Poster J-16

Computational analysis of gene regulatory elements

Authors:
Reddy, D. A. (*University of Hyderabad*)
Mitra, C. K. (*University of Hyderabad*)

Short Abstract: The core promoter regions and also the TFBS have been studied by calculating the average mutual information content. For mouse and human, both TATA-box and TSS-region are likely to play important roles. For plant, the importance of TSS-region in transcriptional initiation is more compared to the TATA box. The TFBS clustering results gives us a new way to look at the protein classification-not based on their structure or function of TFs but by the nature of their TFBS.

Long Abstract:
A major component of gene regulation occurs at the level of transcription. Gene expression and regulation involve the binding of several proteins (transcription factors) to specific non-coding DNA elements. Although non-coding DNA constitutes majority of most eukaryotic genomes, relatively little is known about its function or the nature. The functional identification/ classification of these regions is very difficult due to the lack of proper signals and/or their degenerate nature in the whole genome. We have attempted to identify these regions using information theoretic approach based on the substitution matrices constructed using available databases. The present study involves three parts. First, we have studied the core promoter region in five sets of promoter sequences (sequences obtained from PlantProm and EPD databases) by calculating the average mutual information content. The average mutual information content (H) is the relative entropy of the target and background pair frequencies and can be thought of as a measure of the average amount of information (in bits) available per nucleotide pair. Here we constructed and applied nucleotide substitution matrices (both neighbor independent and neighbor dependent) for the core promoter region and calculated the information content from these substitution matrices to study the Transcription Start Site (TSS) region, TATA-box, and downstream region. The results show that the TSS-region is likely to be 5-10 bases in size. We also notice that both in the case of mouse and humans, both TATA-box and TSS-region are likely to play important roles. However, in case of plant, the results showed the importance of TSS-region for transcriptional initiation (compared to the TATA-box region). Second, once promoter region is identified, the next task is to identify/analyze cis-regulatory elements within such regions. Here we present a new cluster method to classify Transcription Factor Binding Sites (TFBS). The clustering of TFBS (JASPAR database) with information content as a metric suggests that in each group of clustered TFBS with their respective TF-class share any one of TF-class in that clustered TF-class. Thus in JASPAR database, out of the 41 TFBS (in humans), perhaps only 5 -10 or so TFs may be actually needed and in case of mouse instead of 13 TFs, we may have actually 5 or so TFs. The experimental data of TFs of specific gene expression from Transcription Regulatory Regions Database (TRRD) is also coincides with our computational results. This gives us a new way to look at the protein classification-not based on their structure or function of TFs but by the nature of their TFBS.

One interesting pattern that is observed in calculating the information content is that a strong nucleotide neighbor dependency (pairwise) is observed in the TFBSs, but not in the case of core promoter region. Third, we also analyze the complete genome sequences of human and mouse for the gene regulatory regions with the information content obtained from the core promoter and TFBS by scanning the whole genome using the tools developed as mentioned above. References: 1.Reddy, D. A., Prasad, B. V. L. S and Mitra, C. K. Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. Computational Biology and Chemistry, 30, 58-62, 2006. 2.Reddy, D. A., Prasad, B. V. L. S and Mitra, C. K. Functional classification of transcription factor binding sites: Information content as a metric. Journal of Integrative Bioinformatics, 0020, 2006 (Online).