Comparisons of Enhancers Associated Marks Prediction Using K-mer Feature

Sina Nazeri*, Nung Kion Lee*, and Norwati Mustapha‡
*Department of Cognitive Sciences
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak
Email: nklee@fcs.unimas.my
‡Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 Serdang, Selangor
Email: norwati@upm.edu.my

***Abstract*—Epigenetic signatures such as chromatin and histone modification marks are prominent indicator of enhancer motif regions. While many works have been using k-mer as feature of epigenetic sequence, no comprehensive studies has been done to compare and contrast how the different choices of k-mers feature parameter affect machine learning algorithm performances. Furthermore, it is not known how effective is the k-mer feature for representing different epigenetic marks-H3K4me1, DHS and p300. In this paper, a comparative study is performed to determine the accuracy, sensitivity and specificity of using k-mer feature for predicting these marks. Our results found that, classifier perform better when the k-mer length is between 4 to 6. Short k-mer length has poor accuracy, sensitivity and specificity. The k-mer feature works best for DHS sequences and has low accuracy for H3K4me1 sequences prediction. The k-mer feature is also performed poorly on specificity of DHS sequences. It can be concluded that, there are still much room for improvement of identifying better feature for representing epigenetic feature for enhancer prediction.**

I. Introduction

Regulation of gene expression is conducted through constant complex interactions of regulatory regions in DNA and corresponding protein. In other word, a protein called transcription factor binds to specific locations of DNA called binding sites (i.e. motif) in order to active or suppress genes. Identification of locations of these regulatory regions contributes to unravelling the mystery of gene regulation which paves the way for resolving genetic disorders [1].

Motif sequences are categorized into (a)proximal regions those within 500bp to 10kb upstream of a transcription starting site (TSS) and (b) distal regulatory regions like enhancers, silencers and insulators. This paper focused on using epigenetic feature to identify enhancer regulatory elements. Enhancers are distinct genomic regions (or the DNA sequences thereof) that contain binding site sequences for transcription factors (TFs) proteins and that can enhance the transcription of a target gene from its transcription start site (TSS)[2]. Enhancers identification is challenging because there is no single feature that is able to determine they are active, poissed or silenced. In addition, they can be located in any distance from the genes they regulated. Advances in technology like chromatin immunoprecipitation followed by sequence (ChIP-seq) are able to detect locations of enhancers with high precision in genome scale motif analysis [3], [4]. However, enhancers are activated in different stages of developmental cells and their activation are also dependent on cell conditions. It is impossible to setup large combinatorial wet-lab conditions needed to identify all enhancers. In addition, not all cell lines from different species are available to be evaluated.

With more and various additional data are associated with enhancers, generating discriminative features is necessary for effective classifier learning. Typically, DNA sequence where these enhancer associated marks are extracted and then features related to the DNA sequences are generated. One of the most widely employed features is the k-mer feature–a continuous oligonucleotide with length of k. K-mer feature is not only being used for modeling epigenetic marks but popular in motif prediction algorithms as well. For examples, in motif pattern recognition k-mers enable suffix tree to model DNA contents for scoring purposes [5]. In another research it provides similarity profile for identifying regulatory regions is Drosophila [6]. Simple k-mer model is employed to produce comprehensive binding specificity for training linear model of protein binding microarrays (PBMs) [7].

While there are many studies have been using k-mer feature for representing DNA sequences from epigenetic or chromatin remodelling marks, there is no comprehensive studies to compare and contrast the use of k-mer feature representing them. The main aim of this paper to evaluate the performances of using k-mer feature for representing the H3K4me1 histone marks and two chromatin remodelling related marks-P300 a co-activator and DNAse hypersensitivity states (DHS) which is chromatin modification enzymes correlated to regulatory enhancer networks. The evaluation will determine how the length of k-mer affects the performances of those three marks in terms of accuracy, sensitivity and specificity. This study will reveal the limitations and strengths of k-mer feature for prediction as well as provide insights into some k-mer feature design considerations.

II. Background

Early genome-wide enhancer location prediction methods relied on properties of the DNA sequence, such as clusters of TF binding sites called the cis-regulatory module (CRM) [8] and comparative genomic approaches [9]. However, these methods do not determined about the cell-type specificity of the identified enhancers. It is also found that comparative genomic methods missed many non-conserved enhancer