Sequence Model Evaluation Framework for STARR-Seq Peak Calling

Christopher R. Beal
Marquette University
Milwaukee, Wisconsin
christopher.beal@marquette.edu

John G. Peters
Milwaukee School of Engineering
Milwaukee, Wisconsin
petersjg@msoe.edu

Ronald J. Nowling
Milwaukee School of Engineering
Milwaukee, Wisconsin
nowling@msoe.edu

ABSTRACT

Enhancers are short regions of non-coding DNA that increase transcription rates of genes despite being located distantly from the genes themselves [5]. Enhancers are identified through experimental techniques such as ChIP-Seq or CUT&RUN with H3K4me1 and H3K27ac histone modifications, self-transcribing active regulatory region sequencing (STARR-Seq), and massively parallel reporter assays (MPRA). Machine learning models have been used in conjunction with experimental data to identify enhancer activity from sequences [3], predict enhancer-transcription factor interactions [4], and decode the enhancer regulatory language [2].

We describe a framework that connects peak calling errors to the prediction accuracy of sequence models. The key assumptions of our framework are that (1) enhancers have consistent sequence patterns that can be used to separate enhancers from control sequences, (2) errors in the training data impact prediction accuracies in predictable ways, and (3) prediction accuracy is a useful proxy for evaluating peak calling accuracy. In the framework, data sets are constructed from peak (positive) and randomly sampled (control) sequences. Machine learning models are trained and evaluated on the sequences in a cross-chromosome (cross-fold) setup. Lastly, precision of the originating peaks are evaluated by calculating true and false positive rates.

We applied our framework to evaluate peaks for *D. melanogaster* STARR-Seq data [1] called with the MACS software [6]. Although designed for ChIP-Seq data, MACS can be used to process other types of data, but users must be careful about parameter choices. We evaluated different parameter combinations with our framework and visual comparisons of called peaks. True and false positive rates ranged from a high of 88.0% to a low of 74.7% and from a low of 18.6% to a high of 49.4%, respectively. The default MACS parameters produced the highest true and lowest false positive rates, suggesting that the default parameters are also suitable for STARR-Seq data. Our results demonstrate the utility of our framework through a practical application and provide a base for future development.

CCS CONCEPTS

• **Applied computing → Computational genomics**; **Bioinformatics**; **Recognition of genes and regulatory elements**.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

BCB '21, August 1–4, 2021, Gainesville, FL, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8450-6/21/08.
https://doi.org/10.1145/3459930.3469488

KEYWORDS

peak calling, DNA enrichment assays, enhancers, sequence modeld

ACM Reference Format:
Christopher R. Beal, John G. Peters, and Ronald J. Nowling. 2021. Sequence Model Evaluation Framework for STARR-Seq Peak Calling. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3459930.3469488

REFERENCES

[1] Cosmas D Arnold, Daniel Gerlach, Daniel Spies, Jessica A Matts, Yuliya A Sytnikova, Michaela Pagani, Nelson C Lau, and Alexander Stark. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet.* 46, 7 (July 2014), 685–692.

[2] Ling Chen and John A Capra. 2020. Learning and interpreting the gene regulatory grammar in a deep learning framework. *PLoS Comput. Biol.* 16, 11 (Nov. 2020), e1008334.

[3] Ling Chen, Alexandra E Fish, and John A Capra. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput. Biol.* 14, 10 (Oct. 2018), e1006484.

[4] Dina Hafez, Aslihan Karabacak, Sabrina Krueger, Yih-Chii Hwang, Li-San Wang, Robert P Zinzen, and Uwe Ohler. 2017. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biol.* 18, 1 (Oct. 2017), 199.

[5] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 4 (April 2014), 272–286.

[6] 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. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, 9 (Sept. 2008), R137.