DRIMust: a web server for Discovering Rank Imbalanced Motifs Using Suffix Trees

Limor Leibovich[1*], **Inbal Paz**[2*], Zohar Yakhini[1,3] and Yael Mandel-Gutfreund[2]

[1]Department of Computer Science, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

[2]Department of Biology, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

[3]Agilent Laboratories Israel, 94 Em Hamoshavot Road, 49527 Petach-Tikva, Israel

Abstract

Cellular regulation mechanisms that involve proteins and other active molecules interacting with specific targets often involve the recognition of short sequence elements. Studies that focus on measuring and investigating sequence based recognition processes make use of statistical and computational tools that support the identification and understanding of sequence motifs. We present a new web application, named DRIMust, freely accessible through the website: http://drimust.technion.ac.il, for de-novo motif discovery services. The DRIMust algorithm is based on the minimum-hypergeometric (mHG) statistical framework using suffix trees for an efficient enumeration of motif candidates. DRIMust takes as input ranked lists of sequences in FASTA format and returns motifs that are over-represented at the top of the list, where the determination of the threshold that defines top is data driven. The resulting motifs are presented individually with an accurate *p*-value indication and as a Position Specific Scoring Matrix (PSSM). Comparing DRIMust to other state-of-the-art tools demonstrated significant advantage to DRIMust both in result accuracy and in short running times. Overall, DRIMust is unique in combining efficient search on large ranked lists with rigorous *p*-value estimation for the detected motifs.