Improved H3K27ac Histone Mark Prediction using K-mer Proximity Feature

Pui Kwan Fong
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
Email: amandafpk@gmail.com

Nung Kion Lee*
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
Email:nklee@fcs.unimas.my

***Abstract*—Prediction of gene regulatory elements-enhancers is computationally challenging because features associated with them are ill-understood. Several histone marks are known to be associated with enhancers locations and have been successfully used to predict multiple thousands of enhancers approximate locations. The k-mer (a short continuous nucleotides of length k) is one of the most commonly engineered features from histone sequences for machine learning task. However, usually large k-mer (i.e. $5 \leq k \leq 7$) feature set is needed to perform well and no domain knowledge is used. In this study we proposed the k-mer proximity feature which is domain dependent to represent the H3K27ac histone enrichment in DNA sequences. This feature represents the spatial content of DNA sequences. We compare the performances of using the proximity and the k-mer feature for H3K27ac marks prediction and results indicate that the proposed feature gives higher prediction accuracy rates. These findings supported that the proximity feature is a more distinguishing feature of DNA sequences with histone modification enrichment.**

I. Introduction

Application of computational intelligence methods in solving biological problems such as regulatory elements prediction [1]–[3], splice sites identification [4] and epigenetics prediction [5], [6] have been extensively studied. Among various computational methods, supervised machine learning is found to be one of the promising methods due to its capability in discovering patterns and relationships between DNA sequences data [7]. DNA sequences consist of various nucleotides (A, C, G, T) combination which form the basis of all biological function. Each organism has its own set of DNA sequences with increasing complexity for higher level organisms. For example, the length of DNA sequences in the most complex organism, human genome is approximately 3 billion base pairs (bp) [8].

DNA sequences which are initially represented using alphabets are not an appropriate input for supervised machine learning methods thus have to be converted to numerical representations [9]. Through these conversion processes, huge amount of numerical data is produced. It is not possible to interpret or utilize these rich and complex data without proper techniques to extract useful information. Various information or also known as patterns can be inferred from these DNA sequences which play an important function in determining the success of biological prediction. Therefore, feature extraction is crucial to simplify the data before it is feed into supervised machine learning to produce prediction with high accuracy.

*Corresponding Author

Numerical DNA features extracted from DNA sequences function as input variables which can be in the form of continuous, binary or categorical values [7]. Good features have characteristics such as highly informative in order to facilitate the learning of classifier and generalizable for future prediction. In addition, a set of good feature does not contain much redundancy which may increases the computational complexity. In biological prediction, features extracted from DNA sequences are mostly application-specific as different biological functions may consist different DNA patterns.

In this paper, feature extraction is performed on one of the histone modification marks namely H3K27ac. This histone is targeted as it is highly related to active enhancer sites which are found to be an important element in biological functions [10]. In addition, research on predicting this histone mark is not as comprehensive as H3K4me1 which is found to be correlated with both active and poised enhancer. Therefore, this paper focused on extracting the spatial patterns from H3K27ac sequences specifically the proximity between short DNA patterns also known as k-mers in general. It is hypothesized that conventional k-mers frequency has many weaknesses thus insufficient to represent DNA sequences for discriminating between those enriched with and without H3K27ac histone modification.

This paper is organized as follow; Section II discusses on previous works related to the methods and different types of features extracted from DNA sequences data. Section III presents a detailed framework on extraction of k-mer proximity feature from DNA sequences enriched with and without H3K27ac histone modification. Following that, Section IV includes experiments results and analysis on the predictive power of the proposed proximity k-mer feature in comparison with the conventional k-mer frequency features for prediction of H3K27ac histone modification enrichment in human $CD4^+$ T-cells dataset. The last section concludes this study and offers some directions on future work.

II. Related Works

A. Feature Extraction

Feature extraction is one of the pre-processing stage for prediction tasks using supervised machine learning methods. This stage would generate a feature vector $\mathbf{x}$ which stores all feature value of dimension $\mathbf{n}$. In feature extraction, two major steps are involved namely feature generation and feature selection [7]. Feature generation methods have been widely