MotifCombinator: A Web Tool to Search for Combinations of Cis-Regulatory Motifs

Mamoru Kato  
kato@src.riken.jp  
Tatsuhiko Tsunoda  
tsunoda@src.riken.jp

Laboratory for Medical Informatics, SNP Research Center, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, 230-0045, Japan

Keywords: cis-regulatory motif, combinatorial regulation, web tool, regression, genetic algorithm

1 Introduction

Gene expression is controlled by combinatorial regulation of transcription factors and cis-regulatory elements in eukaryotes. Experimental procedures can determine several binding sites for selected transcription factors, but they are too laborious to be applied to large-scale studies. Computational methods are thus required to detect the combinatorial regulation at the genomic level. To reveal the combinatorial regulation, recent studies have developed computational methods to detect significant combinations of patterns (motifs) of cis-regulatory elements, using datasets of upstream sequences and expression levels from DNA microarrays, or binding information from ChIP-on-chip arrays [4]. One widespread type of such computational methods is to find combinations of motifs that specifically appear in upstream sequences of co-regulated genes, which are determined to be expressed by a certain threshold of expression levels.

Another type of computational method is based on regression analysis between expression levels and motif scores (occurrence frequencies or weight matrix scores) in input sequences [2,3,6]. This type measures the goodness-of-fit of the regression for candidate motifs, and selects motifs with the best fitting scores. Well-known regression methods for this objective are the (multivariate) linear regression method [2,6] and the multivariate adaptive regression spline (MARS) method [3]. This type of computational method can take full advantage of information about expression levels because the methods do not compulsively dichotomize expression levels into whether genes are expressed or not. However, there is no web-based tool to systematically search motif combinations based on these regression methods. Moreover, these methods were developed mainly for simple eukaryotes like Saccharomyces cerevisiae; therefore, there are some limitations when they are applied to higher organisms with complex regulatory systems such as mammals. For example, they cannot practically handle combinations composed of more than two or three motifs.

We implemented MotifCombinator, a web-based tool that can systematically search combinations of regulatory motifs based on regression methods. This tool is equipped with the two types of regression methods (the multivariate linear regression and MARS), and moreover it includes the logistic regression. It also employs the genetic algorithm to search combinations composed of more than two or three, or an arbitrary number of motifs. MotifCombinator will serve users to find combinations of regulatory motifs in organisms with complex regulatory systems.

2 Method and Results

MotifCombinator consists of four step-by-step procedures for searching motif combinations. Briefly, the first step is to upload a dataset of upstream sequences and expressions. The second step is to find or upload single motifs that will be used later to constitute candidate motif combinations. At this step, users can use MEME [1], AnnSpec [8], and MDscan [5] to find de novo single motifs from input sequences, and use the pre-installed JASPAR database [7] to upload. Users can also upload their own motifs. The third step is to cut out redundant single motifs or irrelevant ones. At this step, this tool can classify groups of similar motifs and select distinct motifs, using clustering analysis. In addition, this tool can calculate how much variance of
input expressions is explained by scores of a single motif in the input sequences, and can filter out a motif that makes an extremely low contribution to the input expressions.

The fourth step is to search motif combinations that are important for expression levels. MotifCombinator generates candidate motif combinations from motifs that are selected through the previous steps, or from motifs that are selected at this step by users. For each of the motif combinations, the tool calculates scores of the combination in input sequences, and takes regression between the scores and input expression levels to calculate the goodness-of-fit of the regression. It then selects motif combinations with the best fitting. These procedures are iteratively performed through the genetic algorithm. The characteristic points at this step are as follows.

♦ Users can search combinations composed of an arbitrary number of motifs, using the genetic algorithm.
♦ Users can use the three types of regression methods (the multivariate linear regression, MARS, and the multivariate logistic regression) to find motif combinations significant for expression levels.
♦ Users can intuitively understand how the combination search has proceeded by the two visualization systems, which show a history and a brief landscape of the goodness-of-fit scores during the search.
♦ Users can easily store all records by the backup system, and even after shutdown of the computer, they can easily restart the search just by uploading the backup file on the web screen.

References


