CSCI 5622 Final Project Proposal

Yuvraj Arora
Department of Computer Science
University of Colorado
Boulder, CO 80309
yuvraj.arora@colorado.edu

Murad Chowdhury
Department of Computer Science
University of Colorado
Boulder, CO 80309
murad.chowdhury@colorado.edu

Ignacio Tripodi
Department of Computer Science
University of Colorado
Boulder, CO 80309
ignacio.tripodi@colorado.edu

Abstract

Several assays have been developed over the last decade to gain further insights on genetic transcriptional activity, and changes in the chromatin structure. One of them, Assay for Transposase Accessible Chromatin followed by Sequencing (ATAC-Seq [3]) is a relatively novel technique that provides a measure of accessible sites, genome-wide. Many of those open sites are hypothesized to be indicative of transcriptional activity, and overlap with binding sites for regulatory protiens known as transcription factors to enhance or inhibit transcription of a certain gene. The aim for this project is to utilize machine learning as a tool to classify ATAC-Seq peaks that overlap active enhancer sites, against those that do not. Since these peaks can be thought of as a signal independent of the biological context, it would be interesting to explore features extracted using signal processing techniques and apply them to the ATAC-Seq signal.

1 Background

ATAC-Seq is a relatively new assay to obtain information from open chromatin regions of the genome which excels in its simplicity, takes a total of three hours to complete, and costs on the order of a few hundred dollars. This makes the assay highly desirable for inferring different kinds of information about transcriptional activity, as most other assays that operate genome-wide and are not focused to specific antibodies are over an order of magnitude more expensive, take several days to complete, and have a higher error rate. Thus, utilizing ATAC-Seq data to classify different transcriptional conditions and states is a worthwhile endeavor.

2 Dataset

Data files before processing are saved in the BED format. These files can be fed to a peak finding software which outputs the average region that is considered a peak and generates a file which is similar to the BED format. All files will be aligned to the "hg19" human reference genome, and will correspond to assays made using the same cell line (K562) and experimental conditions.

The ATAC-Seq peaks represent approximate regions of the genome that are accessible to transcription factors and other protiens involved in gene regulation. The size of the training data available, although it varies by assay, is generally on the order of 80,000 peaks.

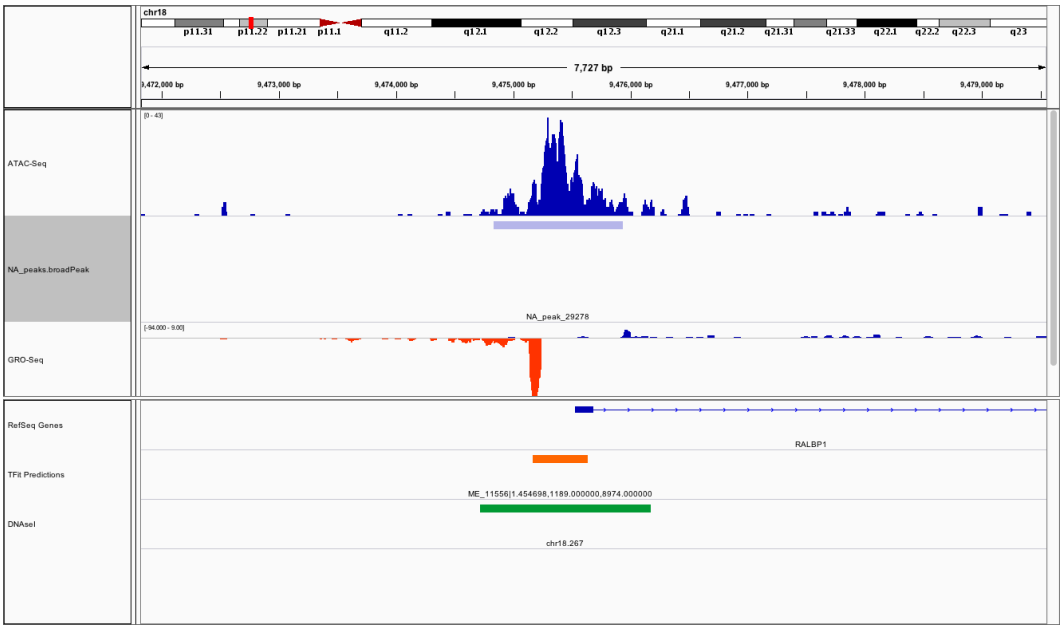

Figure 1: A screenshot from IGV (Integrative Genomics Viewer [1]) showing what an ATAC-Seq signal looks like (in blue top track), in the context of other assays, for a peak that overlaps an active gene enhancer site. Other tracks depicted in the figure are: ATAC-Seq peak boundaries obtained using MACS [5], GRO-Seq signal (note that this nascent transcription assay is aware of DNA strand polarity, so this is a bidirectional signal), DNAse I [4] signal (a different assay to inquire about chromatin accessiblity), and the reference genome annotations depicting gene location and their introns/exons.

A TFit predictions file, also in the BED format, is generated from a software tool developed at the Dowell Lab at the University of Colorado, Boulder. This tool identifies active regions based on GRO-Seq (Global Run-On followed by sequencing [2]) data, an example of a longer, more sophisticated and expensive assay. These regions will be taken as our ground truth. ATAC-Seq peaks overlapping TFit regions will be considered "active" and "inactive" otherwise. The "overlap" may be dependent on the biological context, and may be expanded from a strict overlap to regions which are 1000 base-pairs from an ATAC-Seq peak's median and 1000 base-pairs from the TFit region median overlap.

chr1	566973	566975	8041
chr1	566975	566976	8042
chr1	566976	566977	8049
chr1	566977	566978	8072
chr1	566978	566980	8070
chr1	566980	566981	8069
chr1	566981	566982	8070
chr1	566982	566984	8075
chr1	566984	566985	7995

chr1	10004	11558	NA_peak_1	101	.	4.51481	12.62142	10.14651
chr1	28827	29870	NA_peak_2	40	.	2.51883	6.21283	4.07941
chr1	32468	33155	NA_peak_3	230	.	4.81852	25.92968	23.07220
chr1	38561	38908	NA_peak_4	50	.	3.87921	7.27900	5.03972
chr1	88012	88252	NA_peak_5	52	.	4.94980	7.53575	5.25467
chr1	107080	108618	NA_peak_6	12	.	2.88408	3.17286	1.26707
chr1	115572	115869	NA_peak_7	55	.	4.77988	7.88705	5.59474
chr1	136180	137488	NA_peak_8	29	.	3.15990	5.00642	2.90786
chr1	227554	228785	NA_peak_9	126	.	6.08850	15.16263	12.63885
chr1	235208	236614	NA_peak_10	61	.	4.10723	8.46254	6.16370

Figure 2: **Left**: Example data file in the BED format. The first column denotes the chromosome. The second and third column denote the start and end of a sequence and the final column denotes sequencing depth. **Right:** Example data file after processing to find peaks. The first three columns are the same as in the unprocessed file, and the forth column denotes peak name.

3 Proposed Methods

We propose three approaches to classifying these ATAC-Seq peaks.

Feature Engineering - Use domain knowledge to extract salient features from our dataset, which can be used with traditional supervised learning methods such as feed forward neural networks, SVMs, etc. Some initial proposed features include:

- Peak width/height
- Distance between peaks (previous or next)
- Coefficients obtained from Fourier or wavelet transforms

Peaks as Features - We can treat the peaks as a time varying digital signal and use techniques that can take advantage of the temporal structure of our data. Some proposed methods include:

- 1D convolutional and recurrent neural networks
- Hidden Markov Models

Unsupervised Feature Learning - Instead of hand picking features, use unsupervised methods to learn features from unlabeled input data which can then be used with supervised methods mentioned earlier. Some proposed methods include:

- k-means clustering
- Autoencoders
- Principal Component Analysis

References

[1] Integrative genomics viewer. `http://software.broadinstitute.org/software/igv/home`.

[2] Jose Garcia-Martinez, Augustin Aranda, and Jose E Perez-Ortin. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Molecular Cell*, 15(2):303 – 313, 2004.

[3] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods*, 10(12):1213–1218, 2013.

[4] Wang Y-M, Zhou P, Wang L-Y, Li Z-H, Zhang Y-N, and Zhang Y-X. Correlation between dnase i hypersensitive site distribution and gene expression in hela s3 cells. *PLoS ONE*, 7(8), 2012.

[5] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137, Sep 2008.