Sequence Model Evaluation Framework for STARR-Seq Peak Calling

Christopher R Beal[1], John G. Peters[2], and RJ Nowling[2]

[1]Marquette University [2]Milwaukee School of Engineering

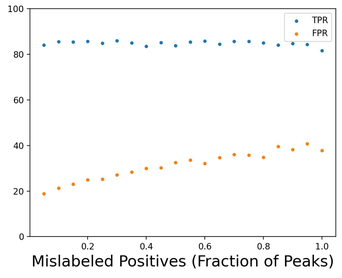
Problem

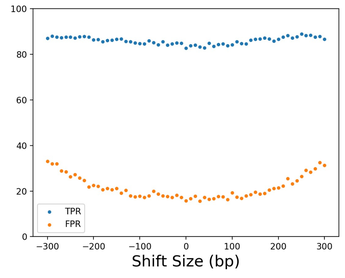
Enhancers are short regions of non-coding DNA that increase transcription rates of genes despite being located distantly from the genes themselves. Enhancers are identified through experimental techniques such as ChIP-Seq or CUT&RUN with H3K4me1 and H3K27ac histone modifications, self-transcribing active regulatory region sequencing (STARR-Seq), and massively parallel reporter assays (MPRA).

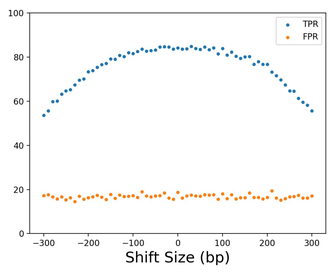
Machine learning models have been used in conjunction with experimental data to identify enhancer activity from sequences, predict enhancer-transcription factor interactions, and decode the enhancer regulatory language. If peak boundaries are not identified accurately, sequence model performance suffers. It is thus imperative that the correct parameters are used for peak calling.

Key Observations

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.

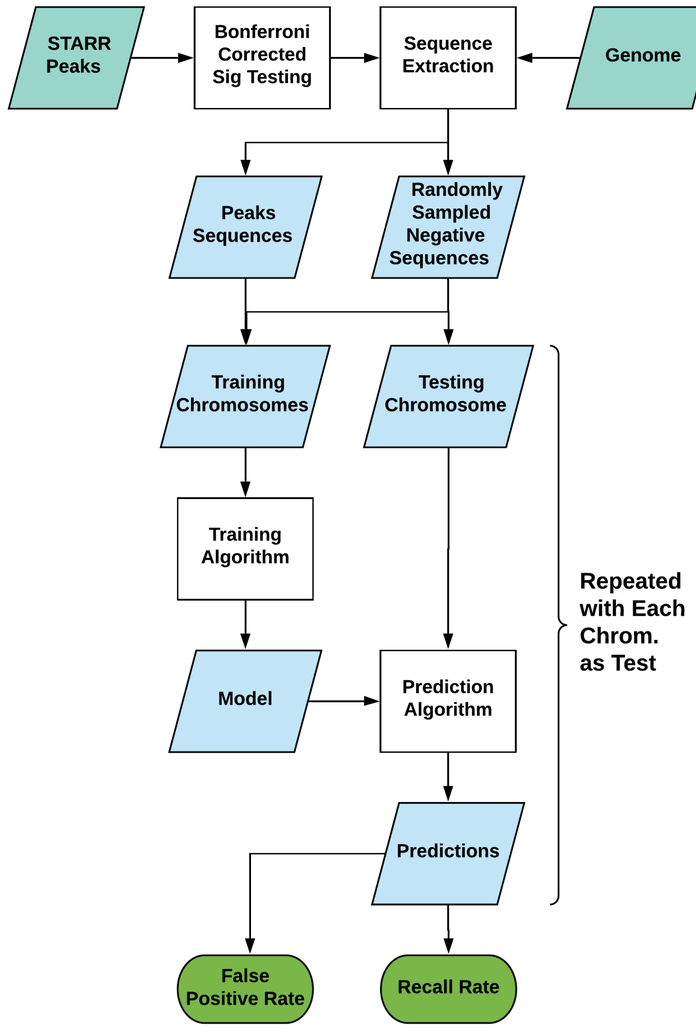






Methods

In the framework, genomic sequences are extracted for peaks (positives) and associated randomly sampled (control) locations. Sequences were divided into folds by chromosome (2L, 2R, 3L, 3R, and X). Features matrices were generated by counting k-mers of length 3 to 8 in each sequence. An ensemble of 50 logistic regression models was trained on four out of five folds and evaluated on the fifth, in a rotating fashion so that predictions were made for each chromosome. Predictions were evaluated using recall rate for the peaks and and false positive rate for the control sequences. The rates were calculated overall predictions from all of the folds.



Results

We applied our framework to evaluate the impact of MACS [2] parameters on peak calling for *D. melanogaster* STARR-Seq data [1]. 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. 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.

Parameters	True Positives (out of 2131)	False Positives (out of 2131)
bam nomodel extsize	85.2% (1816)	17.4% (371)
bam	85.2% (1816)	17.4% (371)
bampe	86.6% (1846)	22.3% (476)
bampe nomodel	86.9% (1851)	19.4% (413)
bam nomodel	85.9% (1830)	23.2% (494)
bam nomodel extsize shift	85.1% (1814)	27.6% (588)
bam keep dups auto	87.9% (1874)	26.8% (572)
bam keep dups all	73.6% (1569)	46.0% (981)

Conclusion

Our results demonstrate that the prediction accuracy of sequence models are sensitive to the precision of the called peak boundaries. Our framework provides a way to optimize peak calling parameters using sequence model prediction metrics. Going forward, we will explore the types of peak calling errors introduced by using the wrong parameters and use our framework to comparatively evaluate experimental enhancer localization techniques in terms of their precision.

References

[1] Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq.” *Science* 339 (6123): 1074–77.

[2] Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137.