Comparative Analysis of Regulatory Motif Discovery Tools for Transcription Factor Binding Sites

Wei Wei and Xiao-Dan Yu*

Department of Pathobiology, Center of Computational Biology, Institute of Basic Medical Sciences, Beijing 100850, China.

In the post-genomic era, identification of specific regulatory motifs or transcription factor binding sites (TFBSs) in non-coding DNA sequences, which is essential to elucidate transcriptional regulatory networks, has emerged as an obstacle that frustrates many researchers. Consequently, numerous motif discovery tools and correlated databases have been applied to solving this problem. However, these existing methods, based on different computational algorithms, show diverse motif prediction efficiency in non-coding DNA sequences. Therefore, understanding the similarities and differences of computational algorithms and enriching the motif discovery literatures are important for users to choose the most appropriate one among the online available tools. Moreover, there still lacks credible criterion to assess motif discovery tools and instructions for researchers to choose the best according to their own projects. Thus integration of the related resources might be a good approach to improve accuracy of the application. Recent studies integrate regulatory motif discovery tools with experimental methods to offer a complementary approach for researchers, and also provide a much-needed model for current researches on transcriptional regulatory networks. Here we present a comparative analysis of regulatory motif discovery tools for TFBSs.

Key words: motif, TFBS, non-coding DNA sequence, computational algorithm, motif discovery tool

Introduction

Biological processes in prokaryotic and eukaryotic organisms are guided by genomic information in coding and non-coding DNA sequences. Both kinds of sequences coordinate the construction of transcriptional regulatory networks to perform gene expression with temporal-spatial variations. Compared with the pre-genomic era, which concentrated on deciphering coding DNA sequences and completed the blueprint of the human genome, the post-genomic era puts more emphases on digging the gold mine hidden in non-coding DNA sequences. Currently the identification of specific motifs or transcription factor binding sites (TFBSs) has become one of the key steps in this task.

As we all know, interaction between transcription factors (TFs) and non-coding DNA sequences is a prerequisite for transcription initiation of genes. The function of TFs is to recognize short conserved regions in non-coding DNA sequences, which are called motifs or TFBSs (*1*). However, it is not enough to find motifs or TFBSs in non-coding DNA sequences only depending on experimental methods. For example, systematic evolution of ligands by exponential enrichment (SELEX), serial analysis of gene expression (SAGE), and DNA microarray are only for transcript profiling *in vitro* (*1*, *2*). Chromatin immunoprecipitation (ChIP) can be combined with DNA microarray, namely ChIP-on-chip, to identify protein-DNA interaction *in vivo* (*3*), but it is limited by antibody performance and availability (*4*). For this reason, a wide range of motif discovery tools and databases have been applied to motif or TFBS prediction in biological studies. Unfortunately, 99.9% of their predictions are shown to be futility theorems (*5*).

Motifs or TFBSs are always represented as consensus IUPAC strings, position frequency matrices (PFMs), position weight matrices (PWMs), or position specific scoring matrices (PSSMs) in databases. Commonly, motifs or TFBSs in non-coding DNA sequences are conserved but still tend to be degenerate, which can influence the interaction between TFs and motifs or TFBSs. Therefore, after motif or TFBS data

***Corresponding author.**
E-mail: yuxd@nic.bmi.ac.cn

are collected and aligned from experimental or computational results, relevant consensus IUPAC strings can be constructed by selecting a degeneracy base pair symbol for each position in the alignment (*5*). The motif or TFBS data can also be modeled as PFM by aligning identified sites and counting the frequency of each base pair at each position of the alignment (*6*). Usually, PFM should be converted into PWM or PSSM according to formulas (*5*, *7*). Site scoring of non-coding DNA sequences can be calculated by computing the values for each position in PWM or PSSM model (*5*). Moreover, by using sequence logos, PWM can be displayed with color and height proportional to the base pair frequency and information content for each position by formulas (*8*).

In 1970s, scientists predicted that the pivotal difference between human and chimpanzee was located in non-coding DNA sequences rather than coding DNA sequences (*9*). Since then many essential elements of transcriptional regulatory networks have been identified in non-coding DNA sequences, including promoters, enhancers, insulators, silencers, and locus control regions (*6*). Nowadays, the discovery of motifs is mainly limited in canonical 5′ termini of known genes, where TFs are generally thought to bind in. Nevertheless, recently some researches have shown that only small proportion of motifs or TFBSs lie in immediate upstream sequences of well-characterized protein-coding genes, while the rest of them exist in either introns or 3′ regions (*6*, *10*, *11*).

A number of algorithms to discover motifs have been applied previously, for example, BE95 (*12*), KYD96 (*13*), DB97 (*14*), vHRCV00 (*15*), BJVU98 (*16*), EP20 (*17*), KFQW99 (*18*), and so on. However, many of these algorithms were designed for finding longer or more common motifs rather than for identifying TFBSs (*19*). The price paid for this generality is that many of the cited algorithms are not guaranteed to find globally optimal solutions, since they employ some forms of local search, such as Gibbs sampling, expectation maximization (EM), and phylogenetic algorithms. In this study, we give a brief introduction to the algorithm design and analysis for TFBSs with a focus on problems in comparative motif discovery.

Results and Discussion

Combinatorial approaches

Among the possible algorithmic approaches, combinatorial approaches try to exhaustively explore all the ways that a molecular process could happen. This leads to hard combinatorial problems for which efficient algorithms are required. Thus this kind of algorithms must make use of complex data representations and techniques.

Sequence-driven or Sample-driven (SD) algorithms

SD algorithms try to find comparative patterns by comparing the given length strings and looking for local similarities between them. They are based on constructing a local multiple alignment of the given non-coding DNA sequences and then extracting the comparative patterns from the alignment by combining the segments, which is common to most of the non-coding DNA sequences (*20*).

Pattern-driven (PD) algorithms

PD algorithms are based on enumerating candidate patterns in a given length string and inputting substrings with high fitness. The advantage of PD algorithms is that they can search the best comparative patterns in some limited sizes (*20*). Compared with SD algorithms, PD algorithms can be performed intelligently so that patterns are not present in the data that are not generated. For example, if a pattern α is not frequently present in the data, then there will be no frequent refinement that makes α more specific (hitting in even fewer places) in the data either (*20*).

Multiprofiler

This algorithm mainly utilizes multi-profiles that generalize a notion of a profile to detect subtle patterns that might escape detection by standard profiles (*21*). It is designed for finding particularly subtle motifs even in the case when real motifs may be blurred by random ones. The advantage of Multiprofiler is that it takes much less time (*21*). Kravchenko *et al* used Multiprofiler to search and statistically assess putative motifs in promoter regions of co-regulated genes, where the discovered over-represented sites could be totally verified by cell transfection experiments (*22*).

Consensus

This approach determines all possible pairwise alignments of matrices and remains words to create two sequence alignments. It scores the two sequence alignments by using information content, and the highest scoring will be saved (*23*). Each of the two sequence matrices is paired with each word that is not already