K-mer Analysis on Developmental and Housekeeping Enhancer Peaks

Yunsi Yang
Yale University, yunsi.yang@yale.edu

Anurag Sethi
Yale University, anurag.sethi@yale.edu

Mark Gerstein
Yale University, mark.gerstein@yale.edu

K-mer Analysis on Developmental and Housekeeping Enhancer Peaks

Yunsi Yang[1,2], Anurag Sethi[1,3], Mark Gerstein[1,3,4]

[1] Department of Molecular Biophysics and Biochemistry, [2] Department of Statistics, [3] Program in Computational Biology and Bioinformatics, [4] Department of Computer Science, Yale University

Abstract

The regulation of gene expression involves interaction between transcriptional enhancers and core promoters. However, the separation between developmental and housekeeping gene regulation remains unknown. Here, we present a method to detect if different core promoters exhibit specificity to certain enhancers within massively parallel assays for enhancer detection. We use k-mers of various length (3-8bp) as sequence features and compare k-mer frequencies between developmental and housekeeping enhancers. This method shows promoter specificity of enhancers in *D. melanogaster*.

Background

- A promoter located in the upstream of protein-coding gene initiates transcription. An Enhancer located in non-coding DNA and far from its promoter activates gene transcription.
- Housekeeping enhancers are active among cells, while developmental enhancers exist in specific cell.
- K-mer is the subsequence of length k in DNA sequence.

Data Description

- STARR-seq, a method for enhancer detection, identifies thousands of cell-type-specific enhancers in Drosophila melanogaster S2 cell.
- Housekeeping core promoter (hkCP): RpS12, eEF1δ, NipB, x16
- Developmental core promoter (dCP): DSCP, eve, evelong, pnr, Hsp70
- Length of each enhancer peak is 500bp.

Methods

Step 1: For each enhancer peak, count all k-mers (k=3…8) to get k-mer frequency

Step 2: For each pair of promoters, run ANOVA test on each k-mer to detect if there is significant difference in means of frequency

Step 3: Narrow down significant k-mers

Results

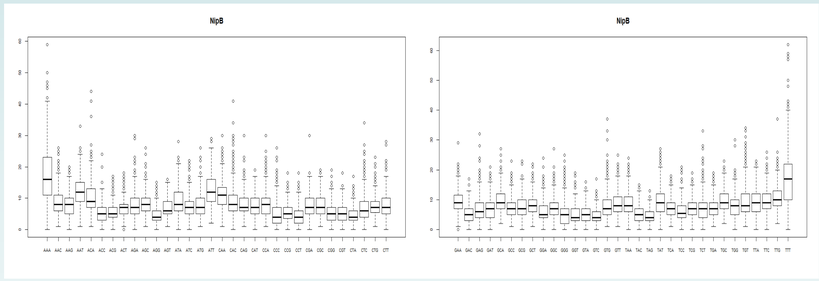


Figure 1: Frequency of 3-mers in enhancer peaks of NipB

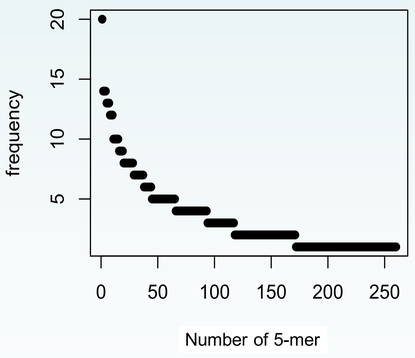
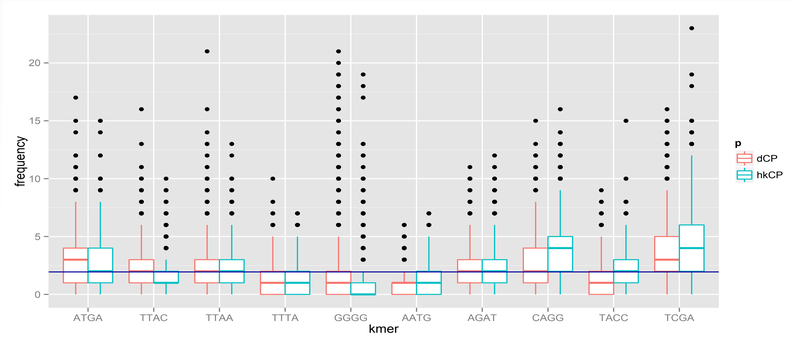


Figure 2 (left): Frequency distribution of 5-mers appearing significant in all pairs

Figure 3 (bottom): Compare frequency of significant 4-mers between merged dCP and hkCP enhancers



Conclusion

Kmer sequence features exhibit a specificity to developmental and housekeeping enhancers

3-mer	ATG AGA TTA
4-mer	ATGA TTAC TTAA TTTA GGGG AATG AGAT CAGG TACC TCGA
5-mer	AGATA CACTG TCGAT ATCGA TATCG TTTAA AATGA GGCGC TTACC AGCTG CATTC TGACC TGGGG TTTAC
6-mer	TATCGA ATCGAT CAGCTG ACAGCT CTATCG TCGATA GTGACC AGATAA CGATAG TTTAAA ATTTAA GGTCAC
7-mer	TATCGAT GGTCACA ATCGATA ACAGCTG ACTATCG TGTGACC CTATCGA GTGTGAC TCGATAG TTATCGA ATATCGA ATCGATT GTTATCG GTCACAC TCACACT
8-mer	GGTCACAC GTCACACT ATATCGAT CGGTCACA AACAGCTG CTATCGAT TATCGATA ACAGCTGA ATCGATAG GTGTGACC TATCGATG TCGATTGT ACTATCGA GTTATCGA TATCGATT

Future Research

- Compare result of significant k-mers with known DNA sequence motifs
- Build machine learning model to classify developmental and housekeeping enhancers

References

Arnold, Cosmas D., et al. "Genome-wide quantitative enhancer activity maps identified by STARR-seq." *Science* 339.6123 (2013): 1074-1077.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." *Annu. Rev. Genomics Hum. Genet.* 7 (2006): 29-59.

Zabidi, Muhammad A., et al. "Enhancer--core-promoter specificity separates developmental and housekeeping gene regulation." *Nature* (2014).