UNIVERSITI MALAYSIA SARAWAK

UNIMAS

Faculty of Cognitive Sciences and Human Development

Ensemble Framework for Motif Discovery Based on Data Partitioning

Allen Choong Chieng Hoon

Doctor of Philosophy
2020

Ensemble Framework for Motif Discovery Based on Data Partitioning

Allen Choong Chieng Hoon

A thesis submitted

In fulfillment of the requirements for the degree of Doctor of Philosophy

(Cognitive Science)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
2020

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

……………………………

Signature

Name: Allen Choong Chieng Hoon

Matric No.: 12010051

Faculty of Cognitive Sciences and Human Development

Universiti Malaysia Sarawak

Date: 11 September 2020

ACKNOWLEDGEMENT

I would like to thank my supervisor Assoc. Prof. Dr. Lee Nung Kion for his helpful advice, guidance and comments throughout my PhD study. This thesis work contributes as part of the Minister of Education Malaysia, Fundamental Research Grant Scheme-FRGS/SG03(01)/1134/2014(01). I would like to thank the funding body for supporting the presentation of the conference paper I co-authored with Norshafrina Omar.

ABSTRACT

Computational DNA motif prediction is a challenging problem because motifs are short, degenerated, and are associated with ill-defined features. With the advances of genome-wide ChIP analysis technology, computational motif discovery tools are necessary to effectively tackle the large-scale datasets for motifs search. Ensemble of DNA motif discovery methods is one of the most successful approaches for motif discovery. Nevertheless, most of the existing works cannot perform motif searches in ChIP datasets because of the limited input sizes of the classical tools employed in the ensemble. Ensemble approach not only uses the results from the classical motif discovery tools, it also combines the discovered results to produce better results. The merging algorithm contributes to the prediction accuracy of the discovered motifs. The primary contribution of this thesis work is the development of an ensemble method called ENSPART with the novelty of using data partitioning technique on ChIP dataset for DNA motif prediction. The idea is to reduce the search space by portioning the input datasets into subsets and tackle by ensemble of classical motif discovery tools separately. Then, using a proposed merging algorithm, the candidate motifs are merged regardless the different lengths. Three experiments are conducted. ChIP datasets have been downloaded to evaluate the performances of the ENSPART with Receiver Operative Curves and Area Under Curve performance metrics. ENSPART was compared with the genome-wide motif discovery tools MEME-ChIP, ChIPMunk, and RSAT peak-motifs using partitioning technique. The results demonstrate that ENSPART performed significantly better than MEME-ChIP and RSAT peak-motifs in terms of the two performance metrics. Another set of datasets are gathered and sampled without partitioning. ENSPART is compared to its employed classifiers: AMD, BioProspector, MDscan, MEME-ChIP, MotifSampler, and Weeder 2. ENSPART is also compared to

MEME-ChIP, ChIPMunk, and RSAT peak-motifs without partitioning. The results show that ENSPART produces significantly better results than its individual classifiers and also MEME-ChIP, ChIPMunk, and RSAT peak-motifs. Finally, an experiment on the simulated datasets is conducted. ENSPART is compared to GimmeMotifs and MotifVoter which both are also ensemble-based tools. The results show that ENSPART produce significantly higher precision and recall rates than GimmeMotifs and MotifVoter. In conclusion, the ensemble technique is effective for DNA motif prediction, while the ChIP dataset can be tackled effectively using data partitioning techniques. The developed merging technique in ENSPART allows effective merging of same motifs from different data partitions. Such methods are generally applicable to any ensemble techniques that utilised classical motif discovery tools, or more recently, ChIP analysis tools.

Keywords: DNA motif discovery, ensemble method, data partitioning

Rangka Kerja "Ensemble" untuk Ramalan Motif DNA Berdasarkan Pembahagian Data

ABSTRAK

Ramalan motif DNA komputasi adalah sesuatu yang mencabarkan kerana motif yang pendek, merosot, dan dikaitkan dengan ciri-ciri yang tidak jelas. Kemajuan teknologi analisis ChIP yang genomik amat memerlukan kemudahan alat penemuan motif komputasi yang dapat mencari motif daripada data berskala besar dengan berkesan. Penemuan motif DNA kaedah "ensemble" adalah salah satus pendekatan yang paling berjaya untuk penemuan motif. Walau bagaimanapun, sebahagian besar penyelidikan yang sedia ada tidak dapat melakukan pencarian motif dalam dataset ChIP kerana alat-alat klasik yang digunakan dalam bersama bersifat dengan saiz input terhad. Pendekatan "ensemble" bukan hanya menggunakan hasil daripada alat penemuan motif klasik, ia juga menggabungkan hasil yang ditemui untuk keputusan yang lebih berkesan. Algoritma penggabungan menyumbang kepada ketepatan ramalan motif yang ditemui. Sumbangan utama kerja tesis ini adalah pembangunan kaedah "ensemble" yang dipanggil ENSPART dengan menggunakan teknik pemecahan data daripada dataset ChIP untuk ramalan motif DNA. Idea ini adalah untuk mengurangkan ruang carian dengan memasangkan dataset input ke dalam subset dan diatasi dengan alat penemuan motif klasik secara berasingan. Dengan menggunakan algoritma penggabungan yang dicadangkan, motif calon disatukan tanpa mengira kepanjangan yang berlainan. Tiga eksperimen telah dijalankan. Dataset ChIP telah dimuat turun untuk menilai prestasi ENSPART dengan metrik "Receiver Operative Curves" dan "Area Under Curve". ENSPART telah dibandingkan dengan alat penemuan motif MEME-ChIP, ChIPMunk, dan RSAT peak-motifs menggunakan teknik pemecahan. Hasilnya menunjukkan bahawa penggunaan ENSPART lebih berkesan daripada MEME-ChIP dan RSAT peak-motifs dari segi dua metrik prestasi. Satu lagi kumpulan dataset dikumpulkan dan disampel tanpa *pemecahan. ENSPART dibandingkan dengan pengelas yang telah digunakan, iaitu AMD, BioProspector, MDscan, MEME-ChIP, MotifSampler, dan Weeder 2. ENSPART juga dibandingkan dengan MEME-ChIP, ChIPMunk, dan RSAT peak-motifs tanpa pemecahan dataset. Hasilnya menunjukkan bahawa ENSPART menghasilkan keputusan yang lebih baik dan mempunyai kesan ketara daripada pengkelas individu dan juga MEME-ChIP, ChIPMunk, dan RSAT peak-motifs. Akhir sekali, eksperimen dengan dataset simulasi telah dijalankan. ENSPART dibandingkan dengan dua alat yang berasaskan kaedah "ensemble" iaitu GimmeMotifs dan MotifVoter. Keputusan menunjukkan bahawa ENSPART mempunyai kadar ketepatan dan pengingatan yang lebih tinggi daripada GimmeMotifs dan MotifVoter. Kesimpulannya, teknik "ensemble" adalah berkesan untuk ramalan motif DNA, manakala ChIP dataset dapat diatasi secara berkesan dengan menggunakan teknik pembahagian data. Teknik penggabungan dalam ENSPART berkesan dalam menggabungkan motif yang sama hasil daripada pembahagian data yang berlainan. Secara amnya, kaedah-kaedah sedemikian dapat digunakan di mana-mana teknik "ensemble" yang menggunakan alat penemuan motif klasik atau alat analisis ChIP yang baru secara amnya.*

Kata kunci: *Penemuan motif DNA, kaedah ensemble, pemecahan dataset*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Proteins are essential biopolymers in cells as the building blocks of various organs or tissues as well as essential component of enzymes. Proteins are produced through a process known as gene-expression, which involves decoding the information stored in protein coding genes in genomes. The two steps involved are transcription and translation. Transcription is a step that replicates the exact copy of genetic codes in the gene into messenger RNA, where the translation step decodes the information in the messenger RNA into proteins. Transcription factor (TF) proteins control when and to what extent each gene is transcribed. The short sequences (i.e. 6–12 bp or base pair) in a genome that are bound by TFs for regulating gene-expression are called transcription factor binding sites (TFBSs) of motifs which are located in the gene's upstream or downstream. There are various types of motifs in the DNA sequences, such as the promoter, silencer, enhancer, insulator, proximal, and distal regulatory motifs. Predicting transcription factors is essential so that biologists are able to study the various diseases such as cancers (Lanchantin, Singh, Wang, & Qi, 2016; Shlyueva, Stampfel, & Stark, 2014; Whitaker, Nguyen, Zhu, Wildberg, & Wang, 2015). Biologists are able to prepare the sequence regions that are anticipated to contain the TFBSs through wet-lab technology (N. K. Lee & Choong, 2013). However, wet-lab experiments to identify the motifs are costly and time-consuming (D. Wang & Do, 2012). Therefore, computational motif analysis techniques are necessary to predict the candidate motifs before further verification (N. K. Lee, Choong, & Omar, 2016). That would allow rapid analysis of transcription factors binding sites in genomic datasets.

Computational motif discovery is a non-deterministic polynomial-time hard (NP-hard) problem (Rigoutsos & Floratos, 1998) because motifs are short (5–20 bp or base pair) and degenerated (Jin, O'Geen, Iyengar, Green, & Farnham, 2007). Dozens of computational tools have been developed to predict the location of TFBSs (Das & Dai, 2007; Lihu & Holban, 2015; Salekin, Zhang, & Huang, 2017; Tran & Huang, 2014). *De novo* motif discovery tools predict novel motifs representing binding sites using certain algorithms. Given an input DNA dataset which contains the binding sites of a TF and its co-factors, computational tools return the most overrepresented repeating sequence patterns or motifs, in the DNA sequences. The predicted candidate motifs can be verified by biologists for functional roles (Bailey, 2011; Satya & Mukherjee, 2004).

A set of motifs which cooperates together is known as *cis*-regulatory module (CRM) (Klepper, Sandve, Abul, Johansen, & Drablos, 2008). There are several types of CRMs: enhancer, silencer, and insulator (Maston, Evans, & Green, 2006). Enhancers are genomic regions that controls the timing, amplitude, and cell-type specific gene expression (Erwin et al., 2014; C. Wang, Zhang, & Zhang, 2013). Because of the role of these enhancers, they are of great interest to understand the evolution and diseases like cancer (Shlyueva et al., 2014). By studying enhancers, biologists are able to understand the development of DNA in order to foresee the tissue specific activity of regulatory elements (Ghandi, Lee, Mohammad-Noori, & Beer, 2014). Enhancers are usually found at distal location from promoter in non-coding regions (C. Wang et al., 2013). They can be located in megabases away from the target genes (Noonan & McCallion, 2010) or located at other chromosomes (Lomvardas et al., 2006). Furthermore, there is no single type of data that is adequate to identify all the enhancers. As a result, enhancers are difficult to be identified (Erwin et al., 2014).

Traditional motif discovery tools are broadly categorised into probabilistic and enumerative approaches. Probabilistic approach uses position weight matrix (PWM) to represent the probability of the nucleotides (A, C, G, and T) occur in the data sequence (Das & Dai, 2007; N. K. Lee & Wang, 2011). Probabilistic approach implements stochastic method such as expectation maximization (EM) and Gibbs sampling (Das & Dai, 2007) while enumerative approach performs exhaustive matching of the nucleotides that are commonly enumerated as A, C, G, and T (Das & Dai, 2007; Kuksa & Pavlovic, 2010; Sandve & Drabløs, 2006).

Artificial intelligence (AI) approaches are also been widely used for DNA motif discovery. Applying AI techniques in motif discovery requires different definition of the problem. For instance, clustering algorithms can be employed to cluster k-mers in a set of DNA sequences based on the k-mers similarities, self-organizing map (SOM) (Mahony, Benos, Smith, & Golden, 2006) has been employed in motif discovery. Furthermore, a motif that is represented as PWM can be assumed as the population in Genetic Algorithm (GA). By using GA, the motifs can be optimised and discovered (L. Li, 2009; L. Li, Liang, & Bass, 2007; F. F. M. Liu, Tsai, Chen, Chen, & Shih, 2004; Z. Wei & Jensen, 2006) in motif discovery.

Motif discovery can be considered as a machine learning task (Brazma, Jonassen, Eidhammer, & Gilbert, 1998) because discovering the motifs is extracting the general rules from the dataset. The sequences that contain the motifs are the positive sequences, while the background sequences are the negative sequences. Thus, the objective of the motif discovery is to identify the motifs through the training from these positive and negative datasets. Recently, supervised learning, especially deep learning, has shown good results in bioinformatics in recent years (Alipanahi, Delong, Weirauch, & Frey, 2015; Eser & Churchman, 2016; Kelley, Snoek, & Rinn, 2015; Qin & Feng, 2017; J. Zhou & Troyanskaya, 2015).

Ensemble approach is a machine learning that uses multiple classifiers to produce a new classifier (Hu, Li, & Kihara, 2005). Unlike hybrid algorithm, ensemble approach does not combine the algorithms to produce a new algorithms. Ensemble approach can be applied in motif discovery by using multiple *de novo* motif discovery tools to discover the candidate motifs. Hence, each individual motif discovery tool or individual classifier can retain its strength. Ensemble approaches have been employed in many previous works and demonstrated excellent performances (Hu et al., 2005; Hu, Yang, & Kihara, 2006; Jin, Apostolos, Nagisetty, & Farnham, 2009; Kuttippurathu et al., 2011; Romer, Kayombya, & Fraenkel, 2007; Wijaya, Yiu, Son, Kanagasabai, & Sung, 2008; Yanover, Singh, & Zaslavsky, 2009). In ensemble approaches, the results from each classifier are combined to produce better results discovered by individual classifier. This allows ensemble learning superior to a single *de novo* motif discovery tool. Furthermore, the ensemble approach is flexible to employ different *de novo* motif discovery tools.

1.2 Problem statements and motivation

1.2.1 Evidences

Motif discovery is a NP-hard problem, because the motifs are short and degenerated (Jin et al., 2007). While many tools have been proposed, the ensemble approaches have shown better overall performances (Hu et al., 2005; Jin et al., 2009; Kuttippurathu et al., 2011; van Heeringen & Veenstra, 2011; Wijaya et al., 2008). One of the reasons is ensemble approaches utilised multiple types of motif discovery algorithms and therefore can predict motifs of different characteristics in dataset. They showed good performance on motif discovery, but existing methods are not designed for large-scale genomic datasets. In addition, these motif discovery tools can only accept limited size of inputs with few hundreds sequences (Zambelli, Pesole, & Pavesi, 2013). In the post ChIP sequencing (ChIP-seq) era, the genome-wide transcription factor binding region datasets have become available. Those datasets typically have hundreds to multiple thousands of sequences. While there are dozens of standalone motif discovery tools have been proposed to enable motif discovery in the large-scale datasets (Haudry, Ramialison, Paten, Wittbrodt, & Ettwiller, 2010; Kulakovskiy, Boeva, Favorov, & Makeev, 2010; Shi et al., 2011; Bailey, 2011), it is hypothesised the ensemble technique has an edge in term of sensitivity and specificity. There are evident in many past studies (Hu et al., 2005, 2006; Jin et al., 2009; Kuttippurathu et al., 2011; Romer et al., 2007; Wijaya et al., 2008; Yanover et al., 2009) that ensemble approaches performed significantly better than any single tool alone. This owing to the fact that different tools might be able to search for motifs with different characteristics, for example, short versus long motifs, conserved versus weakly conserved, or dependent and non-dependent between nucleotides in motifs. Nonetheless, there is a lack of study that demonstrates the potentiality of ensemble approach on motif discovery towards the large-scale genomic datasets. Most existing ensemble approaches developed pre-ChIP-seq era are deemed infeasible due to the limitation of the individual motif discovery tools to search for motifs in the complex, large search space, and the requirement of high memory resource.

1.2.2 Limited input size

While there have been several ensemble methods (e.g. ChIPMotifs, GimmeMotifs, and CompleteMOTIFS) developed for the ChIP-seq dataset motif analysis, they have restricted the input sizes for searching to ensure the result can be completed within reasonable time. Therefore, it is necessary to propose an ensemble method that can accept the input with large sizes and at the same time is able to employ the pre-ChIP-seq motif discovery tools.

1.2.3 Effective merging of large number of intermediate motifs

There are two common models can be used to represent the motifs: profile and consensus. Because of ensemble approach uses multiple tools, it is not restricted to accept only certain motif models produced by the individual tools. This increases the technical challenge of designing an ensemble method, because merging different motif model representation is not a straight forward task. The merging of the motif requires common representation, which involves conversion of one model to another. Besides that, similar or identical motifs would be merged. A measurement is necessary to compute the similarity of the motifs. Furthermore, merging condition needs to be defined, so that only when the condition is fulfilled, the motifs should be merged. The merging algorithms will also determine characteristic of the final output. For example, let $merge(a, b) = (a + b)/2$ as a merging function, where a and b are two similar motifs. In order to merge a group of similar motifs, $merge(a, b)$ will be called repetitively until all similar motifs are merged. This indicates that the motif being merged last has the largest weight on the final output. Contrarily, let $merge(L) = (\sum l)/n$ as a merging function, where L is a list of similar motifs, $\sum l$ is the summation of similar motifs, and n is the number of motifs. By using this formula, every similar motif has equal weight on the final output. Therefore, different merging algorithm will produce motifs differently. Existing merging methods are only suitable for merging small number of motifs from the whole input set. With large dataset, more intermediate motifs could be discovered and hence a more effective merging method is needed.

To address the problems identified in the existing approach, we are motivated by the use of data partitioning approaches in clustering. Data partitioning is a promising solution as it divides the search space into smaller search space. By partitioning the datasets, it not only allows large-scale datasets to be scanned, but it also allows the usage of traditional motif discovery tools on the large-scale datasets. Traditional motif discovery tools were proved to be useful for the pre-ChIP-seq era datasets. This also implies that ensemble approach with data partitioning is potential to solve large-scale datasets motif discovery problem by using any individual motif discovery tools, as long as discovered motifs from the partitioned datasets are able to be merged.

1.3 Research questions

The followings are the research questions derived from problem statements:

i. How to discover the motifs using pre-genomic era or pre-ChIP-seq individual motif discovery tools on the large-scale datasets?

ii. What is the algorithm to combine the discovered motifs from individual motif discovery tools regardless the difference of motif representations?

iii. Does ensemble approach based on data partitioning and multiple merging has better performance comparing to existing ensemble approaches and genome-scale motif discovery approaches?