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

regions[10]. Consequently, additional tissue-specific information is needed for more accurate enhancer prediction and annotation. One that is typically used is the epigenetic feature such as histone modifications and chromatin remodelling data. In addition, the p300 protein-a co-activator which is recruited by TFs during the gene transcription initialization stage.

Epigenetic features consist of chromatin structure, histone modications, DNA-methylation levels and non-coding RNAs. Chromatin structure controls DNA accessibility of TFs to enhancer or other regulatory elements. DNA accessibility can be inferred as DNase I hypersensitivity (DHS)or by Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) technology. The regions detected by DNase I or FAIRE are associated with all known classes of active DNA regulatory elements, including enhancers. Sequence-specic binding TFs often recruit cofactor proteins, such as chromatin-modifying enzymes, for example: histone acetyltransferase p300. The binding of cofactors facilitates chromatin remodeling and DNA looping to form crucial enhancerpromoter interaction. Therefore, genome-wide proling of cofactor occupancy provides a general strategy for detecting enhancers.

Chromatin signatures that consist of histone tail modications are sign of active enhancer activity. These histones include H3 lysine 4 monomethylation (H3K4me1), H3K4me3 and H3K27ac[11]. These chromatin signatures can be identied by clustering analysis of histone modication ChIP-seq data. Enhancers are usually flaked by the H3K4me1 and H3K27ac, but depleted of H3K4me3. Sequences enriched with H3K4me3 are signature of promoter motifs. The different between H3k4me1 and H3k27ac is that H3k4me1 can be associated with poised but H3k27ac with active enhancer. In addition, that are others histone modifications marks that are associated with different types of enhancers depending on the cell-type and cell developmental stage. Furthermore, it is suggested these histones worked coorperatively to mark active enhancers.

III. Materials and Methods

A. K-mers counting

For a given integer k, count the frequency of all k-permutation on the set {A,C,G,T} in each positive and negative DNA sequence of epigenetic marks to generate feature vectors. These k-mers represent the content feature of the sequences. K-mers are counted using the overlapping shifting window method. Only the downloaded strand is counted for every sequence. The frequency values in a feature vector are normalized using the max-min scaling method: $(f_b - f_{min})/(f_{max} - f_{min})$, where f_b, f_{min}, f_{max} is a frequency of k-mer b, minimum k-mer frequency, maximum k-mer frequency in a sequence, respectively. Since the number of k-mers increases exponentially with k, its typical values are between 2-7 for computation feasibility. Literature studies found that k-mers of length 4 to 6 performed well in representing epigenetic features. This study will employ k-mer's length of 2-6 in our comparisons.

It is noted from the literature that besides k-mer feature extraction, the feature selection step is rarely employed. Typically a full set of k-mer feature is needed for classifier to perform well. This implies that the epigenetic sequences do not share much similarity and it may appear to be random and data specific.

B. Datasets

Our datasets are obtained from the ENCODE project which are available at UCSC Genome Browser[12]. Our evaluations are conducted using the GM12878 cell line in human assembly hg19. The positive datasets are extracted from chromosome 1 to 5. The negative data are complementary of the positive regions in the same chromosomes. Four epigenetic marks H3k4me1, H3k27ac, DHS and P300 which have shown to be associated with active enhancers are used in our comparative studies. For each selected chromosome, 2000 both positive and negative regions are randomly picked. Therefore, our database has a total of 20000 datapoints.

C. Performance measure

The support vector machine (LibSVM) [13] is used as classifier for the binary classification task. Support vector machine has been shown to perform well in bioinformatic classification problems [14]. The Radial basis is selected as a mode of kernel which is typically done via Equation 1:

$$K(x, x') = exp(\gamma \left\| x - x' \right\|^2) \tag{1}$$

In our experiment, γ=1/size of training set and cost function=1 is chosen as parameters for SVM model.

The performance measures using P number of positive and N negative test sequences are given as following:

$$Sensitivity = TP/(TP + FN) \tag{2}$$

$$Specificity = TN/(FP + TN) \tag{3}$$

$$Accuracy = (TP + TN)/(P + N) \tag{4}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

IV. Results

A. Performance Evaluation

The accuracy, sensitivity, and specificity of using k-mer feature for representing H3K4me1, p300 and DHS marks are evaluated as follows. For each epigenetic mark type, a five-folds cross validation method is used for evaluation. Sequences from one of the five selected chromosomes (chr1-chr5) are used for training while the other four chromosomes are used for testing the SVM classifier. In every run, the sequences used are randomly selected from the annotated locations in UCSC.

Figure 1 illustrates the average classifiers' performances. It is observed that, generally the average accuracy rates increase when long k-mer lengths (4 to 6) are used (Figure 1(a)). But their differences are not significantly large. The prediction is the best when classifiers are trained using 6-mers from chromosome two. Classifiers' performances using 5 or 6-mers performed better in terms of accuracy and sensitivity compare with other k-mer lengths. 7-mers was also evaluated but the results are not better than or at par to using 6-mers. The accuracy and sensitivity of using 2, 3-mers are low (< 0.75).

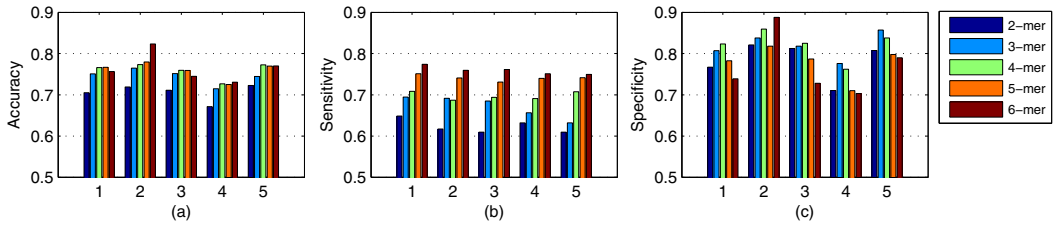

Fig. 1: The average performances of k-mer feature using H3K4me1 sequences over five runs (every run employed different set train and test sequences). The x-axis represents the chromosome number (1-5) which sequences are used for training the classifiers.

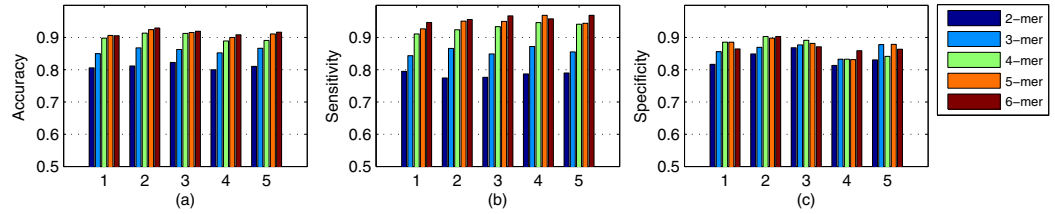

Fig. 2: The performances of k-mer feature using p300 sequences. The x-axis represents the chromosome number (1-5) which sequences are used for training the classifiers.

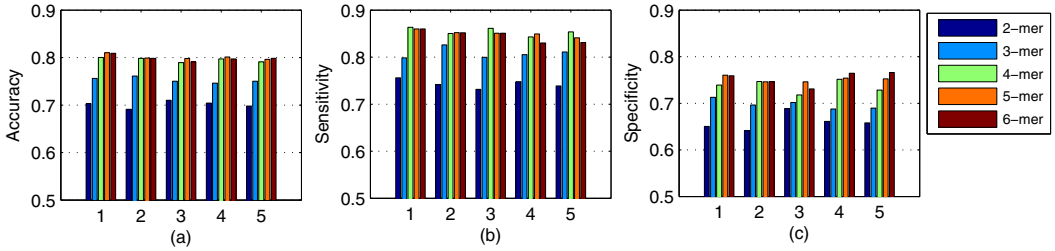

Fig. 3: The performances of k-mer feature using DHS sequences. The x-axis represents the chromosome number (1-5) which sequences are used for training the classifiers.

In terms of sensitivity (recall rates)(Figure 1(b)), the 5, 6-mer feature performed significantly better than 2, 3 and 4-mers. In terms of specificity, the results of using different k-mer lengths are inconclusive. The use of histone marks from different chromosomes have great influence on the classifiers' specificity (see Figure 1(c)).

Experiment on P300 peaks reveals that they are highly predictable by simple k-mer method. K-mer model achieved high accuracy and sensitivity in our experiment on P300 data as shown in Figure 2. Similar to the results of classifiers trained using H3K4me1 data, the 4-,5-and 6-mers feature set achieved superior performance compare with 2- and 3-mers. The average accuracy rates of 2- and 3-mer feature set performed as good as 4 to 5-mers of H3K4me1 data. This indicates that short k-mers are more discriminative for P300 compare with H3K4me1. The best average accuracy and sensitivity rate is over 0.9 using chromosome 2 and chromosome 5 for training, respectively. The specificity rates of predicting p300 are over 0.8 for all the tests.

On evaluating using the DHS sequences the k-mer feature performed the best amongst the three epigenetic marks. The accuracy rates of using 5, 6-mers achieved over 0.9 for this DHS dataset compared with not more than 0.8 for H3K4me1 and P300 (Figure 3(a)). Likewise, the top sensitivity and specificity rates are the best of the three datasets (Figure 3(b-c)). Again the 5 and 6-mer feature gave the best prediction in three of the evaluation criteria.

In the final evaluation, a dataset is prepared from extracting 2000 positive and negative p300 DNA sequences from all five chromosomes (i.e. 10000 in total for both positive and negative dataset). Then five fold cross-validation were performed in which it is ensured that each fold has both 2000 positive and negative. Figure 4 illustrates the classifiers performance. It is observed that mixing the DNA sequences from different chromosomes have not much effect on performances in comparison with using sequences from a single chromosome. For instance, comparing Figure 4(a-c) and Figure 2(a-c), we can see that accuracy, sensitivity and specificity rates for different k-mer lengths have similar performance rates.

V. Discussion

The evaluation results indicate that the k-mer feature's performances are various on the three enhancer associated marks. Overall, k-mer feature performed the best on P300 sequences followed by DHS and H3K4me1. The accuracy rates of k-mer feature for DHS and H3K4me1 sequences are still

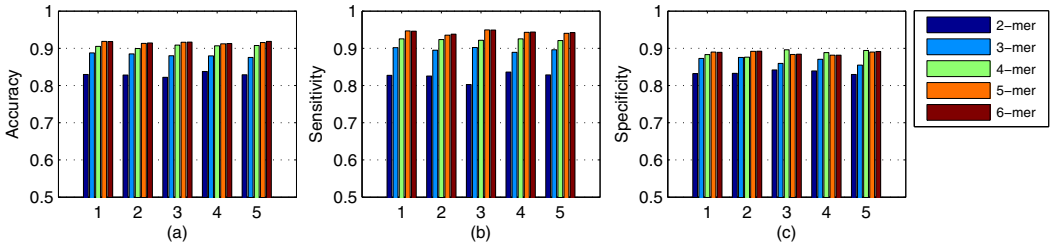

Fig. 4: The performances of k-mers feature using P300 sequences randomly selected from chr 1-5. The x-axis represents the partition numbers in five different non-overlap subsets.

relatively low. It is necessary to generate better discriminative feature set in order to improve prediction accuracy.

It is evident from this study that the suitable lengths of the k-mer feature are between 4 to 6bp. None of these lengths are dominating in all the three evaluation criteria. Short k-mer lengths (e.g., 2 or 3) produced poor accuracy, sensitivity and specificity rates.

The use of DNA sequences from different chromosomes for classifier training do produced rather similar performances implying that DNA sequences of the three marks have similar DNA content characteristic in different chromosomes. Therefore, in building classifiers, the source of training DNA sequences can be from a single chromosome or combination of chromosomes.

VI. Conclusion

Precise detection of potential candidate for enhancer regions contributes to deciphering complex regulatory network through gene expression. Between several features, epigenetic marks are proved to discern locations of functional elements, accurately. Due to cost and time associated problems with experimental data for identification of epigenetic marks, it is desired to apply computational techniques to predict regions efficiently and effectively. K-mer model is obtained to utilizing synergism between experimental and computational techniques. The model extract epigenetic marks in genomic scale. Marks of H3k4me1 are predicted by 76% accuracy on average; other marks including P300 peaks and DHS are predicted by 90% and 79%, accordingly. The experiment demonstrates the robustness of simple k-mer model as feature for classification in predication of enhancer regions coupled with epigenetic marks. Based on the marks, specific lengths of K, 4-,5-, or 6-mer in most cases, are sufficient to obtained compelling results. Nevertheless, there is still much room for improvement on accuracy, sensitivity, and specificity of these epigenetic marks prediction by using better discriminative feature.

Acknowledgment

This project is funded by the KPM grant RACE/b(2)/886/2012(04). SN is partially supported by the Universiti Malaysia Sarawak Zamalah scholarship.

References

[1] M. T. Maurano et al., "Systematic localization of common disease-associated variation in regulatory DNA," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[2] D. Shlyueva et al., "Transcriptional enhancers: from properties to genome-wide predictions," *Nat Rev Genet*, vol. 15, no. 4, pp. 272–286, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1038/nrg3682

[3] R. Jothi et al., "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data," *Nucleic Acids Research*, vol. 36, no. 16, pp. 5221–5231, 2008.

[4] A. Visel et al., "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature*, vol. 457, no. 7231, pp. 854–858, 2009.

[5] A. M. S. Shrestha et al., "A bioinformaticians guide to the forefront of suffix array construction algorithms," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 138–154, 2014.

[6] G. Leung and M. B. Eisen, "Identifying cis-regulatory sequences by word profile similarity," *PLoS ONE*, vol. 4, no. 9, p. e6901, 2009.

[7] M. Annala et al., "A linear model for transcription factor binding affinity prediction in protein binding microarrays," *PLoS ONE*, vol. 6, no. 5, p. e20059, 2011.

[8] J. Su et al., "Assessing computational methods of cis-regulatory module prediction," *PLoS Computational Biology*, vol. 6, no. 12, p. e1001020, 12 2010.

[9] K. Palin et al., "Locating potential enhancer elements by comparative genomics using the EEL software," *Nature Protocols*, vol. 1, no. 1, pp. 368–374, 2006.

[10] R. C. Hardison, "Variable evolutionary signatures at the heart of enhancers," *Nature Genetics*, vol. 42, no. 9, pp. 734–735, Sep. 2010.

[11] R. Andersson et al., "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, pp. 455–461, 2014.

[12] D. Karolchik, et al., "The UCSC genome browser database: 2008 update," *Nucleic Acids Research*, vol. 36, pp. D773–D779, 2008.

[13] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transaction of Intelligent System Technology*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[14] Z. Xing et al., "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.

[15] C.-Y. Chen et al., "Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features," *BMC Genomics*, vol. 13, no. 1, p. 152, 2012.