Advanced Science Letters
Vol. 24, 1163-1167,
2018

Ensemble Prediction of Enhancers Associated Marks Using K-mer Feature

Nung Kion Lee* and Sina Nazeri
Department of Cognitive Sciences, Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia

Epigenetic marks like chromatin remodelers and histone marks are eminent indicator of enhancers' activity. K-mer is a simple representation of DNA sequences that has been useful for computational epigenetic marks prediction. While many studies have been utilizing k-mer as feature of epigenetic marks prediction, no comprehensive studies have been done to show sophistication of k-mers feature of epigenetic sequences with learning models. This study performs a comprehensive evaluation of diverse learning models using k-mer feature to draw comparison between different learning models and employs blended ensemble learning technique to improve overall performance. Our results show that each learning model ranked quite differently the important of different k-mers for discriminative purpose. The blended ensemble increases the performance of enhancer classification significantly compared with using individual classifiers.

Keywords: Enhancers Motifs, Epigenetic Marks, K-mer, Model Prediction, Blended Model.

1. INTRODUCTION

Gene regulation is conducted through constant interactions of specific proteins known as regulatory elements with distinct regions of DNA called regulatory binding sites (i.e. motif). Unrevealing the complexity of gene regulation paves the way for better understanding of circumstances of organ development as well as associated genetic disorders[1]. Motifs are region of DNA which play essential role in gene regulation. Motifs can be divided into two groups; those in proximity of transcription starting sites (TSS) like promoters – known as proximal motifs- and those that are located distal from TSS are known as distal motifs like enhancers.

Enhancers can regulate genes regardless of their location and orientation which makes them notoriously difficult to identify[2]. Besides, based on physical and psychological condition of the cells, enhancers can be in status of being silence, active or poised[2].

Advance combinational techniques like chromatin immunoprecipitation (ChIP) fallowed by sequencing (ChIP-seq) are able to identify enhancers with high precision[3]. However, the techniques are dependent on availability of antibody and associated with high costs. Therefore, it is impossible to establish wet-lab experiment for every stage of cell development to identify all enhancers.

Computational techniques have been developed to address enhancer identification challenge[4]. Early tools based on comparative approaches and profile search performed good on simple species like yeast[4,5]. However, they were unable to scale up when the size of datasets increase, and they returned false positives or missed enhancers[4,5]. Recent approaches rely on additional data which are associated with activity of enhancers. These data known as epigenetic data are including histone marks and chromatin remodelers[2].

Three prominent enhancers' associated epigenetic marks are including H3k4me1 Histone mark and two chromatin remodelers of DNase I hypersensitive sites (DHSs) and P300 coactivator enzymes[2,3].

In order to utilize epigenetic data, it is essential to generate discriminative features for building an effective prediction model. Usually, the features set consist of representation of DNA sequences which related to enhancers associated marks. A prominent approach to

*Email Address: nklee@unimas.my