Improvements over DeepGSR

Daneil Godoy[1][0000−1111−2222−3333] and Luis Barata[2,3][1111−2222−3333−4444]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. We introduce a lightweight Transformer-based framework for recognizing genomic regulatory signals, such as polyadenylation sites (PAS) and translation initiation sites (TIS). Our approach significantly reduces model size while maintaining competitive performance. Using trinucleotide tokenization and models with fewer than 2.2 million parameters, we train on PAS and TIS datasets from human, mouse, bovine, and fruit fly. The models achieve accuracies of up to 83.5% for PAS and 93.1% for TIS. Compared to alternative low-capacity models, the Transformer outperforms in precision and recall, offering a practical solution for bioinformatics applications in resource-limited environments.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

The accurate identification of regulatory signals in genomic sequences is a fundamental problem in computational genomics, as these signals play a critical role in gene expression and protein synthesis. Among the most important of such signals are polyadenylation sites (PAS), which define the cleavage and polyadenylation of messenger RNA (mRNA), and translation initiation sites (TIS), which determine where protein translation begins. Errors in the annotation of these sites can lead to incorrect gene models, misinterpretation of transcript structures, and inaccurate downstream biological analyses. As a result, reliable computational methods for PAS and TIS recognition are essential components of modern genome annotation pipelines.

Traditional approaches to signal recognition have relied on hand-crafted features and probabilistic models, such as position weight matrices and hidden Markov models. While these methods are computationally efficient, they struggle to capture the complex and long-range dependencies present in genomic sequences, particularly in regions surrounding regulatory motifs. The increasing availability of large-scale genomic datasets has therefore motivated the adoption of deep learning techniques, which can automatically learn hierarchical and context-dependent representations directly from raw sequence data.

Convolutional neural networks (CNNs) have been widely used for detecting genomic signals due to their ability to model local sequence patterns. DeepGSR represents a notable example, achieving strong performance by encoding genomic regions as image-like representations and applying a deep 2D-CNN architecture to recognize PAS and TIS motifs. However, this performance comes at the cost of extremely large models.

Transformer architectures are a powerful alternative for modeling sequential data. Originally developed for natural language processing, Transformers use self-attention mechanisms to capture both local and global dependencies within sequences, making them well suited for genomic data.

In this work, we address the challenge of balancing predictive performance with computational efficiency for genomic signal recognition. We propose a lightweight Transformer-based framework for PAS and TIS motif detection that significantly reduces model size while preserving strong classification performance. By employing a compact trinucleotide tokenization scheme and carefully constraining model capacity, we design Transformer encoder models. We evaluate both organism-specific models and models trained jointly across multiple species, allowing us to analyze trade-offs between specialization and generalization.

2 State-of-the-Art in Genomic Signal Prediction

Modern approaches to predicting genomic signals such as polyadenylation sites (PAS) and translation initiation sites (TIS) rely heavily on deep learning models trained on raw DNA sequences. Early work such as DeepGSR [1] used convolutional neural networks (CNNs) to recognize multiple genomic signals simultaneously. DeepGSR dramatically reduced classification error for human PAS and TIS prediction (by up to 29–86%) compared to prior methods. It was evaluated across four species and shown to outperform earlier sequence-based predictors. Nevertheless, DeepGSR was limited to two signals (PAS and TIS) and left room for further improvement.

2.1 Translation Initiation Site (TIS) Prediction

For TIS prediction, recent state-of-the-art models also use deep learning. For example, DeepTIS [2] employs a two-stage architecture combining CNNs and recurrent networks to explicitly model coding-frame features around each ATG start codon. DeepTIS achieves significantly higher accuracy than previous methods on genome-wide human and mouse data. Earlier deep models include TISRover [8], which applied a CNN to learn features like the Kozak sequence and reading-frame context directly from DNA. Other approaches (e.g. NeuroTIS and NetStart) use hybrid CNN/RNN or protein language models, but DeepTIS remains a leading method for genomic TIS prediction. Overall, these modern TIS predictors improve over traditional SVM or motif-scanning methods by learning complex sequence patterns automatically.

2.2 Polyadenylation Signal (PAS) Prediction

Polyadenylation signal (PAS) prediction has similarly benefited from deep networks. Deep learning models now identify PAS variants and cleavage sites with high accuracy. PolyaID [4] is a recent example: it uses a CNN to find putative polyA cleavage sites at nucleotide resolution without requiring a pre-specified signal motif. PolyaID captures position-specific motif interactions and yields an unbiased genome-wide PAS predictor. Other deep models include DeepPASTA [5], which scores sequences for polyA potential, and APARENT or PolyApredictors [6, **?**] that predict PAS strength or usage. However, these earlier models often rely on known PAS motifs (e.g. "AAUAAA") or annotated training sites. PolyaID was designed to overcome this by performing cleavage-site prediction at base-pair resolution. Additional recent tools (e.g. SANPolyA, DeeReCT-PolyA, DeeReCT-APA) use CNN or hybrid networks to classify PAS variants and support alternative polyadenylation predictions. In summary, state-of-the-art PAS predictors combine one-hot encoding of sequence with deep architectures to learn regulatory motifs and their combinations directly.

2.3 Integrated and Transfer-Learning Approaches

Beyond specialized models, new frameworks aim to predict multiple genomic signals jointly. For instance, DeepGenGrep [3] is a hybrid CNN–LSTM network that simultaneously classifies PAS, TIS, and splice sites. It was trained on multi-species data and outperformed both DeepGSR and individual-task predictors on all tasks. DeepGenGrep thus demonstrates improved robustness and cross-task transfer learning: the network's features generalize across signal types and organisms. This trend towards generalized models recognizes that genomic signals interact, and that jointly learning them can improve annotation accuracy.

2.4 Related Genomic Tasks

Similar deep-learning strategies have been applied to other genomic signals as well. For example, promoters and transcription start sites are predicted by CNN models (e.g. DeeReCT-PromID, ReFeaFi), and splice junctions by tools like SpliceRover and SpliceFinder. All these methods share the approach of encoding raw sequence (often via one-hot or k-mer embeddings) and using deep networks to learn discriminative features. Thus, the modern state of the art for GSR prediction across the board is dominated by deep neural architectures that automatically extract regulatory motif patterns, often exceeding the performance of traditional feature-based classifiers.

3 Materials and Methods

The Transformer architecture's primary inputs are sequences. In the case of language modeling the task it was originally developed for—the sequences are

composed of tokens, sub-units of the sequence that must be defined *a priori* [27]. The set of all possible tokens is referred to as the model's vocabulary. Modern language models use a variety of tokenization procedures, mostly based on splitting words into some of their components, such as prefixes and suffixes, even though the tokens need not be grammatically meaningful [23, **?**].

In the context of genomic data, the vocabulary can be reduced to as few as four tokens—adenine, cytosine, guanine, and thymine [**?**]. Alternatively, one can use larger vocabularies by considering small sequences of two, three, or more nucleotides as tokens [15].

3.1 Datasets

The dataset used in this study follows the construction procedure introduced in DeepGSR, which provides polyadenylation site (PAS) and translation initiation site (TIS) sequences for four organisms: human, mouse, bovine, and fruit fly. The data were derived from publicly available cDNA resources from NCBI, the UCSC Genome Browser, the Mammalian Gene Collection (MGC), FlyBase, and Ensembl [9, 14, 24, 26]. cDNA sequences were mapped to the corresponding reference genomes using GMAP [29], and genomic regions surrounding each signal were extracted with `bedtools` [21]. For each PAS and TIS event, 300 nucleotides upstream and downstream were included, yielding final sequence lengths of 606 nucleotides for PAS and 603 nucleotides for TIS.

DeepGSR additionally provides matched negative samples consisting of false PAS and false TIS motifs. These sequences share the same hexamer or trinucleotide patterns as true signals but lack any biological association with polyadenylation or translation. Negative samples were selected to match the number of positive samples and were drawn from chromosomes with GC-content closest to the genome-wide average for each species: chromosomes 21, 13, 28, and X for human, mouse, bovine, and fruit fly, respectively.

The final dataset comprises 20,933, 18,693, 12,082, and 27,203 true PAS sequences across the four species, spanning 16 PAS motif variants. For TIS signals, the dataset contains 28,244, 25,205, 17,558, and 30,283 sequences for human, mouse, bovine, and fruit fly, respectively, all centered on the canonical `ATG` start codon. These counts, along with motif-level distributions, indicate substantial variability across species in both signal frequency and motif usage.

The original dataset was composed of several FASTA files containing the raw sequences, as well as additional text files with preprocessed data. In this study, we compiled the FASTA files into a unified dataset that can be easily queried and browsed. The resulting dataset is publicly available at `https://huggingface.co/datasets/dvgodoy/DeepGSR_sequences`. However, for the experiments presented here, we filtered the data to retain only the `AATAAA` motif for PAS and the canonical `ATG` motif for TIS.

3.2 Proposed Method

Transformer-based models are typically large, ranging from hundreds of millions to tens of billions of trainable parameters [10, 11]. Their size poses challenges for deployment to end users, as browser-based solutions cannot efficiently load and execute such models on consumer hardware. Using large pretrained models as initialization for fine-tuning would therefore require server-side execution, which is consistent with the deployment patterns of many existing bioinformatics tools [**?**].

For this reason, we chose to train compact Transformer encoder models containing up to two million trainable parameters, enabling the possibility of an efficient browser-based solution. Separate Transformer models were trained for each organism and signal (PAS or TIS), along with two additional models trained jointly on all four organisms, one for each signal.

In the following subsections, we outline the most relevant aspects of the model development, including the choice of data representation and the parametrization of the Transformer architecture and training procedure [27].

Data Representation Genomic data are inherently sequential, which naturally constrains the choice of data representation, in particular the vocabulary size. Since PAS and TIS signals are six and three nucleotides long, respectively, we chose trinucleotides as tokens, resulting in a vocabulary of 64 distinct tokens. Genomic sequences were tokenized analogously to word tokenization in language models, that is, without overlaps, as is common in genomic sequence processing through the use of k-mers [**?**,22].

This restriction was mitigated through data augmentation by randomly selecting one of the first three nucleotides of each sequence as the starting point for tokenization. The resulting preprocessed dataset is publicly available at `https://huggingface.co/datasets/dvgodoy/DeepGSR_trinucleotides`. Each sequence was trimmed to include exactly 99 tokens upstream and 99 tokens downstream of the signal, thereby avoiding the use of padding tokens. These subsequences were concatenated to form fixed-length inputs of 198 tokens for the Transformer model.

Model Selection The Transformer architecture is highly flexible, and its parametrization allows for a wide range of model capacities and sizes [27]. To construct a Transformer encoder model, we specify the following hyperparameters: the number of stacked Transformer blocks (`num_layers`), the number of attention heads per block (`n_heads`), the dimensionality of the hidden state (`d_model`), the dimensionality of the feed-forward network within each block (`dim_feedforward`), the dropout probability (`dropout`, fixed at 10% for all models), the number of output classes (`num_classes`), and the optional use of a special classifier token prepended to the input sequence (`use_cls_token`) [11].

All models were trained as binary classifiers to predict whether an input sequence corresponds to a true PAS or TIS motif (positive class) or to a false motif

(negative class). Models trained for individual organisms consist of four Transformer blocks, six attention heads, a hidden dimensionality of 192 (corresponding to 32 dimensions per head), and a feed-forward network of dimensionality 768, resulting in approximately 1.8 million trainable parameters. Models trained jointly on all organisms also use four Transformer blocks but employ four attention heads and a hidden dimensionality of 256 (64 dimensions per head), together with a feed-forward network of dimensionality 512, yielding approximately 2.1 million trainable parameters. These multi-organism models additionally incorporate a prepended classifier token to aggregate sequence-level representations [11].

Trainning Arguments Once model capacity was established, regularization techniques were incorporated during training to improve generalization and reduce the training–validation gap. All models were trained using the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01 [18], and the cross-entropy loss with label smoothing set to 0.05 [25]. Learning rate scheduling combined a linear warm-up phase of 16 steps with cosine decay over 254 steps [27, 17]. Training was performed for up to 100 epochs with early stopping after 10 consecutive epochs without improvement in validation loss [20], using mini-batches of 256 sequences.

Each model was trained multiple times using different random seeds for model initialization, as well as different dataset splits obtained via reshuffling, in order to assess the stability of the observed results [12]. Across all runs, the resulting performance metrics varied by at most one percentage point.

4 Results

The dataset used in this study was balanced, containing an approximately equal number of positive and negative samples for each organism and signal. Therefore, we report results for accuracy, precision (for both classes), and recall (for both classes), which are standard evaluation metrics for binary classification tasks. While accuracy reflects the overall performance of the model, precision and recall enable a more detailed analysis of model behavior with respect to each class. Higher precision, particularly for the positive class corresponding to real motifs, indicates a stricter and more reliable model with a low false-positive rate (most detections are correct), albeit at the cost of a higher false-negative rate (some real motifs will be missed). Conversely, higher recall for the positive class indicates a less strict but more exhaustive model that detects most real motifs, reducing false negatives while increasing false positives (many false motifs will be included as well).

4.1 Signal Recognition

Single-organism models were evaluated on the same organism used for training them. The general models were evaluated on all organisms.

For PAS, the accuracy of the best model (either single-organism or general) for each organism ranged from 82.02% (mouse) to 83.53% (human). In the case of human and fruit fly organisms, the general model trained on all four organisms was slightly more accurate than the corresponding single-organism model. In the case of bovine and mouse organisms, it was the single-organism model that slightly outperformed the general one. The general model consistently outperformed the single-organism models both in the precision of the positive class, ranging from 84.73% (fruit fly) to 87.81% (human), as well as in the recall of the negative class, ranging from 86.04% (fruit fly) to 89.79% (human). Single-organism models outperformed the general model both in the precision of the negative class, ranging from 81.13% (bovine) to 85.56% (human), as well as in the recall of the positive class, ranging from 80.66% (bovine) to 86.52% (human). Figures 1 and 3 show the results obtained for PAS.

For TIS, the accuracy of the best model (either single-organism or general) for each organism ranged from 90.99% (fruit fly) to 93.14% (mouse). In the case of bovine and mouse organisms, the general model trained on all four organisms was slightly more accurate than the corresponding single-organism model. In the case of human and fruit-fly organisms, it was the single-organism model that slightly outperformed the general one. The general model outperformed single-organism models except for the fruit fly both in the precision of the positive class, ranging from 94.14% (bovine) to 95.52% (mouse), as well as in the recall of the negative class, ranging from 94.34% (bovine) to 95.77% (mouse). The single-organism model for the fruit fly exhibited 92.93% precision for the positive class and 93.22% recall for the negative class. Single-organism models, except for the fruit fly, outperformed the general model both in the recall of the positive class, ranging from 91.53% (bovine) to 92.71% (mouse). The general model tested on the fruit fly exhibited 89.11% recall for the positive class. Figures 2 and 4 show the results obtained for TIS.

4.2 Comparison with 2D-CNN and Baseline Methods

The results in Figures 1 and 2 also include a comparison of our Transformer models to two additional models, a small 2D-CNN model inspired in the original DeepGSR architecture (485,794 trainable parameters), and a logistic regression as baseline (65 trainable parameters), both trained on the human organism only. The inputs had to be adapted for each of these models. As for the CNN model, we used the same representation as the original DeepGSR model, except that our image-like inputs had two channels instead, one for the upstream sequence and the other for the downstream sequence. The logistic regression was trained on a bag of one-hot encoded tokens, each token being a trinucleotide as described in Section 2.2.1.

While the performance of the small 2D-CNN model, given its low capacity (the original DeepGSR 2D-CNN model had over 100 million trainable parameters), was below 80% in most metrics for both PAS and TIS, the logistic regression was second to the best Transformer model for each signal on both precision for the positive class (82.10% against 87.81% of the general model on

PAS, 92.88% against 94.77% on TIS) and recall for the negative class (86.16% against 89.79% of the general model on PAS, 94.57% against 94.96% on TIS).

5 Conslusion

In this work, we investigated the effectiveness of small Transformer-based models for the recognition of genomic signals, focusing on polyadenylation sites (PAS) and translation initiation sites (TIS) across four organisms. By constraining model size to approximately two million trainable parameters, we aimed to balance performance with practical deployability, enabling the possibility of efficient browser-based inference without depending on server-side computation.

The results demonstrate that small Transformer models achieve strong and consistent performance on both PAS and TIS recognition tasks. For PAS, accuracies above 81% were obtained across all organisms, while TIS recognition reached accuracies exceeding 89%, confirming that the proposed models capture relevant sequence-level patterns despite their size. The comparison between single-organism and general models highlights a clear trade-off: models trained on multiple organisms tend to be more conservative, exhibiting higher precision for the positive class and stronger recall for the negative class, whereas single-organism models generally favor higher recall for true signals. This behavior suggests that multi-organism training improves robustness and specificity, while organism-specific training enhances sensitivity.

The comparative analysis with a lightweight 2D-CNN and a logistic regression baseline further underscores the advantages of the Transformer architecture, which consistently achieved superior precision for true motifs and better discrimination of false signals. The relatively strong performance of logistic regression indicates that trinucleotide-level representations already encode meaningful information; however, the Transformers' ability to model contextual dependencies yields performance improvements, particularly in reducing false negatives.

Overall, these findings show that small Transformer models constitute a viable and effective alternative to larger deep-learning architectures for genomic signal recognition. They offer competitive performance while significantly reducing computational requirements, making them well suited for interactive and client-side bioinformatics applications. Future work may explore extending this approach to additional genomic signals, using alternative tokenization strategies, or further optimizing model architectures to improve sensitivity without sacrificing specificity.

References

1. M. Kalkatawi et al., DeepGSR: an optimized deep learning structure for the recognition of genomic signals and regions. Bioinformatics 35(7):1125–1132 (2019).
2. C. Wei, J. Zhang, X. Yuan, DeepTIS: Improved translation initiation site prediction in genomic sequence via a two-stage deep learning model. (2021).

3. Q. Liu, F. Li, J. Song, DeepGenGrep: a general deep learning-based predictor for multiple genomic signals and regions. Bioinformatics 38(17):4053–4061 (2022).
4. E. K. Stroup, Z. Ji, Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. Nat. Commun. 14:7378 (2023).
5. A. Arefeen et al., DeepPASTA: A deep learning model for polyadenylation site prediction. Bioinformatics 35(18):3223–3231 (2019).
6. P. Bogard et al., APARENT: A deep learning-based approach for polyadenylation site prediction. BMC Bioinformatics 20(1):1-11 (2019).
7. W. Zhang et al., PolyApredictors: a tool for polyadenylation site prediction using a deep neural network. Nucleic Acids Research 47(9):4290-4298 (2019).
8. J. Zhang et al., TISRover: A deep learning-based method for predicting translation initiation sites. BMC Genomics 18(1):1-12 (2017).
9. B. L. Aken et al.: The Ensembl gene annotation system. *Database*, 2016, baw093. https://doi.org/10.1093/database/baw093
10. T. B. Brown et al.: Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165
11. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805
12. J. Dodge, S. Gururangan, D. Card, R. Schwartz, N. A. Smith: Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. https://arxiv.org/abs/2002.06305
13. G. Eraslan, Ž. Avsec, J. Gagneur, F. J. Theis: Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics*, 20, 389–403, 2019. https://doi.org/10.1038/s41576-019-0122-6
14. L. S. Gramates et al.: FlyBase at 25: Looking to the future. *Nucleic Acids Research*, 45(D1), D663–D671, 2017. https://doi.org/10.1093/nar/gkw1016
15. Y. Ji et al.: DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers for DNA-sequence data. https://academic.oup.com/bioinformatics/article/37/15/2112/6129079
16. M. Kalkatawi et al.: DeepGSR: An optimized deep-learning framework for genomic signal recognition. *Bioinformatics*, 35(7), 1125–1132, 2019. https://doi.org/10.1093/bioinformatics/bty752
17. I. Loshchilov, F. Hutter: SGDR: Stochastic Gradient Descent with Warm Restarts. https://arxiv.org/abs/1608.03983
18. I. Loshchilov, F. Hutter: Decoupled Weight Decay Regularization. https://arxiv.org/abs/1711.05101
19. G. Marçais, C. Kingsford: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6), 764–770, 2011. https://doi.org/10.1093/bioinformatics/btr011
20. L. Prechelt: Early stopping — but when? In: Neural Networks: Tricks of the Trade, 1998. https://link.springer.com/chapter/10.1007/3-540-49430-8_3
21. A. R. Quinlan, I. M. Hall: BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842, 2010. https://doi.org/10.1093/bioinformatics/btq033
22. G. Reinert, D. Chew, F. Sun, M. S. Waterman: Alignment-free sequence comparison (I): Statistics and power. *Journal of Computational Biology*, 16(12), 1615–1634, 2009. https://doi.org/10.1089/cmb.2009.0198
23. R. Sennrich, B. Haddow, A. Birch: Neural Machine Translation of Rare Words with Subword Units. https://arxiv.org/abs/1508.07909

24. R. L. Strausberg et al.: The Mammalian Gene Collection (MGC). *Science*, 286(5439), 455–457, 1999. https://doi.org/10.1126/science.286.5439.455
25. C. Szegedy et al.: Rethinking the Inception Architecture for Computer Vision. https://arxiv.org/abs/1512.00567
26. G. Temple et al.: The completion of the Mammalian Gene Collection (MGC). *Genome Research*, 19(12), 2324–2333, 2009. https://doi.org/10.1101/gr.095976.109
27. A. Vaswani et al.: Attention Is All You Need. https://arxiv.org/abs/1706.03762
28. J. D. Watson, F. H. C. Crick: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171, 737–738, 1953. https://www.nature.com/articles/171737a0
29. T. D. Wu, C. K. Watanabe: GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), 1859–1875, 2005. https://doi.org/10.1093/bioinformatics/bti310
30. Y. Wu et al.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. https://arxiv.org/abs/1609.08144