signals act like thresholds which decide whether the probes on the array were hybridized correctly.
- In the second step, statistical detection and estimation methods are applied to the normalized data to identify POI-enriched regions along the genome. Sometimes, these methods, such as Joint Binding Deconvolution (JBD), use a model for peak profiles derived from experimentally-measured DNA fragment, and require several user-defined prior parameters.
- In the third step, these regions are analyzed further. For example, functional annotation of the genome.

Various problems arise throughout the dry-lab portion, ranging from the suitable approaches to cancel the background noise to reasonable algorithms that normalize the data and make it available for statistical analysis [20-23]. For example, the DNA fragments are of variable length, and often far longer than the spacing of the microarray probes. Thus, each binding event influences the intensities of multiple probes adjacent to the one containing the binding site (TFBS).

In our work, we propose a filter bank scheme and several filter models that can be used to enhance the statistical analysis of ChIP-on-chip experiment data. Preliminary results shows that in prokaryotes these models can effectively identify the TFBS related signals in the genome sequence. In the future we will develop similar algorithms for eukaryotes that will improve the performance of ChIP-on-chip technology in case of mammalian genomes and make high resolution whole mammalian genome maps achievable.

## 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 [24-43]. 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.

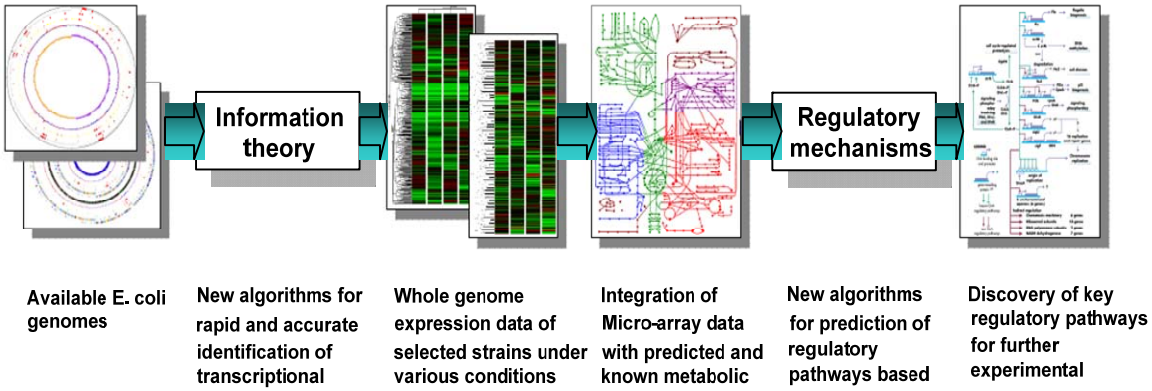


Figure 4: An outline of our computational approach

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 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 [13, 14] 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 [15, 16]. Due to lack of sequence data, previous systems (e.g. [14, 17] 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, it 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 [44], 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

In the future we will expand this to eukaryotic organisms. 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 Research Design and Methods

Analyzing DNA processing in gene expression and chromatin immunoprecipitation("ChIP"), many similarities with the way engineers process 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, such as mutations, and the interference introduced by instruments. 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. Other similar models for gene expression are summarized in [45]. In a similar way, since many steps in wet-lab portion, such as hybridization, are subject to and influenced by external conditions, we can regard the wet-lab portion of ChIP-on-Chip experiment as a random process, where the interested information is interfered by "noise". Also, the probes and light signal detector (camera) can bring about background noise. So, this procedure can be modeled as a channel in a communication system, we can adapt the detection and estimation methods used in communication to identify TFBS related signals and histone modification patterns in the genome. Figure 5 shows a model for ChIP-on-chip experiment based on communications theory approach.

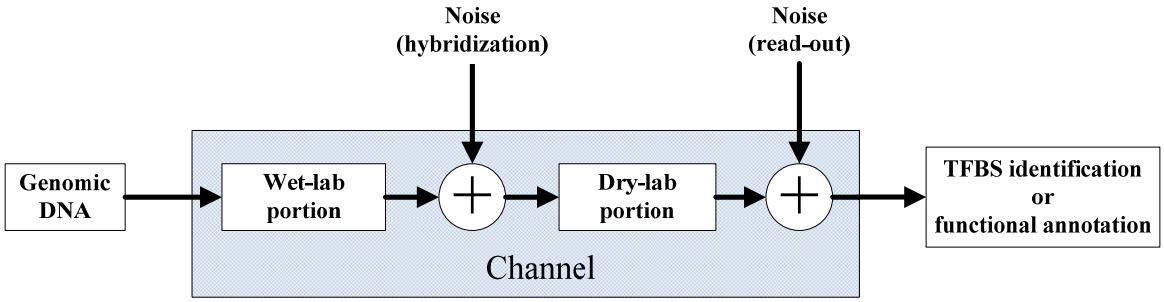


Figure 5: Communication theory model for gene expression

The noise in Figure 5 results in the background noise in ChIP-on-chip data, it is comparable to the AWGN noise in communication channel. On the other hand, Most ChIP-on-chip protocols utilize sonication as a method of breaking up DNA into small pieces. However, sonication is limited to a minimal fragment size of 200 bp. In order for higher resolution maps, this limitation should be overcome to achieve smaller fragments, preferably to single nucleosome resolution. To take advantage of the varied length of protein-DNA complexes (fragment size), we propose a multi-rate filter bank scheme for data analysis in ChIP-on-chip experiment. The multi-rate filter bank is widely used in communication systems to process signals with different resolutions [46]. It turned out to be a very effective way to extract information from noise. The system block diagram for our scheme is shown in Figure 6.

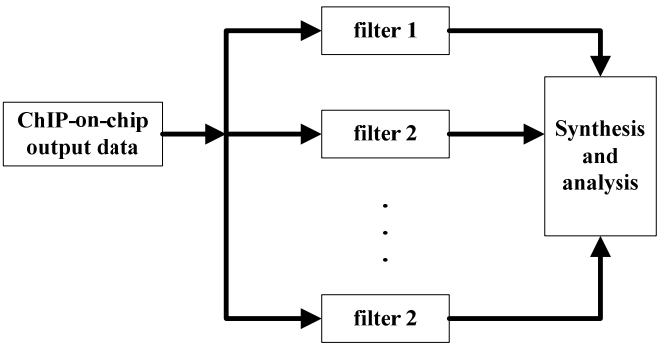


Figure 6: System block diagram for the multi-rate filter bank analysis scheme

What should be noted is that the "filters" used in our scheme are not limited to typical filters, we can generalize them to a wide range of algorithms. These algorithms can be divided into two broad classes: Coding Theory and Communication Based algorithms and Probabilistic algorithms. For

Coding Theory and Communication Based algorithms, we have developed 4 types of models. The filter compares the real input to the expected input based on those models, and output a scoring. Without loss of generality, we can define a scoring function $S(x,y)$, where x is the expected input and y is the real input. According to these scoring given by the filter bank, the synthesis and analysis module can determine the location of the TFBS related signals and histone modification patterns.

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

Table 1: Comparison of Frame Synchronization and Bacterial Transcription and Translation Initiation

| | **Digital Communications** | **Transcription Initiation** |
|---|---|---|
| Data | binary, quaternary or larger alphabet data streams | quaternary DNA sequence (can be a larger alphabet) |
| Marker | binary or quaternary synchronization word | two quaternary promoter regions |
| Detection | Correlator | sigma subunit of RNAP |
| Decision Criteria | correlation between sync word and data | binding energy between sigma factor and DNA |

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 will be considered in this work.

### 3.1 Coding Theory, Communications Based Algorithm

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

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 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 7):

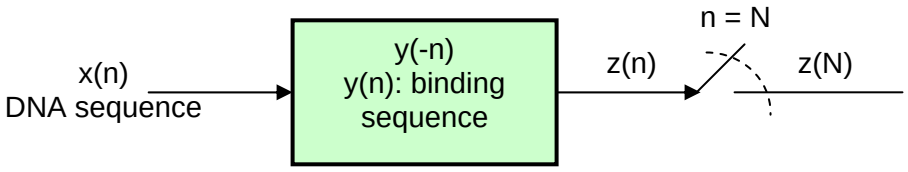


Figure 7: Matched Filter

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

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

   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

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

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 i[th] 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:

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

   where $a$ is an input parameter that controls the exponential growth of the weighting function, and $\sigma$ is the number of matches at each alignment. .
3. Repeat step 2 for all alignments along the DNA sequence to get the weighting vector $\overline{W}$:
   $\overline{W} = [w(1), w(2), \ldots, w(L-N+1)]$,
   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:

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

   (6)

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

   $E(y_n y_{n+1})$ 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:

   $$\delta(n) = \begin{cases} 1, & \text{if } y_n y_{n+1} = x_n x_{n+1} \ (match) \\ 0, & \text{if } y_n y_{n+1} \neq x_n x_{n+1} \ (mismatch) \end{cases} \qquad (7)$$

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 3 as well.

### 3.2 Probabilistic algorithm

The $\chi^2$ distance is originated from correspondence analysis [47]. It is a distance between the statistical profiles of two different sequences or sets. A vector is called a profile when it is composed of numbers greater or equal to zero whose sum is equal to one (such a vector is sometimes called a probabilistic vector). An approximate estimation of the $\chi^2$ distance between the input sequence and the (Center Of Gravity ) COG for the conserved sequence family is [48]

$$d_g^2 \cong \sum_{i=1}^{M} r_i d^2(i) \cong \frac{1}{M}\sum_{i=1}^{M} d^2(i)$$

Where $i \in$ {index of conserved sequences}. Since the index for input sequence can be fixed to M+1, this distance is independent of the index scheme, and can be denoted as $d_g^2$ instead of $d_g^2(M+1)$. Then, the normalized $\chi^2$ distance from the input sequence to the COG of the conserved sequences can be defined by

$$D(n) = A \cdot d_g^2(n) \cong A\sum_{i=1}^{M} r_i d^2(i) = \frac{(M+1)\cdot L}{N'}\sum_{i=1}^{M} r_i d^2(i) \cong \frac{N}{M \cdot N'}\sum_{i=1}^{M} d^2(i)$$

where n denotes the index of the location on the input nucleotide sequence.
To define the dynamic range of the $\chi^2$ distance $D(n)$ for the input nucleotide sequence, we must evaluate lower and upper thresholds based on the conserved sequences. The upper threshold can be obtained with

$Th_{upper} = \max\{D_j(0)\}$, $j \in$ {index of conserved sequences}

The lower threshold is

$Th_{lower} = \min\{D_j(0)\}$, $j \in$ {index of conserved sequences}

The output of the lower threshold function can be defined by

$$T(n) = \begin{cases} D(n), & Th_{lower} \le D(n) \le Th_{upper} \\ Th_{upper}, & \text{otherwise} \end{cases}$$

The metric function is defined as $M(n) = \frac{Th_{upper}}{T(n)} - 1$

The identification of TFBS is based on the peak detection on the value of M(n).

### 3.3 Application and Extension to other Organisms

The proposed models will be extended to other prokaryotic and in the future to 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-on-chip experiments.

## 4 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 [1-4, 46, 48].

### 4.1 Coding Theory and Communications Theory Based Modeling

#### 4.1.1 Coding Theory Based Models

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. 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 3). 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 [49] 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 $\delta_k$ 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 $\sigma$ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches and $\rho$ is an offset variable updated as follows

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

Table 2. 16SrRNA Codebook

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

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

| 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 3. Energy Doublets [17]

For larger values of *a* 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 8 shows average results for the detection of the SD, start and stop codons being compared to previous work [6]; 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.

### 4.1.2 Communications and Coding 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 Coding 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.

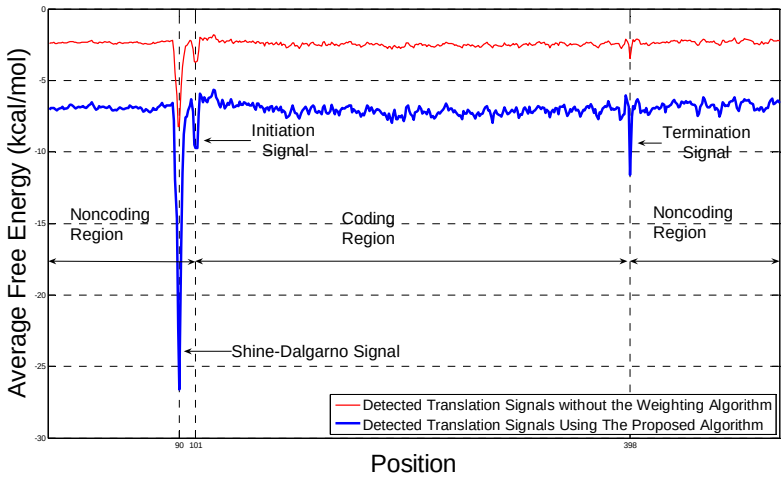


Figure 8: Comparison of SD signal (position 90), start (position 101) and termination (position 398) codon between the algorithm used in [8] and the weighted algorithm (N=5, a=1.5)

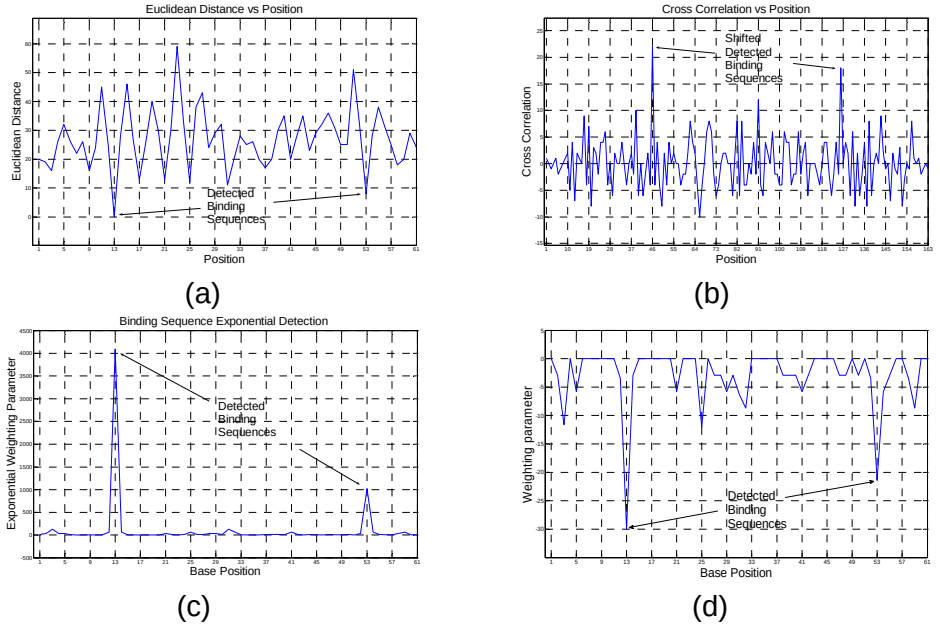


Figure 9: The Four Proposed Methods Results (a) Euclidean Distance (b) Cross Correlation (c) Exponential detection, (d) Free Energy.

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 9, 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.

### 4.2 Probabilistic Modeling

The $\chi^2$ distance approach is more efficient than traditional method based on log-linear model, such as FMM [50], because it waives the need to evaluate highly complicated log likelihood function and objective function [51]. Since this new approach [48] takes into account the position- dependence of TF motifs in computing $\chi^2$ distance, it also brings about significant performance improvement over the PWM model method.

The conserved sequences table used to locate promoters in E. coli sequences is taken from the compilation of such sequences produced by Hawley and McClure [52]. E. coli promoters have been shown to contain 2 regions of conserved sequence located about 10 and 35 bases upstream of the transcription start-site. Their consensuses are TATAAT and TTGACA with an allowed spacing of 15 to 21 bases between. The spacing with maximum probability is 17 bases and all but 12 of the 112 sequences in the Hawley and McClure collection could be aligned with a separation of 17 + or -1 bases. The spacing between the -10 region and the start-site is usually 6 or 7 bases but varies between 4 and 8 bases. Hawley and McClure also show a conserved section to exist around the +1 region. The range definitions for the three regions (the -35, -10 and +1 regions) are in [53]. The input nucleotide sequence (genome sequence) used in simulation can be found in [54].

First, we use the -10 region of the conserved sequence to identify the -10 promoter, the result is shown in Figure 10. The red region marks the real location of the -10 promoter, the center of red region is at 101. It can be observed that the highest peak locates at 125, which is 24 bases from the center of red region; therefore, the predicted location of the -10 promoter should be 115 and the identification error is 14 bases. In -35 promoter identification, we use -35 region of the conserved sequences to form the conserved sequence table, the result is described in the Figure 11.

Similarly, the red region indicates the real location of -35 promoter, it can be discovered that the highest peak appears around the location of 62, so the predicted location for -35 promoter should be 27. Since the center of the red region is at 78, the identification error is 51 bases. However, if we consider the secondly highest peak, which locates at 121 and 43 bases away from the center of the red region, the predicted location is 86 and the identification error is mere 8 bases. In fact, there is a restriction enzyme Taq-I recognition site at location 63, that might be responsible for the highest peak. We are currently researching this interesting result that points out to the ability of the algorithm to identify other signal sequences.

By using both -35 and -10 region of the conserved sequences to form the conserved sequence table, we obtain a more accurate identification result as shown in Figure 12. The red region indicates the real location of the -35 promoter and the green region indicates the -10 promoter. In the same way,

with the highest peak located at 116, it can be discovered that the identification errors for -35 and -10 promoter are 3 and 5 respectively. They are far smaller than those corresponding identification errors in the two previous cases. This is due to the use of longer conserved sequences. The use of longer conserved sequences provides more reliable statistical information, thus a better identification performance.

As a comparison, similar results are described in Figure 5c in [53]. In case of identifying -10 promoter with the PWM model method, since the peak around the real location, which is 111, is not higher than the subsidiary peak 40 base-pairs upstream, the identification result of -10 promoter must combine with that of -35 promoter to achieve the real location. It involves a complex procedure of optimization and the choice of certain criteria. However, our proposed approach does not require any optimization operation, and it is faster and more accurate. Current research is considering other organism and comparisons with the many available identification methods.

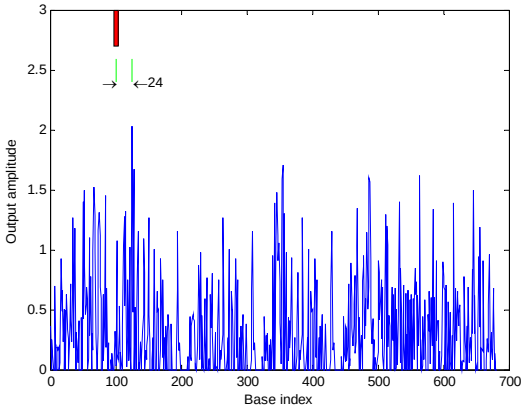


Figure 10. -10 promotor identification

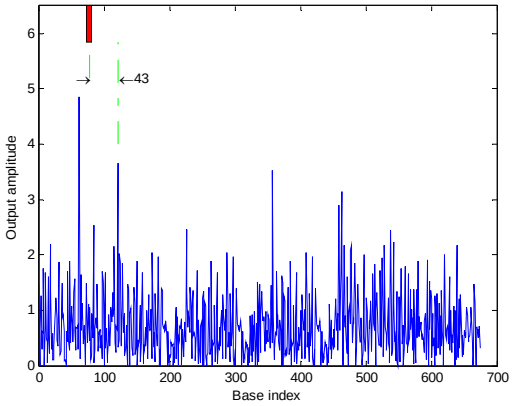


Figure 11. -35 promotor identification

## 5 Timetable

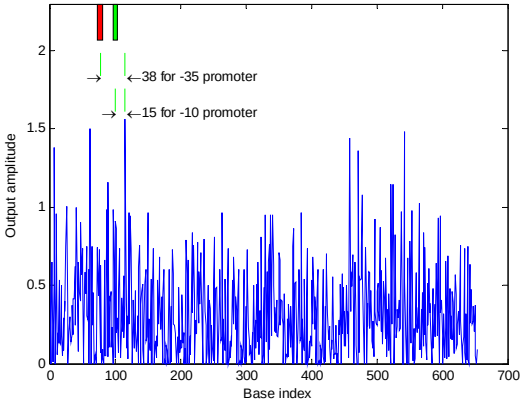


Figure 12. -35 and -10 promotors identification

During the first stage, work will be directed to study prokaryotic genomes using *E. coil* as a test organism to validate the system. We will determine scoring function $S(x,y)$ using models in section 3. This will involve (1) identification and prediction of TFBS and related sequence based on our proposed models and multi-rate filter bank (2) develop the background noise model and normalization algorithm for the captured fluorescence signal data from the array, (3) design the scoring function $S(x,y)$ based on MLSE [55, 56] or other criteria. The relative organization of these signals will then be used to detect specific putative regulatory sequences. The detection of the genes and regulatory elements (REs) will be done using an iterative decoding algorithm analogous to turbo decoders. In later stages, work will be expanded to other organisms (yeast, and establish the basis to study eukaryotes). 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.