CHAPTER I

INTRODUCTION

With the progress of human genome project, huge amount of biological data have been provided in recent years. Nowadays *bioinformatics* is one of the most active research fields which is aims at reduction of the need for laboratory experiments, as there are expensive and time consuming. The discipline of bioinformatics can be described as the intersection of genetics, molecular biology, mathematics, computer science and engineering. Its goal is to understand the relationships between sequence, structure, evolution, biological function, molecular behavior and genetics. One of the main problems that bioinformatics attempts to solve is identify the location of genes in DNA sequences. This problem relates strongly to the recognition and location of promoters that indicate the starting of genes. Promoter is the sequence that is responsible for the transcription from DNA to RNA. There are two main classes of functional information encoded in the genomic DNA of every living organism. One class is the coding regions, which specifies the structure and function of each gene product; another class is the regulatory regions, which controls and regulates when, where, and how the genes are expressed. Promoter is the most important regulatory region that controls and regulates the very first step of gene expression: mRNA transcription.

This dissertation deals with computer-based recognition of promoters, including prokaryotic and eukaryotic organisms. Numerous methods exist for promoter recogntion and location but they are still produce a large number of *false positive* (FP) predictions

especially in eukaryotic organisms. The eukaryotic promoters are far more complex with individual micro-structure than prokaryotic promoters. They have compartmented cells with a nucleus containing the DNA which organizes a number of chromosomes. This distinguish them from prokaryotes, mostly bacteria, which do not have a nucleus and whose genome mostly consist of a single coiled DNA loop.

The accuracy of promoter recogntion is based on two factors, i.e., representation of the given DNA sequence and essential features of the sequence. The goal of this dissertation concerns both issues of the above solutions that can provide a distinct classification between the promoter and non-promoter sequences. A chaos game representation (CGR) is adopted for transforming a DNA sequence having promoters and non-promoters into an image. The essential features of the CGR are selected by applying the concepts of statistical feature selection. It's aim for finding the smallest set of features that can distinguish classes as if with the full set and reduce dimension for the classifier. Classification is subsequently performed by a supervised neural network.

1.1 Problems Identification

Prediction of promoter regions is a challenging problem. There have been many methods and systems to predict promoter region, especially promoter of higher eukaryotes, but few of them are applicable to real DNA sequence mainly owing to high false positive rates. These problems are:

- The start and the end of a promoter region cannot be identified exactly. Hence, an announced promoter sequence might contain some non-promoters which could bias the result of the prediction.

- Eukaryotic promoters have highly diverse primary sequences; it has been very difficult to find generalized patterns or rules by conventional sequence analysis

methods.

- Not all specific subregions such as TATA-BOX, CAAT-box, etc., need to exist in a particular promoter but may appear in non-promoter sequences.

1.2 Objective and Scope of the Research

The goal of this dissertation is to develop a promoter recognition algorithm that provides a distinct classification between promoter and non-promoter sequences. The study focuses on both prokaryotic and eukaryotic organisms.

From a computational point of view, the main features of this dissertation are as follows:

- Chaos Game Representation (CGR) is used for representing DNA sequence patterns in the forms of an image,
- statistical feature selection method is proposed for searching an optimal subset of patterns from all available patterns. Not only the dimensions of training data can be reduced, but also improving the precisions of promoter recogntion, and
- all selected patterns are fed into an Artificial Neural Network (ANN) to verify the performance of prediction.

1.3 Organization of the Dissertation

The organization of this dissertation comprises of the followings: Chapter 2 summarizes background knowledge of the molecular biology, introduction of promoter, methodology of multi-layer feedforward, and backpropagation neural network. Chapter 3 reviews some related works on promoter recognition. Chapter 4 introduces materials and methods that are composed of CGR for representing DNA sequences, statistical feature selection

methods, and the architecture of promoter system used in this dissertation. Chapter 5 describes the experimental parameters and results. Finally, the conclusion of this dissertation is in Chapter 6.