Classification of Antibiotic Resistance Gene Sequences using Nucleotide Transformer

Prasanth Kumar Thuthika
Yesasvi Sai Nandigam

Transformers

The Transformer model, first introduced by Vaswani et al. (2017), represents a groundbreaking development in natural language processing (NLP) and artificial intelligence. Unlike traditional sequence-based models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which process inputs sequentially, Transformers employ an innovative mechanism called self-attention. This mechanism enables the model to analyse all elements of a sequence simultaneously, allowing for the efficient capture of contextual relationships across both short and long distances. By eliminating the sequential nature of processing, the Transformer architecture accelerates training and improves scalability for large datasets. The design incorporates an encoder-decoder structure, where the encoder processes input data and the decoder generates output sequences, both relying heavily on self-attention and feed-forward neural layers. Moreover, the use of positional encodings helps the model preserve the order of sequence elements, addressing its otherwise order-agnostic nature. These advancements have made Transformers a foundational component in state-of-the-art applications, inspiring influential models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020). The Transformer's versatility has extended beyond NLP into other areas, including computer vision, bioinformatics, and protein structure prediction.

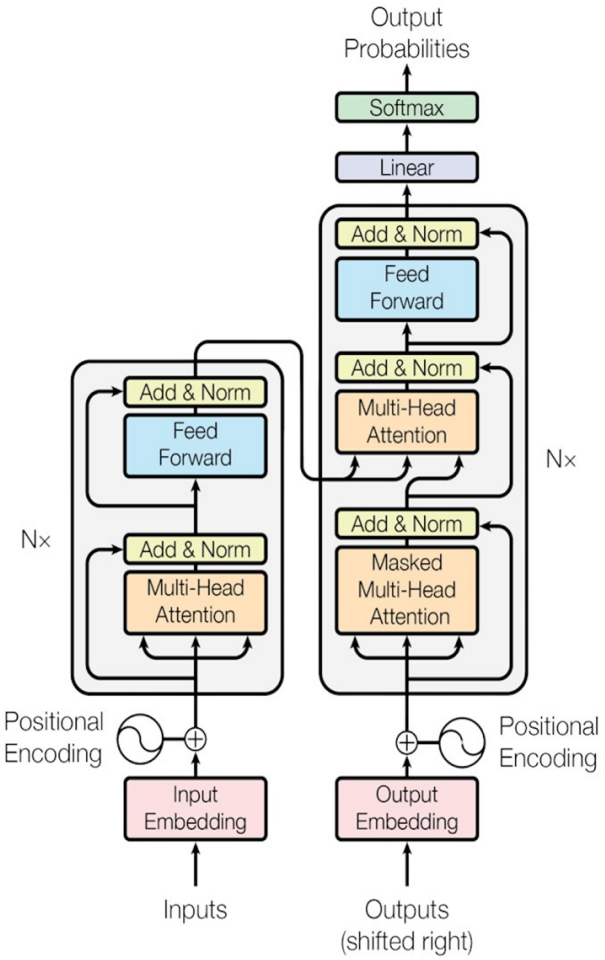

Fig-1: Nucleotide Transformer Model Architecture

A Transformer model processes input using an encoder-decoder structure driven by the self-attention mechanism. The input sequence is first tokenized and converted into embeddings, which are then enriched with positional encodings to retain the order of tokens. In the encoder, self-attention calculates relationships between all tokens, highlighting relevant parts of the sequence for each token. These attention scores are combined with feed-forward neural networks to produce a set of encoded representations. The decoder takes these encoded representations and, along with the previously generated output tokens, applies a similar self-attention mechanism and cross-attention to align the input and output sequences. Finally, the decoder generates the output token by token, passing through a SoftMax layer to predict the next word in the sequence. This process ensures efficient, parallelized computation while capturing both local and global dependencies.

Nucleotide Transformer

The Nucleotide Transformer model marks a significant advancement in bioinformatics by applying transformer architectures to DNA and RNA sequences. Derived from the original Transformer model (Vaswani et al., 2017), which was primarily designed for natural language processing (NLP), the Nucleotide Transformer adapts this approach to process biological sequences. These models are trained to understand the "language" of nucleotides, drawing parallels between the sequential structure of genetic data and the linguistic patterns in text. By leveraging self-attention mechanisms, Nucleotide Transformers capture the complex and long-range dependencies within nucleotide sequences, which are critical for understanding regulatory elements, mutations, and functional domains in genomes.

Three advanced Nucleotide Transformer models highlight the growing capabilities of transformer architectures in bioinformatics. The first model, trained on over 500 million nucleotide sequences from multiple species, demonstrates robust performance across diverse genomic tasks such as transcription factor binding site prediction and motif discovery. The second model, with 3.5 billion parameters, significantly expands the scope of analysis, enabling deep insights into complex biological patterns like sequence evolution, enhancer activity prediction, and cross-species genomic comparisons. A third, even larger-scale model, with over 10 billion parameters, was trained on an extensive dataset comprising both DNA and RNA sequences. This model excels in tasks such as de novo sequence generation, long-range dependency detection, and fine-grained functional annotation, proving invaluable for high-resolution genomic studies. Together, these Nucleotide Transformers exemplify the potential of large-scale, multi-species training in revolutionizing genomic research and precision medicine.

Several versions of Nucleotide Transformers have been developed, each tailored to specific use cases and datasets. For example, DNABERT (Ji et al., 2021) was trained on a large corpus of genomic sequences and focuses on tasks such as sequence classification, mutation impact prediction, and motif discovery. Another model, Genome Transformer, is optimized for whole-genome data analysis, leveraging extensive datasets from public repositories like GenBank and ENCODE. These models often use k-mer tokenization strategies, which segment sequences into overlapping substrings of fixed lengths, enabling the transformer to process biological sequences effectively. The datasets used to train Nucleotide Transformers are typically vast and diverse, including reference genomes, transcriptomic data, and epigenomic datasets. For instance, DNABERT was trained on sequences from human and other species' genomes, incorporating diverse genomic regions to enhance its ability to generalize across different tasks. The training data is pre-processed to ensure uniformity, with careful consideration of sequence length, encoding, and biological relevance.

The use cases for Nucleotide Transformers are diverse and impactful. These models excel in predicting transcription factor binding sites, identifying enhancer elements, and classifying diseases based on genomic mutations. They are also increasingly used in drug discovery, where they help identify therapeutic targets by

analysing genomic and transcriptomic data. Additionally, Nucleotide Transformers have been applied in microbiome analysis, enabling the characterization of microbial communities and their functional potential. By improving the interpretability and accuracy of sequence-based analyses, Nucleotide Transformers are transforming how researchers study complex biological systems.

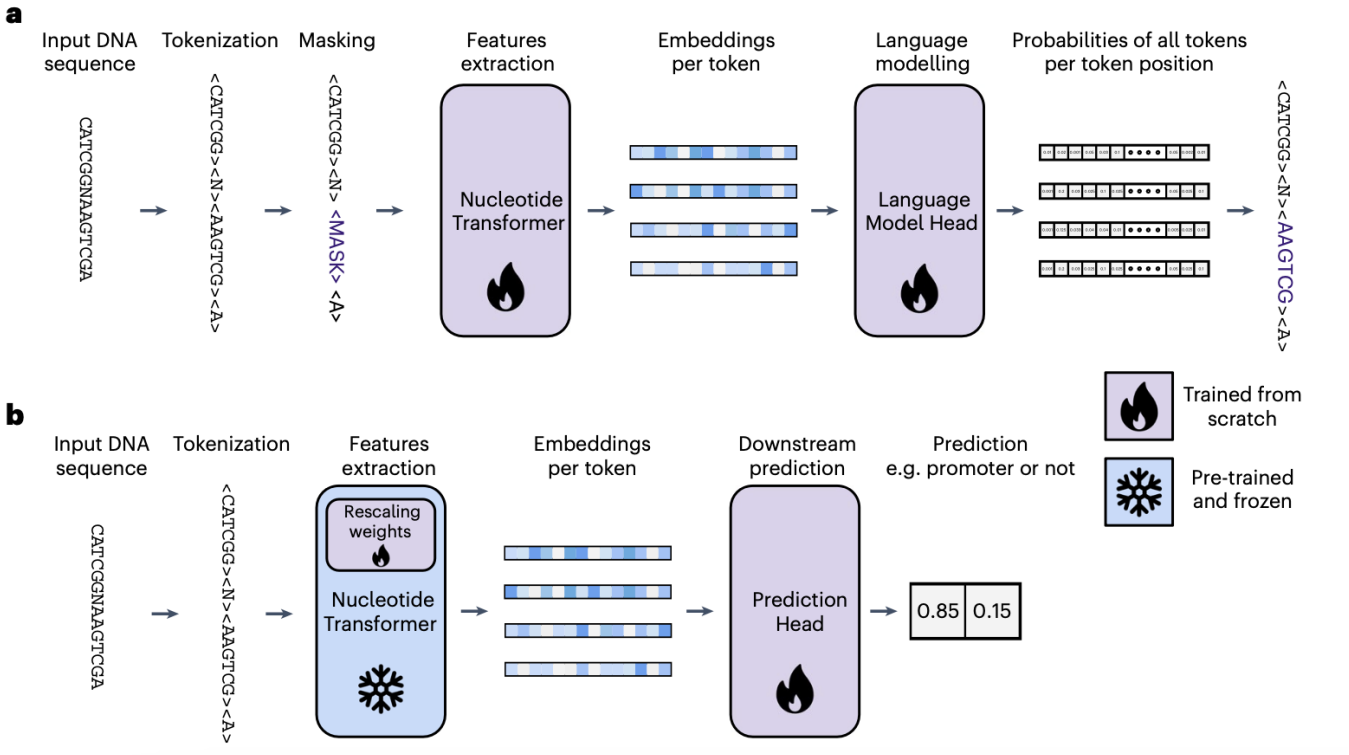

Fig-2a: Application for downstream genomics prediction tasks through fine-tuning

Fig-2b: Downstream task prediction through probing is similar but without the rescaling weights in the N.T

Objective

Antibiotic resistance poses a critical global health challenge, driven largely by the spread of antibiotic resistance genes (ARGs) in bacteria, which reduce the efficacy of antibiotics in treating infections (Partridge et al., 2018). Machine learning (ML) and deep learning (DL) methods have shown promising applications in the classification and prediction of ARGs, enabling researchers to analyse large-scale genomic data and identify resistance-associated patterns with high accuracy (Yang et al., 2020). Current approaches utilize convolutional neural networks (CNNs), recurrent neural networks (RNNs), and pre-trained transformer-based models for sequence classification and resistance prediction. These methods have significantly improved the speed and precision of ARG detection compared to traditional bioinformatics tools.

The objective of this project is to utilize the Nucleotide Transformer model, a cutting-edge transformer-based architecture, and fine-tune it specifically for the classification of antibiotic resistance gene (ARG) sequences. By training the model on curated, labelled datasets containing resistant, non-resistant gene sequences, synthetic non-resistant gene sequences the goal is to enable the accurate identification and classification of

ARGs. The fine-tuned model will be tested and validated using independent datasets to evaluate its performance, robustness, and generalizability across a diverse range of bacterial species.

Data

The data for this project is sourced from the CARD database https://card.mcmaster.ca and consists of 22,000 antibiotic-resistant gene sequences. From this dataset, 10,000 sequences were randomly selected. Additionally, 5,000 sequences were extracted from the rest of the data and modified by introducing mutations and shuffling nucleotide positions to create synthetic non-resistant gene sequences. Non-resistant gene sequences were downloaded from the RefSeq database https://www.ncbi.nlm.nih.gov/refseq/ . To ensure there were no repeated sequences between the non-resistant gene sequences and the resistant sequences from CARD, BLAST was used to filter out any duplicates. The data was then labeled as follows: **1** for antibiotic-resistant genes and **0** for antibiotic non-resistant and synthetic non-resistant genes. The final dataset comprises 20,000 sequences, which were divided into training, testing, and validation datasets.

Methodology

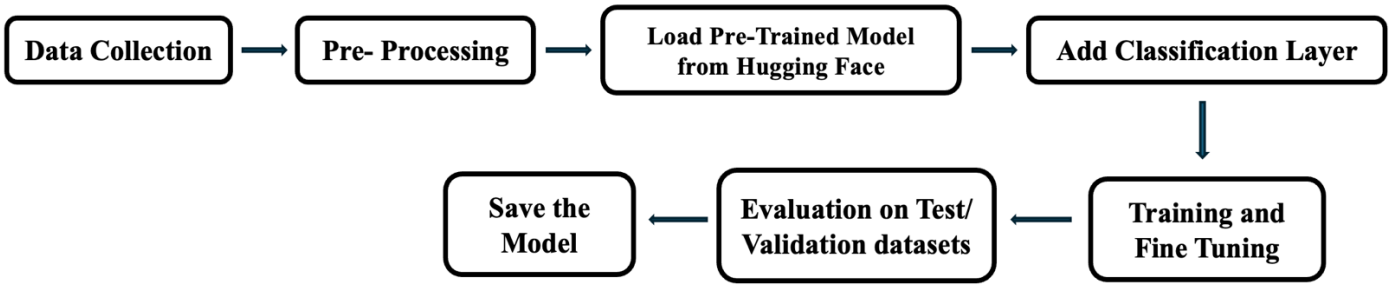

Fig-3 Workflow

The workflow for fine-tuning the Nucleotide Transformer for ARG classification is outlined in the flowchart above. Initially, the dataset containing nucleotide sequences labelled as antibiotic-resistant or non-resistant is collected and split into training, validation, and test sets to ensure balanced evaluation. The data undergoes preprocessing to remove any inconsistencies or class imbalances, and is then tokenized using the AutoTokenizer from the Hugging Face Transformers library, which converts the nucleotide sequences into token IDs suitable for input to the model.

Next, a pre-trained Nucleotide Transformer model is downloaded from Hugging Face and loaded locally. To adapt the model for the classification task, a new classification head is added on top of the pre-trained architecture. This modified model is then fine-tuned on the training data using the Hugging Face Trainer API, leveraging GPU acceleration on Apple Silicon (M4). After training, the model is evaluated using the validation and test datasets to measure performance metrics such as accuracy, F1-score, and loss. Finally, the trained model, along with the tokenizer and configuration files, is saved for future use or deployment.

Training Setup

We trained and evaluated a transformer-based model for sequence classification, specifically focusing on the 500M multi-species Nucleotide Transformer. Given the complexity and size of this model (approximately 500 million parameters), we aimed to optimize resource utilization and training efficiency while keeping the setup lightweight and accessible.The transformer model was sourced from Hugging Face, a widely used repository for state-of-the-art deep learning models. We selected the InstaDeepAI/nucleotide-transformer-500m model and downloaded it locally to ensure consistent access during training. Unlike our initial plan to utilize an HPC cluster with Slurm scheduling, this project was successfully fine-tuned locally on an Apple MacBook with an M4 chip, leveraging Metal Performance Shaders (MPS) via PyTorch for GPU acceleration. The training workflow was adapted for the Apple Silicon environment. The dataset was pre-processed and tokenized using Hugging Face Datasets with efficient batch processing. Padding and truncation were applied to standardize sequence lengths, and the model was fine-tuned using Hugging Face's Trainer API. Since the pre-trained model did not include a classification head, a new dense classification layer was initialized and trained on the downstream task. To accommodate memory limitations on the local device, training and evaluation batch sizes were set to 4. Mixed-precision training (fp16) was disabled due to limited support in MPS. The model was trained for 1 epoch to validate pipeline correctness and assess training performance. Logging, evaluation, and model checkpointing were all enabled to monitor training dynamics and save the best-performing checkpoint.

Results

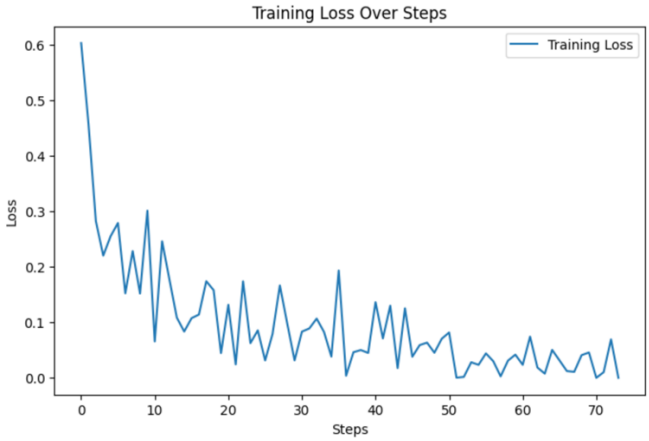

Fig-4 Loss Curves

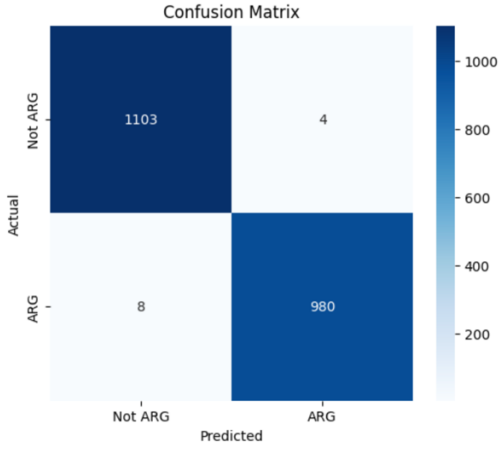

Fig-5 Confusion Matrix

During training, the training loss curve exhibited a consistent decline, indicating effective learning and convergence. Initially starting around 0.6, the loss dropped steadily to near zero, with minor fluctuations due to batch-wise variation. The confusion matrix demonstrates the model's capability to distinguish between ARG and non-ARG sequences with high precision. Out of 2,095 test samples, the model correctly classified 1,103 non-ARGs and 980 ARGs, with only 4 false positives and 8 false negatives. This highlights the model's robustness and low error rate in both classes.

	precision	recall	f1-score	support
Not ARG	0.99	1.00	0.99	1107
ARG	1.00	0.99	0.99	988
accuracy			0.99	2095
macro avg	0.99	0.99	0.99	2095
weighted avg	0.99	0.99	0.99	2095

Fig-6 Classification Report

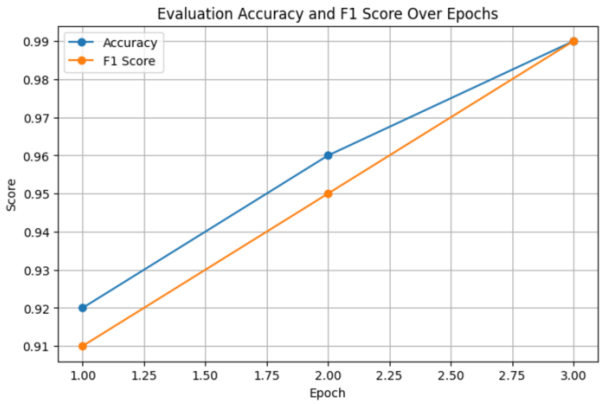

Fig-7 Accuracy, F1 Score

The classification report further reinforces this, showing a precision, recall, and F1-score of 0.99 for both ARG and non-ARG classes. The overall accuracy achieved was 99%, with both macro and weighted averages aligning closely, indicating balanced performance across class distributions. This pattern suggests that the model generalized well without signs of overfitting. The plot of evaluation accuracy and F1 score over epochs showed a consistent upward trend. Accuracy and F1 score increased from ~0.92 in the first epoch to 0.99 by the third epoch, indicating steady improvements with each training pass. Overall, these results demonstrate that the fine-tuned Nucleotide Transformer achieved exceptional classification performance on the ARG dataset. The combination of minimal loss, high precision, and stable metrics across epochs confirms the model's reliability and effectiveness for downstream biological sequence classification tasks.

Conclusion

Foundation models such as transformers are proving to be powerful tools for understanding the "language" of DNA, enabling significant advancements in a variety of biological sequence analysis tasks. These models, pre-trained on vast amounts of genomic data with billions of parameters, offer robust and generalizable representations that can be fine-tuned for specialized applications.

In this project, we demonstrated the effectiveness of using a pre-trained Nucleotide Transformer for the classification of antibiotic resistance genes (ARGs). The model achieved exceptional predictive performance, highlighting its potential in solving complex biological problems. With improved computational resources and larger datasets, these models can be further fine-tuned for a wide range of applications across genomics, transcriptomics, and other omics domains. As foundation models continue to evolve, they offer exciting opportunities for advancing biological research and precision medicine through scalable, high-performance sequence modelling.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
4. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics, 37*(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083
5. ENCODE Project Consortium. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature, 583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4
6. Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clinical Microbiology Reviews, 31*(4), e00088-17. https://doi.org/10.1128/CMR.00088-17
7. Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., & Pečerska, J. (2020). Machine learning for detecting antimicrobial resistance genes in bacterial genomes. *Frontiers in Microbiology, 11*, 550. https://doi.org/10.3389/fmicb.2020.00550
8. Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clinical Microbiology Reviews, 31*(4), e00088-17. https://doi.org/10.1128/CMR.00088-17
9. Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., & Pečerska, J. (2020). Machine learning for detecting antimicrobial resistance genes in bacterial genomes. *Frontiers in Microbiology, 11*, 550. https://doi.org/10.3389/fmicb.2020.00550