A Parameterized Algorithm for Predicting Transcription Factor Binding Sites

Yinglei Song¹, Changbao Wang¹, and Junfeng Qu²

¹ School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China
syinglei2013@163.com, wangchangbao@126.com
² Department of Information Technology, Clayton State University, Morrow, GA 30260, USA
jqu@clayton.edu

Abstract. In this paper, we study the Transcription Factor Binding Sites (TFBS) prediction problem in bioinformatics. We develop a novel parameterized approach that can efficiently explore the space of all possible locations of TFBSs in a set of homologous sequences with high accuracy. The exploration is performed by an ensemble of a few Hidden Markov Models (HMM), where the size of the ensemble is the parameter of the algorithm. The ensemble is initially constructed through the local alignments between two sequences that have the lowest similarity value in the sequence set, the parameters of each HMM in the ensemble are revised when the remaining sequences in the set are scanned through by it one by one. A list of possible TFBSs are generated when all sequences in the set have been processed by the ensemble. Testing results showed that this approach can accurately handle the cases where a single sequence may contain multiple binding sites and thus has advantages over most of the existing approaches when a sequence may contain multiple binding sites.

Keywords: parameterized algorithm, Hidden Markov Model (HMM), Transcription factor binding site, dynamic programming.

1 Introduction

Transcription Factor Binding Sites (TFBS) are subsequences found in the upstream region of genes in DNA genomes. A transcription factor, which is a specialized protein molecule, may bind to the nucleotides in the subsequences and thus may affect some relevant biological processes. Research in molecular biology has revealed that transcription factor binding sites are important for many biological processes, including gene expression and regulation. An accurate identification of TFBSs is thus important for understanding the biological mechanism of gene expression and regulation. Classical experimental methods are time consuming and expensive [6,7]. Recently, a few new experimental methods such as ChIP-chip and ChIP-seq have been developed for TFBS identification [17]. Although the throughput of these methods is high, processing the large amount of complex data generated by...
these methods remains a challenging task [17]. Computational methods that can accurately and efficiently identify TFBSs from homologous sequences are thus still convenient and important alternative approaches to rapid identification of TFBSs.

Since TFBSs for the same transcription factor have similar sequence content in homologous sequences, the most often used computational approaches make the prediction by analyzing a set of homologous sequences and identifying subsequences that are similar in content. The locations of a TFBS may vary in different homologous sequences. To determine the location of a TFBS in each sequence, we need to evaluate all possible starting locations among all sequences to find the optimal solution. The total number of combinations of subsequences that need to be examined is exponential and exhaustively enumerating all of them is obviously impractical when the number or the lengths of the sequences are large. To avoid exhaustive search, a large number of heuristics have been developed to reduce the size of the search space, such as Gibbs sampling based approaches AlignACE [19], BioProspector [16], Gibbs Motif sampler [15], expectation maximization based models [1, 2], greedy approaches such as Consensus [8], and genetic algorithm based approaches such as FMGA [14] and MDGA [4].

Of all these approaches and software tools, Gibbs Motif sampler is a tool based on a stochastic approach. It computes the binding site locations by Gibbs sampling [15, 16, 19]. Consensus uses a greedy algorithm to align functionally related sequences and applies the algorithm to identify the binding sites for the E. coli CRP protein [8]. MEME+ [2] uses Expectation Maximization technique to fit a two component mixture model and the model is then used to find TFBSs. MEME+ achieves higher accuracy than its earlier version MEME [1]. However, the prediction accuracy is still not satisfactory.

Genetic algorithm (GA) simulates the Darwin evolutionary process to find an approximate optimal solution for an optimization problem. GA based approaches have been successfully used to solve the TFBS predicting problem, such as FMGA [14] and MDGA [4]. FMGA was declared to have better performance than Gibbs Motif Sampler [15] in terms of both prediction accuracy and computation efficiency. MDGA [4] is another program that uses genetic algorithms to predict TFBSs in homologous sequences. During the evolutionary process, MDGA uses information content to evaluate each individual in the population. MDGA is able to achieve higher prediction accuracy than Gibbs sampling based approaches while using a less amount of computation time.

So far, most of the existing approaches use heuristics to reduce the size of the search space. However, heuristics employed by these approaches may also adversely affect the prediction accuracy. For example, GA based prediction tools cannot guarantee the prediction results are the same for different runs of the program. A well defined strategy that can be used to efficiently explore the search space and can generate deterministic and highly accurate prediction results is thus necessary to further improve the performance of prediction tools.

Recent work has shown that an ensemble of HMMs can be effectively used to improve the accuracy of protein sequence alignment [21]. In this paper, we develop a new parameterized algorithm that can predict the locations of TFBSs with an ensemble of Hidden Markov Models (HMMs), where the size of the ensemble is the parameter. The approach uses an ensemble of profile HMMs to generate a list of