Employing Genetic Algorithm to Construct Epigenetic Tree-Based Features for Enhancer Region Prediction

Conference Paper · December 2014

DOI: 10.1007/978-3-319-12643-2_48

CITATION
1

READS
83

3 authors, including:

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

SEE PROFILE

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

SEE PROFILE

Employing Genetic Algorithm to Construct Epigenetic Tree-Based Features for Enhancer Region Prediction

Pui Kwan Fong[1], Nung Kion Lee[1,*], and Mohd Tajuddin Abdullah[2]

[1] Department of Cognitive Sciences, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia
[2] Center of Tasik Kenyir EcoSystem, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia
nklee@fcs.unimas.my

Abstract. This paper presents a GA-based method to generate novel logical-based features, represented by parse trees, from DNA sequences enriched with H3K4me1 histone signatures. Current methods which mostly utilize k-mers content features are not able to represent the possible complex interaction of various DNA segments in H3K4me1 regions. We hypothesize that such complex interaction modeling is significant towards recognition of H3K4me1 marks. Our propose method employ the tree structure to model the logical relationship between k-mers from the marks. To benchmark our generated features, we compare it to the typically used k-mer content features using the mouse (mm9) genome dataset. Our results show that the logical rule features improve the performance in terms of f-measure for all the datasets tested.

Keywords: Genetic Algorithm, Feature extraction, Histone modifications.

1 Introduction

Initiation of gene transcription involves variants of regulatory elements whereby locating cis-regulatory elements enlighten the comprehension of complex gene regulation. One of the essential cis-regulatory elements known as enhancer comprises clusters of transcription factor binding sites (TFBS), each spans about 6 to 20 base pair(bp). Enhancer is capable of regulating gene expressions locating ten to hundred thousand bp away regardless of its location [1]. Locating enhancer regions remain a challenging task due to the unusual characteristics of distant-acting and short DNA sequences. In addition,the binding sites of enhancer degenerates easily yet retaining the original function [2]. Thus, it is difficult to find a general pattern of sequence to represent a specific type of enhancer.

Pioneer computational methods focus on implementing motif profiles searching to discover TFBS. Review by [3] highlights that these methods achieve high prediction accuracy for lower organisms only and often produce high false positive hits for complex organisms. Recently, the advancement of chIP-chip and

* Corresponding author.

C.K. Loo et al. (Eds.): ICONIP 2014, Part III, LNCS 8836, pp. 390–397, 2014.