Predicting Drug Resistance in Malaria Using Language Models

Aditya Arun adarun@ucsd.edu
Annie Pham anp013@ucsd.edu
Aryan Shah ars018@ucsd.edu
Jared Simpauco jsimpauco@ucsd.edu

Rob Knight rknight@health.ucsd.edu

Malaria remains one of the most pressing global health challenges, with over 75% of malaria-related deaths occurring in children under the age of five. A key contributor to the persistence and severity of the disease is the growing drug resistance of *Plasmodium falciparum*, the parasite responsible for the most lethal form of malaria. This resistance arises from continuous mutations in the parasite's genome, which diminish the effectiveness of existing treatments. In this study, we present a novel approach that leverages transformer-based language models, to predict potential mutations in *P. falciparum* DNA sequences. Using masked language modeling (MLM), our model learns contextual representations of nucleotide sequences by training on the reference genome of the 3D7 strain. We compare our custom-trained models against DNABERT, a model pretrained on human genomic data, to assess their ability to capture malaria-specific sequence dependencies. Results show that models trained specifically on malaria sequences outperform general DNA models, with prediction accuracy correlating with mutation-prone regions. This approach offers a scalable framework for detecting emerging resistant strains and guiding targeted drug development. Future work includes fine-tuning model architectures and integrating experimental validation to improve predictive power.

Code: https://github.com/jsimpauco/pdr-malaria

1 Introduction

Malaria continues to be a major global health concern, particularly in regions with limited access to healthcare. Each year, it infects hundreds of millions and leads to significant mortality, with nearly 75% of malaria-related deaths occurring in children under five (Medicines for Malaria Venture (2024)). One of the most pressing challenges in combating malaria is the parasite's evolving resistance to antimalarial drugs. *Plasmodium falciparum*, the most lethal malaria parasite, acquires this resistance through genetic mutations—making early detection of these mutations critical for effective treatment planning and public health response.

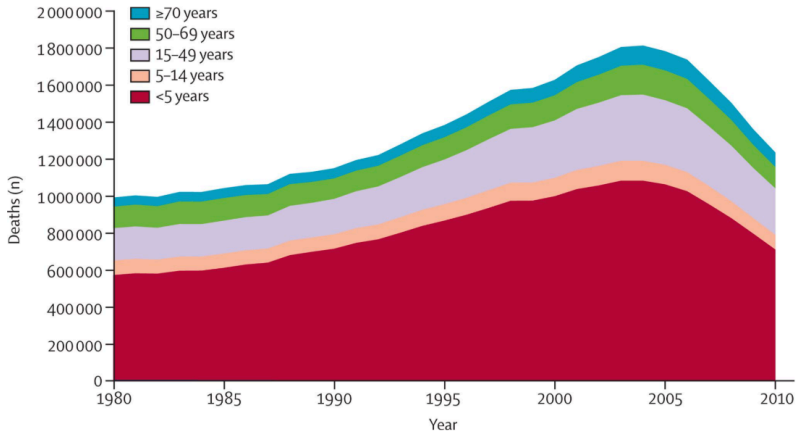

Figure 1: Malaria Cases (Murray et al. (2012))

Genomic surveillance plays a vital role in identifying emerging drug-resistant strains. However, traditional methods for mutation detection often involve time-intensive sequencing and experimental validation, which may not scale well in rapidly evolving public health scenarios. This has created a growing need for scalable computational tools capable of learning from raw genomic data to identify patterns and predict mutations Fidock et al. (2005).

Our project addresses this challenge by applying recent advances in Natural Language Processing (NLP), particularly transformer-based models like BERT, to the field of genomics. Genomic sequences, like natural language, contain context-dependent patterns, making them suitable for models originally developed for text data. Building on our work in the previous quarter, where we explored unsupervised learning for general DNA motif detection, we now shift toward a more targeted and impactful application: predicting mutations in the genome of *P. falciparum* to identify potential regions of drug resistance.

In this work, we preprocess the reference genome of the *P. falciparum* 3D7 strain by segmenting it into fixed-length k-mers and numerically encoding the sequences to make them compatible with transformer architectures. We train a custom BERT-based model using masked language modeling (MLM), a technique that allows the model to learn the underlying distribution of nucleotide sequences by predicting masked tokens. We also compare our model's performance to DNABERT—a pre-trained transformer model trained on human

genomic sequences—to assess the value of domain-specific training in malaria genomics.

Our model aims to capture subtle nucleotide patterns and variations that may signal the emergence of resistance-conferring mutations. By analyzing nucleotide composition, k-mer frequencies, and contextual predictions, we seek to better understand mutation-prone regions and how they diverge from the baseline genome.

The key contributions of our project are as follows:

- Development of a domain-specific BERT-based model trained on the *P. falciparum* 3D7 genome.
- A preprocessing pipeline that segments genomic sequences into k-mers suitable for language modeling.
- Comparative evaluation of our model against DNABERT to demonstrate the value of domain-specific training.
- Insights into the relationship between sequence variability and drug resistance, supporting future use in targeted genomic surveillance.

By integrating transformer-based NLP techniques with infectious disease genomics, this project advances computational approaches for detecting drug resistance mutations in malaria. Our approach holds potential not only for accelerating drug resistance surveillance in malaria but also for broader applications across other pathogens with rapidly mutating genomes.

2 Methods

2.1 Dataset

The genomic data used in this study was provided by the Winzeler Lab at the University of California, San Diego, a research group in the field of malaria genomics. The core dataset consists of the reference genome of *Plasmodium falciparum* 3D7, obtained from PlasmoDB v13, and stored in FASTA format. This reference genome spans approximately **23 million base pairs** and includes **14 nuclear chromosomes**, as well as **mitochondrial** and **apicoplast** sequences, offering comprehensive coverage of the parasite's genetic structure.

To support mutation prediction and model training, the reference genome was used to generate structured inputs for a transformer-based language model. The genome was parsed chromosome by chromosome and segmented into fixed-length chunks of 512 base pairs, allowing the model to learn from localized genomic contexts while conforming to input length constraints typical of BERT-based models. Any malformed or redundant sequences were filtered out to ensure data quality and to prevent information leakage during training.

In addition to the reference genome, the Winzeler Lab also provided access to their resistome dataset, which includes approximately 1,500 whole-genome sequencing (WGS) samples of *P. falciparum* isolates experimentally evolved for drug resistance. Within this set, around 850 mutations (representing 300 unique alleles) have been manually validated or statistically inferred to be associated with resistance phenotypes. While we did not directly make use of the resistome dataset in model training, it plays a critical role in evaluating the

model's capacity to identify mutation-prone regions potentially linked to drug resistance.

	chromosome_id	sequence	length
0	Pf3D7_01_v3	[T, G, A, A, C, C, C, T, A, A, A, A, C, C, T, ...	640851
1	Pf3D7_02_v3	[A, A, C, C, C, T, A, A, A, C, C, C, T, A, A, ...	947102
2	Pf3D7_03_v3	[T, A, A, A, C, C, C, T, A, A, A, T, C, T, C, ...	1067971
3	Pf3D7_04_v3	[A, A, C, C, C, T, A, A, A, C, C, C, T, G, A, ...	1200490

Figure 2: Preview of Genomic Dataset

The processed sequences were randomly split into training (80%), validation (10%), and test (10%) sets to evaluate the model's performance on unseen data. This ensured that the model could generalize beyond specific subsequences within the genome and allowed for robust assessment of mutation prediction capabilities.

The processed dataset provides a structured input for training the BERT-based language model, enabling it to learn the contextual relationships between nucleotide sequences. Future iterations of the project may explore training on additional Plasmodium species or incorporating resistance-associated mutations and using classification methods to further refine predictive performance.

2.2 Model Architecture

To predict mutations in *Plasmodium falciparum* DNA sequences, we implemented a custom BERT-based transformer model tailored to handle genomic data. This model builds on the foundational structure of Bidirectional Encoder Representations from Transformers (BERT) and adapts it for DNA sequence modeling using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks.

In adapting the BERT architecture for genomic sequences, each input DNA sequence is tokenized into overlapping 3-mers, and mapped to a vocabulary of tokens: `A`, `T`, `G`, `C`, along with special tokens like `[PAD]`, `[MASK]`, `[CLS]`, and `[SEP]`. To represent these tokens numerically, the input is encoded using a combination of three embeddings:

- **Token Embeddings**: Maps each nucleotide or special token to a unique vector representation.
- **Positional Embeddings**: Adds sinusoidal encodings to preserve the positional context of nucleotides within the sequence, which is essential since transformers lack inherent sequence order.
- **Segment Embeddings**: Helps differentiate between paired sequence inputs in the NSP task.

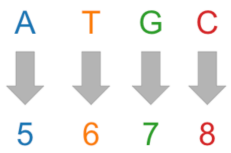

Figure 3: Tokenization

The final embedding for each token is the sum of these three components, followed by dropout regularization.

The model consists of a stack of 12 encoder layers, each composed of the following components:

- **Multi-Head Self-Attention**: Computes contextual relationships between all positions in the sequence simultaneously, allowing the model to capture both local and long-range dependencies.
- **Feed-Forward Neural Networks**: Applies a non-linear transformation across the attention outputs using GELU activations.
- **Layer Normalization and Residual Connections**: Enhances model stability and convergence.

Each encoder layer processes a sequence of length 512 with a hidden dimension of 768. The model operates in a bidirectional manner, enabling it to utilize both upstream and downstream nucleotide context to make predictions.

Two heads are attached to the final transformer output:

- **Masked Language Modeling (MLM)** Head: Predicts the original nucleotide at positions randomly masked during training. This is a multi-class classification task over the nucleotide vocabulary.
- **Next Sentence Prediction (NSP)** Head: Predicts whether two input chunks are adjacent in the original sequence. This task aids the model in understanding inter-segment relationships.

We trained our model on the *P. falciparum* 3D7 reference genome, using 512-base-long chunks of tokenized DNA. During training, **15%** of the tokens in each sequence are randomly selected for masking:

- 80% are replaced with the `[MASK]` token,
- 10% are replaced with a random nucleotide,
- 10% remain unchanged.

The model was trained using Negative Log Likelihood Loss (NLLLoss) applied to both the MLM and NSP outputs. However, for simplicity and improved performance, we focused primarily on optimizing MLM loss during training.

We utilized the Adam optimizer with a warm-up learning rate schedule and weight decay for regularization. Models were trained for up to 50 epochs, and intermediate checkpoints were saved at various intervals (e.g., largeMalariaModelEpoch1, smallMalariaModelEpoch10).

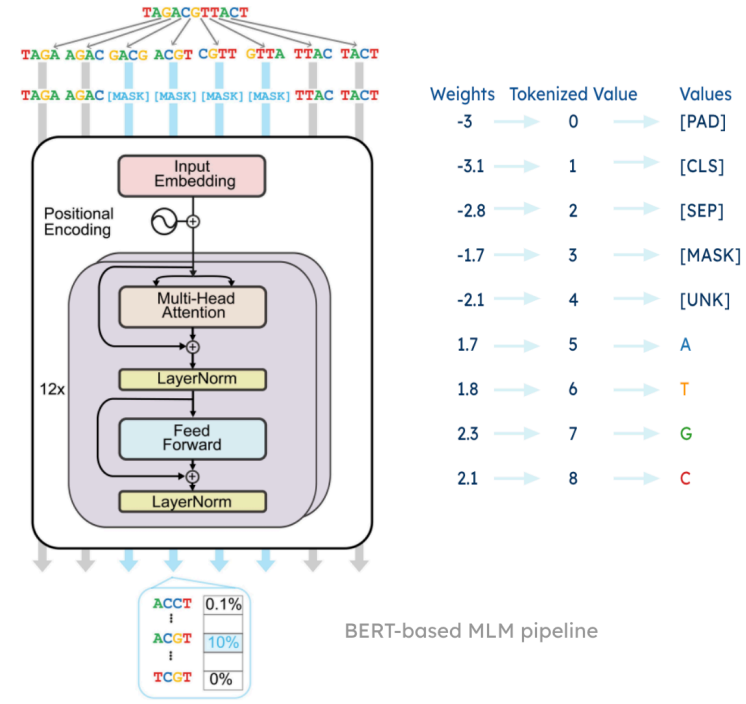

Figure 4: Model Architecture

3 Results

3.1 Evaluation Strategy

To assess the effectiveness of our BERT-based model in learning nucleotide patterns and predicting mutations, we employed a masked language modeling (MLM) evaluation strategy. Specifically, we measured the token-wise prediction accuracy on masked positions within sequences from a held-out validation set. This accuracy represents the proportion of correctly predicted nucleotides at masked positions, providing a direct indicator of the model's ability to capture meaningful genomic patterns.

For benchmarking purposes, we trained and evaluated two primary model configurations:

- **Large models**, trained on approximately 36,000 sequence pairs (80% of the full reference genome), for fewer epochs due to computational constraints.
- **Small models**, trained on only 2,000 sequence pairs, but for extended epochs (10–50), to evaluate the impact of training depth on a limited dataset.

We compared model outputs by calculating average accuracy across the validation set and also examined minimum and maximum accuracy per sequence to identify performance variability. Additionally, we visualized the model's probabilistic predictions across nucleotide positions to better interpret its behavior and identify regions of high or low confidence.

3.2 Model Performance and Observations

Our results highlight notable differences between the large and small model configurations:

- The **small model** trained for 10 epochs achieved an impressive **average accuracy of 99.8%**, with consistent predictions across all positions in the test sequences. This indicates the model learned strong positional dependencies and converged effectively on a small dataset when given sufficient training iterations.
- The **large model** trained for only 1–2 epochs on the full dataset reached a **maximum accuracy of 40.3%**, with significantly more variance in predictions. These results suggest that while the large model had access to more data, it was undertrained, reinforcing the importance of sufficient epoch depth in transformer-based models.

These results are summarized in Figure 5:

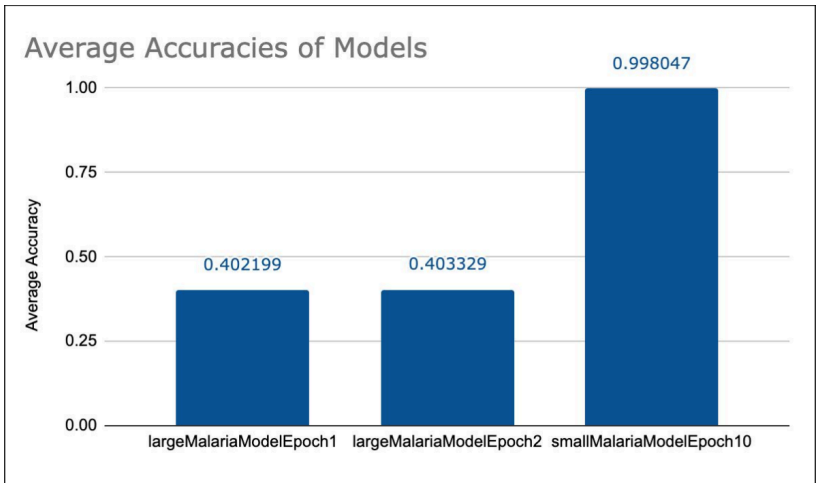

Figure 5: Model Accuracy

3.3 Sequence Prediction and Interpretability

In addition to quantitative metrics, we explored the interpretability of model predictions. Since our model outputs a probability distribution over A/T/G/C at each sequence position, we can visualize which nucleotides the model considers most likely.

In one example, we extracted prediction probabilities for positions 20–29 of a masked sequence and selected the nucleotide with the maximum probability at each position. This yielded the predicted sequence: ”CCCCCCAAAC”, as shown in Figure 5.

A stacked bar plot of the predicted nucleotide probabilities illustrates how confident the model was in each prediction. For instance, early positions showed high confidence in C, while positions toward the end showed more balanced distributions, indicating model uncertainty or natural genomic variability.

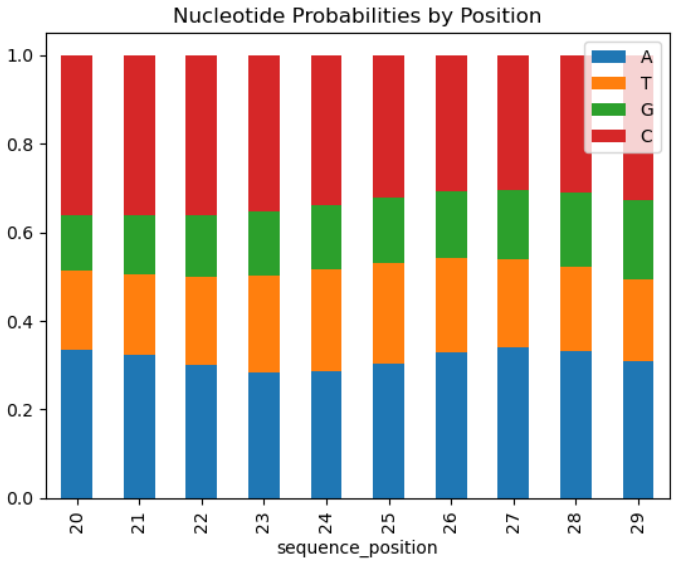

Figure 6: Sequence Prediction

4 Conclusion/Discussion

Drug resistance in malaria, specifically in the species Plasmodium falciparum which we attempted to analyze through this experiment, has exploded in recent decades leading to a resurgence of malaria and millions of deaths, especially in children. This can largely be attributed to not only an overuse of medication, but more importantly for this study, genetic mutation.

By attempting to create a model that can learn what, at a baseline, malaria should look like, we hope that we can pinpoint if new specimens of Plasmodium falciparum are drug resistant. This is done by seeing how accurately the model can guess what the new sequence should look like, having been trained off of the current idea of Plasmodium falciparum, with low accuracy predictions possibly corresponding to DNA mutations that cause drug resistance. Additionally, since the model makes a prediction at each nucleotide position, the model may even be able to pinpoint where the DNA sequence has mutated.

In the future, we hope that this can mean that we are not only able to detect mutations in malaria sooner, thereby starting new drug and medication research earlier, we can also get a better idea of how medication needs to change through locating where in a sequence of DNA has low predictability, thereby locating where the DNA has likely changed and what about the DNA sequence may be causing higher drug resistance.

For example, Patrick G Bray et al. in “Defining the role of PfCRT in Plasmodium falciparum chloroquine resistance” discusses how a specific protein leads to resistance of chloroquine, a common drug that was used in antimalarial medication. It is our hope that our model would be able to track changes in this protein and other proteins present, to get a better understanding of how and why drug resistance may be rising in malaria and how researchers may

need to change current antimalarial medication to better target this new form of malaria.

Additionally, although we trained our model off of Malaria data, this model can be trained off of any type of data which means it may be able to fine-tune to other types of diseases. In the end, by calculating accuracy and by extension the predictability of new DNA sequences we hope that our model will be able to decrease the amount of time it takes to create new effective medication for malaria, thereby decreasing the amount of deaths caused by malaria.

5 Limitations & Future Work

While our model shows promise in identifying drug resistance-related mutations, there are still a few challenges to address.

One key limitation is that our model is trained primarily on the P. falciparum 3D7 strain, which means it might not work as well on malaria strains from different regions. To improve generalization, future versions should incorporate multi-strain datasets, making the model more adaptable to real-world genetic diversity. Another challenge is that we're using sequence predictability as a stand-in for mutation detection. While our results suggest that lower confidence predictions align with drug resistance mutations, this doesn't necessarily prove causation. To confirm whether these mutations actually cause resistance, we would need additional experimental validation, such as lab tests or protein structure analysis.

Computational demands also played a role—training on full-genome datasets is resource-intensive, even with a BERT-based transformer architecture. Future work could explore lighter models, such as using model distillation or combining transformers with CNNs, to balance efficiency and accuracy.

Beyond malaria, this model could be fine-tuned for other pathogens, including viruses and bacteria, expanding its role in infectious disease research. Additionally, incorporating biological and epidemiological data, like patient treatment outcomes or geographic resistance patterns, could improve the model's ability to track and predict drug-resistant strains in real-world settings. By tackling these limitations, we hope to refine our approach, improve global malaria surveillance, and ultimately help in the fight against drug-resistant infections.

References

Fidock, David A. et al. 2005. "Defining the role of PfCRT in Plasmodium falciparum chloroquine resistance." *Molecular Microbiology* 56 (4): 990–1001. [Link]

Medicines for Malaria Venture. 2024. "Malaria Facts & Statistics | Medicines for Malaria Venture." [Link]

Murray, Christopher J., Lisa C. Rosenfeld, Stephen S. Lim, Kathryn G. Andrews, Kyle J. Foreman, Diana Haring, Nancy Fullman, Mohsen Naghavi, Rafael Lozano, and

Alan D. Lopez. 2012. “Global malaria mortality between 1980 and 2010: A systematic analysis.” *The Lancet* 379 (9814): 413–431. [Link]