Computational Discovery of Motifs Using Hierarchical Clustering Techniques

Dianhui Wang and Nung Kion Lee
Department of Computer Science and Computer Engineering
La Trobe University, Melbourne, VIC 3086, Australia
Email: dh.wang@latrobe.edu.au

Abstract

Discovery of motifs plays a key role in understanding gene regulation in organisms. Existing tools for motif discovery demonstrate some weaknesses in dealing with reliability and scalability. Therefore, development of advanced algorithms for resolving this problem will be useful. This paper aims to develop data mining techniques for discovering motifs. A mismatch based hierarchical clustering algorithm is proposed in this paper, where three heuristic rules for classifying clusters and a post-processing for ranking and refining the clusters are employed in the algorithm. Our algorithm is evaluated using two sets of DNA sequences with comparisons. Results demonstrate that the proposed techniques in this paper outperform MEME, AlignACE and SOMBRERO for most of the testing datasets.

1. Introduction

It is known that gene expressions are regulated by transcription factors (TFs) binding to specific transcription factor binding sites (TFBS) or motifs in promoter regions [15]. Therefore, the discovery of binding sites is critically important to study the problem of gene regulation. Experimental approaches for discovering and verifying TFBS are quite time consuming and costly. In recent years, considerable attentions have been paid to computational approaches for resolving this problem [18, 23]. The most common approach for computational discovery of motifs is to collect a set of upstream sequences which are associated with a set of genes having similar functional annotation or gene expression. Subsequently, searching algorithms can be employed to identify the over-represented motifs according to motif model optimization techniques and/or statistical criteria [11, 14]. Due to uncertainties occurring in motif presentation and low signal-to-noise rate, it becomes a challenging task to distinguish true motifs from false positive patterns. Although there are many tools available on-line to extract motifs for a given DNA dataset, it is still difficult to achieve reliable solutions with higher prediction accuracy. Furthermore, most of existing tools for mining motifs demonstrate inability to get feasible solutions for large scope of datasets.

Data mining techniques aim to discover significant or interesting patterns from databases. Clustering techniques are powerful tools for knowledge acquisition from unlabeled data. It has been shown that clustering techniques are effective for pattern discovery in biological sequences [6, 11, 12]. So far, most of the work related to this study are associated with Kohonen's Self-Organizing Map (SOM) [10]. SOM is a powerful tool for clustering data and visualizing the results in low dimensional space. The standard [1] and augmented hierarchical SOM networks [6, 12] have been applied to motif discoveries. The idea behind hierarchical SOM approaches is to successively partition the whole data set into clusters located at different levels according to a similarity metric and criteria for branching and merging clusters. Recently, [12] and [14] made progresses in this direction. In [14], a SOM-based algorithm called SOMBRERO using a new cluster prototype was introduced to replace the standard weight vector representation; whereas in later work, a new hierarchical learning algorithm with top-down subnet-layer network architecture was proposed. The SOMBRERO performs reasonably well for the *Saccharomyces cerevisiae* and *Drosophila melanogaster* datasets, but sometimes results in higher false positive rate due to non-optimal architectural setup and heuristic updating rules. Similarly, the algorithm proposed in [12] is sensitive to the initial setup of the network, the order of inputs applied to the network, and the threshold settings for branching nodes.

The remainder of this paper is organized as follows: Section 2 gives some preliminaries related to computational approaches for motif discoveries; Section 3 details the proposed hierarchical clustering algorithm; and Section 4 reports some experimental results with comparisons, where two sets of DNA sequences and three well-known tools for finding motifs were employed for performance evaluation. A brief discussion is also given in this section.

1550-4786/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDM.2008.21