

*Promoter Recognition using
Local and Global Features
-A Machine Learning Investigation*

A THESIS SUBMITTED
FOR THE DEGREE OF

Doctor of Philosophy

in

Computer Science

by

T. Sobha Rani



Department of Computer and Information Sciences
School of Mathematics and Computer/Information Sciences
University of Hyderabad
Hyderabad - 500 046
INDIA

August 2008



Department of Computer and Information Sciences
School of Mathematics and Computer/Information Sciences
University of Hyderabad, Hyderabad, India

CERTIFICATE

This is to certify that the thesis work entitled '**Promoter Recognition Using Local and Global Features-A Machine Learning Investigation**' being submitted by Ms.T. Sobha Rani (Reg. No. 03MCPC21) in total fulfillment of the requirement for the award of degree of **Doctor of Philosophy (Computer Science)** of University of Hyderabad, is a record of *bona fide* work carried out by her under our supervision. The matter embodied in this thesis has not been submitted for the award of any other research degree.

Dr.S. Bapi Raju
(Supervisor)

Department of Computer and Information Sciences
School of Mathematics and Computer/Information Sciences
University of Hyderabad, Hyderabad, India

Prof. Arun Agarwal
(Head)
Department of Computer and
Information Sciences
University of Hyderabad
Hyderabad, India

Prof T. Amarnath
(Dean)
School of Mathematics and
Computer/Information Sciences
University of Hyderabad
Hyderabad, India

Declaration

I, Sobha Rani, hereby declare that the work presented in this thesis has been carried out by me under the guidance of Dr. S. Bapi Raju, Department of Computer and Information Sciences, University of Hyderabad as per the PhD ordinances of the university. I declare, to the best of my knowledge, that no part of this thesis has been submitted for the award of a research degree of any other university.

(T. Sobha Rani)

Acknowledgements

I take this opportunity to thank all those who have directly or indirectly helped me in this thesis work.

First, I would like to express my gratitude to my supervisor, Dr.S. Bapi Raju for giving me the opportunity to work with him. He is the one who has introduced me to this exciting field of bioinformatics and guided me throughout in my work. I am very glad that I am working in this field.

I thank Prof. Arun Agarwal, Head, Department of Computer and Information Sciences (DCIS) for making available all the facilities required for this research work. I also thank former heads of department Prof. A.K.Pujari and Prof.H.Mohanty for their constant support. I thank Prof. Amarnath, Dean, School of Mathematics and Computer/Information Sciences (MCIS) for supporting me in my study leave to carry out this work.

I am ever grateful to Dr.S.Durga Bhavani for her help and support in doing this work. I thank all my other colleagues especially Dr.Atul Negi, Dr.B.Chakravarthy, Ms.Anupama for their constant support. I would like to thank our office staff and lab staff for the help they provided.

I would like to thank Dr.Leo Gordon, who has given us the promoter and non-promoter data.

Last, but not the least, my family, which has supported me tirelessly in my work. I would like to thank my husband, Dr.C.R.K.Reddy, children Abhijith and Samhitha, in-laws and parents without whose support I couldn't have done even an iota of this work.

August, 2008

Sobha Rani

Abstract

Genes which code for proteins are the most important segments in a genome. A promoter which occurs upstream of a gene acts as a switch in gene transcription. Promoter prediction/identification is helpful in identifying co-regulated genes, unknown function of a gene, gene regulation etc. Various factors contribute to the complexity of the problem of promoter identification. To identify a promoter, signal unique to the promoter is to be extracted from the promoter. Promoters are mostly identified by local motifs/consensus regions present in the promoter. In this thesis, attempt is made to identify a promoter using binary classification approach. The features to the classifier are extracted by considering either *whole promoter* (global) or *binding sites* (local motifs). These features are given as input to feed-forward neural network (NN) classifier. Two sets of data are used for this experimentation. One is *E.coli* and the other is *Drosophila Melanogaster*.

Different features, such as *n-grams* ($n=2,3,4,5$), position weight matrix based features, features extracted using Fourier transform and wavelet transform are explored here. The thesis discusses the extraction and usage of these features in classification of promoters. This approach has been extended in some cases for promoter identification in a whole genome. A study is also made of the interaction between promoter and RNA polymerase through the signal processing techniques. One main contribution from classification results obtained from *n-grams* is the development of whole genome promoter prediction methodology using best *n-gram* features. The results are very good and are better than the results of currently available prediction tools for the forward strand. And an in-depth analysis using 2-grams has given an insight about promoters and non-promoters. It points to the similarities between majority of promoters and a small non-promoter set. Similarly, similarities between a large non-promoter data set and a minor promoter data set are also observed. Experiments with position weight matrix features for genome-wide recognition of promoters reveal that the performance is superior with global features as compared to local features extracted from binding sites. Results of classification using signal processing techniques that use global features show that non-promoters can be identified well with these features. Promoter and RNA polymerase interaction has been modeled using wavelets. The experiments demonstrate that this approach is not well suited for promoter prediction but is very good for non-promoter identification.

It can be concluded from the results that from the viewpoint of classification, features extracted from the promoter as a whole are more representative of the characteristics of the promoter than those acquired from localized motifs. The regions that lie before, in-between and after the binding regions do contribute for the recognition of promoter as is evident from the results obtained using *n-grams* and position weight matrix features.

Table of Contents

Declaration	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
List of Tables	ix
List of Figures	xi
1 Introduction	1
1.1 Promoter Recognition	2
1.1.1 Need for promoter prediction	3
1.1.2 Complexity of the problem	4
1.1.3 Approaches used so far	5
1.1.4 A few drawbacks of the existing methods	5
1.2 Promoter Recognition as a Binary Classification Problem	6
1.2.1 Data Sets	6
1.2.2 Feature Extraction and Classification	7
1.3 Questions Addressed in the Thesis	8
1.4 Organization of the Thesis	8
2 Survey of Promoter Recognition methods	10
2.1 Promoter	10
2.1.1 Prokaryotic Promoters	11
2.1.2 Eukaryotic Promoters	12
2.2 Metrics used in Promoter Recognition	12
2.3 Promoter Recognition Methods	13
2.3.1 Local Signal Recognition Methods	14
2.3.2 Global Signal Recognition Methods	20
2.4 Existing Promoter Prediction Programs	27
2.5 Discussion and Summary	27

3	N-gram Analysis of <i>E.coli</i> and <i>Drosophila</i> Promoters	31
3.1	Introduction	31
3.2	Feature Extraction	33
3.2.1	Neural Network Architecture and Classification Performance for <i>E.coli</i> and <i>Drosophila</i>	36
3.2.2	Classification with Different Negative Data Sets for <i>E.coli</i> and <i>Drosophila</i>	38
3.3	Analysis of Correctly Classified and Misclassified Promoter Sequences	41
3.3.1	A Single Layer Perceptron for Promoter Recognition	45
3.4	Genome-Wide Promoter Recognition	47
3.4.1	A Preliminary study	47
3.4.2	Genome-Wide Promoter Recognition Using Neural Networks Based on Promoter-Coding and Promoter-Noncoding	49
3.5	Discussion	57
3.6	Summary	58
4	Cascaded (Multi-Level) and PWM Based Classification of <i>E.coli</i> Promoters	61
4.1	Multi-level Classification System	62
4.1.1	Unbalanced data method	62
4.1.2	Using a 2-level Multi-layer Feed-forward Network	67
4.2	Introduction to Committee Machines	68
4.2.1	AdaBoost Classifier	69
4.3	Position Weight Matrices (PWM)	70
4.4	PWM from Harley-Reynold's Data for <i>E.coli</i>	71
4.4.1	Mono-gram PWM of -35 and -10 binding sites from Harley's data	71
4.4.2	Bi-gram and tri-gram PWMs of -35 and -10 binding sites from Harley's data	72
4.4.3	PWMs of Binding Sites in Promoter Recognition of <i>E.coli</i>	75
4.4.4	Genome-Wide Promoter Recognition	77
4.4.5	PWM of Whole Promoter in Promoter Classification	78
4.5	PWMs for <i>Drosophila</i> Using Whole Promoter	81
4.6	Discussion and summary	82
5	Investigation of Signal Processing Methods for Promoter Recognition	84
5.1	Introduction to wavelets	86
5.2	Methods	88
5.2.1	Encoding and Decomposition	88
5.2.2	Feature Extraction	90
5.3	Classification	93
5.3.1	Classification Using FFT Coefficients	93

5.3.2	Classification using wavelet coefficients	96
5.3.3	Classification Using Decomposed Signals	97
5.3.4	Classification Using Cross-correlation Between Promoter and RNA-Polymerase	97
5.4	Discussion	101
5.5	Summary	104
6	Conclusions and Future directions	106
6.1	Conclusions	106
6.2	Future directions	110
	References	112
	Appendices	123
A	Promoter: A primer	123
A.1	Sigma subunit of RNA polymerase	126
A.2	Transcription Initiation	127
B	Web sites for Promoter Prediction	130
C	Sample data set of <i>E.coli</i> provided by Dr. Leo Gordon	133
D	<i>Section3</i> of <i>E.coli</i> genome (NCBI)	135
E	Glossary	141
F	List of Publications	147

List of Tables

2.1	Confusion Matrix	13
2.2	Conformational and physico-chemical properties B-DNA dinucleotides [93].)	21
2.3	Structural Properties [56].	22
2.4	Promoter recognition software tools [59]	29
2.5	Eukaryotic promoter recognition [5]	30
2.6	Promoter Recognition Software for prokaryotes	30
3.1	<i>E.coli</i> classification results for different n-gram features. Average of 5-fold cross-validation.	37
3.2	<i>Drosophila</i> classification results for different n-gram features. Average of 5-fold cross-validation.	37
3.3	<i>E.coli</i> promoter Classification for different negative data sets using 2-gram features.	39
3.4	Promoter Classification of <i>Drosophila</i> for different negative data sets using 2-gram features.	40
3.5	Distances between TP,TN,FP and FN	44
3.6	Test data results of neural networks NN_{Maj} and NN_{Min}	46
3.7	Test data results of neural networks NN_{PC} , NN_{PN}	54
3.8	Prediction results of <i>section3</i> of <i>E.coli</i> using different promoter prediction packages.	55
3.9	Summary of promoter recognition results on <i>section3</i> of <i>E.coli</i> using different software tools.	57
4.1	A multi-level classification	62
4.2	<i>E.coli</i> classification results using SMOTE	67
4.3	<i>E.coli</i> classification results using cascading system of networks	68
4.4	<i>E.coli</i> classification results using AdaBoost classifier	70
4.5	<i>E.coli</i> base distribution for -35 binding site (Harley et al. [41])	72
4.6	<i>E.coli</i> base distribution for -10 binding site (Harley et al.[41])	72
4.7	<i>E.coli</i> bi-gram distribution for -35 binding site	73
4.8	<i>E.coli</i> bi-gram distribution for -10 binding site	74
4.9	<i>E.coli</i> classification results using log probability obtained by using PWM for binding sites only.	77
4.10	<i>E.coli</i> classification results using log probability values obtained by using PWM for whole promoter.	80

5.1	Physico-chemical properties of DNA [28].	89
5.2	Classification results using power spectrum values for <i>E.coli</i> using different encoding schemes.	94
5.3	Classification results using power spectrum values for <i>Drosophila</i> using different encoding schemes.	95
5.4	Classification results using wavelet coefficients as features for a neural network classifier for <i>E.coli</i> using EIIP encoding.	96
5.5	Classification results using decomposed waves as features to a neural network classifier for <i>E.coli</i> using EIIP encoding.	97
5.6	Classification results using decomposed waves as features to a neural network classifier for <i>Drosophila</i> binary indicators encoding scheme.	97
5.7	classification results using DNA-RNA Polymerase sigma subunit cross-correlation values as features for a neural network classifier for <i>E.coli</i>	100
5.8	Classification results using DNA-RNAP sigma cross-correlation at various levels.	101
A.1	RNA polymerase subunits [78]	126
A.2	RNA polymerase sigma subunits [78]	128

List of Figures

1.1	Central dogma of molecular biology [48]	2
1.2	<i>E.coli</i> promoter structure [36]	3
3.1	Average distance between promoter and non-promoter sequences using 2-grams for <i>E.coli</i> . On x-axis, 0...15 denote 2-grams AA, AT, AG, AC,, CG,CC.	34
3.2	Average separation between promoter and non-promoter sequences for 3-grams for <i>E.coli</i> . On x-axis, 0...63 denote 3-grams AAA, AAT, AAG, AAC, ... , CCG, CCC.	35
3.3	Average separation between promoter and non-promoter sequences for 4-grams for <i>E.coli</i> . On x-axis, 0...255 denote 4-grams AAAA, AAAT, AAAG, AAAC ,, CCGG, CCCC.	35
3.4	Average separation between promoter and non-promoter sequences using 5-grams for <i>E.coli</i> . On x-axis, 0...1023 denote 5-grams AAAAA, AAAAT, AAAAG, AAAAC,, CCCCCG, CCCCC.	36
3.5	2-gram frequency averages for promoter data set and 50% A+T rich synthetic negative data set.	38
3.6	2-gram frequency averages for correctly classified promoter data set and misclassified negative data set consisting of gene and inter-gene portions.	41
3.7	2-gram frequency averages for misclassified promoter data set and correctly classified negative data set consisting of gene and inter-gene portions.	42
3.8	2-gram frequency averages for promoter data set and negative data set consisting of segments from gene and inter-gene portions of the DNA.	43
3.9	2-gram frequency averages for promoter data set and negative data set consisting of gene segments from DNA.	44
3.10	2-gram frequency averages for promoter data set and 60% A+T rich synthetic negative data set.	45
3.11	Scheme for promoter recognition in whole genome using networks NN_{PC} and NN_{PN} .	50
3.12	The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for <i>section1</i> of <i>E.coli</i> genome.	51

3.13	The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for <i>section1</i> of <i>E.coli</i> genome.	52
3.14	The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for <i>section3</i> of <i>E.coli</i> genome.	52
3.15	The outputs of the the network NN_{PC} and the output generated by SAK versus the moving window for <i>section3</i> of <i>E.coli</i> genome.	56
4.1	Possible output of a multi-level binary classification system. . . .	64
4.2	Outputs of the networks NN_{PC} and NN_{PN} by using features determined by PWM (for binding regions only) versus the moving window for <i>section1</i> of <i>E.coli</i> genome.	78
4.3	Output of the network NN_{PC} using 3-grams and the output generated by using PWM for whole promoter versus the moving window for <i>section1</i> of <i>E.coli</i> genome.	81
5.1	Decomposition and downsampling of signal S using wavelets. . . .	87
5.2	Decomposition of signal S using wavelets.	88
5.3	A sample promoter sequence represented in terms of EIIP values for nucleotides.	90
5.4	A sample promoter sequence decomposed into various levels using Bior3.3	92
5.5	RNA Polymerase subunit sigma, in terms of EIIP values.	98
5.6	Sigma subunit decomposed into various levels.	99
5.7	Cross-correlation between sample promoter and RNA Polymerase subunit sigma.	100
5.8	Cross-correlation between sample promoter and RNA Polymerase subunit sigma at various levels.	102
A.1	Eukaryotic gene structure [48]	124
A.2	Eukaryotic promoter-structure [100]	125
A.3	Transcription process in <i>E.coli</i>	125
A.4	E.Coli promoter-structure	127
A.5	Binding of RNA polymerase to an eukaryote promoter.	129

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

the binding of RNA polymerase to it are described in Appendix A. Next section discusses about the need for promoter recognition and the factors that contribute to the complexity of the problem.

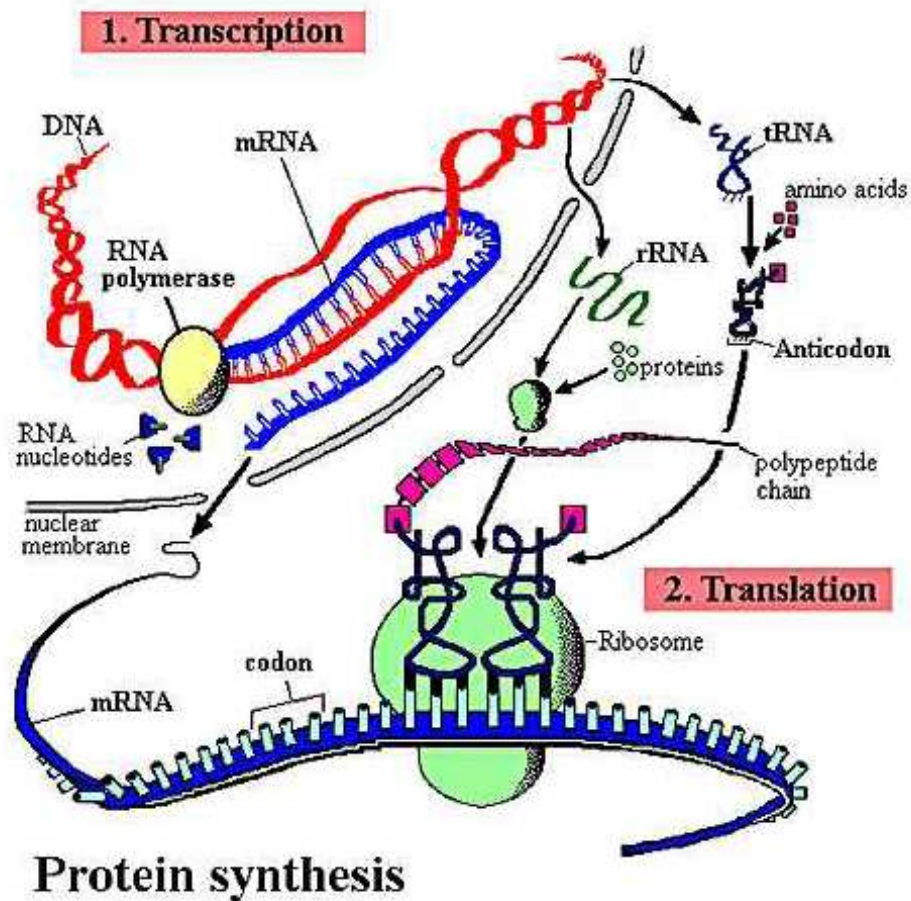
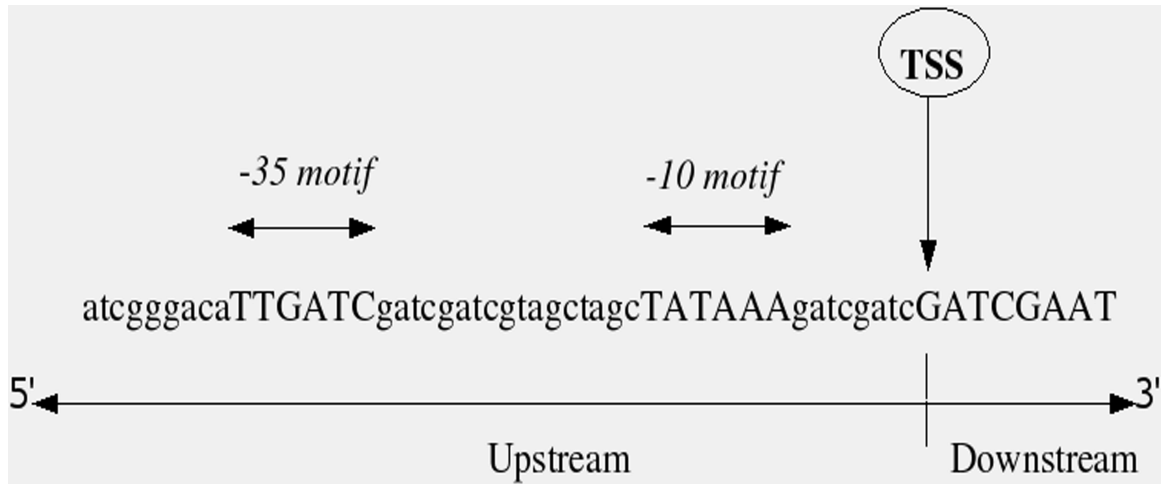


Figure 1.1: Central dogma of molecular biology [48]

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

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

through the co-regulation of the genes.

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 *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.

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.

Chapter 2

Survey of Promoter Recognition methods

This chapter presents a survey of promoter recognition techniques. Here a perspective of methods based upon various criteria is discussed. They include methods that are based on features extracted from certain regions in promoter (local) and the entire promoter (global), prokaryotic versus eukaryotic promoter recognition, and various statistical and non-statistical methods. Different kinds of metrics used to estimate and compare the prediction rates of different classification algorithms in literature are also reviewed.

2.1 Promoter

Organisms are divided into two classes in biology, those that have a nucleus containing DNA isolated from the surrounding plasma by a membrane in a cell and those that have no separate nucleus for the DNA. The ones that have a separate nucleus are called as eukaryotes and those that do not have a separate nucleus for enclosing DNA are known as prokaryotes. Even though the basic mechanism of transcription is the same i.e., binding of RNA polymerase to the promoter, there are differences in the promoter structures. The following sections specify the properties and problems specific to prokaryotic and eukaryotic promoters. Before describing the features that are used in promoter recognition, it is essential to point out the differences and similarities between these two types of promoters.

The role of promoter is to transcribe the gene. This can happen in either an

unregulated or regulated manner through extracellular or intracellular signals. In regulated transcription the regulating factors could occur close to the promoter that it is regulating, hence the promoter is most often taken as region necessary to initiate the transcription [112]. The assumption of the underlying structure is an important factor in the prediction of promoters. Core promoter is the region which initiates the transcription. The region adjacent to the core region on 5' end is called distal promoter. The 5' end of the promoter cannot be demarcated precisely since the distal promoter also occurs adjacent to it. Sequence analysis alone is not sufficient due to these factors. This region can extend from 100 bp to 2 kbp. Binding sites for all known transcription factors, which either activate or repress the transcription can be found in this region. The transcription factor (TF) binding sites can occur anywhere in the promoter and they do not have any consensus pattern or locational invariance. Details of structural features of promoters in a genome are given in Appendix A.

2.1.1 Prokaryotic Promoters

Prokaryotic genes have no introns but only exons. Promoter is supposed to consist of two binding regions, called -35 and -10 binding regions, so named because of the position of occurrence of these regions with respect to the Transcription start site (TSS). Hexamers, i.e. sequences of length 6 nucleotides, in these locations are considered for promoter analysis. Consensus sequence for -10 site is **TATAAT** and for -35 site is **TTGACA** for σ^{70} promoters, which constitute the majority of promoters [41]. There are exceptions to these rules in the sense that there is no guarantee that a promoter will always have these two binding regions. These exceptions can occur in several ways. One of these is extension of -10 box by a very short subsequence 'TG' 1 bp on upstream side of -10 box. Another example is that -35 box is dispensed with in *E.coli*. Also an upstream activator can make the -35 site dispensable or another region called a UP region located 4 bp upstream of -35 box can make the promoter stronger [47]. In case of prokaryotes, there are a group of coupled genes called operons. All the genes in operons will be expressed together. These genes occur consecutively in a genome. Promoters for these genes could be overlapping with each other.

2.1.2 Eukaryotic Promoters

The eukaryotic promoter has in total four promoter binding sites, TATA box, the initiator (Inr) region, an upstream activating element (UPE) and a downstream promoter element (DPE). It is not essential that they are all present in all cases. It is shown that in TATA-less promoters, initiator combined with downstream element is able to initiate the transcription [68]. Same is the case when TATA is present. In that case the the initiator and downstream element are dispensed with. The combination of TATA and initiator is also found in several viral promoters. The last group is where neither the TATA nor the initiator is present, but only upstream and downstream elements do the job. There are variations in TATA box and Inr also. All these facts contribute towards the complexity of the promoter recognition methods, especially those that use specific motifs. Nevertheless detection of transcription factor binding sites is being used extensively. In addition to these, properties such as GpC content, secondary structure elements, cruciform DNA structure are also being used in eukaryotic promoter recognition [112]. With this introduction, first the metrics used to measure classification accuracies are discussed. This is followed by a survey of the methods that use different kinds of features and classifiers.

2.2 Metrics used in Promoter Recognition

Metrics are the measures that characterize classification accuracy. Various kinds of measures are being used in the promoter recognition literature. Specifically in a binary classification problem, four kinds of outputs can be expected. If a positive/negative pattern is correctly classified as positive/negative, then it is termed as TP/TN classification. Similarly, if a positive/negative pattern is classified as negative/positive, it is termed as FN/FP. The following matrix, in Table 2.1, called confusion matrix, illustrates the scenarios clearly. A few measures that are used in literature are defined as follows:

$$Sn (Sensitivity) = \frac{TP}{TP + FN} \quad (2.1)$$

$$Sp (Specificity) = \frac{TN}{TN + FP} \quad (2.2)$$

Table 2.1: Confusion Matrix

	Promoter	Non-promoter
Promoter	TP Output:1	FN Output:0
Non-promoter	FP Output:1	TN Output:0

$$AE \text{ (Average Error)} = \frac{FN + FP}{TP + TN + FP + FN} \quad (2.3)$$

$$CC \text{ (Correlation Coefficient)} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad (2.4)$$

$$P \text{ (Precision)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.5)$$

$$Ppv \text{ (Positive Predictive value)} = \frac{TP}{TP + FP} \quad (2.6)$$

Although there is some variation in terminology used for the metrics in the literature, we will use definitions 2.1-2.6 in this thesis. Sensitivity gives the proportion of the positives that are classified as positives. Similarly, specificity gives the proportion of the negatives that are classified as negatives. Precision gives the number of correctly classified sequences both positive and negative out of the total number of sequences. In the same sense, average error gives the total misclassified sequences in the total number of sequences. Positive predictive value indicates the positives predicted against the total positives that are predicted. These measures are used to compare the results and to estimate the classification accuracy of work of various groups reviewed in this chapter.

2.3 Promoter Recognition Methods

Promoter recognition can be done using different paradigms. It can be sequence-based, where the known consensus sequence patterns/motifs are used. Another

way is to use the interactions between protein and DNA. The literature is organized such that the recognition methods are classified into local signal-based methods and global signal-based methods. Methods that make use of features of specific parts of the promoter sequence are identified as *local signal*-based methods. On the other hand, methods that make use of the entire sequence are termed as *global signal*-based methods. Different feature extraction methods are used to extract features and also different classifiers are used for classification.

2.3.1 Local Signal Recognition Methods

Since genes code for the proteins and proteins hold a vital role in drug designing, emphasis has been on gene recognition most of the time. As it was already pointed out, gene regulation could be done right at the promoter level itself. Recognition of the promoter could be either based upon features extracted from a few regions (could be binding regions/motifs) or otherwise based on whole promoter sequence. These regions as the name suggests would act as beacons for the RNA polymerase to bind in order to start transcription or other special regulatory proteins to initiate the transcription or repression. Most of the methods in promoter recognition are based upon motif recognition. These methods could be called as local signal-based methods. Finding the binding regions is not a trivial task since the content of the binding regions is not conserved and the length of the sub-sequence, called a spacer in some research papers, between the binding regions is also not constant in length. This fact makes the promoter recognition a complex problem. The following section describes the application of weight matrices, neural networks, hidden Markov models, and genetic algorithms for promoter recognition. Most of these techniques can be used for feature extraction as well as classification of promoters. Techniques that are used in local signal detection are all pattern-driven.

Expectation Maximization Algorithm

Lawrence and Reilly [60] have considered a set of *E.coli* promoters and their general characteristics from the data compiled by Harley et al. [41]. They have used Expectation Maximization (EM) algorithm to initially identify the common -35 and -10 binding sites [60]. They considered only binding sites with a fixed

spacer length. Cardon and Stormo extended this idea by considering spacers of variable length 15-21 bp between -35 and -10 binding sites [18]. EM algorithm is a two-step process. First step estimates the parameters of a model and the second step uses these parameters to maximize the likelihood of observing these values and the process iterates until convergence. EM algorithm is applied to find the location and length of the protein binding sites. Their algorithm could identify about 87% of the promoters in the data set they have considered. An addition of a weakly conserved pattern downstream of -10 region has enhanced the recognition capability to 94%. They have also confirmed that promoters need not be differentiated on the basis of a spacer. Eight bases in spacers have positional importance in contributing to promoter specificity. They claim that the EM technique could as well be used in situations where there is little information about the promoters and the number of binding sites with unknown spacing. They have not tested their algorithm on non-promoters and the efficiency of the algorithm can not be estimated in that case.

Ma et al. have used Bayesian Maximum *A Posteriori* (MAP) EM algorithm considering the binding sites as well as the spacers [70]. They have used an *E.coli* promoter data set of 438 sequences and also a synthetically generated negative data set with 60% A+T composition. Here they also considered one more spacer length between TSS and -10 binding site apart from spacer considered by Cardon and Stormo [18]. The spacer length between -10 and TSS was taken to vary between 4 and 11 bases. They have estimated the position weight matrices and the spacers using the MAP algorithm. Using these they have extracted windows around -35, -10 and TSS. The bases in these windows and the spacers are given as input features using orthogonal encoding to neural network. They achieved 96.68% specificity and 91.78% sensitivity. The drawback is that real biological negative data sets are not considered by this group. It will be shown in Chapter 3 that the promoter recognition against non-biological (synthetic) negative set is a relatively simple classification problem.

Position Weight Matrices

A position weight matrix (PWM) is the matrix representation of frequency of occurrence of a particular feature at that position. If there are N features possible at a particular position and if there are M consecutive positions, then the

resulting matrix will be of size $N \times M$. Each element p_{ij} of the matrix represents the frequency of occurrence of feature i at position j .

In case of promoter recognition, position weight matrices are often used to compute the positional occurrence of binding sites. Given a set of sequences with a known reference point, say the Transcription start site (TSS), the frequency of occurrence of each base at a position is determined and a matrix is computed. Once the matrix is computed, it is used to estimate the probability of a binding site in a sequence. Various groups that have worked with PWM to find a promoter are Harley et al. [41], Huerta et al. [47], Gershenzon et al. [34], Bajic et al. [6], Zhong Li et al. [64], Claverie et al [24].

Harley et al. have compiled and analyzed 263 *E.coli* promoters with known TSS [41]. The -35 and -10 hexamers were aligned using the following steps. The initial weight matrices for -35 and -10 were derived from the compilation and promoter alignment of Hawley and McClure [43]. These weight matrices were updated using the new set of sequences. The base frequencies are analyzed at each position in the weight matrices for the binding sites. The PWMs show a clear consensus at these binding sites. It was found that most of the promoters had inter-region spacing of 17 between the binding sites and 75% of the uniquely defined start points are initiated 7 ± 1 bases downstream of the -10 region.

Huerta et al. have created *E.coli* promoter data set called Regulon data base [47]. They extracted and aligned motifs in a given set of unordered sequences producing a frequency matrix. Initially, sequences were aligned with respect to TSS and the motif corresponding to -10 was identified corresponding to corrections of four standard deviations. Corresponding to each -10 box that is determined, -35 box is identified from the realignment of sequences with respect to -10 boxes. A set of 6 sequences of various lengths initiating at 13 bp upstream of -10 are selected. In all, for each -10, six alternative sequence sizes with four different standard deviations are analyzed, resulting in 24 possible -35 matrices. Hence in total 96 matrices (-10 with four standard deviations \times -35 with 6 lengths \times -35 with 4 standard deviations) were generated for the promoter. Similarly, 96 matrices for coding regions and 96 more for non-coding region are also generated. A set of 288 (96 for promoter, 96 for coding region and 96 for noncoding regions) different weight matrices were thus created. Out of these 4 best matrices are chosen based on a particular criterion. These weight matrices are used on a test

set to obtain a score. The score of a particular region is assigned by adding products of frequencies of nucleotides at each position. The global score was defined as the sum of -10 and -35 and the score of the spacer. These scores are used to identify the promoters. The predictive capacity of the method is 86%, however accuracy defined as the average of sensitivity and positive predictive rate is 53%. An important contribution of this work is that there is high number of putative promoters (promoter-like signals) in the vicinity of a true promoter, which show a better score than true promoter. These putative promoters may be trying to bring Ribonucleic polymerase (RNAP) closer to the functional promoter.

Zhong li et al. have considered a set of 683 experimentally verified *E.coli* σ^{70} -promoter sequences [64]. The conservation of sequence segments at various position are calculated. The conservation of hexamer segments is found to be large in 10 positions/sites, namely, -37, -36, -16, -15, -14, -13, -12, -11, -10 and -2. The position correlation scoring function for each of the sites is computed by using 200 sequences as training data set. The position correlation scoring matrix (PCSM) is of size 4096X10 (For a hexamer $4^6=4096$ combinations are possible and ten positions are being considered here). In the same way PCSM for negative set of training sequences is also computed. Given a new testing sequence, the score that results by using PCSM for both positive and negative data sets is used. Based on the score obtained, the test sequence is identified as positive or negative. For test data the sensitivity and specificity are, 91% and 81% respectively for negative data consisting of coding regions alone and 90% and 77% respectively for negative data taken from intergenic portions only. Applying these scores to whole genome to predict the promoters, they have obtained specificity of 30% and sensitivity of 100%. They were able to predict all 683 experimentally verified σ^{70} promoters and also obtained 1567 predictions as probable promoters.

Neural Networks

Artificial neural networks are modeled based on the neuronal system of brain. A neuron takes inputs from external sensory organs or internal other neurons and processes the input and fires if the the processed value exceeds the threshold of a non-linear excitation function of the neuron. In a similar fashion, neural networks require a set of inputs, which are the features extracted from the training examples, and processes the inputs and an output is given if the processed input

has a value, which exceeds the threshold of an excitation function. Different groups that have worked on promoter recognition using neural networks are Ma et al., Mahadevan et al., Xu et al., Reese et al., Wang et al., Bajic et al. [70, 71, 113, 96, 111, 7]. In order to illustrate the technique only two are chosen.

Bajic et al., have designed an algorithm which combines a nonlinear promoter recognition model with signal processing, artificial neural networks (ANNs), and a set of sensors in Dragon fly (*Drosophila melanogaster*) promoter prediction [7]. These sensors are based on the statistical concept of oligonucleotide positional distributions in specific functional regions of DNA. Each sensor models a particular functional region such as promoter, coding-exon, and intron. These distributions are modeled as a set of position weight matrices of the most significant oligonucleotides. Pentamers (regions of length 5) that most significantly contribute to the separation between the promoter and non-promoter regions are chosen by determining the significance using their statistical relevance. The highest ranking 256 (out of possible 1024) pentamers are chosen. The 256 pentamer-position weight matrices are generated for the three functional groups using a sliding window which moves one base at a time. The signals of a sequence using the positional weight matrices for the three functional regions are fed to a signal processing block and the output is fed to ANN, which performs multi-sensor integration. Scores that make the ANN output greater than the selected threshold are to be treated as positive predictions in the promoter region. They have obtained a sensitivity rate of 67%. They have shown that they predict less false positives compared to existing algorithms such as PromoterInspector at that time.

Genetic Algorithms

Genetic algorithm (GA) is an evolutionary algorithm designed after biological evolution. GA finds the solution to a combinatorial optimization problem. GA has three operations, namely, selection, mutation and cross-over which are used to generate new generation of solutions from the existing population of solutions. Levitsky et al., and Beiko et al. [62, 10] have experimented using genetic algorithms.

Levitsky et al. have used the genetic algorithm based on iterative discriminant analysis to classify eukaryotic (*Drosophila*) promoters. The negative set is obtained by shuffling the promoters. As a first step the sequence is partitioned

into a few parts and the optimal non-overlapping partition of the sequence is obtained by using GA. Mutation is obtained by changing the positions of borders between the partitions so that the number of these partitions is constant with a restriction on the constant minimal size of any partition. Cross-over or recombination is done by exchanging fragments between two partitions. The partition most suitable for recognition is determined as a result of successive mutations and recombinations of the partitions analyzed. In the second step, most significant dinucleotide frequencies in these partitions are selected using GA again. A promoter recognition function based on these two steps combined with nucleosome potential is found to increase the accuracy. Two promoter samples TATA-containing and DPE-containing sets are formed. Upon the second stage of application of GA, cross-correlation (CC) for TATA-containing promoters is reported to be 0.92 and CC for DPE promoters is 0.82.

Hidden Markov Models (HMM)

Markov models are used in cases where there is a dependency of the current scenario on the past scenarios. A hidden Markov model can be formally defined as a three tuple (Π, A, B) . Here Π is a vector of initial probabilities of states, A (Transition probability matrix) is probability of going from one state to another state and B (Emission probability matrix) is probability of emitting a symbol b in state k . This model can be used to solve a problem posed as: Given several observation sequences O_n , to estimate the Model Parameters: (Π, A, B) called the Estimation Problem. Baum-Welch Algorithm finds the transition and emission probabilities for a hidden Markov model given some training examples and a structure for the model. Later, given an unknown sequence, the probability of observing that sequence is computed using the estimated model. Depending on a particular threshold, it could be designated as a positive or a negative class item. Pedersen et al. and Ben-gal et al. [91, 11, 85] have used HMM in promoter recognition.

Pedersen et al. have characterized the promoters of prokaryotes (*E.coli*) and eukaryotes (*human*) using self-organizing parallel HMMs [91]. They have considered a set of 3 states-the main states, the delete states, and the insertion states, in addition to start and end states. The set of emissions are the four nucleotides A, T, G, C. Main and insertion states always emit a nucleotide whereas deletion

state is a no-emission state i.e., a mute state. Given a set of K training sequences, the parameters of HMM are iteratively modified to optimize the data fit using a measure based on the log likelihood. A set of HMMs trained on 38 σ^{70} sequences, and 3 σ^{54} sequences are combined in parallel to create a super HMM for *E.coli* promoter recognition. Similarly human promoter sequences are used to train another HMM model. Clear patterns of well known consensus signals such as TATA box etc. could be obtained from the emission probabilities of main states of the HMM model. Their model is able to classify 162 σ_{70} sequences out of 166 sequences as σ^{70} sequences and 3 σ^{54} out of 166 as σ^{54} sequences. Only one σ^{70} sequence out of 166 is misclassified. They have not tested on non-promoter sequences.

Data mining methods

Graph based induction (GBI) method is used by Matsuda et al. for promoter prediction [74]. The data set is same as that of Huang et al [46]. Promoters and non-promoters are formed into two different groups in a directed graph. Graph based induction method is used to extract the patterns in the data sets. If a particular pattern has a frequency threshold more than 4%, then the pattern is replaced with a new node in the graph. This process is repeated until no new patterns can be found. These patterns are extracted as rules to classify promoters and non-promoters. Many rules can be obtained in this process. Importance of the rules is decided by specificity of the rules. More specific rules are given highest importance. Recognition rate of graph based induction method is compared with other standard tree-induction algorithms such as ID3 and C4.5. Error rates of GBI algorithm are found to be slightly lower than either C4.5 or ID3.

2.3.2 Global Signal Recognition Methods

In contrast to the above paradigm, methods that use the whole promoter sequence can be categorized as global signal-based methods. These methods do not distinguish between binding and nonbinding parts of the promoter. Some of the techniques that are proposed using this paradigm are sequence similarity, signal theoretic methods etc.

It is said that DNA encodes two levels of functional information. The first

Table 2.2: Conformational and physico-chemical properties B-DNA dinucleotides [93].)

Parameter name	Units	Min	Max	Ref
Roll in B-DNA	Degree	-6.2	6.0	[107]
Helical twist in B-DNA	Degree	32.6	40.5	[107]
Roll in protein-DNA complexes	Degree	-2.0	6.3	[107]
Twist in protein-DNA complexes	Degree	29.3	39.5	[107]
Slide in protein-DNA complexes	Angstrom	-0.7	0.7	[107]
Wedge	Degree	1.1	8.4	[14]
Direction of Wedge	Degree	-154	180	[14]
Melting temperature	$^{\circ}\text{C}$	36.7	136.1	[38]
Probability of contact with nucleosome core	%	1.1	18.4	[98]
Bend towards major groove	c.u.	0.98	1.18	[33]
Bend towards minor groove	c.u.	1.02	1.27	[33]
Twist for B-DNAs(NDB)	Degree	27.7	40.0	[37]
Minor groove distance for B-DNAs(NDB)	Angstrom	2.79	4.24	[37]
Roll for B-DNAs(NDB)	Degree	-7.0	6.6	[37]
Slide for B-DNAs(NDB)	Angstrom	-0.37	1.46	[37]
Propeller twist for B-DNAs(NDB)	Degree	-17.3	-6.7	[37]
Enthalpy change	kcal/mol	-11.8	-5.6	[106]

level is for protein and targets for activators, enhancers, repressors, transcription factor binders etc. The second level of information is contained in the physical and structural properties of the DNA itself [54, 93]. In literature, several groups have exploited these properties to distinguish between features specific to a particular set of DNA sequences and sequences that do not belong to a particular set. Physico-chemical parameters of DNA double strand available in literature are shown in Table 2.2 [93]. Other groups that have considered the structural properties specific to mammals and plants are given in Table 2.3 [56]. There are some groups who have encoded the DNA independent of these properties in terms of binary values. Whatever is the encoding that is used, the whole sequence is considered for modeling in global signal based methods.

Methods using Structural Properties of DNA

Wang et al., have found from their studies that promoter regions in the genome are more susceptible to stress-induced duplex destabilization (SIDD) [111], and

Table 2.3: Structural Properties [56].

Property	Max	Min	Reference
Stacking energy	14.59 kcal	3.82 kcal	[89]
Propeller twist	18.66	8.11	[42]
Nucleosome	36%	+45%	[98]
Bendability	0.280	+0.194	[16]
A-phility			[49]
Protein-induced deformability	1.6	12.1	[87]
Duplex disrupt energy	0.9 kcal	3.1 kcal	[15]
Duplex free energy	-2.1 kcal/mol	-0.9 kcal	[106]
DNA denaturation	64.35 cal/mol	135.38 cal/mol	[13, 12]
DNA-bending stiffness	20 nm	130 nm	[103]
B-DNA twist	30.6	43.2	[37]
ProteinDNA twist	31.5	37.8	[87]
Stabilizing energy of Z-DNA	5.9 kcal/mol	0.7 kcal/mol	[45]

that the extent of destabilization is bimodally distributed. SIDD is a structural property. In comparison with other structural properties such as DNA curvature, deformability, thermostability or sequence motif scores within the -10 region, SIDD is found to be the most informative DNA property regarding promoter locations in the *E.coli*. They reported that the usage of SIDD to recognize the promoters has very low false positive rate. When SIDD properties are combined with -10 motif scores in a linear classification function, they predict promoter regions with better than 80% accuracy. When these methods were tested with promoter and non-promoter sequences from *Bacillus subtilis*, they achieved similar or higher accuracies.

Aditi et al., have looked at the thermal stability properties of three genomes *E.coli*, *B.subtilis*, *C.glutamicum* with A+T compositions of 0.49, 0.56, 0.46, respectively [1]. The average stability profiles for three sets of bacterial promoter sequences is calculated (using 15 nucleotide moving window). Promoters from all the three bacteria show low stability peak around the -10 region. The other interesting feature in the free energy profiles of all the three bacteria was that the difference in stabilities of the upstream and downstream regions. In all the three groups of promoter sequences, the average stability of upstream region is lower than the average stability of downstream region. Hence it was concluded that the promoter region is less stable and hence more prone to melting as compared to

other genomic regions. They claimed that their analysis shows that the a method of promoter prediction based on the differences in the stability of DNA sequences in the promoter and non-promoter regions gives better results compared to existing prokaryotic promoter prediction programs, which are based on sequence motif searches. This scheme has potential to be extended to automatic promoter recognition in whole genome. But, such attempt has not been reported so far.

Methods Using Signal Processing Techniques

Signal processing techniques are based upon encoding the original signal into a numerical series by one of the encoding methods and then applying the Fourier transform or wavelet transform to the transformed sequence.

Fourier Transform

In case of Fourier transform, the power spectrum is computed. The features are used to recognize a gene versus a non-gene and a promoter versus a non-promoter. Fourier transform is a non-local method, hence to retain the local frequency information wavelets have been used.

Deyneko et al. have applied the physical features of DNA to find similar promoters which correlate with their transcription regulatory responsiveness to different antibiotic and osmotic treatments [28]. They transformed the *E.coli* promoters into numerical sequences using physical parameters such as enthalpy, roll angle etc. They did cross-correlation and auto-correlation between different promoters using FFT of the the transformed sequences. The similarity between cross-correlation and auto-correlation is used to identify co-regulated genes. Two sequences are considered similar if their distance (Euclidean, correlation etc.) is less than a user-defined threshold. In particular, they looked for genes responsible for SOS response. It is to be noted that in these cases the letter similarity (by BLAST used for sequence similrity search) for these promoters is as low as 40-55%, while signal identity using the signal processing methods is more than 85%.

Wavelet Transform

Application of wavelet transform to a signal decomposes the signal into several groups of coefficients. Different coefficient vectors contain information about characteristics of the signal at different scales. If a wavelet with a particular window size matches with the signal in that particular window size then the coefficient will be maximum. Coefficients at coarse scale capture gross and global features of the signal while coefficients at fine scales contain local details. Discrete wavelet analysis (DWT) is more appropriate for samples sampled discretely. A few groups have tried wavelet approach to find the conserved regions in proteins [57] and for protein sequence comparison [21]. But using wavelets promoter recognition has not yet been attempted using wavelets.

Similarity between proteins has been obtained mostly using homology analysis. Such a sequence comparison will have low recognition rates if the sequence similarity falls in the twilight zone (sequence similarity is $< 30\%$). In order to overcome this handicap, Chafia et al. [21] have done protein classification using wavelets. They converted the protein sequence into a numerical sequence using EIIP encoding [25]. They decomposed the original numerical sequence using Bior3.3 mother wavelet into 6 levels. They have proposed a sequence similarity at different spatial resolutions. To obtain the sequence similarity between two proteins, cross-correlation between decomposed waves at the same level for both proteins is computed. When the cross-correlation between the decomposed waves is $> 70\%$, the proteins are supposed to be strongly related to each other. For distantly related proteins, correlation was found to be in the range of 0.5 to 0.7 (weak correlation) even though sequence similarity is only about 15%. Spatial scale similarity was able to predict the similarity at least at one scale in case of distantly related proteins. Krishnan et al. have followed similar scheme to classify mitochondrial proteins [57]. In this case they have used Composition (c), Polarity (p) and Molecular Volume (v) values for each amino acid to convert a protein sequence into a numerical sequence followed by decomposition using wavelets. Then cross-correlation between decomposed waves is performed. They have used a threshold of 0.55 to classify a protein as a mitochondrial protein or a non-mitochondrial protein. In this thesis, an attempt is made to use FT and WT methods for understanding promoter-RNA polymerase interaction in Chapter 5.

Methods Using Sequence Similarity

Similarity between two sequences is defined as how many matches are there between the letters of the two sequences. The higher the matching, the closer the sequences are. To find out the match or mismatch between sequences, they have to be aligned with each other so that a score is computed. This score can be used to signify the degree of resemblance between the sequences. Sequence alignment algorithms are local sequence alignment algorithm, global sequence alignment algorithm, Basic local alignment search tool (BLAST), and Fast alignment (FASTA).

The sequence similarity between the promoters is supposed to be not very high. But, Gordon et al. used the sequence alignment kernel to do the promoter recognition [36]. Here, they considered the *E.coli* promoters from Regulon and Promec data bases. They also considered biologically meaningful data sets taken from the coding regions and non-coding regions. They have built a kernel function as a quantitative measure of sequence similarity between two sequences. This is used in conjunction with Dual SVM to do the classification. This method is important since this is a global method and does not depend upon any prior annotation of the binding regions. They achieved a FP rate of 14.6%, FN rate of 18.5% and sensitivity of 82% for the data set using negative data taken from coding regions alone. In case of negative data taken from non-coding regions, FP rate of 18.2%, FN rate of 19% and sensitivity of 81% were obtained. This scheme is extended later to whole genome promoter prediction.

Position Weight Matrix (PWM)

Huang et al. have used a set of 53 *E.coli* promoter sequences and a set of 53 non-promoter sequences, each of length 57 bp, to train their system [46]. A PWM for hexamers at each position in the sequences for both promoter and non-promoter sequences are computed. That is, the sequences are aligned with respect to the first position. PWM for the first to six positions is computed as the first hexamer weight matrix. Then second to seventh positions as PWM for second hexamer and so on are computed. The score of an unlabeled sequence is computed using these PWMs for promoters and non-promoters as TSCORE and FSCORE respectively. These are fed to a hybrid neural network consisting

of classifiers SVM, Multi-layer feed forward neural network and another neural network based on knowledge base of hierarchically structured rules (KBANN). By employing boosting and bagging, they claim to have obtained a precision of 98%. Extension of this method to whole genome promoter recognition is not tried. In this thesis, PWM method has been combined with n-gram features in order to implement promoter recognition.

N-grams

An *n-gram* is a selection of n contiguous characters from a given character stream [101]. Different n-grams are extracted from a sequence. Various groups have done a studies of n-grams as features in promoter recognition [61, 52].

Leu et al. have developed a vertebrate promoter prediction system based on statistics using cluster computing. They extracted *n-grams* for $n=6$ to 20 [61] to identify possible transcription factor binding sites in promoters and non-transcription factor binding sites in non-promoters. They have not specified the reason for choosing particular n values of 6 to 20. Each n-gram is given a score based on its occurrence in both promoters or only in promoters or only in non-promoters in the training data set. Score for n-gram occurring in promoters is taken as positive and for n-grams occurring in non-promoters is taken as negative. The scores of the n-grams are sorted in descending order. They considered sequences of length 550 bp. Each sequence is segmented into portions of length 200 bp with an overlap of 100 bp with the next segment. The cumulative score of initial n-grams for $n=6$ to 20 is assigned to $s[21]$. Similarly rest of the scores for $s[22]$, $s[23]$, $s[24]$, ... $s[221]$ and so on are computed. Sum of all these 200 scores is stored in $US[X]$. A probable promoter is found if $US[X] > 0$, $US[X + 1] < 0$ and $US[x + 2] < 0$ in the subsequence X. They achieved an accuracy rate of 88%. This algorithm is not extended for whole genome promoter prediction.

Ji et al. have used SVM and *n-grams* ($n=4,5,6,7$) for target gene prediction of *Arabidopsis* [52]. The 'n' values are chosen based on trial and error methods. They considered 500 *n-grams* which include 19 four-grams, 47 five-grams, 121 six-grams, and 313 seven-grams. Overall, a systematic study using n-grams for promoter classification has not been undertaken. This has been attempted in chapters 3 and 4.

2.4 Existing Promoter Prediction Programs

In their report on prediction analysis on whole genome of human, Bajic et al. have considered various promoter prediction programs (PPP) and analyzed their performance [8]. It was found that most of the programs perform well on G+C rich chromosome 22. Prediction on the entire genome is essential to measure the strength of the prediction programs. Eight prediction programs which use different classifiers and features are considered for this task. It was found that the extrapolation of the predictions on the basis of two or three chromosomes data is not sufficient. They attributed this fact to the diversity in the promoters. They also made recommendations in developing and using the PPP. They suggested that the use of combinations of PPPs is more beneficial than using a single PPP. Table 2.4 gives the PPPs and the features that are used in those programs. Appendix B provides links to software programs available on internet. Table 2.5 summarizes the PPPs and their accuracies. In their article they make a recommendation that the combined prediction using gene structure prediction increases the sensitivity considerably [5]. We explore various methods for *E.coli* and *Drosophila* whole genome annotation. The details are given in chapters 3 and 4 and 5.

2.5 Discussion and Summary

In this chapter different promoter recognition techniques have been discussed. These techniques have been divided into two categories: techniques using local signal in promoters and methods based upon global signal in promoters. In local signal-based methods some parts of promoter are only used whereas in global signal-based methods whole promoter is used. Under local signal-based category, position weight matrices, Expectation and Maximization algorithm, neural networks, genetic algorithm and data mining methods are discussed. In global signal methods we have discussed methods based on physico-chemical properties, structural properties and sequence alignment. Some of these properties are used in Fourier transform and wavelet transform application by converting letter sequences into numerical sequences. We have also discussed sequence alignment and n-grams in this category. These methods have been used for *E.coli*, *Drosophila*, human, *B. subtilis*, *C. glutamicum*, and *Arabidopsis*. Most of the methods dis-

cuss experiments done using training and test sets. One important fact that is to be taken into cognizance is that most of these techniques are not evaluated on a genome sequence. A scheme to decide if given sequence is a promoter or a non-promoters cannot be extrapolated for the whole genome in the methods reviewed here.

We have investigated promoter recognition problem based on the assumption that they are functionally similar. There may not be good sequence similarity but there may be a bias towards usage of short segments in promoters. Position of occurrence of a nucleotide in a genome cannot be random. Studies have shown that there is a clear bias towards some triplets at least in genes which code for proteins. There has not been a systematic study of occurrence of bases in a promoter, even though researchers have attempted higher order n -grams for this purpose. As the size of n -grams grows, number of combinations that can be formed with these n bases also would grow exponentially. The N -grams may become more and more specific in that case. As a consequence of these observations we would like to do a systematic study of lower order n -grams for predicting a promoter (see chapters 3 and 4).

Methods using local information extracted from binding regions (sites/motifs) have been most prevalent in literature [47, 70, 46, 34], since the binding sites seem to be conserved. If a binding site is not present then these methods may not work efficiently. We wanted to address the question of whether the binding site recognition is necessary for promoter recognition or not. To analyze this we have chosen position weight matrix method, since annotated experimental data for -35 and -10 sites is available [41]. We have also attempted to analyze the situation in which there is no experimental data available.

Promoter recognition by RNA polymerase sigma subunit has to happen wherever the promoter may be embedded. Promoter has to have a unique signal which RNA polymerase can use to recognise. We would like to explore the interaction between promoter and RNA polymerase in the frequency domain using the signal processing techniques. Promoter recognition has not been attempted using signal processing techniques. We wanted to analyze and find whether a characteristic signal can be found at a particular resolution using wavelets. Results are given in Chapter 5.

Table 2.4: Promoter recognition software tools [59]

Name	Techniques used ^a	Features used
SIGNALSCAN	PWM	TATA box, CAAT box, GCbox, TSS, TFBS
MATRIXSEARCH	PWM	TATA, CAAT, GC, TFBS, TSS
MatInd/ MatInspector	PWM	TATA, CAAT, GC, TFBS, TSS
ConsInspector	Alignment based	TFBS of unlimited length
TFSearch	PWM	TFBS; TSS
TRANSFAC	PWM	
PromoterInspector	PWM	Promoter
PromoterScan	PWM	TATA box, TFBS
TSSG/TSSW	LDA	TATA box, TFBS, hexamer frequency
CpGProD	LDA	TSS CpG island, AT/GC content
CorePromoter/FirstEF	QDA	CpG island
CpG Promoter	QDA	CpG island, TSS
SAMPLER	Gibbs Sampling	TFBS, TSS
AlignACE	Gibbs Sampling	TFBS, TSS
MEME	EM	TSS, TFBS
Promoter2	NN	TATA box, Inr, CAAT box, GC box
DGSF	NN	CpG island, TSS, DPF
DPF	NN	Promoter, Exon, Intron, TSS
McPromoter	NN & interpolated	TAAT box, CAAT box, GC box,
	markov models	nucleosome position
NNPP	Time Delay NN	TATA box, Inr
Eponine	SVM	TATA box, GC box, TSS
Audic/Cleverie approach	HMM	Pol II promoters

^aNN-Neural Network, SVM-Support vector machine, HMM-Hidden Markov model, EM-Expectation and Maximization algorithm, LDA-Linear discriminant analysis, QDA-Quadratic discriminant analysis, PWM-Position weight matrix

Table 2.5: Eukaryotic promoter recognition [5]

Tools	Sensitivity(%)	PPV(%)	CC(%)
CpGProD	37-48	51-70	49-51
DGSF	61-65	62-64	63-64
DPF	53-80	15-32	34-45
FirstEF	79-81	35-40	53-56
Eponine	≈ 40	≈ 67	≈ 52
NNPP2.2	69-93	2.0-4.5	15-17
Promoter 2.0	44-57	≈ 4.5	≈ 14
McPromoter2.0	26-57	70-87	-
TSSG/TSSW	$\approx 29/\approx 42$	$\approx 72/\approx 59$	-
Audic	≈ 24	≈ 82	-

Table 2.6: Promoter Recognition Software for prokaryotes

Name	Techniques used	Features used
BPROM	LDF	Functional motifs and Oligonucleotide composition
PPP	HMM	-
NNPP	Time dealy NN	TATA box, Inr
SAK	SVM	Similarity between two sequences

Chapter 3

N-gram Analysis of *E.coli* and *Drosophila* Promoters

Recent literature postulates the idea that a genome sequence encodes two levels of information. The first level seems to code for proteins and targets for activators, enhancers, repressors, transcription factor binders. The second level of information is contained in the physical and structural properties of the DNA itself. In fact, it is conjectured that the second level is the one which dictates the type of nucleotide in a genome [54, 28]. Hence, it is of interest to find the adjacency property of the bases in a genome which may capture this structural information. And also there are questions like which part of the promoter is most important for realizing the function of a promoter - the binding regions alone or the rest of the promoter sequence also. To address some of these issues, *n-grams* as features are used in the experiments. Since these features are extracted considering whole promoter region, this can be termed as a global signal method for promoter recognition.

3.1 Introduction

Codon usage patterns in coding regions and hexamer conservation (TATA box, CATA box) in promoter region is well known. Techniques that use these concepts are available in abundance in literature. Most of the local content based methods utilize signal assuming there is local conservation of the hexamers [47, 70]. In literature there are a few reports on protein sequence classification, gene identification using *n-grams*, but very few on promoter recognition. An *n-gram* is a

selection of n contiguous characters from a given character stream [101]. Ohler et al. have used interpolated Markov chains on human and *Drosophila* as positive data set achieving a performance accuracy of 53% [85]. Ben-gal et al. have used a variable order Bayesian network which looks at the statistical dependencies between adjacent base pairs to achieve a true positive rate of 47.56% [11]. Leu et al. have developed a vertebrate promoter prediction system with cluster computing, extracting n -grams for $n=6$ to 20 [61]. They have not specified the reason for choosing particular n values 6 to 20. They achieved an accuracy rate of 88%. Ji et al. have used SVM and n -grams ($n=4,5,6,7$) for target gene prediction of *Arabidopsis* [52]. The 'n' values are chosen based on trial and error methods. They considered 500 n -grams which include 19 four-grams, 47 five-grams, 121 six-grams, and 313 seven-grams. Even though the computation of n -grams is straight forward, the efficacy of different n -grams ($n=2,3,4,5$) as features for promoter recognition posed as a binary classification problem has not been done systematically so far. Hence a study is taken up here to explore the possibility of this particular line of thought. As a problem of study we have chosen *E.coli* from prokaryotes and *Drosophila Melanogaster* from eukaryotes.

Extension of a promoter recognition method to find a promoter in whole genome is the natural expectation of a promoter recognition method. There are attempts by various groups to assess the accuracy of the predictions of some of these prediction algorithms for promoter [31, 8, 5]. Bajic et al. described that the prediction is termed as positive if the predicted transcription start site (TSS) falls within a maximum allowed distance from the reference transcription start site [8]. They have assessed performance of some of the prediction algorithms based on the performance measures such as sensitivity and positive predictive value. In their later paper they concluded that the promoter prediction combined with gene prediction yields a better recognition rate [5]. Sensitivity is found to vary from 32% to 58% and positive predictive value in the range of 79% to 93% for the ENCODE regions of human genome as part of the EGASP experiment [5]. They have found that the reduced promoter search space results in a smaller number of false positive predictions. It is imperative to say in this context, that the problem of promoter recognition is far from solved and is still an open.

A simple global feature extraction scheme that extracts an average signal of the entire promoter sequence of length L has been proposed in this chapter. The global features are the n -grams. This signal is used by a neural network to

achieve a comparable performance for the different non-promoter data sets that are proposed in literature. In-depth analysis of the classified and misclassified sequences using 2-grams/bi-grams, in promoter data set against the biological background throws up two distinct kinds of signals in promoter data set. For this analysis the machinery required is simple and uncomplicated. It is shown that simple 2-gram features form an adequate global signal to discriminate the promoter. The fact that this technique can be extended automatically to the eukaryotic promoter recognition is an added advantage. We also report the results obtained with regard to *Drosophila* promoter recognition.

3.2 Feature Extraction

Patterns or features that characterize a promoter/non-promoter are needed to be extracted from the given set of promoter and non-promoter sequences. Extraction of word features of different lengths is quite common in the sequence problems. As the conserved patterns in prokaryotes are of length 6, it would be natural to experiment with words of length 3 i.e. trinucleotide features. Here promoter recognition is addressed by looking at the global signal characterized by the frequency of occurrence of *n-grams* in the promoter region. Section 3.2.1 shows that these features perform well for prokaryotic as well as eukaryotic promoter recognition. To extract the global signal for a promoter, frequency of occurrence of *n-grams* is calculated on the DNA alphabet {A,T,G,C}. The set of *n-grams* for n=2 is 16 possible pairs such as AA, AT, AG, AC, TA, etc and the set of *n-grams* for n=3 are 64 triples such as AAA, AAT, AAG, AAC, ATA etc. Similarly *n-grams* for n=4,5,6 are calculated. Let f_i^n denote the frequency of occurrence of the i^{th} feature of *n-gram* for a particular n value and let $|S|$ denote the length of the sequence. The feature values v_i^n are normalized frequency counts given in Equation 3.1.

$$v_i^n = \frac{f_i^n}{|L| - (n - 1)}, \quad 1 \leq i \leq 4^N, \text{ for } n = 2, 3, 4, 5 \quad (3.1)$$

Here, the denominator denotes the number of *n-grams* that are possible in a sequence of length $|L|$ and hence v_i^N denotes the proportional frequency of occurrence of i^{th} feature for a particular N value. Thus each promoter and non-promoter sequence of the data set is represented as a 16-dimensional feature

vector $(v_1^2, v_2^2, \dots, v_{16}^2)$ for $n=2$, as a 64-dimensional feature vector $(v_1^3, v_2^3, \dots, v_{64}^3)$ for $n=3$, as a 256-dimensional feature vector $(v_1^4, v_2^4, \dots, v_{256}^4)$ for $n=4$, as a 1024-dimensional feature vector $(v_1^5, v_2^5, \dots, v_{1024}^5)$ for $n=5$.

In a binary classification problem, the training set will be a mixture of both positive and negative data sets. Similarly the test set, also consisting of both positive and negative data is used to evaluate the performance of the classifier. A neural network classifier is trained using the n -grams of training set as input feature vectors and then the test set is evaluated using the same network. The positive and negative data sets for *E.coli* as described in Chapter 1.2.1 is used. In the case of non-promoter data set consisting of both gene and inter-gene portions, the proportion of positive data set to the negative data set is taken as 1:2. The average distance between positive and negative data sets for *E.coli* is portrayed in figures 3.1 to 3.4. The average distance is computed by taking sum of a particular n -gram over the entire data set. For *Drosophila*, data set available in BDP(Berkely Drosophila Project) is used as described in the introductory chapter is used.

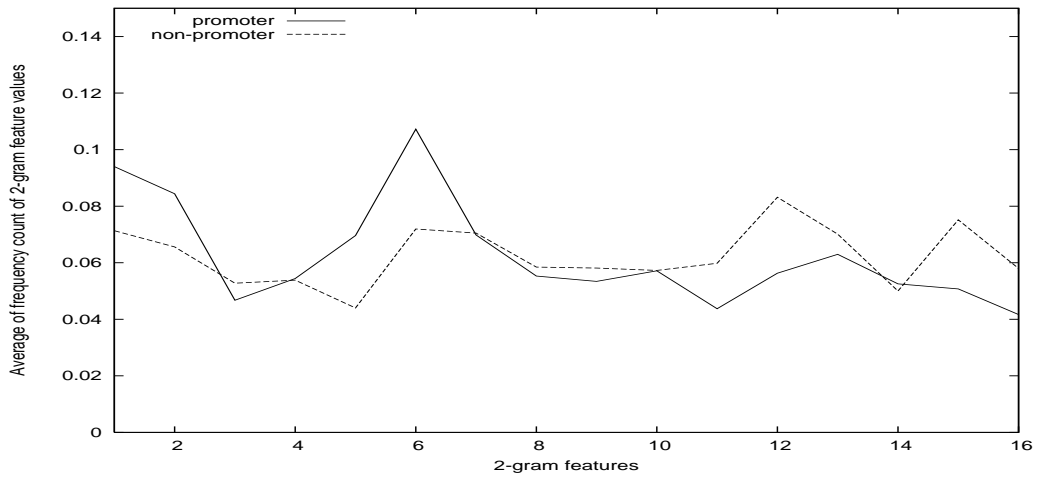


Figure 3.1: Average distance between promoter and non-promoter sequences using 2-grams for *E.coli*. On x-axis, 0...15 denote 2-grams AA, AT, AG, AC,, CG,CC.

Average distance profile shown in figures 3.1 to 3.4 portrays that promoter and non-promoter data sets are separable in the n -gram feature spaces. The efficacy of the features in classification task is presented in the following sections.

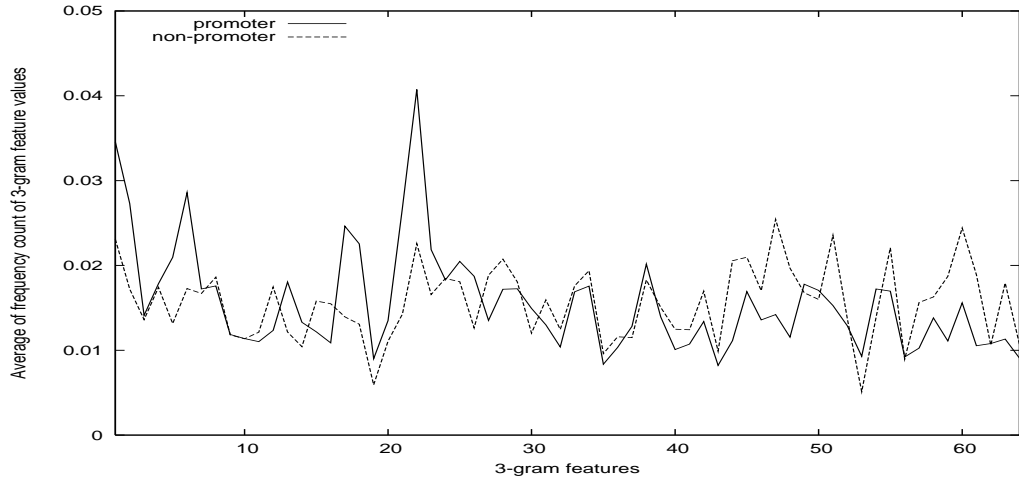


Figure 3.2: Average separation between promoter and non-promoter sequences for 3-grams for *E.coli*. On x-axis, 0...63 denote 3-grams AAA, AAT, AAG, AAC, ... , CCG, CCC.

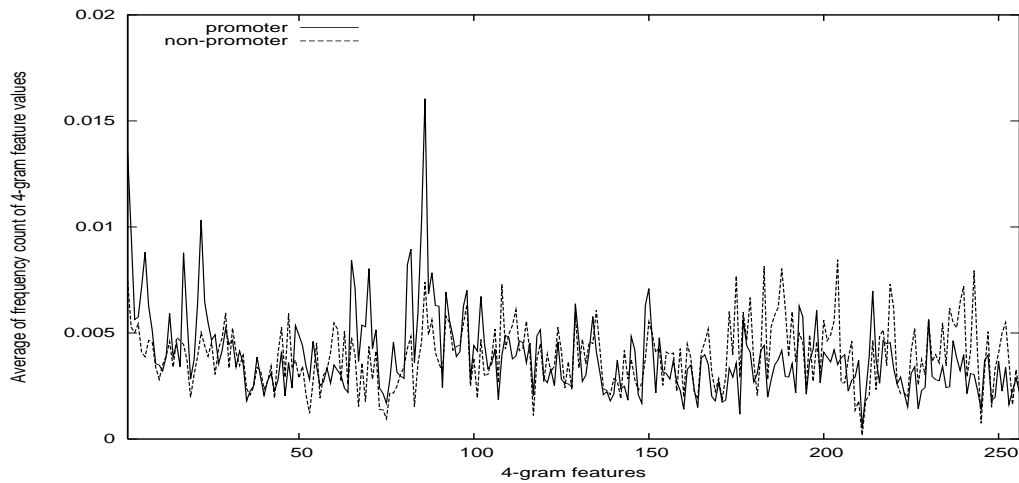


Figure 3.3: Average separation between promoter and non-promoter sequences for 4-grams for *E.coli*. On x-axis, 0...255 denote 4-grams AAAA, AAAT, AAAG, AAAC ,, CCGG, CCCC.

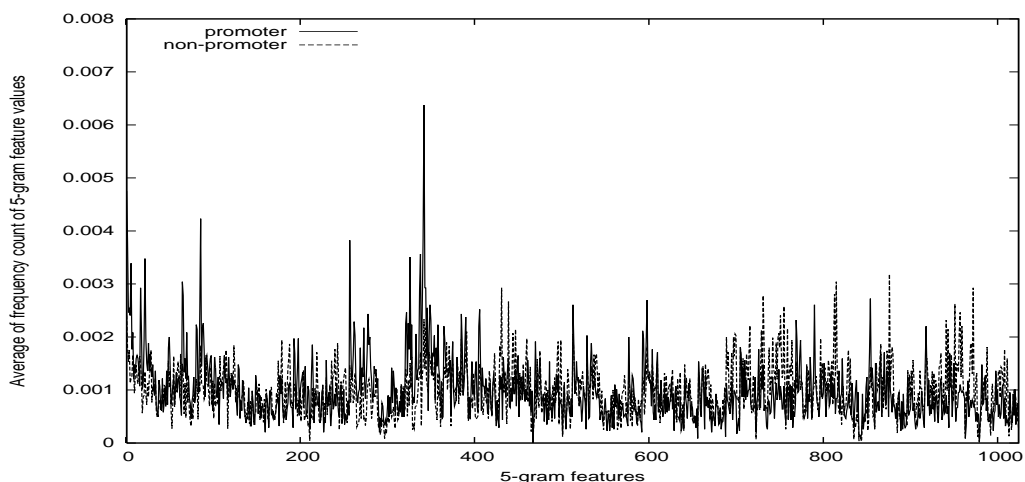


Figure 3.4: Average separation between promoter and non-promoter sequences using 5-grams for *E.coli*. On x-axis, 0...1023 denote 5-grams AAAAA, AAAAT, AAAAG, AAAAC, ..., CCCCCG, CCCCC.

3.2.1 Neural Network Architecture and Classification Performance for *E.coli* and *Drosophila*

A multi-layer feed-forward neural network with three layers, namely, an input layer, one hidden and an output layer is employed for promoter classification. The number of nodes in the input layer is 16, 64, 256, 1024, and 4096 features for $n=2, 3, 4$, and 5 respectively. Experimentation is done with different number of hidden nodes that give an optimal classification performance. The output layer has one node to give a binary decision as to whether the given input sequence is a promoter or non-promoter. 5-fold cross-validation [44, 77] is used to investigate the effect of various n -grams on promoter classification by a neural network. Average performance of all these folds is being reported. These simulations are done using Stuttgart Neural Network Simulator [104]. The classification results are evaluated using the performance measures such as *Precision*, *Specificity* and *Sensitivity* as defined earlier in Chapter 2. The classification results for various n -grams for *E.coli* are presented in Table 3.1.

In order to extend the proposed approach of n -gram features and a multi layer perceptron classifier to eukaryotes, *Drosophila* species is chosen. Eukaryotic promoter recognition is very different from the problem of prokaryotic promoter recognition. Eukaryotic promoters within a species and across the species

Table 3.1: *E.coli* classification results for different n-gram features. Average of 5-fold cross-validation.

<i>n-gram</i>	Precision	Specificity	Sensitivity	Positive predictive value
n=2-gram	76.6	85.44	63.3	67.38
n=3-gram	80.0	86.1	67.75	70.06
n=4-gram	76.8	81.78	72.66	65.6
n=5-gram	77.9	85.6	61.7	67.12

Table 3.2: *Drosophila* classification results for different n-gram features. Average of 5-fold cross-validation.

<i>n-gram</i>	Precision	Specificity	Sensitivity	Positive predictive value
n=2-gram	81.16	89.45	62.46	85.5
n=3-gram	85.83	91.36	72.01	89.28
n=4-gram	87.07	91.0	75.86	89.35
n=5-gram	86.41	93.62	68.4	91.2

have highly divergent promoter sequences [112]. They may or may not contain conserved patterns such as the TATA box. In fact, in case of highly regulated promoters, the extraneous regulation factors like enhancers etc. occur adjacent to promoters, thus making the discrimination difficult. In this context, as a first step to get a base estimate for eukaryotic promoter recognition, n-gram features are extracted from a promoter sequence of 300 base pairs [86]. Using the same classification scheme of *E.coli*, a neural network with backpropagation learning algorithm is used to train and classify a given *Drosophila* promoter and the classification results are given in Table 3.2.

The results show that 3-grams are the best discriminators in *E.coli* whereas, 4-grams are good in discriminating promoters and non-promoters in *Drosophila*. A further detailed analysis of the classification of promoters of *E.coli* is done, by taking 2-gram as basis. Here in addition to the biologically meaningful sequences proposed by Gordon, two more data sets are generated randomly with 60 and 50% A+T composition, used by some research groups [70]. By looking at these synthetic data sets, it can be inferred whether the positional dependence of the nucleotides has any impact on recognition of promoters. Each data set is divided into training and test data sets. Except the case of non-promoters comprising

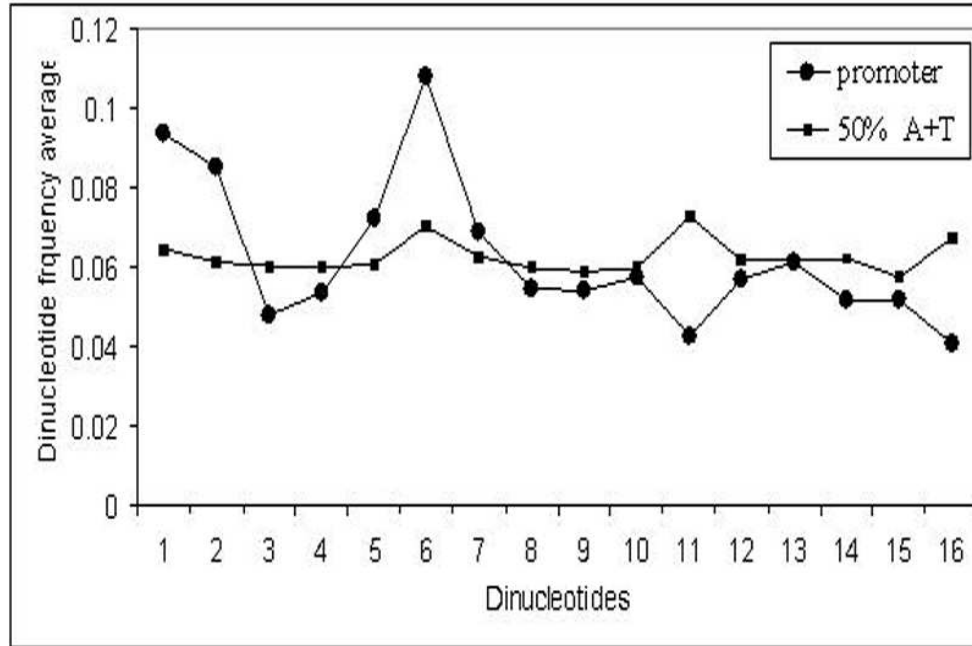


Figure 3.5: 2-gram frequency averages for promoter data set and 50% A+T rich synthetic negative data set.

of both gene and inter-gene portions, in all other cases the positive and negative data sets are taken to be of same size for training. In case of non-promoter data set consisting of both gene and inter-gene portions, the proportion of positive data set to the negative data set is taken as 1:2. Each promoter and the non-promoter sequence of the data set is encoded as a 2-gram feature vector by the method explained in section 3.2. The graph of the 2-gram frequency averages against each of the 2-gram features for the promoter data set and randomly generated 50% A+T rich negative data set is given in figure 3.5. The plot shows clear discrimination of the average signal of the promoter from the non-promoter for more than 50% of the features.

3.2.2 Classification with Different Negative Data Sets for *E.coli* and *Drosophila*

Again a multi-layer feed-forward neural network with three layers an input layer, one hidden and an output layer is used for promoter classification. The number of features in input layer is $(v_1^2, v_2^2, \dots, v_{16}^2)$ corresponding to the 2-gram feature

Table 3.3: *E.coli* promoter Classification for different negative data sets using 2-gram features.

Negative data set	Data set	Precision	Specificity	Sensitivity
Sequences from gene segments	1:1	78.4	79	80
60% A+T rich	1:1	95.5	98.18	93
50% A+T rich	1:1	97.3	99.3	95.8

vector. A hidden layer consisting of 48 hidden nodes is chosen and the output layer has one node to give a binary decision as to whether the given input sequence is a promoter or non-promoter. These simulations are done using Stuttgart Neural Network Simulator [104]. This neural network is trained on the training set and then the classification performance is evaluated on the test set. All the classification experiments are carried out using a 5-fold cross validation procedure [77].

The promoter data set of *Drosophila* is obtained from Eukaryotic promoter database EPD [30]. Positive data set contains sequences of length 241 bp with 200 base pairs upstream of the Transcription Start Site (TSS) and the rest downstream. The choice of length is made so as to include the Downstream Promoter Element which usually occurs around +30 in case of TATA less promoters and an upstream element which occurs at -200 [112]. A biologically meaningful negative data set is constructed by taking sequence fragments outside the promoter region, 300 sequence fragments from the coding region. Two more data sets are generated randomly with 60 and 50% A+T composition.

2-gram features of the data sets are extracted as given in section 3.2. A neural network with these sixteen 2-grams as input vector with a single hidden layer of 48 nodes and an output layer with a single node is constructed. 5-fold cross-validation is done on each of these sets and the results are given in table 3.4. Here also it can be observed that in the case of randomly generated data set the performance of the classifier is as good as 98.8%. When a biologically meaningful data set is considered the performance of the system is about 75%.

The classification precision achieved by the neural network is around 78% with biologically meaningful non-promoter data sets for *E.coli*. With randomly generated negative data sets the classifier achieves a much higher precision of

Table 3.4: Promoter Classification of *Drosophila* for different negative data sets using 2-gram features.

Negative data set	Precision	Specificity	Sensitivity
Sequences from gene segments	74.9	71.5	78.6
60% AT rich	98	98.8	97.4
50% AT rich	97.7	99.8	95.3

about 96% just as reported in the literature. For the sake of comparison a multi-layer perceptron is used in the case of synthetic data sets also. Otherwise same results are obtained without a hidden layer. Hence the promoters against a synthetic background are distinguishable to a high degree of accuracy. Though the use of synthetic data sets as a plausible negative data set is very much in question. This points to the fact that the composition of promoter by A, T, G, C is not a true indicator of a promoter, but the position where they occur do matter. Hence the use of position dependent n-grams as features is justified.

In the case of promoter and non-promoter sequences, negative set consisting of both gene and intergenic portions, in the ratio of 1:1, precision turns out to be 77.1%, specificity 75.69% and sensitivity 80.47%. It can be seen that even though a much better recognition of promoters is achieved, false positives increase compared to the case when the training data set is in the ratio of 1:2. Hence only the 1:2 case is used.

Promoter classification with an accuracy of near 96% is achievable by a single layer perceptron for synthetic negative data sets, potentially indicating the linear separability of the promoter data sets. The issue of linear separability has not been raised previously in the promoter classification literature and hence the problem is investigated further. In the subsequent sections it is shown that a single layer perceptron can achieve adequate classification accuracy with the biological negative data also under certain conditions.

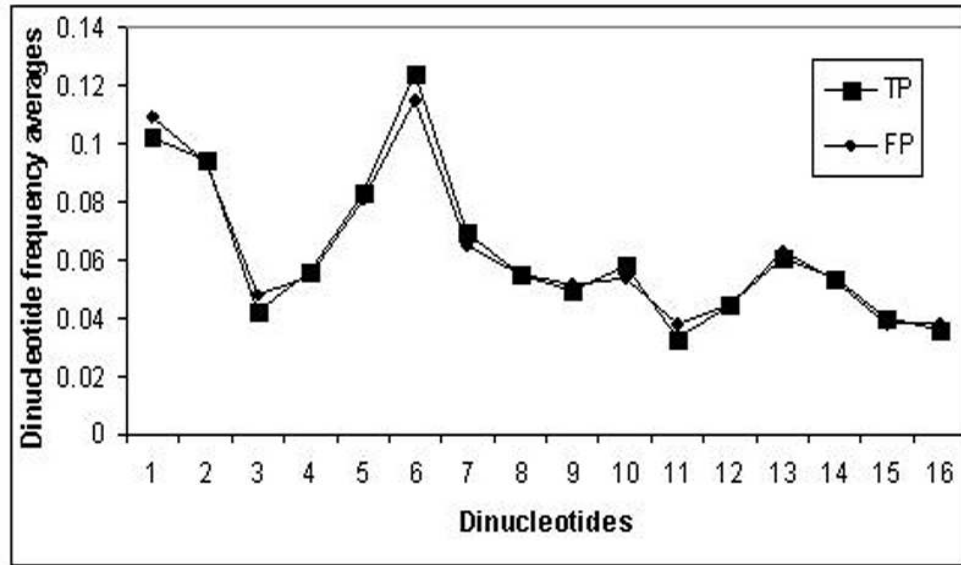


Figure 3.6: 2-gram frequency averages for correctly classified promoter data set and misclassified negative data set consisting of gene and inter-gene portions.

3.3 Analysis of Correctly Classified and Misclassified Promoter Sequences

It was evident from the various experiments carried out in the training phase of the neural network that irrespective of network architecture, such as the number of hidden layers and the number hidden nodes in each hidden layer, the network could not achieve a training performance beyond 85%. In order to carry out a deeper analysis of classification of promoters, a set of sequences are selected randomly from both promoters and non-promoters (consisting of both gene and inter-gene portions) in the ratio of 1:2 respectively for training. That is, a set of 454 sequences are taken from promoter data set as positive set, and 454 sequences are taken from each gene and inter-gene sequence sets. The rest of the data set is used as test data. Feature extraction and classification are done as described in section 3.2. The sets of misclassified sequences and correctly classified sequences are given a closer attention in this section.

To analyze this problem further, the sequences that the neural network finds difficult to learn are isolated from both positive and negative data sets. Promoter sequences that are classified correctly, i.e. true positives (TP), misclassified (FN) and the non-promoter sequences that are correctly classified (TN) and misclassi-

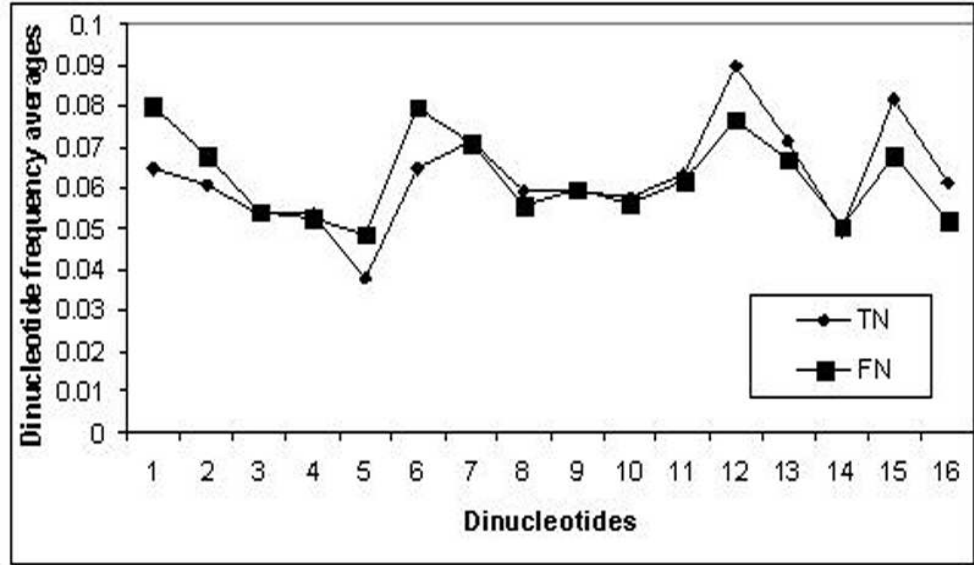


Figure 3.7: 2-gram frequency averages for misclassified promoter data set and correctly classified negative data set consisting of gene and inter-gene portions.

fied (FP) are treated as four classes. A 4-way classifier is designed to check the assumption that there may exist four different signals. But, the results of this classification show that TP comprising of a majority of promoter sequences and a minor portion of non-promoter sequences (FP) are grouped as one class. Similarly TN and FN are grouped as another class. We can clearly illustrate these results by plotting the 2-gram frequency averages of these four sets against the negative data set consisting of gene and inter-gene portions. Figure 3.6 shows the closeness of TP and FP which is measured by the Euclidean distance between corresponding graphs. Figure 3.7 shows the graph of TN versus FN. Additionally figure 3.8 depicts the distances between TP and TN as well as FP and FN. For completeness of discussion we give 2-gram frequency average plots for other negative data sets as well.

For each 2-gram feature i , the average of frequency values v_i over an entire data set D , is computed as $a_{i,D}$, where $|D|$ denotes the size of the data set.

$$a_{i,D} = \frac{1}{|D|} \sum_{j=1}^{|D|} v_{i,j} \quad (3.2)$$

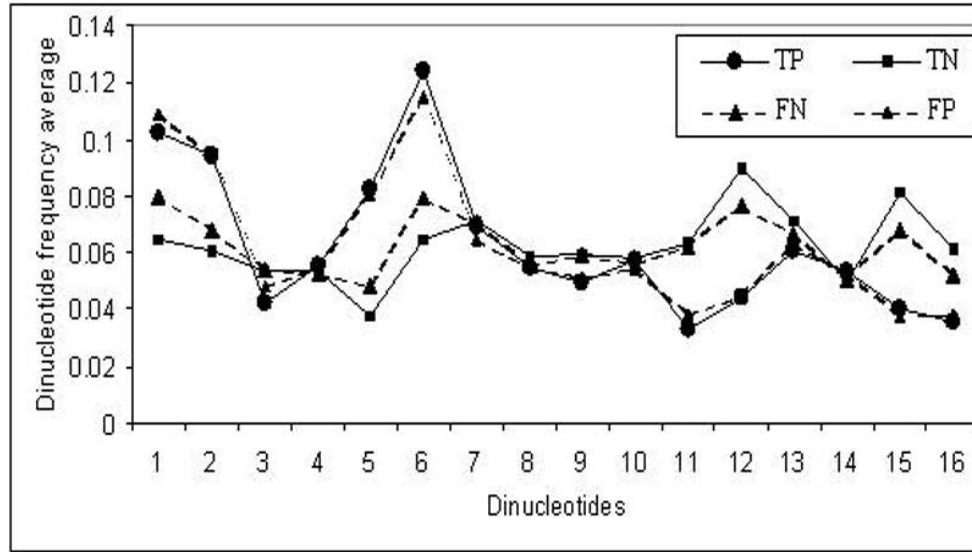


Figure 3.8: 2-gram frequency averages for promoter data set and negative data set consisting of segments from gene and inter-gene portions of the DNA.

The average Euclidean distance between two data sets D_1 and D_2 is given by

$$d(D_1, D_2) = \sqrt{\sum_{i=1}^{16} (a_{i,D_1} - a_{i,D_2})^2} \quad (3.3)$$

For example, the Euclidean distance between TP and FP is denoted as $d(TP, FP)$ etc. Distances between these data sets TPs, TNs, FPs, FNs are presented in table 3.5.

Both the figures 3.8 and 3.9 corresponding to the two biological negative data sets show that the correctly classified promoter sequences and the misclassified non-promoter sequences are close and also misclassified promoter and correctly classified non-promoter sequences are close. It is interesting to note that $d(TP, FP)$ and $d(TN, TN)$ are much smaller than $d(TP, TN)$ and $d(FP, FN)$. These results clearly demonstrate that there is a small confusion set in both the promoter and non-promoter data sets. That is, there exist promoter like non-promoters and non-promoter like promoters. This insight gives us the motivation to dissolve the confusion set by constructing two data sets one which captures 'majority' promoter class and the other that reflects the 'minority' signal. This is taken up in the next section.

It can be noticed that a preliminary evidence for linear separability of pro-

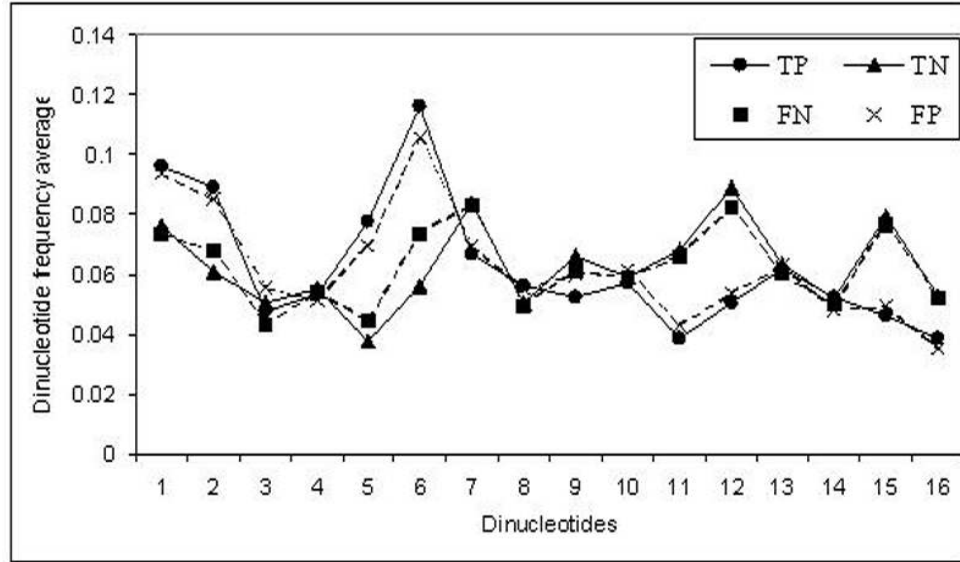


Figure 3.9: 2-gram frequency averages for promoter data set and negative data set consisting of gene segments from DNA.

Table 3.5: Distances between TP,TN,FP and FN

distance	value
d(TP,FP)	0.015962
d(TN,FN)	0.033089
d(TP,TN)	0.112933
d(FP,FN)	0.081533
d(TP,FN)	0.087181
d(TN,FP)	0.112983

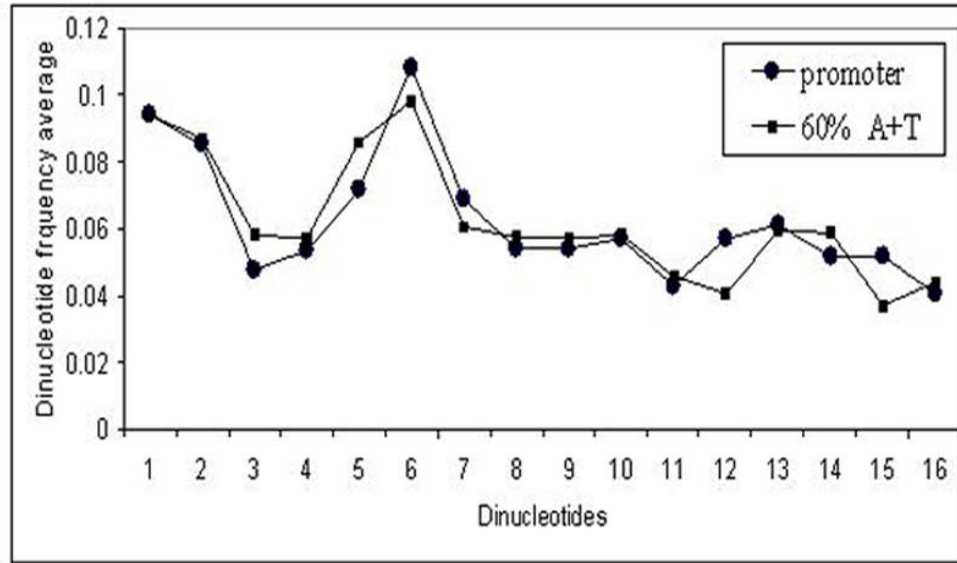


Figure 3.10: 2-gram frequency averages for promoter data set and 60% A+T rich synthetic negative data set.

moter data set against a synthetic negative background is already pointed out in an earlier section 3.2.1. Figure 3.10 depicts the 2-gram frequency averages for promoter and 60% A+T rich negative data set having Euclidean distance as small as 0.031599. The analysis in the next section shows that while biological data sets suffer from a confusion set, the promoters are easier to discriminate against a synthetic negative background. This possibly explains why the results of classification accuracy in the literature are much higher for synthetic negative data sets [70]. In fact this motivates us into regrouping the data set by moving away the confusion set and investigating the possibility of single layer perceptron for classification against biological negative data.

3.3.1 A Single Layer Perceptron for Promoter Recognition

The data sets are regrouped into two sets: sequences that are correctly classified by the neural network are called *Major Set (Maj)*. Similarly the sequences that are misclassified are grouped as *Minor Set (Min)*. Using these two data sets, two separate new neural networks NN_{Maj} and NN_{Min} are constructed. NN_{Maj} is trained only on *Maj* and similarly NN_{Min} is also trained on only *Min*. These

neural networks achieve 100% training performance with *no hidden layer*. That is the *Maj* class constituting the majority promoter and the majority non-promoters is linearly separable. Similarly the *Min* set is also linearly separable.

At this point it is interesting to evaluate the overall classification performance of *Maj* and *Min* classifiers with the original test data without the major and minor segregation. Results of one fold are shown in table 3.6. NN_{Maj} achieves 80% precision while NN_{Min} recognizes the rest of the 20% (up to a max of 3% error) correctly. Moreover, the sequences that are the TPs and TNs of NN_{Maj} become the false negatives and the false positives respectively of NN_{Min} . In this way, NN_{Min} behaves in a kind of complementary fashion to NN_{Maj} on the test set. To summarize, our results demonstrate that by segregating the promoter data into major and minor classes, it is possible to construct linearly separable classifiers.

Table 3.6: Test data results of neural networks NN_{Maj} and NN_{Min}

Positive Test Data (156) TP of NN_{Maj} =107 \supseteq FN of NN_{Min} =101	Negative Test Data (392) TN of NN_{Maj} = 334 \subseteq FP of NN_{Min} = 337
FN of NN_{Maj} =49 \subseteq TP of NN_{Min} =55	FP of NN_{Maj} = 58 \supseteq TN of NN_{Min} = 55

It is to be noted that not only the numbers in each of the boxes of table 3.6 match, but also the exact sequences in each of these boxes match. For example, the set of 101 FNs of NN_{Min} is a subset of the 107 TPs of NN_{Maj} . Thus we see that both promoter and the non-promoter sequences have two distinct patterns, one being recognized by NN_{Maj} and the other by NN_{Min} . But for a confusion of 5 to 7 sequences the NN_{Maj} and NN_{Min} behave in a complementary fashion. A small portion (14%) of the non-promoter data set is similar to a majority (70%) of the promoter data set and also 86% of the non-promoter data set (TN) is closer to 30% of the promoter data set (FN).

Huerta et al. name promoters which have signal close to the consensus pattern at the binding sites as strong promoters and others as weak promoters [47]. Our analysis also gives a clear indication of two kinds of promoters. A majority of

the promoters being easily distinguishable from the background we call as strong promoters, and small portion having a signal closer to the non-promoters as weak promoters. It is to be emphasized that in this process we have successfully built a neural network NN_{Maj} which is a single layer perceptron achieving promoter recognition performance of 80% which is comparable to the powerful classifiers that are presented in the literature [36].

3.4 Genome-Wide Promoter Recognition

One of the main goals of promoter recognition is to locate promoter regions in the genome sequence. In this section, a scheme for locating promoters in a given DNA sequence segment of *E.coli* genome of length N in a particular direction (say, 5' to 3') is proposed. The scheme do not address the issue of locating TSS in the promoter region. Going along with the scheme for classification by the neural network NN_{Maj} , a moving window of length 80 is considered to extract segments from the start of the DNA sequence, that is, 1–80, 2–81, 3–82 and so on. These are represented as the 16-length 2-gram feature vectors which are fed to the neural network classifier. Each of the segments gets classified as promoter (P) or non-promoter(NP). If a segment $m - (m + 79)$ is classified as a promoter, then the nucleotide m is annotated as P and if it is classified as non-promoter then m is annotated as NP . This process of annotation is continued for the entire sequence to get a sequence of P 's and NP 's. We propose that if a contiguous segment of length more than a certain threshold has all P 's then we annotate that region as promoter region otherwise as non-promoter region.

3.4.1 A Preliminary study

This algorithm is implemented on a segment of length 10,000 base pairs taken from the NCBI database as a preliminary exercise [20]. NCBI database has Whole genome sequence of *E.coli* experimentally determined by Blattner et al. It has annotated features such as promoter, TSS, gene, cds, repeat region, binding region, forward or reverse strand. Whole genome of *E.coli* is divided into 400 sections and each is denoted as *section1*, *section2*, *section3*, ..., *section400*. *E.coli* is given the number U00096. Each section is given a separate *gi* number. We have chosen *section1* for our preliminary study. We find that all the promoter

regions that are located around the TSS at 106, 139, 5084, 8209 and 9279 in *section1* [20] are identified by this simple annotation scheme. For example, all the nucleotides from 1 to 132 are annotated as P and the annotation reported in the NCBI database confirms 71–99 and 104–132 as promoter regions. We also obtain short lengths like 2214–2219 as P but if we set an appropriate threshold these short spurious promoters can be rejected. With a threshold set at say, 20 base pairs, all the true positives of the 10,000 length sequence are recognized by our scheme and four false positives are obtained which are not reported in the NCBI site. It is to be emphasized that just a segment of length 80 is not sufficient to say if it contains a promoter region. Since a moving window scheme is being applied one needs a minimum length of 140 if a contiguous segment of P 's of length 20 is to be found even starting at the 60th place in the sequence. The rate of performance in any case is tied up with the classification performance of the classifier and also it is possible that the false positives so obtained are in fact unknown promoters. The number of times the classifier needs to be run on the sequence is $N - 80$.

Satisfactory results are obtained when the threshold is set at 30. Also experimentation is carried out with different kinds of windows including that of consecutive segments of length 80, that is, 1–80, 81–160 and so on. This windowing scheme obtains a false negative by identifying only four out of the five promoter regions correctly. Hence it is reasonable to adopt an intermediate windowing scheme and also thus reducing the time complexity, by leaving 30 out instead of leaving one out for which classifier needs to only run for $N/30$ times and this scheme does achieve satisfactory results as reported above. We considered window centered at 1 whereas it is more common to consider windows actually centered in the middle and the central nucleotide being annotated accordingly. The promoter region is not really symmetric with respect to the TSS which is located at around 60th position and the two binding sites at either end being located approximately at 25 and 50 positions, respectively. Hence the central nucleotide may not actually belong to either of the binding sites. The scheme of annotation certainly needs to be fine-tuned by considering larger lengths of genome sequences and an appropriate threshold chosen after experimentation on these larger data sets.

3.4.2 Genome-Wide Promoter Recognition Using Neural Networks Based on Promoter-Coding and Promoter-Noncoding

When the negative data set is a combination of both coding and non-coding segments, it is advantageous in the sense that the promoter and non-promoter could be classified in one go. But, the classification accuracy is not 100% and there is no way one can eliminate the false positives and false negatives. To overcome this handicap, instead of using the earlier neural networks, a new set of neural networks based on different combinations of the data sets are designed. One network NN_{PC} is trained using promoter and coding data sets as positive and negative data sets respectively and another one NN_{PN} using promoter and noncoding data sets. The main idea behind this segregation is that the network based on promoter and non-coding data NN_{PN} would identify the promoter and non-coding data sets with a certain accuracy. The coding data also would be classified as a promoter or a non-coding sequence by the network.

The outcome of this network, i.e. sequences that are classified as positives are given as input to the network based on promoter and coding data. The assumption is that a small number of coding and non-coding sequences that were classified by the earlier network as positives will be classified as negatives by NN_{PC} . The sequences for which the output of the network NN_{PC} is less than 0.5 would be recognized as coding segments, and the ones that give an output greater than 0.5 would be promoters. The basic idea is illustrated in Figure 3.11. The features chosen for the networks are the 3-grams which were found to give best recognition in *E.coli* case as given in Table 3.1.

The network NN_{PC} is trained by using 3-gram features extracted as explained in Section 3.2. The positive data set is the promoter data of Gordon et al. and negative data set is the coding data of Gordon et al. [36]. 5-fold cross-validation is done for this set of data. And similarly NN_{PN} is trained by using 3-grams as features. In this case, positive data set is the same as earlier, but negative data set is the non-coding data of Gordon et al. The best network architecture is chosen to do the 5-fold cross-validation. 5-fold cross validation for promoter and non-coding gives sensitivity of 74.73%, and specificity of 82.23%. Similarly, network for promoter and coding give sensitivity of 72.18% and specificity of 89.7%. Using these network configurations for one fold, *section1* and *section3*

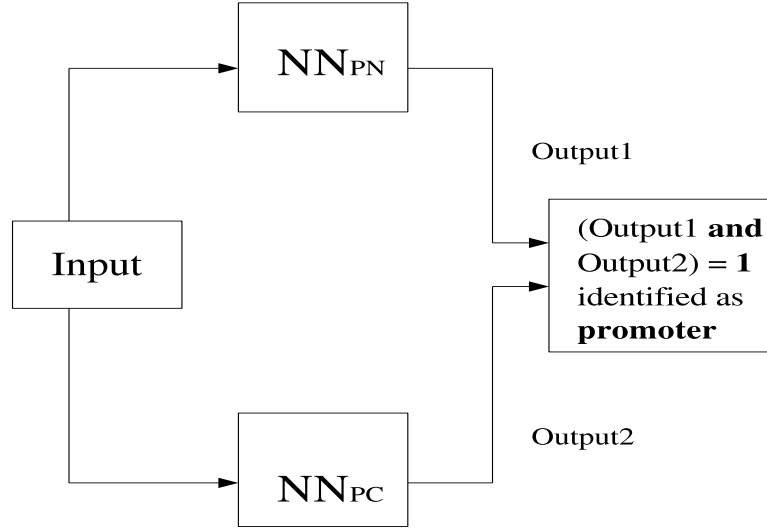


Figure 3.11: Scheme for promoter recognition in whole genome using networks NN_{PC} and NN_{PN} .

segments of *E.coli* are tested. A typical *section* namely, *section3* is given in Appendix D.

To estimate the promoter region in the given genome, first NN_{PN} is applied, then NN_{PC} is applied. The reason for adopting such a cascading system is that as already explained earlier, when NN_{PN} is applied, the classification of coding part is not very definite in the sense that they could be classified as promoters or non-coding parts. Hence the ones which are annotated as positive in this could be promoters, coding or few non-coding also. Similarly the negatives also can be non-coding in majority, coding and few promoters. When the NN_{PC} is applied, again a positive outcome could be a promoter or a non-promoter identified as a promoter. To identify a promoter, a sequence which is identified as positive

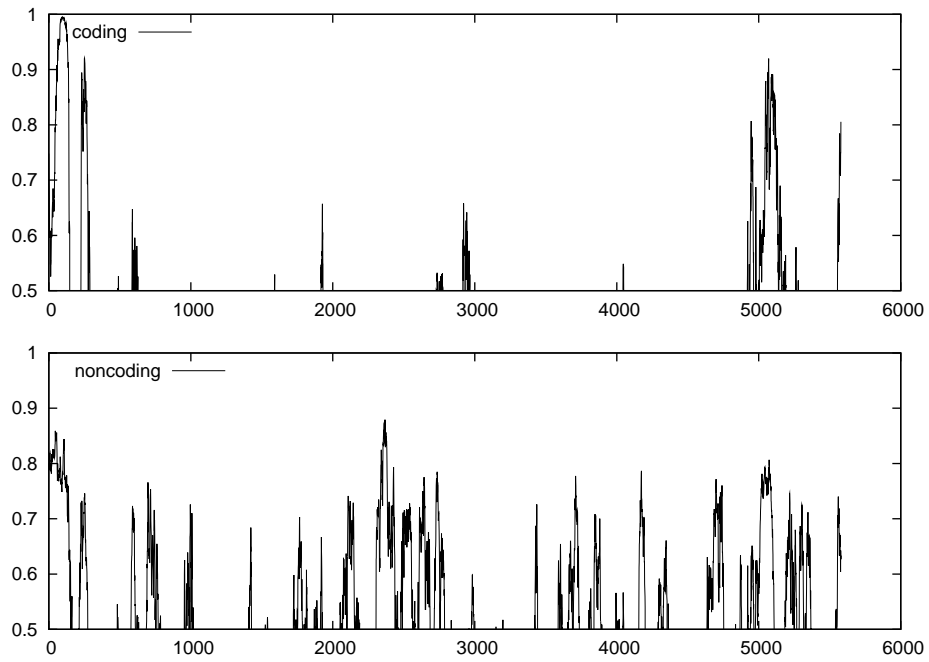


Figure 3.12: The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for *section1* of *E.coli* genome.

by both networks is treated as a positive outcome. A stretch of these outcomes greater than a threshold such as 50 consecutive positive outcomes is treated as a promoter.

Whole genome of *E.coli* is divided into 400 sections. Using the method described above to segment the the genome into portions of length 80 bp, *section1* and *section3* of *E.coli*, out of the 400 sections comprising the whole genome of *E.coli* are considered [20]. This *section* is partitioned into segments of length 80 using the moving window scheme explained above. Using NN_{PN} first and then NN_{PC} on these segments of length of 80 bp gives the results as shown in figures 3.12, 3.13 and 3.14. X-axis gives the 80 bp length segments obtained by using the moving window and y-axis represents the output of the neural network. The computation is applied to forward direction only here. Same can be applied to reverse strand also, after preprocessing wherein a reverse strand is converted to a forward strand.

To estimate the promoter region in the given genome, first NN_{PN} is applied then NN_{PC} is applied. The reason for adopting such a cascading system is as

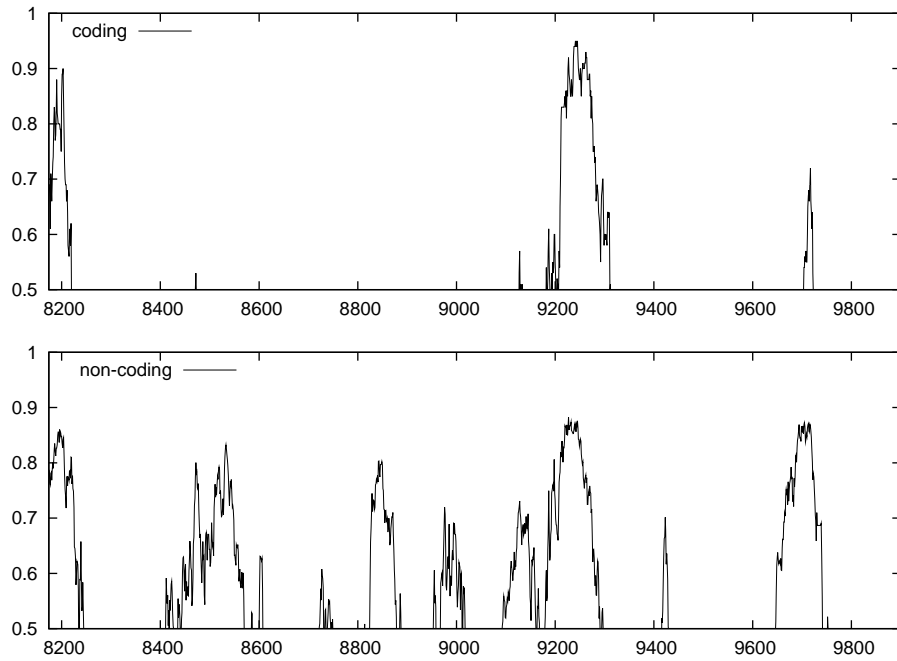


Figure 3.13: The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for *section1* of *E.coli* genome.

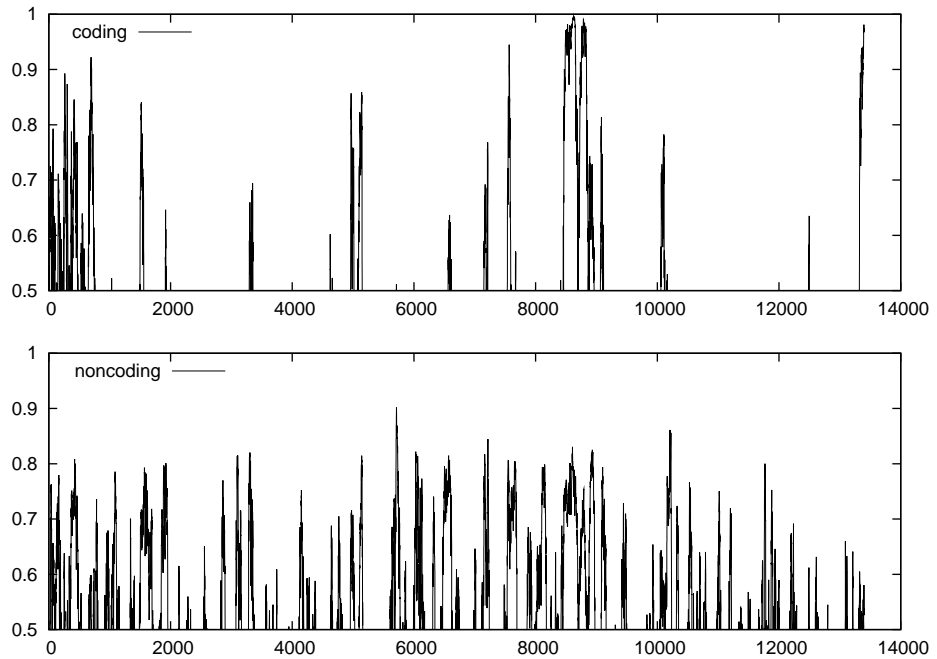


Figure 3.14: The outputs of the the networks NN_{PC} (Top panel), NN_{PN} (Bottom panel) versus the moving window for *section3* of *E.coli* genome.

already explained earlier, when NN_{PN} is applied, the classification of coding part is not very definite in the sense that they could be classified as promoters or non-coding parts. Hence the ones which are annotated as positive in this could be promoters, coding or few non-coding also. Similarly the negatives also can be non-coding in majority, coding and few promoters. When the NN_{PC} is applied to positive outcomes in the earlier network, a promoter is reinforced, and a non-promoter would be repressed. Table 3.7 shows the section number, the extent of promoter region, the sigma factor and the extent of the region identified as promoter from the experiment. There are few stretches where it is positive, but these stretches are very small in size, just 4 or 5 in length, barring these, there is one peak around, 4940-4965. This is not there in the annotation. Whether it is a new promoter is to be determined. Similarly, in *section3*, regions 3296-3355, 7151-7188, 7537-7590, 8454-8708, 9078-9106 are not accounted for in the NCBI data. Proper threshold has to be chosen to consider or for considering them as potential promoters. The main point, that is emphasized here is that the method seems to have 100% promoter recognition rate and some other sigma factors are also being identified. It is deemed to be essential that no true promoter is missed in the genome wide analysis.

Figures 3.12, 3.13 and 3.14 depict the fact that the cascaded network scheme is effective in really identifying the coding regions. For example, *section1* has an operon at the starting of the section, and there are genes spanning the segment from 317 to 5020. In Figure 3.12 plot shows the output of network NN_{PN} against a moving window, identifying lot of regions as promoters in the region 1-5600 bp region of *section1*. When NN_{PC} is tested on these results, it identifies the coding regions and there are very few regions left that would classify as promoters.

There are not many whole genome promoter prediction programs for prokaryotes. Gordon et al. have developed a whole genome promoter prediction using SAK. *section3* of E.coli prediction using 3-grams using NN_{PC} and *section3* predictions using SAK are presented in Figure 3.15 [55]. SAK predicts whether the 61st position in sequence of length 80 is a TSS or not. The legend of SAK says that more positive the outcome is, more probable that segment is a promoter and a promoter is denoted by 'L'. There are no such annotations in the test case. But, it can be seen that most of the peaks coincide. The other promoter prediction packages for bacteria are tested with *section3* data of *E.coli* and the results are shown in table 3.8. In case of BPROM an internal parameter, threshold for

Table 3.7: Test data results of neural networks NN_{PC} , NN_{PN}

Section	factor	Extent from NCBI start..end	from our experiment start..end
<i>section1</i>	Sigma70	71..99	0-148
<i>section1</i>	Sigma70	104..132	0-148
<i>section1</i>	Sigma32	188..212	228-277 ^a
<i>section1</i>	Sigma70	5050..5077	5004-5109
<i>section1</i>	Sigma70	8174..8202	8174-8220
<i>section1</i>	Sigma70	9244..9272	9208-9311
<i>section3</i>	Sigma70	452..483	420-478
<i>section3</i>	Sigma70	1534..1564	1497-1561
<i>section3</i>	Sigma54	5003..5020	4962-5015 ^b
<i>section3</i>	Sigma70	5014..5041	5101-5150
<i>section3</i>	Sigma70	6526..6552	6563-6605
<i>section3</i>	Sigma70	6553..6580	6563-6605
<i>section3</i>	Sigma70	6569..6597	6563-6605
<i>section3</i>	Sigma70	6595..6623	6563-6605
<i>section3</i>	Sigma70	7644..7676	7537-7590
<i>section3</i>	Sigma70	8856..8884	8714-8838
<i>section3</i>	Sigma70	8925..8954	8872-8961
<i>section3</i>	Sigma70	10075..10105	10061-10125
<i>section3</i>	Sigma70	13370..13401	13378-13384, 13390-13401

^aEventhough this is a sigma32 promoter, a peak is observed at this position

^bEventhough this is a sigma54 promoter, a peak is observed at this position

promoters is set as 0.20 [84]. This has predicted 31 promoters in this region. BPROM predicts the TSS and also determines the binding regions. In the table only the TSS is indicated for BPROM. NNPP predictions are made with a cut-off rate 0.80 [82]. The table shows starting and ending of the promoter by the software.

From Table 3.8, we can get an estimation of measures like sensitivity and positive predictive rate. Table 3.9 presents a summary of these results. The results give a clear indication that cascaded networks based on 3-grams outperform the tools. One more factor is that, unless we know, where a promoter region exists, we cannot say an outcome from these other tools as positive since there are so many false positives. In our scheme, wherever we find a stretch of positives, using cascaded method, gives us a positive prediction.

Table 3.8: Prediction results of *section3* of *E.coli* using different promoter prediction packages.

NCBI	SAK	BPROM	NNPP	n-gram
452-483	80-93	325	284-329	420-478*
1534-1564	199-210	725	379-424	1497-1561*
5003-5020	215-228	1112	570-615	3296-3355
5014-5041	324-329	1561*	1533-1578*	4962-5015*
6526-6552	365-374	1861	1819-1864	5101-5150*
6553-6580	388-397	2456	1933-1978	6563-6605*
6569-6597	411-424	2950	3300-3345	7151-7188
6595-6623	442-455*	3612	3571-3616	7537-7590*
7644-7676	476-487	3930	3610-3655	8454-8708
8856-8884	492-500	4355	3889-3934	8714-8838*
8925-8954	503-509	4697	4162-4207	8872-8961*
10075-10105	512-521	5084*	4319-4364	9078-9106
13370-13401	528-534	5469	4773-4818	10061-10125*
	719-728	5791	5009-5054*	13378-13384*
	777-784	6445	5585-5630	13390-13401*
	1567-1581*	6759	5888-5933	
	1588-1592	7069	6404-6449	
	1698-1711	7628*	6519-6564*	
	1927-1942	7932	6553-6598*	
	1998-2003	8608	6616-6661*	
	5052-5055*	8962*	7028-7073	
	6638-6658*	9352	7188-7233	
	7067-7070	9855	7587-7632*	
	7224-7236	10196	7689-7734	
	7244-7255	10667	8084-8129	
	7276-7280	11129	8361-8406	
	7595-7655*	11435	8428-8473	
	7680-7694	11949	8496-8541	
	7701-7705	12448	8514-8559	
	7979-7987	12863	8523-8568	
	8469-8472	13451	8543-8588	
	8525-8530		8569-8614	
	8530-8778		8597-8642	
	8772-8854*		8614-8659	
	8863-8874*		8625-8670	
	8883-8897		8665-8710	
	8928-8939*		8851-8896*	
	8943-8989		8887-8932	
	8999-9010		8921-8966*	
	9135-9186		8942-8987	
	10004-10013*		8962-9007	
	10114-10123		10500-10545	
	10136-10139		10626-10671	
	10146-10151		12247-12292	
	10184-10187		12822-12867	
	13416-13446		13236-13281	
			13382-13427*	
			13410-13455	
	* denotes probable promoters			

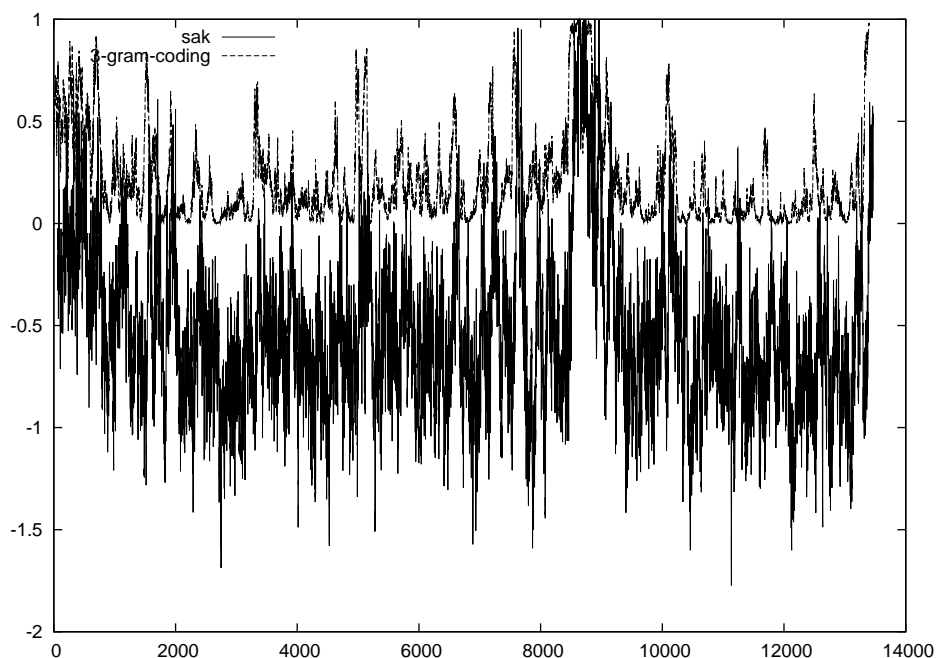


Figure 3.15: The outputs of the the network NN_{PC} and the output generated by SAK versus the moving window for *section3* of *E.coli* genome.

The same exercise is repeated with *Drosophila* genome. A stretch of 10 kbp is used here. This portion starts with a gene followed by two more genes. The same NN_{PC} and NN_{PN} networks are developed for *Drosophila* also. Applying these networks, as described in Section 3.4.2, four promoters are identified. Two promoters before the genes and one intron and one exon. The stretches of intron and exon portions identified as promoters are about 150 bp in length, whereas the portions identified before the genes are of length ≈ 350 bp. The threshold has to be set more rigorously, by more experimentation. NNPP2.2 is used for cross-checking the method, since the training and testing data used by this package is used training and testing by n-grams method. NNPP2.2 software predicted 21 promoters with a cutoff of 0.8, 8 with a cutoff of 0.95 and only one with a cutoff of 0.99. In this, the cutoff plays a crucial role. The other software package FirstEF did not predict any promoter at all.

Table 3.9: Summary of promoter recognition results on *section3* of *E.coli* using different software tools.

Package	sensitivity%	Positive predictive value%
SAK	69.23	19.5
NNPP	76.923	20.83
BPROM	38.46	16.129
3-gram	100	76.47

3.5 Discussion

A positive conclusion that can be drawn from the whole experimentation is that with simple features and using a low-end supervised learning algorithm, the classification results are comparable to that reported in the literature [36, 70]. The emphasis here is on the fact that a simple scheme without any prior knowledge of the data set like the number of binding sites, the range of spacer lengths between the binding sites achieves a rate of recognition comparable to several methods proposed in the literature. It is to be noted that the combination of coding and non-coding sequences together as negative data set has not been tried by Gordon et al. Our recommendation is that to get a realistic background it is better to consider them together rather than individually. To the best of our knowledge such a scheme achieving this kind of precision in recognition has not been reported so far in the literature. This scheme of feature extraction and classification has been extended to the problem of *Drosophila* promoter recognition in section 3.2.1.

The results of *n-gram* analysis of *E.coli* show that 3-grams attain better precision than the other *n-grams* in discriminating the promoter against a background consisting of gene and inter-gene segments. This reinforces the idea that codon usage pattern may play a role in distinguishing a promoter. This fact is even supported by the findings of Li [65]. Periodicity of 3 is observed not only in coding regions but also in the non-coding regions. The higher *n-grams* are not very efficient in classification of promoters. In fact the sensitivity drops as the size of *n-grams* grows to 5. It can be speculated that since the length of the promoter considered is only 80 bp, the 5-grams that occupy an input vector of size 1024 (all possible 5-grams) are few, hence the sensitivity is less. To verify the surmise, Euclidean distance is used as discriminating measure to rank the

features and top ranking features are used as input to the neural network. It was found that the reduction in features hasnot helped the classification any better. It can be concluded that the features are not powerful enough to discriminate promoters and non-promoters. The conclusions that are being drawn are strictly pertinent to the data set and the classifier that is used to address the promoter recognition problem.

The results of *Drosophila* present a different picture. Here it can be seen that the 4-grams are much better in distinguishing the promoters against non-promoters. Trifonov et al. have found a periodicity of 3 component in eukaryotes indicating the codon usage pattern used in translation of the mRNA to amino acids [108]. This has been observed in exons only but not in introns. This is confirmed by taking promoter and cds sequences as positive and negative data sets. In this case a high rate of recognition can be obtained. Now, the question is about the features that would distinguish the promoter and introns. It is indicated in many experiments that the intron portion is similar to the non-coding portion. It is evident that from the results that the 4-grams or for that matter any *n-gram* composition is different for the non-coding and intron portions even though they are supposed to be similar. Hence, this result is of significant importance.

The results of different softwares for a particular section of *E.coli* are presented. The results indicate that other software tools predict lot many promoters than the method proposed in Section 3.4.2.

3.6 Summary

The major thrust here is to estimate the recognition accuracy of a simple classifier for the promoter recognition problem which uses n-gram features as input. A study of *n-grams* ($n=2,3,4,5$) as features for a neural network classifier is done. The preliminary results shows that for *E.coli* 3 – grams give better performance than other *n-grams* whereas for *Drosophila* 4 – grams give an optimal performance. Using the 3-grams which gave best recognition rate out of the n-gram features considered, a genome-wide promoter recognition is attempted in a limited portion. The method seems to identify all promoters in the test cases.

A multi-layer perceptron with one hidden layer using features as small as

2-grams can achieve a binary classification with a precision of around 80% for biologically meaningful non-promoter data set as background, which is comparable to the performance of the high powered classifiers using heavy feature extraction schemes that are presented in the literature. Further, it achieves a precision for promoter recognition of 96% against the background of synthetic non-promoter data sets. This demonstrates the strength of n-grams as features for promoter recognition. This method is generic enough to extend to the eukaryotic promoter recognition also. This is a major advantage since the scheme is not dependent on any prior data and hence it is extendable.

One of the other claims is that, in fact, a single layer perceptron can be constructed which achieves as good a performance as the multi-layer perceptron. It is shown that there are two kinds of promoters: a major set and a minor set. A major set of promoters and a minor set of non-promoters have a similar signal in 2-gram feature space. Similarly a minor set of promoters and a major set of non-promoters have similar signal. This analysis of the data set helps us in building another neural network called NN_{Maj} which is a single layer perceptron that achieves 100% recognition on the major set. Thus the majority class constituting the majority promoter and the majority non-promoter data sets is linearly separable. If the test data is not labeled as belonging to a major or a minor set, NN_{Maj} achieves an overall precision of 80% for the promoter classification problem. This analysis demonstrates the usefulness of sub-classifying promoters before building a classifier.

There are very few promoter annotation techniques for prokaryotes. One of them is SAK developed by Gordon et al. [36]. Others such as NNPP use local information from TATA and Inr [82], BPROM uses functional motifs and oligonucleotide information [84], PPP uses Hidden Markov model [94]. Out of these, only SAK used the signal from the whole promoter, whereas others are basically local content extraction methods. The efficacy of any method would be evident when they are applied to whole genome promoter prediction. Good recognition is not guaranteed even though the recognition on test data set is high. In this light, the proposed extension of the method to identify promoters in a whole genome, has produced satisfactory results. No true promoter is missed. It is to be seen whether the regions that were identified as promoters can be identified as the promoters or not, by setting an appropriate threshold and issues such as how to tackle the segment ends need to be sorted out in future.

Next chapter presents a scheme which tries to remove the imbalance in the data sets *Major Set (Maj)* and *Minor Set (Min)* and enhance the recognition rate through data balancing methods. A multi-level cascading system is proposed for complete recognition of the promoter. Later part of the chapter discusses the position weight matrix based features in promoter identification.

Chapter 4

Cascaded (Multi-Level) and PWM Based Classification of *E.coli* Promoters

In Section 3.3 of chapter 3, in-depth analysis of data sets using 2-grams is performed. The analysis presented a scenario where there is a confusion of the promoter signal with a minor set of non-promoter and vice versa. In an effort to build a complete classification system, using the majority and minority sets in promoters as well as non-promoters, a cascading multi-level system is proposed. Later part of the chapter is devoted to information content in positional weight matrices aligned with respect to the two binding regions and explores the role of local information content in promoter recognition problem.

A multi-level complete classification system to achieve much higher classification performance is proposed in this chapter. This chapter is built upon the method proposed in earlier chapter. An in-depth analysis of classification of *E.coli* promoters using 2-grams is done in Chapter 3. In section 3.3 of Chapter 3, analysis is done by removing the mis-classified sequences from the correctly classified set into *Major (Maj)* and *Minor (Min)* sets and further separate classifiers NN_{Maj} and NN_{Min} are built. Here, to balance the data that was found to be unbalanced, Synthetic Minority Over-sampling Technique (SMOTE) is employed. And also, as the system is complex, a committee machine based classifier, Adaboost classifier is used to enhance the performance.

In Chapter 3 a global signal given by the n-grams is used to determine the promoter. That can be termed as global signal extraction method since whole promoter is considered. In this chapter, a study of local versus the global signal is done by using the position weight matrices (PWM). PWMs are used to identify

Table 4.1: A multi-level classification

	Promoter	Non-promoter
Promoter	TP Output:1	FN Output:0
Non-promoter	FP Output:1	TN Output:0

the motifs present in the promoter and for promoter recognition. The facts that are experimentally determined are used as prior information. Later whole sequence is considered for promoter identification. In case of *Drosophila*, even though prior information is not available, attempt is made to compute position weight matrices and use them in promoter recognition.

4.1 Multi-level Classification System

To build a multi-level classification system unbalanced data method and a 2-level multi-layer feed-forward network systems are used.

4.1.1 Unbalanced data method

In order to build a multi-level (cascaded) classifier, first the basic feed-forward neural network classifier that is used in Chapter 3 is used with bi-grams as the input features. Here, after first-level classification, the correctly classified and misclassified sequences are separated as described in Chapter 3. Now, given an unknown sequence, the following scenarios can emerge: if it is identified as positive, then it can be a true positive (TP) or a false positive (FP) i.e., a non-promoter recognized as a promoter. Similarly, if it is identified as negative, then it can be a true negative, or a false negative(FN) i.e., a promoter identified as a non-promoter. This is first-level classification. Figure 4.1 depicts this.

Here to distinguish a TP from a FP, the second level of classification is proposed. In the second-level the training of the system is done by combining data along the columns in Table 4.1, i.e., TPs and FPs as one set, *Pos-Fpos* and TNs and FNs as one set, *Neg-Fneg*. Here we are re-arranging the data sets in such a way that true promoters recognized as promoters and non-promoters recognized

as promoters form one set *Pos-Fpos*. And the other set *Neg-Fneg* is obtained by grouping promoters misclassified as non-promoters and true non-promoters classified as non-promoters. The distance between $n - gram$ features of promoters and non-promoters in each of these sets is very small. Hence if the neural networks trained on these data sets train well, then recognition can be enhanced further. In order to achieve that goal, two neural networks NN_{P-FP} and NN_{N-FN} are designed using the two data sets *Pos-FPos* and *Neg-FNeg*. For NN_{P-FP} , TP is treated as positive data (*output* : 1) and FP is taken as negative data (*output* : 0). For the other network NN_{N-FN} , FNs are used as positive data and TNs are designated as negative data. If a good amount of learning can be done in these systems, then classification performance can be enhanced. This whole process is depicted in Figure 4.1.

The minority part of *Pos-Fpos* constitutes about half of the majority of *Pos-Fpos*, i.e., 2 : 1 and minority of *Neg-Fneg* is roughly 20% of the majority, i.e., 1 : 4.5. When a multi-layer perceptron is employed to train the system taking positive and negative data as explained above, the system is not able to train well at all. The data is unbalanced because the ratios are not equal. The training of the system is very poor. For example, for NN_{N-FN} , the positive recognition is as low as 3%. Hence, to enhance the recognition further methods of balancing the data ratio are required. In general, the imbalance can be addressed in two ways either by assigning distinct costs to the training examples [90, 29] or re-sampling the original data set by oversampling the minority class or under sampling the majority class [58, 50, 63, 66]. Another approach by Chawla et al. employs both under-sampling of the majority class with a special form of over-sampling of the minority class. This particular method is described as Synthetic Minority Over-sampling Technique (SMOTE) by them [81]. SMOTE has been utilized for balancing the data in this thesis.

Synthetic Minority Over-sampling Technique

An over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement was proposed by Chawla et al. [81]. Here, synthetic examples are generated by using "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the

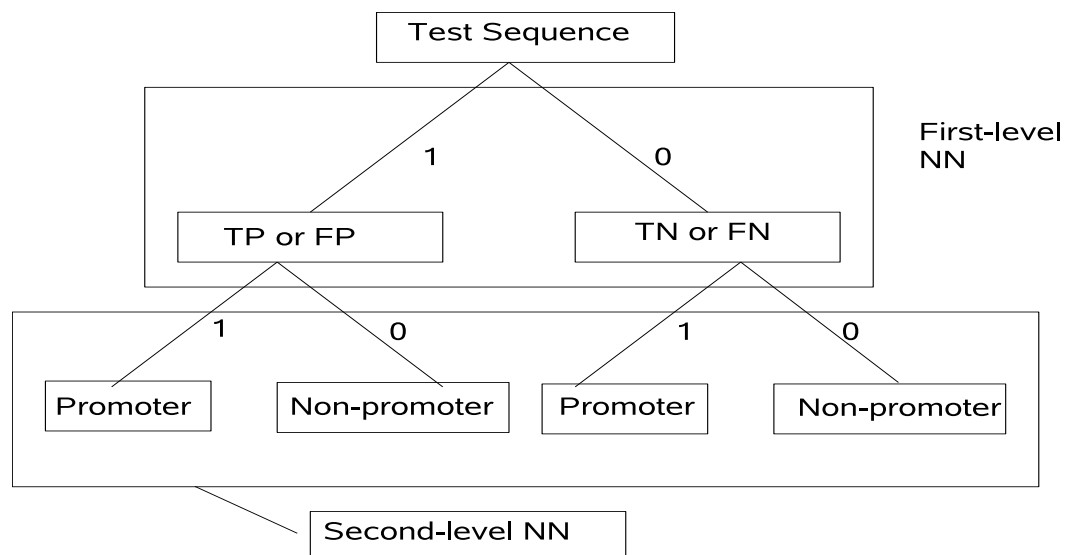


Figure 4.1: Possible output of a multi-level binary classification system.

line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. For instance, if the amount of over-sampling needed is 200%, only two neighbors from the k nearest neighbours are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. Algorithm SMOTE gives the pseudo-code for SMOTE [81].

Algorithm *SMOTE*(T, N, k)

```

Input: Number of minority class samples T ; Amount of SMOTE N ;
        Number of nearest neighbours k
Output: (N/100) * T synthetic minority class samples
1. (* If N is less than 100%, randomize the minority class
    samples as only a random percent of them will be SMOTEd. *)
2. if N < 100
3.     then Randomize the T minority class samples
        T = (N/100) * T
4.
5.     N = 100
6. endif
7. N = (int)(N/100) ( * The amount of SMOTE is assumed to be in
    integral multiples of 100. *)
8. k = Number of nearest neighbors
9. numattrs = Number of attributes
10. Sample[ ][ ]: array for original minority class samples
11. newindex: keeps a count of number of synthetic samples
    generated, initialized to 0
12. Synthetic[ ][ ]: array for synthetic samples
    (* Compute k nearest neighbors for each minority class
    sample only. *)

```

```

13. for i <- 1 to T
14.     Compute k nearest neighbors for i, and save the indices
        in the nnarray
15.     Populate(N , i, nnarray )
16. endfor

    Populate (N , i, nnarray )
    (* Function to generate the synthetic samples.*)
17. while N = 0
18.     Choose a random number between 1 and k, call it nn.
        This step chooses one of the k nearest neighbours of i.
        for attr <- 1 to numattrs
19.
            Compute: dif = Sample[nnarray[nn]][attr]-Sample[i][attr]
20.
21.     Compute: gap = random number between 0 and 1
            Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
22.
23.     endfor
24.     newindex++
        N = N -1
25.
26. endwhile
27. return (* End of Populate. *)

```

Training and Analysis

Using the SMOTE algorithm synthetic minority samples are created for both *Pos-Fpos* and *Neg-Fneg* sets. Using these synthetic samples which amplify the minority samples, two new feed-forward neural networks $NNSMOTE_{P-FP}$ and $NNSMOTE_{N-FN}$ are created. The results of classification of the two networks are combined and are shown in Table 4.2.

The results show that using SMOTE to build synthetic samples in fact hasn't improved the performance of the system any further. It has in fact deteriorated the performance i.e. even though the recognition of the positive data set is

Table 4.2: *E.coli* classification results using SMOTE

Test-data		Ratio of positive to negative	Precision	Specificity	Sensitivity
Test-data without smote	single set	1:2	81.57	88.01	65.38
Test-data using smote	single set	1:1	69.7	71.68	64.74
Test-data using smote	single set	1:2	70.09	83.9	35.25

unaffected, recognition of the negative data set has dropped considerably. One factor that looks likely to have affected the accuracies is the ratio of positive data to negative data. In our past experience, whenever equal proportion of positive and negative data are used, the accuracy of negative data has come down while accuracy of positive has gone up to roughly 70-75%. Here also it could be the same, since SMOTE is used to balance the unbalanced data to equal proportions. It is surprising, that the promoter recognition rate has not gone up. When SMOTE is used to generate data sets same as the original sets, i.e. the proportion of positive to negative to 1:2, there is a considerable drop in positive recognition rate.

4.1.2 Using a 2-level Multi-layer Feed-forward Network

Another complete classification system to recognize the promoters and non-promoters making use of cascaded neural networks along different lines is proposed in this section. A multi-layer feed-forward network with two output nodes is used. Using the same data set that is used in Section 4.1.1, the network is trained with training data. Using the configuration which gives the best results, a set of upper and lower limits are used here to classify the sequences into correctly classified or mis-classified and ambiguously classified. The outputs which fall into the category greater than the lower limit for negative data set and less than the upper limit for positive set of sequences are treated as ambiguously classified sequences. These difficult-to-learn sequences are culled and trained with another multi-layer feed-forward neural network. A different set of upper and lower limits can be set to identify the ambiguously classified sequences. Table 4.3 shows two sets of upper and lower limits.

Table 4.3: *E.coli* classification results using cascading system of networks

Test-data	Lower limit	Upper limit	Precision	Specificity	Sensitivity
single set	0.3	0.6	77	82.4	66
single set	0.35	0.6	76.64	85.96	59.6

Both sets have not improved the results any further than 2-grams. This again can be attributed to close resemblance between promoters and non-promoters.

4.2 Introduction to Committee Machines

A complex task is divided into sub tasks and the solutions of these task are combined in order to get solution of the complex task. In neural networks also similar method is employed to attain computational simplicity. Here the computational simplicity is achieved by dividing the complex task among a set of experts, which divide the input space into a set of subspaces. The set of experts are called a committee machine [80]. The combined efficiency of the committee machine is supposedly higher than individual experts. Committee machines can be basically classified into two major categories: static structures and dynamic structures [44]. In static structure category, the decisions of the experts is combined in such a way that they do not involve the input signal. Ensemble averaging and boosting are such kind of techniques. In dynamic structure category, the input is involved in arriving at the overall output from the outputs obtained from the individual experts. Mixture of experts and hierarchical mixture of experts fall under this class of committee machines.

In boosting, the training of experts is done on data sets with entirely different distributions. There are different kinds of boosting, one of them being Boosting by filtering. Here filtering of the training data set is done by using different versions of a weak learning algorithm. A drawback of this algorithm is that the data set has to be very large. This limitation is overcome by another boosting algorithm called AdaBoost [32]. AdaBoost allows the training data to be reused. A limitation on this is on the performance of the weak learning model on the distributions that are generated during the learning process.

AdaBoost classifier learns by giving a distribution D_n of the training data set

on iteration n to a weak learning algorithm F_n . In the next iteration $n + 1$, in the distribution D_{n+1} , the weight of the example that is classified by learning algorithm in the previous iteration is left unaltered, whereas the weight of the example that is misclassified is multiplied by a number in the interval $(0,1]$. These weights are re-normalized again by dividing by a renormalization constant in such a way that the examples that are harder to classify are given higher weights. This procedure is continued for T iterations. The final hypothesis is arrived at by taking weighted average of all the hypothesis F_1, F_2, \dots, F_T .

4.2.1 AdaBoost Classifier

GML AdaBoost Matlab Toolbox is a set of matlab functions and classes implementing a family of classification algorithms, known as Boosting [109]. This toolbox has 3 different boosting schemes: Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Real AdaBoost is the generalization of a basic AdaBoost algorithm first introduced by Freund and Schapire [99]. Real AdaBoost should be treated as a basic hardcore boosting algorithm. Gentle AdaBoost is a more robust and stable version of real AdaBoost algorithm [51]. Modest AdaBoost is a regularized trade off of AdaBoost, mostly aimed for better generalization capability and resistance to over fitting [109]. A tree learner is the available weak learner in this package. The number of maximum splits that would be done during the training can be defined.

Training and Analysis

GML AdaBoost classifier is used for the promoter data classification using 2-gram features that are used in the earlier experiments. Using 5-fold cross validation facility of the toolbox, the promoter data is classified. In another experiment, data created by SMOTE algorithm explained in Section 4.1.1 of this chapter is the input to the AdaBoost experiments. The results are presented in Table 4.4. The table presents the AdaBoost classification using bi-grams as well as SMOTE data used in Section 4.1.1.

The results of 5-fold cross-validation using AdaBoost classifier are poorer than that of multi-layer feed forward network using 2-grams. One fact could be that the AdaBoost is creating folds which have a kind of distribution which is different

Table 4.4: *E.coli* classification results using AdaBoost classifier

Features	Cross-validation	Precision	Specificity	Sensitivity
2-grams	5-fold	75.96	85.16	56.56
2-grams using SMOTE	single set	78.12	79.85	73.07

from the 5-folds created in earlier experiments using 2-grams.

The preceding results using multi-level classification systems and AdaBoost classifier can be summarized as follows. In order to enhance the classification performance, a multi-level classification scheme is proposed using 2-grams as features. This scheme uses data which is unbalanced by using SMOTE algorithm to address the imbalance by undertaking resampling of data. The results seem to indicate that the confusion in the data set is not reduced but further enhanced by this. In addition a committee machine, AdaBoost is also used to enhance the performance using 2-grams as features. The 5-fold cross-validation results of AdaBoost show that the results did not improve over the 2-grams using multi-layer feed-forward network. Now, we explore the local signal based approach in the next section.

4.3 Position Weight Matrices (PWM)

Previous chapter explores the feasibility of *n-grams* as features in promoter prediction. This approach gives good results in whole-genome promoter prediction as shown in Chapter 3 even though the test data results are not giving very good results. Now, we would like to verify further whether a local signal based on features extracted using position weight matrix is sufficient to predict a promoter as against a global signal extracted using position weight matrices. Efficacy of these global and local features extracted using PWMs is studied for *E.coli* species. This method is extended to *Drosophila* where local motif (binding region) data is not available.

Position weight matrices (PWM) assume that there exist patterns that are position dependent. Aligning these patterns, the frequency of occurrence of each component at a particular position in the pattern is used as the weight of that particular position. Hence, the method would be applicable on the condition

that there exist patterns in the data that is being looked at. If patterns are not conserved, local signal methods would not work.

In *E.coli* -10 binding region has a consensus sequence TATAAT. A search with only 2 mismatches in -10 binding region consensus sequence produces putative promoter approximately every 30 bp in complete genome sequence of *E.coli* [47]. Huerta et al. have considered a set of 288 weight matrices based on the frequency of bases in input sequences, noncoding sequences and the coding regions and chosen a set of best weight matrices which gave good sensitivity measure. They have concluded that even though the model they considered has given good predictability, that it is not a best statistical model. More often they found that the actual promoter has less probability score than the actual promoter. The occurrence of a number of putative promoters could be a pointer to the presence of an actual promoter.

Ma et al. have used Expectation maximization algorithm to identify the binding regions and then the nucleotides around binding regions are extracted and used as features for a neural network classifier [70]. But, they have tried it on synthetic data. Other groups who have used PWM for promoter recognition are Staden et al., Harley and Reynolds, Mulligan and Mclure [105, 41, 79].

4.4 PWM from Harley-Reynold's Data for *E.coli*

4.4.1 Mono-gram PWM of -35 and -10 binding sites from Harley's data

A local signal can be extracted only by identifying the binding regions precisely. To extract the local signal, Harley et al. experimental data is used. Harley et al. have identified -35 and -10 regions in their data set through reiterative alignment of promoter regions in order to select -35 and -10 regions most consistent with the reference list of promoters with known TSS. They have analyzed a set of 263 promoters to arrive at the promoter structure. The spacing between -35 and -10 seems to vary between 15-21 bases and the spacing between -10 and TSS varying between 4-12 bases. Using these sets of binding regions, the base distribution in -35 and -10 is computed. $p(j)(i)$ gives the frequency of occurrence of base at position i for a particular binding region j , which can be treated as probability of

occurrence of a base at a particular position. These probabilities are presented in Table 4.5 and 4.6 [41]. And also the base distribution at TSS is estimated. In case of sequences which have more number of TSS or where TSS is not determined exactly, the first occurrence is taken as the TSS.

Table 4.5: *E.coli* base distribution for -35 binding site (Harley et al. [41])

Position Base	1	2	3	4	5	6
T	0.78	0.82	0.15	0.2	0.1	0.24
G	0.1	0.05	0.68	0.1	0.07	0.17
C	0.09	0.03	0.14	0.13	0.52	0.05
A	0.03	0.1	0.03	0.58	0.32	0.54
Most conserved	T	T	G	A	C	A

Table 4.6: *E.coli* base distribution for -10 binding site (Harley et al.[41])

Position Base	1	2	3	4	5	6
T	0.82	0.07	0.52	0.14	0.19	0.89
G	0.07	0.01	0.12	0.15	0.11	0.02
C	0.08	0.03	0.1	0.12	0.21	0.05
A	0.03	0.89	0.26	0.59	0.49	0.03
Most conserved	T	A	T	A	A	T

4.4.2 Bi-gram and tri-gram PWMs of -35 and -10 binding sites from Harley's data

Using the binding sites identified in Harley's data, we have computed bi-gram distributions for -35 and -10 binding regions. These distributions are shown in Tables 4.7 and 4.8 respectively. The bottom line shows the most conserved bases and bi-grams at each position respectively in these tables.

To predict whether a given sequence is a promoter or a non-promoter, single nucleotide, dinucleotide and trinucleotide distributions are used from the same data sets and these are used in the likelihood computation. P_1 gives the likelihood value of the promoter using single nucleotide base distribution in the two binding

Table 4.7: *E.coli* bi-gram distribution for -35 binding site

Position Bigram	12	23	34	45	56
AA	0.0037	0.0037	0.0221	0.1507	0.1985
AT	0.0294	0.0074	0.0000	0.0588	0.0478
AG	0.0074	0.0772	0.0037	0.0368	0.0551
AC	0.0074	0.0221	0.0037	0.3125	0.0147
TA	0.0625	0.0221	0.0956	0.0625	0.0404
TT	0.6324	0.1324	0.0331	0.0221	0.0368
TG	0.0441	0.5184	0.0147	0.0184	0.0184
TC	0.0221	0.1176	0.0037	0.0809	0.0110
GA	0.0368	0.0037	0.4081	0.0551	0.0221
GT	0.0625	0.0037	0.1029	0.0037	0.0257
GG	0.0000	0.0441	0.0662	0.0000	0.0221
GC	0.0074	0.0074	0.0993	0.0625	0.0074
CA	0.0074	0.0000	0.0331	0.0478	0.2868
CT	0.0662	0.0037	0.0478	0.0221	0.1176
CG	0.0074	0.0368	0.0368	0.0221	0.0882
CC	0.0037	0.0000	0.0294	0.0441	0.0074
Most conserved	TT	TG	GA	AC	CA

sites and TSS. Here, $j = 35$ and 10 denotes the particular binding site and i denotes the position of the base in the hexamer located at the binding site. P_2 gives the probability in -35 binding site using bi-gram distributions, P_3 probability of bi-gram distribution at -10 binding site. Similarly P_4 gives the probability using tri-gram distribution at -35 binding site and P_5 gives probability using tri-gram distribution at -10 site.

Table 4.8: *E.coli* bi-gram distribution for -10 binding site

Position Bigram	12	23	34	45	56
AA	0.026	0.254	0.173	0.298	0.018
AT	0.000	0.449	0.044	0.085	0.419
AG	0.000	0.110	0.033	0.040	0.015
AC	0.000	0.088	0.018	0.162	0.033
TA	0.713	0.015	0.294	0.055	0.004
TT	0.055	0.040	0.051	0.029	0.162
TG	0.007	0.000	0.096	0.026	0.004
TC	0.029	0.007	0.085	0.022	0.022
GA	0.081	0.000	0.055	0.074	0.000
GT	0.000	0.007	0.026	0.055	0.099
GG	0.000	0.000	0.018	0.015	0.004
GA	0.000	0.000	0.011	0.022	0.000
CA	0.081	0.000	0.063	0.059	0.000
CT	0.007	0.029	0.011	0.022	0.221
CG	0.000	0.000	0.018	0.022	0.000
CC	0.000	0.000	0.004	0.015	0.000
Most conserved	TA	AT	TA	AA	AT

$$P_1 = \prod_{j=35, i=1}^6 p(j)(i) \prod_{j=10, i=1}^6 p(j)(i) \prod_{j=TSS} p(j)(i) \quad (4.1)$$

$$P_2 = \prod_{j=35, i=1, k=i+1}^5 p(j)(ik) \quad (4.2)$$

$$P_3 = \prod_{j=10, i=1, k=i+1}^5 p(j)(ik) \quad (4.3)$$

$$P_4 = \prod_{j=35, i=1, k=i+1, l=i+2}^4 p(j)(ikl) \quad (4.4)$$

$$P_5 = \prod_{j=10, i=1, k=i+1, l=i+2}^4 p(j)(ikl) \quad (4.5)$$

4.4.3 PWMs of Binding Sites in Promoter Recognition of *E.coli*

In computing various P_i s, since the probabilities could be low, the multiplication could lead to very low value. It is customary in such cases to compute logarithm of the probabilities which would make it to be additive. The proportion of positive data set to the negative data set is taken as 1:2. The position weight matrices are derived from Harley's data. Using the PWMs the features P_i s are extracted from training and test data sets and used as features. PWMs are used to compute the log-probabilities for a range of values of spacing regions found out in literature. Spacing between -35 and -10 region found to vary between 15-21 bases and spacing between -10 and TSS is found to vary between 4-11 bases. There are 63 combinations that are possible taking into account the spacing between -35 to 10 varies by 7 bases and spacing between -10 and TSS which varies by 9 bases. For each of the combinations log probabilities are calculated. Out of which, the ones with maximum log probability value is chosen as the most probable annotation of the promoter. This would fix the spacing between -35 and -10, and also between -10 and TSS. Various experiments are conducted using these probability values as features to a multi-layer feed-forward neural network.

To illustrate the feature extraction method, consider a subsequence TGGTCA at -35 site, TGTAAT at -10 site and AATTCA at TSS.

P_1 for this will be $p(T) \times p(G) \times p(G) \times p(T) \times p(C) \times p(A)$
 $\times p(T) \times p(G) \times p(T) \times p(A) \times p(A) \times p(T) \times p(A) \times p(A) \times p(T) \times p(T) \times p(C) \times p(A)$.
 That is for -35 site, we consult Table 4.5 and whichever nucleotide is present at that position will determine the probability value as $p(.)$. Here, T is occurring at the first position of the hexamer, hence $p(T)$ will be 0.78. Similarly, G is occurring at the second position, so $p(G)$ is 0.05. G is occurring at third position, $p(G)$ is 0.68 according to the table. In similar fashion probabilities for all the nucleotides for all the subsequences can be determined from the tables. In the same way,

P_2 can be computed as $p(TG) \times p(GG) \times p(GT) \times p(TC) \times p(CA)$,

P_3 as $p(TG) \times p(GT) \times p(TA) \times p(AA) \times p(AT)$,

P_4 as $p(TGG) \times p(GGT) \times p(GTC) \times p(TCA)$ and

P_5 as $p(TGT) \times p(GTA) \times p(TAA) \times p(AAT)$.

Initially, maximum log likelihood of the probabilities of single bases or mono-

grams for both binding regions, dinucleotides for -35 and -10 regions P_1 , P_2 , P_3 are computed for all the 63 combinations mentioned above. P_1 , P_2 and P_3 are added for all 63 combinations, maximum likelihood value is used to determine the combination of spacers between -35 and -10, -10 and TSS. Once the binding sites are fixed, we can find the binding sites themselves. When these binding sites are compared with those of Harley et al. a few of the binding regions determined using the above schema seems to be different from original binding sites. Even though, statistically maximum likelihood criterion is violated, still the binding regions are found to conform to some other configuration which has a less log likelihood. Using these probabilities as features, the data set used in Chapter 3 and section 3.3 is classified. The results are surprisingly close to what we obtained for bi-gram features in Chapter 3, Section 3.3. To improve the results further the trinucleotide probabilities are also computed. Again, using the log likelihood of the probabilities of single base, dinucleotide for -35 and -10 regions, trinucleotide for -35 and -10 regions P_1 , P_2 , P_3 , P_4 , P_5 as the input features to the neural network, the data set used in Chapter 3 and section 3.3 is classified. The classification results are presented in Table 4.9. Similarly the bi-gram frequency and trinucleotide frequency in the non-binding regions i.e. regions before -35 binding site, in between the binding sites and after the -10 binding site are also computed using the same data. In case of non-binding regions, no position specific weights are computed. Just the average bi-gram and trigram values are used as it is difficult to get position specific scores in these regions. Using these, the log-likelihood of these regions is computed and used as additional features to the features obtained from binding regions. Using these, the log-likelihood features of these regions is computed and used as additional features to the features obtained from binding regions. The improvement is marginal with the addition of features from non-binding regions. The results are tabulated in the Table 4.9.

Table 4.9 shows clearly better results with 5 features compared to those obtained with using bi-grams alone. 5-fold cross validation is performed for the set with 11 features as it is the best of all, giving a precision of 83.99, specificity of 89.56 and sensitivity of 72.19. These results are much better compared to 5-fold cross-validation results using bi-grams obtained in Chapter 3. One other important feature here is the identification of the TSS irrespective of sigma factor. This has been tested on RegulonDB 5.7 version. It is found that the TSS identification is 100%. -10 site is also being identified correctly in all the cases.

Table 4.9: *E.coli* classification results using log probability obtained by using PWM for binding sites only.

Features	Number of features	Precision	Specificity	Sensitivity
Bi-grams only	16	81.57	88.01	65.38
Only from binding regions- P_1, P_2, P_3	3	85.95	89.29	77.56
Only from binding regions- P_1, P_2, P_3, P_4, P_5	5	87.23	90.56	80.13
From binding regions- P_1, P_2, P_3, P_4, P_5 and 3 non-binding regions bi-gram, tri-gram probabilities	11	88.50	91.33	81.41

The drawback of using this local signal extraction scheme is that one should have extensive and minute knowledge of the system. One needs to know the number of binding sites the spacing variations, the binding sites themselves. In case of new species this information would not be readily available. The localized signal will be considered as efficient if we get good genome-wide promoter prediction.

4.4.4 Genome-Wide Promoter Recognition

5-fold cross-validation results using features extracted by PWMs from local motifs are encouraging. In order to find the efficacy of the features in identifying the promoters in a genome sequence, same data set used in section 3.4.2 is considered. Using position weight matrix features referred to in the Table 4.9, promoter recognition is done in a genome using two separate neural networks as described in section 3.4.2. One uses promoter, coding segments as positive and negative data sets respectively and another network uses promoter and non-coding segments as positive and negative data sets. The features are computed as explained earlier. Testing these networks with *section1* of *E.coli* [20], a spread of positive signals over the entire range is obtained as shown in Figure 4.2. These results do not show clear cluster behaviour unlike what was seen in Section 3.4.2 using tri-grams. Hence it has been difficult to identify a promoter using PWM features.

This matter will be further explored using whole promoter based PWM features.

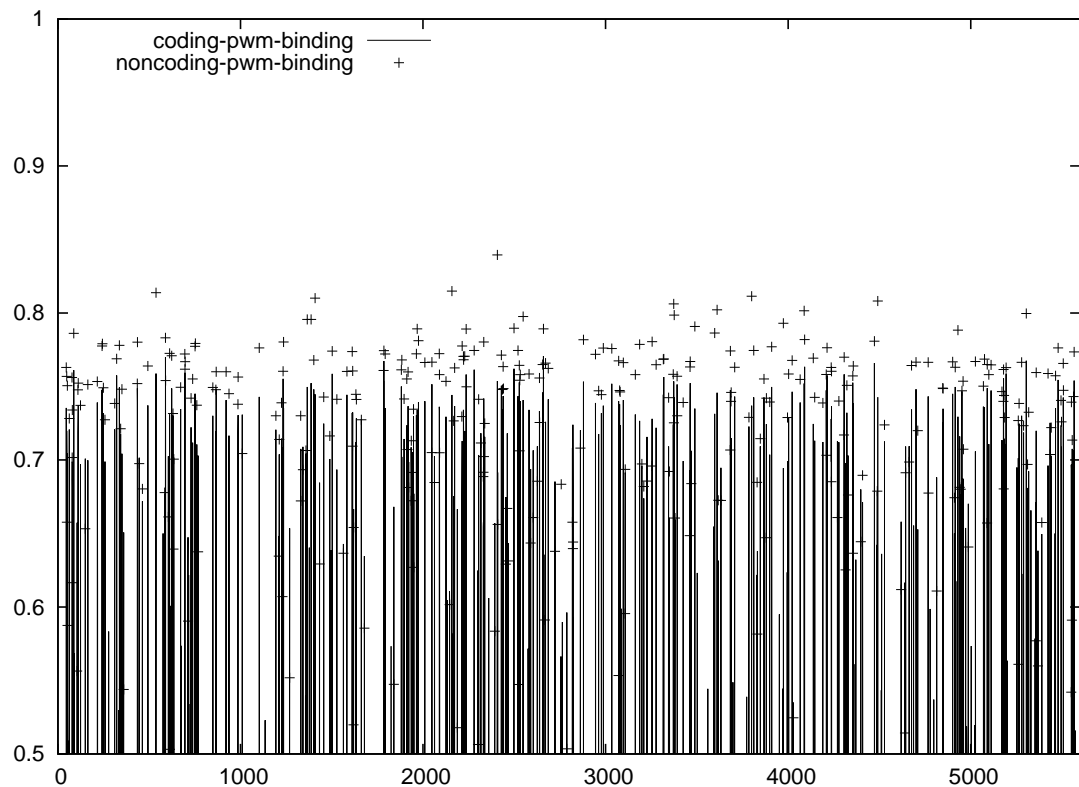


Figure 4.2: Outputs of the networks NN_{PC} and NN_{PN} by using features determined by PWM (for binding regions only) versus the moving window for *section1* of *E.coli* genome.

4.4.5 PWM of Whole Promoter in Promoter Classification

We have conducted two experiments using PWM of whole promoter. First one considers a set of promoters aligned with respect to TSS alone. From this, few regions of interest are identified. The other experiment is done by not identifying any regions at all and considers the whole promoter for PWM based feature extraction.

Aligning sequences of length 80 bp with respect to TSS alone in training data set, roughly three regions of interest can be identified which show large distance between positive and negative sets in these regions. One around the region -40 to -32 and one more around -20 to -7 and another around -5 to +2. These regions

roughly correspond to the two binding regions and the TSS. Now, estimating the log probabilities not only by using position specific weight matrices for promoter set only, but also by using position specific weight matrices for non-promoter set and using them to say whether a given sequence is a promoter or a non-promoter depending upon whether the log probability for promoter is greater than the log probability for non-promoter vice versa. Instead of comparing the values individually, mono-grams for all regions, bi-grams for all regions, tri-grams for all regions, 4-grams for all regions are added together individually for both promoter and non-promoter sets. These values are compared to decide whether a given sequence is a promoter or a non-promoter. This method is giving the same results as that of bi-grams given in Chapter 3. Instead of considering separate regions, using the whole sequence and position weight matrices to compute the log probability, the method gives a better performance, i.e. giving sensitivity rate of 76% and specificity of 86%. Using this to do whole genome promoter recognition is giving a clustered structure as explained in Section 3.4. Still the spurious promoters are spread over the genome segment, but they are fewer compared to what was obtained with features extracted from binding sites only. Surprisingly, the tri-grams, 4-grams log probabilities for promoters in test data are not consistent with what is expected from the training data. Even though the position specific weight matrices of positive data set are computed from the training data, the log probabilities of test data computed using the PWMs of promoter data are lower than that were obtained by using position specific weight matrices for non-promoter data. That is, there is no bias of log-probabilities towards promoter even with PWMs of promoter. For example trigram log probability and 4-gram probability using promoter PWMs are -513.562439 and -3507.333496 respectively for a test sequence. Whereas trigram log probability and 4-gram probability using coding PWMs are -525.244873, -3335.950439 respectively for the same test sequence.

To compare and contrast the results due to whole sequence against the local motif method, a preliminary study is done. Instead of few particular regions of interest the whole sequence is considered for computation. Position weight matrices for mono-grams, bi-grams, tri-grams and 4-grams at each position i ($i = 1$ to 80 for *E.coli*) in the sequence are computed using the training data set. The PWMs for mono-grams, bi-grams, tri-grams and 4-grams are calculated for each group of promoter sequences, coding sequences and noncoding sequences from the training

data set separately. Using the position weight matrices for each group, the log probabilities of mono-grams, bi-grams, tri-grams and 4-grams for test data are estimated. Again, using a cascading scheme, first the log-probabilities obtained by using promoter and non-coding position weight matrices are compared to decide whether a test data sequence is a promoter or a non-coding sequence. The decision is based upon whichever log-probabilities computed by using position weight matrices from promoter and non-coding sequences are higher. Again the outcomes that are positive are tested using position weight matrices for promoter and coding sequences. The test data results gave sensitivity of 71.8% and specificity of 88.8%. The deciding features are the bi-grams and tri-grams. Using the same technique, *section1* of *E.coli* using the moving window scheme described earlier, is tested. This gave results again in clusters, unlike the results obtained by using features obtained from position weight matrices for binding regions alone. Figure 4.3 illustrates the results obtained by using position weight matrices and by tri-grams. The results portray a cluster structure, though with a large number of spikes.

Table 4.10: *E.coli* classification results using log probability values obtained by using PWM for whole promoter.

Features	Number of features	Precision	Specificity	Sensitivity
Bi-grams only	16	81.57	88.01	65.38
From binding regions- P_1 , P_2 , P_3 , P_4 , P_5 and 3 non-binding regions-bi-gram, trigram probabilities	11	88.50	91.33	81.41
From whole promoter (-40 to -32, -20 to -7 and -5 to +2)	mono-gram, bi-gram, tri-gram, 4-gram (4)	82.84	85.7	75.64
From whole promoter (-60 to +19)	mono-gram, bi-gram, tri-gram, 4-gram (4)	83.94	88.8	71.8

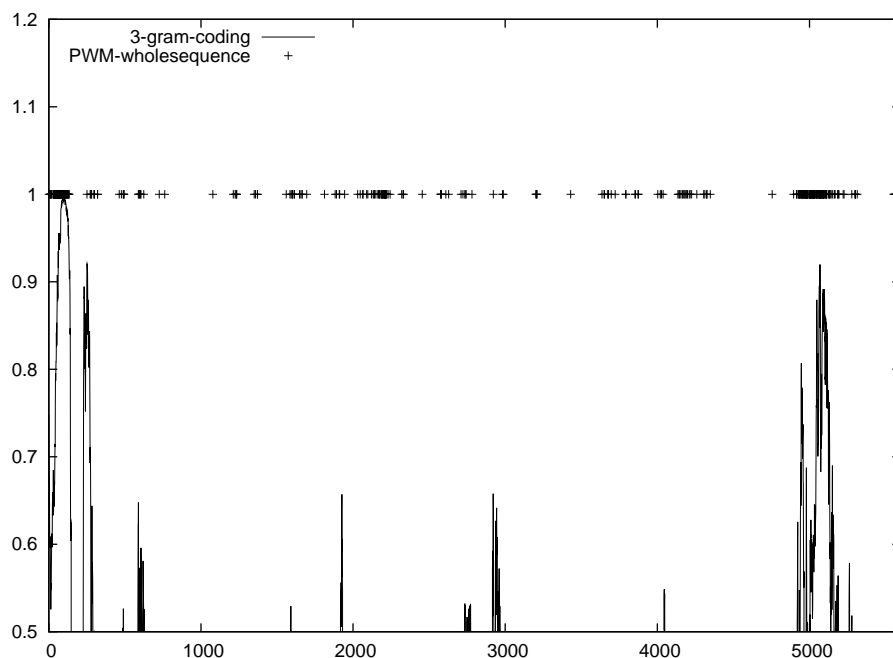


Figure 4.3: Output of the network NN_{PC} using 3-grams and the output generated by using PWM for whole promoter versus the moving window for *section1* of *E.coli* genome.

4.5 PWMs for Drosophila Using Whole Promoter

Now to extend the ideas further to *Drosophila*, where annotated sequences similar to *E.coli* are not available and where only the TSS is a known factor. The data set of Ohler et al. [86] is used where each sequence is of length 300 bp, -250bp upstream and 49 bp downstream and 251 location is TSS in promoters. All the sequences are aligned with respect to TSS. Similar to *E.coli* data, the data set [86] is divided into training and test data sets here. The ratio of positive to negative data sets in both training and test sets is 1:2. Whole sequence data without specifying important and non-important regions is used. With respect to TSS the mono-gram, bi-gram, tri-gram and 4-grams are computed for every position in the sequence. Position weight matrices mono-grams, bi-grams, tri-grams and 4-grams are calculated for promoters, coding sequences and intron sequences separately using the training data set for all positions in the sequence. Using these position weight matrices, log probabilities of mono-grams, bi-grams,

tri-grams and 4-grams for the whole sequences in test data are calculated as $\prod_{i=1}^{300} p(i)$, $\prod_{i=1, k=i+1}^{300} p(ik)$, $\prod_{i=1, k=i+1, l=i+2}^{300} p(ikl)$, $\prod_{i=1, k=i+1, l=i+2, m=i+3}^{300} p(iklm)$ respectively. $p(i)$, $p(ik)$, $p(ikl)$, $p(iklm)$ are obtained from the PWMs computed from training data set. Log probabilities of the test data using PWMs of all three sets (promoter, cds and intron) are calculated. These log probabilities are used to determine whether a sequence is a promoter or a non-promoter in a cascading manner. Here, a neural network is not used for classification instead probability scores are used. Classification of test data is a two step process here. Classification is done by first looking at the log probabilities of mono-grams, bi-grams, tri-grams and 4-grams obtained by using PWMs of promoter and coding data sets. Whichever value is higher, determines the category to which the test sequence belongs to. For example, log probability value of bi-grams and tri-grams obtained by using PWMs of promoter are greater than log probability value of bi-grams and tri-grams obtained by using PWMs of coding, then the sequence is decided as belonging to promoter otherwise a non-promoter. Similarly, in the second step, all those that are positive in the first step are categorized into promoter and intron classes again based on the log probabilities computed by using PWMs for promoter and introns. Whichever value is higher, that determines the class of the test sequence. This scheme has good prediction accuracies. This gave sensitivity of 81.4% and specificity of 90.7% for the test data.

To test the method further to identify promoters in whole genome, the moving window scheme explained in Section 3.4 is employed to segment a portion of the genome sequence into sequence segments of length 300 bp. Using these set of sequences as test data, the log probabilities are estimated and used as explained earlier. The qualitative analysis of the results obtained indicates that the positives are clustered, but lot of spikes are also present camouflaging the signal.

4.6 Discussion and summary

It has been observed that there is data imbalance in data sets *Pos-Fpos* and *Neg-Fneg* to the extent of 50% and 80% respectively. To address this issue and to improve the promoter recognition rate we used data balancing scheme, namely, SMOTE. Although this improves the minority recognition in either set, it is also reducing the majority set recognition hence off-setting any improvement of total

promoter recognition. One more multi-level classification system, using ambiguously classified sequences is experimented with, which doesn't improve the results any further. Comparing the results in 4.4 and 4.2, the results using AdaBoost classifier showed an improvement in case of data enhanced by SMOTE method than using a multi-layer feed forward network in the sense that true positive recognition has improved, but at the same time, recognition of true negatives has come down. We explored the feasibility of data balancing and boosting methods to enhance the recognition rate. While the results seem encouraging, further work is required to make them perform better than the 2-gram classification scheme proposed in Chapter 3.

Comparing the results that were obtained using signal extracted from motifs/binding regions in the promoter and the signal obtained from the whole sequence, it is evident that the local signal has a handicap in applying to a whole genome promoter prediction. The signal from entire sequence without segmenting it into important and non-important portions led to better generalization. The binding regions are important in assisting the RNA polymerase to bind the promoter, but that alone is not sufficient to recognize a promoter. The so called non-important portions also have a bearing on the promoter recognition when PWMs are used.

In recent literature, Zhong li et al. have considered a set of 683 experimentally verified *E.coli* σ^{70} -promoter sequences [64]. They have obtained for test data the sensitivity and specificity of 91% and 81% respectively for negative data consisting of coding regions alone and 90% and 77% respectively for negative data taken from intergenic portions only. Getting sensitivity of 91% is good in the sense that as higher n-grams are used, it becomes more and more specific. By, adding some pseudo counts, they made use of hexamers. Here also, there is no alignment with respect to any pattern, except at the TSS. Their sensitivity is 100% but, specificity is 30% for whole genome. They claim that specificity is low for the reason that all false positives are not yet determined promoters. In our case, for test data the sensitivity and specificity are, 77.56% and 85.2%, respectively for negative data consisting of coding regions alone. And they are 76.3% and 91.3% respectively for negative data taken from intergenic portions. Our global features are better suited to classification of non-promoter data. Next chapter explores another global signal extraction method using wavelets to model the natural mechanism of binding of RNA polymerase to a promoter.

Chapter 5

Investigation of Signal Processing Methods for Promoter Recognition

This chapter explores the application of signal processing techniques such as Fast Fourier Transform (FFT) and wavelet transforms for promoter recognition as well as the possibility of modeling the RNA polymerase-promoter interaction. Various encoding techniques such as electron-ion interaction potential (EIIP), enthalpy, roll angle etc. are also included in the encoding of a sequence into a numerical format.

Traditionally biomedical signals have been analyzed by signal processing techniques such as Fourier transform and wavelet transforms. Biological data sets consist mostly of sequences made up of either nucleotides or amino acids. Hence an encoding system is required to convert these sequences into numerical series. Once a numerical series is obtained, Fourier transform (FT) or wavelet transform (WT) can be applied. Wavelets have been used in the literature to analyze biological signals such as genome sequences, protein structures and gene expression data [92]. Signal analysis has been claimed as one technique which is invariant to letter dissimilarity in a set of target data consisting of DNA sequences [21, 54]. Since the most prevalent search strategy is homology search (sequence similarity search), if there is less sequence similarity, then the homology search strategies do not help. Hence, in that light, signal processing techniques are claimed to be invariant to letter dissimilarity since, they convert original sequence of letters into a numerical sequence. If the numerical values of different nucleotides have closer numerical values, then the sequence would become invariant of letters. It is pointed out earlier that promoters have functionally similar structures, but

have no sequence similarity. This is the motivation behind considering the signal processing techniques for promoter recognition.

Shrish et al. have shown that the FT of gene portion, often has a prominent peak at $\frac{1}{3}$ position confirming the periodicity of codons [102]. But, non-genes do not have any such peak. They have decomposed the original sequence into a set of four indicator sequences for the four nucleotides A, T, G, C [110]. Each sequence is a binary sequence indicating the presence or absence of a particular nucleotide. Deyneko et al. have applied the physical features such as melting enthalpy, roll angle and minor groove depth of DNA to find similar promoters which correlate with their transcription regulatory responsiveness to different antibiotic and osmotic treatments [28]. They transformed the *E.coli* promoters into numerical sequences using physical parameters such as enthalpy, roll angle etc. FT of the the transformed sequences is used in computing cross-correlation and auto- correlation between different promoters. In particular, they looked for genes responsible for SOS response.

In this thesis, promoter recognition is done in two ways. One is by extracting global signal features using FT and WT of the promoter and non-promoter sequences. The other way is to obtain features from promoter and RNA polymerase interaction. Here, the promoter recognition is posed as a binary classification problem. Fourier transform has been used by quite a few groups so far, but there is no attempt to use the wavelets by anybody. Here we not only do classification, but also attempt to understand how RNA polymerase would bind to the promoter. It is assumed that the signal that is responsible for the binding to happen is retained by the promoter irrespective of the place of occurrence of the promoter in the genome, i.e. whether in an inter-genic portion or in a coding region [2]. We hope to extract the interaction between RNA polymerase and a promoter through the cross-correlation between the decomposed wavelets. To start with, FT of the sequences is used to analyze the promoter region to gain the knowledge in the frequency domain. Fourier transform *per se* is not used in promoter recognition. Power spectrum computed using the Fourier coefficients are used as features in promoter identification. Since, in FT positional information is lost, WT is being used to retain that information.

5.1 Introduction to wavelets

Wavelet transformation technique facilitates the analysis of a signal from a global level to a local level similar to zooming the view from forest to individual trees. Wavelets can be used to obtain the information about the localization of a signal in both time and frequency (scale=1/frequency). The ability to vary the scale of the function as it addresses different frequencies also makes wavelets better suited to signals with spikes or discontinuities than traditional transformations such as the FT. Application of wavelet transform to a signal decomposes the signal into several groups of coefficients. Different coefficient vectors contain information about characteristics of the signal at different scales. If a wavelet with a particular window size matches with the signal in that particular window size then the coefficient will be maximum. Coefficients at coarse scale capture gross and global features of the signal while coefficients at fine scales contain local details. Discrete wavelet analysis (DWT) is more appropriate for samples sampled discretely. A DWT denoted by W applied to a vector of observations and decomposes the data into sets of wavelet coefficients as $d = WX$ [72].

$$d = [d_1^T, d_2^T, d_3^T, \dots, d_J^T, c_J^T]^T \quad (5.1)$$

with

$$d_j = W_j X, c_J = V_J X \quad (5.2)$$

Where J is the largest level of the transform.

$$W = [W_1, W_2, \dots, W_J, V_J]^T \quad (5.3)$$

is an orthogonal matrix. W_1, W_2 etc are the dilated and translated versions of a basis function called mother wavelet [26, 23, 72] at different levels. For increasing values of j , the coefficients describe features at lower frequency ranges and larger time periods. Here, this process can also be described as generating low-pass and high-pass filters of a particular wavelet, and convolving the filter with X to generate the approximate and detail waves. Figure 5.1 depicts this process. Here the original signal is convolved with two filters to get two signal of the same length as the original signal. But, in this process we end up with more number of points than we started with. So, a process called *downsampling*

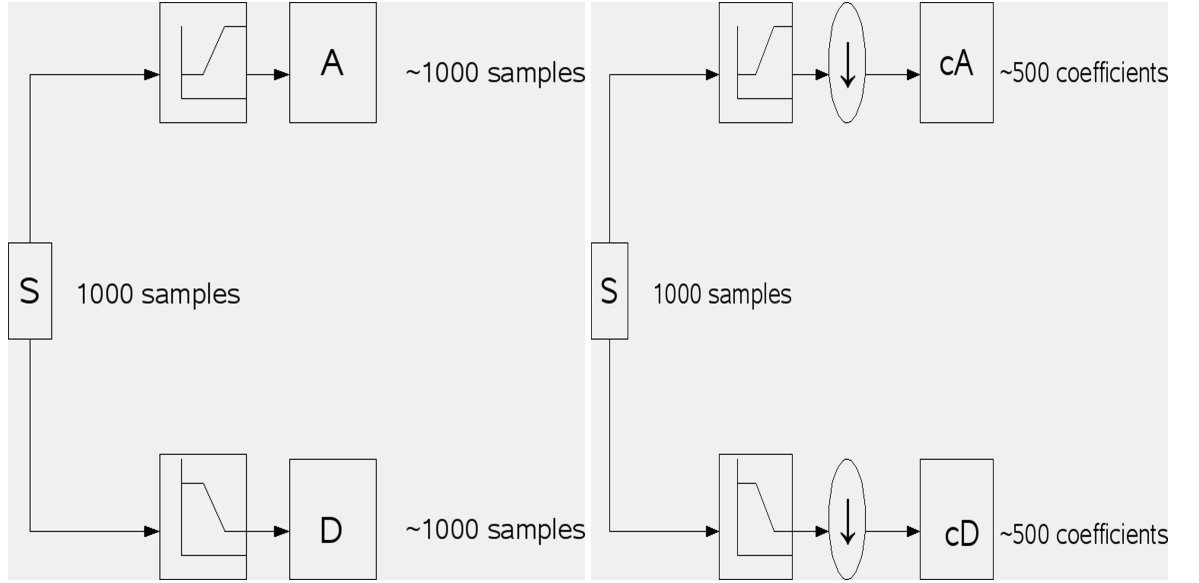


Figure 5.1: Decomposition and downsampling of signal S using wavelets.

is used to reduce the number of samples. In *downsampling* only one point in every two points is retained as illustrated in Figure 5.1. cA and cD are the approximate and detail coefficients. The decomposition process can be iterated, so that the low frequency component is decomposed into low and high frequency components successively as in Figure 5.2. This is called multi-level decomposition. The process can be reversed to obtain the original signal from the decomposed signals by applying *upsampling* and convolution with the filters. Hence, X can be written as an additive decomposition.

$$X = W^T d = \sum_{j=1}^J W_j^T d_j + V_J^T c_J = \sum_{j=1}^J D_j + C_J \quad (5.4)$$

with D_j the detail of the signal describing changes at the level j and C_j the smooth component associated with variations at level $J+1$ and higher [92].

A global signal using wavelet transforms is extracted from both promoter as well as non-promoter sequences and used as input to a classifier. Basically there are two operations that can be performed on a signal, decomposition and reconstruction. One set of experiments are done using wavelet coefficients at various scales as features for a feed-forward neural network to classify the promoter sequence. Another set of experiments are done using the decomposed waves as features to the classifier.

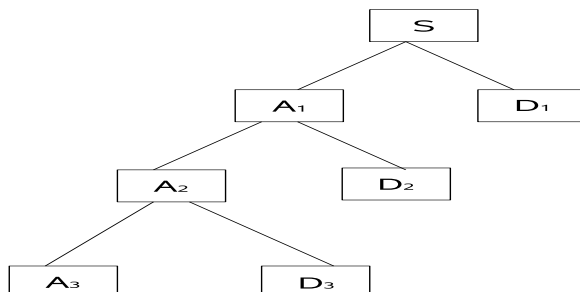


Figure 5.2: Decomposition of signal S using wavelets.

5.2 Methods

5.2.1 Encoding and Decomposition

A DNA sequence is made up of four nucleotides A, T, G and C. Different kinds of encodings have been used by different groups [83, 4, 25]. Nobuyuki et al. have used the values $A=1$, $T=-1$, $G=1$ and $C=-1$ whereas Cosic et al. have encoded the nucleotides by using the electron-ion interaction potential (EIIP) values: A: 0.1260, G: 0.0806, T: 0.1335, C: 0.1340 [25].

There are another set of encodings that are based upon dinucleotides. A large number of physico-chemical parameters of DNA double strand reflecting its specific properties have been collected in a public database [93]. Three parameter sets, melting enthalpy, minor groove depth, roll are given in Table 5.1. DNA enthalpy data describes the melting of DNA double strands. The enthalpy data are dependent on the neighbouring nucleotide and direction $5' \rightarrow 3'$ is of importance here. This is due to the fact that enthalpy is not only attributable to the direction invariant hydrogen-bonds but also to the interactions between electrons of neighbouring base pairs. Van der Waals forces also contribute to the inter-

actions between the immediate base neighbors [107]. This information is not reversible for the strand direction and must therefore be taken into account in the enthalpy-based conversion of the primary structure into a signal [22]. Roll angle is another structural feature that may help in promoter recognition. A dinucleotide step is helically twisted since the distance between sugar-phosphate, backbone is twice the distance between base-stacking distance [17]. If a step is untwisted the basepairs are pushed apart and the rise distance increases. To regain the stacking, i.e. to decrease the rise distance the step then rolls around the major groove [37]. For RNA polymerase to bind to the promoter, an open complex near -10 site is required. Hence this particular structural feature may be important in analyzing the dynamics of DNA segment. The parameters are used to represent the DNA by Kauer et al. [54, 28]. Deyneko et al. contend that the mere symbol computations are misleading since AA in stead of GA numerically is much more significant in terms of melting enthalpy. They claim that by using the physico-chemical parameters, they were able to find much more significant comparison of promoters than with nucleotide comparison.

Table 5.1: Physico-chemical properties of DNA [28].

Dinucleotide	Melting enthalpy (kcal/mol)	Minor groove depth (Å)	Roll angle (degree)
AA	9.05	9.03	0.3
AT	8.60	8.91	-0.8
AG	7.84	8.98	4.5
AC	6.54	8.79	0.5
TA	6.00	9.00	2.8
TT	9.14	9.03	0.3
TG	5.84	9.09	0.5
TC	5.64	9.11	-1.3
GA	5.55	9.11	-1.3
GT	6.45	8.79	0.5
GG	10.95	8.99	6.0
GC	11.10	8.98	-6.2
CA	5.75	9.09	0.5
CT	7.75	8.98	4.5
CG	11.90	9.06	-6.2
CC	11.04	8.99	6.0

Here, we followed the EIIP encoding system of Cosic et al. for promoter-

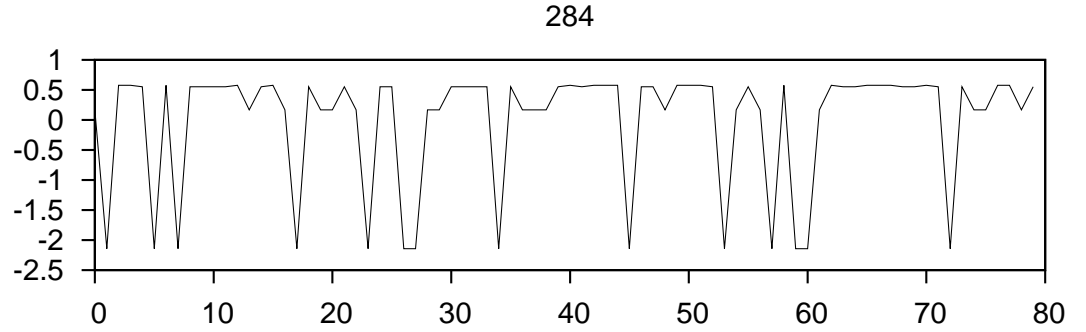


Figure 5.3: A sample promoter sequence represented in terms of EIIP values for nucleotides.

RNA polymerase interaction computations since they also provide EIIP values for amino acids. In case of FT, we used binary indicator sequences [110], enthalpy and roll angle encoding [28] and EIIP encoding [25]. Each sequence is encoded into a numerical sequence by using the encoding scheme. This numerical series is normalized to zero mean and unit standard deviation. Figure 5.3 depicts a sample sequence from the promoter data set which is converted into a numerical sequence using the EIIP values.

5.2.2 Feature Extraction

Using one of these encodings, the original sequence is converted into a numerical series. This numerical series is transformed using Fourier transform and wavelet transforms. In Fourier transform, discrete fourier transform (DFT) is applied to the promoter as well as non-promoter sequences to cull out the dominant components in frequency space. DFT is computed by using Fast fourier transform in MATLAB. The FFT coefficients are complex, hence the power spectrum is computed using the FFT coefficients as given equations 5.6. Here L is a power of 2 greater than or equal to the length of sequence S denoted by $|S|$.

$$Y = fft(sequence, L) \text{ where } L \geq |S| \quad (5.5)$$

$$power = abs(Y) \quad (5.6)$$

In wavelet transform, this series is finally decomposed using a discrete wavelet

transform into J levels [21]. In this thesis Bior3.3 biorthogonal wavelets are used to decompose the numerical promoter sequence. This wavelet has more resemblance to the sequence converted into a numerical sequence. Figure 5.4 gives these decompositions of a sample promoter sequence into 6 levels (i.e. J is 6). The output decomposition structure using wavelets contains the wavelet decomposition vector C and bookkeeping vector L . For example, a signal S can be decomposed into 6 levels by using *wavedec* function of MATLAB as follows:

$$[C, L] = \text{wavedec}(S, 6, 'bior3.3') \quad (5.7)$$

The decomposition structure of signal S into various levels(=3) is depicted in Figure 5.2. Wave reconstruction is done by using reverse transform and the decomposition structure $[C, L]$. The approximation and details coefficients of a particular level are obtained by using *wrcoef* function.

$$A_j = \text{wrcoef}('a', C, L, 'bior3.3', j), j = 1, 2, \dots, J \quad (5.8)$$

$$D_j = \text{wrcoef}('d', C, L, 'bior3.3', j), j = 1, 2, \dots, J \quad (5.9)$$

As described earlier, a major portion of the data set is used for training the classifier and the rest which is not exposed to the classifier is used as the test data set. We denote the set of promoters as positive data set and the set of non-promoters as the negative data set. Wavelet decomposition is done for each positive and negative sequence. This collection of vectors is divided into 5-folds in order to do the standard 5-fold cross-validation. A neural network classifier is then trained using the wavelet feature vectors. The test set is used to evaluate the performance of the classifier.

In case of non-promoter data set consisting of both gene and inter-gene portions, the proportion of positive data set to the negative data set is taken as 1:2. Each promoter and the non-promoter sequence of the data set is encoded by using the coding scheme of Cosic et al [25]. Each sequence is decomposed into 6 levels by using Bior3.3. In total there are 120 decomposition structure values which are required to decompose the original numerical sequence. The original wave is decomposed into 6 detail waves viz., D1, D2, D3, D4, D5, D6 and one

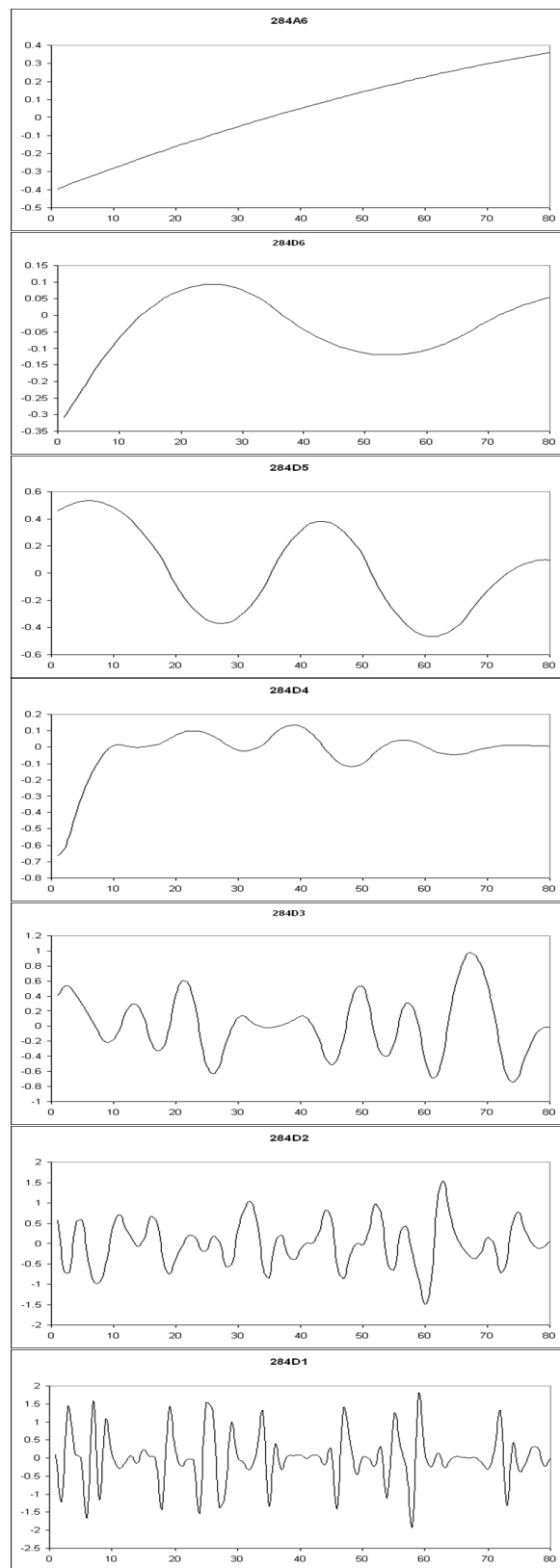


Figure 5.4: A sample promoter sequence decomposed into various levels using Bior3.3

smooth component A6, each of length 80. The classification is based upon various features that are extracted from these decomposition structure values and decomposed wavelet coefficients.

5.3 Classification

A multi-layer feed-forward neural network with three layers, namely, an input layer, one hidden and an output layer is used for promoter classification in the following classification sections with various features based upon signal analysis techniques. The number of nodes in the input layer is dependent on particular features that are used. Hidden layer consists of a certain number of hidden nodes, the number found by trial and error that gives optimal classification performance. The output layer has one node to give a binary decision as to whether the given input sequence is a promoter or a non-promoter. These simulations are done using Stuttgart Neural Network Simulator [104].

Neural network is trained on the training set and then the classification performance is evaluated on the test set. All the classification experiments are carried out using a 5-fold cross validation procedure [77, 3]. The classification results are evaluated using the performance measures such as *Precision*, *Specificity* and *Sensitivity*.

5.3.1 Classification Using FFT Coefficients

Single Nucleotide Binary-indicator Sequences

Earlier research using Fourier transform has shown that the coding region of eukaryotes gave a peak at $\frac{1}{3}$ pointing to a codon bias [102], which could be used in gene recognition in case of eukaryotes. In case of prokaryotes the periodicity of 3 is observed not only in coding regions but also in non-coding regions [108]. If different triplets are responsible for periodicities in coding and non-coding regions, they may become helpful in identifying promoters and non-promoters. Shrish et al. have shown that the gene portion often has a prominent peak at $1/3$ position confirming the periodicity of codons. But, non-genes do not have any such peak. They have decomposed the original sequence U into a set of four indicator sequences viz. U_A, U_G, U_T, U_C for the four nucleotides A, T, G and C

Table 5.2: Classification results using power spectrum values for *E.coli* using different encoding schemes.

Features	Precision%	Specificity%	Sensitivity%
EIIP	75.85	88.72	48.58
Binary indicators	73.98	88.36	43.5
Enthalpy	74.44	86.61	46.91
Roll Angle	No training		

[110]. Each sequence is a binary sequence indicating the presence or absence of a particular nucleotide. Equation 5.11 defines one of the four binary-indicator sequence, U_A . Other binary-indicator sequences can be similarly defined. Table 5.2 displays the classification results of *E.coli* when the power spectrum values as defined in equation 5.11 are used as features for a feed-forward neural network. The positive recognition results are not very encouraging even though negative recognition results are good, pointing possibly to the inseparability of coding versus non-coding sequences in this feature space.

$$U_\alpha(x_j) = \begin{cases} 1 & \alpha = A, x_j \text{ is the position of nucleotide. } 1 \leq x_j \leq 80 \\ 0 & \text{Otherwise} \end{cases} \quad (5.10)$$

$$Power = \sum_{i=1}^4 (P_i) \quad (5.11)$$

Where P_i is the power spectrum of the FT of each binary-indicator sequence.

Experimenting with various lengths starting from 80 bp to 350 bp in steps of 40 bp for coding and non-coding sequences, we found that sequence lengths greater than 200 bp might be needed to get a sizeable distinction near the 1/3rd peak. Hence it is possible that the promoters versus non-promoters are not able to throw up any distinct peak structure, which will be useful in classification of *E.coli* promoters.

In order to check the validity of the ideas, same experimentation is done on *Drosophila* data set [86] used in Chapter 3. Here again the sequence is represented as a set of four binary-indicator sequences, each indicating the presence or

Table 5.3: Classification results using power spectrum values for *Drosophila* using different encoding schemes.

Features	Precision%	Specificity%	Sensitivity%
EIIP	77.13	87.68	50.69
Binary indicators	77.98	86.51	56.66

absence of a particular nucleotide. The sensitivity is about 50-60% even though specificity is about 85%. When the intron data is removed from the total data set, now only consisting of promoter and coding sequences, the sensitivity improves to 86%. Further instead of binary values, if EIIP values are used in place of **1** in binary-indicator sequences, the sensitivity is much higher, it is about 94%, which is supported by Trifonov et al. [108] data. Hence it can be concluded that the intron part is similar to promoter, which is hindering the classification accuracy. In view of the above arguments, in case of *E.coli* there are two factors which are affecting the accuracy, one is the length of the sequence and second is the similarity of non-coding sequences to the promoter sequences. Then FFT of the the DNA sequence encoded using EIIP encoding, gives a slightly better accuracy compared to the other encodings for *E.coli*, and for *Drosophila* binary-indicator sequences give a marginal improvement over EIIP encoding.

Dinucleotide and Trinucleotide Binary-indicator Sequences

In Chapter 3 a global frequency count of 2-grams is done, so spatial information is lost there. It is of interest to see, what happens when the position information is retained and the spatial dependencies that exist between them. Hence, binary-indicator sequence method is extended further to use dinucleotides now. Here, 16 binary-indicator sequences are constructed by replacing a particular dinucleotide with a 1 or with a 0 if that particular dinucleotide is present or absent respectively. This method is depicted in equation 5.13 for dinucleotide AA. Similarly binary-indicator sequences for other 15 dinucleotides can also be defined. This is like counting 2-grams at the position of occurrence without summing them all as in 2-gram computation explained in Chapter 3 averaging and applying FT to each individual binary-indicator sequence and finally, summing them all.

$$\begin{aligned}
U_{\alpha}(x_j) &= 1 & \alpha &= AA \\
&= 0 & \text{Otherwise}
\end{aligned}
\tag{5.12}$$

$$Power = \sum_{i=1}^{16} (P_i) \tag{5.13}$$

Where P_i is the power spectrum of the FT of each binary-indicator sequence.

Now the power spectrum is obtained, using the FFT coefficients for binary-indicator sequences by summing them all for each spectrum value. Power-spectrum for the sequence is obtained as shown in 5.13. Power spectrum values as input features to the neural network to classify promoters and non-promoters. Similarly, trinucleotide binary-indicator sequences are also computed and the results have not improved any further.

5.3.2 Classification using wavelet coefficients

Simple Fourier transform is not enough to discriminate a promoter against coding and non-coding backgrounds as seen in the earlier section. The time or positional information is lost in a Fourier transform. Wavelet transform retains the positional as well as frequency information. The decomposition structure C has different values for each detail and approximate coefficients. The values are 8, 8, 9, 11, 16, 25, 43 for A6, D6, D5, D4, D3, D2 and D6 respectively. Total coefficients are 120. The classification accuracy of promoter recognition problem is computed using wavelet coefficients as input to the neural network. The results are given in Table 5.4. The results show that non-promoter recognition is good compared to promoter recognition.

Table 5.4: Classification results using wavelet coefficients as features for a neural network classifier for *E.coli* using EIIP encoding.

Features	Precision%	Specificity%	Sensitivity%
All 120 values	63.6	86.52	31.7

Table 5.5: Classification results using decomposed waves as features to a neural network classifier for *E.coli* using EIIP encoding.

Features	Precision%	Specificity%	Sensitivity%
All decomposed waves (560)	69.23	87.59	30.21

Table 5.6: Classification results using decomposed waves as features to a neural network classifier for *Drosophila* binary indicators encoding scheme.

Features	Precision%	Specificity%	Sensitivity%
All decomposed waves (2100)	77.61	89.14	50.62

5.3.3 Classification Using Decomposed Signals

Each decomposed wave is rebuilt using decomposed structure values into a wave of length 80. In total there are 7 waves viz., A6, D6, D5, D4, D3, D2, D1 resulting in 560 values (7×80). All waves are used to see whether more information is imparted by transforming one initial signal wave into so many decomposed waves. These 560 values are used as input features to a neural network classifier to identify the promoters. Table 5.5 presents the results of the classifier for these feature values. The results again are showing good non-promoter recognition than promoter recognition. It can be concluded that both experiments using wavelet coefficients and decomposed waves are good for non-promoter recognition. Increase in number of features has not helped in gaining more information to classify promoters better. Experiments on *Drosophila* data set also present similar kind of results. The sensitivity is about 50% for *Drosophila* using binary indicator sequences encoding scheme. The results are shown in Table 5.6.

5.3.4 Classification Using Cross-correlation Between Promoter and RNA-Polymerase

In this section we will look at the interaction between promoter and RNA polymerase as means of promoter recognition. Basically, interactions between protein and DNA can be categorized into 4 classes: DNA backbone - protein backbone (18%), DNA backbone - protein side chain (51%), DNA side chain - protein

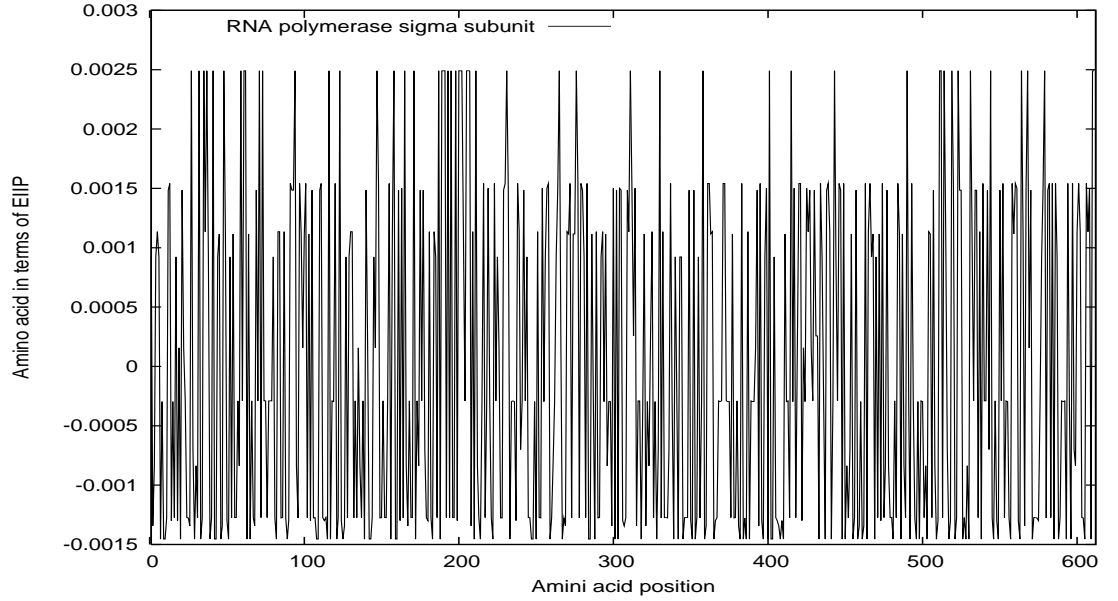


Figure 5.5: RNA Polymerase subunit sigma, in terms of EIIP values.

backbone (1%) DNA side chain - protein side chain (30%) [73]. Protein-DNA interactions are chemically the same as protein-protein interactions. They consist of electrostatic interactions, hydrogen bonds and hydrophobic interaction. However, hydrogen bonds constitute the major term for recognition and specificity and a large portion of the binding energy [69, 67]. It has been proposed that matching of periodicities within the distribution of energies of free electrons along the interacting proteins or protein and DNA can be regarded as resonant recognition [25]. The whole process can be observed as the interaction between transmitting and receiving antennae of a radio system.

The sigma subunit of the RNA polymerase is of 612 aa (amino acids) length. The subunit is converted into a numerical sequence using the EIIP values for the aminoacids [21]. Figure 5.5 shows the EIIP conversion of the sigma subunit. This particular subunit is also decomposed into 6 levels using the Bior3.3 biorthogonal wavelet. Figure 5.6 gives the decomposed sigma into various levels.

Cross-correlation between the waves is given in equation 5.14 [88, 21]. The cross-correlation coefficient for signals s_1 and s_2 of length n is defined as

$$\rho^{12}(j) = \frac{r^{12}(j)}{\frac{1}{N} \sqrt{[\sum_{n=0}^{N-1} s_1^2(n) \sum_{n=0}^{N-1} s_2^2(n)]}} \quad j = 0, \pm 1, \pm 2, \pm 3, \dots \quad (5.14)$$

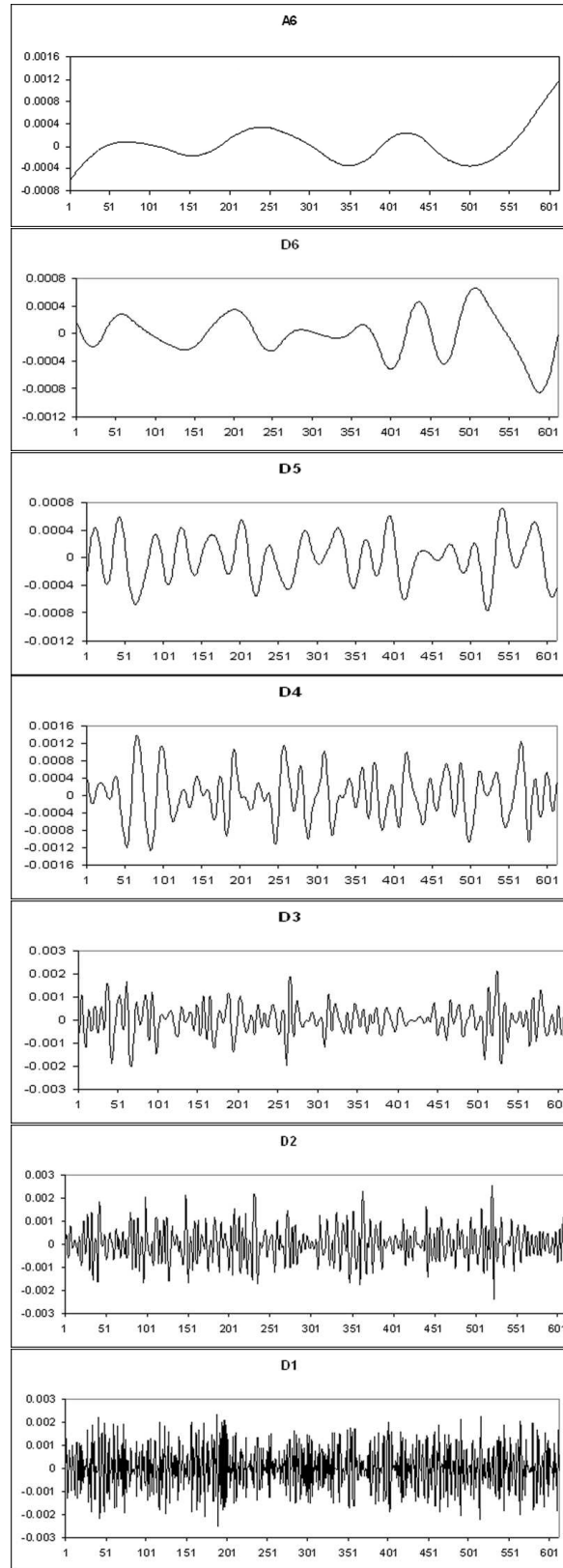


Figure 5.6: Sigma subunit decomposed into various levels.

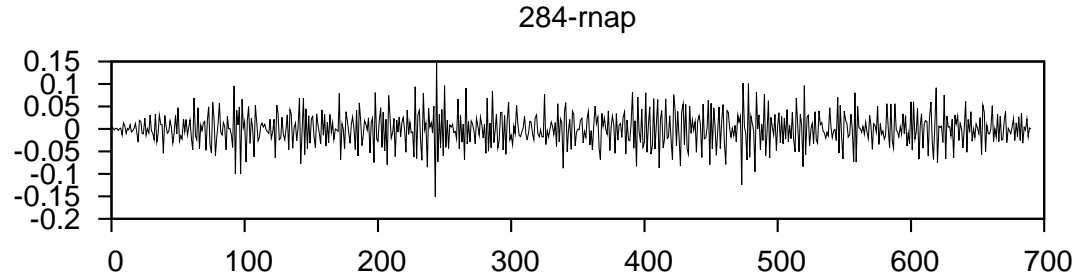


Figure 5.7: Cross-correlation between sample promoter and RNA Polymerase subunit sigma.

Table 5.7: classification results using DNA-RNA Polymerase sigma subunit cross-correlation values as features for a neural network classifier for *E.coli*.

Features	Precision%	Specificity%	Sensitivity%
EIIP encoding	66.46	86.18	24.65
Binary encoding	69.67	89.64	27.34

$$r^{12}(j) = \frac{1}{N} \sum_{n=0}^{N-1} s_2(n) s_1(n-j) \quad (5.15)$$

The maximum absolute value of the correlation coefficient at each decomposition level can be treated as the similarity score between the signals. Cross-correlation between RNA polymerase and sample sequence is given Figure 5.7. The number of correlation coefficients is 691. In case of binary encoding, wavelet at a particular level is obtained by taking the norm of 4 vectors for each i, j where i is the location and j is the level number as

$$W(j, i) = \sqrt{W_A(j, i)^2 + W_T(j, i)^2 + W_G(j, i)^2 + W_C(j, i)^2} \quad (5.16)$$

Now the interaction between each of the decomposed waves of promoter and the RNA sigma subunit is computed by using the equation 5.14 are used as the input features. The total number of features turns out to be 4842 in this case. The correlation between promoter and RNA polymerase sigma subunit is shown in Figure 5.7. Similarly the cross-correlation between decomposed levels of promoter and RNA polymerase are shown in Figure 5.8. Table 5.8 depicts the classification performance of the neural network classifier. DNA-polymerase

Table 5.8: Classification results using DNA-RNAP sigma cross-correlation at various levels.

Features	Precision%	Specificity%	Sensitivity%
EIIP encoding	No training at all		
Binary encoding	70.01	89.36	28.08

interaction using melting enthalpy and roll angle found not to be giving expected performance, when correlation computed by taking inverse fourier transform of the cross spectrum is used as the set of features for classification.

The results using features of cross-correlation between promoter and RNA polymerase sigma subunit, and cross-correlation between decomposed waves of both promoter and RNA polymerase sigma subunit have shown a remarkable ability to identify non-promoters. Hence this neural network using FFT or wavelet features may be used in conjunction with neural network using n-gram features to reinforce the identification of non-promoters. Interaction between anti-sense strand of RNA polymerase and promoter also gives similar results.

5.4 Discussion

Promoter recognition is addressed in this chapter by using features obtained from FT and wavelet transforms. Different encoding schemes EIIP, enthalpy, binary indicator, roll angle have been used in order to convert letter sequences to numerical sequence. The results using FT features for *E.coli* gave very good non-promoter recognition of $\sim 88\%$ and sensitivity of $\sim 48\%$. The results of FT, confirms the idea that mere frequency information is not enough to recognize a promoter since both gene (non-promoter) and non-gene (promoter or a intergenic portion which is not a promoter) in *E.coli* have a triplet nature. This may be the reason why, best recognition using n-grams was obtained for $n=3$ (extension of the work using bi-grams [95]). Different triplets could be dominating the gene and non-gene portions. Both promoter and non-promoters use different triplets, hence discrimination using 3-grams is efficient. As was pointed out in the classification section, length also could be one more factor contributing to this. And also, the encoding scheme that is used does not have much impact except

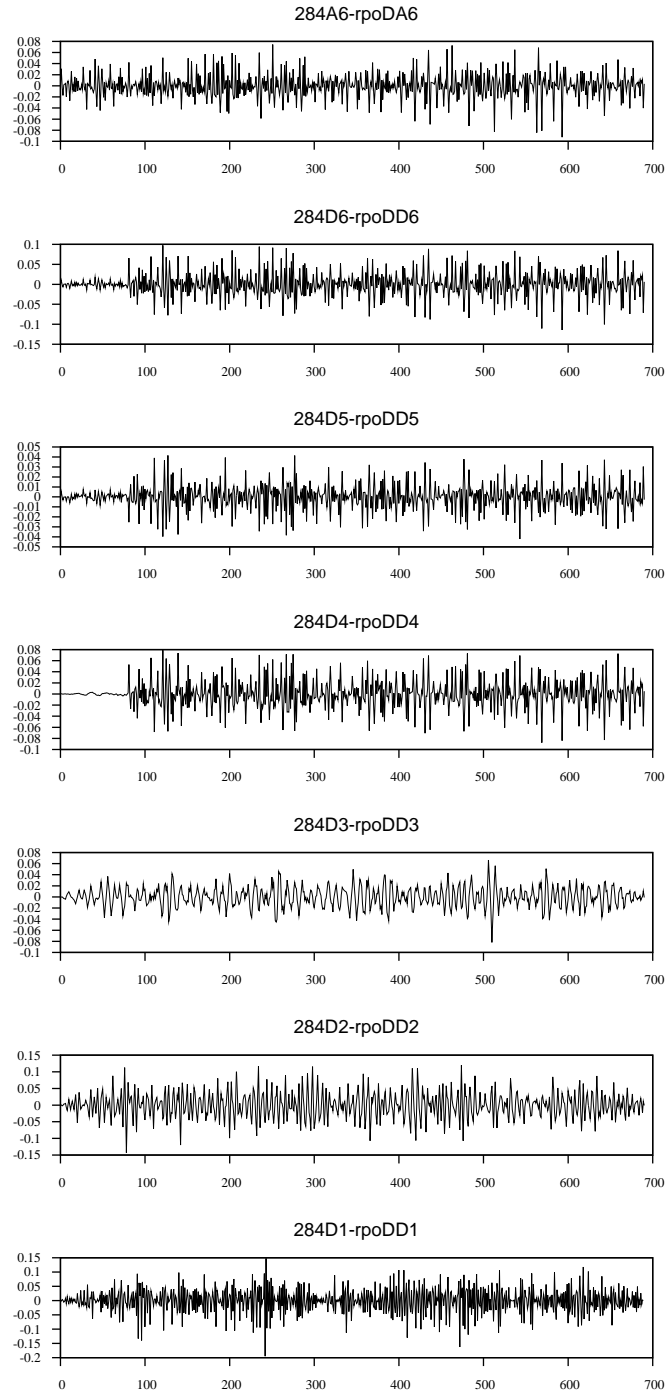


Figure 5.8: Cross-correlation between sample promoter and RNA Polymerase subunit sigma at various levels.

the roll angle for which no training is happening at all. Roll angle seems to be very sensitive to the letter changes in the promoter and non-promoter sequences, since training couldn't be achieved for this encoding. The results of Deyneko et al. have only displayed the ability to recognize a certain group of promoters whose genes respond to a particular stimulant but not as a generic tool for promoter recognition. In case of *Drosophila* the similarity between inter-gene of which promoter is a part and intron portions is hampering the classification process.

Kinetic and structural aspects of promoter and RNA polymerase interaction in prokaryotes are indicated in the study made by DeHaseth et al. [27]. Still, several other aspects were found to be unclear. It was indicated that the amino acid residue-nucleotide interactions which occur in various intermediate processes are to be identified to get clear picture of the process. In this context, we proposed promoter recognition using wavelet transform. In case of wavelet transform, both the experiments using coefficients and decomposed waves show a remarkable ability to recognize non-promoters and a reduction in recognizing promoters. Non-promoter recognition is very high at $\sim 88\%$. This turns out to have same result as in the case of position weight matrices which are dependent on the frequency of occurrence of a particular parameter at that position. It looks as though that by trying to incorporate the positional information, the recognition machinery is getting bogged down by the local information at that position. It can be surmised that the signal processing methods have not been successful in extracting the signal of a promoter.

Another approach of representing resonance effect of interaction between promoter and RNA polymerase has been attempted. The features are extracted in two ways here. First one by extracting features from cross-correlation between promoter/non-promoter and RNA polymerase sigma subunit. Second one by decomposing both promoter and RNA polymerase sigma subunit into waves at various levels using wavelets. Cross-correlation between the decomposed waves as well as the original sequences is computed. These are used as features to the neural network classifier. The results are again very good for non-promoter recognition but not encouraging for promoter recognition. Since RNA polymerase is an enzyme its 3-dimensional structure may expose amino acids in a different order (alpha helix, beta sheet and turns). In that case interaction between promoter and RNA polymerase would be quite different compared to mere promoter and sequential RNA polymerase interaction. This line of thought can be

further explored using RNA polymerase structural information.

5.5 Summary

A study is made to classify promoters using neural networks based on features obtained from Fourier transform and wavelet transforms. In order to obtain Fourier and wavelet transforms, the letter sequence is to be converted into numerical sequence. Different encoding schemes such as EIIP, melting enthalpy, binary indicators are used to convert the letter sequence into numerical sequence. Two approaches are used to recognize promoters. First one uses the features obtained by Fourier and wavelet transforms. Second approach uses the interaction between promoter and RNA polymerase sigma subunit modeled through wavelets.

The experiments are carried out to find a signature of the promoter present at any particular resolution. If the characteristic signal of a promoter which is supposed to be conserved irrespective of place of occurrence of the promoter in the genome is not preserved, then promoter recognition based on the signal processing technique will not be very helpful. This fact is evident from the results where the negative data seems to be recognized better than a promoter. The promoter data set seems to have no pattern in the FT and WT feature space which can be exploited by the classifier.

The results of *E.coli* and *Drosophila* show that the features using FFT and wavelets are not sufficient to recognize a promoter against a coding and non-coding background. The length could be an impediment for *E.coli*. For *Drosophila* the similarity or resemblance of intron region with noncoding region could be a negative factor. Different encoding schemes do not seem to make much difference to promoter recognition except marginally.

The assumption that the signal processing methods can capture the interaction between RNA polymerase sigma subunit and promoter has not fructified well. The particular encoding schemes and classification approach do not seem to exploit any resonance signal that might exist between RNA polymerase and DNA. Future work can build on these experiments using RNA polymerase structural information.

A positive conclusion from these experiments is that non-promoter recognition

is very good using signal processing techniques FT and WT. Hence, the classifier using these features can be used in conjunction with classifier using n-grams to reinforce the non-promoter identification.

Chapter 6

Conclusions and Future directions

6.1 Conclusions

Huge amount of genome data is available currently due to fast sequencing methods. But similar fast annotation methods of the genome are not available and the current technologies consume a lot of time. Hence machine annotation methods are required to tackle the major problems of promoter recognition and gene recognition.

In the literature, the techniques proposed for feature extraction can be broadly classified as those that exploit biological information and those that may not be biologically relevant. In the former category, the features are extracted from the binding regions or local motifs whereas in the latter, global signal methods, features are derived utilizing the physico-chemical and structural properties of the whole promoter region. The recognition methods that exploit the promoter signal like the position weight matrices (PWM), the expectation maximization algorithm and the techniques like Fourier transforms are proposed in the literature. On the other hand, one can explore n-gram based features which may not necessarily be biologically relevant for the problem of promoter recognition. In this thesis, the promoter recognition problem is researched from these two different perspectives and the following conclusions are arrived at from this study.

Conclusion 1: n-gram based features perform the best for the whole genome annotation.

In Chapter 3, a systematic study of n-grams with $n=2,3,4,5$ as features for a neural network classifier is carried out. A set of experiments is set up to estimate

the recognition accuracy of a neural network classifier for the promoter recognition problem which uses n -gram features as input. The preliminary results show that for *E.coli*, 3 – grams give better performance than other n -grams whereas for *Drosophila* 4 – grams give an optimal performance. It can be concluded from these results that different adjacency preferences are shown for promoter and non-promoter regions. Using 3-grams which gave best recognition rate out of the n -gram features considered, a genome-wide promoter recognition is attempted in a limited portion of the genome available in NCBI database using a set of neural networks in a cascaded manner. This method which identifies promoters in a whole genome has produced very satisfactory results. No true promoter is missed. Wet lab experiments need to validate if the additional promoter regions are in fact true promoters. The results of this scheme are compared with other software tools. One of them is SAK developed by Gordon et al. [36]. Others are NNPP which uses local information from TATA and Inr [82], BPROM which uses functional motifs and oligonucleotide information [84] and PPP that uses Hidden Markov model [94]. Out of these, only SAK uses the signal from the whole promoter, whereas others adopt basically local content extraction methods and we show that the performance of n -gram based classifier is far superior. The proposed annotation scheme should be further fine-tuned in order to tackle issues of setting up an appropriate threshold and identifying the segment ends. One major advantage of using n -gram based whole genome promoter annotation is that one does not need any prior information about the promoter, the location of binding sites, the spacer lengths etc. Vector size of n -grams will not change irrespective of length of promoter sequence.

Conclusion 2: Existing Schemes proposed in the literature are not really extendable to whole genome annotation.

The performance of an algorithm on a limited training and test data set may not be really a performance indicator of how well it may identify promoters in a whole genome. Methods that are proposed giving good accuracies on training and test data sets, may or may not perform better on the whole genome. In Section 4.4.3 of Chapter 4 it is shown that the performance rates of classifier using position weight matrix based features from binding sites are better than the results obtained with 3-grams. Interestingly whole genome annotation results give a diametrically opposite picture. In comparison, 3-grams perform much better on the whole genome annotation than the extension of the position-weight

matrix based features scheme for the whole genome annotation.

Conclusion 3: Whole promoter based features are more meaningful in promoter classification.

The binding regions are important in assisting the RNA polymerase to bind the promoter, but that alone is not sufficient to recognize a promoter. Local features are calculated from binding sites which are available in Harley's data [41]. Global features are extracted from the whole promoter sequences aligned with respect to TSS and non-promoter data sets. If we compare the results that are obtained using signal extracted from motifs/binding regions in the promoter with the signal obtained from the whole promoter sequence, it is evident that the local signal has a handicap in extending to a whole genome promoter prediction. The signal from the entire promoter sequence without segmenting it into important and non-important portions will lead to better generalization in case of *E.coli*. PWM based features as well as n-grams from the whole promoter give better genome promoter annotation results than the other local signal extraction schemes. These results are presented in Section 4.4.5 and Section 3.4.2 respectively. This result fortifies the idea that the whole promoter sequence is required for promoter prediction. One good outcome of the results using local features is accurate prediction of TSS. Position weight matrix based global feature extraction method is extendable to *Drosophila* data set also.

Conclusion 4: N-gram feature space seems best suited for eukaryotic promoter classification.

In n-gram feature space, recognition rate of *Drosophila* promoters is better than *E.coli* promoter recognition. Best performance for *Drosophila* is 87% compared to *E.coli*'s 80% with a very good positive predictive rate is shown in Section 3.2.1. Promoter architecture of eukaryotes is in general much more complex than a prokaryote promoter architecture. It is found that n-gram preferences for *Drosophila* is stronger in discriminating promoter versus a non-promoter than for *E.coli*. The prediction of the negative data set is higher possibly because the intron portions are similar to non-gene segments in eukaryotes. The results of promoter recognition using Fourier transform in fact give only about 50% positive identification. In comparison with this result, n-grams are performing very well.

Conclusion 5: Synthetic negative data set is not appropriate as negative data

set in n-gram feature space.

Machine learning approaches are dependent on data sets that are used for training and testing. In the literature, usually 60% A+T rich sequences are generated to constitute a synthetic negative data set. Performance of the classifier using a negative data set extracted from the genome is not as good as the performance obtained with the synthetic data set as negative data set. This fact is evident from the experiment, where the training is done with promoter as positive data and 60% A+T rich synthetic data as negative data which gives a precision rate of 95% using a *single layer perceptron* as given in Section 3.2.2. When this classifier is used on a test data consisting of promoter as positive and coding data as negative data, about 93% promoters are recognized and 0% non-promoters are recognized. Hence it can be concluded that the adjacency properties definitely are very different in promoter and non-promoter sets in the n-gram feature space. Synthetic negative data sets only ensure that they have composition similar to promoters but the adjacency cannot be dictated in this case.

Conclusion 6: Investigation of biological-mechanism inspired promoter recognition method was not fruitful.

The assumption that signal processing methods can capture the interaction between RNA polymerase sigma subunit and promoter has not fructified well. Chapter 5 explores the objective of applicability of signal processing techniques as well as analyzing promoter and RNA polymerase interaction through signal processing techniques. A study is made to classify promoters using neural networks based on features obtained from Fourier transform and wavelet transforms. Two approaches are used to recognize promoters. First one uses the features obtained by Fourier and wavelet transforms. Second approach uses the interaction between promoter and RNA polymerase sigma subunit modeled through wavelets. The different encoding schemes including those that use the structural properties of the genome do not influence the classification performance.

Experiments are carried out to find a signature of the promoter present at any particular resolution. If the characteristic signal of a promoter which is supposed to be conserved irrespective of place of occurrence of the promoter in the genome is not preserved, then promoter recognition based on the signal processing techniques will not be very helpful. The results of *E.coli* and *Drosophila* show that features using FFT and wavelets are not sufficient to recognize a promoter

against a coding and non-coding background. This fact is evident from the results where the negative data seems to be recognized better than a promoter. The promoter data set seems to have no pattern in the FT, WT feature space which can be learned by the classifier. In summary, it turns out that signal processing techniques cannot be used as general classification algorithm. The conjecture that structural properties are underlying principles for base selection is not evident from the current experiments.

Conclusion 7: Promoter recognition problem is a hard problem.

Carninci et al [19] reveal that Eukaryotes have two distinct promoter signals one which are TATA box enriched promoters and the other that are CpG rich promoters. In Section 3.3 we found that there exist two distinct promoter signals in E.coli, we call them, majority signal and minority signal. We found that in the data set of the promoter sequences, nearly 20% of the sequences lie in the coding region and hence seem to be closer to the negative data set. Similarly it was found that the negative data set too has a majority and a minority signal where the minority signal is very close to the majority promoter signal. This intrinsic confusion may lead to a hard limiter on the performance results of classification.

6.2 Future directions

The techniques described in the thesis can be developed to build a full-fledged automated whole genome promoter annotation tool. Whole genome promoter prediction using 3-grams has been applied to forward strand only. The same scheme needs to be applied to the reverse strand with appropriate preprocessing so that the method can achieve whole genome promoter recognition.

Signal processing techniques need to be explored further to enhance the recognition rates. Since RNA polymerase is an enzyme, which is a protein, 3-dimensional structure information of RNA polymerase can be used to characterize promoter-RNA polymerase interaction. It is to be explored if instead of using Bior3.3 as the mother wavelet, RNA polymerase itself can be used as a mother wavelet to imitate the biological mechanism closely. This line of application of wavelets would be quite novel.

Mitochondrial RNA polymerase (human genome) and chloroplast RNA poly-

merase (plant genome) are similar to bacterial RNA polymerase and their promoters resemble the promoter structure of *E.coli*. It would be meaningful to apply the feature extraction methods discussed in the thesis to these other classes of genomes.

Proteins are believed to be responsible for most of the genetically important functions in all cells. Hence the focus has been entirely on gene recognition. Recent studies indicate that ncRNAs (noncoding RNAs), which do not code for proteins, affect transcription and the chromosome structure, in RNA processing and modification, regulation of mRNA stability and translation, and also affect protein stability and transport [75, 35]. An effort to look for them in typically 95% of the total DNA is a huge task. By suitably modeling the promoters of these ncRNAs, these can be predicted much more easily. Some of the techniques that are developed here can be extended to do this work.

References

- [1] K. Aditi and B. Manju. A novel method for prokaryotic promoter prediction based on dna stability. *BMC Bioinformatics*, 6(1):doi: 10.1186/1471-2105-6-1, 2005.
- [2] J.D. Alicia, D.J. Bradley, L. Michael, and A.G. Carol. The sigma subunit of escherichia coli rna polymerase senses promoter spacing. *Proceedings of the National Academy of Sciences*, 93:8858–8862, 1996.
- [3] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, India, 2004.
- [4] A. Arneodo, E. Bacry, P.V. Graves, and F. Muzy. Characterizing long-range correlations in dna sequences from wavelet analysis. *Physical Review Letters*, 74(16):3293–3297, 1995.
- [5] V.B. Bajic, M.R. Brent, R.H. Brown, A. Frankish, J. Harrow, U. Ohler, V.V. Solovyev, and S.L. Tan. Performance assessment of promoter predictions on encode regions in the egasp experiment. *Genome Biology*, 7:Online, 2006.
- [6] V.B. Bajic, A. Chong, S.H. Seah, and V. Brusic. An intelligent system for vertebrate promoter recognition. *IEEE Intelligent Systems*, pages 64–70, 2002.
- [7] V.B. Bajic, S.H. Seah, A. Chong, G. Zhang, J.L.Y. Koh, and V. Brusic. Dragon promoter finder: recognition of vertebrate rna polymerase ii promoters. *Bioinformatics*, 18(1):198–199, 2002.
- [8] V.B. Bajic, S.L. Tan, Y. Suzuki, and S. Sugano. Promoter prediction analysis on the whole human genome. *Nature Biotechnology*, 22:1467–1473, 2004.
- [9] Berkeley Drosophila Genome Project (BDGP). <http://www.fruitfly.org>.

-
- [10] R.G. Beiko and R.L. Charlebois. Gann(genetic algorithm neural networks for the detection of conserved combinations of features in dna). *BMC Bioinformatics*, 6(36):doi:10.1186/1471-2105-6-36, 2005.
 - [11] I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse. Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics*, 21:2657–2666, 2005.
 - [12] R.D. Blake, J.W. Bizzaro, J.D. Blake, G.R. Day, S.G. Delcourt, J. Knowles, K.A. Marx, and J.Jr SantaLucia. Statistical mechanical simulation of polymeric dna melting with meltsim. *Bioinformatics*, 15:370–375, 1999.
 - [13] R.D. Blake and S.G. Delcourt. Thermal stability of dna. *Nucleic Acids Research*, 26:3323–3332, 1998.
 - [14] A. Bolshoy, P. McNamara, R.E. Harrington, and E.N. Trifonov. Curved dna without a-a: experimental estimation of all 16 dna with experimentally determined tata wedge angles. *Proceedings of the National Academy of Sciences*, 88:2312–2316, 1991.
 - [15] K.J. Breslauer, R. Frank, H. Blocker, and L.A. Marky. Predicting dna duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*, 83:3746–3750, 1986.
 - [16] I. Brukner, R. Sanchez, D. Suck, and S. Ponger. Trinucleotide models for dna bending propoensity: Comparison of models for dnasei digestion and nucleosome packaging data. *Journal of Biomoleculat Structure & Dynamics*, 13:309–317, 1995.
 - [17] C.R. Calladine and D.R. Drew. *Molecular Structure and Life*. CRC Press, 1992.
 - [18] L.R. Cardon and G.D. Stormo. Expectation maximization algorithm of identifying protein-binding sites with variable lengths from unaligned dna fragments. *Journal of Molecular Biology*, 223:159–170, 1992.
 - [19] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P.G. Engstrom, M.C. Frith, A.R.R. Forrest, W.B. Alkema, S.L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M.K. Katayama, Y. Kitazume, H. Kawaji, C. Kai,

- M. Nakamura, H. Konno, K. Nakano, S.M. Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S.M. Grimmond, C.A. Wells, V. Orlando, C. Wahlestedt, E.T. Liu, M. Harbers, J. Kawai, V.B. Bajic, D.A. Hume, and Y. Hayashizakia. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38:626–636, 2006.
- [20] National center for Biotechnology Information (NCBI). <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=1786181>.
- [21] H.T. Chafia, F. Qian, and I. Cosic. Protein sequence comparison based on the wavelet transform approach. *Protein Engineering*, 15(3):193–203, 2002.
- [22] T.V. Chalikian, J. Volkner, G.E. Plum, and K.J. Breslauer. A more unified picture for the thermodynamics of nucleic acid duplex melting: A characterization by calorimetric and volumetric techniques. *Proceedings of the National Academy of Sciences*, 96:7853–7858, 1999.
- [23] C.K. Chui. *An introduction to wavelets*. Academic press, New York, 1992.
- [24] J.M. Claverie and S. Audic. The statistical significance of nucleotide position-weight matrix matches. *CABIOS*, 12(5):431–439, 1996.
- [25] I. Cosic. Macromolecular bioactivity(is it resonant interaction between macromolecules?-theory and applications. *IEEE Transactions on Biomedical Engineering*, 41(12):1101–1114, 1994.
- [26] I. Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, 1992.
- [27] P.L. DeHaseh, M.L. Zupancic, and Jr.M.T. Record. Rna polymerase-promoter interactions: the comings and goings of rna polymerase. *Journal of Bacteriology*, 180:3019–3025, 1998.
- [28] I.V. Deyneko, E.K. Alexander, B. Helmut, and G. Kauer. Signal-theoretical dna similarity measure revealing unexpected similarities of e. coli promoters. *In Silico Biology*, 5, 2005.
- [29] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.

-
- [30] Eukaryotic Promoter Database (EPD). <http://www.epd.isb-sib.ch/index.html>.
- [31] J.W. Fickett and A.G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Research*, 7:861–878, 1997.
- [32] T. Freund and R.E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 325–332, 1996.
- [33] M.R. Gartenberg and D.M. Crothers. Dna sequence determinants of cap-induced bending and protein binding affinity. *Nature*, 333:824–829, 1988.
- [34] N.I. Gershenzon, G.D. Stormo, and I.P. Ioshikhes. Computational technique for improvement of the position-weight matrices for the dna/protein binding sites. *Nucleic Acids Research*, 33(7):22902301, 2005.
- [35] S. Gisela. An expanding universe of noncoding rnas. *Science*, 296:1260–1263, 2002.
- [36] L. Gordon, A. Y. Chervonenkis, A. J. Gammerman, I. A. Shahmurradov, and V. Solovyev. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, 19:1964–1971, 2003.
- [37] A.A. Gorin, V.B. Zhurkin, and W.K. Olson. B-dna twisting approach correlates with base-pair morphology. *Journal of Molecular Biology*, 247:34–48, 1995.
- [38] O. Gotoh and Y. Tagashira. Stabilities of nearest-neighbor doublets in double-helical dna determined by fitting calculated melting profiles to observed profiles. *Biopolymers*, 20:1033–1042, 1981.
- [39] R. Grosschedl and M.L. Birnstiel. Identification of regulatory sequences in the prelude sequences of an h2a histone gene by the study of specific deletion mutants in vivo. *Proceedings of the National Academy of Sciences*, 77(12):7102–7106, 1980.
- [40] FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, 2005.

-
- [41] C.B. Harley and R.P. Reynolds. Analysis of e.coli promoter sequences. *Nucleic Acids Research*, 15(5):2343–2361, 1987.
 - [42] M.A. El Hassan and C.R. Calladine. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in dna. *Journal of Molecular Biology*, 259:95–103, 1996.
 - [43] D.K. Hawley and W.R. McClure. Compilation and analysis of escherichia coli promoter dna sequences. *Nucleic Acids Research*, 11:2237–2255, 1983.
 - [44] S. Haykin. *Neural networks A comprehensive foundation*. Pearson education, 2001.
 - [45] P.S. Ho, G.W. Zhou, and L.B. Clark. Polarized electronic spectra of z-dna single crystals. *Biopolymers*, 30:151–163, 1990.
 - [46] Y.F. Huang and C.M. Wang. Integration of knowledge discovery and artificial intelligence approaches for promoter recognition in dna sequences. In *Proceedings of the Third International Conference on Information Technology and Applications (ICITA05)*, 2005.
 - [47] A.M. Huerta and J. Collado-Vides. Sigma70 promoters in escherichia coli: Specific transcription in dense regions of overlapping promoter-like signals. *Journal of Molecular Biology*, 333:261–278, 2003.
 - [48] L. Hunter. *Introduction to Bioinformatics, Lecture series*. <http://compbio.uchsc.edu/hunter/bioi7711>.
 - [49] V.I. Ivanov and L.E. Minchenkova. The a-form of dna (in search of the biological role). *Molecular Biology*, 28:780–788, 1995.
 - [50] N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI2000), Special Track on Inductive Learning*, pages 111–117, 2000.
 - [51] F. Jerome, H. Trevor, and T. Robert. Additive logistic regression(a statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, 2000.
 - [52] H. Ji, D. Xinbin, and Z. Xuechun. A systematic computational approach for transcription factor target gene prediction. In *IEEE Symposium on*

-
- Computational Intelligence and Bioinformatics and Computational Biology CIBCB '06*, pages 1–7, 2006.
- [53] H. Jim. *General transcription in initiation factors (Molecular biology Lecture notes)*.
- [54] G. Kauer and B. Helmut. Applying signal theory to the analysis of biomolecules. *Bioinformatics*, 19(16):2016–2021, 2003.
- [55] Sequence Alignment Kernel. http://nostradamus.cs.rhul.ac.uk/leo/sak_demo.
- [56] F. Kobe, S. Yvan, D. Sven, R. Pierre, and Y.P. Peer. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Research*, 33(13):4255–4264, 2005.
- [57] A. Krishnan, K.B. Li, and P. Issac. Rapid detection of conserved regions in protein sequences using wavelets. *In Silico Biology*, 4, 2004.
- [58] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186, 1997.
- [59] Udyant Kumar. *Promoter and other TFBS recognition - in silico*. www.cs.helsinki.fi/u/skaski/bioinf_semin05/slides_lect4.pdf.
- [60] C.E. Lawrence and A.A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function and Genetics*, 7:41–51, 1990.
- [61] F. Leu, N. Lo, and L. Yang. Predicting vertebrate promoters with homogeneous cluster computing. In *Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, pages 143–148, 2005.
- [62] V.G. Levitsky and A.V. Katokhin. Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis. *In silico Biology*, 3, 2003.

-
- [63] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference of Machine Learning*, pages 148–156, 1994.
- [64] Q.Z. Li and H. Lina. The recognition and prediction of σ^{70} promoters in escherichia coli k-12. *Journal of Theoretical Biology*, 242(1):135–141, 2006.
- [65] W. Li. The study of correlation structures of dna sequences: a critical review. *Computers & Chemistry*, 21:257–271, 1997.
- [66] C. Ling and C. Li. Data mining for direct marketing problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 73–79, 1998.
- [67] K. Luger, A.W. Mder, R.K. Richmond, D.F. Sargent, and T.J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389:251–260, 1997.
- [68] H. Luo, G. Gilinger, D. Mukherjee, and V. Bellofatto. Transcription initiation at the tata-less spliced leader rna gene promoter requires at least two dna-binding proteins and a tripartite architecture that includes an initiator element. *Journal of BioChemistry*, 274(45):31957–54, 1999.
- [69] N.M. Luscombe, R.A. Laskowski, and J.M. Thornton. Amino-acid base interactions a three-dimensional analysis of protein-dna interactions at atomic level. *Nucleic Acids Research*, 29:2860–2874, 2001.
- [70] Q. Ma, J.T.L. Wang, D. Shasha, and C.H. Wu. Dna sequence classification via an expectation maximization algorithm and neural networks: a case study. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, Special Issue on Knowledge Management*, 31:468–475, 2001.
- [71] I. Mahadevan and I. Ghosh. Analysis of e.coli promoter structures using neural networks. *Nucleic Acids Research*, 22:2158–2165, 1994.
- [72] S.G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on pattern analysis and machine intelligence*, 1(7):684–704, 1989.

-
- [73] Y. Mandel-Gutfreund, O. Schueler, and H. Margalit. Comprehensive analysis of hydrogen bonds in regulatory protein dna-complexes: In search of common principles. *Journal of Molecular Biology*, 253:370–382, 1995.
- [74] T. Matsuda, H. Motoda, and T. Washio. Graph based induction and its applications. *Advanced Engineering Informatics*, 16:135–143, 2002.
- [75] J.S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding rnas in complex organisms. *BioEssays*, 25:930–939, 2003.
- [76] S.L. McKnight and K.R. Yamamoto. Transcriptional regulation. *Cold Spring Harbor Laboratory Press*, 1992.
- [77] T.M. Mitchell. *Machine Learning*. McGraw Hill, Singapore, 1997.
- [78] M.E. Mulligan. *Bacterial RNA Polymerase (Lecture notes)*. www.mun.ca/biochem/courses/3107/Lectures/Topics/RNAP_bacterial.html.
- [79] M.E. Mulligan and W.R. McClure. Analysis of the occurrence of promoter-sites in dna. *Nucleic Acids Research*, 14(1):109–126, 1986.
- [80] N.J. Nilsson. *Learning machines: Foundations of the trainable pattern classifying systems*. McGraw Hill, New York, 1965.
- [81] V.C. Nitesh, W.B. Kevin, O.H. Lawrence, and K.W. Philip. Smote(synthetic minority over-sampling technique). *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- [82] Neural Network Promoter Predictor (NNPP). http://www.fruitfly.org/seq_tools/promoter.html.
- [83] K. Nobuyuki, O. Yashurio, M. Kazuo, M. Kenichi, K. Jun, C. Piero, H. Yoshihide, and K. Shoshi. Wavelet profiles: Their application in oryza sativa dna sequence analysis. In *Proceedings of the IEEE computer society Bioinformatics conference(CSB02)*, pages 345–348, 2002.
- [84] Prediction of Bacterial promoters (BPRM). <http://www.softberry.com/berry.phtml?topic=bprom>.
- [85] U. Ohler. Promoter prediction on a genomic scale—the adh experience. *Genome Research*, 10(4):539–542, 2000.

-
- [86] U. Ohler, G.C. Liao, H. Niemann, and G.M. Rubin. Computational analysis of core promoters in the drosophila genome. *Genome Biology*, 3:doi:10.1186/gb-2002-3-12-research0087, 2002.
- [87] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin. Dna sequence-dependent deformability deduced from protein-dna crystal complexes. *Proceedings of the National Academy of Sciences*, 95:11163–11168, 1998.
- [88] A.V. Oppenheim and R.W. Schaffer. *Discrete-time signal processing*. Prentice-Hall, 1993.
- [89] R.L. Ornstein, R. Rein, D.L. Breen, and R.D. Macelroy. An optimized potential function for the calculation of nucleic acid interaction energies: base stacking. *Biopolymers*, 17:23412360, 1987.
- [90] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning San Francisco*, pages 217–225, 1994.
- [91] A.G. Pedersen, P. Baldi, Y. Chauvinb, and S. Brunak. The biology of eukaryotic promoter prediction - a review. *Computers & Chemistry*, 23:191–207, 1999.
- [92] L. Pietro. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.
- [93] J.V. Ponomarenko, M.P. Ponomarenko, A.S. Frolov, D.G. Vorobyev, G.C. Overton, and N.A. Kolchanov. Conformational and physicochemical dna features specific for transcription factor binding sites. *Bioinformatics*, 15:654–668, 1999.
- [94] Prokaryotic Promoter Prediction (PPP).
http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp_start.php.
- [95] T. Sobha Rani, S. Durga Bhavani, and S.B. Raju. Analysis of e.coli promoter recognition problem in dinucleotide feature space. *Bioinformatics*, 23:582–588, 2007.

-
- [96] M.G. Reese. Application of time-delay neural networks to promoter annotation in drosophila melanogaster genome. *Computers & Chemistry*, 26(1):51–56, 2001.
- [97] Genome Science Group (Genome Network Project Core Group) RIKEN Genome Exploration Research Group and FANTOM Consortium. Anti-sense transcription in the mammalian transcriptome. *Science*, 309:1564–1566, 2005.
- [98] S.C. Satchwell, H.R. Drew, and A.A. Travers. Sequence periodicities in chicken nucleosome core dna. *Journal of Molecular Biology*, 191:659–675, 1986.
- [99] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.
- [100] F.G. Scott. *A companion to developmental biology (Online Book)*.
- [101] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.
- [102] T. Shrish, S. Ramachandran, B. Alok, B. Sudha, and R. Ramakrishna. Prediction of probable genes by fourier analysis of genomic sequences. *CABIOS*, 13(3):263–270, 1997.
- [103] A.V. Sivolob and S.N. Khrapunov. Translational positioning of nucleosomes on dna: the role of sequence-dependent isotropic dna bending stiffness. *Journal of Molecular Biology*, 247:918–931, 1995.
- [104] Stuttgart Neural Network Simulator (SNNS). <http://www-ra.informatik.uni-tuebingen.de/SNNS/>.
- [105] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.
- [106] N. Sugimoto, S. Nakano, M. Yoneyama, and K. Honda. Improved thermodynamic parameters and helix initiation factor to predict stability of dna duplexes. *Nucleic Acids Research*, 24:4501–4505, 1996.

-
- [107] M. Suzuki, N. Yagi, and J.T. Finch. Role of base-backbone and base-base interactions in alternating dna conformations. *FEBS Letters*, 379:148–152, 1996.
 - [108] E.N. Trifonov and J.L. Sussman. The pitch of chromatin dna is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences*, 77:3816–3820, 1980.
 - [109] A. Vezhnevets and V. Vezhnevets. Modest adaboost - teaching adaboost to generalize better. In *Graphicon*, 2005.
 - [110] R.F. Voss. Evolution of long-range fractal correlations and 1/f noise in dna base sequences. *Physical Review Letters*, 68(5):3805–3808, 1992.
 - [111] H. Wang and C.J. Benham. Promoter prediction and annotation of microbial genomes based on dna sequence and structural responses to superhelical stress. *BMC Bioinformatics*, 7:doi: 10.1186/1471-2105-7-248, 2006.
 - [112] T. Werner. Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome*, 10:168–175, 1999.
 - [113] Y. Xu, R.J. Mural, J.R. Einstein, M.B. Shah, and E.C. Uberbacher. Grail: A multi-agent neural network system for gene identification. *Proceedings of IEEE*, 84(10):1544–1552, 1996.

Appendix A

Promoter: A primer

The central dogma is crucial for biology. Any protein that is to be synthesized, is to be synthesized only using this dogma. Here the gene is transcribed onto a RNA and after various processing steps is translated into a linear amino-acid chain. This amino-acid chain folds into a three-dimensional protein structure. This whole process happens only after RNA polymerase binds to a switch like sequence segment that occurs upstream of a gene called promoter. This process is depicted in the Figure 1.1. The corresponding gene structure is shown Figure A.1 [48].

Promoters of genes that transcribe relatively large amounts of mRNA have similarities. They have a TATA sequence (sometimes called the TATA box or Goldberg-Hogness box) about 30 base pairs upstream from the site where transcription begins, as well as one or more promoter elements further upstream as shown in Figure A.2 [39, 76]. Figure A.2 depicts the promoter structure in general [100].

Transcription requires the interaction of RNA polymerase with promoter DNA. In eukaryotic cells, there are three different types of RNA polymerase, each having a particular functions and properties (Rutter et al. 1976). RNA polymerase I is found in the nucleolar region of the nucleus and is responsible for transcribing the large ribosomal RNAs; RNA polymerase II transcribes messenger RNA precursors; and RNA polymerase III transcribes small RNAs such as transfer RNA, 5S ribosomal RNA, and other small DNA sequences. None of the eukaryotic RNA polymerases can bind efficiently to DNA. Rather, there are families of DNA-binding proteins that first bind to DNA and, once bound, inter-

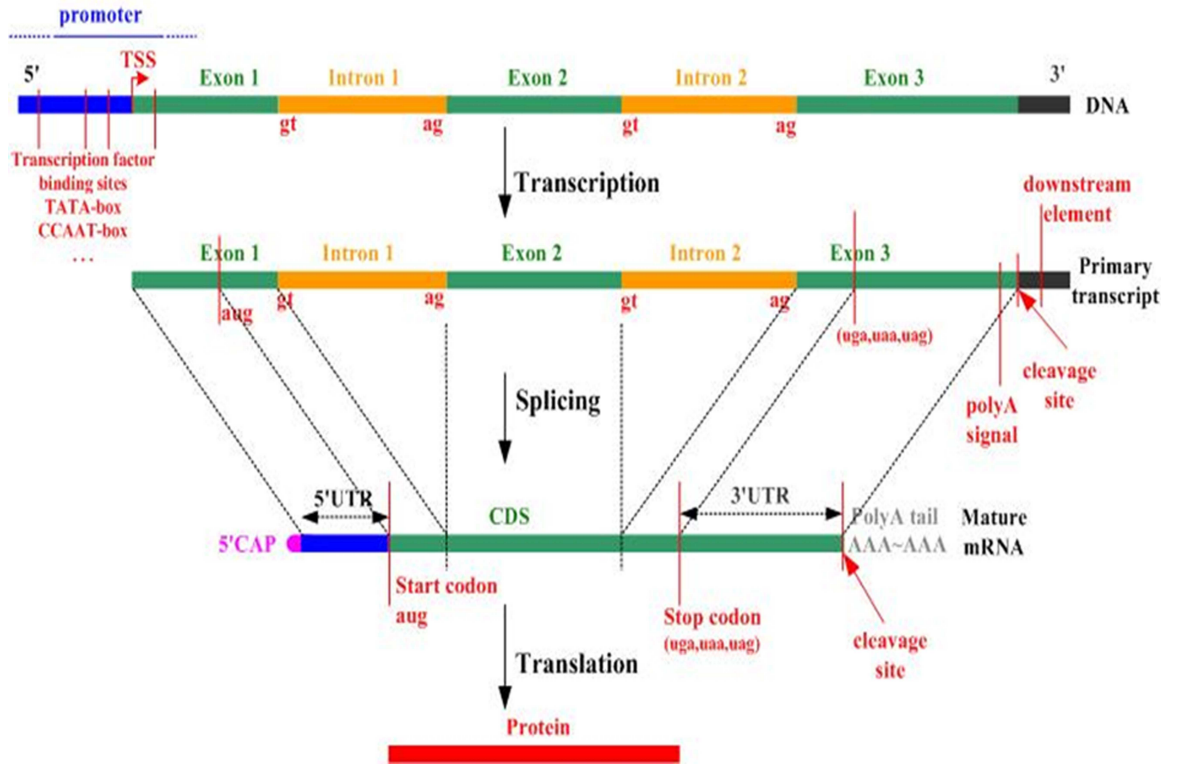


Figure A.1: Eukaryotic gene structure [48]

act with the RNA polymerase to initiate RNA synthesis. Whereas, in prokaryotes, RNA is synthesized by a single polymerase. RNA polymerase of *E.coli* is a multisubunit enzyme. This complex enzyme contains four kinds of subunits. The subunit composition of the entire enzyme is called holoenzyme consisting of $\alpha_2\beta\beta'\omega\sigma$. The sigma subunit finds a promoter site where transcription has to begin, helps in initiating RNA synthesis, and then dissociates from the rest of the enzyme. The process is depicted in Figure A.3 [53]. α_2 : the two α subunits assemble the enzyme and recognize regulatory factors. Each subunit has two domains: α CTD (C-Terminal domain) binds the UP element of the extended promoter, and α NTD (N-terminal domain) binds the rest of the polymerase. β : this has the polymerase activity (catalyzes the synthesis of RNA) which includes chain initiation and elongation. β' : binds to DNA (nonspecifically). ω : restores denatured RNA polymerase to its functional form *in vitro*.

The promoter structure of *E.coli* is basically thought of consisting of two binding regions, -35 binding box and -10 binding box separated by 17 bp on average. The consensus sequence of -35 box is supposed to be **TTGACA** and

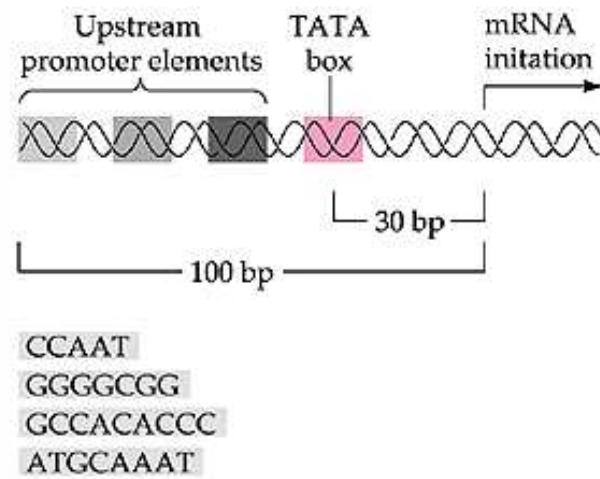


Figure A.2: Eukaryotic promoter-structure [100]

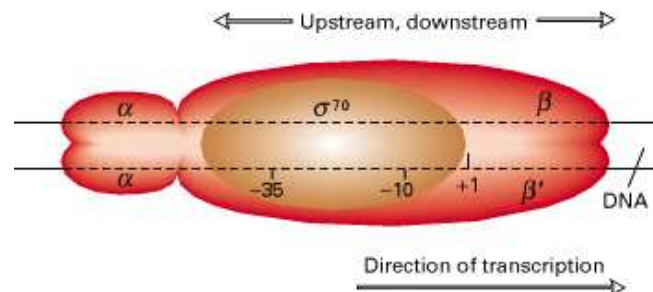


Figure A.3: Transcription process in *E.coli*

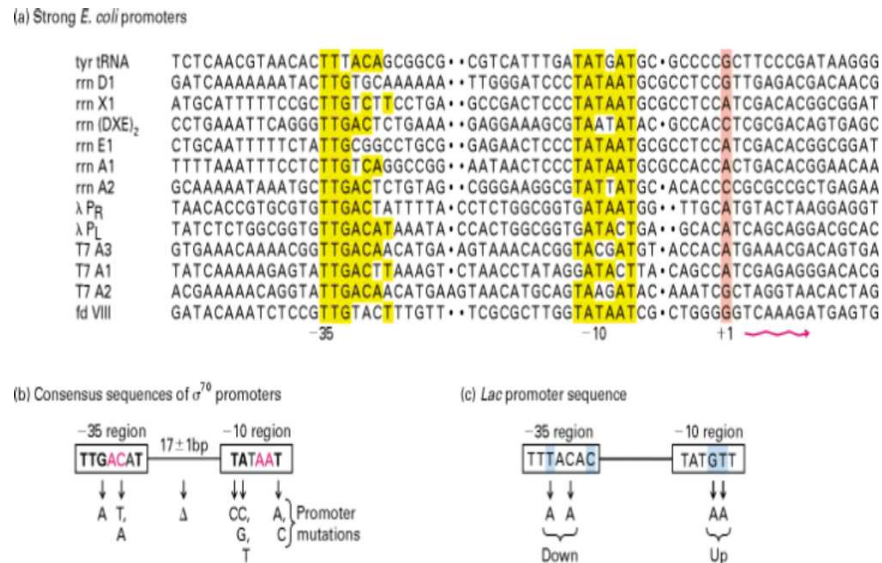
Table A.1: RNA polymerase subunits [78]

Subunit	Size(aa) ¹	Size(Kd) ²	Gene	Function
alpha (b)	329	36511	rpoA	required for assembly of the enzyme; interacts with some regulatory proteins; also involved in catalysis
beta (b)	1342	150616	rpoB	involved in catalysis: chain initiation and elongation
beta (b')	1407	155159	rpoC	binds to the DNA template
sigma (s)	613	70263	rpoD	directs enzyme to the promoter
omega (w)	91	10237	rpoZ	required to restore denatured RNA polymerase in vitro to its fully functional form

the consensus sequence of -10 box is **TATAAT** [41]. This information is shown in Figure A [53]. Mutational studies of *E.coli* promoters have shown that the changes to the -35 region affect the ability of RNA polymerase to bind, whereas changes to the -10 box affect the conversion of the closed promoter complex into the open form. The consensus sequences of *E.coli* are quite variable. These variations, together with less well-defined feature around the transcription start site and in the first 50 or so nucleotides of the transcription start site (TSS), affect the efficiency of the promoter. Efficiency is defined as the number of productive initiations that are promoted per second. A productive initiation is the one that results in the RNA polymerase clearing the promoter and beginning synthesis of a full length transcript. Most efficient promoters (called strong promoters) initiate 1000 times as many productive initiations as the weakest ones. Promoters whose binding regions conform to the consensus sequences are termed as strong promoters.

A.1 Sigma subunit of RNA polymerase

Sigma is a specificity factor. It directs RNA polymerase to the promoter and ensures that transcription is initiated only where it is supposed to be initiated. The very fact that RNA polymerase depends upon a specificity factor to direct



The principal sigma factor in *E.coli* is σ^{70} - so called because the protein is 70 kilo Daltons (kD) in size. The corresponding holoenzyme containing this sigma factor is sometimes abbreviated: Es70. *E.coli* also has six alternative sigma factors that are used in special circumstances. The sigma factors and their activation conditions are tabulated in Table A.2 [78].

Transcription of gene onto a RNA is a three step process. First step in this process is recognition and binding to a promoter. Second is copying the gene code (elongation) and third is termination of the process. Transcription process starts only when RNA polymerase comes and binds to a promoter. For the binding to happen first the promoter is to be identified. Sigma subunit of RNA polymerase can identify the promoter. Then RNA polymerase can start the transcription process.

Table A.2: RNA polymerase sigma subunits [78]

sigma factor	Gene	Function
σ^{70}	rpoD	principal sigma factor
σ^{54}	rpoN (ntrA, glnF)	nitrogen-regulated gene transcription
σ^{32}	rpoH	heat-shock gene transcription
σ^S	rpoS	gene expression in stationary phase cells
σ^F	rpoF	expression of flagellar operons
σ^E	rpoE	involved in heat shock and oxidative stress responses; regulates expression of extracytoplasmic proteins
σ^{FecI}	fecI	regulates the fec genes for iron dicitrate transport

There is a slight variation in this mechanism for prokaryotes and eukaryotes. In prokaryotes only one RNA polymerase exists. It directly binds to the promoter. In eukaryotes, there are three RNA polymerases. And a RNA polymerase cannot directly bind to a promoter. A number of transcriptional factors come and form an initiation complex to which RNA polymerase comes and binds. Initiation complex is formed initially by the binding to TATA binding protein (TBP) to the TATA box. Later, other transcriptional factors such as TFIID, TFIIB, TFIIA, TFIIF and TFIIIE come and assemble at this site of initiation complex. Then RNA polymerase comes and binds to the promoter through this initiation complex to start the transcription process. This process is depicted in Figure A.2.

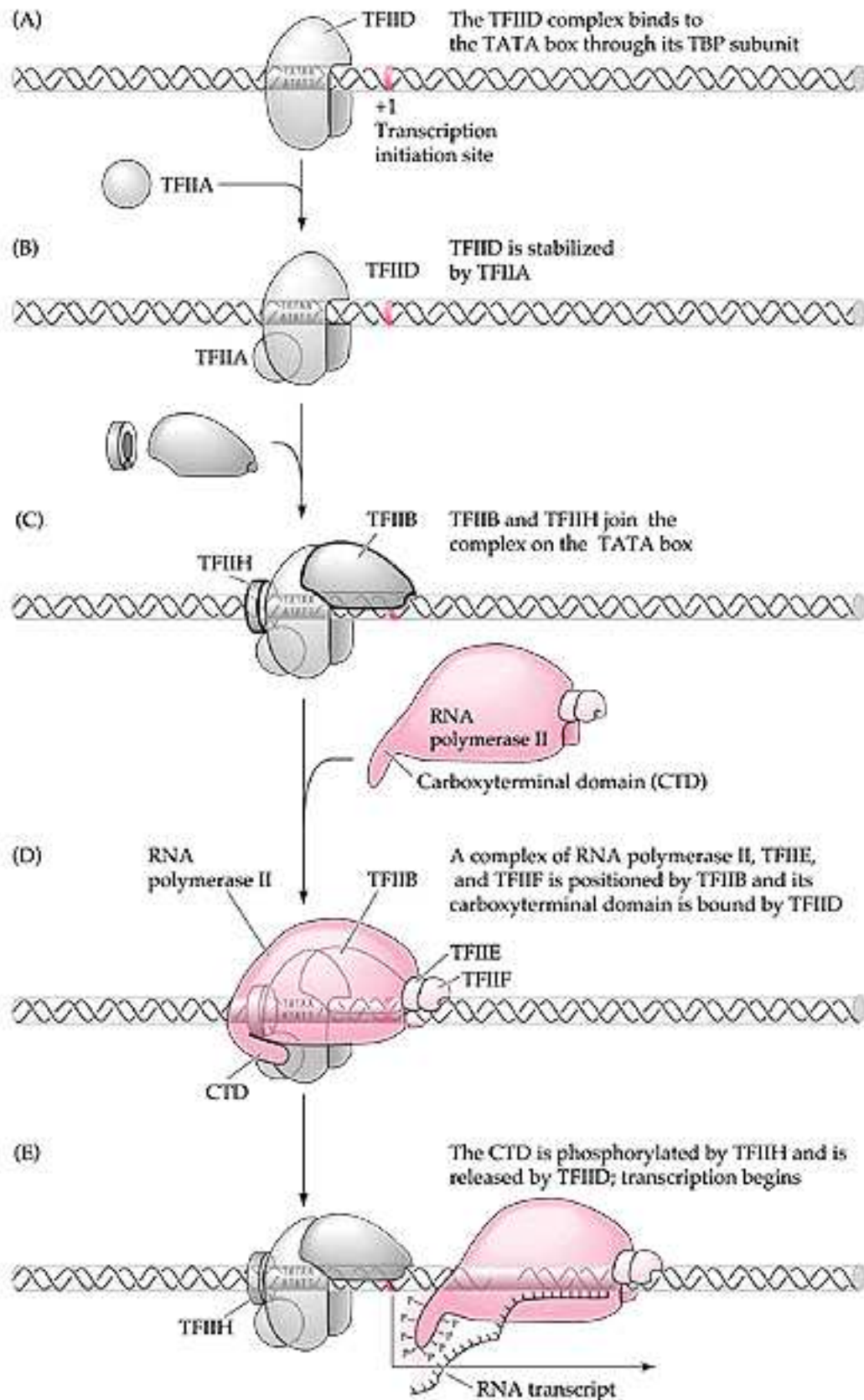


Figure A.5: Binding of RNA polymerase to an eukaryote promoter.

Appendix B

Web sites for Promoter Prediction

Audic - <http://igs-server.cnrs-mrs.fr/audic/selfid.html>

Self-identification of coding regions in microbial genomes.

Dragon Promoter Finder -

<http://research.i2r.a-star.edu.sg/promoter/promoter1.5/DPF.htm>

Eponine Transcription Start Site finder -

<http://servlet.sanger.ac.uk:8080/eponine/>

Eponine is a program for detecting transcription start sites in mammalian genomic DNA sequences.

NNPP - <http://searchlauncher.bcm.tmc.edu/seq-search/gene-search.html>

Prokaryotic promoter prediction at BCM.

Promoter Prediction - http://www.fruitfly.org/seq_tools/promoter.html

Eukaryote and Prokaryote Promoter Search Tool, by Neural Network at LBNL.

PromoterInspector - <http://www.genomatix.de/shop/index.html>

PromoterInspector. Need to register first.

PromoterScan 2 - <http://bimas.dcrt.nih.gov/molbio/proscan/>

Regulatory Sequence Analysis Tools - <http://rsat.ulb.ac.be/rsat/>

Hundreds of species - Regulatory Sequence Elements.

The Markov Chain Promoter Prediction Server -

<http://genes.mit.edu/McPromoter.html>

McPromoter is a program aiming at the exact localization of eukaryotic RNA

polymerase II transcription start sites. This service is free for non-commercial use and restricted to sequences up to 20 kb. The results are e-mailed.

Promoter Prediction by Neural Network -

http://www.fruitfly.org/seq_tools/promoter.html

(Martin Reese, Lawrence Berkeley Laboratory, CA, U.S.A.)

Prokaryotic promoter analysis using SAK -

http://nostradamus.cs.rhul.ac.uk/leo/sak_demo/

Allows analysis of moderately long sequences for prokaryotic Sigma-70 promoters using Sequence Alignment Kernel method. It was tested on a set of 669 known sigma-70 promoters of Escherichia coli. Error rate 16.5% on promoter-positive sequence data and 18.6% on negative data.

PPP - Prokaryotic Promoter Prediction -

<http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp-start.php>

determines from input sequences the putative promoter sequences by using multiple HMM models and presents the results.

FPPROM - <http://www.softberry.com/berry.phtml>

Human promoter prediction.

PATTERN - <http://www.softberry.com/berry.phtml>

pattern search .

TSSP - <http://www.softberry.com/berry.phtml>

Prediction of PLANT Promoters (Using RegSite Plant DB, Softberry Inc.)

TSSG - <http://www.softberry.com/berry.phtml>

Recognition of human PolII promoter region and start of transcription

TSSW - <http://www.softberry.com/berry.phtml>

Recognition of human PolII promoter region and start of transcription (Transfac DB, Biobase GmbH, ONLY for academic use).

NSITE -PL - <http://www.softberry.com/berry.phtml>

Recognition of PLANT Regulatory motifs with statistics (RegSite Plant DB, Softberry Inc.)

NSITEM -PL - <http://www.softberry.com/berry.phtml>

Recognition of PLANT Regulatory motifs conserved in several sequences (RegSite

Plant DB).

NSITE - <http://www.softberry.com/berry.phtml>

Recognition of Regulatory motifs (Transfac DB, Biobase GmbH, ONLY for academic use.)

NSITEM - <http://www.softberry.com/berry.phtml>

Recognition of Conserved Regulatory motifs (Transfac DB, ONLY for academic use).

NSITEH - <http://www.softberry.com/berry.phtml>

Search for functional motifs conserved in a pair of orthologous sequences.

POLYAH - <http://www.softberry.com/berry.phtml>

Recognition of 3' -end cleavage and polyadenylation region.

BPROM - <http://www.softberry.com/berry.phtml>

Prediction of bacterial promoters.

PromH(G) - <http://www.softberry.com/berry.phtml>

Promoter prediction using orthologous sequences in eukaryotic genomes.

PromH(W) - <http://www.softberry.com/berry.phtml>

Promoter prediction using orthologous sequences in eukaryotic genomes (only for academic usage).

CpGFinder - *GC* - <http://www.softberry.com/berry.phtml>

islands finding.

ScanWM - *P* - <http://www.softberry.com/berry.phtml>

Search for weight matrix patterns of plant regulatory sequences.

Motif Explorer - <http://www.softberry.com/berry.phtml>

Motif and promoter visualization.

Appendix C

Sample data set of *E.coli* provided by Dr. Leo Gordon



```
> arcap7
taacgtaagtcgcagaaaaagccctttacttagcttaaaaaaggctaaactatttcctgaCTGTACTAAC
GGTTGAGTTG
> arcap6
ggataattttataaaaaataaatctcgacaattggattcaccacgtttattagttgtatgATGCAACTAG
TTGGATTATT
> arcap5
tttataaaaaataaatctcgacaattggattcaccacgtttattagttgtatgatgcaacTAGTTGGATT
ATTAAAATAA
> arcap4
tgacgaaagctagcatttagatacgatgatttcatcaaactgttaacgtgctacaattgaACTTGATATA
TGTCAACGAA
> arcap3
gatatatgtcaacgaagcgtagttttattgggtgtccggccctcttagcctgttatgttGCTGTTAAAA
TGGTTAGGAT
> arcap2
tttattgggtgtccggccctcttagcctgttatgttgctgttaaaatggtaggatgacAGCCGTTTTT
GACACTGTCG
> arcap1
ttgggtgtccggccctcttagcctgttatgttgctgttaaaatggtaggatgacagccGTTTTTGACA
CTGTCTGGGTC
> smp
taacgcatagaggctaccttgtatccattgcttctggcaacattaagctctcaaattttcaaAGGGTGGAAG
ATGGCTCGCA
> leuq
cacctctgtcgataattaactattgacgaaaagctgaaaaccactagaatgcgcctccgtGGTAGCAATT
CTTTTTAAGA
> feci
acaacatgttaaaaatgtctattggaaacaattttatttccaattgtaatgataaccattCTCATATTAA
TATGACTACG
```

```
> valsp1
tctgcgaacaagctttgcagattttgccaccgctttcacagaagtggtagacttcgttccTTATGAAGA
TTCTCTGAAAC
> valsp2
caagctttgcagattttgccaccgctttcacagaagtggtagacttcgttccttatgaagATTCTCTGAA
ACAACCTGGCG
> argi
gtgacaaagatttatgcttttagacttgcaaatgaataatcatccatataaattgaattttAATTCATTGA
GGCGTTAGCC
> pyrbp1
tgtagccgttcgctttcacactccgccctataagtcggatgaatggaataaaatgcataTCTGATTGCG
TGAAAGTGAA
> pyrbp2
gcgctgacaaaatattgcatcaaattgcttgccgcttctgacgatgagtataatgccggACAATTTGCC
GGGAGGATGT
> trer
aaccaacgataaaccagactttaccattgctgaatgcacgggtaacgttaggctcaaataATTAAACAA
CACGTTACAG
> treb
tcgctgcgtttcggaacgttcccgtttttaattttccgcgcaatatattctgcagccAACCAAAAAT
GTCATCTGCC
> ilerp2
tgcgttaatagatatggaaagcggctcggagagaaaaagcaaaaggtgagggaattacaaaACAGAATGCG
AGTCGTCTCG
> cpdb
tgcgccaactgtgatagtgcatcattttcaaagcgtaaaattgtggcattcttcactgtTCTATAAGTA
AGACGTTTAT
```

Appendix D

Section3 of E.coli genome (NCBI)

Here we give the details about *Section3* of *E.coli* from NCBI. Total genome of *E.coli* is divided into 400 sections. Each of size about 10-13 kbp. This section lists the promoter regions, coding sequences of proteins (cds), information related to coding sequences, repeat regions and the actual genome sequence as well. This particular section genome is used in Chapter 3 to test the efficacy of n-gram method to determine the promoters.

[PubMed](#)
[Nucleotide](#)
[Protein](#)
[Genome](#)
[Structure](#)
[PMC](#)
[Taxonomy](#)
[OMIM](#)
[Books](#)

Search for

[Limits](#)
[Preview/Index](#)
[History](#)
[Clipboard](#)
[Details](#)

Display Show Hide: ☐ sequence ☐ all but gene, CDS and mRNA features

Range: from to ☐ Reverse complemented strand Features:

1: [AE000113](#). Reports ...[gi:2367095] The record has been replaced by [U00096](#)

[Comment](#) [Features](#) [Sequence](#)

LOCUS AE000113 13485 bp DNA linear BCT 01-DEC-2000
 DEFINITION Escherichia coli K12 MG1655 section 3 of 400 of the complete genome.
 ACCESSION AE000113 U00096
 VERSION AE000113.1 GI:2367095
 KEYWORDS .
 SOURCE Escherichia coli K12
 ORGANISM [Escherichia coli K12](#)
 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.
 REFERENCE 1 (bases 1 to 13485)
 AUTHORS Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y.
 TITLE The complete genome sequence of Escherichia coli K-12
 JOURNAL Science 277 (5331), 1453-1474 (1997)
 PUBMED [9278503](#)
 REFERENCE 2 (bases 1 to 13485)
 AUTHORS Blattner,F.R.
 TITLE Direct Submission
 JOURNAL Submitted (16-JAN-1997) Guy Plunkett III, Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA. Email: ecoli@genetics.wisc.edu Phone: 608-262-2534 Fax: 608-263-7459
 REFERENCE 3 (bases 1 to 13485)
 AUTHORS Blattner,F.R.
 TITLE Direct Submission
 JOURNAL Submitted (02-SEP-1997) Guy Plunkett III, Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA. Email: ecoli@genetics.wisc.edu Phone: 608-262-2534 Fax: 608-263-7459
 REFERENCE 4 (bases 1 to 13485)
 AUTHORS Plunkett,G. III.
 TITLE Direct Submission
 JOURNAL Submitted (13-OCT-1998) Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA
 COMMENT On Sep 9, 1997 this sequence version replaced gi:[1786205](#). This sequence was determined by the E. coli Genome Project at the University of Wisconsin-Madison (Frederick R. Blattner, director). Supported by NIH grants HG00301 and HG01428 (from the Human Genome Project and NCHGR). The entire sequence was independently determined from E. coli K12 strain MG1655. Predicted open reading frames were determined using GeneMark software, kindly supplied by Mark Borodovsky, Georgia Institute of Technology, Atlanta, GA, 30332 [e-mail: mark@amherstgatech.edu]. Open reading frames that have been correlated with genetic loci are being annotated with CG Site Nos., unique ID nos. for the genes in the E. coli Genetic Stock Center (CGSC) database at Yale University, kindly supplied by Mary Berlyn. A public version of the database is accessible (<http://cgsc.biology.yale.edu>). Annotation of the genome is an ongoing task whose goal is to make the genome sequence more useful by correlating it with other data. Comments to the authors are appreciated. Updated information will be available at the E. coli Genome Project's World Wide Web site (<http://www.genetics.wisc.edu>). *** The E. coli K12 sequence and its annotations are periodically updated; this is version M54. No sequence changes. Annotation updates: updated gene identifications and products; all new functional assignments courtesy of Monica Riley; added promoters, protein binding sites, and repeated sequences described in reference 1. The unique numeric identifiers beginning with a lowercase 'b' assigned to each gene (protein- or RNA-encoding) are now designated as gene synonyms instead of labels. This should allow them to be searched for in Entrez as gene

	names.	Location/Qualifiers
FEATURES		
source		1..13485 /organism="Escherichia coli K12" /mol_type="genomic DNA" /strain="K12" /sub_strain="MG1655" /db_xref="taxon:83333"
<u>gene</u>		complement(156..419) /gene="rpsT"
<u>CDS</u>		/note="synonym: b0023" complement(156..419) /gene="rpsT" /function="structural component; Ribosomal proteins - synthesis, modification" /note="f87; 100 pct identical RS20_ECOLI SW: P02378 but includes initiator met; TTG start" /codon_start=1 /transl_table=11 /product="30S ribosomal subunit protein S20" /protein_id="AAC73134.1" /db_xref="GI:1786206" /translation="MANIKSAKKRAIQSEKARKHNASRRSMMRTFIKKVYAAIEAGDK AAQKAFNEMQPIVDRQAAGLIHKNAARHKANLTAQINKLA"
<u>protein_bind</u>		complement(442..453) /note="central position to predicted promoter:23" /bound_moiety="AraC predicted site"
<u>protein_bind</u>		443..454 /note="central position to predicted promoter: -40" /bound_moiety="AraC predicted site"
<u>protein_bind</u>		443..454 /note="No predicted promoter" /bound_moiety="AraC predicted site"
<u>promoter</u>		452..483 /note="factor Sigma70; predicted +1 start at 21149"
<u>promoter</u>		complement(467..495) /note="factor Sigma70; predicted +1 start at 21119"
<u>gene</u>		522..740 /gene="b0024"
<u>CDS</u>		522..740 /gene="b0024" /function="orf; Unknown" /note="o72; 43 pct identical (1 gap) to 32 residues from cation-transporting P-type ATPase A, CTPA_MYCLE SW: P46839 (780 aa)" /codon_start=1 /transl_table=11 /product="orf, hypothetical protein" /protein_id="AAC73135.1" /db_xref="GI:1786207" /translation="MCRHSLRSDGAGFYQLAGCEYSFSAIKIAAGGQFLPVICAMAMK SHFFLISVLNRRLLTAVQGILGRFSLF"
<u>gene</u>		748..1689 /gene="ribF"
<u>CDS</u>		/note="synonym: b0025" 748..1689 /gene="ribF" /function="putative regulator; Not classified" /note="o312; formerly designated yaaC" /codon_start=1 /transl_table=11 /product="putative regulator" /protein_id="AAC73136.1" /db_xref="GI:1786208" /translation="MKLIRGIHNLSQAPQEGCVLTIGNFDGVHRGHRALLQGLQEEGR KRNLPMVMVLFEPQPLELATDKAPARLTRLEKRLRYLAECGVYVLCVRFDRRFAAL TAQNFISDLLVKHLRVKFLAVGDDFRFGAGREGDFLLQKAGMEYGFDTSTQTFCEG GVRISSTAVRQALADDNLALAESLLGHPFAISGRVVHGDDELGRITIGFPTANVPLRRQV SPVKGVYAVEVLGLGEKPLPGVANIGTRPTVAGIRQQLEVHLLDVAMDLYGRHIQVVL RKKIRNEQRFASLDELKAQIARDELTAFFGLTKPA"
<u>promoter</u>		1534..1564 /gene="ribF" /note="factor Sigma70; predicted +1 start at 22230"
<u>gene</u>		1732..4548 /gene="ileS"
<u>CDS</u>		/note="synonym: b0026" 1732..4548 /gene="ileS" /EC_number="6.1.1.5" /function="enzyme; Aminoacyl tRNA synthetases, tRNA modification" /note="o938; 100 pct identical to SYI_ECOLI SW: P00956; alternate gene name ilvS" /codon_start=1 /transl_table=11

```

/product="isoleucine tRNA synthetase"
/protein_id="AAC73137.1"
/db_xref="GI:2367096"
/translation="MSDYKSTLNLPEGTGFPMRGDLAKREPGMLARWTDLDDLYGIRAA
KKGKKTFFILHDGPPYANGSIHIGHSVKNILKDIIVKSKGLSGYDSPYVPGWDCGHLPI
ELKVEQEYKGKPEKFTAAEFRAKCREYAATQVDGQRKDFIRLGVLDGWSHPYLTMDFK
TEANIIRALGKIIIGNHGLHKGAKPVHWCDCRSALAEAEVEYYDKTSPSIDVAFQAVD
QDALKAKFAVSNVNGPISLVIWTTTPWTLNRAISIAPDFDYALVQIDGQAVILAKD
LVESVMQRIGVTDYITLGTVKGAELELLRFTHPFMGFDVPAILGDHVTLDAGTGAHT
APGHGPDYVIGQKYGLETANVPGPDGTYLPGTYPTLDGVNVFKANDIVALLQEKGA
LLHVEKMQHSYPCCWRHKTPIIFRATPQWFVSMQKGLRAQSLKEIKGVQWIPDWGQA
RIESMVANRPDWCISRQRTWGVPMSLFVHKDTEELHPRTLELMEEVAKRVEVDGIQAW
WDLDAKEILGDEADQYVKVPDLDVWFDGSGTHSSVVDVRPEFAGHAADMYLEGSQDQ
RGWFMSSLMISTAMKKGKAPYRQVLTHGFTVDGQGRKMSKSI GNTVSPQDVMNKLAD I
LRLWVASTDYTGEMAVSDEILKRAADSYRRIRNTRARFLANLNGFDPKAKDMVKPEEMV
VLDRWAVGCAKAAQEDILKAYEAYDFHEVVQRLMRFCSVEMGSFYLDI IKDRQYTKA
DSVARRSCQTALYHIAEALVRWMAPILSFTADEVWGYLPGEREKYVFTGEWYEGFLGL
ADSEAMNDAFWDELKVRGEVNKVIEQARADKKVGGSEAAVTLTLYAEPELSAKLTALG
DELRFLVLLTSGATVADYNDAPADAQQSEVLKGLKVALSKAEGEKCPRCWHYTQDVGVK
AEHAEICGRCVSNVAGDGKKRFA"
gene 4548..5042
      /gene="lspA"
      /note="synonym: b0027"
CDS 4548..5042
     /gene="lspA"
     /EC_number="3.4.23.36"
     /function="enzyme; Protein, peptide secretion"
     /note="o164; 100 pct identical to LSPA_ECOLI SW: P00804"
     /codon_start=1
     /transl_table=11
     /product="prolipoprotein signal peptidase (SPase II)"
     /protein_id="AAC73138.1"
     /db_xref="GI:1786210"
     /translation="MSQSICTGLRWLWLVVVLIIDLGSKYLILQNFALGDTVPLFP
SLNLHYARNYGAAFSLADSGGWQRFWFFAGIAIGISVILAVMMYRSKATQKLNNIAYA
LIIGGALGNLFDRLWHGFVVDIMDFYVGDWHFATFNLDATAICVGAALIVLEGFLPSR
AKKQ"
promoter 5003..5020
         /gene="lspA"
         /note="factor Sigma54; predicted +1 start at 25686"
promoter 5014..5041
         /gene="lspA"
         /note="factor Sigma70; predicted +1 start at 25707"
repeat_region 5051..5136
              /note="REP (repetitive extragenic palindromic) element;
              contains 2 REP sequences"
gene 5167..5616
     /gene="slpA"
     /note="synonym: b0028"
CDS 5167..5616
     /gene="slpA"
     /function="putative enzyme; Proteins - translation and
     modification"
     /note="o149; 100 pct identical to FKBX_ECOLI SW: P22563"
     /codon_start=1
     /transl_table=11
     /product="probable FKBX-type 16KD peptidyl-prolyl
     cis-trans isomerase (a rotamase)"
     /protein_id="AAC73139.1"
     /db_xref="GI:1786211"
     /translation="MSESVQNSAVLVHFTLKLDDGTTAESTRNNGKPAFLRGDASL
     SEGLEQHLLGLKVGDKTTFSLPEDAFVGPSPDLIQYFSRREFMDAGEPEIGAIMLFT
     AMDGSEMPGVIREINGDSITVDFNHLPLAGQTVHFDIEVLEIDPALEA"
gene 5618..6568
     /gene="lytB"
     /note="synonym: b0029"
CDS 5618..6568
     /gene="lytB"
     /function="regulator; Global regulatory functions"
     /note="o316; 100 pct identical to LYTB_ECOLI SW: P22565"
     /codon_start=1
     /transl_table=11
     /product="control of stringent response; involved in
     penicillin tolerance"
     /protein_id="AAC73140.1"
     /db_xref="GI:1786212"
     /translation="MQILLANPRGFCAGVDRAISIVENALAIYGAPIYVRHEVVHNR
     VVDSLRERGAIFIEQISEVPDGAIIIFSAHGVSQAVRNEAKSRDLTVFDATCPLVTKV
     HMEVARASRRGEESILIGHAGHPEVEGTMGQYSNPEGGMYLVEFPDDVWKLTVKNEEK
     LSFMTQTTLSDVDDTSDVIDALRKRFPKIVGPRKDDICYATTNRQEAVALAEQAEVVL
     VVGSKNSSNRLAELAQRMGKRAFLIDDAKDIQEEWVKEVKCVGVTAGASAPDILVQ
     NVVARLQQLGGGEAIPLEGREENIVFEVPKELRVDIREVD"
promoter 6526..6552
         /gene="lytB"
         /note="factor Sigma70; predicted +1 start at 27218"

```

promoter 6553..6580
/note="factor Sigma70; predicted +1 start at 27246"

promoter 6569..6597
/note="factor Sigma70; predicted +1 start at 27263"

promoter 6595..6623
/note="factor Sigma70; predicted +1 start at 27289"

gene 6634..7548
/gene="yaaF"
/note="synonym: b0030"

CDS 6634..7548
/gene="yaaF"
/function="orf; Unknown"
/note="o304; 100 pct identical to YAAF_ECOLI SW: P22564"
/codon start=1
/transl table=11
/product="orf, hypothetical protein"
/protein_id="AAC73141.1"
/db_xref="GI:1786213"
/translation="MRLPIFLDTPGIDDAVAIAAIFAPELDLQLMTTVAGNVSVET
TTRNALQLLHFWNAEIPLAQGAAPLVRAPRDAASVHGESGMAGYDFVEHNRKPLGIP
AFLAIRDALMRAPEPVTVAIGPLTNIALLLSQCECKPYIRRLVIMGGSAGRGNCCTP
NAEFNIAADPEAAACVFRSGIEIVMGLDVTNQAILTPDYLSTLPQLNRTGKMLHALF
SHYRSGSMQSGLRMHDLCAIAWLVRPDLFTLKPCFVAVETQGEFTSGTTVVDIDGCLG
KPANVQVALDLDVKGFGQWVAEVLALAS"

promoter 7644..7676
/note="factor Sigma70; promoter dapB; documented +1
at28343"

gene 7715..8536
/gene="dapB"
/note="synonym: b0031"

CDS 7715..8536
/gene="dapB"
/EC_number="1.3.1.26"
/function="enzyme; Amino acid biosynthesis: Lysine"
/note="o273; 100 pct identical to DAPB_ECOLI SW: P04036"
/codon start=1
/transl table=11
/product="dihydrodipicolinate reductase"
/protein_id="AAC73142.1"
/db_xref="GI:1786214"
/translation="MHDANIRVAIAGAGGRMGRLIQAAALALEGVOLGAALEREGLSS
LGSDAGELAGAGKTGVTVQSSLDVAKDDFDVFDTRPEGLNLHLAFRCRHGKGMVIG
TTGFDEAGKQAIRDAADIAIVFAANFSGVNVMLKLEKAAKVMGDDYTDIEIEAHH
RHKVDAPSGTALAMGEAIAHALDKDLKDCAVYSREHGTGERVPGTIGFATVRAGDLVG
EHTAMFADIGERLEITHKASSRMTFANGAVRSALWLSGKESGLFDMRDVLDLNNL"

promoter 8856..8884
/note="factor Sigma70; promoter carABp1; documented +1 at
29551"

promoter 8925..8954
/note="factor Sigma70; promoter carABp2; documented +1 at
29619"

protein_bind 8939..8961
/note="central position to predicted promoter:58.5;genetic
evidence for the site"
/bound moiety="ArgR predicted site"

protein_bind 8960..8982
/note="central position to predicted promoter:79.5;genetic
evidence for the site"
/bound moiety="ArgR predicted site"

gene 8992..10140
/gene="carA"
/note="synonym: b0032"

CDS 8992..10140
/gene="carA"
/EC_number="6.3.5.5"
/function="enzyme; Pyrimidine ribonucleotide biosynthesis"
/note="o382; 100 pct identical to CARA_ECOLI SW:
P00907;TTG start"
/codon start=1
/transl table=11
/product="carbamoyl-phosphate synthetase, glutamine
(small) subunit"
/protein_id="AAC73143.1"
/db_xref="GI:1786215"
/translation="MIKSALLVLEDGTQFHGRAIGATGSAGVGVFNTSMTGYQEILT
DPSYSRQIVTLTYPHIGNVTNDADESSQVHAQGLVIRDPLIASNFRNTEDLSSYL
KRHNIVAIADIDTRKLTRLREKGAQNGCIIAGDNPDAALALEKARAFPLNGMDLAK
EVTTAESAEDVLTQGSWLTGGLPEAKKEDELPHVVAIDFGAKRNILRMLVDRGCRLLTI
VPAQTSADVDLKMNDGIFLSNGPGDPAPCDYAITAIQKFLTDIPVFGICLGHQLLA
LASGAKTVKMKFGHHGGNHPVKDVEKNVVMITAQNHGFVDEATLPANLRVTHKSLFD
GTLQGIHRTDKPAFSFQGHPEASPGPHDAAPLFDHFIELIEQYRKTAK"

promoter 10075..10105
/gene="carA"
/note="factor Sigma70; promoter carB; documented +1
at30775"

NCBI Sequence Viewer v2.0

<http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=...>

```

gene      10158..13379
          /gene="carB"
          /note="synonym: b0033"
CDS       10158..13379
          /gene="carB"
          /EC_number="6.3.5.5"
          /function="enzyme; Pyrimidine ribonucleotide biosynthesis"
          /note="o1073; 100 pct identical to CARB_ECOLI SW: P00968"
          /codon_start=1
          /transl_table=11
          /product="carbamoyl-phosphate synthase large subunit"
          /protein_id="AAC73144.1"
          /db_xref="GI:1786216"
          /translation="MPKRTDIKSILILGAGPIVIGQACEFDYSGAQACKALREEGYRV
ILVNSNPATIMTDPEMADATYIEPIHWEVVVKIIEKERPDVLPMTGGQTALNCALEL
ERQGVLEEFVGTMGATADAIKDAEDRRRFDVAMKKIGLETARSGIAHTMEEALAVAA
DVGFPCIIRPSFTMGSGGGIAYNREEFEEICARGLDLSPTKELIDESLIGWKEYEM
EVVRDKNDCNIIIVCSINFDAMGIHTGDSITVAPAQTLTDKEYQIMRNASMAVLRIG
VETGGSNVQFAVNPKNRGLIVIEMNPRVSRSSALASKATGFPKAKVAAKLAVGYTLDE
LMNDITGGRTPASFEPSIDYVVTIKIPRFNFEKAGANDRLTTQMKSVGEVMAIGRTQQ
ESLQKALRGLEVGTGTFDPKVSLDDPEALTKIRRELKADAGADRIWYIADAFRAGLSVD
GVFNLTNIDRWFLVQIEELVRLLEKVAEVGITGLNADFLRLKRGKGFADARLAKLAGV
REAEIRKLQDQYDLHPVYKRVDTCAAEFATDTAYMYSTYEECEANPSTDREKIMVLG
GGPNRIGQGIEFDYCCVHASLALREDGYETIMVNCNPETVSTDYDTSDRLYFEPVTL
DVLEIVRIEKPGVIVQYGGQTPKLARALEAAGVPVIGTSPDAIDRAEDRERFQHAV
ERLKLQPANATVTAIEMAVEKAKEIGYPLVVRPSYVLGGRAMEIVYDEADLRRYQT
AVSVDNDAPVLDHFLDDAVEVDVAICDGMVLIGGIMEHIEQAGVHSGDSACSPLA
YTLSQEIQDVMRQOVQKLAFAELQVRGLMNVQFAVKNNEVYLIEVNPRAARTVPFVSKA
TGVPLAKVAARVMAGKSLAEQGVTKVEIPPYYSVKEVVLFPNKFPGVDPLLGPEMRST
GEVMGVGRTFAEAFKAQLGSNSNMTKKHGRALLSVREGDKERVVDLAAKLLKQGFELD
ATHGTAIVLGEAGINPRLVNKVHEGRPHIQDRIKNGEYTYIINTSGRRRAIEDSRVIR
RSALQYKHVYDITLNGGFATAMALNADATEKVISVQEMHAQIK"
          13370..13401
          /note="factor Sigma70; predicted +1 start at 34067"

promoter

ORIGIN
1  gattcttaag ccacgaagag ttcagatagt acaacggcat gtctcttttg actatctggc
61  aaccggcagt gtgttctctc acgcatcaca aaagcagcag gcataaaaaa acccgcttgc
121 gcgggctttt tcacaaagct tcagcaaaatt ggcgattaa ggcagtttgt gatctgtgca
181 gtcaggttag ccttatgacg tgcagctttg tttttgtgga tcagaccttt agcagcctga
241 cggtccacga tcggttgcat ttcgttaaat gctttctgtg cagcagcttt gtcgccagct
301 tcgatatagtg cgtatacttt cttgatgaaa gtacgcacat tagagcgacg gcttgcgttg
361 tgcttacgag ccttttcaga ctgaatggcg cgtctcttag ctgatttgat attagccaag
421 gtccaaactcc caaatgtgtt ctatatggac aattcaaaag ccgaggaata tgcctcttta
481 gccttctttt gtcaatggat ttgtgcaaat aagcgccgtt aatgtgccgg cactcgttac
541 gtagtgtatg cgcaggatgc taccagcttg cggggtgtga atacagcttt tccgcgataa
601 aaattgcagc aggcggtcag tttcttcccg tgatttgcgc catggcgaatg aaaagccact
661 tctttctgat ttcggtactc aatcgccggt taaccttgac cgcgtacaaa ggtatactcg
721 gacgattttc actgttttga gccagacatg aagctgatac gcggcataca taactctgac
781 caggcccccgc aagaagggtg tgtgctgact attgtaatt tcgacggcgt gcactcgcgt
841 catcgccgcg tgttacaggg cttgcaggaa gaaggcgcca agcgcaactt accgctgatg
901 gtgatgcttt ttgaacctca accactggaa ctgtttgcta ccgataaaag cccggcaaga
961 ctgacccggc tgcgggaaaa actgcgttac cttgcagagt gtggcgttga ttactgtctg
1021 tgcgtgcgtt tcgacaggcg tttcgcggcg ttaaccggcg aaatttcat cagcgatctt
1081 ctggtgaagc atttgcgcgt aaaatttctt gccgtaggtg atgatttccg ctttggcgct
1141 ggtcgtgaag cgcatttctt gttattacag aaagctggca tggaaatcag cttcgatatac
1201 accagtcacg aaactttttg cgaagggtgc gtgcgcacat gcagcaccgc cgtgcgtcag
1261 gcccttgcgc atgacaatct ggctctggca gagagtttac tggggcaccc gtttgccatc
1321 tccggcgctg tagtccacgg tgatgaatta gggcgacata taggtttccc gacggcgaat
1381 gtaccgctgc gccgtcaggt ttcccgggtg aaagggttt atcggtaga agtgctgggc
1441 ctcggtgaaa agccgttacc cggcgtggca aacatcgaaa caccgccaac ggttgcggt
1501 attcgccagc agctggaagt gcatttgtta gatgttgcaa tggaccttta cgttcgccat
1561 atacaagtag tgctgcgtaa aaaaatacgc aatgagcagc gatttgcgtc gctggacgaa
1621 ctgaaagcgc agattgcgcg tgatgaatta accgcccgcg aattttttgg gatcaaaaaa
1681 ccggcttaag cctgttatgt aatcaaacgc aaatcaggaa ccgagaatct gatgagtac
1741 tataaatcaa cctgaattt gccggaaaaa gggttccga tcggtggcga tctcgccaag
1801 cgcgaacccg gaatgctggc gcgttggact gatgatgac tgcacggcat catccgttgc
1861 gctaaaaaag gcaaaaaaac cttcattctg catgatggcc ctcttatgc gaatggcagc
1921 attcatattg gtcaactcgt taacaagatt ctgaaagaca ttatcgtgaa gtccaaaggg
1981 ctttccggtt atgactcgcc gtatgtgcct ggctgggact gccacggtct gccgatcgag
2041 ctgaaagtcg agcaagaata cggtaagccg ggtgagaaat tcaccgcccg cgagttcccg
2101 gccaaagtgc gcgaatacgc ggcgaccagc gttgacggtc aacgcaaaag ctttatccgt
2161 ctggcgctgc tgggcgactg gtcgcacccg tacctgacca tggacttcaa aactgaagcc
2221 aacatcatcc gcgcgctggg caaatcatc ggcaacggtc acctgcacaa aggcgcgaag
2281 ccagttcact ggtgcgttga ctgcgcttct gcgctggcgg aagcggaagt tgagtattac
2341 gacaaaactt ctccgtccat cgacgttgct ttccaggcag tcgatcagga tgcactgaaa
2401 gcaaaatttg ccgtaagcaa cgttaacggc ccaatctcgc ttgtaattct gaccaccagc
2461 ccgtggactc tgcctgccaa ccgcgcaatc tctattgcac cagatttcca ctatgcgctg
2521 gtgcagatcg acggtcaggc cgtgattctg gcgaagaatc ttggtgaaag cgttaatgac
2581 cgtatcggcg taccgatta caccattctc ggcacggtaa aaggtgcgga gcttgagctg
2641 ctgcgcttta cccatccgtt tatgggcttc gacgttccgg caatcctcgg cgatcacgtt
2701 accctggatg ccggtaccgg tgcggttcac accgcgctg gccacggccc gcacgactat
2761 gtgatcggtc agaaatcagg cctggaaacc gctaaccggg ttggcccgga cgacacttat
2821 ctgccgggca cttatccgac gctggatggc gtgaacgtct tcaaaagcaa gcacatcgtc
2881 gttgcgctgc tgcaggaaaa aggcgcgctg ctgcacgttg aaaaaatgca gcacagctat
2941 ccgtgctgct ggcgtcacia aacgccgac atcttccgcg cgacgccgca gtggttcgtc

```


Appendix E

Glossary

Glossary taken from website <http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/mbglossary/mbgloss.html> and "Genomes": Second edition by T.A.Brown (Bios Scientific Publishers Limited 2002)

3' end/5' end: A nucleic acid strand is inherently directional, and the "5 prime end" has a free hydroxyl (or phosphate) on a 5' carbon and the "3 prime end" has a free hydroxyl (or phosphate) on a 3' carbon (carbon atoms in the sugar ring are numbered from 1' to 5'). That's simple enough for an RNA strand or for single-stranded (ss) DNA. However, for double-stranded (ds) DNA it's not so obvious - each strand has a 5' end and a 3' end, and the 5' end of one strand is paired with the 3' end of the other strand (it is "antiparallel"). One would talk about the 5' end of ds DNA only if there was some reason to emphasize one strand over the other - for example if one strand is the sense strand of a gene. In that case, the orientation of the sense strand establishes the direction.

5' flanking region: A region of DNA which is not transcribed into RNA, but rather is adjacent to 5' end of the gene. The 5'-flanking region contains the promoter, and may also contain enhancers or other protein binding sites.

Base pair: The hydrogen-bonded structure formed by two complementary nucleotides. When abbreviated to 'bp', the shortest unit of length for a double stranded DNA molecule.

Binding site: A place on cellular DNA to which a protein (such as a transcription factor) can bind. Typically, binding sites might be found in the vicinity of genes, and would be involved in activating transcription of that gene (promoter elements), in enhancing the transcription of that gene (enhancer elements), or in reducing the transcription of that gene (silencers). Note that whether the protein in fact performs these functions may depend on some condition, such as the pres-

ence of a hormone, or the tissue in which the gene is being examined. Binding sites could also be involved in the regulation of chromosome structure or of DNA replication.

Closed promoter complex: The structure formed during the initial step in assembly of the transcription initiation complex. The closed promoter complex consists of the RNA polymerase and /or accessory proteins attached to the promoter, before the DNA has been opened up by breaking of base pairs.

Coding sequence: The portion of a gene or an mRNA which actually codes for a protein. Introns are not coding sequences; nor are the 5' or 3' untranslated regions (or the flanking regions, for that matter - they are not even transcribed into mRNA). The coding sequence in a cDNA or mature mRNA includes everything from the AUG (or ATG) initiation codon through to the stop codon, inclusive.

Coding strand: an ambiguous term intended to refer to one specific strand in a double-stranded gene.

Codon: In an mRNA, a codon is a sequence of three nucleotides which codes for the incorporation of a specific amino acid into the growing protein. The sequence of codons in the mRNA unambiguously defines the primary structure of the final protein. Of course, the codons in the mRNA were also present in the genomic DNA, but the sequence may be interrupted by introns.

Codon bias: Refers to the fact that not all codons are used equally frequently in the genes of a particular organism.

Consensus sequence: A nominal sequence inferred from multiple, imperfect examples. Multiple lanes of shotgun sequence can be merged to show a consensus sequence. The optimal sequence of nucleotides recognized by some factor. A DNA binding site for a protein may vary substantially, but one can infer the consensus sequence for the binding site by comparing numerous examples. For example, the (fictitious) transcription factor ZQ1 usually binds to the sequences AAAGTT, AAGGTT or AAGATT. The consensus sequence for that factor is said to be AARRTT (where R is any purine, i.e. A or G). ZQ1 may also be able to weakly bind to ACAGTT (which differs by one base from the consensus).

CpG island: A GC-rich DNA region located upstream of approximately 56% of the genes in the human genome.

DNase: Deoxyribonuclease, a class of enzymes which digest DNA. The most common is DNase I, an endonuclease which digests both single and double-stranded DNA.

E. coli: A common Gram-negative bacterium useful for cloning experiments. Present in human intestinal tract. Hundreds of strains of E. coli exist. One strain, K-12, has been completely sequenced.

Enhancer: A regulatory sequence that increases the rate of transcription of a gene or genes located some distance away in either direction.

Exon: A coding region within a discontinuous gene.

Footprinting: A technique by which one identifies a protein binding site on cellular DNA. The presence of a bound protein prevents DNase from "nicking" that region, which can be detected by an appropriately designed gel.

Gene: A DNA segment containing biological information and hence coding for an RNA and/or polypeptide molecule. A gene contains coding regions, introns, untranslated regions and control regions.

Genome: The total DNA contained in each cell of an organism. Mammalian genomic DNA (including that of humans) contains 6×10^9 base pairs of DNA per diploid cell. There are somewhere in the order of a hundred thousand genes, including coding regions, 5' and 3' untranslated regions, introns, 5' and 3' flanking DNA. Also present in the genome are structural segments such as telomeric and centromeric DNAs and replication origins, and intergenic DNA.

Helix-loop-helix: A protein structural motif characteristic of certain DNA-binding proteins.

Homology searching: A technique in which genes with sequences similar to that of an unknown genes are sought, the objective being to gain an insight into the function of the unknown gene.

Intergenic: Between two genes; e.g. intergenic DNA is the DNA found between two genes. The term is often used to mean non-functional DNA (or at least DNA with no known importance to the two genes flanking it). Alternatively, one might speak of the "intergenic distance" between two genes as the number of base pairs from the polyA site of the first gene to the cap site of the second. This usage might therefore include the promoter region of the second gene.

Intergenic region: The regions of a genome that do not contain genes.

Intron: A non-coding region within a discontinuous gene. Introns are portions of genomic DNA which are transcribed (and thus present in the primary transcript) but which are later spliced out. They thus are not present in the mature mRNA. Note that although the 3' flanking region is often transcribed, it is removed by endonucleolytic cleavage and not by splicing. It is not an intron.

Kilobase pair (kb): one thousand base pairs.

Leucine zipper: A motif found in certain proteins in which Leu residues are evenly spaced through an α -helical region, such that they would end up on the same face of the helix. Dimers can form between two such proteins. The Leu zipper is important in the function of transcription factors such as Fos and Jun and related proteins.

mRNA: "messenger RNA" or sometimes just "message"; an RNA which contains sequences coding for a protein. The term mRNA is used only for a mature transcript with polyA tail and with all introns removed, rather than the primary transcript in the nucleus. As such, an mRNA will have a 5' untranslated region, a coding region, a 3' untranslated region and (almost always) a poly(A) tail. Typically about 2% of the total cellular RNA is mRNA.

Oligonucleotide: A short synthetic single-stranded DNA molecule.

Operon: A set of adjacent genes in a bacterial genome, transcribed from a single promoter and subject to the same regulatory regime.

Polymerase: An enzyme which links individual nucleotides together into a long strand, using another strand as a template. There are two general types of polymerase: DNA polymerases (which synthesize DNA) and RNA polymerase (which makes RNA). Within these two classes, there are numerous sub-types of polymerase, depending on what type of nucleic acid can function as template and what type of nucleic acid is formed. A DNA-dependant DNA polymerase will copy one DNA strand starting from a primer, and the product will be the complementary DNA strand. A DNA-dependant RNA polymerase will use DNA as a template to synthesize an RNA strand.

Post-transcriptional regulation: Any process occurring after transcription which affects the amount of protein a gene produces. Includes RNA processing efficiency, RNA stability, translation efficiency, protein stability. For example,

the rapid degradation of an mRNA will reduce the amount of protein arising from it. Increasing the rate at which an mRNA is translated will increase the amount of protein product.

Post-translational processing: The reactions which alter a protein's covalent structure, such as phosphorylation, glycosylation or proteolytic cleavage.

Post-translational regulation: Any process which affects the amount of protein produced from a gene, and which occurs AFTER translation in the grand scheme of genetic expression. Actually, this is often just a buzz-word for regulation of the stability of the protein. The more stable a protein is, the more it will accumulate.

Primary transcript: When a gene is transcribed in the nucleus, the initial product is the primary transcript, an RNA containing copies of all exons and introns. This primary transcript is then processed by the cell to remove the introns, to cleave off unwanted 3' sequence, and to polyadenylate the 5' end. The mature message thus formed is then exported to the cytoplasm for translation.

Promoter: The first few hundred nucleotides of DNA "upstream" (on the 5' side) of a gene, which control the transcription of that gene. The promoter is part of the 5' flanking DNA, i.e. it is not transcribed into RNA, but without the promoter, the gene is not functional. Note that the definition is a bit hazy as far as the size of the region encompassed, but the "promoter" of a gene starts with the nucleotide immediately upstream from the cap site, and includes binding sites for one or more transcription factors which can not work if moved farther away from the gene.

Sequence: As a noun, the sequence of a DNA is a buzz word for the structure of a DNA molecule, in terms of the sequence of bases it contains. As a verb, "to sequence" is to determine the structure of a piece of DNA; i.e. the sequence of nucleotides it contains.

Splicing: Removal of introns from the primary transcript of a discontinuous gene.

TATA box: A sequence found in the promoter (part of the 5' flanking region) of many genes. Deletion of this site (the binding site of transcription factor TFIID) causes a marked reduction in transcription, and gives rise to heterogeneous transcription initiation sites.

Transcription factor: A protein which is involved in the transcription of genes. These usually bind to DNA as part of their function (but not necessarily). A transcription factor may be general (i.e. acting on many or all genes in all tissues), or tissue-specific (i.e. present only in a particular cell type, and activating the genes restricted to that cell type). Its activity may be constitutive, or may depend on the presence of some stimulus; for example, the glucocorticoid receptor is a transcription factor which is active only when glucocorticoids are present.

Transcription: The process of copying DNA to produce an RNA transcript. This is the first step in the expression of any gene. The resulting RNA, if it codes for a protein, will be spliced, polyadenylated, transported to the cytoplasm, and by the process of translation will produce the desired protein molecule.

Translation: The process of decoding a strand of mRNA, thereby producing a protein based on the code. This process requires ribosomes (which are composed of rRNA along with various proteins) to perform the synthesis, and tRNA to bring in the amino acids. Sometimes, however, people speak of "translating" the DNA or RNA when they are merely reading the nucleotide sequence and predicting from it the sequence of the encoded protein. This might be more accurately termed "conceptual translation".

Upstream/Downstream: In an RNA, anything towards the 5' end of a reference point is "upstream" of that point. This orientation reflects the direction of both the synthesis of mRNA, and its translation - from the 5' end to the 3' end. In DNA, the situation is a bit more complicated. In the vicinity of a gene (or in a cDNA), the DNA has two strands, but one strand is virtually a duplicate of the RNA, so its 5' and 3' ends determine upstream and downstream, respectively. Note that in genomic DNA, two adjacent genes may be on different strands and thus oriented in opposite directions. Upstream or downstream is only used in conjunction with a given gene.

Upstream activator sequence: A binding site for transcription factors, generally part of a promoter region. A UAS may be found upstream of the TATA sequence (if there is one), and its function is (like an enhancer) to increase transcription. Unlike an enhancer, it can not be positioned just anywhere or in any orientation.

Zinc finger: A protein structural motif common in DNA binding proteins. Four Cys residues are found for each "finger" and one finger can bind a molecule of zinc. A typical configuration is: CysXxxXxxCys-(intervening 12 or so aa's)-CysXxxXxxCys.

Appendix F

List of Publications

Journals

- T. Sobha Rani, S. Durga Bhavani, Raju S. Bapi. Analysis of E.coli promoter recognition problem in dinucleotide feature space, *Bioinformatics* 23(5): 582-588 (2007).

Proceedings of the Conferences

- T. Sobha Rani, S. Bapi Raju. *E.coli* promoter recognition through wavelets, Proceedings of BIOCOMP'08 *The 2008 International Conference on Bioinformatics & Computational Biology*, 14-17 July 2008, Las vegas, U.S.A.
- T. Sobha Rani, S. Durga Bhavani, S. Bapi Raju. Promoter Recognition using dinucleotide Features: A case study for E.coli, Proceedings of ICIT 2006 (*9th International Conference on Information Technology*), 18-21 December 2006, Bhubhaneswar, India.

Communicated

- T. Sobha Rani, S. Bapi Raju. Analysis of *n-gram* based promoter recognition methods and application to whole genome promoter prediction *communicated to In Silico Biology*.
- T. Sobha Rani, S. Bapi Raju. Position weight matrices role in the context of local and global signal extraction and classification of *E.coli* promoters *communicated to Journal of Bioinformatics and Computational biology*.

- T. Sobha Rani, S. Bapi Raju. Cascaded multi-level promoter recognition of *E.coli* using dinucleotide Features *communicated to ICIT08*.

Publications related to the thesis

- S. Durga Bhavani, T. Sobharani, S. Bapi Raju. Feature Selection Using Correlation Fractal Dimension: Issues and Applications in Binary Classification Problems. *Applied Soft Computing*, pp 555-563, Vol. 8, 2008.
- Anuj Kumar, U. Pramod Kumar, T. Sobha Rani, S. Durga Bhavani, S. Bapi Raju. Identification of Promoter Region in a DNA sequence using EM algorithm and Neural Networks. Proceedings of the First Indian International Conference on AI (IICAI), 2003.