CHAPTER III

LITERATURE REVIEWS

In this chapter, I classify the serveyed papers into two categories of organisms—*Prokaryotic* and *Eukaryotic* organisms. Fickett et al.[1] published an excellent overview of the promoter recognition algorithms. Besides giving a great introduction and overview of the biological process, they compared various programs on a standardized eukaryote data set. From the result, they concluded that the problem of eukaryotic promoter recognition was complex and far from being solved.

3.1 Promoter Recognition in Prokaryotes

The first in silico promoter studies concentrated on prokaryotic promoters, which have less complex structures than their eukaryotic counterparts.

Mahadevan and Ghosh [2] invented a three modules for predicting the consensus boxes. In the first module, two neural networks learned the -10 and the -30 bp conserved regions. The second module aligned the sequence, relying on the spacer length. The third module which had one neural network, learned how to predict the aligned sequences.

Pedersen [3, 4] used the Hidden Markov Model [3] and the neural network [4] to find some distinct features of promoters.

Ma et al. [5] combined expectation maximization (EM) algorithms with NN. The EM algorithm was used for locating the -35 and -10 binding sites. Then, features in each training promoter were choosen according to their information content and fed to

an ANN for promoter recognition.

Takashi Matsuda [6] predicted the promoter using data mining method, which was a graph-based induction method. The accuracy of their methods is about 84.91% based on their test data.

Matsuyama and Kawamura [7] proposed Independent Component Analysis (ICA) algorithm that included a position-dependent conversion based on symbol frequencies.

Huang and Wang [8] proposed a hybrid learning system to calculate the distribution of oligo-neucleotides statistics as position weight matrices and fed as inputs to the support vector machine (SVM) for discriminate promoters and non-promoters. Their result is better than other prokaryotic promoter prediction methods with 97.2% accuracy.

3.2 Promoter Recognition in Eukaryotes

One of the first statistical studies of RNA Polymerase II promoter regions in eukaryotes was performed by Bucher [9], who analyzed functional promoter sites from different eukaryotes and built statistical weight matrices for each individual element, such as the TATA box, Inr site, CAAT box and the GC box. The weight matrices were based on counts of a specific nucleotide at a fixed position.

PromoterScan [10] recognizes primate promoters by means of (1) the TATA PWM from Bucher [9], and (2) the density of specific transcription factor binding sites.

Hutchinson [11] proposed an algorithm which employed a simple frequency analysis of differential hexamers (sequence with 6 nucleotides). The true positive accuracy of the testing is sets over 62%.

Audic and Claverie [12] used two Markov transition matrices of promoter, a non-promoter, and an objective Baye's theorem function to determine whether a non-characterized DNA sequence was a promoter or a non-promoter. The results showed nearly 50% true

positive for the testing set. This was due to the Markov transition matrix of a promoter set that it did not seem to learn the features of promoter well.

PromFind [13] is not based on any collection of putative transcription factor binding sites but rather on the differences in nucleotide hexamer frequencies between promoters, protein coding regions, and noncoding regions downstream of the first coding exon.

TSSG and TSSW [14] both applies the same underlying algorithm, which uses a linear discriminant function combining (1) a TATA box score, (2) triplet preferences around the TSS, (3) haxamer preferences relative to the TSS, and (4) potential transcription factor binding sites.

NNPP2.1 [15] is constructed from time-delay neural networks to recognize *Drosophila melanogaster* promoters. It is based on the recognition of two specific signals within the promoter regions : the TATA-box and the initiator (Inr), as well as their mutual distance. This system uses three time-delay ANNs, one for recognition of the TATA-box, one for Inr, and one that combines the outputs of the two and accounts for the spatial distance between these signals.

PromoterInspector [16], one of the most well-known content-based promoter prediction tools which gives attention to analyzing genetic context instead of context location. The main idea is to extract common sequence features from training sequences and generates a set of context features called IUPAC word dictionaries. The PromoterInspector introduces not only a promoter region as the training set but also three non-promoter sequences, namely exon, intron and 3'UTR. The greatest advantage of PromoterInspector seem to be that of dramatically reducing false positives.

McPromoter [17] integrates physical proterties of DNA, such as DNA bendability or GC content into probabilistic promoter recognition system. In the new model, a promoter is represented as a sequence of consecutive segments represented by joint likelihoods for DNA sequence and profiles of physical properties. Sequence likelihoods are

modelled with interpolated Markov chains and physical properties with Gaussian distributions.

Promoter 2.0 [18] combines several neural networks. Each neural network model used perceptron algorithms. There are four networks responded to TATA-box, cap site, CCAAT-box and GC box. To optimize these neural networks, genetic algorithms are used to randomly choose and change an indiviual weight. Promoter 2.0 reported 63% true positives for its test data.

Dragon Promoter Finder 1.2 (DPF) [19] is an integrate promoter prediction model that predicts promoters of vertebrates. The DPF consists of a nonlinear promoter recognition model, sensors for recognizing specific functional regions of DNA, signal processing and artificial neural networks. The data window contents pass through three sensors i.e., promoter, exon and intron. A non-linear signal processing model further analyzes the sensors' outputs and feeds them into to a neural network. The DPF 1.3 extends the capability for recognizing a GC-rich or GC-poor DNA sequence.

Daniel and Karl [20] used Genetic Programming (GP) to build a classifier for recognizing promoter regions in the primary sequence data of eukaryotes. The basic idea is to be able to look at the model that is built and to identify specific motifs and their locations.

PromPredictor [21] recognizes promoter regions in the human genome. PromPredictor extracts compositional features and CpG islands information from genomic sequence, feeding these features as inputs for a hybrid neural network system (HNN) and then applies the HNN for subsequent prediction. It combines a novel promoter recognition model, coding theory, feature selection and dimensionality reduction with machine learning algorithm. Evaluation on Human chromosome 22 was 66% in sensitivity and 48% in specificity.

Prometheus [22] uses non-linear time series descriptors along with nonlinear machine-

learning algorithms, support vector machine (SVM), are used to discriminate between promoter and non-promoter regions. It specifically deals with the application of non-linear dynamics and statistical thermodynamics descriptors, such as Lyapunov component and Tsallis entropy along with non-linear machine-learning algorithms. Prometheus is found to perform significantly better than some other promoter finding programs.