A Comparative Study of Models Predicting Transcription Start Sites

Mirudhula Mukundan, Wrootchit Mishra, Xueke Jin, David Luo
Carnegie Mellon University
Pittsburgh, PA 15213
{mirudhum, wrootchm, xuekej, zhichenl}@andrew.cmu.edu

1 Introduction

Identifying transcription start sites (TSS) is a critical step in understanding gene regulation, as they mark the initiation points of transcription, where DNA is transcribed into RNA. Accurate TSS identification enables researchers to unravel complex gene regulatory networks and can provide valuable insights into various biological processes, diseases, and potential therapeutic targets [8].

In this project, our aim is to explore the latest advancements in TSS modeling and evaluate the efficacy of these models. We recognize that each of these models is designed to work with specific data types. Therefore, we plan to enhance the predictive power of these models by integrating multiple data types such as CAGE-seq, RNA-seq, ChIP-seq, and DNase-seq. Our approach will be integrative, and we will develop a pipeline that incorporates the strengths of each data type, enabling us to build a comprehensive and robust model for the prediction of TSS.

2 Background

Predicting TSSs is a challenging task due to the presence of numerous promoters and other regulatory elements in the genome, as well as variations in transcription initiation patterns across different cell types and conditions [1]. Computational methods for TSS prediction have been developed, but their accuracy can be limited by factors such as insufficient training data and the complexity of the genomic landscape [10].

Despite these obstacles, advances in high-throughput sequencing technologies and computational approaches have led to improvements in TSS prediction, emphasizing the importance of this research area in advancing our understanding of gene regulation and its implications for human health [2]. As we continue to refine these techniques and expand our knowledge of the genome, the ability to accurately predict TSSs will become increasingly valuable in both basic research and clinical applications.

3 Data Sources

Below is a list of available database containing various high-throughput sequencing data with annotated TSS and CAGE information to enable the development of the computational models and the progression of our project:

- FANTOM5: The FANTOM5 project generated a comprehensive atlas of CAGE profiles across various cell types and tissues.
- ENCODE: The ENCODE project has generated a large amount of CAGE data as part of its efforts to annotate the human genome.

- GEO: The Gene Expression Omnibus (GEO) is a public repository for gene expression data, which will have board data including TSS.
- ArrayExpress: ArrayExpress is a public repository for gene expression data.
- UCSC Genome Browser: The UCSC Genome Browser provides a comprehensive view of the genome, including annotated TSSs.
- RefSeq: The RefSeq database is a comprehensive collection of annotated genomic sequences, including information on TSSs.
- DBTSS: The DataBase of Transcriptional Start Sites (DBTSS) provides information on TSSs for human and mouse genes.

4 Softwares & Algorithms

Research studies focused on predicting transcription start sites (TSS) have traditionally relied on CAGE-seq data due to its specificity for identifying TSS. However, the noise and complexity of the human genome can present challenges when working with CAGE-seq data. With the emergence of machine and deep learning techniques, several recent studies have leveraged these approaches to improve TSS prediction. The following models from recent papers represent examples of such efforts:

- ADAPT-CAGE [5]: This novel machine learning model utilizes an unsupervised learning approach to identify TSS. The authors report that the model was able to perform better than many of the older and more recognized models like Paraclu [4] and Reclu [9].
- DeepTSS [6]: This method makes use of Convolutional Neural Network that takes DNA sequence, information of the DNA structure and evolutionary conservative features as input to predict TSS.
- DeeReCT-TSS [12]: Another deep learning approach that makes use of cponventional RNA-seq and DNA sequence data to predict TSS.

Comparing these recent findings would be valuable in providing up-to-date insights, as opposed to older studies which have already been explored in the papers mentioned above.

Apart from these algorithms, since we might work with both RNA-seq and CAGE-seq data, we would need pre-processing techniques like STAR [3] or HISAT2 [7] that helps to align reads from RNA-seq data to a reference genome, and peak-calling techniques like MACS [11] for CAGE-seq data.

5 Questions about the project & Future Goals

5.1 Questions

Q1: Is it possible to obtain a shared outcome across various algorithms and use that to reassemble complete transcripts where we could then potentially juxtapose these findings with transcription assembly tools such as StringTie and Cufflinks?

Q2: How does the variability of the TSS region compared to other promoter elements impact the accuracy and effectiveness of predictive models?

Q3: What specific factors contribute to the low performance of some computational methods in predicting TSSs/promoters, and how can these factors be addressed to improve prediction accuracy?

Q4: Are there any consistent patterns or unique features in the composition of subsequences around TSSs that can be leveraged to improve model performance across different genomes?

Q5: What is the role of information content in measuring the variability of TSS regions, and how can this information be incorporated into predictive models to enhance their performance?

5.2 Future goals

The following are potential topics for future exploration. In this project will focus on investigating some of these areas:

Enhancing prediction accuracy: Improve the performance of computational methods by addressing the factors that contribute to their low accuracy and developing new strategies for better TSS/promoter identification.

Understanding TSS variability: Investigate the greater variability of TSS regions compared to other promoter elements, using measures such as information content, to gain insights into the mechanisms driving this variability.

Expanding model generalizability: Examine the relationship between prediction accuracy and the length of subsequences across different genomes, such as human, rat, and mouse, to develop models that can be applied more broadly and effectively across various organisms.

Integrating multi-omics data: Incorporate additional types of data, such as epigenomic and transcriptomic information, to enhance the predictive power and comprehensiveness of the models.

Comparing with transcription assembly tools: Assess the performance of TSS prediction models against transcription assembly tools like StringTie and Cufflinks to identify the strengths and weaknesses of each approach and inform the development of more effective methodologies.

6 Team Roles

Program manager: Wrootchit Mishra

Technical lead: Mirudhula Mukundan

Lead technical writer: David Luo

Communications lead: Xueke Jin

Q & A lead: Every group member

References

[1] ABEEL, T., SAEYS, Y., BONNET, E., ROUZÉ, P., AND VAN DE PEER, Y. Generic eukaryotic core promoter prediction using structural features of dna. *Genome Research 18*, 2 (2008), 310–323.

[2] DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., ET AL. Landscape of transcription in human cells. *Nature 489*, 7414 (2012), 101–108.

[3] DOBIN, A., AND GINGERAS, T. R. Mapping rna-seq reads with star. *Current protocols in bioinformatics 51*, 1 (2015), 11–14.

[4] FRITH, M. C., VALEN, E., KROGH, A., HAYASHIZAKI, Y., CARNINCI, P., AND SANDELIN, A. A code for transcription initiation in mammalian genomes. *Genome research 18*, 1 (2008), 1–12.

[5] GEORGAKILAS, G. K., PERDIKOPANIS, N., AND HATZIGEORGIOU, A. Solving the transcription start site identification problem with adapt-cage: a machine learning algorithm for the analysis of cage data. *Scientific Reports 10*, 1 (2020), 877.

[6] GRIGORIADIS, D., PERDIKOPANIS, N., GEORGAKILAS, G. K., AND HATZIGEORGIOU, A. G. Deeptss: multi-branch convolutional neural network for transcription start site identification from cage data. *BMC bioinformatics 23*, 2 (2022), 1–17.

[7] KIM, D., PAGGI, J. M., PARK, C., BENNETT, C., AND SALZBERG, S. L. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology 37*, 8 (2019), 907–915.

[8] LENHARD, B., SANDELIN, A., AND CARNINCI, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics 13*, 4 (2012), 233–245.

[9] OHMIYA, H., VITEZIC, M., FRITH, M. C., ITOH, M., CARNINCI, P., FORREST, A. R., HAYASHIZAKI, Y., AND LASSMANN, T. Reclu: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (cage). *BMC genomics 15*, 1 (2014), 1–15.

[10] PONGER, L., AND MOUCHIROUD, D. Cpgprod: identifying cpg islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics 18*, 4 (2002), 631–633.

[11] ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W., ET AL. Model-based analysis of chip-seq (macs). *Genome biology 9*, 9 (2008), 1–9.

[12] ZHOU, J., ZHANG, B., LI, H., ZHOU, L., LI, Z., LONG, Y., HAN, W., WANG, M., CUI, H., LI, J., ET AL. Annotating tsss in multiple cell types based on dna sequence and rna-seq data via deerect-tss. *Genomics, Proteomics & Bioinformatics* (2022).