Poster L-15

MDframe: a computational framework for the high-throughput identification of human regulatory motifs

Authors:
Linyong Mao (*Medical University of South Carolina*)
Omar Moussa (*Medical University of South Carolina*)
John S. Yordy (*Medical University of South Carolina*)
Dennis K. Watson (*Medical University of South Carolina*)
W. Jim Zheng (*Medical University of South Carolina*)

Short Abstract: We developed a computational framework, MDframe, for the high-throughput identification of potential human regulatory motifs. The framework integrates microarray/transcription profiling, comparative genomics, and de novo motif discovery to identify overrepresented DNA motifs in the sequence set upstream of co-regulated genes by employing an exhaustive oligomer enumeration technique

Long Abstract:

I. Introduction

As more and more genomes are sequenced, it becomes feasible to employ comparative genomics approaches to identifying conserved regulatory elements that may be involved in gene regulation. In addition, rapidly accumulating transcriptional profiling data allows investigators to identify gene clusters that are potentially co-regulated. By integrating transcriptional profiling, comparative genomics, and de novo motif discovery, we developed a computational framework, MDframe, for the high-throughput identification of potential human transcription factor (TF) DNA binding motifs.

MDframe includes two databases and four modules. Database I contains the conserved upstream regions of 20835 annotated human RefSeq genes, where Database II contains known TF DNA-binding consensus sequences. The functionality of the first module is to get co-regulated gene clusters from microarray data analysis. The conserved upstream regions of these genes can then be retrieved from Database I by the second module. The third module uses this set of conserved upstream regions to discover the overrepresented DNA motifs. From these overrepresented DNA motifs, the fourth module probes their identities by comparing to the known TF DNA-binding profiles stored in Database II. The computational framework was implemented in Linux using C/C++ and shell script.

II. Methods

Database I: conserved upstream regions (Homo sapiens)

Multiple alignments of the upstream regions of RefSeq genes were downloaded from University of California, Santa Cruz. The alignment was made against the following seven species: chimp(panTro1), mouse(mm5), rat(rn3), dog(canFam1), chicken(galGal2), fugu(fr1)

and zebrafish(danRer1). Two methods were developed to extract conserved upstream regions from these alignments, respectively. The first method, Stringent Masking Method (SMM), kept or masked each base in human sequence depending on how conserved it is in other species (by applying a threshold value t). The second method, referred to as Window-Based Masking Method (WBMM), utilized less stringent conservation criteria. It first assigned a score for each base in a human gene's upstream sequence. If a base was in the non-repeat region and conserved in at least t of the seven species, it has a score of one. Zero otherwise. Then a Window-Based Value (WBV) for each base was calculated as the sum of scores over a user-specified window size centered at that base. The sum excluded the sore of that base. After WBV was calculated, a base was retained if it met either one of the following two conditions: (i) the score for the base is 1; (ii) the base is in the non-repeat region and the WBV for the base exceeds a certain threshold specified by a user. Otherwise, the base was masked by 'N'.

Module III

This module is a modification of the Weeder program to cope with the masked bases and gaps in upstream sequences, and utilize an exhaustive oligomer enumeration technique to identify overrepresented sequence motifs.

III. Results

We applied MDframe to identify TF binding motifs for the well characterized muscle specific gene set [3]. The results of MDframe with different masking parameters were listed (Table 1). When t was set at 3 for SMM, five TF DNA-binding motifs were detected. When WBMM was applied and t was set at 4, four TF DNA binding motifs were detected with the combined sensitivity equal to 0.41 and positive prediction rate (PPR - the fraction of predicted sites that are known binding sites) equal to 0.56. For comparison, CompareProspector [4] and Toucan [2] were also applied to detect DNA binding motifs for the muscle set. CompareProspector detected four TF DNA binding motifs with the combined sensitivity equal to 0.24 and PPR 0.5. Toucan only detected two motifs, and the combined sensitivity and PPR is 0.09 and 0.18, respectively.

IV. References

1. Pavesi, G., et al., Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res, 2004. 32(Web Server issue): p. W199-203.
2. Aerts, S., et al., Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucl. Acids Res., 2003. 31(6): p. 1753-1764.
3. Wasserman, W.W., et al., Human-mouse genome comparisons to locate regulatory sites. Nat Genet, 2000. 26(2): p. 225-8.
4. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res 2004, 14(3):451-458.