Transformer-Encoders

The Author

January 6, 2026

Contents

List of Figures

List of Tables

1 Abstract

10pages no ref

2 Introduction

Antimicrobial Resistance of pathogens (AMR) poses a critical and escalating threat to global public health. Recognized by the World Health Organization (WHO) as a major concern, AMR minimizes the effectiveness of antibiotics, leading to far more severe illnesses, increased mortality rate and higher healthcare needs (Hu et al. 2024). The currently used culture-based antimicrobial susceptibility testing (AST) is time-consuming, making the use of a broader spectrum antibiotic more attractive, favouring the emergence of resistent pathogens (Barbosa et al. 2000, apparently in)[**?**]. This reinforces the need for faster AMR prediction.
Whole-genome sequencing (WGS) offers an option for this, by providing a quick way to assess genetic information, but it's speed and accuracy is dependent on advanced computational tools.(Su18) This paper?seminar?report? explores the role of Encoder architectures in the prediction of AMR directly from genomic sequences. As part of the Transformers architectures, Encoders represent a paradigm shift since their introduction in 2017 (Mowbray25). In particular the self-attention characteristic of these Transformer models will be examined in the learning of biological patterns within DNA sequences. This capability is crucial for the identification of resistent mutations and thereby predicting AMR phenotypes. To achieve this, this report will start with an overview of machine learning models for sequence analysis, highlighting the advancements from traditional models to modern neural networks. Following this, the architectural principles of encoder only Transformer models will be showcased, explaining their key components. Subsequently the practical implementation using a pre-trained encoder model to predict the resistance of cefoxitin in Staphylococcus aureus strands, will be described. Finally, the results of various models will be presented with a discussion of encoder models for AMR prediction. and possible future works

3 neuralnetworks other models

3.1 Data Representation

To apply machine learning to the Straphylococcus aureus gene, biological sequences must be transformed to numerical representations. Traditional approaches work with the One-Hot Encoding approach, which maps nucleotides to binary vectors or k-mer frequency analysis, which counts the occurence of short subsequences (Angermüller). These methods however often result in high-dimensional, sparse data which fails to capture the relationships between nucleotides. Modern Deep Learning approaches address this by utilizing Learned Embeddings, where nucleotides are mapped to dense vectors in a continuous space, allwoing the model to learn mathematical similarities between biologically similar sequences (Zou et all 2019)

3.2 Deep Learning

Early computational approaches to AMR prediction relied on supervised algorithms like Support Vector machines (SVM) and Random Forests. These models usually require extensive feature engineering and prior domain knowledge regarding specific resistance markers (Yang 2020) Deep Learning introduced architectures capable of automatic feature extraction. Convolutional Neural Networks (CNNs) apply sliding filters to sequence data, but struggle to model dependencies between distant parts of sequences, due to limited receptive fields. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, on the other hand process data sequentially, making the theoretically suitable for analysis of genomic sequences. RNNs often fail to retain context over long sequences due to the vanishing gradient problem (Vaswani et al 2017) The limitations on these models, especially with long-range dependencies in DNA, necessitated the development of the Transformer architecture.

4 Transformers

heremaybesoimething

4.1 Architecture

The limitations of sequential processing in RNNs were addressed by the introductio of the Transformer architecture. (Vaswani 2017). For the task of AMR prediciton, the Encoder part of this architecture was utilized.
Unlike neural networks, as CNNs, which compare local motifs, or RNNs which compare the previous steps, the Encoder models use a self-attention mechanism, which allows the model to weigh the relevance of every nucleotide in the sequence against every other nucleotide simultaneously. This enables the capture long-range dependencies, which are crucial in genomics, since the resisstent mutation at one site may depend structurally on a distant region within the same genome. (Devlin 2018)(Thomas STructural Bases (noaccess)).
Because the encoder processes the entire sequence in parallel, there is no order to follow (Weiss2021). To resolve this, positional encodings are added to the input embeddings to retain information about the relative positions of nulceotides. The model is typically trained using Masked Language Modeling checkhere, were a percentage of the input sequence is hidden and the model learns to reconstruct the missing nucleotides based on the contest.

4.2 Biological application

Encoders facilitate Transfer Learning. In a biological context, labeled data is scarce. The provided data set contains around 800 training samples. A deep network trained from scracth in a data set of this size, would lead to overfitting of the model.

To overcome this, a pre-trained, and then fine-tuned approach was chosen. Models such as DNABERT (Ji2021) are pretrained on unlabeled bacterial genome databases, learning the fundamental properties of DNA. This pre-trained model is then fine-tuned to specific tasks, in this case the prediction of genomic sequences resistent to Cefoxitin.
There are models trained on whole amino acid sequences, called Protein Language Models (PLMs), like ProtBERT, the data provided was in form of raw nucleotide inputs, for which genomic models, such as the chosen DNABERT, are specifically optimized.

5 Work done

5.1 different architectures

5.2 measures

5.3 code inclusion

what was done

6 Results

6.1 various arch result comparisons

what we got including comparisons different dropout values freeze unfreeze base model train a few epochs
try to get good results
show difference between different models imbalance and statistics

7 Discussion

meaning of the results

8 related works

9 References

impact00 Teresa M. Barbosa, Stuart B. Levy Addison Wesley, Massachusetts, 2nd Edition, 1994.