Chapter 1

Introduction

A living organism has innumerable cells, each containing a set of genes made of deoxyribonucleic acid (DNA). DNA is a polymer in which the monomeric subunits are four distinct nucleotides Adenine (A), Guanine (G), cytosine (C) and Thymine (T). The central principle (dogma) of biology states that the translation of gene into a three dimensional protein structure leads to the synthesis of proteins which are essential ingredients in any life building and life sustaining processes. The first step in this process starts with replicating the copy of a gene, that is transcribing the gene, from DNA onto RNA by RNA polymerase. Splicing of introns from RNA, in case of eukaryotes, and creating messengerRNA (mRNA) is the intermediate step in the process. Second step is the translation of mRNA into an amino acid sequence which folds into a three dimensional protein structure as illustrated in Figure 1.1. For the transcription to happen, RNA polymerase has to bind to the promoter, which occurs upstream of a gene. Promoters can function not only to bind RNA polymerase, but also can specify the places and times at which transcription can occur from that gene. Promoters of genes that transcribe relatively large amounts of mRNA have a set of binding sites/regions [39, 76]. One of these sites is a TATA sequence, a hexamer, upstream from the site where transcription begins and this location is known as the transcription start site (TSS). Promoter also contains one or more binding regions further upstream and also downstream as shown in Figure 1.2 (capital letters indicate conserved subsequences). Figure 1.2 depicts the promoter structure for E.coli with two binding regions present at -10 and -35 positions with respect to TSS (position of which is taken as +1). These are indicated as -35 motif and -10 motif there. The detailed structure of a promoter, the role it plays in gene transcription via
1.1 Promoter Recognition

Recently, there has been a deluge of sequencing information due to efficient sequencing methods. Several mammalian, bacterial and plant species have been sequenced. One can use experimental methods such as DNA footprinting, DNA protein crosslinking, X-ray crystallography and NMR spectroscopy to identify a promoter or a gene. Typically, there are millions of protein sequences, but
1.1. Promoter Recognition

Figure 1.2: *E. coli* promoter structure [36]

experimentally determined protein structures are only of the order of thousand. Experimental methods to determine a promoter, a gene or a protein structure are time consuming processes. Hence, annotation of important regions such as genes is not very fast. To overcome this handicap, techniques or algorithms that can automatically identify these regions are required. In order to identify a gene, either gene recognition methods can be used directly, or a promoter can be identified and used for gene recognition indirectly.

1.1.1 Need for promoter prediction

Prediction of a promoter has a variety of applications. One of them could be gene regulation. Regulation of expression of a gene occurs at various stages and places, along the pathway from genome to proteome. Regulation at promoter site is the most important way of regulating the gene expression since transcription occurs only when RNA polymerase binds to the promoter. Transcription can be inhibited or enhanced by the binding of certain transcription binding factors which can bind to a promoter. Hence, through a promoter a gene can be regulated. Prediction of a promoter can also be used for gene prediction since a promoter precedes a gene. Genes which are functionally coupled can have promoters of similar structures. This fact can be used to find genes co-regulated by a promoter through promoter prediction. Extending this idea further, the promoter may also be used to give clues as to the function of a completely unknown protein.
1.1.2 Complexity of the problem

Promoter recognition is not trivial due to several reasons. Promoter recognition unlike other recognition problems such as exon prediction and gene recognition, does not yield good results with methods of alignment or sequence similarity searches since they have very low sequence similarity. In general, patterns (TATA box, CAAT box, Initiator etc.) in the promoter sequences within a species and across species in some cases are known to be conserved. But there exist many exceptions to this rule such as the presence or absence of a particular region (TATA box) that makes the promoter recognition a difficult problem. Also the occurrence of a promoter is not restricted to 5’ end of a gene alone, but could in fact be found in an exon, intron, untranslated region of 3’, or overlap with another promoter [40]. Hence the problem of recognition of promoter against various backgrounds gains importance computationally. That is to say that a promoter which behaves like a coding region, when it occurs in the coding region and behaves like a non-coding region when it occurs in a non-coding region but still retains its special characteristics has to be extracted against these backgrounds. In addition, spacing between the patterns, presence or absence of the patterns, non-conservation of the patterns in a promoter make the task of promoter prediction an even more complex problem. To cap it all there could be several promoters for a gene and a promoter can have many TSSs located closely [97].

The distinct feature in case of eukaryotic transcription is that the RNA polymerase do not bind to the promoter directly. A number of transcription binding proteins come and bind to the binding sites and form a complex before RNA polymerase binds. And also, there are three kinds of RNA polymerase in eukaryotes unlike the prokaryotes. For the proteins to bind to DNA, the DNA has a physical structure wherein the proteins can come and bind. Special proteins that are used for this purpose are Helix turn Helix, and Zn$^{++}$ fingers. The generalization of promoter prediction becomes a non-trivial process, because of these factors.

The crux of the problem is to identify a promoter irrespective of place of occurrence in genome, by extracting features that are unique to it. Different research groups have been trying to identify these patterns or features specific
for promoters by various feature extraction methods and different classifiers.

1.1.3 Approaches used so far

Machine learning techniques can be used to address the issues mentioned above by modeling the recognition/prediction problem as a pattern recognition problem. To properly classify the promoter sequences \textit{in-silico}, one should get features which capture the essence of promoters. Promoter prediction/recognition methods can be broadly categorized into groups such as genetic algorithms, statistical models such as hidden Markov models, position weight matrices, syntactic recognition algorithms, automatic motif discovery methods, neural networks.

There are many techniques from those cited above, which deal with only specific regions such as binding regions/sites (deemed as crucial) in the promoter. These methods can be categorized as local signal methods, since they do not use whole promoter but the binding sites. That is, techniques that are based upon the features extracted from the binding sites alone can be categorized as local signal based methods. Position weight matrices, expectation and maximization algorithm, hidden Markov models etc fall under this category. In contrast, techniques that use whole promoter sequence can be termed as global signal based methods. In global signal based approaches, features are extracted from the whole promoter sequence. Methods like Fourier transform, sequence alignment etc come under this category. The next section gives an overview of the modeling process, data sets that are used and the general flow of the classification technique used in the thesis.

1.1.4 A few drawbacks of the existing methods

We can identify some drawbacks that are there in the existing methods as follows. It was pointed out earlier that all promoters may not contain all the binding regions, hence the methods that specifically use information from these regions will not have good recognition rates. One more factor is that not all the promoter prediction methods have been tested on promoter prediction performance on genome-wide experiments. That is, it is not clear how some of the existing methods could be extended to whole genome promoter prediction and their performance results are also not known. In practice, it is not feasible to obtain a
representative set of promoters and non-promoters (experimentally verified) using which pattern recognition algorithms can be designed and subsequently used to classify the promoters in the general context of whole genome.

1.2 Promoter Recognition as a Binary Classification Problem

In this thesis, promoter recognition is modeled as a binary classification problem. A promoter is taken as a segment of DNA sequence where a known Transcription start site exists, with a certain length before TSS (upstream) and a certain length of sequence after the TSS (downstream) as in Figure 1.2. Depending on the species the length of the sequence upstream can vary from 250 basepairs (bp) to 60 basepairs and the downstream subsequence can be 20 to 50 bp. For eukaryotes, it is generally believed that CAAT box exists around 200 bp upstream from TSS, and some promoter elements are present at 30 bp downstream from TSS. For prokaryotes, specifically *E.coli*, both the binding sites are within 60 bp upstream from a TSS. Various features are extracted by considering the whole promoter and they are given as input features to the binary classifier. Supervised training methods are utilized in this thesis for promoter prediction. The problem of promoter recognition as a binary classification problem can be formally defined as identifying a sequence $S$ of length $n$ with a known TSS as a promoter or as a non-promoter using a particular classifier. The classifier will output 1 if $S$ is a promoter or a 0 if it is a non-promoter. Each sequence is represented as a vector of size $m$, where $m$ represents the number of features. Various feature extraction schemes are proposed in this thesis in order to recognize a promoter against various backgrounds.

1.2.1 Data Sets

Experiments are carried out on two types of data set, one from the prokaryotic species (*E.coli*) and the other from the eukaryotic species (*Drosophila Melanogaster* (*Drosophila* for short). In the thesis, we take *E.coli* as a model organism and develop feature extraction and recognition schemes. We test these themes on *Drosophila* for the generality of proposed schemes.
1.2. Promoter Recognition as a Binary Classification Problem

Positive data set in case of *E.coli* is built by taking 669 promoter sequences of length 80 from RegulonDB and Promec data bases by Gordon et al. [36]. Promoter data set of Gordon et al. is considered as positive set [36]. There is no standard negative data set available. We consider negative data sets of Gordon et al. who have chosen these in a biologically meaningful way by taking sequence fragments outside the promoter region. They consider 709 sequence fragments from the coding region (coding) and 709 sequence segments from intergenic portions (non-coding) [36]. Sample data is given in Appendix C.

In case of *E.coli*, we also consider synthetic negative data set in Chapter 3. These are randomly generated sequences of length 80 bp consisting 60% A+T. And we also consider Harley’s experimentally determined *E.coli* data set to construct position weight matrices [41] in Chapter 4. They have identified the -35 and -10 motifs in this data set.

The promoter data set of *Drosophila Melanogaster* is obtained by Ohler et al. [86], from Eukaryotic promoter database (EPD) [30]. Negative data set is collected by them from the same genome. Sequences from both positive and negative data sets are of length 300 bp with 250 base pairs upstream of the Transcription Start Site (TSS) and the rest downstream. The data set contains 1864 promoter sequences, 2859 sequences from coding (cds) and 1799 sequences from intron portions [9].

1.2.2 Feature Extraction and Classification

Throughout the thesis, the general flow consists of two stages. First one is about extracting features using different techniques and the next one is about using these features as inputs to a neural network classifier. Data set is partitioned into training and test sets. A neural network is trained using these inputs and 5-fold cross-validation is used on the test data set. The output of the neural network is used to classify a given sequence as a promoter or a non-promoter. Same algorithm can be extended to recognize a promoter in a whole genome. All neural network simulations are carried out using Stuttgart Neural Network Simulator (SNNS) [104]. The following section lists a set of questions that were addressed in the thesis.
1.3 Questions Addressed in the Thesis

- How global signals extracted from 2-grams or n-grams are useful in promoter recognition? What are the advantages of global schemes over local schemes? (addressed in Chapter 3).

- Can the interaction between promoter and RNA polymerase be simulated through signal processing techniques? (addressed in Chapter 5).

- Can we combine local signals and prior structural data efficiently to identify a promoter? Is a global signal sufficient when prior structural data is not available? Can the same methodology be extended to promoters with unknown structural information? (addressed in Chapter 4).

- Can we understand the similarity/dissimilarity between promoters of a particular sigma unit using the features extracted? (addressed in all chapters)

1.4 Organization of the Thesis

The organization of the thesis is as follows.

Chapter 2 reviews various promoter recognition techniques and results available in literature. Main similarities and differences between promoters of prokaryotes and eukaryotes are pointed out here. This gives an insight in understanding the complexity of the problem in general and why it may be feasible or infeasible to extend the promoter recognition methods developed for prokaryotes to eukaryotes. Later part discusses the recognition/identification techniques that can be categorized into local motif recognition and contrasting global feature techniques that use whole promoter sequence. Under local signal recognition methods Expectation maximization algorithm, position weight matrices, neural networks, hidden Markov models are discussed. Global signal recognition methods are based upon physical, structural properties of promoters, Fourier transform, wavelet transform and sequence alignment. A set of metrics used to compare the results of these techniques are also discussed here.

Chapter 3 contains the introduction to n-gram features and extraction. Classification results obtained using various n-grams (n=2,3,4,5) as features to a neural network are given here. A scheme to recognize promoters using the best n-grams
in a genome sequence is also proposed in this chapter. Efficacy of this particular scheme is compared with other software tools in use such as Neural network promoter prediction (NNPP), Bacterial promoter prediction (BPROM) etc. In addition a detailed analysis of correctly classified and misclassified sequences is done using 2-grams.

Chapter 4 extends results obtained in chapter 3. A multi-level classifier is proposed as a complete classifier system to recognize a promoter. In addition, AdaBoost classifier is also tried in an attempt to enhance the results. Later part of the chapter is entirely devoted to position weight matrices for binding regions of a promoter as well as those constructed from the whole promoter sequence. Here, we try to verify whether the identification of binding sites is essential or not for good recognition rates. In order to achieve that goal, position weight matrices in local and global context are analyzed. An attempt is also made to apply position weight matrices for Drosophila promoter recognition where no structural features (location of binding regions) are available.

Chapter 5 is devoted to signal processing techniques. Here, two lines of thought are explored. First one is about the idea that the RNA polymerase binds to promoter using some kind of resonance formalism and is explored through wavelets in this chapter. Second one is about classification in frequency domain. Fast Fourier transform (FFT) and wavelet transform are used to verify this fact and extraction of features using Fourier transform (FFT) and wavelet transform and their suitability to promoter recognition is studied.

Chapter 6 presents the overall discussion of results obtained from the n-grams, position weight matrices and signal processing techniques. This chapter will also point out possible interpretation of our results. Future directions are also indicated.

Appendix A gives quick molecular biology primer on promoter, binding of RNA polymerase to promoter.

Appendix B gives links to various software packages for promoter recognition.

Appendix C gives view of typical promoter data provided by Dr. Leo Gordon.

Appendix D gives one of the typical sections of E.coli genome.

Appendix E gives the glossary for the biological terms used in this thesis.