Benchmarking of Motif-Finding Algorithms

Alona Chubatiuk

Department of Mathematics, USC

There are several approaches to detect motifs. The most common approach is to find a “word” common to all or most of the promoters in a given set of co-expressed genes. The limitation of this approach is the subjectivity in selecting the gene set. An alternative approach is to carry genome-wide search for statistically significant association between “words” in biological sequence and phenotypic data. Methods from both schools are commonly used.

We compare three motif-finding algorithms that use gene expression and sequence information to find regulatory elements in promoters. The programs include MatrixREDUCE [1], Allegro [2], and MotifExpress [3]. We have developed a comprehensive benchmarking approach using 16000 promoters of *A. thaliana* and diverse experimental conditions from NCBI GEO. All promoters were randomly divided into training and testing sets. Most significant motifs for each experimental condition were identified in the training set. We detected these motifs in the testing set and compared the gene expression between genes with and without the motif. We discovered that no single program can be considered to be the best: performance varies from one dataset to another.

1. *Foat B, Morozov V, Bussemaker H. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixReduce. Bioinformatics. 2006.*
2. *Halperin Y, ChaimL, Ulitsky I, Shamir R. Allegro: Analyzing expression and sequence in concert to discover regulatory programs. NAR. 2009;37(5).*
3. *Triska M, Southern J, Grocutt D, Tatarinova T. MotifExpress: Motif Detection in DNA Sequences. 2012.*