Abstract

Genes which code for proteins are the most important segments in a genome. A promoter which occurs upstream of a gene acts as a switch in gene transcription. Promoter prediction/identification is helpful in identifying co-regulated genes, unknown function of a gene, gene regulation etc. Various factors contribute to the complexity of the problem of promoter identification. To identify a promoter, signal unique to the promoter is to be extracted from the promoter. Promoters are mostly identified by local motifs/consensus regions present in the promoter. In this thesis, attempt is made to identify a promoter using binary classification approach. The features to the classifier are extracted by considering either whole promoter (global) or binding sites (local motifs). These features are given as input to feed-forward neural network (NN) classifier. Two sets of data are used for this experimentation. One is *E.coli* and the other is *Drosophila Melanogaster*.

Different features, such as n-grams (n=2,3,4,5), position weight matrix based features, features extracted using Fourier transform and wavelet transform are explored here. The thesis discusses the extraction and usage of these features in classification of promoters. This approach has been extended in some cases for promoter identification in a whole genome. A study is also made of the interaction between promoter and RNA polymerase through the signal processing techniques. One main contribution from classification results obtained from n-grams is the development of whole genome promoter prediction methodology using best n-gram features. The results are very good and are better than the results of currently available prediction tools for the forward strand. And an in-depth analysis using 2-grams has given an insight about promoters and non-promoters. It points to the similarities between majority of promoters and a small non-promoter set. Similarly, similarities between a large non-promoter data set and a minor promoter data set are also observed. Experiments with position weight matrix features for genome-wide recognition of promoters reveal that the performance is superior with global features as compared to local features extracted from binding sites. Results of classification using signal processing techniques that use global features show that non-promoters can be identified well with these features. Promoter and RNA polymerase interaction has been modeled using wavelets. The experiments demonstrate that this approach is not well suited for promoter prediction but is very good for non-promoter identification.

It can be concluded from the results that from the viewpoint of classification, features extracted from the promoter as a whole are more representative of the characteristics of the promoter than those acquired from localized motifs. The regions that lie before, in-between and after the binding regions do contribute for the recognition of promoter as is evident from the results obtained using n-grams and position weight matrix features.