BIOTECHNOLOGY & BIOTECHNOLOGICAL EQUIPMENT, 2018
https://doi.org/10.1080/13102818.2018.1438209

Taylor & Francis
Taylor & Francis Group

ARTICLE; BIOINFORMATICS

DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery

Nung Kion Lee[a], Farah Liyana Azizan[b], Yu Shiong Wong[a] and Norshafarina Omar[a]

[a]Department of Cognitive Sciences, Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia; [b]Centre For Pre-University Studies, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia

ABSTRACT

We propose an improved solution to the three-stage DNA motif prediction approach. The three-stage approach uses only a subset of input sequences for initial motif prediction, and the initial motifs obtained are employed for site detection in the remaining input subset of non-overlaps. The currently available solution is not robust because motifs obtained from the initial subset are represented as a position weight matrices, which results in high false positives. Our approach, called DeepFinder, employs deep learning neural networks with features associated with binding sites to construct a motif model. Furthermore, multiple prediction tools are used in the initial motif prediction process to obtain a higher number of positive hits. Our features are engineered from the context of binding sites, which are assumed to be enriched with specificity information of sites recognized by transcription factor proteins. DeepFinder is evaluated using several performance metrics on ten chromatin immunoprecipitation (ChIP) datasets. The results show marked improvement of our solution in comparison with the existing solution. This indicates the effectiveness and potential of our proposed DeepFinder for large-scale motif analysis.

ARTICLE HISTORY
Received 17 February 2017
Accepted 3 February 2018

KEYWORDS
Deep learning neural network; motif discovery; DNA sequence feature; chromatin immunoprecipiation-sequencing analysis

Introduction

The ability to identify transcription factor binding sites or motifs in the genome is one of the keys to decipher gene regulation mechanisms. Motifs are recurring sequence patterns in a genome and are the binding sites of transcription factors crucial for the regulation of protein production in cells. Analysis of motifs is important for advancements of medical treatment and understanding of cell processes [1]. Both wet-lab and computational techniques have been widely employed for location identification and analysis of motifs.

Motif analyses with chromatin immunoprecipitation (ChIP) combined with massive parallel DNA sequencing (ChIP-seq) followed by computational prediction have enabled rapid genome-wide location prediction of thousands of high-confidence candidate motif locations. Genome-wide datasets have posed several challenges to the computational algorithm design because of increasing complexities of the sequence search space and the requirement of a large amount of memory space. Early methods for genome-wide motif discovery are based on comparative genomic [2] and motif profile search. The comparative genomic method is based on the principle that functional elements (e.g. motifs) evolved from the common ancestors are conserved, compared to their surrounding non-functional bases. Therefore, such conserved functional elements can be identified by performing conservation analysis between sequences of orthologous or paralogous species using pair-wise and multiple sequence alignment techniques. GenomeVISTA [3], LAGAN/MLAGAN [4], MUMmer [5], AVID [6] and MULAN [7] are examples of such tools. They are mostly based on the dynamic programming algorithm such as Smith–Waterman [8] for the local alignment and Needleman–Wunsch [9] for the global alignment. To speed up the alignment of genomes, heuristic techniques such as anchoring [6], threaded blockset [10] or greedy search [11] have been employed. Although comparative genomic methods enabled identification of conserved motifs, these methods missed many functional motifs that are not conserved [12]. The second group of methods uses a database of annotated motif profiles to detect associated sites in input datasets [13–16]. Motifs are typically represented as a position weight matrix (PWM) [17] or its variants [18]. MATCH [13] combines the matrix and core similarity score for scoring a sequence; MISCORE [14] computes the average mismatch score between a sequence and motif instances for scoring; the MAST [19] score of a sequence is simply the sum of the PWM's

CONTACT Nung Kion Lee nklee@unimas.my