Translation Initiation Site Prediction Using Deep Learning and Synthetic Datasets

Yunseol Park[1], Espoir Kabanga[1], Jasper Zuallaert[2,3], Hyunjin Shim[1], Arnout Van Messem[4], Wesley De Neve[1,5]*

[1] Center for Biosystems & Biotech Data Science, Ghent University Global Campus, Incheon, South Korea

[2] Center for Medical Biotechnology, VIB, Zwijnaarde, Belgium

[3] Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium

[4] Department of Mathematics, Universit\'e de Li\`ege, Li\`ege, Belgium

[5] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

* Wesley.DeNeve@ghent.ac.kr

Machine learning-based prediction has become an essential tool in genomics, seeing its adoption for structural and functional annotation of biological sequences. Furthermore, the understanding of prediction models are also important as interpretable insights can be gained into the underlying patterns in the data and the corresponding biological processes (Raghu & Schmidt, 2020). As such, building a prediction model for translation initiation site (TIS) and determining its important features may aid in uncovering new translation mechanisms and give emphasis to already existing ones. However, interpretation is difficult, as many machine learning models are black box in nature. Therefore, to better understand the relevant feature interpretability, we investigate the use of synthetic data in the context of TIS prediction for *A. thaliana* and, through transfer learning, for *H. sapiens*.

As shown in **Figure 1**, our synthetic dataset for TIS prediction was generated *de novo* with five features, taken from previous feature analysis results (Zeng et al., 2002; Zuallaert et al., 2018): (1) *consensus sequence*, (2) *upstream ATG*, (3) *downstream stop codon*, (4) *donor splice site*, and (5) *nucleotide frequency*. Furthermore, the real dataset (Magana-Mora et al., 2012) and the synthetic dataset were used in a 1:1 ratio to generate the combined dataset, thus resulting in three datasets of the same size. We used the synthetic, real, and combined datasets of *A. thaliana* genomes to train a synthetic black-box model (SBBM), a real black-box model (RBBM), and a combined black-box model (CBBM), respectively. Furthermore, we conducted feature and noise analysis via occlusion in the synthetic dataset. Here, we leveraged another feature, namely *codon usage*, replacing *nucleotide frequency* with *codon frequency*. Finally, using the above models, we investigated the effectiveness of transfer learning using a small human dataset (Chen et al., 2014).

Through our experiments, we found that SBBM learns similar features as RBBM, although the magnitude of the features was different. Furthermore, we noticed that CBBM has similar prediction effectiveness as RBBM, achieving an accuracy of 90.68% and 90.74%, respectively. Among the features studied, we observed that *consensus sequence* and *nucleotide frequency* have the most significant positive influence on TIS prediction, with *downstream stop codon* also having a positive influence. We found *donor splice site* and *upstream ATG* to be a less influential feature, where the case of *upstream ATG* may point to the capability of the model to learn leaky scanning. On the other hand, we observed *codon usage* to be a negatively influencing feature, possibly because of the random way in which we added codons, resulting in non-meaningful k-mers ($k \neq 3$). Finally, through transfer learning, we found that RBBM obtains the highest prediction effectiveness, at 85.40% accuracy, when used as a pre-trained model. Nonetheless, the prediction effectiveness of CBBM was close to that of RBBM, at 83.75% accuracy. We hypothesize that we could create a more general and effective TIS prediction model for all eukaryotes if more information could be gathered on the TISs for different species.

We can conclude that SBBM and RBBM learn from similar features, with CBBM obtaining a similar effectiveness as RBBM. We also found that CBBM could be used to reduce overfitting when training with small datasets. Furthermore, we found that *consensus sequence* and *nucleotide frequency* are the most positively influencing features, while *codon usage* was found to be a negatively influencing feature. On the other hand, the models seemed to learn leaky scanning, as shown by the less influential nature of *upstream ATG*.

In summary, through this case study on TIS prediction for *A. thaliana* and *H. sapiens*, we were able to gain insight into (1) the potential of leveraging synthetic data for the interpretation of black-box prediction models and (2) the prediction potential of models trained using a combination of synthetic and real data.

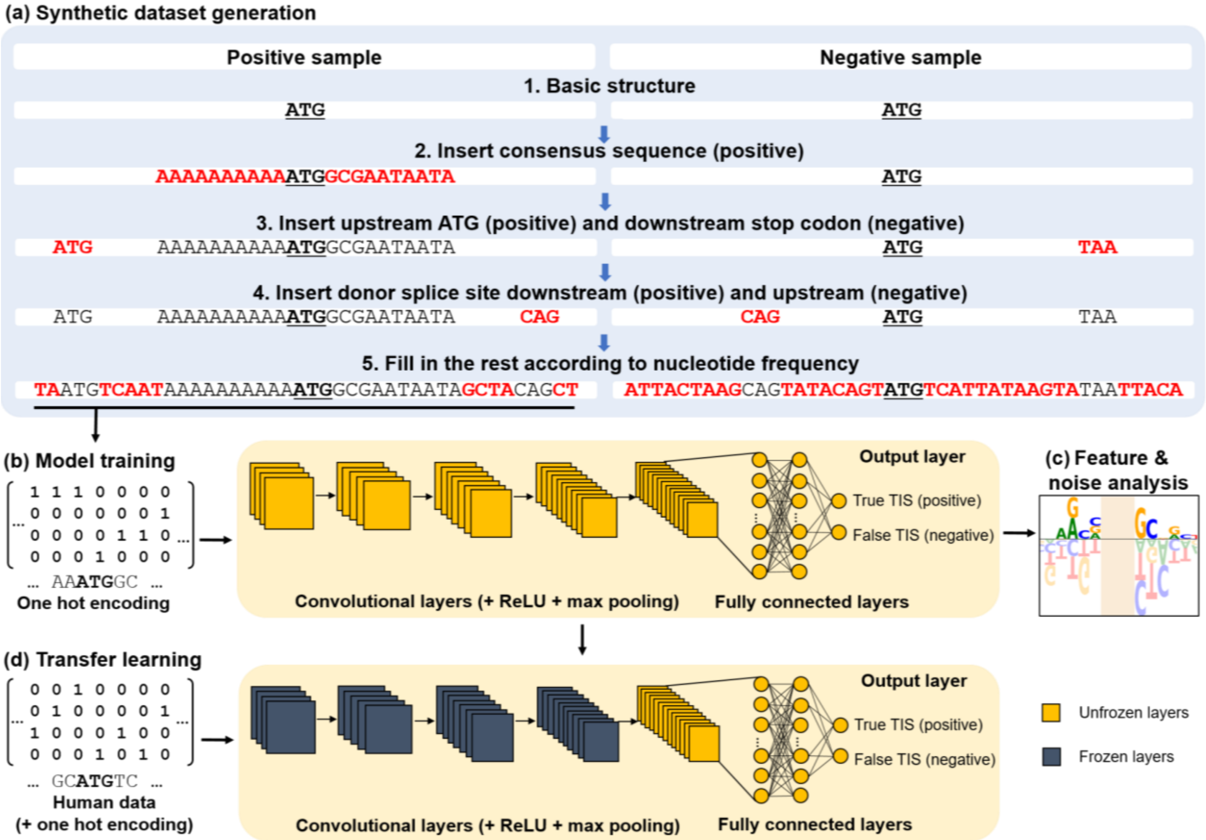

Figure 1 – Overview of the methodology used

(a) Generation of the synthetic dataset. The candidate TIS is shown in bold and underlined, while the changes are indicated in red. (b) Creation of TIS prediction models, with the synthetic, real, and combined datasets all undergoing the same training procedure. (c) Feature and noise analysis of the results and interpretation. (d) Transfer learning with human data.

REFERENCES

Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H., & Chou, K.-C. (2014). iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry, 462*, 76–83. doi: 10.1016/j.ab.2014.06.022.

Magana-Mora, A., Ashoor, H., Jankovic, B. R., Kamau, A., Awara, K., Chowdhary, R., Archer, J. A., & Bajic, V. B. (2012). Dragon TIS Spotter: an Arabidopsis-derived predictor of translation initiation sites in plants. *Bioinformatics, 29(1)*, 117–118. doi: 10.1093/bioinformatics/bts638.

Raghu, M. & Schmidt, E. (2020). A Survey of Deep Learning for Scientific Discovery. CoRR, arXiv:2003.11755.

Zeng, F., Yap, R. H., & Wong, L. (2002). Using Feature Generation and Feature Selection for Accurate Prediction Of Translation Initiation Sites. *Genome Informatics, 13,* 192–200. doi: 10.11234/gi1990.13.192.

Zuallaert, J., Kim, M., Soete, A., Saeys, Y., & De Neve, W. (2018). TISRover: ConvNets learn biologically relevant features for effective translation initiation site prediction. *International Journal Of Data Mining and Bioinformatics, 20(3),* 267–284. doi: 10.1504/ijdmb.2018.10016079.