GMFR-CNN: An Integration of Gapped Motif Feature Representation and Deep Learning Approach for Enhancer Prediction

Yu Shiong Wong
Department of Cognitive Sciences,
Universiti Malaysia Sarawak
Sarawak, Malaysia 94300
allyson1115ar@gmail.com

Nung Kion Lee*
Department of Cognitive Sciences,
Universiti Malaysia Sarawak
Sarawak, Malaysia 94300
nklee@unimas.my

Norshafarina Omar
Department of Cognitive Sciences,
Universiti Malaysia Sarawak
Sarawak, Malaysia 94300
norshafarina.omar@gmail.com

ABSTRACT

Unravelling gene expression has become a critical procedure in bioinformatics world today and required continuous efforts to form a complete picture of enhancers. Enhancers are explicit patterns of gene expression that bound by activators to stimulate transcription. It could reside in upstream or downstream thousands of base pairs away without any fixed position. Therefore, the identification task of enhancers is extremely challenging. The inclusion of gaps in motif identification improved the overall accuracy and sensitivity, however, this feature is not fully utilised in deep learning method yet. Deep learning, is a powerful machine learning technique that has been actively used in image recognition and this technique has begun to shed light in bioinformatics. The expressiveness of deep learning enables higher feature learning from lower level ones. As a result, an integration of gapped motif feature representation (GMFR) and deep learning approach called deep convolutional neural networks (CNNs) is introduced to improve the predictive power of enhancer prediction. We called this method as GMFR-CNN. Comparative studies indicate that GMFR-CNN outperforms the other deep learning and gapped *k*-mer SVM tools with average 98% prediction accuracy. Breakthrough in deep learning technique certainly improves the performance in the near future.

CCS Concepts

• Computing methodologies→Supervised learning by classification • Computing methodologies→Motif discovery

Keywords

Convolution neural network, enhancer motifs, gapped motif feature representation.

1. INTRODUCTION

The mechanism regulates gene expression entails the biological information that is needed to unravel the system of biology. In large vertebrate genomes, it is extremely difficult to accurately predict enhancers and other regulatory elements. Three commonly cis-acting elements involved in gene expression are promoters, enhancers, and silencers. In this study, we focus on the prediction of enhancers where its orientations are not fixed and it can be located in upstream or downstream thousands of base pairs away from the gene where it regulates [1]. The activation of gene transcription can be remotely controlled by enhancers, thus, it is an extremely challenging task for biologists to identify the exact location of enhancers. The initial discovery of enhancers has more than 30 years of history and their characteristics have been enlightened but still missing the complete picture of enhancers especially in eukaryote genomes [2]. Enhancer identification is a vital task because it plays a crucial role in the medical field especially cell development, disease, and gene therapy. Experimental works involved enhancer identification are costly and time-consuming. Fortunately, with the advent of ChIP-seq technology, the properties of enhancer bound regions can be studied and develop a platform to build prediction tools [3]. Therefore, many computational methods have been introduced in recent years such as GKM-SVM [4] and DeepBind [5]. Mammalian enhancer sequences can be differentiated from random genomic loci using sequence features. Our model, GMFR-CNN is solely depended on sequence characteristics to maximise the discrimination between enhancers and non-enhancers. *K*-mer is often referring to the possible substrings of length *k* and it is commonly used to classify regulatory elements. As for DNA sequence, the nucleic acid A, C, G, and T are substituents of *k*-mer. Previous proposed *k*-mer frequency vector [6] and its refined version called gapped *k*-mer vector were implemented in SVM models to predict enhancers. *K*-mer sequence features have been used extensively by Lee et al. to predict enhancers bound by P300 in 500-2500 sequences in mouse forebrain, midbrain, and limb. However, this approach suffers from overfitting when the binding sites involved are longer than 6-8 base pairs. It is of interest to consider gapped sequence motifs when most motif finding algorithms consider continuous sequence motifs. Gaps in the motif enable the modelling of short internal loops in RNA structures and allow for a small number of mutations in the conserved regions [7]. Thus, we believe the implementation of gapped motif feature could greatly enhance motif prediction by capturing motifs not discovered in the earlier continuous sequence motifs.

The adoption of deep learning in biology has started to gain popularity especially in predicting the sequence specificities. This has been shown by the work of Alipanahi, B. et al. and Kelly, D. R. et al [8]. Both tools have produced promising results that deep

CSBio '16, December 19-22, 2016, Macau, Macao
© 2016 ACM. ISBN 978-1-4503-4794-5/16/12...$15.00
DOI: http://dx.doi.org/10.1145/3029375.3029380