



EAST WEST UNIVERSITY



Identification of Genetic Promoter through Stochastic Approach

by

Qazi Adnan Ghyas

Id: 2005-2-96-008

M.Sc. (CSE), East West University

A Thesis submitted to

Department of Computer Science and Engineering

Faculty of Sciences and Engineering

East West University

Dhaka, Bangladesh

As partial fulfilment of the requirements for the degree of
Master of Science in Computer Science and Engineering



EAST WEST UNIVERSITY



Identification of Genetic Promoter through Stochastic Approach

by

Qazi Adnan Ghyas

Id: 2005-2-96-008

M.Sc. (CSE), East West University

A Thesis submitted to

Department of Computer Science and Engineering

Faculty of Sciences and Engineering

East West University

Dhaka, Bangladesh

As partial fulfilment of the requirements for the degree of
Master of Science in Computer Science and Engineering

DECLARATION

This is certified that this thesis is an original work and was done by me and it has not been submitted elsewhere for the requirement of any degree or diploma or for any other purposes except for publication.

Signature of the candidate

Qazi Adnan Ghyas

(Qazi Adnan Ghyas)



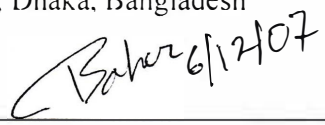
Donated by.....



Acceptance

The Thesis entitled *Identification of Genetic Promoter through Stochastic Approach* submitted by Qazi Adnan Ghyas, ID No. 2005-2-96-008 to the Department of Computer Science and Engineering, East West University, Dhaka 1212, Bangladesh is accepted as satisfactory for partial fulfilment of requirements for the degree of Masters of Science (MS) in Computer Science and Engineering on December 06, 2007.

BOARD OF EXAMINERS

1. 
Mr. Syed Akhter Hossain
Associate Professor and Chairperson
Department of Computer
East West University
Dhaka, Bangladesh
Chairperson, CSE Department
2. 
06.12.2007
Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh
External Member
3. 
06/12/07
Mr. Syed Murtuza Baker
Lecturer
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
Supervisor

Acknowledgement

I must acknowledge the inspiration, effort and time given by my supervisor Mr. Syed Murtuza Baker, Lecturer, Computer Science and Engineering Department, East West University, Dhaka, Bangladesh. Truly speaking, without his support and inspiration, I could not make it possible. I would also like to thanks to my honourable teacher Mr. Syed Akhter Hossain, Chairperson and Associate Professor, Computer Science and Engineering Department, East West University, Dhaka, Bangladesh, without his direction I could not complete my work.



Analysis of a gene sequence, which is transcribed into RNA and then translated into protein, is a difficult task. If this could be achieved, it would make possible better understand how the organisms are developed from DNA information. The behavior of gene is highly influenced by promoter sequences residing upstream or downstream of the Transcription Start Site (TSS). The promoter recognition process is a part of the complex process where genes interact with each other over time and actually regulates the whole working process of a cell. This paper attempts to develop an efficient algorithm that can successfully distinguish promoters and non promoters by analyzing statistical data. A learning model is developed from the known dataset to predict the unknown ones.

Results: We have developed an efficient algorithm that can successfully distinguish genes from non-gene sequences by analyzing statistical data. A learning model is initially developed to train the Support Vector Machine (SVM) to identify distinctive features between gene and non gene. Then this context was used to predict other foreign sequence by the SVM. Our system has been tested using standard plant prom data sequence from the EMBL and the performances are: 0.86 for the Sensitivity and 0.90 for the specificity.

Table of Contents

1.	Introduction	-----	1
1.1	Background	-----	1
1.2	Motivation	-----	2
1.3	Objectives	-----	3
1.4	Related Work	-----	3
1.5	Layout Synthesis	-----	4
2.	Literature Review	-----	6
2.1	What is Bio Informatics	-----	6
2.2	Introduction to Molecular Biology	-----	7
2.2.1	Cells	-----	7
2.2.2	Deoxyribonucleic Acid (DNA)	-----	8
2.2.3	Ribonucleic Acid (RNA)	-----	10
2.2.4	mRNA	-----	11
2.2.5	tRNA	-----	11
2.2.6	Amino Acids	-----	11
2.2.7	Genome, Transcriptome, Proteome	-----	14
2.3	Protein Metabolism	-----	15
2.3.1	The Genetic Code	-----	16
2.3.2	Protein Synthesis	-----	18
2.3.3	Protein Targeting and Degradation	-----	20
2.4	Introduction to Promoter and Gene	-----	20
2.4.1	Promoter	-----	20
2.4.2	Gene	-----	22
2.5	Support Vector Machine (SVM)	-----	24
3.	Problem Definition	-----	26
3.1	Necessity of Promoter Identification	-----	26
4.	Proposed Method and Data Preparation	-----	28
4.1	Data Preparation	-----	28
4.2	Proposed Method to Identify Promoter	-----	28
4.3	SVM Mathematical Model	-----	29
4.4	Training with SVM	-----	30
5.	Experimental Results and Comparative Analysis	-----	32
5.1	Prediction Accuracy	-----	32
5.2	Comparison with Existing Methods	-----	33

Table of Contents

6. Conclusion and Further Scope	-----	35
6.1 Conclusion	-----	35
6.2 Further Scope	-----	35
References	-----	36
List of Figures	-----	40
List of Tables	-----	41



Introduction

1.1 BACKGROUND

The biological technology becomes popular science in recent years. Biologists try to investigate the secrets of life by going into gene sequences. However the gene sequence data grow too huge recently. Though some mathematicians have presented mathematical or statistical method to discover features of gene sequences, it is still time consuming and inefficient if we study gene scientists get into the biological technology, and give some methods which take advantages of computer power to see into gene sequences.

Proteins perform a biological function by interacting with other proteins, compounds, RNA, and DNA. Understanding the characteristics of interfacial sites is a requirement for understanding the molecular recognition process. In addition, the ability to predict interfacial sites is important in mutant design and drug design. The physical and chemical aspects of the protein interface have been investigated in a number of studies. As a result, general interfacial sites are widely recognized as being more hydrophobic, flat, and protruding than outer surfaces.

The *promoter* plays an important role in DNA transcription. It is defined as the sequence in the region of the upstream of the *transcriptional start site (TSS)* and responsible for the transcription from DNA to RNA. The related position of the promoter in a DNA sequence is illustrated in Figure 1.1.

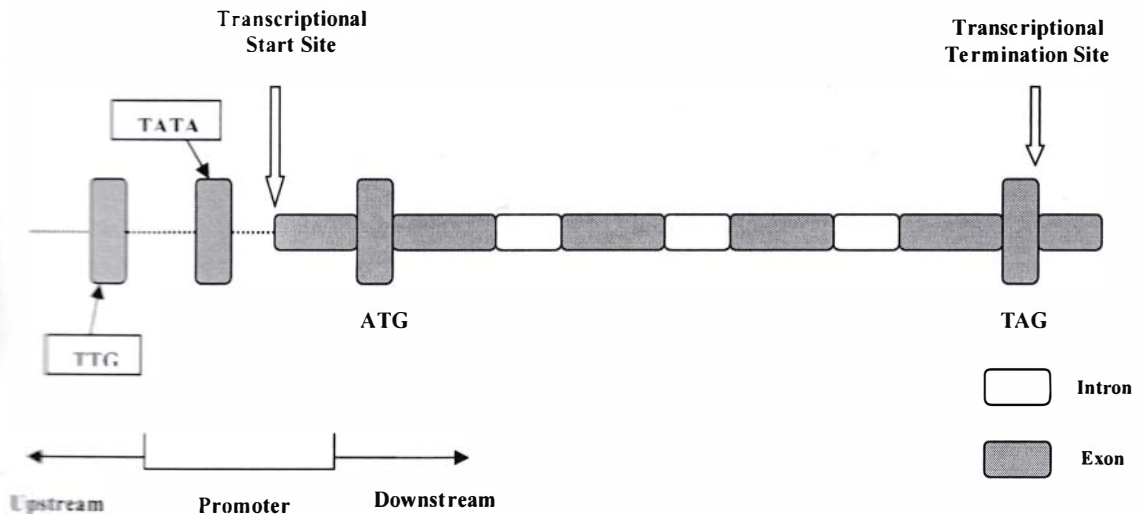


Figure 1.1: The promoter region in a DNA sequence

A promoter is required for a DNA sequence to be transcribed. In a DNA sequence transcription, there must be a promoter in the sequence. When the promoter sequence is bound with the RNA Polymerase II enzyme, the DNA sequence can be transcribed

into mRNA sequence. The central dogma of molecular biology is shown in Figure 1.2.

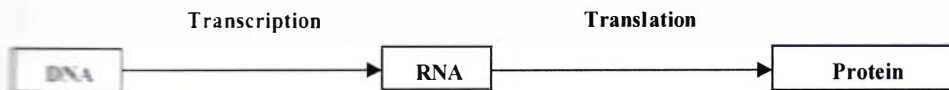


Figure 1.2: The central dogma of molecular biology

Since the promoter is located around the upstream of TSS in a DNA sequence, and the RNA Polymerase II is always binding in that region. The transcription starts from the end of 5' of the DNA sequence, the 5' UTR (upstream of TSS) contains promoter sites (such as TATA-box), and the 3' UTR (downstream of TSS) contains stop codon. The translation stops when the stop codon is met.

However, some times even the upstream of TSS of a DNA sequence contains some transcriptional features, the promoter may not exist. Whether a DNA sequence transcribed or not can be verified biological experiments, but experiments are usually time consuming and take high cost. With the promoter prediction method, we may be able to narrow down the promoter regions among massive DNA sequence. A farther experiment then can be designed and tested. Therefore, much more time and cost will be saved.

1.2 MOTIVATION

The predominate quest of biological science is to understand how nature functions. This discovery of the genetic code—the language a living organisms uses to produce proteins from a nucleic-acid template- was a major step toward understanding a complex and intriguing biological process [1]. The use of informatics to organize, manage and analyze these data has consequently become an important element of biology and medical research.

Genetic code resides in the gene of a living organism which actually produces protein through a process of Transcription and Translation. To understand the transcriptional process it is necessary to identify and characterize the promoter as the motifs. The promoter region plays a role in triggering the transcriptional process.

Promoter actually resides mostly in the upstream of a Transcription Start Site (TSS) which ensure the DNA sequence to be transcribed in to mRNA. In prokaryotes some features have been recognized in the upstream of a gene to be the indicator of a promoter. Finding promoter sequences in eukaryotes are much more difficult than finding promoter sequences in prokaryotes. As a result researchers mostly depend on statistical data.

Despite the important role of promoters the numbers of genes whose promoters have been identified are limited [2]. Traditional biological methods are not enough to maintain and annotate the ever growing vast genomic data. Many researchers are now working to develop good and efficient computer algorithms to identify promoter region from DNA sequence.

A number of studies have been carried out on promoter prediction using Hidden Markov Model (HMM) and Artificial Neural Network (ANN). These different methods show different success rates with different dataset [3, 4]. Most of these methods identified promoter regions by analyzing various promoter features. A machine is then trained with these features to identify an unknown promoter. This work attempts to propose a method to identify promoter by examining certain features which proliferantly prevails in promoter. A Support Vector Machine (SVM) [5] working as a binary classifier is then trained with these features to individually identify promoter and non promoter.

1.3 OBJECTIVE

The principal objective of this project was to develop an efficient tool that can discriminate between promoter and non-promoter in an unknown sequence with proper accuracy. The results of promoter prediction with our approach in Plant, Human, Drosophila, Mouse and Rat have clearly proved the validity of using frequency distribution of 4mers in discrimination between promoter and non promoter.

1.4 RELATED WORK

The main problem in working with biological data is that they do not produce a symbolic pattern recognition that provide control signal to the cellular protein-production machinery. So a symbolic pattern-recognition task for which computers are particularly well suited cannot be applied. Total gene of even a small virus can be several hundreds to thousand bases long, so discovering patterns, is tedious and repetitive. Researches are continuously trying to find some mechanism to discover these patterns and associate it with protein production. In this research promoter identification plays a vital role as they provide the control signal to the genes which ultimately triggers the translation and transcription of protein. Following are some distinct features of promoter sequences that have been identified and mentioned in several literatures.

TATA-Box CAAT box: A TATA box is DNA sequence found in the core promoter region of prokaryotes and eukaryotes. The TATA box assists in directing RNA polymerase II to the initiation site downstream on DNA [6, 7]. The two identified promoter sequence are the -10 box and -35 boxes. -10 and -35 indicates that these elements always appear around the position of -10 and -35 considering Transcription Start Site (TSS) is at +1. The -10 box is TATA -box [3, 8, 9] and -35 is the CAAT box [9].

CpG Island: There are regions of the DNA in a gene which have a higher concentration of CpG sites, known as CpG Island. Roughly half of all genes in mammalian genomes have CpG islands associated with the start of the gene. They exist in approximately 40% of promoters of mammalian genes (about 70 % in human promoters). Because of this, the presence of a CpG island is used to help in the

prediction and annotation of genes. "CpG" stands for cytosine and guanine separated by a phosphate, which links the two nucleotides together in DNA [10, 11, 12].

Pederson and Engelbrecht used an artificial neural network to discover signals in the upstream of the TSS [4]. They attempted to predict whether a given DNA sequence has a TSS or not.

Pederson et al. also used Hidden Markov Model (HMM) to characterize the prokaryotic and eukaryotic promoters. They used promoters from two species to train the HMM a found that HMMs after training can be used to help to classify the unknown promoters in prokaryotic [3].

Promoter has been identified using protein subunit composition [9]. Promoter has been predicted from Homo sapiens DNA sequence by identifying the RNA-poll transcribe site in Gill and Tijan work, where they tried to find the comparative measurement of this site by looking at the partly homologous bacterial subunits. Different data mining methods are also used for this purpose such as Positional Specific Matrix and Weight Matrix. The GBI (graph-Based Induction) method [13] is one kind of data mining methods. GBI is applied to minimize the size of the graph by replacing identical pattern and assigning new nodes.

1.5 LAYOUT SYNTHESIS

An outline for the rest of this thesis will be structured as follows:

Chapter 1 – Introduction

Already have discussed above.

Chapter 2 – Literature Review

This chapter will contain all the relevant topics of the different areas related to Bio Informatics and Molecular Biology, Promoter and Gene and different promoter Identification Methods. The Support Vector Machine is also discussed in this chapter. These studied areas will help establish a general understanding and motivation for the thesis.

Chapter 3 – Problem Definition

This chapter will cover the Necessity of Promoter Identification Method and the analysis of existing Different Promoter Identification Method.

Chapter 4 – Proposed Method and Data Preparation

In this chapter I discuss about my propose solution and the preparation of required data.

Chapter 5 – Experimental Results and Comparative Analysis

This chapter will present all of the results and performance produced throughout the project. This will include results from analysis, design, implementation and testing stages.

Chapter 6 – Conclusion and Further Scope

This concludes with an evaluation and discussion of the success or failure of the projects' outcome and also tried to give a direction for future.



Literature Review

2.1 WHAT IS BIOINFORMATICS?

Defining the terms bioinformatics and computational biology is not necessarily an easy task, in the past few years, as the areas have grown, a greater confusion into these two terms has prevailed. For some, the terms bioinformatics and computational biology have become completely interchangeable terms, while for others, there is a great distinction [14].

Computational biology and bioinformatics are multidisciplinary fields, involving researchers from different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics. The goal of these two fields [14] is as follows:

- **Bioinformatics:** Typically refers to the field concerned with the collection and storage of biological information. All matters concerned with biological databases are considered bioinformatics.
- **Computational biology:** Refers to the aspect of developing algorithms and statistical models necessary to analyze biological data through the aid of computers.

In this respect, the understanding of bioinformatics and computational biology follows the definitions listed below:

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

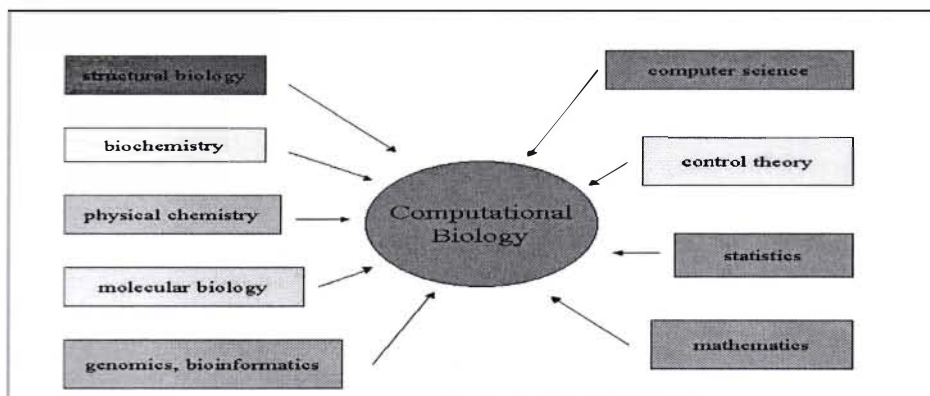


Figure 2.1: Different fields of Bioinformatics

Now a days Bio informatics is becoming a very hot field. One reason behind is that it is tied to the human genome project which has generated a lot of popular interest. Various advances in molecular biology techniques (such as genome sequencing and microarrays) have led to a large amount of data that needs to be analyzed. Now that we are close to having the human genome finished, what does it all mean? That's where bioinformatics steps in. Bioinformatics can lead to important discoveries as well as help companies save time and money in the long run. In addition, there needs to be methods to manage large amounts of data. One of the biggest reasons for bioinformatics being a hot field is the old supply and demand adage. There just are too few people adequately trained in both biology and computer science to solve the problems that biologists need to have solved.

2.2 INTRODUCTION TO MOLECULAR BIOLOGY

In order to be a good computational biologist, it is important to understand the terminology and basic processes behind the biological problems. Many interesting problems arise out of sequence analysis. There are two different types of biological sequences studied in this class: DNA/RNA and amino acids. But first I want to make sure the basics are covered.

2.2.1 CELLS

Every organism is made up of tiny structures called cells. Often these cells are too small to be seen with the naked eye. Each cell is in itself a complex system enclosed in a membrane. Some organisms, such as bacteria and baker's yeast are composed of only a single cell (i.e. they are unicellular). Other organisms are made up of many different cells (i.e. they are multi cellular). For instance, the human body is composed of around 60 trillion cells. Humans have about 320 different cell types, each having a different type of function or structural property.

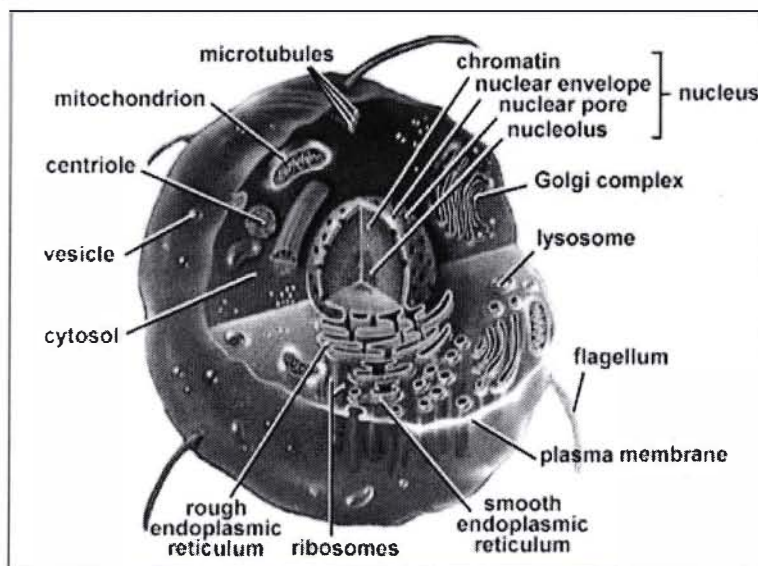


Figure 2.2: Structure of an animal cell [14]

There are two types of organisms: eukaryotes and prokaryotes [14]. Eukaryotes (or as Bruce Rice from the University of Oklahoma calls them the “You and I” Karyotes) represent most of the organisms which we can see, including plants and animals. Prokaryotes (such as bacteria) are smaller than eukaryotic cells and have simpler structure. Prokaryotes are single cellular organisms (but not all single-celled organisms are prokaryotes!)

So what is the difference between the two types of cells? A eukaryotic cell has a nucleus, which is separated from the rest of the cell by a membrane. Inside the nucleus are the chromosomes, where all of the genetic information for the organism is stored. In addition, eukaryotic cells contain membrane bound organelles with various functions, including centrioles, lysosomes, mitochondria, ribosomes, etc.

Contained within the nucleus are one or several long double stranded DNA molecules organized as chromosomes. For humans, there are 22 pairs of autosomes, as well as one pair of sex chromosomes [14]. One copy of each pair is inherited from each parent.

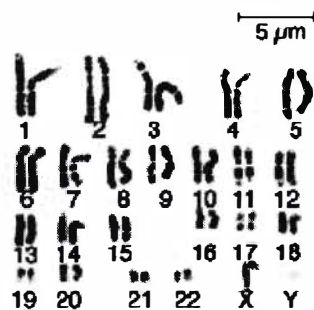


Figure 2.3: Karyotype showing the 23 pairs of human chromosomes [14]

2.2.2 DNA

Deoxyribonucleic Acid (DNA) is the basis for the building blocks encoding the information of life. A single stranded DNA molecule, called a *polynucleotide* or *oligomer*, is a chain of small molecules called nucleotides. There are four different nucleotides, or bases: adenosine (A), cytosine (C), guanine (G) and thymine (T) [14].

The bases can be separated into two different types: purines (A and G) and pyrimidines (C and T). The difference between purines and pyrimidines is in the base structure.

Stringing together a simple alphabet of four characters together we can get enough information to create a complex organism! Different nucleotides can be strung together to form a polynucleotide. However, the ends of the polynucleotide are different, meaning that each polynucleotide sequence will have directionality. The ends of the polynucleotide are marked either 3' or 5'. The general convention is to label the coding strand from 5' to 3' (left to right).

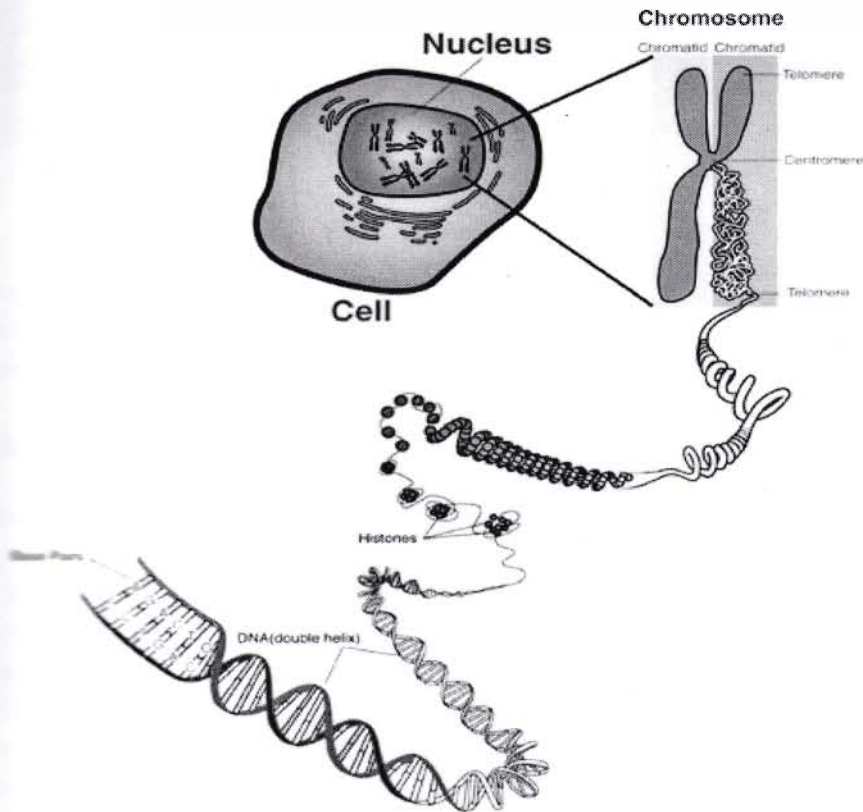


Figure 2.4: Snap of a DNA sequence within a cell [14]

For instance, the following is a polynucleotide:

5' G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3'

DNA can be either single-stranded or double stranded. When DNA is double-stranded, the second strand is referred to as the reverse complement strand. This name is derived from the fact that the directionality of this second strand runs in the opposite direction as the first, and the fact that the bases in the second strand are complementary to the bases in the first. Complementary bases are determined by which pairs of nucleotides can form bonds between them. In the case of DNA, A binds to T, and C binds to G. For the polynucleotide given above, the double-stranded polynucleotide is as follows:

5' G→T→A→A→A→G→T→C→C→C→G→T→T→A→G→C 3'

|||||

3' C←A←T←T←T←C←A←G←G←G←C←A←A←T←C←G 5'

Two complementary polynucleotide chains form a stable structure known as the DNA double helix. This spring represents the 50th anniversary of the discovery of the double helix structure of DNA by Watson, Crick and Franklin.

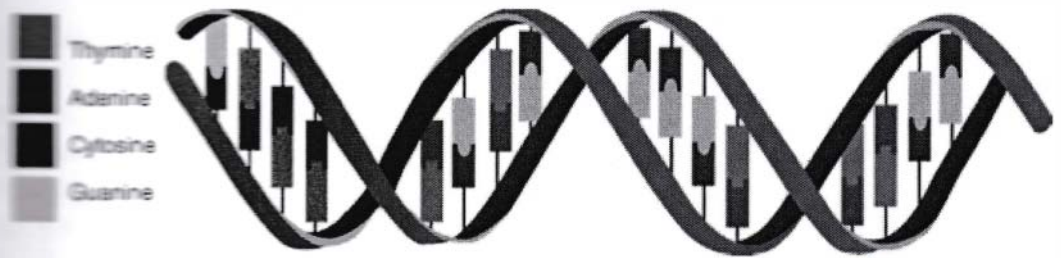


Figure 2.5: DNA double helix structure [14]

Note that in this image, there appear to be two types of grooves: A larger one, which is called the major groove and a smaller one, known as the minor groove. In addition, there are roughly 10.5 base pairs in one complete turn of the helix.

2.2.3 RNA

Ribonucleic Acid (RNA) is similar to DNA in the fact that it is constructed from nucleotides. However, instead of thymine (T), an alternative base uracil (U) is found in RNA. RNA can be found as double-stranded or single-stranded, and can also be part of a hybrid helix where one strand is an RNA strand and the other is a DNA strand. RNA is generally found as a single stranded molecule that may form a secondary structure or tertiary structures due to the complementary bases between parts of the same strand [15]. RNA folding will be discussed in detail during a later class period. RNA is important in the cell and contributes in a variety of ways. One of the most important roles of RNA is in protein synthesis. Two of the major RNA molecules involved in protein synthesis are messenger RNA (mRNA) and transfer RNA (tRNA).

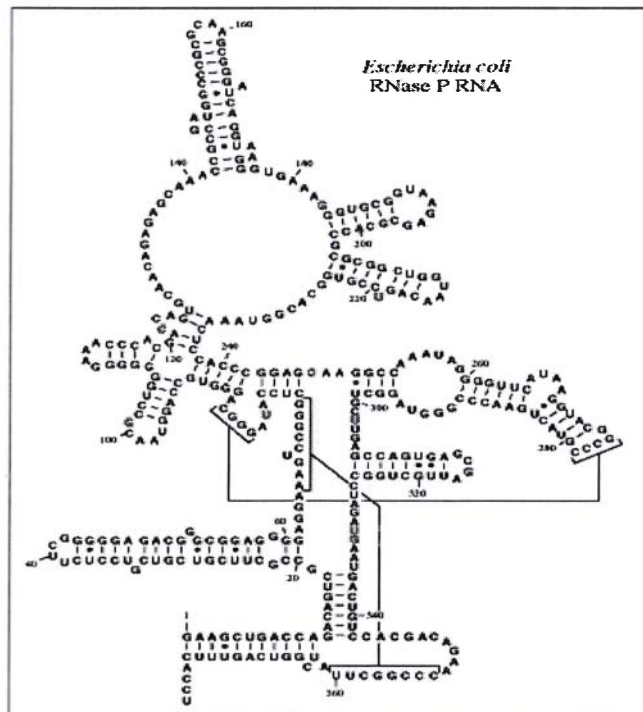


Figure 2.6: Secondary structure for *E. coli* RNase P RNA [15]

2.2.4 mRNA

mRNA encodes the genetic information as copied from the DNA molecules. *Transcription* is the process in which DNA is copied into an RNA molecule. The resulting linear molecule is an mRNA transcript. In eukaryotic cells, before the mRNA can be translated into a protein, it needs to be modified [15]. The nature of most eukaryotic genes is that the genes are created in pieces, where coding regions, called *exons*, are interspersed with noncoding regions, called *introns*. One of the steps in processing the mRNA is to remove the intronic regions and to splice together the coding, or exonic regions. The processed mRNA can then be transported from the nucleus and translated into a protein sequence.

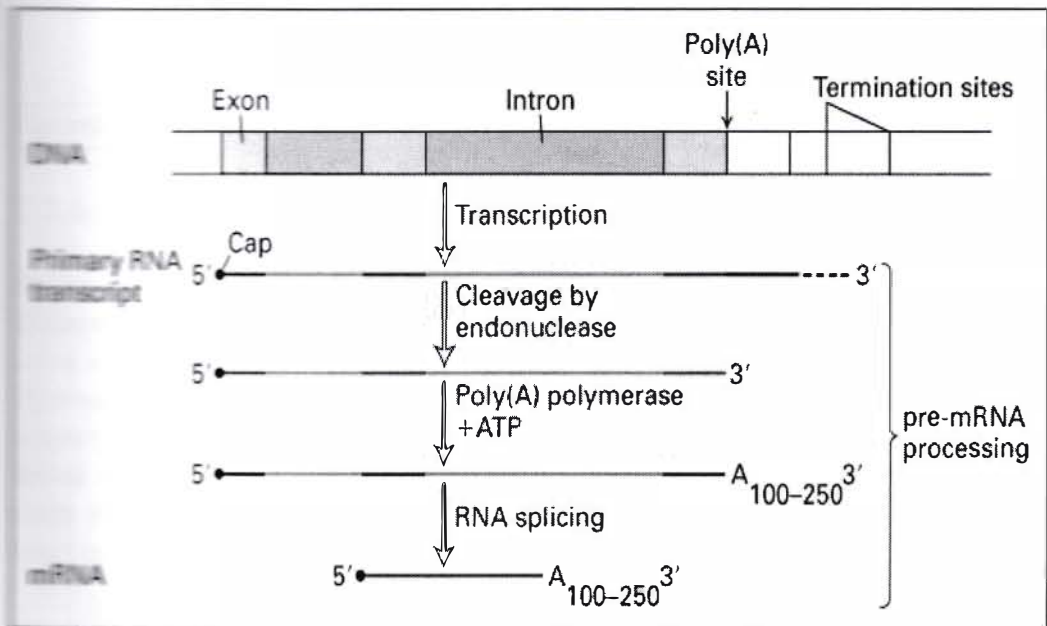


Figure 2.7: mRNA processing

2.2.5 tRNA

tRNA molecules develop a well-defined three-dimensional structure which is critical in the creation of proteins. Attached to each tRNA molecule is an amino acid (which will be discussed momentarily). The amino acid to be attached is determined by a three base sequence called an anticodon sequence, which is complementary to the sequence in the mRNA. *Translation* is the process in which the nucleotide base sequence of the processed mRNA is used to order and join the amino acids into a protein with the help of ribosomes and tRNA [15].

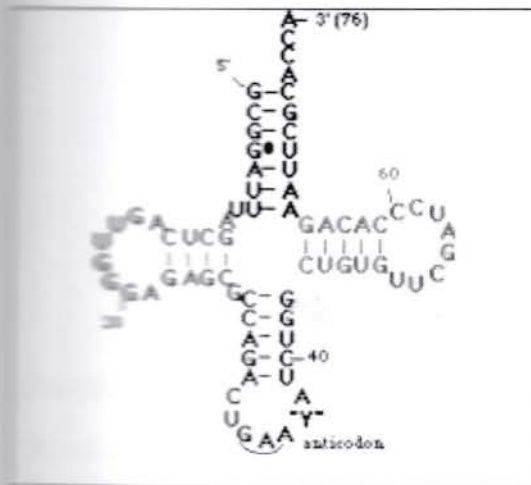


Figure 2.8: tRNA secondary structure [15]



Figure 2.9: tRNA tertiary structure [15]

2.2.6 AMINO ACIDS

Amino acids are the building blocks from which proteins are made. There are 20 different amino acids that vary from each other by their side chain groups. Amino acids can be classified into different groups based on their solubility in water. *Hydrophilic* amino acids are water soluble, while *hydrophobic* are not. This property becomes important when a protein sequence is made. Amino acids are linked to one another via a single chemical bond, called a *peptide bond*. A linear chain of amino acids can be referred to as a *peptide* (if it is short – less than 30 a.a. long) or *polypeptide* (which can be upwards of 4000 residues long).

Table 2.1: Amino Acid Codes

One-letter	Three-letter	Full name
G	GLY	Glycine
A	ALA	Alanine
V	VAL	Valine
L	LEU	Leucine
I	ILE	Isoleucine
F	PHE	Phenylalanine
P	PRO	Proline
S	SER	Serine
T	THR	Threonine
C	CYS	Cysteine
M	MET	Methionine
W	TRP	Tryptophan

Y	TYR	Tyrosine
N	ASN	Asparagine
Q	GLN	Glutamine
D	ASP	Aspartic acid
E	GLU	Glutamic acid
K	LYS	Lysine
R	ARG	Arginine
H	HIS	Histidine



Proteins

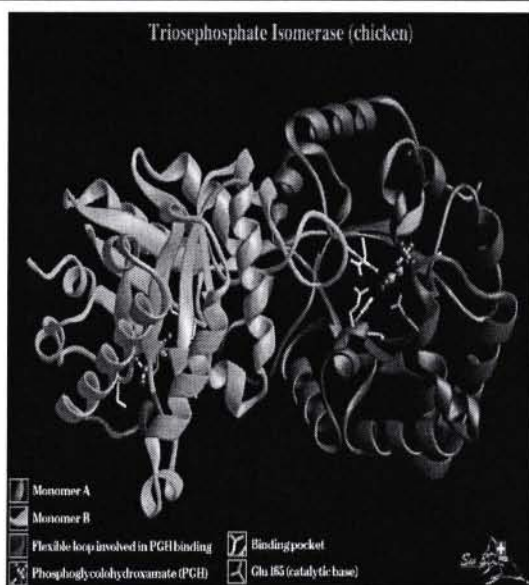
Proteins are polypeptides that have a three dimensional structure. They can be described through four different hierarchical levels:

- **Primary structure** – the sequence of amino acids constituting the polypeptide chain.
- **Secondary structure** – the local organization of the parts of the polypeptide chain into secondary structures such as α helices and β sheets.
- **Tertiary structure** – the three dimensional arrangements of the amino acids as they react to one another due to the polarity and resulting interactions between their side chains.
- **Quaternary structure** – if a protein consists of several protein subunits held together, then the protein can be described as well by the number and relative positions of the subunits.

Visualization of Protein Structures.



Magenta: alpha helix
Gold: Beta Sheets



Blue: Monomer A
Orange: Monomer B

Figure 2.10: Protein Structure [15]

Calculating the secondary and tertiary structure of a protein given its primary structure is not an easy task. Protein folding prediction will be covered at some point close to the end of the semester.

Monomer – Any small molecule that can be linked with others of the same type to form a polymer. For the purpose of this class, the molecules could be nucleic acids, amino acids, or proteins.

Dimer – Two small molecules of the same type linked together.

Trimer – Three small molecules of the same type linked together.

Oligomer – General term for a short polymer most commonly consisting of nucleic acids or amino acids.

Polymer – Any large molecule consisting of multiple identical or similar subunits linked by covalent bonds.

Putting it all together, we get the flow of genetic information. That is, DNA directs the synthesis of RNA, and RNA then in turn directs the synthesis of protein. This flow of genetic information from nucleic acids to protein has been called the Central Dogma of Molecular Biology.

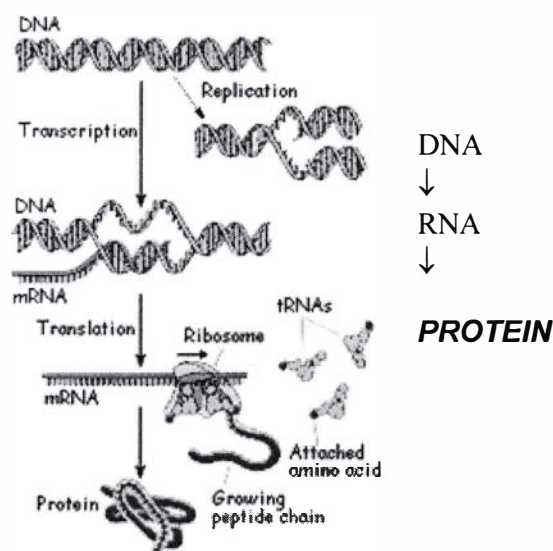


Figure 2.11: Central Dogma of Molecular Biology [15]

2.2.7 GENOME, TRANSCRIPTOME, PROTEOME

Whenever the term *genome* is used, it typically refers to the chromosomal DNA of an organism, or as far as sequencing is concerned, the heterochromatic regions of the chromosomal DNA. The number of chromosomes and genome size varies quite significantly from one organism to another. An example list of genome sizes is given below. Don't be fooled by this table that the size of the genome and the number of

genes determines the complexity of an organism. In fact, many plant genomes are much greater in size than the human genome!

Table 2.2: Chromosomes, Genes and Genome Sizes in different Organisms

ORGANISM	CHROMOSOMES	GENOME SIZE	GENES
Homo sapiens (Humans)	23	3,200,000,000	~ 30,000
Mus musculus (Mouse)	20	2,600,000,000	~30,000
Drosophila melanogaster (Fruit Fly)	4	180,000,000	~18,000
Saccharomyces cerevisiae (Yeast)	16	14,000,000	~6,000
Zea mays (Corn)	10	2,400,000,000	???

The term *transcriptome* refers to the complete collection of all possible mRNAs (including splice variants) of an organism. This can be thought of as the regions of an organism's genome that get *transcribed* into messenger RNA. In some cases, the transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes.

The term *proteome* refers to the complete collection of proteins that can be produced by an organism. The proteome can be studied either as a static (sum of all proteins possible) or a dynamic (all proteins found at a specific time point) entity.

2.3 PROTEIN METABOLISM

Proteins are the end products of most information pathways. A typical cell requires thousands of different proteins at any given moment. These must be synthesized in response to the cell's current needs, transported (targeted) to their appropriate cellular locations, and degraded when no longer needed.

An understanding of protein synthesis, the most complex biosynthetic process, has been one of the greatest challenges in biochemistry. Eukaryotic protein synthesis involves more than 70 different ribosomal proteins; 20 or more enzymes to activate the amino acid precursors; a dozen or more auxiliary enzymes and other protein factors for the initiation, elongation, and termination of polypeptides; perhaps 100 additional enzymes for the final processing of different proteins; and 40 or more kinds of transfer and ribosomal RNAs. Overall, almost 300 different macromolecules cooperate to synthesize polypeptides. Many of these macromolecules are organized into the complex three-dimensional structure of the ribosome [15].

To appreciate the central importance of protein synthesis, consider the cellular resources devoted to this process. Protein synthesis can account for up to 90% of the chemical energy used by a cell for all biosynthetic reactions. Every prokaryotic and eukaryotic cell contains from several to thousands of copies of many different

proteins and RNAs. The 15,000 ribosomes, 100,000 molecules of protein synthesis-related protein factors and enzymes, and 200,000 tRNA molecules in a typical bacterial cell can account for more than 35% of the cell's dry weight.

Despite the great complexity of protein synthesis, proteins are made at exceedingly high rates. A polypeptide of 100 residues is synthesized in an *Escherichia coli* cell (at 37 °C) in about 5 seconds. Synthesis of the thousands of different proteins in a cell is tightly regulated, so that just enough copies are made to match the current metabolic circumstances. To maintain the appropriate mix and concentration of proteins, the targeting and degradative processes must keep pace with synthesis. Research is gradually uncovering the finely coordinated cellular choreography that guides each protein to its proper cellular location and selectively degrades it when it is no longer required.

2.3.1 THE GENETIC CODE

Three major advances set the stage for our present knowledge of protein biosynthesis. First, in the early 1950s, Paul Zamecnik and his colleagues designed a set of experiments to investigate where in the cell proteins are synthesized. They injected radioactive amino acids into rats and, at different time intervals after the injection removed the liver, homogenized it, fractionated the homogenate by centrifugation, and examined the subcellular fractions for the presence of radioactive protein. When hours or days were allowed to elapse after injection of the labeled amino acids, *all* the subcellular fractions contained labeled proteins. However, when only minutes had elapsed, labeled protein appeared only in a fraction containing small ribonucleoprotein particles. These particles, visible in animal tissues by electron microscopy, were therefore identified as the site of protein synthesis from amino acids, and later were named ribosomes.

The second key advance was made by Mahlon Hoagland and Zamecnik, when they found that amino acids were “activated” when incubated with ATP and the cytosolic fraction of liver cells. The amino acids became attached to a heat-stable soluble RNA of the type that had been discovered and characterized by Robert Holley and later called transfer RNA (tRNA), to form **aminoacyl-tRNAs**. The enzymes that catalyze this process are the **aminoacyl-tRNA synthetases** [15].

The third advance resulted from Francis Crick's reasoning on how the genetic information encoded in the 4-letter language of nucleic acids could be translated into the 20-letter language of proteins. A small nucleic acid (perhaps RNA) could serve the role of an adaptor, one part of the adaptor molecule binding a specific amino acid and another part recognizing the nucleotide sequence encoding that amino acid in an mRNA (Fig. 27-2). This idea was soon verified. The tRNA adaptor “translates” the nucleotide sequence of an mRNA into the amino acid sequence of a polypeptide. The overall process of mRNA-guided protein synthesis is often referred to simply as **translation** [15].

These three developments soon led to recognition of the major stages of protein synthesis and ultimately to the elucidation of the genetic code that specifies each amino acid.

Since there are 4 possible bases (A, C, G, U) and 3 bases in the codon, there are $4 * 4 * 4 = 64$ possible codon sequences. However, the codon AUG can also be used as a signal to initiate translation, while the codons UAA, UAG, and UGA are terminal codons signaling the end of translation. That leaves a 61 codon sequences that can code for amino acids (AUG can also code for an amino acid). However, there are only 20 amino acids. Therefore the genetic code is redundant, meaning that a single amino acid could be coded for by several different codons.

Table 2.3: Genetic Code

		Second Position of Codon					
		U	C	A	G		
F i r s t P o s i t i o n	U	UUU Phe [F]	UCU Ser [S]	UAU Tyr [Y]	UGU Cys [C]	U	T h i r d P o s i t i o n
		UUC Phe [F]	UCC Ser [S]	UAC Tyr [Y]	UGC Cys [C]	C	
		UUA Leu [L]	UCA Ser [S]	UAA STOP	UGA STOP	A	
		UUG Leu [L]	UCG Ser [S]	UAG STOP	UGG Trp [W]	G	
	C	CUU Leu [L]	CCU Pro [P]	CAU His [H]	CGU Arg [R]	U	
		CUC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C	
		CUA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CUG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G	
	A	AUU Ile [I]	ACU Thr [T]	AAU Asn [N]	AGU Ser [S]	U	
		AUC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
		AUA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
		AUG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
	G	GUU Val [V]	GCU Ala [A]	GAU Asp [D]	GGU Gly [G]	U	
		GUC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GUA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GUG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	

Note that the initiator codon is labeled in green, and the terminal codons are labeled in red. The first column gives the triplet base; the second the three letter amino acid label, and the third the one letter amino acid label [15].

2.3.2 PROTEIN SYNTHESIS

As we have seen for DNA and, the synthesis of polymeric biomolecules can be considered in terms of initiation, elongation, and termination stages. These fundamental processes are typically bracketed by two additional stages: activation of precursors before synthesis and post synthetic processing of the completed polymer. Protein synthesis follows the same pattern. The activation of amino acids before their incorporation into polypeptides and the posttranslational processing of the completed polypeptide play particularly important roles in ensuring both the fidelity of synthesis and the proper function of the protein product. The cellular components involved in the five stages [15] of protein synthesis in *E. coli* and other bacteria are listed in Table 2.4; the requirements in eukaryotic cells are quite similar, although the components are in some cases more numerous. An initial overview of the stages of protein synthesis provides a useful outline for the discussion that follows.

Protein Biosynthesis Takes Place in Five Stages

Stage 1: Activation of Amino Acids For the synthesis of a polypeptide with a defined sequence, two fundamental chemical requirements must be met: (1) the carboxyl group of each amino acid must be activated to facilitate formation of a peptide bond, and (2) a link must be established between each new amino acid and the information in the mRNA that encodes it. Both these requirements are met by attaching the amino acid to a tRNA in the first stage of protein synthesis. Attaching the right amino acid to the right tRNA is critical. This reaction takes place in the cytosol, not on the ribosome. Each of the 20 amino acids is covalently attached to a specific tRNA at the expense of ATP energy, using Mg²⁺-dependent activating enzymes known as aminoacyl-tRNA synthetases. When attached to their amino acid (aminoacylated) the tRNAs are said to be “charged.”

Stage 2: Initiation The mRNA bearing the code for the polypeptide to be made binds to the smaller of two ribosomal subunits and to the initiating aminoacyl-tRNA. The large ribosomal subunit then binds to form an initiation complex. The initiating aminoacyl-tRNA basepairs with the mRNA codon AUG that signals the beginning of the polypeptide. This process, which requires GTP, is promoted by cytosolic proteins called initiation factors.

Stage 3: Elongation The nascent polypeptide is lengthened by covalent attachment of successive amino acid units, each carried to the ribosome and correctly positioned by its tRNA, which base-pairs to its corresponding codon in the mRNA. Elongation requires cytosolic proteins known as elongation factors. The binding of each incoming aminoacyl-tRNA and the movement of the ribosome along the mRNA are facilitated by the hydrolysis of GTP as each residue is added to the growing polypeptide.



Table 2.4: Components Required for the Five Major Stages of Protein Synthesis in *E. coli*

Stage	Essential Components
1. Activation of amino acids	20 amino acid 20 aminoacyl-tRNA synthetases 32 or more tRNAs Mg ²⁺
2. Initiation	mRNA N-Formylmethionyl-tRNA ^{fmet} Initiation codon in mRNA (AUG) 30S ribosomal subunit 50S ribosomal subunit Initiation factors (IF-1, IF-2, IF-3) GTP Mg ²⁺
3. Elongation	Functional 70S ribosome (initiation complex) Aminoacyl-tRNAs specified by codons Elongation factors (EF-Tu, EF-Ts, EF-G) GTP Mg ²⁺
4. Termination and Release	Termination codon in mRNA Release factors (RF-1, RF-2, RF-3)
5. Folding and posttranslational processing	Specific enzymes, cofactors, and other components for removal of initiating residues and signal sequences, additional proteolytic processing, modification of terminal residues, and attachment of phosphate, methyl, carboxyl, carbohydrate, or prosthetic groups

Stage 4: Termination and Release Completion of the polypeptide chain is signaled by a termination codon in the mRNA. The new polypeptide is released from the ribosome, aided by proteins called release factors.

Stage 5: Folding and Posttranslational Processing In order to achieve its biologically active form, the new polypeptide must fold into its proper three-dimensional conformation. Before or after folding, the new polypeptide may undergo enzymatic processing, including removal of one or more amino acids (usually from the amino terminus); addition of acetyl, phosphoryl, methyl, carboxyl, or other groups to certain amino acid residues; proteolytic cleavage; and/or attachment of oligosaccharides or prosthetic groups.

2.3.3 PROTEIN TARGETING AND DEGRADATION

The eukaryotic cell is made up of many structures, compartments, and organelles, each with specific functions that require distinct sets of proteins and enzymes. These proteins (with the exception of those produced in mitochondria and plastids) are synthesized on ribosomes in the cytosol, so how are they directed to their final cellular destinations.

Proteins destined for secretion, integration in the plasma membrane, or inclusion in lysosomes generally share the first few steps of a pathway that begins in the endoplasmic reticulum. Proteins destined for mitochondria, chloroplasts, or the nucleus use three separate mechanisms. And proteins destined for the cytosol simply remain where they are synthesized.

The most important element in many of these targeting pathways is a short sequence of amino acids called a **signal sequence** [15]. The signal sequence directs a protein to its appropriate location in the cell and, for many proteins, is removed during transport or after the protein has reached its final destination. In proteins slated for transport into mitochondria, chloroplasts, or the ER, the signal sequence is at the amino terminus of a newly synthesized polypeptide. In many cases, the targeting capacity of particular signal sequences has been confirmed by fusing the signal sequence from one protein to a second protein and showing that the signal directs the second protein to the location where the first protein is normally found. The selective degradation of proteins no longer needed by the cell also relies largely on a set of molecular signals embedded in each protein's structure.

2.4 INTRODUCTION TO PROMOTER AND GENE

2.4.1 PROMOTER

In biology, a promoter is a regulatory region of DNA located upstream (towards the 5' region) of a gene, providing a control point for regulated gene transcription [16].

The promoter contains specific DNA sequences that are recognized by proteins known as transcription factors. These factors bind to the promoter sequences, recruiting RNA polymerase, the enzyme that synthesizes the RNA from the coding region of the gene [17].

In prokaryotes, the promoter is recognized by RNA polymerase and an associated sigma factor, which in turn are brought to the promoter DNA by an activator protein binding to its own DNA sequence nearby [18].

In eukaryotes, the process is more complicated, and at least seven different factors are necessary for the transcription of an RNA polymerase II promoter.

Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene [19].

It is worth noting that promoters are not DNA specific, and can in fact locate upstream towards the 3' end of a RNA genome, e.g. Respiratory Syncytial Virus (RSV).

Identification of relative location:

As promoters are typically immediately adjacent to the gene in question, positions in the promoter are designated relative to the transcriptional start site, where transcription of RNA begins for a particular gene (i.e., positions upstream are negative numbers counting back from -1, for example -100 is a position 100 base pairs upstream) [20].

Promoter Elements

Core promoter - the minimal portion of the promoter required to properly initiate transcription [17].

Transcription Start Site (TSS): Approximately -34, A binding site for RNA polymerase

RNA polymerase I: transcribes genes encoding ribosomal RNA.

RNA polymerase II: transcribes genes encoding messenger RNA and certain small nuclear RNAs.

RNA polymerase III: transcribes genes encoding tRNAs and other small RNAs.

This type of promoter has General transcription factor binding sites.

Proximal promoter - the proximal sequence upstream of the gene that tends to contain primary regulatory elements [19, 17].

Transcription Start Site (TSS): Approximately -34

This type of promoter has Specific transcription factor binding sites

Distal promoter - the distal sequence upstream of the gene that may contain additional regulatory elements, often with a weaker influence than the proximal promoter [20].

Transcription Start Site (TSS): Anything further upstream (but not an enhancer or other regulatory region whose influence is positional/orientation independent).

This type of promoter has Specific transcription factor binding sites.

Prokaryotic promoters

In prokaryotes, the promoter consists of two short sequences at -10 and -35 positions upstream from the transcription start site. Sigma factors not only help in enhancing RNAP binding to the promoter but helps RNAP target which genes to transcribe [18]. The sequence at -10 is called the Pribnow box, or the -10 element, and usually consists of the six nucleotides TATAAT. The Pribnow box is absolutely essential to start transcription in prokaryotes [21].

The other sequence at -35 (the -35 element) usually consists of the six nucleotides TTGACA. Its presence allows a very high transcription rate.

Both of the above consensus sequences, while conserved on average, are not found intact in most promoters. On average only 3 of the 6 base pairs in each consensus

sequence is found in any given promoter. No promoter has been identified to date that has intact consensus sequences at both the -10 and -35; it is thought that this would lead to such tight binding by the sigma factor that the polymerase would be unable to initiate productive transcription [19].

It should be noted that the above promoter sequences are only recognized by the sigma-70 protein that interacts with the prokaryotic RNA polymerase. Complexes of prokaryotic RNA polymerase with other sigma factors recognize totally different core promoter sequences.

Eukaryotic promoters

Eukaryotic Promoters are extremely diverse and are difficult to characterize. They typically lie upstream of the gene and can have regulatory elements several kilobases away from the transcriptional start site. In eukaryotes, the transcriptional complex can cause the DNA to bend back on itself, which allows for placement of regulatory sequences far from the actual site of transcription. Many eukaryotic promoters, but by no means all, contain a TATA box (sequence TATAAA), which in turn binds a TATA binding protein which assists in the formation of the RNA polymerase transcriptional complex. The TATA box typically lies very close to the transcriptional start site (often within 50 bases) [18].

Eukaryotic promoter regulatory sequences typically bind proteins called transcription factors which are involved in the formation of the transcriptional complex. An example is the E-box (sequence CACGTG), which binds transcription factors in the basic-helix-loop-helix (bHLH) family (e.g. BMAL1-Clock, cMyc) [21].

2.4.2 GENE

A gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions [22,23]. The physical development and phenotype of organisms can be thought of as a product of genes interacting with each other and with the environment [24], and genes can be considered as units of inheritance. A concise definition of gene taking into account complex patterns of regulation and transcription, genic conservation and non-coding RNA genes has been proposed by Gerstein et al "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" [25].

In cells, genes consist of a long strand of DNA that contains a promoter, which controls the activity of a gene, and a coding sequence, which determines what the gene produces. When a gene is active, the coding sequence is copied in a process called transcription, producing an RNA copy of the gene's information. This RNA can then direct the synthesis of proteins via the genetic code. However, RNAs can also be used directly, for example as part of the ribosome. These molecules resulting from gene expression, whether RNA or protein, are known as gene products.

Most genes contain non-coding regions that do not code for the gene products, but regulate gene expression. The genes of eukaryotic organisms can contain non-coding regions called introns that are removed from the messenger RNA in a process known as splicing. The regions that actually encode the gene product, which can be much smaller than the introns, are known as exons. In eukaryotes, one single gene can lead to the synthesis of multiple proteins through the different arrangements of exons produced by alternative splicing. In prokaryotes, such as bacteria and archaea, genes are arranged in operons with promoter and operator sequences regulating transcription of an RNA that contains multiple coding sequences that produce multiple proteins.

The total complement of genes in an organism or cell is known as its genome. An organism's genome size is generally lower in prokaryotes such as bacteria and archaea have generally smaller genomes, both in number of base pairs and number of genes, than even single-celled eukaryotes. However, there is no clear relationship between genome sizes and perceived complexity of eukaryotic organisms. One of the largest known genomes belongs to the single-celled amoeba *Amoeba dubia*, with over 670 billion base pairs, some 200 times larger than the human genome [26]. The estimated number of genes in the human genome has been repeatedly revised downward since the completion of the Human Genome Project; current estimates place the human genome at just under 3 billion base pairs and about 20,000–25,000 genes [27]. A recent Science article gives a final number of 20,488, with perhaps 100 more yet to be discovered [28]. The gene density of a genome is a measure of the number of genes per million base pairs (called a megabase, Mb); prokaryotic genomes have much higher gene densities than eukaryotes. The gene density of the human genome is roughly 12–15 genes/Mb [29].

Functional structure of a gene

All genes have regulatory regions in addition to regions that explicitly code for a protein or RNA product. A universal regulatory region shared by all genes is known as the promoter [16], which provides a position that is recognized by the transcription machinery when a gene is about to be transcribed and expressed. Although promoter regions have a consensus sequence that is the most common sequence at this position, some genes have "strong" promoters that bind the transcription machinery well, and others have "weak" promoters that bind poorly. These weak promoters usually permit a lower rate of transcription than the strong promoters, because the transcription machinery binds to them and initiates transcription less frequently. Other possible regulatory regions include enhancers, which can compensate for a weak promoter. Most regulatory regions are "upstream" — that is, before or toward the 5' end of the transcription initiation site. Eukaryotic promoter [18, 21] regions are much more complex and difficult to identify than prokaryotic promoters [16].

Many prokaryotic genes are organized into operons, or groups of genes whose products have related functions and which are transcribed as a unit. By contrast, eukaryotic genes are transcribed only one at a time [30], but may include long

stretches of DNA called introns which are transcribed but never translated into protein (they are spliced out before translation) [31].

2.5 SUPPORT VECTOR MACHINE (SVM)

Support vector machines (SVM) are supervised learning algorithms proposed by Vapnik (Vapnick, 1995). Data examples labeled as positive or negative are projected into a high-dimensional feature space using a kernel, and the hyper-plane in the feature space is optimized to maximize the margin between the positive and negative examples.

We used LibSVM [32]. Only user-defined kernel subroutines were implemented. In this application, linear, polynomial, Radial Bias Function (RBF), and Gaussian kernels and their sum and product kernels are used. The RBF performs the best when vector data is to be mapped from linear to multidimensional hyperplane margin and we therefore report only the results for the RBF kernel.

Since the SVM optimizes the success ratio for whole sequences but does not optimize the recall and precision (defined below) of interaction sites, prediction performance depends on the ratio of negative and positive data in the learning process. According to the definition, only about 20% of a whole sequence is interaction site residues. If all data are used as learning samples, the prediction result at the default discriminant value (=zero) shows high precision and low recall. Accordingly, half the negative data (non-interaction site residues) were randomly removed from the learning sets when whole sequence residues were used as feature vectors, while a third of the negative data was randomly removed when only surface residues were used as feature vectors in Table 2.5 [33]. Basically, when the recall-FP/(FP+TP) curves were generated, all the data were used.

Since there was sufficient data for homo-hetero mixed validation (if three-fold cross validation was used, the learning time is too long), leave 375 (=2/3*563) cross validation was used. For homo and hetero complex validation, five-fold and three-fold cross validation were used, respectively. In predicting interaction site ratios, ten-fold cross validation was used for mixed homo and hetero validation data. When no explicit statement is made, "homo-hetero mixed data" was used.

For homo-hetero mixed validation data, "filtering by boosting" (Schapire, 1990), which converts a weak learning algorithm into a stronger learning machine, was also applied. This consisted of the following steps. First, the SVM learned using N samples (abbreviated as SVM-1). Using SVM-1 and a random number, $N/2$: wrongly predicted (false negative or false positive) samples and $N/2$: correctly predicted (true positive or true negative) samples were gathered. They became the learning set for SVM-2 [for details see Schapire, 1990]. Next, the N samples that were predicted differently by SVM-1 and SVM-2 were collected and these became the learning set for SVM-3. The predictions were decided according to the majority of SVM-1, SVM-2, and SVM-3 predictions. Using this method, ten-fold cross validation was carried

Table 2.5: The recall and precision of each feature vector

Data-type: feature vector ¹	recall ² (%)	precision ³ (%)	success rate at whole sequence ⁴ (%)	success rate at surface ⁵ (%)
mix ⁶ : whole sequence ⁷ (window 11)	28.8 (22.3) ¹³	26.4 (20.0)	69.1 (66.6)	63.5 (60.9)
mix: whole sequence + boosting by filtering (window 5)	28.8 (21.9)	27.0 (20.0)	69.1 (66.9)	63.7 (61.0)
mix: whole sequence + actual interaction site ratio (window 5)	35.2 (20.0)	35.8 (20.0)	74.0 (68.0)	68.0 (61.8)
mix: whole sequence + prediction interaction site ratio (window 5)	28.3 (18.3)	30.7 (20.0)	72.4 (69.0)	65.6 (62.7)
mix: sequence at surface (window 11)	39.6 (30.4)	40.2 (30.4)	-	63.2 (57.6)
mix: sequence + ASA (window 9)	41.5 (23.3)	54.9 (30.4)	-	71.4 (60.5)
mix: spatially neighboring ⁸ + ASA (15 residues)	44.6 (24.5)	56.1 (30.4)	-	71.0 (60.0)
mix: spatially neighboring ⁸ + ASA + actual interaction site ratio (9 residues)	50.4 (26.8)	58.1 (30.4)	-	73.5 (59.1)
mix: spatially neighboring ⁸ + ASA + predicted interaction site ratio (9 residues)	42.8 (22.3)	57.8 (30.4)	-	73.3 (60.9)
mix: sequence + ASA + Flatness (window 11)	43.2 (24.3)	55.8 (30.4)	-	70.1 (60.1)
hetero ⁹ : sequence + ASA (window 9)	45.0 (26.9)	55.9 (32.8)	-	69.7 (57.9)
homo ¹⁰ : sequence + ASA (window 9)	40.3 (21.0)	55.8 (28.9)	-	73.4 (62.2)
Hetero-mixed ¹¹ : sequence + ASA (window 9)	42.4 (24.1)	54.9 (32.8)	-	71.2 (58.9)
Homo-mixed ¹² : sequence + ASA (window 9)	38.4 (21.2)	55.0 (28.9)	-	72.0 (62.1)
feature vector ¹ = input feature vector of SVM recall ² = True_Positive/(True_Positive+False_Negative), precision ³ = True_Positive/(True_Positive+False_Positive) The “success rate at whole sequence ⁴ ” and the “success rate at surface ⁵ ” mean the average per residue prediction (interaction site or non-interaction site) accuracy				

out. The number of learning samples for each SVM with boosting (ten-fold cross validation) was set to be the almost the same as that for SVMs without boosting (leave 1/3 data set cross validation).

Problem Definition

3.1 NECESSITY OF PROMOTER IDENTIFICATION

A promoter is a signal element on a DNA molecule that specifies controlling region of a gene where RNA polymerase binds to initiate the transcription of the gene. RNA polymerase II (RNA pol II) in Eukaryotic cell binds the promoter signals of all protein coding sequences. But there is no universal occurrence of these signals. So, it is not possible to predict the promoter efficiently using some mere predetermined signals. An approach that will consider as many signals as possible which are not mutually exclusive will increase the chance of finding a promoter in a given sequence significantly.

A number of methods for the prediction of promoters, TSS (Transcription Start Signals) and TF (Transcription Factor) binding sites in eukaryotic DNA sequence presently exist [34,35]. Although contemporary algorithms have much elevated predictive ability than the prior approaches, it is almost certainly fair-haired to state that performance is still out lying from satisfactory.

Many general purpose promoter prediction implementations of primary level could recognize just ~50% of the promoters with a false positive (FP) rate of ~1 per 700–1000 bp [35]. Then the use of Markov chain in promoter prediction tools by Ohler et al. [36] improved the results slightly but they acknowledged the same 50% of promoters from the dataset analyzed by Fickett and Hatzigeorgiou [35], while having a false positive prediction rate of 1/ 849 bp. An additional promoter identification program, Promoter 2.0, designed by Knudsen [37] applied a combination of neural networks and genetic algorithms.

After the human genome had been sequenced, the efficiency of promoter prediction tools faced a major challenge. Promoter Inspector program [38] was the first software tool used to identify the promoters in human chromosome 22. It could identify ~50% of known promoters as genomic regions up to 1 kb in length by discriminating them from the exon, intron and 3' untranslated region (3'UTR) sequences. Recently, Bajic et al. [39] has reported the Dragon Promoter Finder (DBF) program, which uses sensors for three functional regions: promoters, exons and introns. Judging by the authors' estimates, that approach has a higher accuracy than three other promoter finding programs which it was compared: NNPP 2.1 [40] Promoter2.0 [37] and Promoter Inspector [41]. Another tool developed by Down and Hubbard [38] reported a novel hybrid machine learning method capable of predicting >50% of human TSS with a specificity of >70%.

Moreover, one promoter prediction tool TSSPTCM has been trained and adapted for plants [42].

In the present study, with the aid of PromMachine machine learning tool, promoters are distinguished from the non-promoter sequences on the basis of abundance of some

characteristic 4mer motifs. The PromMachine is trained with 128 distinguishing 4mers that can discriminate between promoter and non promoter. Using this knowledge the machine learning tool can efficiently decide whether a given sequence contains a promoter or not. With high sensitivity, specificity and accuracy this approach promises very high efficiency in promoter prediction in Eukaryotic genomic sequences. Being applicable for any reasonable length of given sequences this approach becomes a dynamic tool for finding promoters.



Proposed Method and Data Preparation

4.1 DATA PREPARATION

The promoter sequences are obtained from the Eukaryotic Promoter Database (EPD) [43] as positive dataset. We have used 1800 different vertebrate promoter sequences of length 300 bps contained in EPD Rel.65 and covering the region of 300 bps upstream of the transcription start site (TSS). We also collected a set of non overlapping human gene sequences of length 250 bps each as negative dataset, from the GenBank [44]. The training of this system is made on these data.

Dataset

There were two datasets used in the development and testing of the promoter recognition algorithm: (1) the plant promoter sequence database and (2) the non-promoter sequence database.

Promoter sequence database

To accomplish the task of RNA polymerase II promoter prediction, the plant promoter dataset was taken from PlantProm database [45] which contains plant promoter sequences. The database serves as learning set in developing plant promoter prediction programs. A total of 305 entries of plant promoter sequence with window size of 200 bp upstream and 51 downstream of TSS were obtained from PlantProm DB. To make the model applicable to predict promoters in *Drosophila*, Human, Mouse and Rat, promoter sequences are collected from EPD as training data. The test data are collected from the same source but these are independent of the training data.

Non-promoter sequence database

The non-promoter sequence database is extracted from Unigene database that belongs to EMBL. CDS are the best non promoter sequence. So, for every organism specific model, equal number of CDS sequence as the promoter sequence are used in the training process.

4.2 PROPOSED METHOD TO IDENTIFY PROMOTER

In this work I tried to find out the transcriptional elements which appear more frequently in promoters and less frequently in the non promoters. Promoter identification method can be divided into two categories, one knowing the absolute position of Transcriptional Start site and the other without any information about the absolute position of Transcriptional Start Site. The first method is relatively simple as one knows that the promoter sequences generally lies within the upstream of a Gene, but the later one is much more complex as one doesn't know the exact position of Gene. So to use different positional features such as TATA box and CpG Island are inherently difficult in these scenarios. My method doesn't take the absolute position of Transcriptional Start Site (TSS) into consideration. As a result I depended on the statistical analysis of different features in known promoter sequences.

Promoter and non promoter both contain A, C, T, G Taking this 4 nucleotide tetramer can be created. To develop a learning model we applied inductive inference where a model is derived from data and this model is further applied on new data. We used Support Vector Machine (SVM) to develop this model as historically SVM is proved to be better than other techniques in the analyzing biological data [46].

My method is described as below:

Step 1: Find all possible combination of ‘A’, ‘T’, ‘C’ and ‘G’ taking all four at a time, this creates in total 256 different combinations.

Step 2:

- i. Find f_{ij} where f_{ij} is the frequency of i^{th} combination in j^{th} known promoter sequence.
- ii. Find fn_{ij} where fn_{ij} is the frequency of i^{th} combination in j^{th} sequence.

Step 3:

i. Calculate $P_i = \sum_{i=1}^{256} \sum_{j=1}^n f_{ij}$ where n is the total number of promoter

ii. Calculate $NP_i = \sum_{i=1}^{256} \sum_{j=1}^n fn_{ij}$ where n is the total number of non – promoter

Step 4: Calculate the absolute difference between the number of occurrences of those 256 possible combinations of nucleotides in known promoter and non-promoter sequences. So, we took $Diff_i = | P_i - NP_i |$, here $Diff_i$ is the absolute difference of the occurrence of a certain sequence in known promoter and non promoter.

Step 5: Sort $Diff_i$, in descending order and took 128 combinations of nucleotides for which the absolute difference is maximum.

Step 6: Use the number of occurrences of these 128 combinations of nucleotides in the particular promoter and non-promoter sequence as a feature to train SVM on known promoter and non-promoter sequences.

4.3 SVM MATHEMATICAL MODEL

SVM, a supervised machine learning technique has been used for discriminating between promoter and non-promoter sequences. SVM classifiers solve multiclass classification problems using the structural minimization principle. Given a training set in a vector space, SVMs can find the best decision hyperplane, which separates two classes [47].

For a typical learning task $P(X, y) = P(y | X) P(X)$, an inductive SVM learner aims to build a decision function

$f_L: X \rightarrow \{-1, +1\}$ based on a training set S_{train} , which is
 $f_L = L(S_{\text{train}})$

Where:

$S_{\text{train}} = (X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$

SVM are trained in a supervised manner on a collection of promoter and non-promoter sequences.

It is necessary to select a kernel function and the regularization parameter in each Binary Classifier. Radial Basis Function (RBF) is selected as the kernel function.

The SVM classification problem can be formulated in terms of a convex quadratic optimization problem as

$$Max \left[\sum_{i=1}^N \alpha_i - \left(\frac{1}{2} \right) \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right] \dots\dots\dots (1)$$

In the above equation N is the total number of input vectors, x_i is any real number as the input vectors and y_i be their corresponding target class, which is either -1 or 1 in binary classifier.

A radial basis function (RBF) is a real valued function whose value depends only on the distance from the origin, so that

$$\sigma(x) = \sigma(\|x\|) \dots\dots\dots (2)$$

or alternatively on the distance from some other point c , called a center, so that

$$\sigma(x, c) = \sigma(\|x - c\|) \dots\dots\dots (3)$$

Radial basis functions are typically used to build up function approximations of the form

$$y(x) = \sum w_i \sigma(\|x - c_i\|) \dots\dots\dots (4)$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions, each associated with a different center c_i , and weighted by an appropriate coefficient w_i .

4.4 TRAINING WITH SVM

The frequencies (f_i) of these 128 characteristic 4-mer motifs are used to find the promoter (or Non-promoter) in a given sequence. As the values of f_i are dependant on the length of a sequence, the values of f_i will be different depending on the length of the sequence and this complicates the comparison between two heterogeneous sequences. To solve the problem a new parameter, d_i is defined by normalizing the f_i with the following equation

$$d_i = \{f_i - \min (f_1, f_2, \dots\dots\dots f_{128})\} / \max (f_1 \dots\dots\dots f_{128}) \dots\dots\dots(\text{equ5})$$

Here, the numerical value of d_i ranges from 0 to 1.

To use these features for promoter prediction, SVM the most perfect supervised learning algorithm, is trained. For this purpose, equal numbers of promoter and non-promoter sequences, collected from databases mentioned above are used as the training dataset. Being trained with the frequency patterns of the 128 characteristic 4-mer motifs in known promoters and non-promoters, a model is built that can distinguish between promoter and non promoter in test sequences.

Testing:

To substantiate the machine learning model, jackknife validation could be the ideal process. But it is more time consuming and so not used in this model. Instead a 7fold cross validation is performed. Then, prediction test is done using independent test data.



Experimental Results and Comparative Analysis

5.1 PREDICTION ACCURACY

In order to present the significance of n-mer sequences, a model is built using 305 promoter and 305 non-promoter sequences.

In order to test the prediction accuracy of the proposed model 100 sequences are selected randomly from Plant Prom DB and form EMBL that constitutes with 50 known promoter sequences and 50 known CDS. These dataset are completely independent from the training set (Table 5.1).

Table 5.1: Prediction done using the proposed model

Predicted Sequences	Total no. of Sequences	True Positive	False Positive	False Negative	True Negative	Sensitivity ^a	Specificity ^b
Promoter	50	43	Nil	7	Nil	0.86	0.9
Non-Promoter	50	Nil	5	Nil	45		

Sensitivity^a = $100 \cdot TP / (TP + FN)$

Specificity^b = $100 \cdot TN / (TN + FP)$

True Positive: When SVM detect a promoter as a promoter during training.

True Negative: When SVM detect a non-promoter as a non-promoter during training.

False Positive: When SVM detect a non-promoter as a promoter during training.

False Negative: When SVM detect a promoter as a non-promoter during training.

The model confidently predicts the promoters in these test data as it is less prone to false positive and false negative prediction. The model shows a high level of sensitivity and specificity (Table 5.1). The primary experimental results are summarized in Table 5.2 in which percentage of correct value (for both cross validation test and prediction test) and correlation coefficient value is given.

Table 5.2: Result of Model built for promoter prediction

Input data: Promoter and non-promoter sequences	Correctly classified instances on cross-validation data (%)	Correctly classified instances on 100 test dataset (%)	Classifier used for the proposed Algorithm	Correlation coefficient
Proposed Model	84.92	89	SVM	0.77

The correlation coefficient is also very high. When the proposed model is trained with the 128 discriminating 4-mer sequences of Human, Drosophila, Mouse and Rat and applied to predict the promoter in test data of corresponding organisms, we find very satisfactory results again in Table 5.3.

Table 5.3: Cross validation accuracy of the model for various organisms

Proposed Algorithm for	Cross validation accuracy (7 fold)
Plant	83.81 %
Drosophila	94.82 %
Human	91.25 %
Mouse	90.77 %
Rat	82.35 %

The 7 fold cross validation accuracies are very significant in all the test organisms. It indicates the universality of the model for all eukaryotes. The high percentage of correct value, correlation coefficient for this proposed model clearly indicates that calculated frequencies of 4-mer sequences are capable of discriminating between promoter and non-promoter regions.

Complexity of the proposed model is $O(4^{dM}N)$

5.2 COMPARISON WITH EXISTING METHODS

There are various algorithms used for promoter prediction. Using these algorithms, some widely used promoter prediction tools have been developed, e.g. Soft Berry, Dragon Promoter Finder, Neural Network Promoter Prediction, Promoter 2.0 Prediction Server and Promoter Scan.

However, the model proposed here promises even better performance than the most successful tools of this day. The results shown in Table 5.4 clearly indicate that the prediction accuracy of the model is relatively very high in comparison with other tools.

Table 5.4: Program accuracy: comparisons with existing methods

Program Name	NNPP (threshold 0.8)	Soft Berry (TSSP)	ProScan version 1.7	Dragon Promoter Finder version 1.4	Promoter 2.0 Prediction Server	Prom-Machine
Sensitivity (%) ^a	68	88	0	12	0	86
Specificity (%) ^b	76	90	100	100	78	90
Correlation coefficient ^c	0.44	0.78	**	0.25	**	0.77

** Infinity

$${}^c \text{Correlation coefficient} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

There are a number of superior features that have made our proposed model better performer than any other tools such as the use of large numbers (128) of discriminating features between promoters and non promoters and the exploitation of the most successful supervised learning system SVM. Other prominent promoter prediction tools use either statistical approach or Neural Network. However, SVM has outperformed both of these approaches in pattern matching and supervised learning. Besides, these approaches use only limited numbers of features as the promoter signals such as TATA box or Inr etc. On the other hand, the proposed model uses SVM to detect the most number of signals to date to declare whether a sequence is promoter or not. These superior components of my approach have made it a better tool.



Conclusion & Further Scope

6.1 CONCLUSION

The triumphant prediction of promoters with high accuracy using frequency distribution of n-mer sequences noticeably designates that the novel method has an assurance as an approach for successful Eukaryotic promoter prediction. The principal objective of this project was to develop an efficient tool that can discriminate between promoter and non-promoter in an unknown sequence with proper accuracy. The highly accurate results of promoter prediction with our approach in Plant, Human, Drosophila, Mouse and Rat (Table 5.3) have clearly proved the validity of using frequency distribution of 4-mers in discrimination between promoter and non promoter.

However, though the approach is very much efficient in predicting the presence of promoter in a given sequence, it cannot locate the position when TATA box is not present. But this challenge will be met very soon if other signals can be characterized for specific position. So, it is expected that the approach proposed here would be a highly useful and efficient tool to meet the demand of the molecular biologists.

6.2 FURTHER SCOPE

This paper focused on identifying promoter sequences. Same algorithm can be applied to identify the GENE sequence within a DNA. Combining the novel concept of identifying promoter with the gene identification will surely be of high research interest in the biological research domain. Our future work will try to focus on this research area.



- [1]. (<http://old-www.idiap.ch/learning/SVMTorch.html>)
- [2]. Yutaka Suzuki, Tatsuhiko Tsunoda, Jun Sese, Hirotohi Taira, Junko Mizushima-Sugano, Hiroko Hata, 1 Toshio Ota, 6 Takao Isogai, Toshihiro Tanaka, Yusuke Nakamura, Akira Suyama, Yoshiyuki Sakaki, Shinchi Morishita, Kousaku Okubo and Sumio Sugano, "Identification and Characterization of the potential promoter regions of 1031 kinds of Human genes" 11:677-684 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01; www.genome.org.
- [3]. A.G. Pederson, P. Baladi, S. Brunak and Y. Chauvin, "Characterization of Prokaryotic and Eukaryotic promoters using Hidden Markov Model" International Proceeding of the Third International Conference on International System on Molecular Biology.
- [4] A. G. Pederson and J. Engelbrecht, "Investigations of Escherichia coli promoter sequences with artificial neural network: New signals discovered upstream of the transcriptional start point." Proceedings of the Third International Conference on International System on Molecular Biology.
- [5]. B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, 1992. ACM Press.
- [6]. Lifton RP, Goldberg ML, Karp RW, Hogness DS, (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harb Symp Quant Biol. 42, 1047-1051.
- [7]. Smale ST, Kadonaga JT (2003). The RNA Polymerase II core promoter. Annu Rev Biochem. 72, 449-479.
- [8]. S. Audic And J. M. Claverie, "Visualizing the competitive recognition of TATA-boxes in vertebrate promoters", trends Genet, Vol. 14, pp. 10-11, 1998.
- [9]. D.S. Prestridge, "Predicting pol II promoter sequences using transcription factor binding sites", J. Mo. Biol.
- [10]. Fatemi, M; Pao, MM, jeong,S, Gal-Yam, EN, Egger, G, Weisenberger, DJ, Jones, PA. (November Nov 27; 33(20): e 176). "Vectors and delivery systems in gene therapy". Medical Science Monitor 33 (20). Retrieved on 2006-06-28.
- [11]. Saxonov, S; Berg, P, Brutlag, DL (January 2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters PNAS 103 (5)".
- [12]. Feil, R; Berger, F. (May 2007). "Convergent evolution of genomic imprinting in plants and mammals". Trends in Genetics 23 (4).
- [13]. T. Batsuda, H. Motoda and T. Washio, "graph-based induction and its application". Advanced Engineering informatics, Bol. 16, pp. 135-143, 2002.
- [14]. Eric C. Rouchka University of Louisville.
- [15]. Lehninger Principles of Biochemistry, Fourth Edition - David L. Nelson, Michael M. Cox
- [16]. Population genetics study: Hobbs, K.; Negri, J.; Klinnert, M.; Rosenwasser, L.J.; and Borish, L. (1998). Interleukin-10 and transforming growth factor-beta

Reference

- promoter polymorphisms in allergies and asthma. *Am J Respir Crit Care Med.* 158 (6), 1958-1962. PMID 9847292
- [17]. Kulozik, A.E.; Bellan-Koch, A.; Bail, S.; Kohne, E.; and Kleihauer, E. (1991). Thalassemia intermedia: moderate reduction of beta globin gene transcriptional activity by a novel mutation of the proximal CACCC promoter element. *Blood.* 77 (9), 2054-2058. PMID 2018842.
- [18]. Smale ST, Kadonaga JT (2003). The RNA polymerase II core promoter. *Annu Rev Biochem.* 72, 449-479. PMID 12651739 PDF.
- [19]. population genetics study: Burchard, E.G.; Silverman, E.K.; Rosenwasser, L.J.; Borish, L.; Yandava, C.; Pillari, A.; Weiss, S.T.; Hasday, J.; Lilly, C.M.; Ford, J.G.; and Drazen, J.M. (1999). Association between a sequence variant in the IL-4 gene promoter and FEV(1) in asthma. *Am J Respir Crit Care Med.* 160 (3), 919-922. PMID 10471619.
- [20]. Petrij F, Giles RH, Dauwerse HG, Saris JJ, Hennekam RC, Masuno M, Tommerup N, van Ommen GJ, Goodman RH, Peters DJ, et al. (1995). Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature.* 376 (6538), 348-351. PMID 7630403.
- [21]. Levine M, Tjian R (2003). Transcription regulation and animal diversity. *Nature.* 424(6945), 147-151. PMID 12853946 PDF.
- [22]. Pearson H (2006). "Genetics: what is a gene?". *Nature* 441 (7092): 398-401. PMID 16724031.
- [23]. Elizabeth Pennisi (2007). "DNA Study Forces Rethink of What It Means to Be a Gene". *Science* 316 (5831): 1556-1557.
- [24]. see e.g. Martin Nowak's *Evolutionary Dynamics*
- [25]. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007). "What is a gene, post-ENCODE? History and updated definition". *Genome Research* 17 (6): 669-681. PMID 17567988.
- [26]. Cavalier-Smith T. (1985). Eukaryotic gene numbers, non-coding DNA, and genome size. In Cavalier-Smith T, ed. *The Evolution of Genome Size* Chichester: John Wiley.
- [27]. International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome.". *Nature* 431 (7011): 931-45. PMID 15496913.
- [28]. Pennisi, Elizabeth (2007). "Working the (Gene Count) Numbers_ Finally, a Firm Answer". *Science* 316 (5828): 1113.
- [29]. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R (2004). *Molecular Biology of the Gene*, 5th ed., Peason Benjamin Cummings (Cold Spring Harbor Laboratory Press). ISBN 080534635X.
- [30]. Mark B. Gerstein et al., "What is a gene, post-ENCODE? History and updated definition," *Genome Research* 17(6) (2007): 669-681

- [31]. Min Jou W, Haegeman G, Ysebaert M, Fiers W (1972). "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein". *Nature* 237 (5350): 82-8. PMID 4555447.
- [32]. (<http://old-www.idiap.ch/learning/SVMTorch.html>)
- [33]. "Prediction of protein-protein interaction sites using support vector machines" Asako Koike and Toshihisa Takagi, December 2003.
- [34]. Bucher,P.: Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 1990, 212, 563–578.
- [35]. Fickett,J.W. and Hatzigeorgiou,A.C.: Eukaryotic promoter recognition. *Genome Res.* 1997, 7, 861–878.
- [36]. Ohler,U., Harbeck,S., Niemann,H., Noth,E. and Reese,M.: Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* 1999, 15, 362–369.
- [37]. Knudsen,S.: Promoter 2.0: for recognition of Pol II promoter sequences. *Biotechnologies* 1999, 15, 356–361.
- [38]. Scherf,M., Klingenhoff,A., Frech,K., Quandt,K., Schneider,R., Grote,K., Frisch,M., GailusDurner, V., Seidel,A., BrackWerner, R. and Werner,T.: First pass annotation of promoters of human chromosome 22. *Genome Res.* 2001, 11, 333–340.
- [39]. Bajic,V.B., Seah,S.H., Chong,A., Zhang,G., Koh,J.L.Y. and Brusica,V.: Dragon promoter finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* 2002, 18, 198–199.
- [40]. Reese,M., Harris,N.L. and Eeckman,F.H. (1996) Large scale sequencing specific neural networks for promoter and splice site recognition. In Hunter,L. and Klein,T.E. (eds), *Biocomputing Proceedings of the 1996 Pacific Symposium*, 2–7 January, World Scientific Co., Singapore.
- [41]. Scherf,M., Klingenhoff,A. and Werner,T.: Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* 2000, 297, 599–606.
- [42]. Shahmuradov,I. A., Solovyev,V. V. and Gammerman1,A.J.: Plant promoter prediction with confidence estimation *Nucleic Acids Research* 2005, Vol. 33, No. 3 1069–1076.
- [43]. <http://www.epd.isb-sib.ch/>
- [44]. <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>
- [45]. Shahmuradov, I. , Gammerman,A., Hancock,J.M., Bramley, P.M. and Solovyev,V.V.: PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.* 2003, 31, 114–117.
- [46]. Nikola Kasabov and Shaoning Pang, (May 2004) "Transductive Support Vector Machines and Applications in Bioinformatics for Promoter Recognition" *Neural Information Processing - Letters and Reviews*, Vol. 3, No. 2.

- [47]. Yuan,X., Buckles,B.P. and Zhang,J.: A comparison study of decision tree and SVM to classify gene sequence. Electrical Engineering and Computer Science Department, Tulane University; 2003.



Figure No.	Description	Page No.
Figure 1.1	The promoter region in a DNA sequence	1
Figure 1.2	The central dogma of molecular biology	2
Figure 2.1	Different fields of Bioinformatics	6
Figure 2.2	Structure of an animal cell	7
Figure 2.3	Karyotype showing the 23 pairs of human chromosomes	8
Figure 2.4	Snap of a DNA sequence within a cell	9
Figure 2.5	DNA double helix structure	10
Figure 2.6	Secondary structure for E. coli RNase P RNA	10
Figure 2.7	mRNA processing	11
Figure 2.8	tRNA secondary structure	12
Figure 2.9	tRNA tertiary structure	12
Figure 2.10	Protein Structure	13
Figure 2.11	Central Dogma of Molecular Biology	14

List of Tables

Table No.	Description	Page No.
Table 2.1	Amino Acid Codes	12
Table 2.2	Chromosomes, Genes and Genome Sizes in different Organisms	15
Table 2.3	Genetic Code	17
Table 2.4	Components Required for the Five Major Stages of Protein Synthesis in E. coli	19
Table 2.5	The recall and precision of each feature vector	25
Table 5.1	Prediction done using the proposed model	32
Table 5.2	Result of Model built for promoter prediction	32
Table 5.3	Cross validation accuracy of the model for various organisms	33
Table 5.4	Program accuracy: comparisons with existing methods	33

