Chapter 3: Methodology

The number and sizes of genome databases have grown rapidly over the last few years. A huge amount of information requires new ways for processing them and using them in efficient ways. Two of the most important problems are classification and clustering of genomes, i.e., automatically determining the group to which a previously unseen genome sequence belongs and grouping the genome sequences into a tree structure according to their similarity. For example, distinguishing virus subspecies, strains and isolates is important in vaccine development, diagnostics, and other fields of biological and medical research and practice. Classification is the task of automatically determining the class of an unseen object, based typically on a model trained on a set of objects with known class memberships (Andrija et al., 2006).

3.1 Research Design:

The positive and negative sequences (promoters and non-promoters) are downloaded from the curated databases available on the Internet. Table 3.1 describes the name of database, name of the species and number of sequences considered. These sequences are converted into numeric profile (feature vector) using the composition of the nucleotides, physical, chemical and electrical properties of the DNA sequence. The variations of these features in the positive and negative sequences were studied. As the features considered are more, to avoid curse of dimensionality, dimensionality of the feature vector is reduced using dimensionality reduction methods. After dimensionality reduction, using Grey Relational Analysis method a classifier is applied to classify the given target sequences. Performance of the classifier is analyzed through confusion
matrix analysis, specificity, sensitivity, precision and through Receiver Operating Characteristics (ROC) Curve.

Clean annotations from well-studied regions in genomes are absolutely essential in considering the data sets. In order to accomplish the task of promoter prediction, positive datasets and negative datasets were taken from different curated databases available on the World Wide Web. These databases contain non-redundant collection of promoters, for which the transcription start site has been determined experimentally. Table 3.1 describes the database name, species and number of sequences in that class, that are used in this study.

<table>
<thead>
<tr>
<th>Class</th>
<th>Database</th>
<th>Taxonomic group/organism</th>
<th>No. of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EPD</td>
<td>Homo Sapiens</td>
<td>608</td>
</tr>
<tr>
<td>2</td>
<td>Essential Genes</td>
<td>Homo Sapiens</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>726</td>
</tr>
</tbody>
</table>

Table: 3.1 Details of dataset 1

3.2 Data Format:

The primary structure of DNA i.e., sequence of nucleotides are retrieved in FASTA format.

>EP17030 (-) Hs snRNA U1 (pU1-6); range -499 to 100.
GCGGAGGTGCAAGGAGCCGAGATCATGCAGCTGCACTTCAGCCTGGCCAGAGAGCGAG
ACGGCCGAAAACAAAACAAAACAAAACAAAACAAAACAAAACAAAACAAAACAAAACAAAACAA
GACCATGGCAGGGGAGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGA
CTGGCCACCATGTGGTGCCGTAGTGCCGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGA
AGGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG

>EP17031 (-) Hs snRNA U1 (pHU1-1); range -499 to 100.
TTTTCTCTGGATATCCGAAACATTCTAGTCTGCGAATTAAAAGCCATTATTTGAAGAAGGAG
TGCCCCGGCTCCATCTGGCCACCGAAAGGTTGCTCCTTAACACAGGCTAAGGACCAGCTT59
Fig. 3.1 The primary structure of DNA (http://epd.vital-it.ch/)

3.3 Data Validity:

As the DNA is represented using IUPCA codes (Table 2.1), it is ensured that the DNA sequences considered in this research are consisting of the nucleotide bases Adenine(A), Cytosine(C), Guanine(G) and Thymine(T) only.

3.4 Procedure:

In genetics, a promoter is a region of DNA that facilitates the transcription of a particular gene. Promoters are located near the genes they regulate, on the same strand and typically upstream (towards the 5' region of the sense strand).
The five prime untranslated region (5' UTR) starts at the +1 position, where transcription begins, the Transcription Start Site (TSS), and ends one nucleotide before the start codon (usually AUG) of the coding region. The 5' UTR, also known as the leader sequence, is a particular section of messenger RNA (mRNA) and the DNA that codes for it.

Hence the positive sequences (promoters) are collected from -499 to 100 on the upstream of the positive strand. For non promoters, sequences of length 600 were collected from the inter-genetic region of the genes from the respective databases as shown in table 3.1.

3.4.1 Compositional Analysis of Nucleotides:

All the positive sequences were aligned against the TSS. A Position specific Score Matrix (PSM) is prepared. A PSM is a matrix of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol of the alphabet, and one column for each position in the pattern. The score assigned by a PSM to a substring $S = (S_j)_{j=1}^N$ is defined as $\sum_{j=1}^N m_{s_j,j}$, where $j$ represents position in the substring, $s_j$ is the symbol at position $j$ in the substring, and $m_{a,j}$ is the score in row $a$, column $j$ of the matrix. In other words, a PSM score is the sum of position-specific scores for each symbol in the substring.

A PSM assumes independence between positions in the pattern, as it calculates scores at each position independently from the symbols at other positions. The score of a substring aligned with a PSM can be interpreted as the log-likelihood of the substring under a product multinomial distribution. Since each column defines log-likelihoods for
each of the different symbols, where the sum of likelihoods in a column equals one, the
PWM corresponds to a Multinomial distribution. A PSM’s score is the sum of log-
likelihoods, which corresponds to the product of likelihoods, meaning that the score of a
PSM is then a product-multinomial distribution. The PSM scores can also be interpreted
in a physical framework in identifying the promoter regions in DNA sequence. The PSM
scores are plotted against the corresponding positions.(Figure 3.2).
Fig. 3.2 Compositional Analysis of Nucleotides near Promoter Regions of Homo Sapiens

Fig. 3.3 Compositional Analysis of Nucleotides in Genes of Homo Sapiens
Fig. 3.4 Composition of Adenine near the promoter regions of Homo Sapiens

Fig. 3.5 Composition of Adenine in genes of Homo Sapiens
Fig. 3.6 Composition of Thymine near the promoter regions of Homo Sapiens

Fig. 3.7 Composition of Thymine in genes of Homo Sapiens
Fig 3.8 Composition of Guanine near the promoter regions of Homo Sapiens

Fig 3.9 Composition of Guanine in genes of Homo Sapiens
Fig. 3.10 Composition of Cytosine near the promoter regions of Homo Sapiens

Fig. 3.11 Composition of Cytosine in genes of Homo Sapiens
From the above graphs (Fig. 3.2 -3.1 l) it is observed that the properties of the promoter regions are considerably different from those of other parts of the genome.

As clear cut discrimination appears between the promoter region and non-promoter region through compositional analysis through position specific frequency matrix method, nucleotide composition plays an important role in detecting the promoter from the DNA sequence.

3.4.2 Physico-Chemical Properties:

As the DNA naturally exists in tertiary format, the physico-chemical properties that stabilize the DNA were also studied. Structural properties can be divided into two broad categories: those for which the numerical profile shows a peak around the TSS (e.g., GC content, bendability, and DNA denaturation), and those for which the profile shows a cleft around the TSS (e.g., AT content, bDNA twist, and Duplex Stability-Free Energy).

Base Stacking Energy:

The purine and pyrimidine bases are hydrophobic and relatively insoluble in water at the near-neutral pH of the cell. At acidic or alkaline pH the bases become charged and their solubility in water increases. Hydrophobic stacking interactions in which two or more bases are positioned with the planes of their rings parallel (like a stack of coins) are one of two important modes of interaction between bases in nucleic acids. The stacking also involves a combination of vander Waals and dipole-dipole interactions between the bases. Base stacking helps to minimize contact of the bases with water, and base-stacking interactions are very important in stabilizing the three-dimensional
structure of nucleic acids. Base Stacking energy values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions are shown in Fig. 3.12 and Fig. 3.13.

Fig. 3.12 Basestacking Energy near the promoter regions of Homo sapiens

Fig. 3.13 Basestacking Energy in genes of Homo sapiens
DNA Denaturation:

It is also called DNA melting, a process by which double-stranded deoxyribonucleic acid unwinds and separates into single-stranded strands through the breaking of hydrogen bonding between the bases (Fig. 3.14). Both terms are used to refer to the process as it occurs when a mixture is heated, although "denaturation" can also refer to the separation of DNA strands induced by chemicals like urea. The process of DNA denaturation can be used to analyze some aspects of DNA. Because cytosine/guanine base-pairing is generally stronger than adenosine/thymine base-pairing, the amount of cytosine and guanine in a genome (called the "GC content") can be estimated by measuring the temperature at which the genomic DNA melts.

![Denaturation of DNA](image)

Fig: 3.14 Denaturation of DNA

DNA denaturation can also be used to detect sequence differences between two different DNA sequences. DNA is heated and denatured into single-stranded state, and the mixture is cooled to allow strands to re-hybridize. Hybrid molecules are formed between similar sequences and any differences between those sequences will result in a disruption of the base-pairing. On a genomic scale, the method has been used by researchers to estimate the genetic distance between two species, a process known as DNA-DNA hybridization. DNA denaturation values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions are shown in Fig. 3.15 and Fig. 3.16.
B-DNA:

A right-handed double-helical conformation of DNA is normally seen in solution (Watson & Crick, 1953). A second DNA conformation (A form) is seen in unhydrated DNA (fibres or crystals of oligonucleotides). B-DNA also differs from other DNAs (i.e.,
DNA exists in many possible conformations that include A-DNA, B-DNA, and Z-DNA forms, although, only B-DNA and Z-DNA have been directly observed in functional organisms. The conformation that DNA adopts depends on the hydration level, DNA sequence, the amount and direction of supercoiling, chemical modifications of the bases, the type and concentration of metal ions, as well as the presence of polyamines in solution. BDNA Twist values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions are shown in Fig.3.17 and Fig. 3.18.
Z-DNA

Z-DNA was the first single-crystal X-ray structure of a DNA fragment (a self-complementary DNA hexamer d(CG)$_3$). It was resolved as a left-handed double helix with two anti-parallel chains that were held together by Watson-Crick base pairs. It was solved by Andrew Wang, Alexander Rich, and co-workers in 1979 at MIT (Wang et al., 1979.).
Z-DNA is adopted by short oligonucleotides that have sequences of alternating pyrimidines and purines. Early studies of oligonucleotides with alternating purine-pyrimidine sequences revealed the left-handed helical conformation of Z-DNA. This structure is characterized by alternating helical parameters and torsion angles with a 2-base pair periodicity, causing the backbone of the helix to zig-zag (hence the name Z-DNA). Zig-zagging is thus a consequence of the fact that the repeating unit is a dinucleotide (not a mononucleotide), especially a sequence in which pyrimidine alternate with purines, e.g., alternating C and G or 5-methyl cytosine and G residues. Although, alternating purine-pyrimidine tracts such as oligo-dGdC and oligo-dAdC provide a good substrate for Z-DNA, this sequence specificity is now known to be neither necessary nor sufficient for its formation. Methylation of C-5 of cytosyl residues in alternating CG sequences (e.g. CGCGCG) facilitates the transition of B-DNA to Z-DNA, because the added hydrophobic methyl groups stabilize the Z-DNA structure.

Z-DNA is formed when purine residue flip in syn conformation while alternating pyrimidine is in anti conformation. Phosphate groups of backbone are closer to each other as compared to that in A or B forms, hence high salt concentration is required to minimize electrostatic repulsion between the backbone phosphates. It contains one deep helical groove. The Z-DNA form occurs under physiological conditions in certain cases only. The biological role of Z-DNA is uncertain, however, its existence graphically shows that DNA is a flexible, dynamic molecule. Z-DNA structure tend to form in torsionally stressed DNA and are stabilized by dehydration, they may play an important role in control of gene expression. Fig. 3.19 depicts the common structural variants of DNA, i.e. A and Z, along with B form of DNA. Fig. 3.20 and 3.21 plots the ZDNA
values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions.

Fig. 3.19 Common structural variants of DNA

A DNA        B DNA        Z DNA

Fig. 3.19 Common structural variants of DNA
PROPELLOR TWIST:

DNA exhibits variation in its precise structure from base pair to base pair. This was revealed by comparison of the crystal structures of individual DNAs of different sequences. For example, the two members of each base pair do not always lie exactly in the same plane. Rather they can display a "propeller twist" arrangement in which the two
flat bases counter rotate relative to each other along the long axis of the base pair, giving the base pair a propeller-like character. Moreover, the precise rotation per base pair is not a constant. As a result, the width of the major and minor grooves varies locally. Thus, DNA molecules are never perfectly regular double helices, instead, their exact conformation depends on which base pair (A:T, T:A, G:C, or C:G) is present at each position along the double helix and on the identity of neighboring base pairs (Watson et al., 2006).

Fig. 3.22 Rotation of the bases about their long axis due to Propeller Twist

a. The structure shows a sequence of three consecutive A:T base pairs with normal Watson-Crick bonding (Fig. 3.22).

b. Propeller twist causes rotation of the bases about their long axis. (Aggarwal et al., 1988).

Fig. 3.23 & 3.24 plots the propeller twist values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions.
Fig. 3.23 Propellor Twist near the promoter regions in Homo sapiens

Fig. 3.24 Propellor Twist in genes of Homo sapiens
Bendability:

**Fig. 3.25** Bendability near the promoter regions of homo sapiens

**Fig. 3.26** Bendability in the intergenic region of homosapiens

**Fig. 3.27** Bendability in genes of homo sapiens

CPG ISLANDS:

In genetics, CpG islands are genomic regions that contain a high frequency of CpG sites. In mammalian genomes, CpG islands are typically 300-3,000 base pairs in length. They are in and near approximately 40% of promoters of mammalian genes. The
"p" in CpG refers to the phosphodiester bond between the cytosine and the guanine, which indicates that the C and the G are next to each other on the sequence strand be it a single or double stranded. For example, both C and G would be on the same strand connected by phosphodiester connection and not triple-bonded across two strands of DNA.

The usual formal definition of a CpG island is a region with at least 200 bp and with a GC percentage that is greater than 50% and with an observed/expected CpG ratio that is greater than 60%. Another recent study revised the rules of CpG island prediction in order to exclude other GC-rich genomic sequences such as Alu repeats. Based on an extensive search on the complete sequences of human chromosomes 21 and 22, DNA regions >500 bp with a GC content >55% and observed CpG/expected CpG of 0.65 were more likely to be the true CpG islands associated with the 5' regions of genes. CpG islands typically occur at or near the transcription start site of genes, particularly housekeeping genes, in vertebrates.

The regions of genes containing a relatively high concentration of CpG pairs, are collectively referred to along a chromosome as CpG islands, which typically vary in length from a few hundred to a few thousand nucleotides long. CpGs are very useful in genome mapping projects. They are often associated with the promoters of most housekeeping genes and many tissue-specific genes. Due to their involvement in the regulation of gene expression, they can be used as gene markers. Moreover, methylation of promoter-related CGIs is very common in all types of cancer cells, thus the hypermethylated CGIs in promoter regions can be used as molecular tumor markers, which makes the early detection of cancer possible. Consequently, identification of
potential CGIs helps not only to find the promoters of many specific genes but also to find candidate regions for aberrant DNA methylation and therefore contributes to the understanding of the epigenetic causes of cancer. CpG values of the DNA sequence of Homo sapiens near the promoter and non-promoter regions are shown in Fig.3.27 and Fig. 3.28.

Fig. 3.27 CpG Islands near the promoter regions of Homo Sapiens

Fig. 3.28 CpG Islands in genes of Homo sapiens
From the above graphs it is observed that the nature of the physico-chemical properties of the promoter regions are considerably different from those of other parts of the genome.

As clear cut discrimination appears between the promoter region and non-promoter region with respect to physico-chemical properties, they also play an important role in detecting the promoter from the DNA sequence.

3.5 Datasets

The database, organism and number of sequences of promoter and non-promoters of different species are shown in the table 3.2.

<table>
<thead>
<tr>
<th>Class</th>
<th>Database</th>
<th>Taxonomic group/organism</th>
<th>No. of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EPD</td>
<td>Homo Sapiens</td>
<td>608</td>
</tr>
<tr>
<td>2</td>
<td>DCPD</td>
<td>Drosophila Melanogaster</td>
<td>192</td>
</tr>
<tr>
<td>3</td>
<td>PromEC</td>
<td>E.Coli.</td>
<td>471</td>
</tr>
<tr>
<td>4</td>
<td>SCPD</td>
<td>Saccharomyces Cerevisiae</td>
<td>232</td>
</tr>
<tr>
<td>5</td>
<td>Essential Genes</td>
<td>Homo Sapiens</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>Essential Genes</td>
<td>Drosophila Melanogaster</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>Essential Genes</td>
<td>E.Coli</td>
<td>300</td>
</tr>
<tr>
<td>8</td>
<td>Essential Genes</td>
<td>Saccharomyces Cerevisiae</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>2111</td>
</tr>
</tbody>
</table>

Table: 3.2 Details of dataset 2

The Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection of experimentally characterized eukaryotic POL II promoters for which the transcription start site has been determined experimentally. It is also strictly non-redundant database. The general rule is that one entry corresponds to one transcription initiation site in a genome. Organisms are distinguished at the taxonomic level of the species. According to this policy, data from different literature sources pertaining to the
same transcription initiation sites are represented by the same entry. Likewise, promoters belonging to different alleles of the same gene, or to the same gene in different subspecies, are covered by the same entry regardless of whether they differ in sequence (Perier et al., 2000). EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets. Information presented in EPD results from an independent evaluation of primary experimental data shown in the biological literature. Sequences flanking transcription initiation sites are indirectly given by pointers to EMBL sequences. The annotation part of a promoter entry includes description of the promoter-defining evidence, cross-references to other databases, and bibliographic references. Being designed as a resource for comparative sequence analysis, EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets (Cavin et al., 1998).

PromEC is an updated compilation of Escherichia coli mRNA promoter sequences. It includes documentation on the location of experimentally identified mRNA transcriptional start sites on the E.coli chromosome, as well as the actual sequences in the promote region. The database is currently updated as of July 2000 and includes 472 entries. The added promoter sequences are based on three sources: the documentation in the E-coli genome file, the Regulon DB database, and an updated literature search (covering the years of 1993-2000). Promoters that were compiled following the literature search were those for which the transcriptional start sites were identified experimentally either by primer extension or by S1 nuclease mapping (Hershberg et al., 2001).
Drosophila Core Promoter Database (DCPD) Created by Alan K. Kutach and Scott Iyama in the laboratory of James T. Kadonaga. 205 Drosophila melanogaster core promoters aligned by their empirically determined transcription start site. Promoters were included only if their start site was determined by RNase protection, primer extension, or multiple clones from 5' RACE.

A Drosophila Promoter Database containing 252 independent Drosophila melanogaster promoter entries has been compiled. The database and its subsets have been searched for overrepresented sequences. The database and its subsets have been searched for overrepresented sequences. The analysis reveals that the proximal promoter regions play the most dramatic nucleotide sequence irregularities and exhibits a tripartite structure, consisting of TATA at -25/-30 bp, initiator (Inr) at +/-5 bp and a novel class of downstream elements at +20/+30 bp from the RNA start site.

However, they differ from TATA and Inr in several aspects. Viz., (1) they are represented not by a single, but by multiple sequences, (2) they are shorter, (3) their position is less strictly fixed with respect to the RNA start site, (4) they emerge as a characteristic feature of Drosophila promoters and (5) some of them are strongly overrepresented in the TATA-less, but not TATA-containing, subset. About one-half of known Drosophila promoters can be classified as TATA less. The overall sequence organization of the promoter region is characterized by an extended region with an increase in GC content and a decrease in A, which contains a number of binding sites for drosophila transcription factors (Irina, 1995).
In order to facilitate a systematic study of the promoters and transcriptionally regulatory cis-elements of the yeast Saccharomyces cerevisiae on a genomic scale, a comprehensive yeast-specific promoter database, SCPD was developed (Zhu & Zhang, 1999). Currently SCPD contains 580 experimentally mapped transcription factor (TF) binding sites and 425 transcriptional start sites (TSS) as its primary data entries. It also contains relevant binding affinity and expression data where available. In addition to mechanisms for promoter information (including sequence) retrieval and a data submission form, SCPD also provides some simple but useful tools for promoter sequence analysis.

SCPD is based on published results of individual genes. It can be used either to complement or substantiate large-scale genomic expression data. It provides more up-to-date information specific to yeast than other databases such as TRANSFAC (Wingender et al., 1996), TRRD (Heinemeyer et al., 1998), EPD (Cavin et al., 1998) and TFD (Ghosh, 1998). All sequences in SCPD refer to the corresponding genomic records in SGD (Cherry et al., 1998) and GENBANK (Benson et al., 1998). Since for most yeast genes the transcriptional start site (TSS) is not mapped, the location of a DNA element in the promoter region is defined relative to the translational start site (A of ATG is at +1) (Zhu & Zhang, 1999).

Database of Essential Genes (DEG 6.5), which contains all the essential genes that are currently available. Each entry of essential genes has a unique DEG identification number, gene reference number, gene function and sequence. A Database of Essential Genes (DEG) contains essential genes in more than 10 bacteria, such as E. coli, B. subtilis, H. pylori, S. pneumoniae, M. genitalium and H. influenzae, whereas for
eukaryotes, DEG contains those in yeast, humans, mice, worms, fruit flies, zebra fish and the plant A. thaliana (Zhang et al., 2004).
### ABF1, BAF1 (Retrieved Sequence)

<table>
<thead>
<tr>
<th>Get mapped sites</th>
<th>Get putative sites</th>
<th>Get intergenic region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve sequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start &lt;- ATG</td>
<td>-499</td>
<td>ATG -&gt; End 100</td>
</tr>
</tbody>
</table>

> YKL112W ABF1 225717 226316
TTAAAGTAAATAAATGGAAATAAATAGATGAAAAACTGTTTACACT
AATTATAACGTAAAAATACAGTATAATACAAATTATGCGCATGTA
AGCCCAATAGTCAGCCGTGGCTTTCTATCAAATTTACCAGGGCGGTCCC
CAATCAGTGTATCTGTACACGGAAAAGTCGTTAAAAGGTGATGGATTAG
CGGTTACTGTGCTGACATTTTTTTTCGCTTTTTTTCTGGAATAAAAACGA
AAAGTGCGCAAGCAACTTAATTACACCTTTTCTACTCTATTTCTCTGTCCTTTT
TGTTATTTGGCAATGGCCACTACACCACATCCCCATATAACGAAATTCAACAAA
AGTGTGTCAATTAAAATTACAGGTTAGTGACAGCAATTGTTAGATCTCA
ACTAGAGAGATTATCGATATCGCAAAGCGCTTTACACATATTTCGATGATCC
TGCTGCACATGGAAAAAGTCTGGAAAATCCCAATTATTTGCAATTTCACAAAGGA
TGAGAAATATAGCTGGAATTATTGCAAATACAGCAACCTATATAATTAA
AAAGACCTGGCCATTGGAGCCCATGGAGGCAAAAAATTTCCCACCTGG

### ABF1, BAF1 (Mapped sites)

<table>
<thead>
<tr>
<th>ORF</th>
<th>YKL112W</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE</td>
<td>ABF1, BAF1</td>
</tr>
<tr>
<td>XX</td>
<td></td>
</tr>
<tr>
<td>FACTOR</td>
<td>ABF1, BAF1</td>
</tr>
<tr>
<td>COORDINATE</td>
<td>(-172, -161)</td>
</tr>
<tr>
<td>SEQUENCE</td>
<td>TCCCCATTTACG</td>
</tr>
</tbody>
</table>

XX
> YKL112W ABF1
ABF1
TTGTTTATATGGCAATGGCCACTACACCACATCCCCATTTAAGGAATTCACCA -151
AAGTTGTATATTAATACAGTATTAGTGACAGCATTTTGGTTTAGATCTC -101
AAGTACAGAGGATATTACATCTGAAGCTTTTACACGTTTTATAGTATC -51
CTGCGTCAACTAGGAAAGTCTGAAAATTATCACCATATTTGCAATTTCACAAAGG -1
ATGGACAAATATAGCTGGAATTATTGCAAATACAGCAACCTATATAATTAA +50

---

Fig.3.29 Web interface to retrieve Saccharomyces cerevisiae:
http://rulai.cshl.edu/cgi-bin/SCPD/getgene?ABF1.BAF1
3.6 Feature Extraction From DNA Sequence:

n-mers (or x-mers where x can be virtually any consonant of choice) usually refer to specific n-tuples or n-grams of nucleic acid or amino acid sequences that can be used to identify certain regions within biomolecules like DNA (e.g. For gene prediction) or proteins. Either n-mer strings as such can be used for finding regions of interest, or n-mer statistics giving discrete probability distributions of a number of possible n-mer combinations are also used in this research.

Good input representations make it easier for the classifier to recognize underlying regularities. Therefore, good input representations are crucial to the success of classifier learning. The sequence or primary structure of a nucleic acid is the exact specification of its atomic composition, electrical properties and the chemical bonds connecting those atoms. From a one dimensional point of view, a DNA sequence contains characters from the 4-letter nucleic acid alphabet A,C,G or T.

Let a given DNA sequence either promoter or non-promoter be:

\[ S = s_1, s_2, s_3, \ldots s_{L-1}, s_L, \text{ where } s_i \in \{A,C,G,T\}, 1 \leq i \leq L. \]

The feature values of the sequence are calculated on the composition of the trimers. i.e., \( \Delta 3 = \{AAA, AAC, AAG, AAT, ACA, ACC, \ldots, TTT\} \).

The 64 compositional frequencies are calculated as:

\[ v_i = \frac{f_i}{|S|-2}, \quad 1 \leq i \leq 64 \quad --- \text{Eq. 3.1} \]

\( f_i \) denote the frequency of occurrence of the \( i^{th} \) feature and \( |S| \) denote the length of the sequence. The feature values \( v_i \) are normalized frequency counts (Nageswara Rao et al., 88).
Feature extraction is transforming the data in the high-dimensional space to a space of fewer dimensions.

**3.7 Feature Selection:**

To select the relevant features (i.e., n-mer) by employing a distance measure to calculate the relevance of each feature.

Let X be a feature and let x be its value. Let \( P(x|\text{Class} = 1) \) and \( P(x|\text{Class} = 0) \) denote the class conditional density function for feature X, where Class 1 represents the target class and class 0 is the nontarget class. Let \( D(X) \) denote the distance function between \( P(x|\text{Class} = 1) \) and \( P(x|\text{Class} = 0) \), defined as:

\[
D(X) = \int |P(x|\text{Class} = 1) - P(x|\text{Class} = 0)| \, dx
\]

--- Eq. 3.2

The distance measure performs feature X to feature Y if \( D(X) > D(Y) \). Intuitively, this means it is easier to distinguish between Class 1 and Class 0 by observing feature X than feature Y. That is, X approaches often in Class 1 and seldom in Class 0 or vice versa. In this work, each feature X is a 2-gram (3-gram). Let c denote the occurrence number of the feature X in a sequence S. Let l denote the total number of 2-grams in S and let \( l = \text{len}(S) - 1 \). Define the feature value x for the 2-gram (3-gram) X with respect to the sequence S as:

\[
X = \frac{c}{\text{len}(S) - 1}
\]

--- Eq. 3.3

Example: Suppose \( S = \text{AACCAAAGGTT} \). Then the value of feature AA with respect to S is \( 3/(11-1) = 0.33 \).
Because the DNA segment (Promoter or gene) considered is short with respect to Genome, random pairings may have effect on this result. Therefore $D(X)$ is approximated as:

$$D(X) = \frac{m_1-m_0^2}{d_1^2+d_0^2}$$  \quad \text{--- Eq. 3.4}$$

Where $m_1$ and $d_1$ ($m_0$ and $d_0$, respectively) are the mean value and the standard deviation of the feature $X$ in the positive (negative, respectively) training data set.

In the eq.3.4, the larger the numerator is (or the smaller the denominator is), the larger the interclass distance is and therefore the easier to separate class 1 from class 0 (and vice versa).

The mean value $m$ and the standard deviation $d$ of the feature $X$ in a set $T$ of sequences are defined as

$$M=\frac{1}{N}\sum_{i=1}^{N} x_i$$  \quad \text{--- Eq. 3.5}$$

$$d=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - m)^2}$$  \quad \text{--- Eq. 3.6}$$

where $x_i$ is the value of the feature $X$ with respect to sequence $S_i \in T$, and $N$ is the total number of sequences in $T$.

Now $N_g$ features with largest $D(X)$ values $X_1, X_2, \ldots, X_{N_g}, N_g<<(\text{number of 2-tuples / 3-tuples})$ are selected as the key-features.
To compensate for the possible loss of information due to ignoring the other features, a Linear Correlation Coefficient (LCC) between the all the features is calculated as is used as feature in the reduced set of features. The LCC is calculated as:

\[
\text{LCC} (S) = \frac{n \sum_{j=1}^{n} x_j m_j - \sum_{j=1}^{n} x_j \sum_{j=1}^{n} m_j}{\sqrt{n \sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2} \sqrt{n \sum_{j=1}^{n} m_j^2 - (\sum_{j=1}^{n} m_j)^2}} 
\]

---Eq.3.7

Where \(m_j\) is the mean value of the \(j^{th}\) k-tuple, \(1 < j < n\), in the positive training data set, and \(x_j\) is the feature value of the \(j^{th}\) k-tuple with respect to \(S\) as defined by Eq.3.3(Tan et al.,2006).

3.8 Grey Relational Analysis Procedure:

The black box is used to indicate a system lacking interior information ( Ashby, 1945). The black is represented as lack of information and white is used to represent full of information. Thus, the information that is either incomplete or undetermined, is called Grey. A system having incomplete information is called Grey system. The Grey number in Grey system represents a number with less complete information. The Grey element represents an element with incomplete information. The Grey relation is the relation with incomplete information. Those three terms are the typical symbols and features for Grey system and Grey phenomenon(Wu, 1996). There are several aspects for the theory of Grey system:

1. Grey generation: This is data processing to supplement information. It is aimed to process those complicate and tedious data to gain a clear rule, which is the whitening of a sequence of numbers.
2. Grey modeling: This is done by step 1 to establish a set of Grey variation equations and Grey differential equations, which is the whitening of the model.

3. Grey prediction: By using the Grey model to conduct a qualitative prediction, this is called the whitening of development.

4. Grey decision: A decision is made under imperfect countermeasure and unclear situation, which is called the whitening of status.

5. Grey relational analysis: Quantify all influences of various factors and their relation, which is called the whitening of factor relation.

6. Grey control: Work on the data of system behavior and look for any rules of behavior development to predict future's behavior, the prediction value can be fed back into the system in order to control the system.

Grey system theory is an interdisciplinary scientific area that was first introduced in early 1980s by Deng in 1982. Since then, the theory has become quite popular with its ability to deal with the systems that have partially unknown parameters. As a superiority to conventional statistical models, grey models require only a limited amount of data to estimate the behavior of unknown systems (Deng, 1989).

During the last two decades, the grey system theory has been developed rapidly and caught the attention of many researchers. It has been widely and successfully applied to various systems such as social, economic, financial, scientific and technological, agricultural, industrial, transportation, mechanical, meteorological, ecological, hydrological, geological, medical, military, etc., systems. Some research studies in financial area are as follows: In another study, the moving average autoregressive exogenous (ARX) prediction model is combined with grey predictors for time series prediction in Huang & Jane (2009), and it is proved that the hybrid method has a greater forecasting accuracy than the GM(1,1) method. Another study (Chang & Tsai, 2008)
introduces a support vector regression grey model (SVRGM) which combines support vector regression (SVR) learning algorithm and grey system theory to obtain a better approach to time series prediction. In these studies and the others, it is seen that grey system theory-based approaches can achieve good performance characteristics when applied to real-time systems, since grey predictors adapt their parameters to new conditions as new outputs become available. Because of this reason, grey predictors are more robust with respect to noise and lack of modeling information when compared to conventional methods.

Grey Relational Analysis:

Grey Relational Analysis (GRA) has been one of the most practical analytical tools (Deng, 1988; Liu & Lin, 2005; Nagai and Yamaguchi, 2004; Wen, 2004; Wen et al., 2006; Yamaguchi et al., 2006b). Several GRA models are developed and well summarized in (Liu & Lin, 2005; Wen et al., 2006; Yamaguchi et al., 2007; You et al., 2006). A supervised algorithm has been developed by using grey relational grade as the distance measure for data classification of electrocardiogram (ECG) heartbeat recognition and showed that it has achieved a good performance (Lin, 2006). The GRA models provide appropriate tools for examining a rank of order of multiple objects with resemblance from an objective. In the recent years, GRA models have been applied to a lot of applications, such as decision making in computer science (Akabane et al., 2005; Yamaguchi et al., 2006a), system modeling, social science, geometry, chemistry, management (Lin et al., 2009; Nagai et al., 2005; Rui & Wunshch, 2005; Yamaguchi et al., 2005; Yamaguchi et al. 2006a), economics, marketing research (Yamaguchi et al., 2004).
Owing to the usefulness and robustness of GRA, a new approach for predicting the promoter is proposed, based on GRA. This study applies GRA model to predict promoter from the data set of 2111 samples from 4 species. The results of the current model are compared to those of traditional and hybrid models, which include conventional logistic regression, logarithm logistic regression and ANN approaches. The result shows that in predicting the promoter, GRA model gave better performance and demonstrates stronger prediction power than conventional logistic regression and ANN approaches. A system that has no information is defined as a black system, while a system that is full of information is called white. Thus, when the information of a system is either incomplete or undetermined, it is defined as grey system. The grey number in grey system represents a number with incomplete information. The grey element represents an element with incomplete information. The grey relation is the relation with incomplete information. This section describes the basic definitions of grey relational analysis, GRA. The inner product and metric of two vectors are first defined. What follows are properties of norm space, grey relational space, grey relational grade for both globalized and localized grey relationships. Fig. 3.30 shows the work flow of the thesis.
Data collection from Various Sources

Trimer features

Physico-chemical features

Feature Extraction

Feature selection

Grey relational Generating

Reference Sequence Definition

Grey Relational coefficient Calculation

Grey Relational Grade Calculation

Class Determination

Fig. 3.30 Workflow of Thesis
3.9 Design of Classifier using Grey Relational Analysis:

The GRA includes local relation and global relation analysis. Grey relational-based classifier is used to classify promoter/non-promoters. GRA is a method to determine the relation of a discrete data to other sequence data (Chang, 2000; Wu and Chen, 1999). The novel grey relational procedure is introduced. Suppose the test sequence $\varphi_i(0)$, $i=1,2,3,...,n$, and $K$ comparative sequences $\Phi(k)=[\varphi_1(k), \varphi_2(k), \varphi_3(k), ..., \varphi_i(k), ..., \varphi_n(k)]$, $k=1, 2, 3, ..., K$, can be represented as

$$\Phi_{test}=[\varphi_1(0)\varphi_2(0)\varphi_3(0) ... \varphi_i(0) ... \varphi_n(0)]$$

--- Eq. 3.8

$$\Phi_{comp}=\begin{bmatrix}
\Phi(1) \\
\Phi(2) \\
\vdots \\
\Phi(k) \\
\vdots \\
\Phi(K)
\end{bmatrix} = \begin{bmatrix}
\varphi_1(1)\varphi_2(1)\varphi_3(1) ... \varphi_i(1) ... \varphi_n(1) \\
\varphi_1(2)\varphi_2(2)\varphi_3(2) ... \varphi_i(2) ... \varphi_n(2) \\
\vdots \\
\varphi_1(k)\varphi_2(k)\varphi_3(k) ... \varphi_i(k) ... \varphi_n(k) \\
\vdots \\
\varphi_1(K)\varphi_2(K)\varphi_3(K) ... \varphi_i(K) ... \varphi_n(K)
\end{bmatrix}$$

--- Eq. 3.9

Compute the absolute deviation of the test sequence $\Phi_{test}$ and $k$ comparative sequence $\Phi(k)$ by

$$\Delta \varphi_i(k) = [\varphi_i(0) - \varphi_i(k)]$$

--- Eq. 3.10

The deviation matrix $\Delta \Phi$ can be represented as
The grey relational grades $r(k)$ can be calculated as:

$$r(k) = \exp \left[ -\xi \left( \frac{ED(k)}{\Delta \varphi_{\text{max}} - \Delta \varphi_{\text{min}}} \right)^2 \right], \quad \xi \in (0, 25)$$  \hspace{1cm} \text{--- Eq. 3.12}$$

$$= \exp \left[ -\xi \left( \frac{\sqrt{\sum_{t=1}^{n} (\Delta \varphi_t(k))^2}}{\Delta \varphi_{\text{max}} - \Delta \varphi_{\text{min}}} \right)^2 \right]$$  \hspace{1cm} \text{--- Eq. 3.13}$$

$$\Delta \varphi_{\text{max}} = \max_{v_k} \max_{v_l} \Delta \varphi_t(k)$$  \hspace{1cm} \text{--- Eq. 3.14}$$

$$\Delta \varphi_{\text{min}} = \min_{v_k} \min_{v_l} \Delta \varphi_t(k)$$  \hspace{1cm} \text{--- Eq. 3.15}$$

The Grey Relational Grade (GRG) is a globalized measure adopted for the Grey Relational Analysis. It is used to describe and explain the relation between two sets under a certain background. The degree of similarity between them is easily determined using the GRG. The greater GRG between two tuples/objects, the closer the relationship.
between these tuples/objects. The similarity of GRA measures the relative distance between tuple’s/object’s features and maximal or minimal distance differences.

Where ED(\(k\)) is the Euclidean Distance(ED) between vector \(\Phi_{\text{test}}\) and vector \(\Phi(k)\); \(\Delta \varphi_{\text{min}}\) and \(\Delta \varphi_{\text{max}}\) are the minimum and maximum values of the matrix \(\Delta \Phi\), respectively; \(\xi\) is a recognition coefficient with parameter interval (0,25), \(\xi=15\) was chosen in this study. The grey relational grades \(r(k)\) are inversely proportional to the distances. If the test vector \(\Phi_{\text{test}}\) is similar to any comparative vector \(\Phi(k)\), the grade \(r(k)\) will be a maximum value. GRA uses the grey relational grade to measure the relationship between the reference sequence data and comparative sequences data. The dimension of grey relational vector \(\Gamma=[r(1),r(2),...,r(k),...,r(K)]\) can be reduced from \(K\)-dimension to \(m\)-dimension by

\[
\gamma_j = \sum_{k=1}^{K} r(k)w_{kj}, \quad j = 1,2,3,...m
\]

--- Eq. 3.16

\[
w_{kj} = \begin{cases} 
1, & \text{if } k \in \text{Class } j, \\
0, & \text{if } k \notin \text{Class } j. 
\end{cases}
\]

The final grey grade \(g_j\) that an unknown vector \(\Phi_{\text{test}}\) belongs to Class \(j\) can be derived from the following equation:

\[
g_j = \frac{\gamma_j}{\sum_{j=1}^{m} \gamma_j}, \quad j = 1,2,3,...,m
\]

--- Eq. 3.17

which defines the decision for classifying an unknown vector \(\Phi_{\text{test}}\).

The performance of the classifier is analyzed by calculating sensitivity, specificity and precision through confusion matrix analysis.