CHAPTER VI

CONCLUSION

This dissertation proposes a new method for solving the promoter recognition problems on prokaryotic and eukaryotic organsims. Existing methods for promoter recogntion and location still produce a large number of *false positive* (FP) predictions especially in eukaryotic organisms.

The accuracy of promoter prediction is based on two factors, i.e., the representation of the given DNA sequence and the essential features of the sequence. 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 concept of statistical feature selection. It is aimed at finding the smallest set of features that can distinguish the classes over the full set and reduce the dimension of the classifier. Recognition can then be performed by a supervised neural network. In this dissertation, I do not consider some well-known patterns around TSS, such as TATAAT-box and TTGACA-box, which were previously used by many researchers. The fact that the patterns of these signals vary a lot, they may appear in different combinations. Their relative locations with respect to the TSS are different for different promoters, thus not all of these signals need to exist in a particular promoter. The experiment considers the content in the sequence composition of promoter and non-promoter examples. So the method does not require any specific knowledge about a particular promoter to make a prediction and thus has a big advantage especially when nothing is known about the

promoter to be predicted.

The promoter sequences in different species have some distinct features. For example, the CpG island looks obvious in eukaryotic promoter regions, but these cannot be applied in prokaryotic promoter regions. The method in this dissertation can be applied to both organisms and used for promoter search in long and continuous of DNA sequences. The result in *E.coli* sequences performs better than the other methods. The results in large genome sequences produce acceptable ratios of TP and FP predictions (maximize the TP recognition while minimize the FP recognition) when compare with other methods.