C.La.P.: Leveraging ATAC-seq with transformers for context-sensitive cis-regulatory modeling

Panos Firbas Nisantzis[×], Carolina Gonçalves, Gonzalo G. de Polavieja 🔮

panos.firbas@research.fchampalimaud.org

carolina.goncalves@research.fchampalimaud.org

gonzalo.polavieja@neuro.fchampalimaud.org



The typical modeling approach

Inputs large windows of reference genomic sequence to simulate large numbers of trained

Our approach

Uses ATAC-seq to isolate CREs as samples and includes the ATAC-seq signal in the model.



Champalimaud

features.

- \rightarrow no biological context awareness
- \rightarrow CRE regions are always included
- \rightarrow non-CRE regions are also included
- \rightarrow even the largest windows are too small





- \rightarrow inactive CRE regions are not included
- → non-CRE regions are not included
- → foundational model



The model, which consists of a convolutional tokenizer, followed by a transformer-encoder body and the task heads, was pre-trained with a masking task and fine-tuned with a nucleotide classification task for CTCF, H3k27ac, and H3k4me3. It achieves F1 scores of over 0.85 per nucleotide for all three classes.





In the case of the CTCF class, the model accurately predicts CTCF binding sites, but has a harder time in predicting the edges of the target ranges. Analysis of its attention mechanism reveals its ability to pinpoint CTCF binding sites with nucleotide-level precision.





Left: By aggregating the ATAC-seq signal over multiple mA positions, we visualize the characteristic imbalance of + and - sequencing reads around events of protein binding.

Right: We can visualize the imbalance in a single line by subtracting the two signals. The imbalance around PWMs that overlap mA positions is much higher than those that don't because C.La.P. identifies real protein binding events.

Left: C.La.P.'s mA positions at times overlap PWMs outside of ground truth CTCF regions. The ATAC-seq signal imbalance shows that the model detects protein binding events even when those have not been captured by the ChIP-seq assay.

Right: PWM hits with very low score (here with score lower than 6) exibit higher ATAC-seq signal imbalance when overlapping C.La.P.'s mA positions. The model relies on more than the genomic sequence to model protein binding.

Future directions



