Enhancer Prediction in Proboscis Monkey Genome: A Comparative Study

Conference Paper · May 2017

CITATION
1

READS
83

6 authors, including:

Norshafarina Omar
University Malaysia Sarawak
7 PUBLICATIONS **30** CITATIONS
SEE PROFILE

YEE LING Chong
30 PUBLICATIONS **169** CITATIONS
SEE PROFILE

Mohd Tajuddin Abdullah
Universiti Malaysia Terengganu
437 PUBLICATIONS **1,079** CITATIONS
SEE PROFILE

Nung Kion Lee
University Malaysia Sarawak
49 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Enhancer Prediction in Proboscis Monkey Genome: A Comparative Study

Norshafarina Omar[1], Yu Shiong, Wong[2], Xi, Li[3] and Yee Ling, Chong[4], Mohd Tajuddin Abdullah[5] and Nung Kion, Lee[6,*]

[1,2,6] *Department of Cognitive Sciences, University Malaysia Sarawak*
[3]*Life Science Informatics, Data 61, CSIRO*
[4]*Department of , University Malaysia Sarawak*
[5]*Kenyir Research Institute, University Malaysia Terengganu*
[*]*Corresponding author: nklee@unimas.my*

***Abstract*—Genome annotation is an essential task for understanding and analyzing the whole genome and its function. We have sequenced the complete proboscis Monkey (*Nasalis larvatus*) genome due to its important for medical and evolutionary studies. We have performed an initial annotation of the genes genome using the MAKER gene annotation pipeline. 3084 genes were predicted from chromosome 18 of the genome using six eukaryotic model species. Intergenic regions possibly enriched with enhancers are then predicted using five different tools: DeepBind, LS-GKM, GMFR-CNN, CSI-ANN and iEnhancer-2L. These tools find the enhancers of the complex intergenic regions based on epigenetic features, in which intergenic regions are seen as a potential region for enhancers with a certain epigenetic features bound to it. Empirical results demonstrate competitive performance using different prediction tools with multiple epigenetic features to predict the enhancers for chromosome 18 in *proboscis monkey*. Based on the finding of this study, predicted enhancers can be used for the purpose of scientific and genomic discoveries.**

***Index Terms*—enhancer annotation; enhancer prediction; motif discovery; proboscis monkey**

I. INTRODUCTION

Annotation is the first step in understanding the biological functions, identifying functional elements, and for performing scientific inquiries using the genome of a species. Fundamental annotation tasks including identifying coding and non coding DNA regions in a genome. Regulatory elements are important functional DNA sequences located in non coding region of a genome. They play a major role in regulating gene expression for the production of RNA and proteins. Regulatory elements include promoters, enhancers, proximal regulatory and distal regulatory elements. Predicting enhancer is one of the important tasks since enhancer has a capability to regulate gene expression. However, experimental approaches are costly and time consuming, therefore, a reliable and effective computational approach is needed for annotation of enhancers.

There were several studies of experimental approaches and computational approaches which have been done with enhancer prediction. Liu et al. [1] aimed to identify enhancers along with their strength by using the pseudo *k*-tuple nucelotide composition in order to formulate the DNA sequences. Meanwhile, Dai et al. [2] investigate the relationship between low methylated regions (LMRs) that derived from whole genome bisulfite sequencing (WGBS) with the enhancer prediction. Some studies learned enhancers from DNA sequence features by capturing the combination of binding sites [3]. Since enhancer tend to be bound on certain epigenetic features, [4,5,6] combined transcription factors, and chromatin histone modifications to identify enhancers and it has been found to improve the accuracy of enhancer predictions. According to Zhu et al. [7] enhancers are generalized as the peaks of the H3K4me1 enriched regions, and its been supported by [6,8] where the presence of this histone modification along with H3K4 methylation, H3K27ac and few transcription factors (TF) such as EP300, CTCF, TAL1, GAT1 have been used to predict enhancers [5].

Different enhancer prediction tools have been developed and widely used. [1] used SVM to distinguish enhancers from the whole genome sequences. In [9], they proposed an enhancer predictor called DELTA by integrating shape features of histone modifications with AdaBoost algorithm. The DeepBind [10] is one of the promising pattern discovery, a tool that is based on deep convolutional neural networks. [6] identified functional DNA features by making use of chromatin signatures and applied artificial neural network on it. Wong et al. [11] proposed an integrated enhancer predictor based on gapped motif features representation (GMFR) and deep convolutional neural network (CNN).

II. RELATED WORK

MAKER is an automated gene annotation pipeline that mainly include masking repetitive elements, *ab initio* gene prediction using programs such as: SNAP, AUGUSTUS and GeneMark, aligning the predicted *ab initio* gene models together with reference protein sequences and transcript sequence (EST/RNA) from closely related species, applying certain refinement metrics to produce the final annotated gene models. For more details about MAKER pipeline and how it works, readers can refer to refs [12,13]. Coombe et al. [14] used MAKER to annotate the coding and non-coding genes of Sitka spruce and used gene sequences of Norway spruce as evidences. MAKER also has been used to annotate the whole