EXPLOITING CROSS-SPECIES CONSERVATION TO IMPROVE PREDICTION ACCURACY OF DISCOVERING TRANSCRIPTION FACTOR BINDING SITES

Po-Chun Wu, Mei-Ju May Chen, and Chien-Yu Chen*

Dept. of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan

**Email: cychen@mars.csie.ntu.edu.tw*

Comparative genomics has been widely used to identify functional regions in the genome. To facilitate discovery of human transcription factor binding sites (TFBSs), we designed a web server to explore sequence conservation across related species to re-score motif candidates derived from computational methods. First, data of multiple sequence alignment (MSA) for seven species, including human, rhesus, horse, dog, mouse, rat, and chicken, was downloaded from VISTA database (released on 14-May-2008). In addition, human promoter regions were collected from UCSC genome database maintained by University of California, Santa Cruz (version: hg18, released on Mar. 2006). For each Reference Sequence ID of human, we extracted the region of upstream 2000 to downstream 2000 b.p. with respect to the transcription start site (TSS) as the promoters. A gene is associated with one or more MSA regions from VISTA according to the annotation of genome positions. When given a set of motif instances, the conservation score is calculated by counting the proportion of conserved instances present within a specified range in MSA with respect to the binding site reported in human. We demonstrated in this study that the prediction accuracy of TFBS discovery can be improved when conservation scores are incorporated in the final ranking procedure. A validation set was created to evaluate the performance of employing conservation scores for ranking motif candidates. Among the 48 test cases, the new method successfully ranks a correct motif as the top-1 or the top-2 motif in 44 cases, better than the original ranking scheme, which only succeeds on 39 cases. We concluded that the conservation information derived from the related species indeed helps to discover human TFBSs with higher accuracies.

* Corresponding author.