**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 ≤ k ≤ 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