Bio-Medical Materials and Engineering 24 (2014) 3807–3814
DOI 10.3233/BME-141210
IOS Press

Modelling complex features from histone modification signatures using genetic algorithm for the prediction of enhancer region

Nung Kion Lee[a,*], Pui Kwan Fong[a], and Mohd Tajuddin Abdullah[b]
[a] *Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia*
[b] *Center of Tasik Kenyir Ecosystem, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia*

Abstract. Using Genetic Algorithm, this paper presents a modelling method to generate novel logical-based features from DNA sequences enriched with H3K4mel histone signatures. Current histone signature is mostly represented using k-mers content features incapable of representing all the possible complex interactions of various DNA segments. The main contributions are, among others: (a) demonstrating that there are complex interactions among sequence segments in the histone regions; (b) developing a parse tree representation of the logical complex features. The proposed novel feature is compared to the k-mers content features using datasets from the mouse (mm9) genome. Evaluation results show that the new feature improves the prediction performance as shown by f-measure for all datasets tested. Also, it is discovered that tree-based features generated from a single chromosome can be generalized to predict histone marks in other chromosomes not used in the training. These findings have a great impact on feature design considerations for histone signatures as well as other classifier design features.

Keywords: Genetic algorithm, tree-based feature, histone feature

1. Introduction

Comprehension of gene regulation involves locating the cis-acting regulatory elements comprising clusters of transcription factor binding sites (TFBS) that initiate the mechanism of gene transcription. Enhancers are a type of cis-regulatory element that promote gene expression and often are essential for eliciting complex expression patterns of developmental genes. An enhancer region typically spans a few hundred base pairs (bp) comprising clusters of TFBSs (at multiple sites) that work in cis--each site is about 6 to 20bp in length.

The first two authors contributed equally to this paper.

*Corresponding author: Nung Kion Lee, Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia. E-mail: nklee@fcs.unimas.my, Tel:+6082-584152, Fax:+6082-581567

Enhancer regions can also span from ten to hundreds of thousand bp in length–in the upper or the lower region of the genes it regulates. Thus, this becomes a challenging task for prediction because the genome-wide sequence search space is large. In addition, its DNA characteristics are ill-defined and non-specific to enhancers at large. Traditional enhancer prediction methods employ motif profiles searching for individual enhancer sites. This approach is useful but produces high false positive hits because they fail to characterize the specificity of the sites. Other approaches use additional features to reduce the occurrences of false positives such as cross-species sequence conversation analysis, identification of the presence of cis-regulatory module, and correlational analysis using epigenetic marks. This paper focuses on generating features from epigenetic marks to enhance prediction accuracy of the enhancer regions.

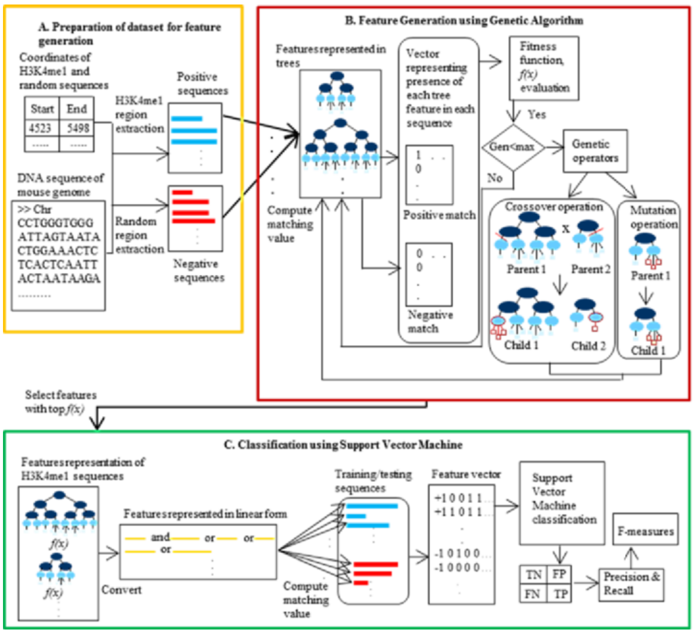
Epigenetic marks refer to elements that change in tandem with cellular activities such as gene expressions while its DNA sequences remain unchanged [1]. Currently, one of the most widely studied elements is histone modification where genomic locations are known to correlate strongly with specific enhancer regions. Existing methods using histone marks largely employs the supervised learning methods [2, 3]. Good feature representation from histone marks is definitely one of the key factors in producing good prediction results. The k-mers frequency [3] feature is typically used to represent histone features. Nevertheless, these features only capture the content composition of the histone regions but not the co-existence of DNA features and their possible interactions.

This paper hypothesizes the existence of complex and higher-order features in DNA sequences. A method is proposed to describe these complex features using parse trees generated by Genetic Algorithms (GA) [4]. The effectiveness of these tree features is determined by scrutinizing and characterizing the enrichment of H3K4me1 epigenetic marks in DNA sequences.

2. Related works

Computational enhancer prediction can generally be categorized into direct or indirect methods depending on whether the exact or the approximate locations of enhancer are inferred. Direct methods use motif profiles that employ machine learning algorithms (supervised or otherwise) to predict candidate enhancer regions or TFBSs locations. Meanwhile, indirect methods use correlational analysis of enhancer regions with some landmark DNA features for the inference of approximate locations—e.g., CpG island, chromatin or histone marks. Extensive studies on indirect methods are focused mainly in the generation and modelling of discriminative features from landmarks of supervised learning [2, 3]. Clearly, the key to success for these methods requires good representation of the DNA landmarks so that the non-landmarks could be clearly differentiated.

Genome-wide mapping of epigenetic marks presents an unprecedented opportunity for indirect enhancer prediction using epigenetic features. Significant findings from [3,5,6] conclude that the distance between H3K4me1 enrichment and enhancer regions is approximately 100 to 2000bp in length. For example, one study [3] has successfully predicted the 7361 and 7788 melanocyte enhancers in the mouse and human genomes respectively using the H3K4me1 marks.


A. Preparation of dataset for feature generation
Coordinates of H3K4me1 and random sequences
Start
End
4523
5498

H3K4me1 region extraction
Positive sequences
DNA sequence of mouse genome
>> Chr
CCTGGGTGGG
ATTAGTAATA
CTGGAAACTC
TCACTCAATT
ACTAATAAGA
....
Random region extraction
Negative sequences
B. Feature Generation using Genetic Algorithm
Features represented in trees
Compute matching value
Vector representing presence of each tree feature in each sequence
1
0
Positive match
0
0
Negative match
Fitness function, f(x) evaluation
Yes
Gen<max
No
Genetic operators
Crossover operation
X
Parent 1
Parent 2
Child 1
Child 2
Mutation operation
Parent 1
Child 1
Select features with top f(x)
C. Classification using Support Vector Machine
Features representation of H3K4me1 sequences
f(x)
f(x)
Convert
Features represented in linear form
and
or
or
or
Compute matching value
Training/testing sequences
Feature vector
+1 0 0 1 1
+1 1 0 1 1
-1 0 1 0 0
-1 0 0 0 0
Support Vector Machine classification
TN
FP
FN
TP
Precision & Recall
F-measures

Fig. 1. Schematic diagram of complex tree feature generation.

Conventionally, either ChIP-chip or ChIP-seq is used to create a high resolution mapping and profiling of these histone modification distributions [7, 5]. However, these experimental techniques are costly and incapable of culturing all possible cell conditions needed to identify active histone modification marks. Consequently, data related to histone modifications is unavailable for most organisms. Therefore, there is a pressing need to utilize computational approaches to identify and characterize DNA sequences enriched with histone modification marks.

Meanwhile, another study [8] uses content (k-mer frequency) and context (distance from the gene) based features to locate histone modification signatures using Support Vector Machine (SVM). The prediction of H3K4me1 using 20344 positive and 20401 negative sequences of the yeast genome achieves an accuracy of 90.86%. High accuracy in prediction is achieved when 9-mer frequency with and distance from the nearest annotated genes are used as features for SVM prediction. However, the performance is drastically reduced to 72.61% when only 9-mer features are involved and deteriorates further when shorter nucleotides are used. Another study [9] successfully predicts the sequences enriched with H3K4me1 in the human genome with a high area under the ROC score of 0.9. The prediction is performed on CD4$^+$ T cells where histone modification information is obtained from [7]. Instead of using k-mer features only, they integrate dinucleotide (2-mer) frequencies with wavelet features to develop a modified N-score model for histone prediction.

3. Methods and materials

3.1. GA-based feature generation

The overall workflow of the proposed method is shown in Figure 1. This approach is motivated by the discovery that functional elements are characterized by complex and possible logical interactions among the features themselves. In fact, a large body of research works could be used to support this claim, e.g., works on predicting the initializing sites for transcription, modelling of family motifs and binding sites of proteins [10]. It is also observed that features of functional elements are usually ill-defined and not well understood. It usually requires a large set of features (in the range of thousands) to represent complex properties. In brief, the proposed approach employs GA to generate logical relationship of short DNA segments (i.e., k-mers) represented by a complex tree. The short DNA segment, i.e., k-mers, comes in two patterns: length *k* continuous DNA letters (A, C, G, T) or a couple of short-length *l* k-mers separated by gaps of various sizes (< 5). The formal is labeled *pattern-1* while the latter is called *pattern-2*. Assuming that the H3K4me1 content features take up one of these patterns, logical interactions between the content features are then modelled using the logical 'AND' and 'OR' operators. The generated logical features are used to construct feature vectors for SVM training. The tree structure is used to represent the logical interactions between the two patterns (Supplementary Figure S1-C). The logical operators become the parent node of any of the two patterns in a tree (Supplementary Figure S1-C). Thus creating a logical and hierarchical relationship of important features found in the H3K4me1 sequence. The nodes and the patterns in a tree are evolved by applying the customized genetic operators (see Supplementary Figure S2-S3). Tree features (i.e. chromosome) are evaluated using a fitness function value to determine their ability to discriminate positive from negative sequences. Tree features are ranked and the top N will be selected for binary feature vector generation (Supplementary Section 1.4). SVM is subsequently trained using the binary feature vectors for classification purpose.

3.2. Fitness function

Every tree feature is assigned a fitness value where features with the larger values have higher chances to be selected for genetic operations. Eq. (1) shows the fitness function for GA comprising two sections.

$$fitness(T) = \underset{j=1\ldots N}{\arg\max} \left\{ \frac{\sum_{i=1}^{a} f(p_j, s_i^+)}{\sum_{k=1}^{b} f(p_j, s_k^-)} \right\} + \frac{\sum_{i=1}^{a} f(T, s_i^+)}{\sum_{k=1}^{b} f(T, s_k^-)} \tag{1}$$

where T represents a tree feature; N, a and b represent the total number of patterns in T, the number of positive sequences and the number of negative sequences, respectively; s_i^+ is the ith positive sequence. Whereas s_i^- is the ith negative sequence. $f(x,y)$ is a binary indicator function which returns a '1' if a feature x (i.e., a pattern or tree feature) is present in the DNA sequence y. Otherwise, it returns a '0'. The first section of the equation aims to capture individual patterns in a tree with the highest discriminative value. That is, a discriminative candidate pattern in a tree should be over-represented in a positive sequence set but is rare in a negative sequence set. This will lead to selection of trees with patterns that discriminate between a positively and negatively histone-marked DNA sequences. Nevertheless, this part of the equation only selects pattern scores with the maximum value to ensure diversity in the population. On the other hand, the second part of Eq. (1) is used to evaluate the rareness of tree features as a whole.

3.3. Datasets

Coordinate of sequences enriched with H3K4me1 are obtained from [3], cisGenome is then used to identify the peak of H3K4me1 from the ChIP-seq data of melan-a cells in the mouse genome. One thousand (from 3794) histone-marked sequences is randomly selected from Chromosome 1 for feature generation. Subsequently, negative sequences (1000) are randomly selected from coordinate of sequences flanked by H3K4me1 enrichments [3]. Another 1000 histone-marked sequences from Chromosome 2 till 6 are also downloaded for testing purposes. Meanwhile, a different negative set (1000) is randomly selected from sequences flanked by H3K4me1 enrichments for testing.

4. Results

In this section, comparison results of the tree and the k-mer feature will be presented using real dataset from the mouse genome. Experiments are also performed to determine how the number of top tree features cut-off and GA generation influence the classification performance. The precision, recall and f-measure rates are used as performance measure for all the experiments.

4.1. Parameters setting

Experiments were conducted to determine how the number of selected top tree features to generate feature vector affects the classifier's performance (Supplementary Section 2). The result found that 500 top features performed the best prediction in 5 out of 8 of the tested chromosomes (Supplementary Table S2, Figure S4). In another experiment, we determine how the number of GA generations would affect the classifier's performances. It is found that a reasonable GA generation is 30 (Supplementary Table S3, Figure S5).

4.2. Comparison with k-mer feature

To evaluate the proposed feature representation, it is compared to the widely used k-mer feature using the datasets that have been prepared. Selected 500 top tree features from chromosome 1 are used to train SVM while prediction is carried out on different chromosomes to discover the generality of these features. GA was trained with 1000 DNA sequences of positive set while the negative set consists of 1000 DNA sequences. Different sets of positive and negative sequences, 1000 in each set, are used for testing. Results from the average of 5-fold cross-validation are shown in Table 1. For the k-mer feature, 4, 5 and 6-mer are used in the evaluation. Since the total number of k-mer features depends on its length, we select at least 50% of the k-mers from each set (i.e, 4, 5, 6) as inputs to the SVM classifier. In addition to that, we also select the top 50 k-mers for benchmarking. Normalized k-mer frequencies in the input DNA dataset are calculated to serve as inputs to the classifier.

Table 1 shows the results of the comparison between the tree and the k-mer feature. The best result for each test case is highlighted in bold. It can be seen that classifiers constructed by using the tree feature achieved higher precision rates in comparison to all the k-mer features used. 5-mer and 6-mer feature performed slightly better than the tree feature in terms of recall rates.

Table 1

Comparison of precision and recall rates using tree and k-mer feature for prediction

Feature type	**Chr 2**		**Chr 3**		**Chr 4**		**Chr 5**		**Chr 6**	
	P	R	P	R	P	R	P	R	P	R
500 tree features	0.818	0.595	0.839	0.687	0.827	0.632	0.840	0.691	0.830	0.643
128 4-mer	0.775	0.631	0.79	0.689	0.784	0.666	0.792	0.695	0.78	0.650
512 5-mer	0.738	0.630	0.764	0.727	0.753	0.684	0.759	0.706	0.748	0.664
2048 6-mer	0.721	0.638	0.740	0.702	0.738	0.695	0.740	0.703	0.735	0.684
50 4-mer	0.675	0.567	0.765	0.651	0.695	0.621	0.695	0.621	0.608	0.608
50 5-mer	0.719	0.600	0.741	0.674	0.730	0.635	0.743	0.680	0.726	0.622
50 6-mer	0.674	0.613	0.677	0.620	0.678	0.624	0.687	0.649	0.668	0.595

Nevertheless, the tree feature attained better balance between precision and recall rates, given by the f-measure as depicted in Figure 2. For example, the average f-measure using the 500 tree features reached 0.729 which is significantly outperformed top 50 4-mer, 5-mer and 6-mer features by 0.081, 0.045 and 0.082 as well as top 50% 4-mer, 5-mer and 6-mer by 0.009, 0.013 and 0.020, respectively.

To further verify the proposed feature representation, classifiers are constructed using tree features produced from chromosome 2 and chromosome 7. Again, the top 500 top features are selected for classifier training with five folds cross-validation. The evaluation results in Tables 2(a) and 2(b) are consistent in terms of performance level in comparison to results obtained using features from chromosome 1. This implies that the tree feature representation is non-specific to single but other chromosomes as well.

An analysis is also performed to analyze the characteristics of the generated tree-features. It is found that most of the tree features are composed of combinations of the two patterns (Supplementary Section 4)

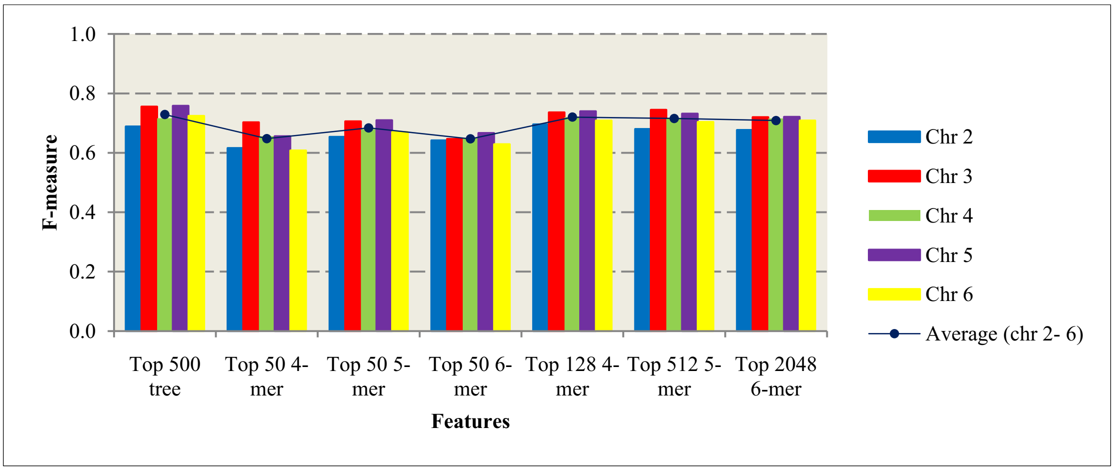
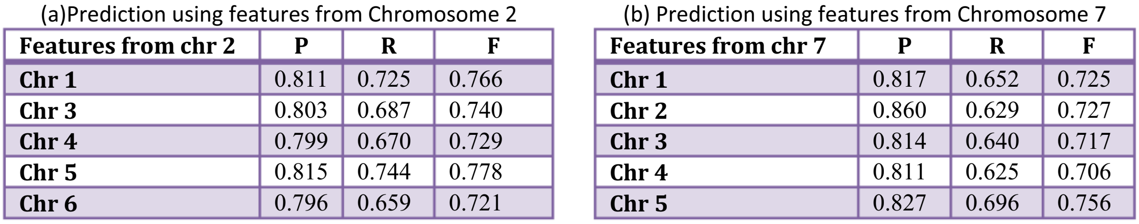


Fig. 2. Comparison of F-measure using different features for classification.

Table 2

Performance of classification using different chromosomes to generate feature



(a)Prediction using features from Chromosome 2

Features from chr 2	P	R	F
Chr 1	0.811	0.725	0.766
Chr 3	0.803	0.687	0.740
Chr 4	0.799	0.670	0.729
Chr 5	0.815	0.744	0.778
Chr 6	0.796	0.659	0.721

(b) Prediction using features from Chromosome 7

Features from chr 7	P	R	F
Chr 1	0.817	0.652	0.725
Chr 2	0.860	0.629	0.727
Chr 3	0.814	0.640	0.717
Chr 4	0.811	0.625	0.706
Chr 5	0.827	0.696	0.756

5. Discussion

The objective of this paper is to demonstrate the good performance of the proposed tree feature representation of histone marks for classification. A method for generating logical rule-based features for H3K4me1 histone mark is proposed. It is shown for the first time that complex feature modelling is necessary to effectively model histone modification sequence signatures. These complex tree features not only are able to capture a fixed number of nucleotide frequencies in DNA sequences it also represents the nucleotides logical interactions. Empirical results show that the diverse combinatorial patterns (with logical operators) perform better than the most widely used k-mer feature.

Nonetheless, the model certainly needs to be improved as the features generated are low in sensitivity. This could be attributed to the lack of diversity in the GA population in which the problem of early convergence might need to be addressed. In particular, the fitness function and the selection procedure of GA used in producing tree features could be made more robust. Thus, future works should focus on fine-tuning the GA parameters to generate features diverse enough to represent the whole search space. Last but not least, though the paper is primarily concerned with modelling features from H3K4me1 enriched sequences, it can be potentially made applicable to other types of epigenetic marks such as H3K4me3, H3Ac and P300. Combining them for generating discriminative features would be challenging but has been shown to improve sensitivity and specificity in motif prediction.

Acknowledgement

PK Fong is supported by the Malaysian MyBrain15 (MyPhD) scholarship. We thank KW Tan for proof read this article. This project is supported by the MOSTI grant, RACE/B(2)/886/2012(04). Supplementary information is available at http://www.fcshd.unimas.my/Department-of-CS/Academicians/syafiq-lee-nung-kion.html.

References

[1] A. Goldberg, C. Allis and E. Bernstein, Epigenetics: A landscape takes shape, Cell **128** (2007), 635–638.

[2] H.A. Firpi, D. Ucar and K. Tan, Discover regulatory DNA elements using chromatin signatures and artificial neural network, Bioinformatics **26** (2010), 1579–1586.

[3] D.U. Gorkin, D. Lee, X. Reed, C. Fletez-Brant, S.L. Bessling, S.K. Loftus, M.A. Beer, W.J. Pavan and A.S. Mccallion, Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes, Genome Research **22** (2012), 2290–2301.

[4] J. Holland, Adaptation in Natural and Artificial Systems, The University of Michigan Press, Ann. Arbor., 1975.

[5] N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, W.C. Ching, L.O. Barrera, S. Van Calcar, C. Qu, K. Ching, W. Wang, Z. Weng, R.D. Green and G.E. Crawford, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, Nature Genetics **39** (2007), 311–318.

[6] Y. Nie, H. Liu and X. Sun, The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines, PLoS One **8** (2013), e60002.

[7] A. Barski, S. Cuddapah, K. Cui, T. Roh, D.E. Schones, Z. Wang, G. Wei, L. Chepelev and K. Zhao, High-Resolution profiling of histone methylations in the human genome, Cell **129** (2007), 823–837.

[8] T.H. Pham, T.B. Ho, D.H. Tran and K. Satou, Prediction of histone modifications in DNA sequences, In: Yang JY, Yang MQ, Zhang, MMZY, Arabnia, HR, Deng YP, and Bourbakis, N, (editors). BIBE 2007. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering; 2007 Oct 14-17; Boston MA, 959–966.

[9] A. Kel, M. Niehof, V. Matys, R. Zemlin and J. Borlak, Genome-wide prediction of HNF4alpha functional binding sites by the use of local and global sequence context, Genome Biology **9** (2008), R36.

[10] G. Yuan, Targeted recruitment of histone modifications in humans predicted by genomic sequences, Journal of Computational Biology **16** (2009), 341–355.