Expert Systems with Applications 38 (2011) 11160–11166

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Assessment of clustering algorithms for unsupervised transcription factor binding site discovery

Mustafa Karabulut [a], Turgay Ibrikci [b,*]

[a] Vocational High School, Gaziantep University, 27310 Gaziantep, Turkey
[b] Department of Electrical-Electronics Engineering, Çukurova University, 01330 Adana, Turkey

ARTICLE INFO

Keywords:
Motif finding
Clustering
Comparison
Transcription factor
Machine learning

ABSTRACT

Identification of transcription factor binding sites is a key task to understand gene regulation mechanism to discover gene networks and functions. Clustering approach is proved to be useful when finding such patterns residing in promoter regions of co-regulated genes. Four clustering algorithms, Self-Organizing Map, K-Means, Fuzzy C-Means and Expectation-Maximization are studied in this paper to discover motifs in datasets extracted from *Saccharomyces cerevisiae*, *Escherichia coli*, *Droshophila melanogaster* and *Homo sapiens* DNA sequences. Required modifications to clustering algorithms in order to adapt them to motif finding task are presented through the paper. Then, their motif-finding performances are discussed carefully and evaluated against a popular motif-finding method, MEME.

1. Introduction

Transcription factors (TFs) are key elements that play important role in gene regulation mechanism. During the regulation process these factors bind to specific regions of DNA sequence which are called transcription factor binding sites (TFBS) that reside in the upstream regions of the genes to be regulated. A TF may bind to several TFBS of a well conserved sequence pattern, which is generally called a motif, with some variability. The identification of such motifs is crucial to understand the complex gene regulatory networks and gene functions specifically for putatively co-regulated genes.

The determination of TFBS via computational methods is preferable to experiments in laboratory conditions since the latter is labor intensive and usually more expensive to implement than the former one. Thus, in the literature, numerous computational tools have been developed for the motif finding task. As far as their underlying search strategies are concerned, these tools are generally classified into two major categories (Das & Dai, 2007): probabilistic methods that attempt to find optimal motifs by usually optimizing a Position Weight Matrix (PWM) and word-based methods that mainly rely on exhaustive enumeration or counting frequencies. Probabilistic algorithms are reported to be more appropriate for longer motifs (e.g., 10–22 nucleotides long) and sequences, whereas word-based methods are experimentally proved to be successful at shorter motifs specifically for DNA sequences of prokaryotic organisms (Das & Dai, 2007). Nonetheless, regardless of its adopted model, no algorithm alone is reported to be sufficient for predicting accurate motifs for every condition (Hu, Li, & Kihara, 2005; Osada, Zaslavsky, & Singh, 2004) Thus, the motif discovery problem still remains to be challenging for researchers.

In addition to the methods of above mentioned two major categories, machine learning algorithms have recently shown a promising direction to practice unsupervised motif discovery. In some studies (Ganesh, Siegele, & Ioerger, 2003; Kankainen & Loytynoja, 2007), machine learning methods such as clustering algorithms are utilized as post-processing or pre-processing tools to detect motifs in given sequences. Moreover, a paper (Kankainen & Loytynoja, 2007) has shown that a clustering algorithm alone can be used to find optimal motifs in promoter regions of co-regulated genes, and also to find protein sequence motifs (Hardik & Sun, 2005). For instance the Self-Organizing Map (SOM) is employed by Liu, Xiong, DasGupta, and Zhang (2006) and Mahony, Hendrix, Smith, and Golden (2005) in their studies and the SOM is reported to be appropriate to find TFBS reside in promoter regions of co-regulated genes of prokaryotic *Saccharomyces cerevisiae* organism. In general, it is understood from such papers that subsequence clustering is an efficient strategy to compose a motif pattern, or in other words to build a PWM, from TFBS being sought. Consequently, it is clear that the machine learning methods, specifically clustering algorithms, are suitable means to support unsupervised identification of TFBS in DNA sequences.

This study is primarily inspired from the study (Mahony et al., 2005) that reports a pure clustering strategy is efficient and satisfactory for *de novo* motif discovery. The main goal of the paper is to compare and evaluate performances of some well-known

* Corresponding author. Tel.: +90 5377720027.
E-mail address: ibrikci@cu.edu.tr (T. Ibrikci).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.161

clustering algorithms at motif-finding and DNA subsequence clustering. The clustering algorithms, however, are mainly targeted to work within high-dimensional vector space of numerical values and not suitable to process DNA subsequences with the consideration of background model. Thus, some novel contribution to each considered clustering algorithm in order to make them work with strings of DNA alphabet will be presented in the paper. In general, such modifications include adaptation of distance function and forming a cluster with given inputs where the cluster is actually a PWM and the input is a DNA subsequence.

For the mentioned goals, the paper will initially introduce the relevant clustering algorithms, SOM, K-Means, Fuzzy C-Means (FCM) and Expectation-Maximization/Gaussian Mixture Models (EM/GMM). This essential information and the required modifications to the algorithms will be available to the reader through Section 2. Experiments over promoter DNA sequences of Yeast, *Escherichia coli*, Fly and Human species and the evaluation of the experimental results are given in Section 3. The performances of the presented clustering algorithms will also be compared to the results of a well known motif discovery tool, MEME (Bailey & Elkan, 1994).

2. Methods

This paper considers the motif finding problem as finding n number of l-length DNA subsequences from a set of promoter sequences of putatively co-regulated genes. The sought subsequences, i.e., transcription binding sites, are overrepresented in the given sequences and have a common pattern of highly conserved nucleotide positions with some variability. Thus, main idea of the paper is clustering subsequences of the same motif pattern, in other words aligning or locally aligning sequences will reveal the sought motif. However, in real-life scenarios low signal-to-noise ratio existing in the given input sequences prevents a straightforward clustering scheme from extracting potential motifs. In order to overcome such an issue, a common way is to utilize a background model that generally provides adequate information to distinguish statistically important sequence features from trivial ones. A background model is organism specific statistical information of nucleotide occurrences and it is extracted with the consideration of the whole genome of the organism.

Under these circumstances, DNA motif finding problem can be summarized more formally as uncovering a PWM of a sought motif pattern. Such a PWM will be based upon locally aligned l-length DNA subsequences. A PWM is thus an $l \times 4$ matrix of probabilities of nucleotides at each position i where $1 < i < l$. As stated before, removing the bias towards the background noise requires taking background model into account and hence PWM is characterized as:

$$m_{i,k} = \log\left(\frac{f_{i,k}}{p_b^i}\right), \tag{1}$$

where m is the $l \times 4$ PWM, f is the $l \times 4$ position frequency matrix, p_b^i denotes the background frequency of letter $b \in \{A, C, G, T\}$. Third order Hidden Markov Model (HMM) of the whole genomic sequence of the organism is utilized to obtain the background frequencies of the letters, i.e., nucleotides, is utilized as background model.

This l-length subsequence clustering process will produce a set of clusters where each cluster center is a PWM to represent a motif model. In this system, potential PWMs that represent a statistically interesting sequence pattern can be distinguished by ranking PWMs with a scoring scheme. Most frequent scoring schemes from the literature (Stormo, 2000) focus upon the point that a motif pattern should be distinct from the background model where the distinction comes with a simple characteristics: high information content. In order to reveal potential motifs, this paper calculates *z-score* of each PWM after the clustering phase is accomplished and then PWMs with highest *z-score* are taken into account as motif candidates:

$$z\text{-}score = \frac{O - E}{\sigma}, \tag{2}$$

where O stands for the number of subsequences associated with the PWM. In order to calculate parameters E and σ, artificial sequences are generated by using the background model and given to the clustering scheme. After all artificial subsequences are associated with the final clusters, E represents the number of artificial subsequences coincide to the node and σ shows the standard-deviation of the coincidence. In conclusion, the PWMs are ranked with *z-score*s and the highest *z-score*s represent the most probable motif candidates of the algorithm. Please note that information content value in this case is directly proportional with *z-score*.

Clustering is at the core of described method given in this section. So, this paper implements Self-Organizing Map (SOM, Fuzzy C-Means (FCM), K-Means and Gaussian Mixture Models (GMM) based Expectation-Maximization (EM) clustering schemes. Following subsections will provide implementation details of each clustering algorithm utilized in this study.

2.1. Self-Organizing Map (SOM)

Self-Organizing Map (SOM) is a sort of neural-network that is mainly used for visualization, dimension reduction and data compression (Kohonen, 1998). SOM generally takes inputs of high-dimensional space and consequently maps these inputs into a lower dimensional space. In order to accomplish such a projection of inputs, SOM employs an input layer of vectors with the same dimensionality of inputs and an output layer of nodes interconnected with input layer. In most of the applications the output layer is chosen as low dimensionality as 2D planar grid of nodes to provide easy interpretation of transformation products.

The basic SOM algorithm can be summarized as: (a) randomly initialized weight vectors in the input layer are fed with one input at a time, (b) closest weight vector to the input, in other words winner node or best matching unit (BMU) is chosen, (c) winner node and its topological neighbors are updated with the input, (d) steps a–c are repeated for each input for a number of times which is called training. More formally, the index of BMU is determined via:

$$c = \arg\, min(dist(x_i, n_k)), \tag{3}$$

where x_i represents an input to be compared to n_k which is a node on SOM output layer; *dist* stands for an appropriate distance function, most commonly Euclidean.

2.1.1. Algorithm adaptations for motif finding strategy

In the application of motif finding the basic SOM algorithm flow remains almost the same; SOM, however, is mostly designed for numerical inputs and thus some modifications should be applied in order to make it work for subsequence clustering with PWMs. For motif finding, where the input space consists of subsequences extracted from the given promoter sequences of putatively co-regulated genes, each node at the output layer of SOM is associated with a randomly initialized PWM (a). Comparing an input subsequence x, which is an l-length string of nucleotides A, C, G and T in an arbitrary order, with an $l \times 4$ PWM in order to find closest PWM to the given input requires a likelihood function to adapt Eq. (3) to motif finding procedure: