Chapter 2: Review of Literature

This chapter describes theoretical concepts of bioinformatics, previous works on promoter prediction methods and introduction to Grey Relational Analysis.

2.1 Review of Literature on Bioinformatics:

*Bioinformatics* is the application of information technology and computer science to the field of molecular biology (Hogeweg & Hesper, 1979). The term *bioinformatics* was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems.

The most fascinating aspect of bioinformatics is the computational investigation, discovery and prediction of biological functions of different parts of DNA/RNA and protein sequences. One of the important practical goals of bioinformatics to reduce the need for laboratory experiments, as they are expensive and time consuming.

The field of bioinformatics has evolved to analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- the development and implementation of tools that enable efficient access to, and use and management of various types of information.
- the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods.
to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

There are two fundamental ways of modeling a Biological system (e.g. living cell) both coming under Bioinformatic approaches.

- **Static**
  - Sequences - Proteins, Nucleic acids and Peptides.
  - Structures - Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides.
  - Interaction data among the above entities including microarray data and Networks of proteins, metabolites.

- **Dynamic**
  - Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites.
  - Multi-Agent Based modeling approaches capturing cellular events such as signaling, transcription and reaction dynamics.

Ten years ago, two fingers were enough to count the number of sequenced human genomes. Until last year, the fingers on two hands were enough. Today, the rate of such sequencing is escalating so fast it is hard to keep track. *Nature* requested more than 90 genomics centres and labs to estimate the number of human genome sequences they have in the works. Although far from comprehensive, the tally indicates that at least 2,700 human genomes will have been completed by the end of this month, and that the total will rise to more than 30,000 by the end of 2011 (Human Genome 2010).
2.1.1 The Human Genome Project:

"Science is essentially a cultural activity. It generates pure knowledge about ourselves and about the universe we live in, knowledge that continually reshapes our thinking" (Sulston, 2010).

In 1953 James Watson and Francis Crick discovered the structure of DNA - the code of instructions for all life on earth...

...in 2003 - just 50 years later - humankind had developed and exploited the technology, the computing capability and the financial and social thrust to record one whole human DNA sequence: nearly 3.2 billion letters of genetic code.

Surprisingly the final sequence was found to contain only 25,000 genes - a quarter of the number originally suspected. The worldwide effort quickly turned to establishing what these genes and their code - the "book of life" - would mean for human health and disease. Determined that the medical fruits of that "book of life" would be shared equally, irrespective of wealth or nationality, the Wellcome Trust extended its funding for the Sanger Institute and other biomedical projects around the world to work on developing genetic knowledge of some of the most devastating and human diseases: from cancer to malaria; from heart disease to typhoid.

The cell is the functional basic unit of life (Fig.2.1). It was discovered by Robert Hooke and is the functional unit of all known living organisms. It is the smallest unit of life that is classified as a living thing, and is often called the building block of life. In most cellular organisms, DNA is organized on chromosomes located in the nucleus of the cell (Bruce 2002).

DNA is a double-stranded helix of nucleotides which carries the genetic information of a cell. It encodes the information for the proteins and is able to self-
replicate. A molecule of DNA consists of two chains or strands, which are composed of a large number of chemical compounds called nucleotides which are linked together to form a chain. These chains are arranged like a ladder that has been twisted into the shape of a winding staircase, called a double helix.

![Fig. 2.1. The cell, chromosome and DNA. (Image Credit: www.virtualmedicalcentre.com)](image)

Nucleic acids are biological molecules essential for life, and include deoxyribonucleic acid and ribonucleic acid (Dahm, 2008). Nucleic acids consist of a chain of linked units called nucleotides. Each nucleotide consists of three subunits: a phosphate group and a sugar (ribose in the case of RNA, deoxyribose in DNA) make up the backbone of the nucleic acid strand, and attached to the sugar is one of a set of nucleobases. The nucleobases are important in base pairing of strands to form higher-level secondary and tertiary structure such as the famed double helix (Fig. 2.2).
The possible letters are $A$, $C$, $G$, and $T$, representing the four nucleotide bases of a DNA strand—Adenine($A$), Cytosine($C$), Guanine($G$), Thymine($T$)—covalently linked to a phospho-diester backbone. In the typical case, the sequences are printed adjoining one another without gaps, as in the sequence $AAAGTCTGAC$, read left to right in the $5'$ to $3'$ direction. With regards to transcription, a sequence is on the coding strand if it has the same order as the transcribed RNA.

One sequence can be complementary to another sequence, meaning that they have the base on each position is the complementary (i.e. $A$ to $T$, $C$ to $G$) and in the reverse order. For example, the complementary sequence to $TTAC$ is $GTAA$. If one strand of the double-stranded DNA is considered the sense strand, then the other strand, considered the antisense strand, will have the complementary sequence to the sense strand.

2.1.2 Nucleic Acid Notation:

While $A$, $T$, $C$, and $G$ represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and
Applied Chemistry (IUPAC) (Moss, 1984) are given in Table 2.1. These symbols are also valid for RNA, except with U (Uracil) replacing T(Thymine).

<table>
<thead>
<tr>
<th>IUPAC Code</th>
<th>Amino Acid(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>R</td>
<td>GA</td>
<td>PuRine</td>
</tr>
<tr>
<td>Y</td>
<td>TC</td>
<td>PYrimidine</td>
</tr>
<tr>
<td>K</td>
<td>GT</td>
<td>Keto</td>
</tr>
<tr>
<td>M</td>
<td>AC</td>
<td>AMino</td>
</tr>
<tr>
<td>S</td>
<td>GC</td>
<td>Strong Bonds</td>
</tr>
<tr>
<td>W</td>
<td>AT</td>
<td>Weak Bonds</td>
</tr>
<tr>
<td>B</td>
<td>GTC</td>
<td>All but A</td>
</tr>
<tr>
<td>D</td>
<td>GAT</td>
<td>All but C</td>
</tr>
<tr>
<td>H</td>
<td>ACT</td>
<td>All but G</td>
</tr>
<tr>
<td>V</td>
<td>GCA</td>
<td>All but T</td>
</tr>
<tr>
<td>N</td>
<td>AGCT</td>
<td>Any</td>
</tr>
</tbody>
</table>

Table 2.1 Amino Acid, IUPAC code and Description

2.1.3 Structure of DNA

Nucleic acid structure refers to the structure of nucleic acids such as DNA and RNA. It is often divided into four different levels: Primary structure: the raw sequence of nucleobases of each of the component DNA strands; Secondary structure: the set of interactions between bases, i.e., which parts of which strands are bound to each other; Tertiary structure: the locations of the atoms in three-dimensional space, taking into consideration geometrical and steric constraints; and Quaternary structure: the higher-level organization of DNA in chromatin, or to the interactions between separate RNA units in the ribosome or spliceosome. (Nucleic Acid Structure, 2010)

**Primary Structure:** The sequence or primary structure of a nucleic acid is the composition of atoms that make up the nucleic acid and the chemical bonds of those
atoms. Because nucleic acids, such as DNA and RNA, are unbranched polymers, this specification is equivalent to specifying the sequence of nucleotides that comprise the molecule. This sequence is written as a succession of letters representing a real or hypothetical nucleic acid molecule or strand. By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end. The sequence has capacity to represent information. Biological DNA represents the information which directs the functions of a living thing. In that context, the term genetic sequence is often used. Sequences can be read from the biological raw material through DNA sequencing methods (Fig. 2.3).

![Fig. 2.3 The Primary Structure of DNA](image)

**Secondary Structure:** The secondary structure of a nucleic acid molecule refers to the base-pairing interactions within a single molecule or set of interacting molecules, and can be represented as a list of bases which are paired in a nucleic acid molecule (Dirks et al, 2004) (Fig. 2.4). The secondary structures of DNA and RNA tend to be different:
biological DNA mostly exists as fully base-paired double helices, while biological RNA is single stranded and often forms complicated base-pairing interactions due to its increased ability to form hydrogen bonds stemming from the extra hydroxyl group in the ribose sugar. In a non-biological context, secondary structure is a vital consideration in the rational design of nucleic acid structures for DNA nanotechnology and DNA computing, since the pattern of base-pairing ultimately determines the overall structure of the molecules.

**Tertiary Structure:** The tertiary structure of a nucleic acid is its precise three-dimensional structure, as defined by the atomic coordinates (McNaught & Wilkinson, 1997) (Fig. 2.5). RNA and DNA molecules are capable of diverse functions ranging from molecular recognition to catalysis. Such functions require a precise three-dimensional tertiary structure. While such structures are diverse and seemingly complex, they are
composed of recurring, easily recognizable tertiary structure motifs that serve as molecular building blocks. Some of the most common motifs for RNA and DNA tertiary structure are described below, but it is important to remember that this information is based on a limited number of solved structures. Many more tertiary structural motifs will be revealed as new RNA and DNA molecules are structurally characterized.

2.1.4 The Central Dogma of Molecular Biology:

The central dogma of molecular biology was first articulated by Francis Crick in 1958 and re-stated in *Nature* (Crick, 1970).

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that information cannot be transferred back from protein to either protein or nucleic acid."

The Central Dogma of Molecular Biology specifies the relationship between the DNA, RNA and Protein in terms of three major processes (Limsoon Woong, 2006) (Fig. 2.6 and Fig. 2.7).
1. Replication: The process by which the information in the DNA molecule in one cell is passed on to new cells as the cell divides and the organism grows. The double-stranded complementary structure of the DNA molecule, together with the nucleotide pairing rules provide the framework for single DNA chain to serve as templates for creating complementary DNA molecules. Entire genetic blueprint can thus be passed on from cell to cell through DNA replication. In this way, virtually all the cells in our body have the full set of recipes for making all the proteins necessary to sustain life’s many different functions.

Fig. 2.6 Central Dogma of molecular biology

Fig. 2.7 Central Dogma and Splicing
2. Transcription: The process by which the relevant information encoded in DNA is transferred into the copies of messenger RNA molecules during the synthesis of the messenger RNA molecules. Just as in the DNA replication process, the DNA chains also serve as templates for synthesizing complementary RNA molecules. Transcription allows the amount of the corresponding proteins synthesized by the protein factories—the ribosome in the cytoplasm—to be regulated by the rate at which the respective mRNA is synthesized in the nucleus. Microarray experiments measure gene expression with respect to the amount of corresponding mRNAs present in the cell, and indirectly infer the amount of the corresponding proteins—the gene products—present in the cell.

3. Translation: The process by which genetic information on the mRNA is transferred into actual proteins. Protein synthesis is carried out by the ribosome and it involves translating the genetic code transcribed on the mRNA into a corresponding amino-acid string which can then fold into the functional protein.

In other words, 'once information gets into protein, it can't flow back to nucleic acid'. The dogma is a framework for understanding the transfer of sequence information between sequential information-carrying biopolymers, in the most common or general case, in living organisms. There are 3 major classes of such biopolymers: DNA and RNA (both nucleic acids), and protein. There are $3 \times 3 = 9$ conceivable direct transfers of information that can occur between these. The dogma classes these into 3 groups of 3: 3 general transfers (believed to occur normally in most cells), 3 special transfers (known to occur, but only under specific conditions in case of some viruses or in a laboratory), and 3 unknown transfers (believed never to occur). The general transfers describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA
information can be copied into mRNA, (transcription), and proteins can be synthesized using the information in mRNA as a template (translation) (Crick, 1970).

2.1.5 The standard Genetic Code:

The genetic code is the set of rules by which information encoded in genetic material (DNA or mRNA sequences) is translated into proteins (amino acid sequences) by living cells. The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids. With some exceptions (Turanov et al., 2009), a triplet codon in a nucleic acid sequence specifies a single amino acid. Because the vast majority of genes are encoded with exactly the same code (Fig 2.8), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact there are many variant codes. For example, protein synthesis in human mitochondria relies on a genetic code that differs from the standard genetic code.

![Fig. 2.8 The standard Genetic Code](image-url)
One codon, AUG serves two related functions:

- it signals the start of translation
- it codes for the incorporation of the amino acid methionine (Met) into the growing polypeptide chain

The genetic code can be expressed as either RNA codons or DNA codons. RNA codons occur in messenger RNA (mRNA) and are the codons that are actually "read" during the synthesis of polypeptides (the process called translation). But each mRNA molecule acquires its sequence of nucleotides by transcription from the corresponding gene. Because DNA sequencing has become so rapid and because most genes are now being discovered at the level of DNA before they are discovered as mRNA or as a protein product, it is extremely useful to have a table of codons expressed as DNA.

2.1.6 Gene:

A gene is a unit of heredity in a living organism. It normally resides on some stretches of DNA and RNA that codes for a type of protein or for an RNA chain that has a function in the organism (Fig. 2.9). DNA is the genetic material of all cellular organisms and most viruses. DNA carries the information needed to direct protein synthesis and replication. Protein synthesis is the production of the proteins needed by the cell or virus for its activities and development.

Replication is the process by which DNA copies itself for each descendant cell or virus, passing on the information needed for protein synthesis. The self-replicating genetic structures of cells contain the cellular DNA that bears in its nucleotide sequence the linear array of genes.
A prokaryotic cell is a cell which contains no nucleus, while eukaryotic cells have nuclei. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on a single chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.

A gene in relation to the double helix structure of DNA and to a chromosome is shown in Fig. 2.9. The chromosome is X-shaped because it is dividing. Introns are regions often found in eukaryote genes that are removed in the splicing process (after the DNA is transcribed into RNA). Only the exons encode the protein.

The first product of transcription differs in prokaryotic cells from that of eukaryotic cells. In prokaryotic cells the product is mRNA, which needs no post-transcriptional modification. In eukaryotic cells, the first product is called primary transcript, that needs post-transcriptional modification to give hnRNA (heterophil nuclear RNA). hnRNA then undergoes splicing of introns (noncoding parts of the gene) via spliceosome to produce the final mRNA.
In prokaryotic cells most of the DNA sequence is coding for protein. For example, almost 70% of the genome of the bacterium *H. influenza* is coding. These organisms replicate very fast; therefore less time is 'wasted' on clever mechanisms. Each gene is one continuous stretch of bases. That is, there are no introns in the coding region.

In Prokaryotes, translation occurs in the cell's cytoplasm, where the large and small subunits of the ribosome are located, and bind to the mRNA. In Eukaryotes, translation occurs across the membrane of the endoplasmic reticulum in a process called vectorial synthesis. The ribosome facilitates decoding by inducing the binding of transfer RNAs (tRNA) with anticodon sequences complementary to that of the mRNA. The tRNAs carry specific amino acids that are chained together into a polypeptide as the mRNA passes.

2.1.7 Promoter:

Promoters are modular DNA structures containing complex regulatory elements required for gene transcription initiation (Lin and Li, 2010). In genetics, a promoter is a region of DNA that facilitates the transcription of a particular gene. Promoters are typically located near the genes they regulate, on the same strand and upstream (towards the 5' region of the sense strand) (Promoter (Biology), 2007). Hence, the identification of promoters using machine learning approach (*insilico*) is very important for improving genome annotation and understanding transcriptional regulation.

In genetics the promoter is commonly referred to as the region upstream of a gene that contains the information permitting the proper activation or repression of the gene that it controls (Pedersen et al., 1999; Smale & Kadonaga 2003). Promoters are located
near the genes they regulate, on the same strand and typically upstream of protein coding regions. (towards the 5' region of the sense strand). They are usually short sequences and are bound by specific transcription factors.

The promoter region itself is typically divided into three parts: (1) The core promoter is the minimal portion of the promoter required to properly initiate gene transcription. It contains a binding site for RNA polymerase (RNA polymerase I, RNA polymerase II, or RNA polymerase III). A vast network of regulatory factors that contribute to the initiation of transcription by RNA polymerase ultimately target any specific gene's core promoter. (2) A proximal promoter is a proximal sequence upstream of the gene, specifically the transcription start site of the gene, that tends to contain primary regulatory elements. It is approximately 250 base pairs (bp) upstream (signified by a negative sign before the number of base pairs, e.g. -250 bp) of the TSS and has specific transcription factor binding sites and (3) A 'distal promoter' is a distal sequence upstream of the gene, specifically the transcription start site, that may contain additional regulatory elements, often with a weaker influence than the proximal promoter.

However, the accurate identification and delineation of promoter regions is important for several reasons, such as improving genome annotation and devising experiments to study and understand transcriptional regulation. Experimentally (biochemical method) finding promoters from huge genome like human is near impossible for researchers (Alfred et al., 2009). Hence, prediction of promoters by computational method is a highly regarded area of interest.
In order for the transcription to take place, the enzyme that synthesizes RNA, known as RNA polymerase, must attach to the DNA near a gene. Promoters contain specific DNA sequences and response elements which provide a secure initial binding site for RNA polymerase and for proteins called transcription factors that recruit RNA polymerase. These transcription factors have specific activator or repressor sequences of corresponding nucleotides that attach to specific promoters and regulate gene expressions. This transcription yields mRNA, which is the first step of the protein production process.

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).

This thesis uses this method for the identification of genes in genomic DNA. The proposed system works on the premise that the identification of a promoter region leads to the successful identification of the start position of a gene.

2.1.8 Prokaryotic Gene Structure:

Fig. 2.10 Structure of Prokaryotic DNA near TSS
Figure 2.10 and 2.11 displays a simple model of a Prokaryotic gene. The '+1' position marks the beginning of the gene and the 'T' signifies the 'terminator' which marks the end of the gene. It is the DNA between these two positions which is transcribed into mRNA, which is used for the production of the protein. In prokaryotes, the sequence of a promoter is recognized by the Sigma(σ) factor of the RNA polymerase. These sites are represented by the '-35' and '-10' symbols in Fig. 2.12. E.Coli promoters are composed of two of these sites. These are the locations to which E.Coli polymerase, the RNAP used by E.Coli, binds in order to begin the transcription of protein. These two binding sites are always located at the points known as the -35 hexamer box (TTGACA) and the -10 hexamer box(TATAAT). Consensus sequences have been identified for each of these two locations (Rosenberg & Court, 1979; Hawley & McClure, 1983; Hertz & Stormo, 1996). A consensus sequence is the most probable sequence to occur at a certain position. The spacer between the -10 hexamer box and the transcriptional has a variable length, the most probable length being 7. The spacer between the -10 site and the -35 site is also of variable length and can vary between 15 and 21 bases (Hawley & McClure, 1983).
An ideal E.Coli promoter would look like the sequence displayed in Figure 2.12. It displays the most probable sequences that could occur at the two hexamer boxes. It is this variation that can make the recognition of these promoters difficult with traditional methodologies. Many promoter sequences have the pyrimidine (i.e. C or T) at position +1 (one nucleotide upstream of the transcription start site or the gene), and purine (i.e. A or G) at the transcriptional start site. In addition to the more obvious features described, a few more non-obvious features have been observed by (Mengeritsky & Smith, 1987; Galas et al., 1985).

2.1.9 Eukaryote Gene Structure:

An eukaryote is an organism whose cells contain complex structures enclosed within membranes. The defining membrane-bound structure that sets eukaryotic cells apart from prokaryotic cells is the nucleus, or nuclear envelope, within which the genetic material is carried (Youngson, 2006; Nelson & Cox, 2005; Martin, 1983).

The gene structure and the gene expression mechanism in eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of exons and introns. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called splicing takes place, in which the intron sequences are excised and discarded from the RNA sequence. The remaining RNA
segments, the ones corresponding to the exons, are ligated to form the mature RNA strand. A typical multi-exon gene has the following structure (Fig. 2.13). It starts with the promoter region, which is followed by a transcribed but non-coding region called 5’ untranslated region (5’ UTR). Then follows the initial exon which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon. It is followed by another non-coding region called the 3’ UTR. Ending the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signaled by specific short (2bp long) sequences. The 5’(3’) end of an intron(exon) is called the donor site, and the 3’(5’) end of an intron(exon) is called the acceptor site. Discovering the location of genes on the genome is the first step towards building such a body of knowledge can be located using Hidden Markov Model (Uma Devi & Nageswara Rac, 2005).

![Fig. 2.13 Structure of Eukaryotic DNA near TSS](image)

The transcriptional signals most often used in gene finding are the initiator or cap signal, located at the transcription start site (TSS), and the A+T rich, TATA-box signal, typically
located about 30 bp upstream of the TSS (Bucher, 1990). These core promoter elements are, however, present in only about 70% of human promoters and, even when present, are not sufficiently precise to allow reliable prediction of promoter locations (Fickett & Hatzigeorgiou, 1997). Even when the full spectrum of characterized transcription factor-binding sites is used in a promoter recognition algorithm (Prestridge, 1995), there does not appear to be a significant improvement in the prediction of precise promoter locations when tested on novel promoter sequences (Fickett & Hatzigeorgiou, 1997). This somewhat disappointing result is probably related to the variability in the location of transcription factor-binding sites relative to the location of the TSS and to the difficulty of accounting for their combinatorial activity. Other features known to play a role in promoter function, such as transcriptional enhancers and silencers, CpG methylation, chromatin structure and DNA curvature, could prove useful in prediction when they are better understood.

2.2 Review of Promoter Prediction Methods:

The problem of promoter recognition has many facets such as:

- Determination of the promoter region, without any attempt to find out what such regions contain.

- Determination of the location of different binding sites for numerous Transcription Factors (TFs) that participate in the initiation of the transcription process.

- Determination of the TSS, which is an important reference point in the context of transcription initiation.

- Determination of the functional classes of promoters, etc.
Extracting new biological knowledge in a computational manner from recorded biological sequence databases is one of the key issues of bioinformatics. One of the most important general problems in bioinformatics is the annotation of uncharacterized biological sequences. At present, the primary target of annotation of sequences originated from eukaryotes is the location of protein coding genes. Correctly recognizing the starting and ending points of different genes is not a simple task, and methods for this purpose are not sufficiently accurate yet.

The starting end of genes can be more precisely determined through the location of promoters, since promoters are usually located before the respective gene, so that recognizing a promoter allows for a more precise determination of the gene's 5' end.

Computational promoter finding has received more attention in the last decade. In Eukaryotes, the promoter region is usually located upstream of the transcription start site or overlaps the transcription start site. A gene has at least one promoter. (Pedersen et al., 1999; Weinzierl, 1999) Thus finding a gene is by first locating its promoter (Werner, 1999). Discovering new genes through the identification of their promoters in anonymous DNA and the study of transcriptional control make promoter recognition an extremely important issue for Bioinformatics.

2.2.1 Literature Review Based on Signal features

The promoter region of protein coding genes of eukaryotes—shortly called eukaryotic PolII promoters, represents a section of DNA to which RNA Polymerase II enzyme and different transcription factors (TFs) bind, forming the so-called transcription pre-initiation complex that makes the initiation of the DNA transcription possible.
Different transcription factors bind to different subsections of the promoter region. These docking sites, called transcription factor binding sites, are recognized by transcription factors via the bio-chemical machinery of the cell. One of the most transcription factor binding sites in eukaryotes is the so-called TATA-box (Benoist et al., 1980; Bucher 1990; Corden et al., 1980; Hahn et al., 1989; Nussinov et al., 1986; Penotti, 1990; Singer et al., 1990; Wobbe & Struhl 1990).

Although there are a number of transcription factor binding sites in eukaryotic PolII promoters – such as Initiator(Inc), CCAAT-box and GC-box that are shared among larger promoter groups, the most common transcription factor binding site among them seem to be the TATA-box.

Recognition of this element is heavily dependent on finding good matches to the TATA-like motif. Position specific Weight Matrix (PWM) is a statistical motif descriptor. It is derived from the base frequency matrices that represent the probabilities of a given nucleotide occurring at a given position in a motif. Since it basically relates to probabilities, the PWM attempts to describe in a statistical fashion some characteristics of the motifs found in a collection of sequences. For the proper determination of PWM enough data should be available.

To accurately predict promoter regions, finding discriminative and informative features is the first and key step. As far as feature choice is concerned, there are two distinct types of features used in the area of promoter prediction: signal and context structure features. The most important signal features include CpG islands, transcription factor binding sites(TFBSs) such as TATA-box and CAAT-box, and initiator. PWM
(Bucher, 1990) derives four weight matrices of TATA-box, cap signal, CCAAT-box and GC-box respectively.

Bucher has enhanced the basic algorithm for the determination of the PWM by introducing an optimization criterion based on a measure of local over representation. The PWM for DNA motifs are represented by a rectangular matrix that has 4 rows, where each row corresponds to one of the 4 bases, and its number of columns is equal to the length of the motif. The PWM of the TATA-box motif from Bucher, is of size 4 x 15, giving the motif length of 15 nucleotides, with the start of the core motif at column 2 of the PWM. This PWM can be used to scan a sequence for the presence of the TATA motif. A window of length 15 slides along the sequence. The matching score for the motif is calculated based on the nucleotides found and the PWM. The matching score for the window is given by

\[ x = \sum_{i=1}^{15} w_{b_i,i} \]  
--- Eq. 2.1

where \( w_{b_i,i} \) is the weight of base \( b_i \) at the \( i^{th} \) position of the motif, and \( b_i \) is the \( i^{th} \) base of the sequence in a window; \( b_i \) relates to A, C, G, or T. The presence of the TATA-like motifs, is inferred by comparing the matching score with threshold \( \tau \). The cut-off score is determined experimentally or statistically. The data window that exceeds the threshold value indicates the presence of the TATA-box.

The eukaryotic promoters are far more complex than the prokaryotic ones, and possess very individual structures which are not common to large groups of promoters. It is thus difficult to design a general algorithm for recognition of eukaryotic promoters. The recognition of eukaryotic promoters cannot be made very accurate if it relies on the
detection of the presence of only one of the transcription factor binding sites—even if that is the TATA-box—and that other promoter characteristics or binding sites should be used simultaneously in promoter recognition algorithms. Promoter prediction can be done using string matching like Horspool’s algorithm. (Appa Rao et al., 2006, UmaDevi et al., 2008). However, the recognition performance based on PWM score alone is not very good and too much false recognition is produced if a high sensitivity level of recognition is required.

Most of the methods used to find E-coli promoters perform the alignments of all the given sequences at the transcriptional start site (i.e. the +1 position) and then scanning them for the consensus sequences. They look for the primary consensus sequences at positions ‘-35’ and ‘-10’. Some of the more successful methods scan for some of the weaker motifs that have been identified at positions: ‘+1’, '-22' and '-44' (Galas et al., 1985; Mengeritsky & Smith, 1987; Robison et al., 1998; Salgado et al., 1999).

Each of these systems requires known knowledge of the E. Coli., promoters in order to correctly classify unknown sequences. The main problem associated with these systems, and other systems which use this technique, is the uncertainty involved in the placement of the consensus sequence within the promoter. This can be caused by the consensus sequence varying from the typical or by the consensus sequence being placed at a slightly different position.

In the previous studies on the promoter predictions, Hidden Markov Model (HMM), Artificial Neural Network(ANN), graph based model (Matsuda et al., 2002) and some data mining and weight matrix methods were used. Most of them tried to find the features of the promoters.
A number of computational methods have been proposed for detecting transcription start sites. The simplest approach is to use a DNA weight matrix to detect the TATA box motif (Bucher 1990), thought to be the core of most eukaryotic promoters. PromoterScan (Prestridge, 1995) uses a weight matrix to score TATA-box. More sophisticated methods use hidden Markov models (Audic & Claverie, 1997) or neural networks (Knudsen, 1999)(Appa Rao et al., 2007). Many of these methods were reviewed and evaluated by Fickett and Hatziegeorgiou (1997). All methods were shown to suffer from poor sensitivity, many false positives, and poor positional accuracy. A recent development in promoter recognition is the PromoterInspector program (Scherf et al. 2000). This was trained using a brute-force algorithm to discover a set of sequence motifs overrepresented in promoter regions. It has a much lower false-positive rate than any of the programs reviewed above. However, it only attempts to detect ‘promoter regions’ (defined as regions of the genome containing promoter-like motifs), rather than locating transcription start sites.

A weight matrix is a simple generative model for a short, ungapped sequence motif (Down & Hubbard, 2002). PWM is used extensively in signal feature extraction processing, as it can create a profile that represents the common feature across the training sequence. This profile can be used to scan new sequences and make a decision as to whether these sequences are related to the training group or not (Raychaudhuri, 2006).

Hidden Markov Model (HMM) (Krogh & Brown, 1994) is another method for feature extraction from sequences compared to PWM. HMM can represent spacer-included motifs (Murakami et al., 2000) of a sequence family.
Generalized Hidden Markov Model (GHMM) (Stormo & Haussler, 1994) is used for generating multi-symbol strings in gene finding systems (Kulp et al., 1996).

The Pol II promoter prediction program (Murakami et al., 2000) is built based on PromFD (Chen et al., 1997) and utilizes HMM to acquire additional motifs.

Pedersen et al.,(1998) took the HMM (hidden markov model) to characterize the prokaryotic and eukaryotic promoters. They use promoters from two species to train the HMM. One is for prokaryotic promoters, using E.coli promters and the other is for eukaryotic promoters, using human promoters(Lisser & Margalit,1993).

McPromoter is developed based on GenScan(Burge & Karlin, 1997), and uses stochastic segment models (SSMs) (Ostendorf et al., 1996) which is a generalization of HMM to represent six segments of the promoter sequence from -250 to +50bp: upstream 1 and 2, TATA box, spacer, initiator and downstream(Ohler, 2006).

Even though promising solutions have been proposed - by Bajic et al.,(2002a); Bajic et al.,(2002b); Bajic et al.,(2003); Davuluri et al.(2001); Down & Hubbard(2002); Hannenhalli & Levy(2001); Ioshikhes & Zhang(2000); and Scherf et al.,(2000)—computational recognition of promoters still has not yet achieved a satisfactory level of confidence (Fickett & Hatzigeorgiou, 1997; Stormo, 2000). The reason is that the biological process of transcription activation is very complex and hierarchical, and is not completely understood (Weinzierl 1999). There are numerous and functionally diverse transcription factors that individually bind to specific DNA consensus sequences — called transcription factor binding sites — in the promoter to activate the transcriptional machinery in concert with RNA polymerases.
While the boundaries of promoter region are loosely defined, each promoter has at least one strong reference site: the transcription start site. Promoter search can thus focus either on locating the promoter region (Davuluri et al., 2001; Hannenhalli & Levy, 2001; Ioshikhes & Zhang, 2000; Scherf et al., 2000) or on pinpointing the TSS (Bajic et al., 2003; Bajic et al., 2002a; Bajic et al., 2002b; Down & Hubbard, 2002; Knudsen, 1999; Ohler et al., 2001; Reese et al., 1996). Existing TSS-finders like NNPP2.1 (Reese et al., 1996), Promoter 2.0 (Knudsen, 1999), McPromoter (Ohler et al., 2001) produce a lot of false positive predictions (Ohler et al., 2001; Pedersen et al., 1999; Prestridge, 2000; Reese et al., 2000) making them unsuitable for locating promoters in large genomic sequences.

A recently reported system, Eponine (Down & Hubbard, 2002) has demonstrated very good performance. Its predictions are very much related to CpG-island associated promoters and to G+C rich promoters (Bird et al., 1986; Cross & Bird, 1995; Gardiner-Garden & Frommer, 1987; Larsen et al., 1992). The G+C content of a DNA segment is the proportion of the total number of G and C nucleotides relative to the length of that segment. CpG islands are found around gene starts in approximately half of mammalian promoters and are estimated to be associated with about 60% of human promoters. CpG islands could represent a good global signal to locate promoters across genomes. At least in mammalian genomes, CpG islands are good indicator of gene presence (Pedersen et al., 1999).

Poonam Singhal proposed gene prediction in prokaryotic genomes based on physicochemical characteristics of codons calculated from molecular dynamics (MD) simulations. It is based on three calculated quantities for each codon: the double-helical trinucleotide base pairing energy, the base pair stacking energy, and an index of the
propensity of a codon for protein-nucleic acid interactions. Thus gene finding program, called ChemGenome2(CG2), differentiated genes from non-genes at a level better than previously reported gene-finding methods, underlining the possibility of a unique and useful *ab initio* characterization of DNA sequences from codons to genomes. (Poonam et al., 2008)

Davuluri et al., (2001) identified promoters and first exons using quadratic discriminant functions. They presented a set of discriminant functions that can recognize structural and compositional features such as CpG islands, promoter regions and first splice-donor sites.

Implementation of the discriminant functions into a decision tree constituted a new program called FirstEF. Frequency of CpG di-nucleotides is calculated for stretches of DNA sequences greater than 200 bp in length with a high G+C content. The CpG di-nucleotide percentage for each window defines the maximum of these CpG percentages as the CpG score and the corresponding window as the CpG window. FirstEF program predicts the first exons and promoter regions in the human genome. FirstEF consists of different discriminant functions structured as a decision tree. The probabilistic models are designed to find potential first splice-donor sites and CpG-related and non-CpG-related promoter regions based on discriminant analysis.

Biologically speaking, there are many signals that help us detect promoters (Akan & Deloukas, 2008). One is the Initiator(Inr) element, located around the TSS and can work independently or synergistically with the TATA box. However, the occurrence of the Inr element in mammals is not clear, but it seems to be quite abundant in
Drosophila. TATA boxes and CpG islands are very clear signals that have been studied thoroughly, and they will be reviewed. Still, the presence of a TATA box or a CpG island is not a requirement for the presence of a promoter, since there exist promoters without them. So, their absence does not imply that there is no promoter.

2.2.2 Need of n-gram:

Regulation of expression of a gene occurs at various stages and places, along the pathway from genome to proteome. Regulation at promoter is the most important way of regulating the gene expression since transcription occurs only when RNA polymerase binds to the promoter. The occurrence of a promoter is not restricted to the 5' end of a gene alone, but could in fact be found in an exon, intron, untranslated region of 3', or may overlap with another promoter (Carninci et al., 2005). Hence the problem of the recognition of a promoter against various backgrounds becomes a challenging computational problem. In addition, presence or absence of the patterns in a promoter makes the task of promoter prediction an even more complex problem.

2.2.3 Literature Review based on Context Features

Compared to sequence signal features, context features are more complex. Context features are basically extracted from training genomic sequences and analyzed by statistical methods. The word "n-mers" can cover all of the context features. Codons (3-mer) are used as the genetic code and they can be translated into 20 distinct amino acids. From 3-mer (codon) to 8-mer, the number of variables increases exponentially. A better promoter prediction may result based on finding more discriminative features rather than improving model building methodology. Therefore,
many systems start to focus on how to select the most effective and informative features among the massive feature-candidate pool (Ohler, 2001).

A coding region always begins with a start codon and ends with a stop codon. ATG is usually regarded as an initial codon, and TGA, TAA or TAG are usually known as stop codons. It is found that codon patterns in coding and non-coding regions are different (Attwood & Parry-Smith 1999), thus it is supposed that codon-usage statistics can be used to analyze context features of promoters. Context features have been used in many well-known promoter prediction models recently: Promoter Explorer (Xie et al., 2006) selects informative pentamers (5-mer) as context features; DragonGSF (Bajic & Seah, 2003) generates the positional weight matrix (PWM) of pentamers to calculate scores of fixed length sequence. The Kullback-Leibler divergence (Wu et al. 2007) based classifier chooses the hexamers (6-mer) as promoter features to balance the discriminant power of classifiers and the computational speed. Additionally, it has also been shown that 8-mers have distinct pattern relative to TSS (FitzGerald et al. 2004). The great power of sequence context features is testified by the improved performance of these new promoter models.

All the n-mer scores are weighted averages of the n-mer frequencies, the weights being the log-likelihood ratios of the n-mer frequencies of the first exons divided by the n-mer frequencies of the pseudo-exons. Conditional weight matrices of splice-donor sites were calculated based on a sub-classification of splice-donor sites similar to the maximal dependence method explained by Burge and Karlin using a larger window (-5 to +8) than the usual one (-3 to +6); +1 indicated the position of G in the splice-donor site GT and -1 indicated the position of the nucleotide just before GT (Burge & Karlin, 1997).
The pentamer selection method used by PromPredictor (Li & Chen 2005) is also based on calculating the relevance of the same features between different data sets. It refers to the distance function (Solovyev & Makarova 1993):

\[ D(X) = \frac{(m_1 - m_0)}{d_1^2 + d_0^2} \quad -- Eq. 2.2 \]

where \( m_1 \) and \( m_0 \) are the mean value of the feature \( X \) in the promoter and non-promoter sequence data sets respectively. \( d_1 \) and \( d_0 \) are the standard deviations of feature \( X \) in positive and negative training sequence data sets separately. They can be calculated by the following function:

\[ m = \frac{1}{N} \sum_{i=1}^{N} x_i \quad -- Eq. 2.3 \]
\[ d = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - m)^2} \quad -- Eq. 2.4 \]

where \( x_i \) is the value of feature \( X \) appearing in sequence \( S_i \) and \( N \) is the total number of sample sequences in a data set. Subsequently, the top \( N_a \) pentamers (\( N_a < 1024 \)) with the largest values of \( D(X) \) are selected.

Promoter Explorer involves posterior probability in context feature selection. A function for selecting most informative pentamers is defined as:

\[ \eta = \frac{P(I=1|a_l)}{P(I=0|a_l)} , \quad I=1,2,.....1024 \quad -- Eq. 2.5 \]

Where \( P(I=1|a_l) \) is the posterior probability of \( I \) given a pentamer \( a_l \), \( I=1 \) representing the input sequence is the promoter, otherwise \( I = 0 \). The pentamers are ranked according to their \( \eta \) values and then 250 pentamers with the highest values are selected. More recently, the promoter prediction system has adopted the concept of relative entropy in the feature selection process (Wu et al., 2007).
Kullback-Leibler divergence is used to calculate the distance, which is defined as follows:

$$ \delta(P^k_{\text{promoter}}, P^k_{\text{non-promoter}}) = \sum_{i=1}^{4^k} P^k_{\text{promoter}}(i) \ln \frac{P^k_{\text{promoter}}(i)}{P^k_{\text{non-promoter}}(i)} \quad \text{----- Eq. 2.6} $$

where $P^k_{\text{promoter}}$ and $P^k_{\text{non-promoter}}$ represent the probability density functions of words (the combination of A, C, T and G) in promoter and non-promoter sequences. K(K=4,5,6,7) indicates the fixed word length and the total number of words. One subgroup of the most effective words can be obtained by maximizing the following criterion function:

$$ s = \arg \left\{ \max \delta(P^k_{\text{promoter}}, P^k_{\text{non-promoter}}) \right\} \quad \text{----- Eq. 2.7} $$

$$ = \{ i | P^k_{\text{promoter}}(i) > P^k_{\text{non-promoter}}(i) \} \quad \text{----- Eq. 2.8} $$

where $i$ represents the set of subscripts of all the words in the subgroup that are selected.

The desirable number of words within a subgroup can be selected by sorting

$$ \left\{ P^k_{\text{promoter}}(i) \ln \frac{P^k_{\text{promoter}}(i)}{P^k_{\text{non-promoter}}(i)}, i \in S \right\} \text{ in descending order.} $$

Shannon in 1948 was the first to employ $N$-grams for characterizing texts (he also proposed the term $N$-gram). He was speaking about a "discrete source as generating the message, symbol by symbol" (Shannon, 1948). By Shannon, this could be a message written in a natural language, continuous information sources that have been rendered discrete (for example, speech) and, more generally, an abstract stochastic process which generates a sequence of symbols.

A sequence $X$ of length $k$ is a linear succession of $k$ symbols from a finite alphabet, $A$, of cardinality $|A| = r$. A segment of $n$ consecutive symbols from the
sequence $X(n \leq k)$ is an $n$-gram ($n$-tuple, $n$-word, $n$-plet, $n$-mer) of the sequence $X$. There are $L = r^n$ different $n$-grams over the alphabet $A$, $\{w_1, w_2, \ldots, w_l\}$. There are $k-n+1$ overlapping $n$-grams in the sequence $X$. Some authors use $n$-grams in a broader sense not assuming contingency of symbols but a distance of a given length between them (Vinga & Almeida, 2003).

If $c_i$ denotes the number of occurrences of the $n$-gram $w_i$ ($i=1, 2, \ldots, L$) in the sequence $X$, and $f_i$ denotes relative frequency of the $n$-gram $w_i$ in the sequence $X$ ($f_i = c_i/(k-n+1)$), then a vector of $n$-gram counts, $C^X_n = (c_1, c_2, \ldots, c_L)$, as well as a vector of $n$-gram frequencies, $f^X_n = (f_1, f_2, \ldots, f_i)$ may be associated with the sequence $X$. For example, for DNA sequences, $A=\{A, C, G, T\}$, $r=4$; for $n=2$, number of all possible bigrams is $L=r^n=16$ and the set of all possible bigrams will be $\{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}$.

In particular, Shannon’s $N$-grams were defined as formal words (i.e. not related to their real values). Informally speaking, if an object can be represented by a sequence over the taken alphabet $A$, then one way of performing a feature extraction is to describe it in terms of its subsequences. An $N$ gram is a subsequence of length $N$. Despite its initial narrow usage in the theory of communication, $N$-grams were later applied more widely including such fields as classification of different “texts” like messages in natural and artificial languages, music, images, etc. Depending on the application field, this approach may use $N$-grams of different length—for example, from 2 to 100 letters.

$n$-Grams have been successfully used for a long time in a wide variety of problems and domains, including: text compression (Wisniewski, 1997) spelling error detection and correction (Angell et al., 1983; Zamora et al., 1981), optical character
recognition (Adnan et al., 2003), information retrieval (Heer, 1974), language identification (Schmitt, 1991), automatic text categorization (Cavnar & Trenkle, 1994), music representation (Downie, 1999), computational immunology (Marceau, 2001), analysis of whole-genome protein sequences (Ganapathiraju et al., 2002) authorship attribution Ke’ selj et al., 2003) protein classification (Solovyev & Makarova, 1993), protein classification (Cheng et al., 2005) and phylogenetic tree reconstruction (Qi et al., 2004).

2.3 Review of Literature on Grey Systems Theory:

This part of thesis provides information about Grey System Theory's Scientific background, characteristics of unascertained systems, incomplete information, inaccuracies in data, comparison of several studies of uncertain systems, position of Grey Systems Theory in cross-disciplinary areas, the development, history and current state.

2.3.1 The Scientific Background:

On the basis of dividing the spectrum of scientific and technological accomplishments into finer sections, the overall development of modern science has shown a tendency of synthesis at a higher level. This higher level synthesis has caused the appearance of the various studies of systems science with their specific methodological and epistemological ("the study or a theory of the nature and grounds of knowledge especially with reference to its limits and validity") significance.

Because of the emergence of various new areas in systems science, our understanding of nature and the laws that govern objective evolutions has been gradually
deepened. At the end of the 1940s, there appeared systems theory, information theory, cybernetics. Toward the end of 1960s and the start of 1970s, there appeared the theory of dissipative structures and bifurcations. During the middle and toward the end of the 1970s, there appeared one by one such new trans-field and interfiled theories of systems science as the ultra-circular theory, dynamic systems, pan-systems, etc. When investigating systems, due to both the existence of internal and external disturbances and the limitation of our understanding, the available information tends to contain various kinds of uncertainty and noises (Liu & Lin, 2011).

Along with the development of science and technology and the progress of the mankind, our understanding of uncertainties of systems has been gradually deepened and the research of uncertain systems has reached at a new height. During the second half of the 20th century, in the areas of systems science and systems engineering, the seemingly non-stoppable emergence of various theories and methodologies of unascertained systems has been a great scene.

The grey systems theory is a new methodology that focuses on the study of problems involving small samples and poor information (Deng, 1982). It deals with uncertain systems with partially known information through generating, excavating, and extracting useful information from what is available. So, systems’ operational behaviors and their laws of evolution can be correctly described and effectively monitored. In the natural world, uncertain systems with small samples and poor information exist commonly. That fact determines the wide range of applicability of grey systems theory.
2.3.2 Characteristics of Unascertained Systems:

The fundamental characteristic of uncertain systems is the incompleteness and inadequacy in their information. Due to the dynamics of system evolutions, the biological limitations of the human sensing organs, and the constraints of relevant economic conditions and technological availabilities, uncertain systems exist commonly.

2.3.2.1 Incomplete Information

Incompleteness in information is one of the fundamental characteristics of uncertain systems. The situation involving incomplete system information can have the following four cases:

1. The information about the elements (parameters) is incomplete;
2. The information about the structure of the system is incomplete;
3. The information about the boundary of the system is incomplete; and
4. The information on the system’s behaviors is incomplete.

The situation of incomplete information is often seen in our social, economic, and scientific research activities. For instance, in agricultural production, even if we know all the exact information regarding the areas of plantation, seeds, fertilizers, irrigation, due to the uncertainties in areas like labor quality, natural environment, weather conditions, the commodity markets, etc., it is still extremely difficult to precisely predict the production output and the consequent economic values.

For biological prevention systems, even if we clearly know the relationship between insects and their natural enemies, it is still very difficult for us to achieve the expected prevention effects due to our ignorance of the knowledge on the relationships
between the insects and the baits, their natural enemies and the baits, and a specific kind of natural enemy with another kind of natural enemy.

As for the adjustment and reform of the price system, it is often difficult for the policy makers to take action because of the lack of the information about how much psychological pressure the consumers could bear and how price change on a certain commodity would affect the prices of other commodities.

On the security markets, even the brightest market analysts can’t be assured of winning constantly due to their inability to correctly predict economic policy and interest rate changes, management changes at various companies, the direction of the political winds, investors’ behavioral changes in the international markets, and price changes in one block of commodities on other blocks. For the general social-economic system, because there are no clear relationships between the “inside” and the “outside” and between the system itself and its environment, and because the boundary between the inside of the system and the outside is difficult to define, it is hard to analyze the effect of input on the output.

Incompleteness in available information is absolute, while completeness in information is relative. Man employs his limited cognitive ability to observe the infinite universe because it is impossible for him to obtain the so-called complete information. The concept of large samples in statistics in fact represents the degree of tolerance man has to give to incompleteness. In theory, when a sample contains at least 30 objects, it is considered “large.” However, for some situations, even when the sample contains
thousands or several tens of thousands of objects, the true statistical laws still cannot be successfully uncovered.

2.3.2.2 Inaccuracies in Data

Another fundamental characteristic of uncertain systems is the inaccuracy naturally existing in the available data. The meanings of uncertain and inaccurate are roughly the same. They both stand for errors or deviations from the actual data values. From the essence of how uncertainties are caused, they can be categorized into three types: the conceptual, level, and prediction types.

The Conceptual Type: Inaccuracies of the conceptual type item from the expression about a certain event, object, concept, or wish. For instance, all such frequently used concepts as “large,” “small,” “many,” “few,” “high,” “low,” “fat,” “thin,” “good,” “bad,” “young,” “beautiful,” etc., are inaccurate due to the lack of clear definition. It is very difficult to use exact quantities to express these concepts.

The Level Type: This kind of inaccuracy of data is caused by a change in the level of research or observation. The available data, when seen on the level of the system of concern, that is the macroscopic level, or the level of the whole, or the cognitive conceptual level, might be accurate. However, when they are seen on a lower level, that is a microscopic level or a partial localized level of the system, they generally become inaccurate. For example, the height of a person can be measured accurately to the unit of centimeters or millimeters. However, if the measurement has to be accurate to one ten-thousandth level, the earlier accurate reading will become extremely inaccurate.

The Prediction Type (The Estimation Type): Because it is difficult to completely understand the laws of evolution, prediction of the future tends to be inaccurate. As a
matter of fact, no matter what method is used, it is very difficult for anyone to obtain the absolutely accurate (estimated) value.

2.3.3 Comparison of Several Studies of Uncertain Systems:

Probability and statistics, fuzzy mathematics, and grey systems theory are three mostly seen research methods employed for the investigation of uncertain systems. Their research objects all contain certain kinds of uncertainty, which represents their commonality. It is exactly the differences among the uncertainties in the research objects that these three theories of uncertainty are different from each other with their respective characteristics.

Fuzzy mathematics emphasizes on the investigation of problems with cognitive uncertainty, where the research objects possess the characteristic of clear intension and unclear extension.

Probability and statistics study the phenomena of stochastic uncertainty with emphasis placed on revealing the historical statistical laws. They investigate the chance for each possible outcome of the stochastic uncertain phenomenon to occur. Their starting point is the availability of large samples that are required to satisfy a certain typical form of distribution.

The focus of grey systems theory is on the uncertainty problems of small samples and poor information that are difficult for probability and fuzzy mathematics to handle. It explores and uncovers the realistic laws of evolution and motion of events and materials through information coverage and through the works of sequence operators. One of its characteristics is construct models with small amounts of data. What is clearly
different of fuzzy mathematics is that grey systems theory emphasizes on the investigation of such objects that process clear extension and unclear intension.

2.3.4 Position of Grey Systems Theory in cross-disciplinary areas:

Corresponding to differences in how we look at the objective world and matters, the human knowledge is divided into three major blocks: history, poetry and arts, and philosophy, based on his belief that such division should correspond to human capacity of memory, imagination and judgment.

Later the totality of knowledge tends to be divided into either two major blocks with one block of humanities and the other science, or into three areas of natural science, mathematics, and social science.

The totality of knowledge is divided on the basis of the classification of the problems addressed. Firstly, we classify problems according to their complexity and uncertainty. And then, we point out the corresponding disciplines of specific epistemological significance based on the problems’ characteristics so that the position of grey system theory in the spectrum of cross-disciplinary studies is clarified.

Let us use a rectangle $\Omega$ to represent the totality of all matters in the world and circles A, B, C, and D to represent respectively the sets of simple matters, complex matters, deterministic matters, and indeterminate matters. So, we obtain the Venn diagram for the classification of scientific problems in Figure 2.14, where when the methods of resolving the corresponding problems are labeled, we obtain the Venn diagram for cross-disciplinary studies in Figure 2.15.
From a comparison of Figures 2.14 and 2.15, it can be seen that as the scientific methodology of resolving indeterminate semi-complex problems, grey systems theory is established as a jump from probability and statistics that are developed to resolve indeterminate problems.

Fig. 2.14 The Venn diagram for classification of scientific problems

Fig. 2.15 The Venn diagram for classification of cross-disciplines

From a comparison of Figures 2.14 and 2.15, it can be seen that as the scientific methodology of resolving indeterminate semi-complex problems, grey systems theory is established as a jump from probability and statistics that are developed to resolve indeterminate problems.
GRA provides a ranking scheme that gives the order of grey relationship among the dependent and independent factors which leads to essential information such as which input factor need to be considered (Roselina et al., 2008).

2.3.5 The Development History and Current State:


As soon as these works appeared, they immediately caught the attention of many scholars and scientific practitioners from across the world. Numerous well-known scientists strongly supported the validity and livelihood of such research. Many scholars actively participated in the investigation of grey systems theory.

With great enthusiasm these researchers carried the theoretical aspects of the theory to new heights and employed their exciting results to various fields of application. In particular, successful applications in great many fields have won the attention of the international world of learning. Currently, a great number of scholars from many countries have been involved in the research and application of grey systems theory.

In 1989, the British journal, *The Journal of Grey System*, was launched. Currently, this publication is indexed by INSPEC (formerly Science Abstracts) of England, Mathematical Review of the United States, Science Citation Index, and other important indexing agencies from around the world. In 1997, a Chinese publication,

The GRA in the grey system theory is a problem-solving method that is used when dealing with similarity measures of complex relations.

The black box is used to indicate a system lacking interior information (Ashby, 1945). Nowadays, the black is represented, as lack of information, but the white is 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).

Grey system theory is an interdisciplinary scientific area that was first introduced in early 1980s by Deng (1982). Since then, the theory has become quite popular with its ability to deal with the systems that have partially unknown parameters. As superior 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.

There are currently many professional journals in the world that have accepted and published papers in grey systems theory. As of this writing, a Journal of the Association for Computing Machinery (USA), Communications in Fuzzy Mathematics (Taiwan), Kybernetes: The International Journal of Cybernetics, Systems and Management Science, have respectively published special issues on grey systems theory.

In a civil application Lin et al., (2008) introduces grey number and grey relational analysis to develop a new grey model for qualitative identification of the origin of hydraulic cement clinker. The grey model using grey number analysis is an objective and effective technique for the operation by giving an interval number. Grey number will be easier than giving a precise number when an evaluation of an object is based on uncertain information.

Research studies in medical area are as follows: Andy et al., (2005) used grey relational analysis for classification of shapes from the given shape database with dimensionality of two or complex topologies using expectation maximization algorithm. This is useful for medical applications like Emphysema (lung disease), Dandy-Walker Syndrome (congenital brain malformation).

Ming-Feng Yeh, Kuang-Chiung & Chang proposed a new ECG beat classifier based on GreyART networks by incorporating the modified grey relational grade into Adaptive Resonance Theory (ART2) network. This is based on pattern comparison, which need not extract features of the ECG beat and it is proved that Grey ART network
has a good classification accuracy and is applicable to real-time ECG beat analysis. (Yeh et al., 2007).

Some research studies in financial area are as follows: In one study (Wang, 2002), the combination of fuzzification techniques and the grey system theory (GM(1,1) model with adaptive step size) is proposed to predict stock prices and it is shown that the approach is very efficient.

A financial study by Kung & Wen used grey relational analysis to evaluate the relationship between company attributes and its financial relationships. In this study they used six financial indicators to classify twenty items of financial ratios as research variables through the Globalization Grey Relational Analysis (GRA) to extract the significant financial ratio variables and financial indicators which affect the financial performance of venture capital in Taiwan. Then they applied Grey Decision-Making to evaluate financial performance in the sampled enterprises and arrange them in order and then they used GM(0, N) model to analyze different company attributes, which influence the financial performance among enterprises (Kung & Wen, 2007).

Grey relational analysis is integrated with GA method in soft effort estimation. As most software projects have incomplete information and uncertain relations between effort drivers and the required development effort, the grey relational analysis (GRA) method has been applied in building a formal software effort estimation model. A software effort estimation model using the GRA identifies one or more historical projects that are similar to the project that is to be estimated and derives an estimate from them. A globalized measure GRG is obtained from GRA. GRG can present the degree of the
relationship between the project to be estimated and its historical projects. The estimated effort is the average of efforts retrieved from the historical project with the largest weighted GRG among all historical projects. All the projects are ranked in accordance to their weighted GRGs (Huang et al., 2007).

Hsu & Huang (2010) proposed six weighted methods, including non-weighted, distance-based, correlative, linear, nonlinear, and maximal weights, to be integrated into GRA for software effort estimation. The relative importance between the project features and development effort should be considered within the similarity measure. When the weighted similarity of the GRA method is taken into account, equation can be modified as follows:

\[ r_{0i} = \sum_{k=1}^{M} \beta_k \gamma(X_0(k), X_i(k)) \quad \text{where} \quad \sum_{k=1}^{M} \beta_k = 1. \]

He proved that the weighted GRAs perform better than the nonweighted GRA. The linearly weighted GRA can mainly improve accuracy and reliability of estimates. Increasing distinguishing coefficients and choosing smaller analogous numbers enhanced the accuracy of prediction results, but the analogous numbers are much more influential than the distinguishing coefficients. The performance of weighted GRAs is better or close to other estimation techniques.

Nihat Tosun & Hasim Pihtili used grey relational analysis in MQL milling of 7075 Al alloy. The optimization of the face milling process of 7075 aluminum alloy is proposed by using the grey relational analysis for both cooling techniques of conventional cooling and minimum quantity lubrication (MQL). For this the performance characteristics in the milling process such as surface roughness, material removal rate (MRR) in the milling process are considered. The experiments were conducted under
different settings of spindle speed, feed rate, cooling technique, and cutting tool material.

For data preprocessing in the grey relational analysis process, surface roughness and MRR were taken as the “lower is better” and the “higher is better”, respectively. The grey relational grade values for each level of the milling parameters were calculated. The optimal level of the machining parameters is the level with the greatest grey relational grade value. The importance of the controllable factors on the multi performance characteristics was in order of feed rate > cutting speed > tool material > cooling technique. The study indicated clearly that the grey relational analysis accomplished effectively the optimization of surface roughness and MRR in the milling operation at multiple quality requests (Tosun & Pihtili, 2010).

In another study, the moving average autoregressive exogenous (ARX) prediction model is combined with grey predictors for time series prediction and it is proved that the hybrid method has a greater Stock market forecasting accuracy using portfolio selection based on ARX, grey system and RS theories than the GM(1,1) method. (Huang & Jane (2009).

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 (GRA) has been widely applied in analyzing Multivariate Time Series data (MTS). In a complex and multivariate time series system, such as social and economic systems, there are many factors that can influence the state of the system. Evaluation model utilizes GRA to establish a ranking scheme that ranks the order of the grey relationship among dependent and independent factors based on Grey Relational Grade (GRG) values. GRA gives a favorable solution as analyzing tool for small and incomplete multivariate time series data. Some research study in agriculture: GRA can provide a ranking scheme that gives the order of the grey relationship among the dependent and independent factors which leads to essential information such as which input factor need to be considered to forecast grain crop yield. The performance of GRA in ANN model has outperformed the Multiple Linear Regression (MR) model with 99.0% in forecasting accuracy (Roselina et al., 2008).

Chih-Hung Tsai et al., used grey relational analysis to establish a complete and accurate evaluation model for selecting proper vendor to meet production demand. This methodology will significantly reduce the purchasing cost and increase the production efficiency and overall competitiveness. A good vendor must supply high-quality materials, deliver proper quantities of materials at proper time, reduce costs, and provide excellent services in order to satisfy customers' demands (Tsai et al., 2003).

Grey relational analysis used in solving multiple attribute decision-making problems (MADM). There are many cases in daily life and in the workplace which pose a decision problem. Some of them involve picking the best from among multiple available alternatives. GRA has been successfully applied in solving a variety of MADM problems, such as the hiring decision (Olson & Wu, 2006), the restoration planning for power
distribution systems (Chen, 2005), the inspection of integrated-circuit marking process (Jiang et al., 2002), the modeling of quality function deployment (Wu, 2002), the detection of silicon wafer slicing defects (Lin et al., 2006), facility layout and dispatching rules selection (Yiyo et al., 2008) etc.