Analysis and classification of long terminal repeat sequences from plant LTR-retrotransposons - Leaflet

Author: Jakub Horváth[1*] Supervisor: Matej Lexa[1†]

Faculty of Informatics, Masaryk University, Botanická 68A,
602 00 Brno-Královo Pole, Czech Republic

Long Terminal Repeats (LTRs) are repetitive DNA sequences widely distributed throughout eukaryotic genomes, found in particular abundance in retrotransposons. A deeper understanding of the structural and functional characteristics of LTRs is crucial for deciphering their role in genome evolution and their involvement in disease mechanisms.

The problem with the detection and analysis of these sequences, however, lies in their inherently high repetitiveness and mutability, making them difficult to observe using traditional methods.

This thesis presents a comprehensive analysis of frequently co-occurring transcription factor binding site motifs within LTRs (using the ECLAT algorithm) and employs increasingly complex classification methods for identifying LTR sequences.

The first classification approach focuses on implementing a more simple machine learning model trained on the presence of transcription factor binding sites as features. Here, three different classifiers have been tested: Multilayer Perceptron, Random forest classifier and Gradient boosting classifier.

In the second approach, a combination of convolutional neural networks and LSTM nodes is trained on a vector representing the bases within the DNA sequence as one-hot encoded vectors.

The third and last approach features the DNABERT[1] pre-trained model fine-tuned on LTR sequences.

In order to gain a deeper insight into the structure of LTRs and avoid treating the created models as black boxes, several model interpretation techniques are employed with the aim of uncovering significant regions and features within these sequences. These include the SHAP[2], Random forest feature importance and attention analysis in the DNABERT model.

Results Out of the three models, trained the fine-tuned DNABERT model achieves the highest score with around 84% on the testing set. During the analysis of the model's predictions it was found that the most significant features which are highly specific to LTRs are located towards the beginning and end of the LTR sequence. When analyzing TF binding sites transcription factors connected to cellular stress and circadian clock regulation have emerged as the most significant in the context of LTR sequences

References

[1] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predi pdf.

*jakubhorvath119@gmail.com
†lexa@fi.muni.cz